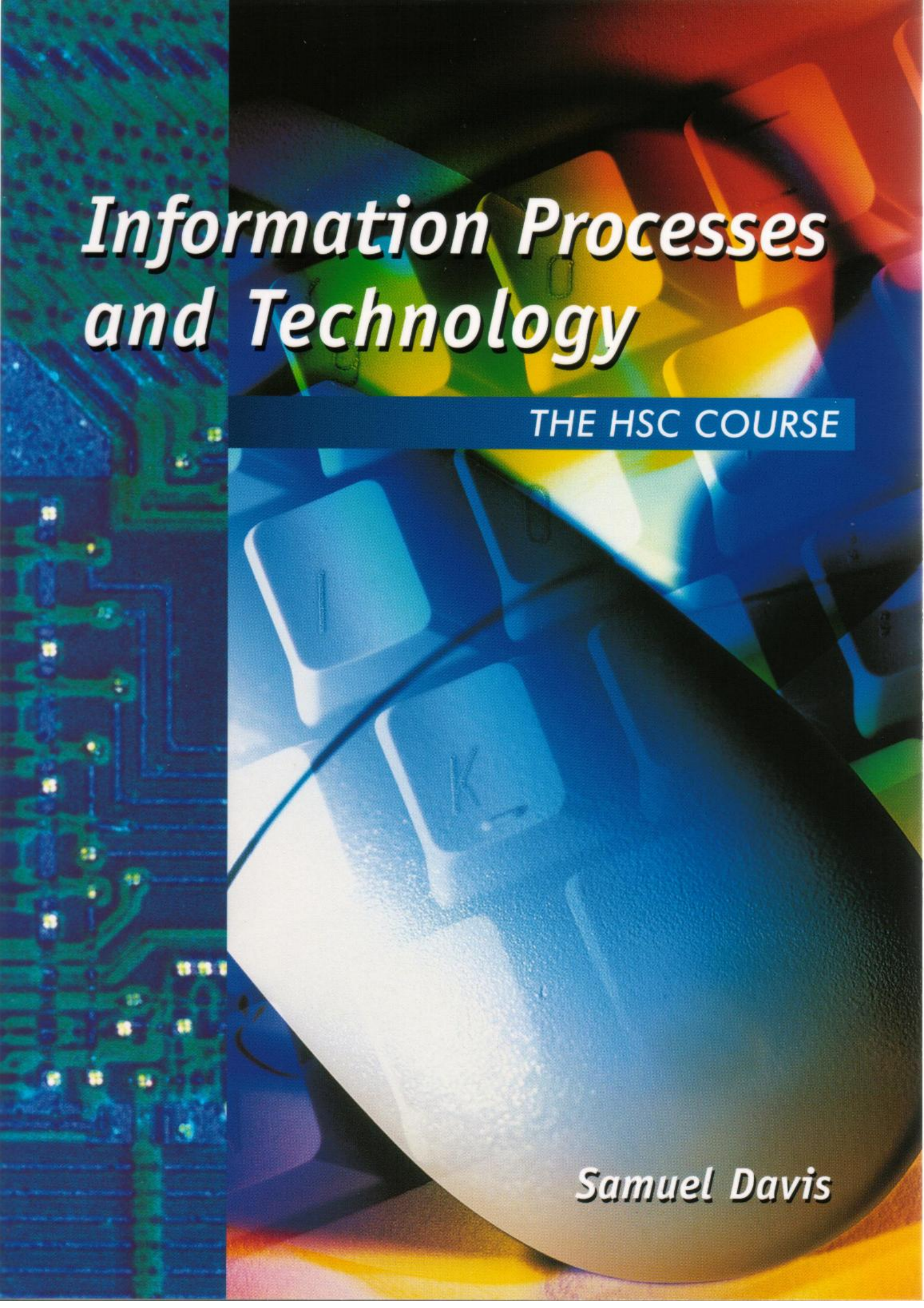

Information Processes and Technology
THE HSC COURSE
Samuel Davis

First published 2007 by
Parramatta Education Centre
PO Box 26, Douglas Park NSW 2569
Tel: (02) 4632 7987 Fax: (02) 4632 8002

Visit our website at www.pedc.com.au

Copyright © Samuel Davis 2007

All rights reserved.

Copying for educational purposes
The Australian Copyright Act 1968 (the Act) allows a maximum of one chapter or 10% of this book, whichever is the greater, to be copied by any educational institution for its educational purposes provided that that educational institution (or the body that administers it) has given a remuneration notice to the Copyright Agency Limited (CAL) under the Act.

Copying for other purposes
Except under the conditions described in the Australian Copyright Act 1968 (the Act) and subsequent amendments, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the copyright owner.

National Library of Australia
Cataloguing in publication data

Davis, Samuel, 1964-.
Information processes and technology: the HSC course.

Includes index.
ISBN 9780957891036 (pbk.).

1. Information technology - Textbooks. 2. Information storage and retrieval systems - Textbooks. 3. Electronic data processing - Textbooks. I. Title.

004

Cover design: Great Minds
Printed in Australia by Southwood Press

CONTENTS

CONTENTS iii
DETAILED CONTENTS v
ACKNOWLEDGEMENTS xiii
TO THE TEACHER xiv
TO THE STUDENT xiv

1. PROJECT MANAGEMENT 3
Techniques for managing a project 4
Introduction to system development 21
Understanding the problem 26
Planning 46
Designing 64
Implementing 84
Testing, evaluating and maintaining 90

2. INFORMATION SYSTEMS AND DATABASES 107
Examples of database information systems 108
Organisation methods 119
Storage and retrieval 162
Collecting and displaying for database systems 203
Issues related to information systems and databases 215

3. COMMUNICATION SYSTEMS 229
Characteristics of communication systems 231
Examples of communication systems 260
Network communication concepts 305
Network hardware 325
Network software 349
Issues related to communication systems 355

OPTION STRANDS

4. TRANSACTION PROCESSING SYSTEMS 365
Characteristics of transaction processing systems 366
Real time (on-line) transaction processing 381
Batch transaction processing systems 400
Backup and recovery 414
Collecting in transaction processing systems 425
Analysing data output from transaction processing systems 435
Issues related to transaction processing systems 441

5. DECISION SUPPORT SYSTEMS **449**
Characteristics of decision support systems ... 451
Examples of decision support systems ... 452
Tools that support decision making ... 465
Spreadsheets ... 479
Analysing using spreadsheets ... 497
Expert Systems ... 506
Artificial neural networks ... 527
Issues related to decision support systems ... 538

6. MULTIMEDIA SYSTEMS **547**
Characteristics of each of the media types ... 548
Hardware for creating and displaying multimedia ... 565
Software for creating and displaying multimedia ... 583
Examples of multimedia systems ... 601
Expertise required during the development of multimedia systems ... 610
Other information processes when designing multimedia systems ... 615
Issues related to multimedia systems ... 638

GLOSSARY 643

INDEX 655

DETAILED CONTENTS

CONTENTS iii
DETAILED CONTENTS v
ACKNOWLEDGEMENTS xiii
TO THE TEACHER xiv
TO THE STUDENT xiv

1. PROJECT MANAGEMENT 3
Techniques for managing a project 4
Communication skills 5
Active listening 5
Conflict resolution 7
Negotiation skills 9
Interview techniques 10
Team building 11
Project management tools 14
Gantt charts for scheduling of tasks 15
Journals and diaries 16
Funding management plan 16
Communication management plan 17
Social and ethical issues related to project management 18
Set 1A 20
Introduction to system development 21
Understanding the problem 26
Interview/survey users of the existing system 27
Interview/survey participants in the existing system 30
Requirements prototypes 33
Define the requirements for a new system 35
How requirements reports are used during the SDLC 36
The content of a typical requirements report when using the traditional approach 37
Set 1B 45
Planning 46
Feasibility study 46
Technical feasibility 47
Economic feasibility 47
Schedule feasibility 49
Operational feasibility 50
Choosing a system development approach 53
Traditional 53
Outsourcing 54
Prototyping 55
Customisation 56
Participant development 57
Agile methods 58
Determine how the project will be managed and update the requirements report 59
Set 1C 63
Designing 64
System design tools for understanding, explaining and documenting the operation of the system 65
Context diagrams 65
Data dictionaries 66
Data flow diagrams 68
Decision trees and decision tables 71
Storyboards 73

Designing the information technology 75
Building/creating the system 79
Refining existing prototypes 79
Guided processes in application packages 80
Set 1D 82
Implementing ... 84
Implementation plan 84
Methods of conversion 85
Direct conversion 85
Parallel conversion 86
Phased conversion 86
Pilot conversion 86
Implementing training for users and participants 87
Testing, evaluating and maintaining ... 90
Testing to ensure the system meets requirements 90
Volume data 91
Simulated data 91
Live data 92
Trialling and using the operation manual 93
Ongoing evaluation to monitor performance 95
Ongoing evaluation to review the effect on users, participants and people within the environment 96
Maintaining the system to ensure it continues to meet requirements 98
Modifying parts of the system where problems are identified 99
Set 1E 102
Chapter 1 review 103

2. INFORMATION SYSTEMS AND DATABASES 107
Examples of database information systems ... 108
School timetable system 108
The roads and traffic authority holding information on vehicles and driver's licences 113
Video stores holding information on borrowers and videos 116
Set 2A 118
Organisation methods ... 119
Organisation of flat file databases 120
Choosing appropriate field data types 121
Non-computer examples of flat files 125
Set 2B 126
Relational databases 127
The logical organisation of relational databases 128
Tables 128
Primary keys 129
Relationships 131
Referential integrity 134
Set 2C 138
Normalising databases 139
First normal form (1NF) 140
Second normal form (2NF) 141
Third normal form (3NF) 144
Set 2D 149
Hypertext/hypermedia 150
The logical organisation of hypertext/hypermedia 151
Storyboards 151
Hypertext markup language (HTML) 154
Meta tag 156
Anchor tags 156
Uniform resource locators (URLs) 157
Set 2E 161

Storage and retrieval 162
Storage hardware 164
Direct and sequential access 164
On-line and off-line storage 165
Magnetic storage 165
Optical storage 169
Securing data 170
Backup and recovery 170
Physical security measures 171
Usernames and passwords 172
Encryption and decryption 172
Restricting access using DBMS views (user views) 174
Record locks in DBMSs 175
Set 2F **177**
Overview of searching, selecting and sorting 178
Tools for database searching and retrieval 179
Searching and sorting single tables (including flat files) 179
Query by example (QBE) 183
Searching and sorting multiple tables 184
Set 2G **191**
Centralised and distributed databases 192
Types of distributed databases 193
Tools for hypermedia searching and retrieval 197
Operation of search engines 198
Set 2H **202**
Collecting and displaying for database systems 203
Screen and report design principles 204
Consistency of design 204
Grouping of information 205
Use of white space 207
Judicious use of colour and graphics 208
Legibility of text 208
Data validation 210
Effective prompts 211
Set 2I **214**
Issues related to information systems and databases 215
Acknowledgement of data sources 215
Access, ownership and control of data 216
Freedom of information (FOI) acts 217
Privacy principles 218
Accuracy and reliability of data 219
Current and emerging trends 222
Data warehouses 222
Data mining 223
Online analytical processing (OLAP) 224
Online transaction processing (OLTP) 224
Set 2J **225**
Chapter 2 review **226**

3. COMMUNICATION SYSTEMS **229**
Characteristics of communication systems 231
Overview of protocol levels 231
IPT presentation level 231
IPT communication control and addressing level 231
IPT transmission level 232
Overview of how messages are passed between source and destination 232

Protocols 237
Hypertext transfer protocol (HTTP) 238
Transmission control protocol (TCP) 239
Internet protocol (IP) 241
Ethernet 243
Set 3A 245
Measurements of speed 246
Error checking methods 249
Parity bit check 249
Checksums 251
Cyclic redundancy check (CRC) 253
Hamming distances and error correction (extension) 256
Set 3B 259
Examples of communication systems 260
Internet 260
Public switched telephone network 260
Intranet and extranet 261
Teleconferencing 261
Business meeting system, sharing audio over the PSTN 262
Distance education system, sharing audio, video and other data using both PSTN and the Internet 266
Set 3C 274
Messaging systems 275
Traditional phone and fax 275
Voice mail and phone information services 276
Voice over Internet protocol (VoIP) 282
Electronic mail 284
- Email contents component 285
- Transmitting and receiving email messages 289
Set 3D 293
Electronic commerce 294
Automatic teller machine 294
Electronic funds transfer at point of sale (EFTPOS) 296
Internet banking 298
Trading over the Internet 301
Set 3E 304
Network communication concepts 305
Client-server architecture 305
Network topologies 307
Physical topologies 307
Logical topologies 311
- Logical bus topology 311
- Logical ring topology 314
- Logical star topology 316
Set 3F 319
Encoding and decoding analog and digital signals 320
Analog data to analog signal 320
Digital data to digital signal 321
Digital data to analog signal 323
Analog data to digital signal 324
Network hardware 325
Transmission media 325
Wired transmission media 326
Wireless transmission media 330
Set 3G 338
Network connection devices 339
Servers 346
Network software 349
Network operating system (NOS) 349
Network administration tasks 349
Set 3H 354

Issues related to communication systems 355
Internet fraud 355
Power and control 356
Removal of physical boundaries 357
Interpersonal issues 357
Work and employment issues 358
Current and emerging trends in communication 359
Chapter 3 review **361**

OPTION STRANDS

4. TRANSACTION PROCESSING SYSTEMS _____ 365
Characteristics of transaction processing systems 366
Historical significance of transaction processing 366
Automation of manual transaction processing 368
Components of transaction processing systems 371
Data integrity 375
Data validation 376
Data verification 376
Referential integrity 377
ACID properties 377
Set 4A **380**
Real time (on-line) transaction processing 381
Reservation systems 382
Point of sale (POS) systems 387
Library loans systems 392
Set 4B **399**
Batch transaction processing systems 400
Cheque clearance 402
Bill generation 404
Credit card transactions (real time or batch?) 406
Set 4C **413**
Backup and recovery 414
Full and partial backups 415
Transaction logs, mirroring and rollback 416
Backup media 417
Backup procedures 419
Set 4D **424**
Collecting in transaction processing systems 425
Collection hardware 425
Collection from forms 429
Analysing data output from transaction processing systems 435
Data warehouse 435
Management information systems 436
Decision support systems 437
Enterprise systems 439
Set 4E **440**
Issues related to transaction processing systems 441
The changing nature of work 441
The need for alternative non-computer procedures 441
Bias in data collection 442
Data security issues 443
Data integrity issues 443

Data quality issues 443
Control and its implications for participants 444
Chapter 4 review **445**

5. DECISION SUPPORT SYSTEMS 449
Characteristics of decision support systems 451
Examples of decision support systems 452
Semi-structured situations 452
Approving bank loans 452
Fingerprint matching 455
Unstructured situations 457
Predicting stock (share) prices 457
Disaster relief management 459
Set 5A **464**
Tools that support decision making 465
Spreadsheets 466
Expert systems 466
Artificial neural networks 467
Databases 467
Data warehouses and data marts 468
Data mining 469
Decision tree algorithms 470
Rule induction 470
Non-linear regression 471
K-nearest neighbour 471
Online analytical processing (OLAP) 472
Data visualisation 472
Drill downs 473
Online transaction processing (OLTP) 475
Group decision support systems (GDSS) 475
Intelligent agents 476
Geographic information systems (GIS) 477
Management information systems (MIS) 479
Spreadsheets 479
Identifying inputs and data sources 480
Developing formulas to be used 481
Planning the user interface 482
Extracting information from a database for analysis using a spreadsheet 483
Spreadsheet formulas 485
Linking multiple worksheets 486
Naming ranges 487
Absolute and relative references 487
Set 5B **490**
Charts and graphs 492
Spreadsheet macros 494
Spreadsheet templates 496
Analysing using spreadsheets 497
What-if analysis and scenarios 497
Goal seeking 498
Statistical analysis 501
Set 5C **505**

Expert Systems 506
Human experts and expert systems compared 506
Structure of expert systems 507
Knowledge base 508
Database of facts 513
Inference engine 513
- Backward chaining 514
- Forward chaining 516
Explanation mechanism 518
Developing expert systems (knowledge engineering) 519
Set 5D **526**
Artificial neural networks 527
Biological neurons and artificial neurons 527
Structure of artificial neural networks 529
How biological and artificial neural networks learn 532
Back propagation 533
Genetic algorithms 533
Set 5E **537**
Issues related to decision support systems 538
Reasons for intelligent decision support systems 538
Participants in decision support systems 540
Chapter 5 review **543**

6. MULTIMEDIA SYSTEMS **547**
Characteristics of each of the media types 548
Text and numbers 548
Hyperlinks 550
Audio 551
Images 554
Animation 557
Video 561
Set 6A **564**
Hardware for creating and displaying multimedia 565
Screens (or displays) 565
Video cards (display adapters) 565
CRT (cathode ray tube) based monitors 565
LCD (liquid crystal display) based monitors 566
Plasma screens 569
Touch screens 570
Digital projectors 571
Head-up display 574
Audio display 575
Sound card 575
Speakers 576
Head-sets 576
Optical storage 578
Set 6B **582**
Software for creating and displaying multimedia 583
Presentation software 583
Applications such as word processors with sound and video 585
Authoring software 587
Animation software 590
Web browsers and HTML editors 592
Set 6C **600**
Examples of multimedia systems 601

Education and training 601
Leisure and entertainment 603
Provision of information 606
Virtual reality and simulation 607
Expertise required during the development of multimedia systems 610
Set 6D **614**
Other information processes when designing multimedia systems 615
Organising presentations using storyboards 616
Collecting multimedia content 618
Flatbed scanner 618
Digital camera 620
Microphone and sound card 622
Video camera 623
Analog to digital conversion 625
Storing and retrieving multimedia content 626
Bitmap image file formats 626
Vector image file formats 628
Audio file formats 628
Video and animation file formats 630
Processing to integrate multimedia content 632
Set 6E **637**
Issues related to multimedia systems 638
Copyright issues 638
Integrity of source data 639
Current and emerging trends in multimedia systems 640
Virtual worlds 641
Chapter 6 review **642**

GLOSSARY 643

INDEX 655

ACKNOWLEDGEMENTS

First a vote of thanks to my wife Janine for her valuable contribution and support during the writing process and in particular during the final editing and production phase. Janine's experience in the IT industry and her various professional contacts have greatly improved the relevance and accuracy of the content.

Thanks to all the many computer teachers who have made comments and suggestions, hopefully these have been included to your satisfaction. In particular, thanks to Stephanie Schwarz who reviewed much of the content. Stephanie's comments are always accurate, pertinent and insightful.

My children, Luke, Kim, Melissa and Louise, together with my wife Janine have all made sacrifices so I can disappear to research and write. At time it seemed this text would never be completed. Thanks for your patience – at last I'm back!

Thanks also to the many companies and individuals who willingly assisted with the provision of screen shots and other copyrighted material. Every effort has been made to contact and trace the original source of copyright material in this book. I would be pleased to hear from copyright holders to rectify any errors or omissions.

Samuel Davis

TO THE TEACHER

This text provides a thorough and detailed coverage of the revised NSW *Information Processes and Technology (*IPT) Higher School Certificate course syllabus first examined as part of the 2009 HSC. The revised syllabus adds new content and also clarifies the existing content within the original IPT syllabus. The IPT syllabus is written such that it is suitable to a broad range of abilities. The better students will want to know the how and why – this text includes such detail.

Numerous group tasks and question sets are included throughout the text. These exercises aim to build on both the theoretical and practical aspects of the course. A teacher resource kit is available that provides further detail, including discussion points for all group tasks and full answers for all question sets. The teacher resource kit also includes many blackline masters and a CD-ROM containing a variety of other relevant resources.

Students often have difficulty determining the level of detail required in examination responses. To assist in this regard, a variety of HSC Style questions together with suggested solutions and comments are integrated within the text. Many of these questions are sourced from past Trial HSC examinations.

Every effort has been made to include the most up-to-date information in this text. However computer technologies are changing almost by the minute, which makes the writing task somewhat difficult. Technologies that are emerging today will be commonplace tomorrow.

TO THE STUDENT

Information systems are all around us; we use them routinely to meet our daily needs. The *Information Processes and Technology* HSC course focuses on the underlying processes and technologies within information systems. Throughout the course you will learn about information systems and how they are developed. IPT is not about learning to use software applications; rather it concerns the study of complete information systems, including hardware, software, processes and people. It's a course about systems that process data into information for people; information systems!

In the HSC course you must complete all three core topics – Project Management, Information Systems and Databases, and Communication Systems. In addition two of the option topics must be completed. In the final HSC examination sixty marks are allocated to the core topics and twenty marks to each of the two options you complete. To assist your preparation for the HSC examination numerous HSC Style questions and suggested solutions are included throughout the text. These questions are largely sourced from past Trial HSC examinations and provide an excellent guide to the detail required in HSC exam responses.

Best wishes with your *Information Processes and Technology* studies and the HSC in general.

In this chapter you will learn to:

- understand the communication skills required to manage a system development project, such as
  - active listening
  - conflict resolution
  - negotiation skills
  - interview techniques
  - team building
- understand the need to apply project management tools to develop a system using a team approach
- appreciate the advantages of groups that function as a team, including
  - increased productivity
  - enhanced job satisfaction
  - the development of a quality system
- appreciate the need for complete documentation throughout all aspects of the system
- assess the social and ethical implications of the solution throughout the project
- apply appropriate techniques in understanding the problem
- interpret a requirements report which includes:
  - the purpose of the systems
  - an analysis of an existing system
  - definition of extra requirements
- diagrammatically represent existing systems using context diagrams and data flow diagrams
- identify, communicate with and involve participants of the current system
- create a requirements prototype from applications packages that provide:
  - screen generators
  - report generators
- use a prototype to clarify participants' understanding of the problem
- conduct a feasibility study and report on the on the benefits, costs and risks of the project
- compare traditional, iterative and agile system development approaches
- create Gantt charts to show the implementation time frame
- investigate/research new information technologies that could form part of the system
- develop a solution to a problem from a prototype
- use a guided process in an application to create all or part of a solution
- use system design tools to:
  - better understand the system
  - assist in explaining the operation of the new system
  - document the new system
- determine training needs arising from the creation of a new system
- compare and contrast conversion methods
- justify the selected conversion method for a given situation
- convert from the old system to the new
- implement the appropriate information technology
- develop an implementation plan for the project
- compare the new system to the old and evaluate whether the requirements have been met
- update system documentation

Which will make you more able to:

- apply and explain an understanding of the nature and function of information technologies to a specific practical situation
- explain and justify the way in which information systems relate to information processes in a specific context
- analyse and describe a system in terms of the information processes involved
- develop solutions for an identified need which address all of the information processes
- evaluate and discuss the effect of information systems on the individual, society and the environment
- demonstrate and explain ethical practice in the use of information systems, technologies and processes
- propose and justify ways in which information systems will meet emerging needs
- justify the selection and use of appropriate resources and tools to effectively develop and manage projects
- assess the ethical implications of selecting and using specific resources and tools, recommends and justifies the choices
- analyse situations, identify needs, propose and then develop solutions
- select, justify and apply methodical approaches to planning, designing or implementing solutions
- implement effective management techniques
- use methods to thoroughly document the development of individual or team projects.

In this chapter you will learn about:

Techniques for managing a project

- communication skills necessary for dealing with others
- the consequences for groups that fail to function as a team, including:
  - financial loss
  - employment loss
  - missed opportunities
- project management tools including:
  - Gantt charts
  - scheduling of tasks
  - journal and diaries
  - funding management plan
  - communication management plan
- identifying social and ethical issues

Understanding the problem

- approaches to identify problems with existing systems, including
  - interview/survey users of the information system
  - interview/survey participants
  - analysing the existing system by determining how it works, what it does and who uses it
- requirements reports
- requirements prototype - a working model of an information system, built in order to understand the requirements of the system
  - used when the problem is not easily understood
  - repetitive process of prototype modification and participants' feedback until the problem is understood
  - can be the basis for further system development

Planning

- a feasibility study of proposed solutions, including:
  - economic feasibility
  - technical feasibility
  - operational feasibility
  - scheduling
- choosing the most appropriate solution
- choosing the appropriate development approaches
  - traditional
  - outsourcing
  - prototyping
  - customisation
  - participant development
  - agile methods
- the requirements report that:
  - details the time frame
  - details the subprojects and the time frame for them
  - identifies participants
  - identifies relevant information technology
  - identifies data/information
  - identifies the needs of users

Designing

- clarifying with users the benefits of the new information system
- designing the information system for ease of maintenance
- clarifying each of the relevant information processes within the system
- detailing the role of participants, the data and the information technology used in the system
- refining existing prototypes
- participant development, when people within the information system develop the solution
  - participant designed solutions
  - tools for participant development such as guided processes in application packages
- tools used in designing, including:
  - context diagrams
  - data flow diagrams
  - decision trees
  - decision tables
  - data dictionaries
  - storyboards

Implementing

- acquiring information technology and making it operational
  - hardware
  - software, customised or developed
- an implementation plan that details:
  - participant training
  - the method for conversion
    - parallel conversion
    - direct conversion
    - phased conversion
    - pilot conversion
  - how the system will be tested
  - conversion of data for the new system
- the need for an operation manual detailing procedures participants follow when using the new system

Testing, evaluating and maintaining

- testing and evaluating the solution with test data such as
  - volume data
  - simulated data
  - live data
- checking to see that the original system requirements have been achieved
- trialling and using the operation manual
- reviewing the effect on users of the information system, participants and people within the environment
- modifying parts of the system where problems are identified

1

PROJECT MANAGEMENT

Project management is a methodical and planned approach used to guide all the tasks and resources required to develop projects. It is an ongoing process that monitors and manages all aspects of a project's development. The overriding aim is to produce a high quality system that meets its objectives and requirements. In order to achieve this aim requires significant planning, including defining the systems requirements, setting and controlling the budget, scheduling and assigning tasks, and specifying the lines of communication between all stakeholders. To implement such project plans requires leadership skills with a particular emphasis on ongoing two-way communication between all parties, including the client, users, participants and members of the development team. It is a virtual certainty that problems will be encountered, hence maintaining an ongoing dialogue is critical if such problems are to be foreseen and their consequences avoided or at least minimised.

Project Management
A methodical, planned and ongoing process that guides all the development tasks and resources throughout a project's development.

GROUP TASK Discussion
Explain why project management should be an ongoing process that occurs throughout the whole system development lifecycle.

In many references project management is described using the 'project triangle', where time, money and scope form the three sides (see *Fig 1.1*). If any one side of the triangle is altered, the remaining two sides are affected. For example, if the time available for development is reduced then it is likely that costs will increase and the ability to achieve all requirements will decrease. Similarly the addition of extra requirements widens the project's scope and as a consequence both costs and time are likely to increase. Project management establishes and maintains a balance between money, time and scope in an effort to develop a system of the highest quality. To maintain this balance is an ongoing process throughout the system development lifecycle. Notice that in *Fig 1.1* quality is centred within the triangle – the implication being that the quality of the final system is affected by each of the three sides.

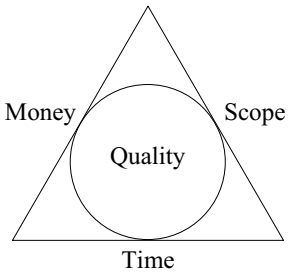


Fig 1.1
The Project Triangle

In this chapter we first examine techniques for managing projects. We then introduce system development and work through the stages of the system development lifecycle (SDLC), namely; understanding the problem, planning, design, implementing and finally testing, evaluating and maintaining the system. Clearly in this course we are concerned with the development of information systems, however many of the project management tasks and processes we shall examine are common to all types of projects and systems. For example the traditional structured approach to system

development mirrors the strategy used for most other engineering projects. However, information systems are significantly and fundamentally different to most other engineering projects and hence new and different methods of development are possible and appropriate. In the Preliminary course we focussed on the traditional approach to system development, in the HSC course we introduce other development approaches, such as outsourcing, prototyping, customisation, end-user and agile development. These approaches can be used in isolation or combined and integrated to suit the specific needs of each project.

Consider the following:

When designing and building a new bridge, the design stage is by necessity quite separate and consumes far less time and cost compared to the bridge's construction – typically design consumes just 10 to 15 percent of the total budget. The bridge design must be finalised in intricate detail prior to the construction stage commencing, once construction begins even minor design alterations will prove costly. Such projects are well suited to the traditional structured approach. In contrast the design of most information systems centres on the creation or customisation of software and the use of existing hardware components. The design stage for new information systems consumes the large majority of the budget and time. In fact in IPT we do not even consider construction or building as a separate stage. Rather we build our software components during the design stage and purchase and install the hardware during the implementation stage.

GROUP TASK Discussion
Based on the above discussion distinguish between the development of large construction projects and large information system projects.

GROUP TASK Discussion
Reflect on an information system you have developed. Did you use a strict structured approach much like the bridge project described above, or did the requirements change during design? Discuss using examples.

GROUP TASK Discussion
Realistically some requirements will be added or changed during the design phase of most projects. Should such additions and changes be encouraged or discouraged? Discuss.

TECHNIQUES FOR MANAGING A PROJECT

When developing large systems a specialist project manager or even a team of project managers will be appointed to perform project management tasks. All projects require project managers; for small projects a single individual may develop the system and also take on the role of project manager.

Successful project managers possess excellent communication and planning skills. They must motivate the development team, negotiate with all stakeholders, resolve conflict and at the same time ensure the project progresses within budget and time constraints. A variety of different project management tools are available to describe and document the various techniques that will be used to manage the project. In this section we consider relevant communication skills for project managers and then describe examples of common project management tools.

COMMUNICATION SKILLS

The project manager is a leader as well as a manager. There are many different leadership and communication styles and strategies; each individual must find a mix that suits their personality but also elicits the maximum performance from each team member. Most successful managers and leaders have a range of strategies at their disposal and they adjust their style in response to feedback – even during a single interview or meeting and often in response to non-verbal clues.

Despite differences in individual management styles there are various widely used and accepted communications strategies that should be considered and incorporated into all management styles. In this section we introduce some of these strategies. Furthermore the communication management plan (which is one of the project management tools we examine in the next section) should specify methods that support rather than hinder the use of these communication strategies. For instance, large lecture style meetings stifle feedback from participants while smaller round table sessions encourage feedback.

Active Listening

A significant portion of a project manager's time is spent listening to people. This is their main source of critical information required for a project to run smoothly. Listening is not the same as hearing; to listen well requires attention and involvement. In contrast hearing is an automatic, passive and often selective process. We notice some noises and sounds whilst ignoring others – we continually hear but without effort we don't comprehend or understand.

Many of us have developed techniques for "faking" listening. For instance we maintain eye contact, nod appropriately and even respond with "Oh yeah" and "I see", we try to give the impression we are listening when in fact we are barely hearing. Most of us can accurately detect such "fake listening" using non-verbal clues. If it occurs often then our view of the person diminishes and communication suffers – not something anyone wants and certainly a negative in terms of project managers.

Effective listening skills do not come naturally for most of us; we tend to focus on the message we wish to deliver rather than understanding messages we receive. Active listening is a strategy for improving listening skills – the aim being to better receive and understand the speaker's intended message and importantly for the speaker to know that the listener has received and understood their message. Each of these strategies requires the listener to verbally respond using words that directly relate to the speakers message. You must listen to the speaker to formulate such responses.

Active listening techniques include:

- **Mirroring**

Mirroring involves repeating back some of the speaker's key words. This technique indicates to the speaker that you are interested and would like to know and understand more. In addition the speaker hears the words they have just spoken, which allows them to reflect on the appropriateness and accuracy of their message. Consider the following brief exchange:

Speaker: *I doubt we'll be able to finish by Friday.*

Listener: *You don't think you'll be able to finish by Friday?*

The listener, presumably the project manager, has not made a judgement rather they have confirmed and encouraged further information. The speaker knows the message was received and in addition they have been encouraged to elaborate. Mirroring simply repeats back the speaker's words; it does little to confirm the message has been actually understood. Therefore mirroring should be used sparingly and in

conjunction with other active listening techniques. If overused it can appear repetitive and condescending – particularly when the listener holds a position of authority over the speaker.

- **Paraphrasing**

Paraphrasing is when the listener uses their own words to explain what they think the speaker has just said. In addition the listener reflects feelings as well as meaning within their response. Paraphrasing helps the speaker understand how their message sounds to others. The listener is communicating their desire to understand what the speaker feels about the content. This encourages the speaker to continue in an attempt to refine their message. Consider the following exchange:

Speaker: *There's a lot going on at the moment, I've got relatives staying so I really can't work any overtime, two of my team are out training on another job and well, finishing by Friday, I just can't see it happening.*

Listener: *You're feeling stressed as you can't see how to finish on time because two team members are out and you can't work late.*

The listener acknowledges the speakers feelings and reflects their words. It is important not to tell the speaker what they mean, for instance avoid phrases such as "What you mean is..." or "You're trying to say…". Rather the response should reflect what you honestly think the speaker feels in a way that allows them to correct or refine any inaccuracies.

- **Summarising**

Summarising responses are commonly used to refocus or direct the speaker to some important topic or to reach agreement so the conversation can end. A summary of an important point will cause the speaker to elaborate in more detail on that point. A complete summary confirms your understanding in the speaker's mind and hence helps to bring the conversation to an end. Typical summarising statements commence with:

Listener: *"If I understand correctly, your idea is…"*

Listener: *"So we agree that…"*

Listener: *"I believe you're saying…"*

- **Clarifying questions**

Often speakers will neglect or gloss over important details. This is natural as the speaker understands their points and can often assume the listener does also. The listener asks questions or makes statements that encourage the speaker to provide more detailed explanations.

Open-ended questions are used where a free and extended response is required rather than a simple answer. Examples include:

Listener: *"What do you think about…"*

Listener: *"Can you tell me more about…"*

Listener: *"I'm interested to understand your view on…"*

On the other hand, closed questions encourage single word or short answers – often either yes or no and should be used with caution. There are times when seeking a specific answer is necessary to provide detail. Try to limit such questions to factual information gathering or final confirmation of details rather than areas where opinions and feelings are involved. For instance asking, *"When will they return to work?"* requests factual information, while questions such as *"So you won't finish on time?"* or *"So you agree, don't you?"* are somewhat confronting and hence they may discourage further discussion.

- **Motivational responses**

The purpose is to encourage the speaker and reinforce in their mind that you are indeed listening and interested in what they have to say. One common technique is to use simple neutral words such as *"I understand", "Tell me more"* or *"That's interesting"* often combined with a nod of the head.

Another technique is to show that you relate to or have experienced what they are saying. In effect you place yourself in their situation in order to reinforce your acceptance of their words. This can involve some form of self-disclosure, where the listener briefly relates a similar experience. Such responses show you accept the speaker and are sympathetic or at least understanding of their situation. Possible example responses include:

Listener: *"I know what you mean, I felt like that when…"*

Listener: *"I too would be upset if…"*

Listener: *"That must make you feel great…"*

In each example the listener is seeing the situation from the speaker's point of view. This encourages the speaker to continue and also helps to establish and reinforce good relationships.

GROUP TASK Practical Activity
Split into pairs, one person being the speaker and the other the active listener. The speaker is to describe a hobby, sport or other interest whilst the listener uses active listening techniques.

Conflict Resolution

When groups or teams of people work together some amount of conflict is inevitable. This is not always a bad thing, indeed some amount of conflict is to be expected and can actually be beneficial. It is when conflicts become personal or remain unresolved that they cause problems. Team members, and in particular project managers, need to manage conflict so that issues are resolved appropriately for all concerned and in the best interests of the project.

Throughout the development of information systems decisions are constantly being made. Each decision involves a choice between different alternatives. Often different people will support different alternatives for a variety of different reasons. Understandably this is likely to cause conflict. Common areas where conflict occurs include:

- Allocating limited resources to development tasks. For example the total funds and time allocated to a project must be split equitably amongst each subtask. Increasing funding or time for one task often requires a corresponding reduction for other tasks. Conflict will arise as team members attempt to argue their case for a larger share of the limited resource.
- Different goals of team members. Individuals quite naturally formulate goals based on their interests, experience and area of expertise. For instance a graphic designer may rate the visual appeal of the user interface over functionality, whilst a software developer has little regard for visual appeal when it reduces functionality.
- Scheduling of tasks. During development many tasks must be performed in sequence. The ability to commence or complete one task relying on the completion of another task. It is often difficult to precisely specify in advance how long each task will take. As a result tasks later in the development process often suffer delays and can easily become the scapegoats for time overruns.

- Personal differences between people are a significant cause of conflict and can often be the most difficult to resolve effectively. Such differences include cultural, educational, religious, age and experience. The result being different feelings, attitudes and opinions.
- Internal conflict within individuals. People can have mixed feelings about how to perform their work or they can experience conflict between their personal and work commitments. Such internal conflict often results in high levels of stress, frustration and decreased productivity. Much like personal differences between people, internal conflict is often difficult to resolve.

To resolve conflict requires more than just a decision, it requires that the decision be accepted by each of the conflicting parties. This is not to say that all parties must feel they have won, in some conflict situations it may be appropriate for neither party to win or for one to win and the other lose. The overriding aim of conflict resolution is for all parties to participate, understand and then accept the final outcome.

Some strategies that assist when resolving conflict include:

- Attack the problem not the person. First try to define the problem and explore each person's perception of the problem. Try to understand people's point of view without judging them. Active listening techniques can be of assistance.
- Brainstorming where each person expresses ideas as they come to mind. No discussion takes place at this time. Often new and innovative solutions can emerge.
- Mediation involves a third party who is removed from the conflict acting as a sounding board for the conflicting persons. Such mediators are peacemakers, whose aim is to ensure opposing parties understand and appreciate the other's feelings and point of view. The conflicting parties express their thoughts and ideas through the mediator who is then able to steer the resolution process, ensuring it remains focussed on the problem and its resolution.
- Group problem solving requires a setting where all involved are on an equal footing and are encouraged to contribute equally. Commonly the group is arranged in a circle to promote equality. Each person expresses their point of view in turn whilst other group members listen without criticism. Often new and creative solutions will emerge. Even decisions that do not result in a "win" situation for all members are more easily accepted when all points of view are understood.

Consider the following situations:

- John has just been promoted to the position of project manager. He must now manage and lead a project team that includes many of his close friends with whom he once worked as an equal.
- To develop a new information system a large group is split into a series of teams, each led by a team leader. The team leaders meet with the project manager on a weekly basis. Some team leaders are highly experienced, others are young with limited experience and others are new to the company.
- A project manager just received cost and time estimates from each of his team members. He finds the total cost and time of all the estimates far exceeds the total budget and time allocated to the project.

GROUP TASK Discussion
Identify potential causes and areas of conflict in each of the above situations. Discuss suitable strategies for resolving such conflict.

Negotiation Skills

Negotiation is something we all do as part of our day-to-day lives. For instance negotiating who will cook dinner and who will wash up. We negotiate with others to reach a compromise situation that suits both parties. The parties communicate their needs and wishes whilst listening and understanding the others needs. Negotiation should be a friendly exchange where differences are argued logically and in a reasoned manner. Successful negotiation prevents situations escalating into conflict.

Many business negotiations occur in an environment where both parties already have a vested interest in reaching agreement. For example, negotiating the cost and terms for the purchase of goods or services. Both buyer and seller wish to reach agreement. The buyer needs the product or service and the seller needs to make a sale. The negotiation process is about agreeing on price and terms. In general, negotiations commence with both parties arguing for more than they ultimately expect – in our purchasing example the buyer starts at a low price and the seller at a high price. During negotiations the parties progressively alter their positions until agreement is reached. Skilled negotiators influence the negotiation process such that they achieve the best possible deal.

The skills and techniques discussed previously for conflict resolution are also valuable during negotiations. However there are recognised techniques used by most skilled negotiators, such techniques include:

- Knowing in advance all you can about the person, product, service and/or organisation prior to negotiations commencing can prove invaluable. When negotiating with outside organisations, research the worth or market value of the product or service they offer and assess other viable alternatives. Set limits in advance so that should the negotiations begin to break down you know in advance when to back off and reassess the situation.
- Consider a range of possible acceptable arrangements in advance. Try to think of options that will appeal to the other party or that they may well bring to the negotiation table. The aim is to anticipate the other party's position and prepare a reaction in advance. For instance perhaps a seller will not compromise sufficiently on purchase price alone, however they may offer low interest terms where payments are made over time or perhaps they will include extended warranties and guarantees. It is far better to assess such alternatives in advance rather than attempting making a quick decision in the heat of negotiations.
- Approach the other party directly to make an appointment in advance. At this time ensure the other party understands the agenda; this will ensure they are able to prepare sufficiently so that negotiation and agreement will be possible. Don't get drawn into detailed discussion at this time, try to leave your comments for the actual appointment. Remember the aim is to negotiate the best deal – don't give away detail that may allow the other party to pre-empt your position.
- During negotiations it is always easier to lower your expectations than it is to raise them. In general, start the negotiations at a point that exceeds your expected outcome. This improves your bargaining power as you have room to compromise during negotiation. Furthermore the other party will feel they have negotiated a better deal when they have lowered your initial expectations.
- Successful negotiators are confident and assertive, which allows them to maintain control during the negotiation process. This is where prior research and planning is critical. If you honestly know and understand the situation then being assertive is much easier. The points you make will be delivered more confidently and you will be able to formulate logical reasoned responses more effectively.

- Establish trust and credibility before negotiations commence. Negotiation is largely about persuading the other party to compromise their position in favour of your position. A climate where each party trusts the other and feels they are credible is a cooperative one that is more likely to encourage compromise. Furthermore it is rare for negotiations to be one off situations, more likely the parties will be negotiating agreements on a regular basis.

Consider the following negotiations:

- A company has used the same outside contractor to install electrical and LAN cabling for each information system they develop. Although happy with the quality of the contractor's work, they find that quotes from competing contractors are significantly less expensive.
- Diana is an experienced database professional who has been offered a new job by a larger competitor. The competitor is offering a much higher salary and the option of working from home. Diana would prefer to stay with her current employer if they can match the offer. Her current employer does not wish to lose her. However raising her salary would present problems as other employees on the same level as Diana would justifiably expect a similar raise.
- The contract for the development of an information system specifies financial penalties should the project extend beyond the stated completion date. The project manager, after discussion with members of the project team determines that it is unlikely they will finish on time. The project manager intends to arrange a meeting with his senior management in an attempt to negotiate a solution.

GROUP TASK Discussion
For each of the above situations, identify the issues and the parties involved. Discuss how each party could best prepare prior to negotiations commencing.

Interview Techniques

Interviews are used to identify problems with existing systems, obtain feedback during development and also to recruit and assess staff performance. We will consider interviews and surveys of a system's users later in this chapter as part of the Understanding the Problem stage of the system development lifecycle. In this section we concentrate on general interview techniques and in particular on techniques used when interviewing staff. Interviews with system users and participants have a different focus – they are used to collect and then summarise information about a systems operation. Staff interviews are generally used to gather information specific to the individual team member. Such interviews occur when recruiting new staff, assessing the performance of existing staff and also as part of disciplinary procedures.

Planing and preparation is the key to successful interviews. Questions should be formulated in advance and if a panel of interviewers is used then the questions should be shared out appropriately. One commonly used technique is to prepare pairs of questions. The first asks for specific information and often begins with words such as who, what, where, which or when. The second follow-up question is more open-ended and often asks how or why. For example, asking, "What was your last project?" followed by "How did you assist in achieving the project's goals?" The first question is relatively simple to answer and aims to focus and prepare the interviewee for the follow-up question.

When scheduling an interview the interviewee should be made aware of the purpose of the interview and they should also be given sufficient time to prepare. Interviews should be relaxed, professional and private – interruptions should be discouraged. When the interviewee arrives try to put them at ease; shake hands and perhaps engage in some informal chitchat. Commence by clearly stating the purpose of the interview and its likely duration. In a job interview a brief yet accurate description of the job and the company is worthwhile. An overview of the areas to be addressed in the interview may also be beneficial. Use a conversational tone throughout, however the interviewer should control the topics and direction of the interview. Many interviewees will be nervous or shy. The first few questions should be designed to be relatively easy for the interviewee to answer. Use active listening techniques and be prepared to adjust the speed of the interview to suit the interviewee.

There are many factors that influence the success of the interview process. Most of these factors revolve around how the interviewer conducts him or her self during the interview. Following are lists of positive and negative attributes worth considering when conducting interviews:

Positive interviewer attributes:

- Well-prepared questions.
- Attention and careful listening.
- Personal warmth and an engaging manner.
- The ability to sell ideas and communicate enthusiasm.
- Putting the interviewee at ease.
- Politeness and generosity.
- Focus on the topics that need to be covered.

Negative interviewer attributes:

- Lack of preparation.
- Not allowing enough time for the interview.
- Talking too much.
- Losing focus.
- Letting the interviewee direct the conversation.
- Biased towards people with similar ideas and styles to their own.
- The tendency to remember most positively the person last interviewed.

GROUP TASK Discussion
Recall an interview where you were the interviewee – perhaps a job interview or an interview with a teacher. Analyse the interviewer in terms of the above lists of positive and negative interviewer attributes.

Team Building

Team
Two or more people with complimentary skills, behaviours and personalities who are committed to achieving a common goal.

A team is more than a group of people. Successful teams are able to achieve more when working together than would be possible if each member operated alone – that is, "the whole is greater than the sum of the parts". Teams members focus on and are jointly responsible for achieving a shared goal. To build successful teams requires careful selection and ongoing training of people with different yet complimentary behaviour and personality traits. Clearly a team must include personnel with all the necessary skills to complete the work, however this should not be the sole selection criterion. In this section we first consider advantages of groups that function as a team and then consequences for groups that fail to function as a team, we then discuss popular techniques for building teams.

- **Advantages of groups that function as a team**

Groups that function as a team are more productive and the systems they develop are of higher quality. When team members co-operate they exchange ideas and formulate solutions together. The different skills, experiences, attitudes and behaviours of individuals complement each other rather than causing conflict. This joint sharing approach means more is achieved in less time. The team is more productive when working together than would have been the case if each member worked independently. Furthermore such collaboration results in higher quality systems – systems that exceed their requirements, have fewer bugs, are more tolerant of faults and are easier to maintain. No individual owns any single part of the system's design, rather each part is a joint effort that encompasses design ideas from the entire team.

There also advantages for the individual team members. There is less conflict within a collaborative team environment and responsibility for task completion is shared. This positive atmosphere increases job satisfaction. As job satisfaction increases then so too does productivity and pride in the quality of one's work. Increasing job satisfaction leads to higher productivity and quality, which in turn further improves job satisfaction – a positive cycle of improvement evolves.

- **Consequences for groups that fail to function as a team**

Groups that fail to function as teams can result in financial loss, employment loss and missed opportunities. Such groups are unable to reliably meet deadlines, produce quality work and operate within financial constraints. The group becomes a liability that lowers productivity and profit levels. If a company is unable to perform it cannot compete and hence it will have difficulty attracting clients, its profits will fall and staff will need to be retrenched.

Individuals also suffer when team performance is poor. Teams operate cooperatively such that each member learns and grows through their interactions with other team members. When real teamwork is not occurring each individual's skills will stagnate – a particular issue in the IT field where new technologies are constantly emerging. Furthermore the poor performance of a team reflects poorly on each of its members. Such issues reduce opportunities for promotion and advancement.

GROUP TASK Discussion
Sports teams composed of many international star players regularly get beaten by teams with no such star players. Discuss likely attributes of each of these teams that allow this to occur.

- **Team building skills and techniques**

To build strong and productive teams requires an understanding of how teams form and develop and also the composition of successful teams. We will briefly describe Tuckman's (1965) widely accepted stages of team development. Understanding these stages allows team leaders, such as project managers, to better understand and manage behaviour and performance. We then examine Belbin's Nine Team Roles. Belbin's powerful model is used extensively as the basis for building successful corporate work teams.

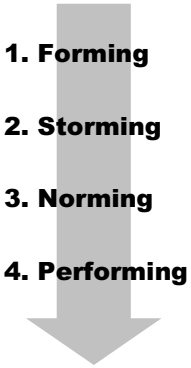


Fig 1.2
Tuckman's four stages of team development.

Tuckman describes four stages of team development, namely forming, storming, norming and performing. A brief description of each stage together with typical behaviours associated with each stage follows.

1. Forming. This is when team members are getting to know each other. Much like when you first started school, everyone is cautious and doesn't really know what to expect. People are trying to get to know each other and establish what role they and others will play. During the forming stage managers should help team members get to know each other, they should set the overall purpose and goals of the team and set expectations.
2. Storming. People are beginning to feel comfortable with each other. They now start to question issues and fight for position. Commonly this is the most difficult stage for a team to endure. Members will question procedures, disagree and even irritate each other as they jostle to establish their roles. Managers should ensure the team acknowledges this is quite normal, without ignoring conflicts that arise.
3. Norming. Team members now recognise their differences. Roles are fairly well established and settled and the team starts to work together. They consider how to adjust procedures and work flows to suit their particular way of operating. Personal differences have been resolved and emotions are more stable. Managers need to re-establish the teams goals, whilst accepting and responding to feedback.
4. Performing. The team is now operating as an effective productive unit. They are able to solve problems easily and even prevent problems arising in the first place. Team members are loyal and supportive of each other and they all share a common commitment to achieve the team's goals. Performing teams require little management; they largely regulate and manage themselves.

GROUP TASK Discussion
Reflect on the initial formation of your IPT class. Can you identify the forming, storming, norming and performing stages? Does your class currently operate as a "performing" team? Discuss.

The Belbin model is one popular technique used to build and develop productive management and work teams. The model has been extensively tested and is now used by many of the world's major corporations – including McDonalds, Nike, Nokia, Rolls Royce and Starbuck's Coffee. The main objective is to construct a team containing a balance of complimentary yet different behavioural and personality types. Research and experience indicates that such teams out perform those built based on skills alone. There are numerous training organisations across the world who specialise in the provision of team building courses based on the Belbin model. Belbin Associates also produces its own training material including e-Interplace, a software application for automating much of the analysis required to use the model.

The first step is to classify potential team members using Belbin's nine team role types. To do this each person completes a self-assessment questionnaire and also completes similar questionnaires with regard to other people with which they have worked in the past. The results are compiled and used as the basis for categorising each person according to Belbin's nine team role types (see *Fig 1.3* on the next page). Each role type describes a particular way of behaving, contributing and relating to others. Most people display characteristics of more than one team role and are able to select from these roles appropriately based on their current situation.

The e-Interplace software developed by Belbin Associates is able to produce a variety of reports that comment on individuals and also on the compatibility and detailed characteristics of different team combinations. In general a productive team should include members that include all nine team roles in roughly equal proportions.

Belbin Team-Role Descriptions

Team Role	Contribution	Allowable Weaknesses
Plant	Creative, imaginative, unorthodox. Solves difficult problems.	Ignores incidentals. Too preoccupied to communicate effectively.
Resource Investigator	Extrovert, enthusiastic, communicative. Explores opportunities. Develops contacts.	Over-optimistic. Loses interest once initial enthusiasm has passed.
Co-ordinator	Mature, confident, a good chairperson. Clarifies goals, promotes decision making, delegates well.	Can be seen as manipulative. Offloads personal work.
Shaper	Challenging, dynamic, thrives on pressure. The drive and courage to overcome obstacles.	Prone to provocation. Offends people's feelings.
Monitor Evaluator	Sober, strategic and discerning. Sees all options. Judges accurately.	Lacks drive and ability to inspire others.
Team Worker	Co-operative, mild, perceptive and diplomatic. Listens, builds, averts friction.	Indecisive in crunch situations.
Implementer	Disciplined, reliable, conservative and efficient. Turns ideas into practical actions.	Somewhat inflexible. Slow to respond to new possibilities.
Completer Finisher	Painstaking, conscientious, anxious. Searches out errors and omissions. Delivers on time.	Inclined to worry unduly. Reluctant to delegate.
Specialist	Single minded, self starting, dedicated. Provides knowledge and skills in rare supply.	Contributes on only a narrow front. Dwells on technicalities.

Reproduced with permission. Copyright © e-Interplace® Belbin Associates, UK. 1991-2006+

Fig 1.3
The Belbin model's nine team role types.

During training sessions various scenarios, often in the form of team games, are played out. Based on the reports from the e-Interplace software the trainers can deliberately choose an unbalanced team for some scenarios and a well-balanced team for others. Participants are therefore able to confirm the validity of the model before implementation in the work environment.

GROUP TASK Activity
Read through the nine team role descriptions in *Fig 1.3*. Note any team roles that you feel apply to you. Ask your friends if they agree.

GROUP TASK Research
Using the Internet, research training organisations that specialise in team building activities. List the team building techniques you discover.

PROJECT MANAGEMENT TOOLS

Project management tools are used to document and communicate:

- what each task is,
- who completes each task,
- when each task is to be completed,
- how much time is available to complete each task, and
- how much money is available to complete each task.

Without such documentation and planning, time and budget overruns are likely and furthermore the problems leading to such overruns are difficult to detect until it is too late. Lack of planning is a major reason for project failure; indeed poor planning can lead to projects being abandoned altogether. Project management documentation must recognise that virtually all projects encounter problems at some stage. As a consequence they cannot be static documents, they must adapt and change to reallocate tasks, resources, money and time in an effort to overcome problems.

Project managers use a variety of project management tools including:

- Gantt charts for scheduling of tasks.
- Journals and diaries for recording the completion of tasks and other details.
- Funding management plan for allocating money to tasks.
- Communication management plans to specify how all stakeholders will communicate with each other during the development of the new system.

Let us briefly consider each of these project management tools.

Gantt charts for scheduling of tasks

A Gantt chart is a horizontal bar chart used to graphically schedule and track individual tasks within a project. The horizontal axis represents the total time for the project and is broken down into appropriate time intervals – days, weeks or even months. The vertical axis represents each of the project tasks. Horizontal bars of varying lengths show the sequence, timing and length of each task. *Fig 1.4* below shows a Gantt chart produced with Microsoft Project.

Project checkpoints or milestones should be planned to signify the completion of significant tasks. Milestones are particular points in time; they have no duration and do not require work. Rather they are flags indicating that work should or has been completed. During development, reaching each milestone is an indication to management that the project is progressing as intended. Milestones are also times when the overall progress is assessed, which may result in changes to the schedule or various other aspects of the project's management.

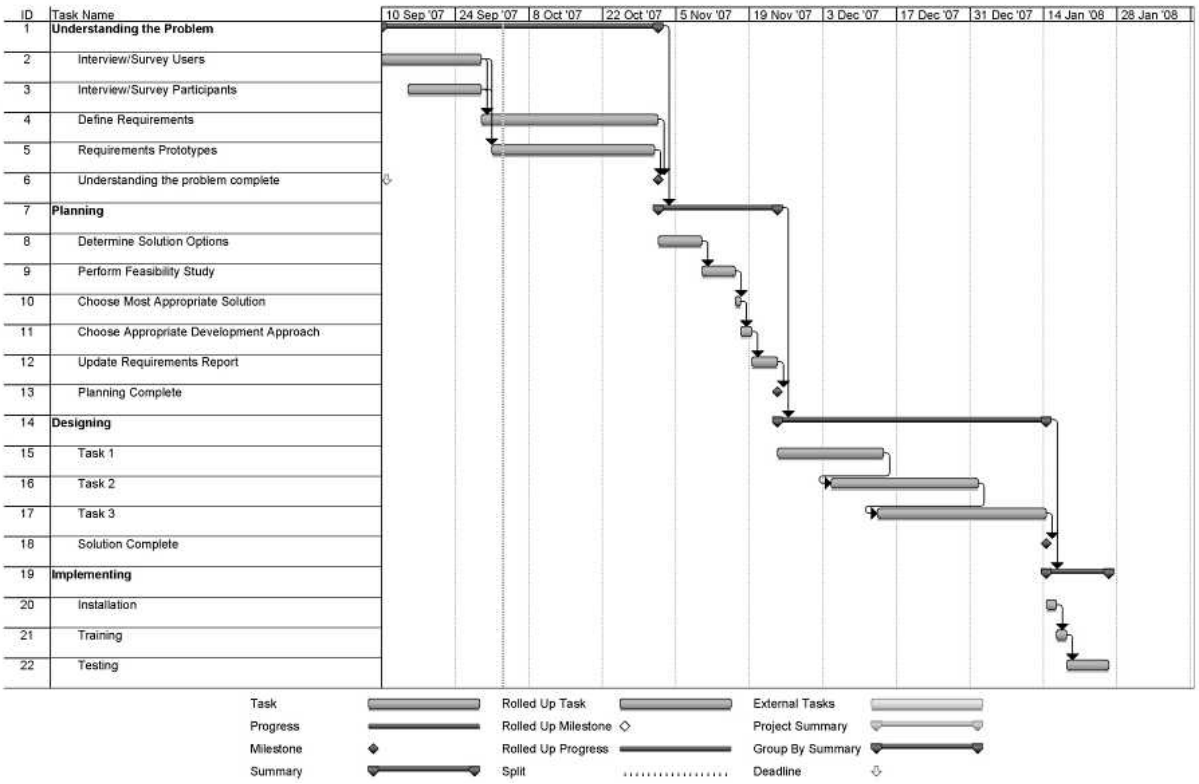


Fig 1.4
Example Gantt chart produced with Microsoft Project.

GROUP TASK Practical Activity
Create a Gantt chart to describe Tuckman's four stages of team development.

Journals and diaries

Journals and diaries are tools for recording the day-to-day progress and detail of completed tasks. Diaries are arranged in chronological order with a page or section for each day's events. Meetings, appointments, tasks and any other events are recorded in advance. Both diaries and journals are used to record details of events that have recently occurred. After or during an event diaries tend to be used to record factual information, whilst journals include a more detailed analysis and reflection on recent events. However in terms of recording past events the distinction between the two is unclear.

Diaries are an organisational tool and a memory aid. Most individuals, teams and organisations maintain diaries. For example, your school administration probably has a school diary where teachers record all future events that will affect other members of the school community, such as excursions, meetings and exam periods. The school diary is used by administration to generate the daily notices that are read out each morning or printed and distributed to staff. Teachers refer to the school diary prior to booking events to ensure they don't clash with other events. Project teams maintain diaries for similar reasons. For instance, the project manager records when meetings will occur and team members record appointments that will take them out of the office. Such information is critical if the team is to operate smoothly and effectively.

Most people also keep a personal diary where they record future events and deadlines relevant to themselves. Examining our personal diary allows us to prioritise tasks and prepare for meetings and other appointments. For instance, entries in your homework diary are used to determine the best order in which to complete your homework tasks. The IPT assessment task that's due tomorrow should be completed before you study for next week's Maths test. At school your daily schedule is largely organised for you courtesy of the school timetable. As part of a work team the order in which tasks are completed is often much more flexible. Each individual is largely responsible for determining his or her daily work routine. Their personal diary is the primary source of information for making such decisions.

Journals, and sometimes diaries, document work completed by team members during the project's development. As tasks are completed team members write down what was done and any issues they encountered. In addition such comments can include ideas and comments on possible future improvements. Project managers refer to journals as they monitor the completion of tasks and identify possible issues. Also, journals are valuable tools when evaluating a system's development. Details of problems encountered and their effect on the development process can be analysed. New ideas and other comments can be discussed and considered when planning future projects.

GROUP TASK Discussion
Diaries and journals can be hand-written paper documents, however for many project teams networked software applications are used. Contrast paper-based journals and diaries with software-based journals and diaries.

Funding management plan

A funding management plan aims to ensure the project is developed within budget. For this to occur requires that each development task be allocated sufficient funds at the correct time and that these funds are spent wisely.

Funding management plans should specify:

- How funds will be allocated to tasks. Will the funds be released in full before the task commences, progressively during the task or after the task is complete? Answers to such questions will depend on the individual nature of the task, the development approach being used and whether the task is completed in-house or is outsourced for completion by an external party.
- Mechanisms to ensure money is spent wisely throughout the SDLC. The plan should specify the procedures to be followed each time a product or service is ordered during the system's development. For example, often three quotations are required for all significant purchases and full payment is not to be made until after the product has been received and checked.
- Accountability for each task's budget. Ultimately the money spent on every detail of the system's development contributes to the total development process remaining within budget. Therefore someone should be accountable for ensuring each task is completed within its own slice of the total budget. Funding management plans should detail who this person is for each task, together with procedures they must follow to allow management to monitor their use of funds.
- The procedure for reallocating funds during development. Funding plans should include sufficient flexibility so that funds can be redirected should problems occur. Unforseen circumstances almost always occur, the funding plan should recognise and plan for such occurrences.

Communication management plan

It is vital that all parties involved in the system's development communicate with each other effectively. Communication management plans specify how this communication is to take place. Strategies documented in the communication management plan provide a structure that supports and reinforces effective ongoing communication between all team members throughout the project's development.

A typical communication management plan should specify:

- The communication medium to be used, for example e-mail, newsgroups, facsimile, meetings, weekly bulletins or even telephone calls. Different types of communication are likely to be effective under different circumstances. For instance facsimile may be specified for quotations whilst email is likely to be a more suitable medium for informal communication between software developers.
- The lines of communication. The communication management plan should specify how each party is able to obtain answers to questions or communicate other details to and from other project team members and the client. For example, it may well be appropriate for the systems analyst to contact the client directly, however it may not be appropriate for a programmer working on a specific part of the solution to do so. The communication management plan should specify the lines of communication the programmer must negotiate to obtain answers from the client.
- Methods for monitoring the progress of the system's development. This includes completion of tasks, monitoring costs and also verifying requirements as part of ongoing testing. For example, meetings can be scheduled to check critical tasks will be completed on time.
- Changing or emerging requirements. During the development of most projects new requirements will emerge or existing requirements will require alteration. The communication management plan should pre-empt such occurrences so that these new or changed requirements can be effectively communicated to all parties.

GROUP TASK Discussion
List different methods or mediums of communication, such as email, weekly bulletins, meetings, etc. Discuss when each would be appropriate during a project's development.

SOCIAL AND ETHICAL ISSUES RELATED TO PROJECT MANAGEMENT

Social and ethical issues should be considered when managing the development of information systems. The total work environment of the development team has a significant effect on productivity, commitment and also the moral of individual team members. Honest and open lines of communication, including mechanisms for identifying and resolving potential conflict, encourage a positive and cooperative climate. Team members should receive positive feedback and acknowledgement for work completed.

Privacy and copyright issues should also be considered. Often existing system data is required to assist the development. Team members must respect the confidentiality of such data and not divulge its content to others. Often parts of existing systems are utilised within new or modified systems. Permission should be obtained from copyright holders and documented before such components are used or modified. Furthermore there are copyright issues surrounding the creation of new systems. Does the individual team member retain the copyrights for work they complete or will the development company or even the client hold all copyrights? Such issues should be negotiated and documented in advance.

Some social and ethical issues related to managing the project team include:

- The work environment including health and safety issues such as ergonomic design of furniture, appropriate lighting and noise levels, varied work routines, and also procedures for reporting and resolving potential OHS problems.
- Security of data and information during development. This includes mechanisms to protect against loss such as regular offsite backups and physical barriers. It also includes techniques to restrict access to authorised personnel, such as passwords, encryption and assigning different levels of access. Development systems should also be protected against virus attack.
- Copyright issues including who will retain the copyrights for the new system. Often team members are required sign a contract that hands over all copyrights to the development company. Procedures should also be in place for obtaining permission and documenting the use of copyrighted material during development. This includes software used to assist development and also software that is incorporated within the solution.
- Respect for the rights and needs of individual team members. This includes respecting a person's right to privacy such as individuals deciding how much of their private life they wish to reveal. Also supporting team members as they complete courses to improve their work skills – many companies assist financially or are flexible about work hours prior to examinations.

GROUP TASK Discussion
Outline ergonomic issues that should be addressed when designing the work environment for development teams.

GROUP TASK Discussion
Teams work cooperatively to develop new systems, so who should hold the copyrights over the systems they develop? Discuss.

HSC style question:

Funding management plans and communication management plans are examples of project management tools.

(a) Outline the content of a typical funding management plan.

(b) Outline the content of a typical communication management plan.

(c) Predict likely consequences for projects developed without funding or communication management plans.

Suggested Solution

(a) A funding management plan documents how the total cost of a project's development will be distributed to each of the subtasks required to implement the new information system. It details when and how funds will be released and who is responsible for each part of the total budget. For instance, it could specify that a contract be signed for all purchases over some set amount. It may also specify the percentage that can be paid as a deposit with the balance only being paid once the product has been received. Funding management plans also explain what happens if a task needs more money to be completed. There could be regular financial meetings planned where all the people responsible for allocating money can discuss and negotiate changes to the budget.

(b) A communication management plan specifies how all the people involved in the development of an information system should contact each other. This includes whether they should use email, fax, telephone, meetings or some other means of communication. The communication plan explains how and who team members should contact when they have questions. For instance, if a requirement is unclear, should the developer contact the client directly by phone or do they need to communicate through management. Regular team meetings would be specified to monitor progress and discuss any issues that may arise.

(c) Likely consequences for projects without funding management plans include:

- Earlier tasks will use too much of the budget, so later development tasks are left with insufficient funds.
- The whole project runs over budget because there are no controls on who spends what. This results in little or no profit or even a loss.
- Conflict occurs between team members as they argue about how much money they need without any overall guidance.
- Money is not spent wisely or appropriately. For instance buying more expensive computers when less expensive ones would be sufficient.

Likely consequences for projects without communication management plans include:

- Team members will be unclear about which tasks they should complete, hence the development process may become "ad hoc" and disorganised.
- The client will be contacted and asked questions that they have already answered, causing them to lose faith in the ability of the development team.
- Management will be unclear of the progress of the project and its sub-tasks, which is likely to result in time and cost overruns.
- Individuals will evolve their own preferred methods of communication, which may well meet their needs but not the needs of the total development effort.

SET 1A

1. Active listening is a technique for:
   (A) faking listening.
   (B) improving understanding of a speaker's message.
   (C) ensuring the speaker knows they have been understood.
   (D) Both (B) and (C).

2. In terms of the "project triangle":
   (A) Quality improves when money, scope or time is increased.
   (B) Quality is compromised when money, scope or time is increased.
   (C) Quality improves when money, scope or time is decreased.
   (D) Quality is only compromised when money, scope and time is decreased.

3. The Belbin model is a:
   (A) tool for managing high performance teams.
   (B) theory describing the stages teams go through when first created.
   (C) strategy for selecting team members who compliment each other.
   (D) series of techniques for resolving conflict.

4. On a Gantt chart the size of each horizontal bar is used to indicate:
   (A) when the project starts and ends.
   (B) the length of time allocated to each task.
   (C) the sequence of tasks that need to be completed.
   (D) the relative importance of each task.

5. Which of the following best describes a team?
   (A) Multiple people who cooperated to achieve a common shared goal.
   (B) Multiple people who complete similar tasks in a work environment.
   (C) Co-workers whose jobs overlap or influence the work of others.
   (D) People with different skills who all contribute to a project's development effort.

6. Forming, storming, norming and peforming are stages of:
   (A) project development.
   (B) system development.
   (C) team development.
   (D) human development.

7. According to Belbin, effective teams include:
   (A) members with a balance of complimentary yet different behavioural and personality types.
   (B) people with the required skills but who have similar personalities.
   (C) people with a common goal who are able to organise and prioritise their own work routines.
   (D) members who require little leadership but do accept directions without questioning authority.

8. How development funds are allocated to tasks and who is responsible for each task's budget would be detailed within:
   (A) the funding management plan.
   (B) the communication management plan.
   (C) journals and diaries.
   (D) Gantt charts.

9. Reaching a compromise situation using logical discussion and that suits both parties requires:
   (A) team building skills.
   (B) conflict resolution skills.
   (C) negotiation skills.
   (D) interview skills.

10. Which of the following best describes the purpose of project management?
    (A) To document the system's information technology and information processes.
    (B) To document the technical details of each task required to develop the new system.
    (C) To manage the people and other resources used to develop a system.
    (D) To identify problems occurring during the development of systems.

11. Define each of the following.
    (a) Project management (b) Gantt chart (c) Team (d) Project triangle

12. Explain active listening. Use specific examples to illustrate your response.

13. Discuss suitable communication skills and strategies for:
    (a) Resolving conflict resulting from personal differences between two team members.
    (b) Negotiating the cost and terms for the purchase of hardware for the new system.

14. Explain techniques for building strong and productive development teams.

15. Outline the content and purpose of each of the following project management tools.
    (a) Gantt charts (b) Funding management plans (c) Communication management plans

INTRODUCTION TO SYSTEM DEVELOPMENT

New information systems are developed when either an existing system no longer meets the needs of its users or new needs are identified that could be met by an information system. In the Preliminary course we introduced the traditional or structured approach to developing information systems. In the HSC course we extend our discussion of the traditional approach and also introduce a variety of other system development approaches.

The development of many information systems is substantially different to the development of most other engineered systems. We touched on these differences at the start of this chapter. The most fundamental differences are due to the nature of software. Design and construction of software is integrated – we actually construct software as it is being designed. Hence, the design of software can and is often altered significantly whilst it is being built and even after it is installed and operating. Although a complete redesign during or after development is still costly, it is still a relatively minor issue compared to redesigning many other products when half built or already complete. For instance it really would be a disaster to completely change the design of a building when it's half built, making such a change after the building is complete would be difficult to even contemplate. For buildings and most other engineering projects the traditional structured approach to development makes logical sense. The requirements and the design must be determined precisely prior to construction commencing. The need for such precise requirements and accurate design is less critical when developing information systems. Indeed many argue that accurately determining requirements in advance is not a realistic possibility for most information systems. Furthermore, for most operational information systems correcting design errors and implementing new requirements is a routine maintenance task. It is for these reasons that various different system development approaches have emerged and are appropriate when developing information systems.

GROUP TASK Discussion

Design errors do occur with all types of products. Contrast the recall of a motor vehicle to correct a small problem with an update to a software application to correct a small bug or security flaw.

In this chapter we consider various approaches for developing information systems including the traditional approach, outsourcing, prototyping, customisation, participant development and also agile methods of system development. In general, the traditional or structured approach requires each stage to be completed before the next commences. Outsourcing is where external specialists are contracted to develop part of the system. Prototyping is when an existing prototype is refined over time and evolves into the final system. Customisation is where existing information technology is modified to meet different requirements. Participant development is when people who are or will be part of the system develop the system. Agile methods are used to refine a system whilst it is operational.

An appropriate selection of approaches should be selected and integrated to suit the particular needs of each project. For some projects a strict traditional approach may well be suitable, whilst for others an integrated combination of approaches is appropriate. Regardless of the final approach used, a similar set of development activities will still be present, however they will likely be performed in different sequences and with different emphasis. In this chapter we work through these development activities in the order dictated by the traditional structured approach whilst pointing out differences when using other approaches.

GROUP TASK Discussion
Consider each of the development approaches mentioned above and decide whether it could be suitable for use as part of the development approach for other products and projects. Use examples of possible products or projects to justify your decisions.

The traditional structured approach to system development specifies distinct stages or phases. These stages combine to describe all the activities or processes needed to develop an information system from an initial idea through to its final implementation and ongoing maintenance. The complete development process is known as the 'System Development Life Cycle (SDLC)' or simply the 'System Development Cycle (SDC)'. In this text we will use the abbreviation SDLC. The SDLC is closely linked to the concept of structured systems analysis and design, where a series of distinct steps are undertaken in sequence during the development of systems.

During each traditional stage of the SDLC a specific set of activities is performed and each stage produces a specific set of outputs. These outputs are commonly called 'deliverables'. For example, a funding management plan is an example of a deliverable that describes the management of the project's budget. In general the deliverables from each stage of the SDLC form the inputs to the subsequent stage. For example, the initial requirements report provides crucial input data when formulating the cost feasibility of a solution.

The particular stages or phases within the SDLC differ depending on the needs of the organisation and also on the nature of the system being developed. As a consequence different references split the SDLC into slightly different stages. In the IPT syllabus the SDLC is split into five stages, namely

1. Understanding the problem,
2. Planning,
3. Designing,
4. Implementing, and
5. Testing, evaluating and maintaining.

In the remainder of this chapter we discuss the activities occurring during each stage. The overall activities performed are similar regardless of the number of distinct stages. The five stages specified in the IPT syllabus describe one method of splitting the SDLC, but of course there are numerous other legitimate ways of splitting the SDLC into stages.

Consider the following sets of SDLC stages

The SDLC policy (1999) of the U.S. House of Representatives specifies and describes the following seven phases:

1. Project Definition
2. User Requirements Definition
3. System/Data Requirements Definition
4. Analysis and Design
5. System Build
6. Implementation and Training
7. Sustainment

The HSC Software Design and Development (SDD) course focuses on the creation of software rather than total information systems. In terms of information systems the development of software is just one part of the solution. In the SDD syllabus the version of the SDLC used is called the Software Development Cycle and is split into the following five stages:

1. Defining and understanding the problem
2. Planning and design of software solutions
3. Implementation of software solutions
4. Testing and evaluation of software solutions
5. Maintenance of software solutions

Many Systems Analysis and Design references use SDLC stages similar to one of the following:

1. Investigation	1. Planning	1. Requirements
2. Design	2. Analysis	2. Analysis
3. Construction	3. Design	3. Design
4. Implementation	4. Build	4. Construction
	5. Implementation	5. Testing
	6. Operation	6. Acceptance

GROUP TASK Discussion
Compare and contrast each of the above lists of SDLC stages with the stages specified in the IPT syllabus.

GROUP TASK Research
Use the Internet or other references to obtain at least two further examples of SDLC stages. Do the IPT stages agree in principle with the stages from your examples?

GROUP TASK Discussion
In most examples of the SDLC, including IPT, the word 'implementing' refers to the installation of the final system. However in the SDD course 'implementing' refers to building or coding the software.
Can you explain this anomaly? Discuss.

Consider the following:

David Yoffie of Harvard University and Michael Cusumano of MIT studied how Microsoft developed Internet Explorer and Netscape developed Communicator. They discovered that both companies did a nightly compilation (called a build) of the entire project, bringing together all the current components. They established milestone release dates and enforced them. At some point before each release, new work was halted and the remaining time spent fixing bugs. Both companies built contingency time into their schedules, and when release deadlines got close, both chose to scale back product features rather than let milestone dates slip.

GROUP TASK Discussion
Identify project management techniques apparent in this development scenario. Is this system development approach suitable for developing all types of information systems? Discuss.

Before we begin examining each stage of the SDLC in detail let us briefly identify the activities occurring and the major deliverables produced during each stage of the IPT syllabus version of the SDLC. The data flow diagram in *Fig 1.5* shows each stage as a process, and the deliverables as the data output from each process. The deliverables from all previous stages are used during the activities of each subsequent stage. To improve readability these data flows have not been included on the diagram. For example the Requirements report is produced when Understanding the problem and is then used and perhaps updated during all subsequent stage, not just the next Planning stage. The grey circular arrow behind the diagram indicates the traditional sequence in which the stages are completed. Project management efforts are ongoing throughout the SDLC.

Users are included on the diagram as their input is central to the successful development of almost all information systems. Indeed it is often ideas from users, and in particular participants, that initiate the system development process in the first place. Furthermore, the needs of users largely determine the requirements of the new system. As a consequence feedback from users is vital during the SDLC if the requirements are to be met and are to continue to be met.

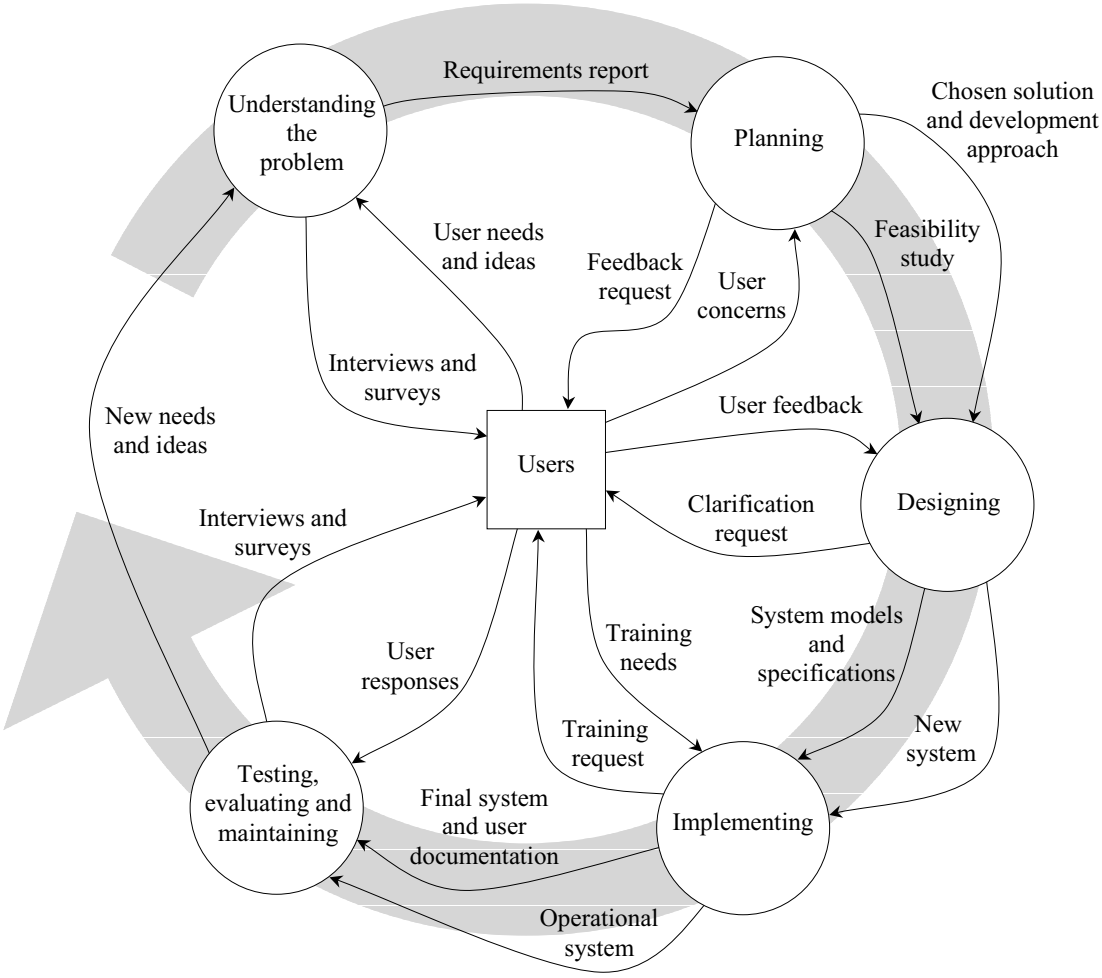
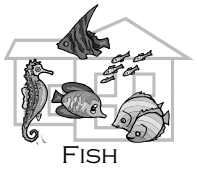


Fig 1.5
The version of the System Development Lifecycle (SDLC) used in IPT.

GROUP TASK Discussion
The above diagram implies some activities during the SDLC. Identify and discuss the general nature of the activities occurring during each stage.

Consider Pet Buddies Pty. Ltd.

To illustrate the activities occurring and the deliverables produced during the SDLC we will use a pet care business called 'Pet Buddies Pty. Ltd.'. This example scenario will be referred to throughout this chapter as we develop an information system for the business. A brief introduction to Pet Buddies follows:



PET BUDDIES PTY. LTD.

EXPERT HOME PET CARE – BREEDER SPECIALISTS

Company Background

Iris and Tom Cracker have been breeding exotic, and valuable, parrots for more than 20 years. It had always been difficult for Iris and Tom to find suitable people to care for their birds when they went away on business trips or holidays. Numerous businesses existed that provided satisfactory home care for pet dogs and cats, however exotic birds were another matter. In 1999 Iris and Tom formed the business 'Bird Buddies' to fulfil this need.

Initially Bird Buddies concentrated on providing expert home care to aviculturists (bird breeders) – most of their business being generated through local avicultural clubs. It soon came to their attention that similar problems existed for breeders of reptiles, fish and also dogs and cats. In early 2001 the name Bird Buddies was changed to Pet Buddies. As Iris and Tom had limited experience with these other species they began to contract expert reptile, fish, dog and cat personnel. Each of the experts employed is a successful and experienced breeder in their own right.

Pet Buddies has grown substantially since 1999 to the point where in 2004 they employed more than 25 different experts and serviced some 600 customers. Currently Iris and Tom are unable to provide home care services themselves as their entire day is more than filled with the administrative and management aspects of running this thriving business.

Customer Service Guarantee

- All experts are honest, genuine and motivated specialists with extensive experience keeping and/or breeding similar animals to your own.
- A specialist veterinarian for your species is on call at all times.
- We are aware of the value of many exotic animals, hence we guarantee confidentiality in regard to the number and type of animals you keep. (Optional insurance is available upon request.)
- We guarantee to perform all activities (e.g. feeding, medication, cleaning, exercise regime) specified on your accepted application form.
- Direct contact between customers and experts is encouraged. We believe quality of service and peace of mind is closely linked to frequent communication between each of our experts and our customers.

Fig 1.6
Pet Buddies Pty. Ltd. company background and customer service guarantee.

GROUP TASK Discussion
Identify the central needs that are fulfilled by Pet Buddies Pty. Ltd. How are these needs being met? Discuss.

GROUP TASK Discussion
Brainstorm a list of possible ideas that could be implemented within a new information system for Pet Buddies.

UNDERSTANDING THE PROBLEM

The primary aim of this first stage of the SDLC is to determine the purpose and requirements of a new system. Once the requirements have been established then an accurate Requirements Report can be produced. The Requirements Report is therefore the primary deliverable produced by this stage – it defines the precise nature of the problem to be solved. In essence this stage determines what needs to be done.

Systems Analyst
A person who analyses systems, determines requirements and designs new information systems.

A systems analyst is responsible for analysing existing systems, determining requirements and then designing the new information system. They are problem solvers who possess strong analytical and communication skills. In relation to 'understanding the problem to be solved' the systems analyst completes and/or manages each of the activities specified in *Fig 1.7*. Notice that each of these activities contributes to the creation of information needed to define the requirements for the new or modified system. For example, interviewing/surveying existing system users and participants provides the information required so that the systems analyst can produce models of the existing system. Requirements prototypes can be used to obtain further information relevant to the production of the Requirements Report. Note that we are concentrating on a traditional structured approach, hence each of the activities and deliverables provides additional input needed to create the subsequent deliverable. There is a logical sequence to the order in which the activities and the production of deliverables occurs.

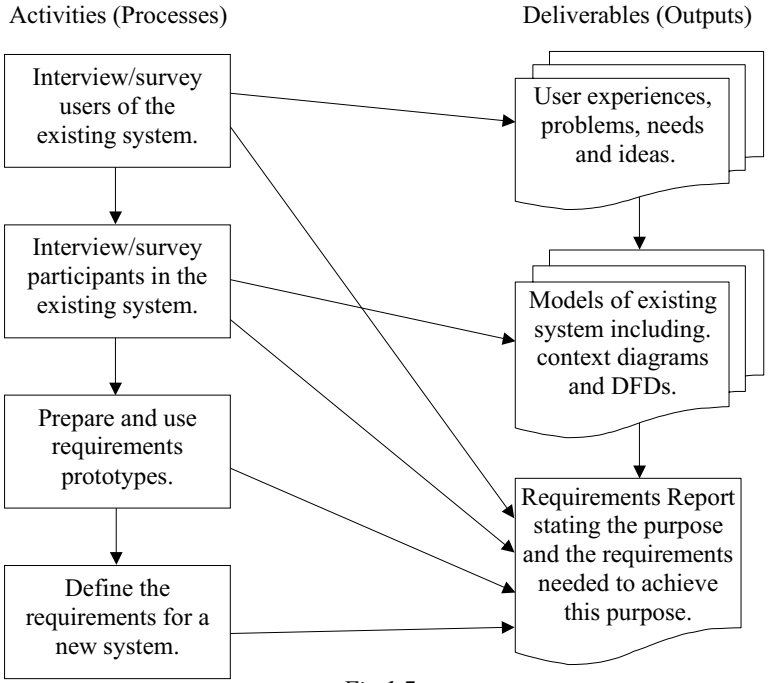


Fig 1.7
Activities performed and deliverables produced during the 'Understanding the problem' stage of the SDLC.

GROUP TASK Discussion
A lot of effort is directed towards understanding the operation of the existing system. Why do you think this is necessary? Discuss.

Before we commence discussing the detail of each activity specified in *Fig 1.7* it is worthwhile discussing what a requirement is, and how requirements relate to the system's purpose. In general terms, a requirement is a feature, property or behaviour that a system must have. If a system satisfies all its requirements then the system's purpose will be achieved. In practice a system's requirements are a refinement of the system's purpose into a list of achievable criteria.

Requirements
Features, properties or behaviours a system must have to achieve its purpose. Each requirement must be verifiable.

A successful project achieves its purpose, and furthermore this purpose is achieved when each requirement has been met. Therefore it is necessary to verify that all requirements have been met if we are to evaluate the success of the project. For this to occur all requirements must be expressed in such a way that they can be verified or tested. Consider the statement 'Customers should receive a response in a reasonable amount of time after submitting a request'. This is a satisfactory objective and may well form part of the system's purpose, however it is difficult to verify if it has been achieved. It is a subjective statement and is therefore unsuitable as a requirement. Now consider the statement 'The system shall generate a customer quotation within 24 hours of the system receiving a customer's quotation request'; this statement can easily be tested and is therefore a suitable requirement. In essence it must be possible to test and verify that a requirement has or has not been met.

GROUP TASK Discussion
System requirements should address aspects of all the components of an information system, including participants, data/information, information technology and also information processes.
Why do you think this is necessary? Discuss.

INTERVIEW/SURVEY USERS OF THE EXISTING SYSTEM

In the majority of information systems the purpose of the system is primarily concerned with fulfilling the needs of its users – users being the people who utilise the information created by the system. For example, the objective 'Customers should receive a response in a reasonable amount of time after submitting a request' aims to fulfil the need of users, who are customers, to receive timely responses. It follows that such knowledge is critical when trying to understand the problem to be solved.

Interviews and surveys are the primary tools for collecting user experiences and problems with the existing system, and also for identifying their needs and any new ideas they may have to improve the system. It is common to conduct a survey of a sample of users – the larger the sample the more statistically reliable the results will be. Unfortunately surveys, by their very nature, must be constructed in advance. This means the questions tend to draw out particular information that the survey designer feels is relevant. Furthermore, it is likely that even open-ended questions will only be answered within the context of the existing system. For example, when modifying an existing website the open-ended question 'Do you have any suggestions for inclusion in the new website?' is included in a user survey. The intention of the question is to gather new user needs. In reality many people will not respond at all to open ended questions and those that do respond are likely to address improvements to the current website rather than suggestions outside the scope of the existing system. The results of surveys are often more useful for highlighting existing problems rather than revealing new needs and ideas that are not currently being addressed.

New needs and ideas are more likely to reveal themselves via personal and informal interviews conducted with users in their own environment. Unfortunately conducting such interviews is time consuming and expensive. Interviews can also be conducted with small focus groups of users where particular aspects of the system critical to these users can be informally discussed.

Be aware that what people say they need and what they actually need is often different. Furthermore, users often express the relative significance of their needs incorrectly. For example, a user may express a strong need for a particular report to be generated more rapidly. In reality this report may only be used on a weekly basis, hence saving a minute or so becomes relatively insignificant. Such issues are potential problems with both surveys and interviews. In an attempt to verify user needs, many systems analysts directly observe sample users whilst they work with the existing system. This can only occur when an existing system is already in use and operating. For completely new systems requirements prototypes can be built so that possible user needs can be verified using a simplified version of the new system. Requirements prototypes are more often used with system participants rather than general users. We discuss requirements prototypes in more detail later in this section.

Once the collection of data from users has been completed the systems analyst must organise the data into a form suitable for analysis; spreadsheets or simple databases are common tools. The data is then analysed to determine and prioritise problems with the new system, identify user needs and also to document any new ideas. A report summarising all this information can then be produced. This report forms the essential deliverable resulting from the interviewing/surveying of users.

Consider Pet Buddies Pty. Ltd.

Iris and Tom, the owners of Pet Buddies, have contracted Fred to advise them about possible options in regard to improving the efficiency of their existing information systems. Fred, who is a systems analyst, explains the sequence of activities he will perform, beginning with identifying the experiences and needs of their users. In this case the users are comprised of two distinct groups, the customers and the experts.

The customers are indirect users of the system, whilst the experts are direct users who are also system participants. Each group will have different experiences and needs and hence requires separate consideration. Iris, Tom and Fred agree that it makes sense to consult the experts once the needs of the customers have been established.

After consultation with Iris and Tom, Fred creates the one page 'Customer Satisfaction Survey' reproduced in *Fig 1.8*. A copy is mailed to all 600 of Pet Buddies existing customers. A stamped self-addressed envelope is included with each survey in an attempt to increase the response rate.

GROUP TASK Discussion

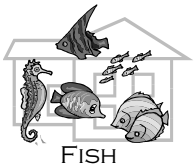
The survey created by Fred (see *Fig 1.8*) aims to encourage each customer to provide comments. Identify features on the survey that encourage comments and explain why Fred would wish to encourage comments.

After 2 weeks Iris and Tom have received a total of 315 completed surveys. Iris feels this is a rather poor response rate, however Fred informs her, that in his view the response rate is exceptional as he anticipated approximately 30% would be returned – he also mentions that response rates for emailed surveys are usually less than 10%.

CATS & DOGS

BIRDS

REPTILES



FISH

PET BUDDIES PTY. LTD.

EXPERT HOME PET CARE – BREEDER SPECIALISTS

Customer Satisfaction Survey

Dear Jack and Jill,

We are constantly looking for ways to improve the quality of our services. To do that, we need to know what you think. As a valued customer, we'd really appreciate it if you would take just a few minutes to respond to the handful of questions below.

Please return your completed survey in the included stamped self-addressed envelope or fax to 9912 3456.

Please tick "Outstanding" or "Needs Improvement" and then comment:

	Outstanding	Needs Improvement
Booking your home care service		
Feedback and communication with your expert		
Confidence in your expert's abilities		
All activities were accomplished well		
Flexibility of home care service		
Confidentiality and privacy		
Value for money		

Thankyou!

Fig 1.8
Customer Satisfaction Survey for Pet Buddies Pty. Ltd.

Fred's task is to organise the survey responses in such a way that they can be analysed to identify a list of customer needs. He enters the responses into a database that is linked to a copy of Pet Buddies existing customer database. This enables Fred to analyse the survey responses according to animal type, location, expert, length of home care, frequency of home care, cost and so on. The aim is to identify if particular customer problems and needs are specific to particular aspects of Pet Buddies' services. For example, "Are repeat customers' needs and problems different to the needs and problems experienced by first time customers?" or "Do keepers of reptiles have different experiences and needs compared to those keeping birds?"

GROUP TASK Discussion
Identify the information technology needed by Fred to perform the analysis detailed above.

During his analysis Fred intends to telephone some of the customers who responded to the survey, his aim being to confirm any problems they mention and also to obtain further specific details.

GROUP TASK Discussion
Identify reasons why Fred would choose to telephone some customers to confirm and obtain more specific details.

Fred will use the information to establish a set of user needs, which will then form the basis for the creation of a set of achievable user requirements. Let us assume Fred has created a list of user needs and he is now formulating user requirements. One of these needs together with the associated user requirements follows:

Customers need reassurance that all specified activities are indeed being completed.

- The system shall ensure experts have a complete list of required activities for each customer.
- The system shall generate 'completion of activities' reports for customers.
- The system shall maintain a record of how often a customer is to receive a 'completion of activities' report.
- The system shall alert management if a 'completion of activities' report cannot be generated on time.

GROUP TASK Discussion
Notice how the above need includes the word 'need', similarly each requirement commences with the words 'The system shall'. The use of these specific words is not necessary, however it is a technique Fred finds useful. Why do you think Fred uses this technique? Discuss.

INTERVIEW/SURVEY PARTICIPANTS IN THE EXISTING SYSTEM

Participants within existing systems will have an understanding of the part of the system with which they primarily interact. They are able to identify problems and often they also have ideas in regard to solving these problems. Furthermore, participants are a vital source of information in regard to the detail of the information processes occurring within the existing system. Notice that in *Fig 1.7* the results of participant interviews and surveys are used to create models of the existing system as well as to create the final requirements report for the new system.

System analysts often perform task analysis activities with participants. Task analysis involves writing down each step performed to complete a particular task. The time taken to complete each step is noted together with the inputs required and the outputs produced during the task's completion. Such information provides a basis for the creation of system models, such as data flow diagrams.

Although system participants are familiar with the procedures required to perform their specific tasks, they are often not aware of how the system actually performs these tasks or how these tasks contribute to the larger information system. For example, data entry operators are unlikely to understand the various information processes that utilise the data they enter. As a consequence data entry operators may comment that some data items have no relevance to the information system. It is the job of the systems analyst to determine the correctness of participants' responses.

Consider Pet Buddies Pty. Ltd.

Pet Buddies is a small business where the two owners, Iris and Tom, either initiate or carry out virtually all of the information processes. From past experience Fred knows that this is true of most small businesses. Obviously Iris and Tom are the main system participants. During discussions with Iris and Tom it is clear the business is growing, and soon it will simply be impossible for them to complete all these tasks themselves.

Fred suspects that currently Iris and Tom are controlling all information processing – he needs to confirm this suspicion. Fred feels part of the solution is likely to revolve around passing control, and perhaps even responsibility for some processes to the experts contracted by Pet Buddies. Currently the experts' primary task is to perform the actual home care activities. These activities are absolutely central to Pet Buddies operation. But do the experts currently initiate or carry out any of the existing system's information processes? Fred needs to answer this question, and furthermore he wishes to identify possible information processes the experts could perform or initiate without compromising their ability to perform the home care activities.

Fred decides to spend a day observing and questioning Iris and Tom while they work. During this time he will concentrate on the movement of data through the system, together with the identification of the information processes occurring. Fred also intends to note the time Iris and Tom spend on each task. Fred's aim is to gather enough data to understand the operation of the existing information system and also to identify tasks where significant amounts of time can be saved.

GROUP TASK Discussion
Is it really necessary for Fred to understand the details of the existing system? Surely he should just focus on the new system. Discuss.

Some of the data collected by Fred during his day with Iris and Tom is reproduced in *Fig 1.9* on the next page. Much of the data was compiled during his observations of Iris and Tom at work. At the end of the day Iris, Tom and Fred spend about an hour discussing Fred's observations. Various changes are made to compensate for the fact that this was just a single day, and therefore not entirely typical.

GROUP TASK Discussion
Consider the organisation of the data collected by Fred (see *Fig 1.9*). Identify reasons why Fred has used this method for organising the data.

Iris, Tom and Fred's discussion then turns to the experts. Fred indicates he wishes to identify their needs, together with their experiences as participants working with the current system. Furthermore he feels it is vital to include them in the development process as early as is possible. He proposes to create a questionnaire, which he will use as the basis of a phone survey/interview with at least half of the experts. Once the results have been analysed, a meeting with all the experts will take place to confirm and communicate his findings. Iris and Tom suggest an informal meeting, combined with a social barbeque. Fred agrees and a date is set.

GROUP TASK Discussion
Fred is continually verifying the data he collects about the existing system. Is this really necessary? Justify your response.

Pet Buddies Existing System — Task Analysis (C-Computer, M-Manual)

Inputs (source)	Information Processes	Outputs (sink)	Time (Min)	Frequency (per day)
Prepare new job				
• Customer Application (Customer via mail) • Confirmation (Customer) • Confirmation (Expert)	• Enter/edit customer details (C) • Schedule job to an expert (M) • Confirm with expert (M – Phone) • Confirm job and activities with customer (M – Phone) • Enter job details (C) • Print activity report pro-forma (C) • Print and fax job card to expert (C) • Print customer confirmation and job quotation (C)	• Job card (Expert) • Customer confirmation and job quotation (Customer)	45	1-3 (up to 15 prior to Christmas)
New customer enquiry				
• Questions, name and address (New customer via phone)	• Discussion (M) • Collect name address details (M) • Mail marketing pack (M)	• Marketing pack (Customer)	6	5-10
Create activity report				
• Activity details (Expert via phone)	• Photocopy customer activity report pro-forma (M) • Complete activity report pro-forma with expert (M) • Telephone, fax or mail activity report to customer (M)	• Completed activity report (Customer)	40	10 (up to 30 around Christmas)
Prepare invoice				
• Job quotation (Database) • Additional charges (Job card from expert)	• Enter invoice charges (C) • Print and mail invoice (M)	• Invoice (Customer)	5	3

Fig 1.9
Some of the data collected by Fred to understand the existing system.

Over the next few days Fred develops a context diagram (a simplified version is reproduced in *Fig 1.10*) and begins to create a series of data flow diagrams to model the operation of the existing system. As Fred creates the data flow diagrams he gains a deeper understanding of the operation and flow of data through the existing system. As a consequence new ideas begin to emerge in regard to possibilities for inclusion in the new system.

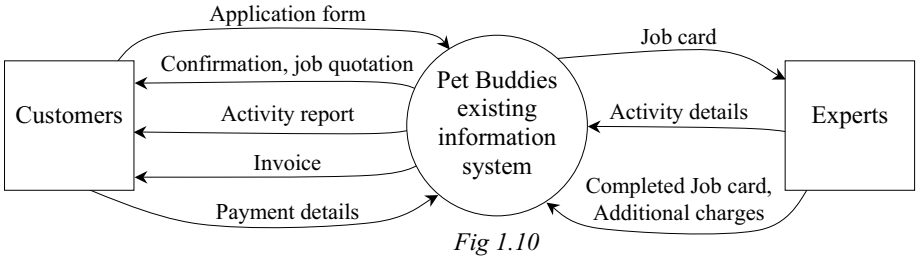


Fig 1.10
Simplified context diagram for Pet Buddies existing information system.

Fred now creates the questionnaire he will use during his telephone surveys/interviews with the experts. Some of the questions emerge from the context and data flow diagrams he has just created. For example, he notices that the activity details from the experts are not significantly altered by the system prior to their delivery to customers, rather their format is simply altered. Fred is particularly interested in each expert's response to the following question: "How do you record the results of each home care activity report prior to phoning Iris and Tom?"

GROUP TASK Discussion
Why is Fred particularly interested in each expert's response to the question 'How do you record the results of each home care activity report prior to phoning Iris and Tom?' Discuss.

Let us assume Fred has phoned the experts and completed his surveys/interviews. He produces a summary of the expert's needs and faxes a copy to Iris and Tom. Although Iris and Tom agree with most of the identified needs, there are two with which they disagree, namely:

- Experts need to deal directly with customers.
- Experts need to be able to alter the length of time of each home care visit after their initial visit.

Iris and Tom feel many of the experts do not possess the necessary communication skills to contact customers directly. Fred points out that most of the comments leading to this need came from either fish or reptile experts, however a number of others also implied such a need. After further discussion, Fred agrees to question the expert's need to deal directly with customers in some detail during the informal experts meeting.

Iris and Tom express concern over how they will charge customers if the length of the home visits is altered after the customer has signed their application and subsequently agreed to their quotation. Fred assures them there are many techniques that will emerge to solve this issue.

The informal meeting takes place with 20 of Pet Buddies experts in attendance. Fred delivers his presentation, followed by a question and answer session. The experts are split down the middle in regard to contacting customers directly. Half see it as the logical thing to do – some of them comment that they already know many of their customers through clubs and shows. The other half is reluctant to alter the current system, they feel it is not part of their job and furthermore they simply do not have the time. Fred, together with Iris and Tom, assure the experts that any changes will take account of both points of view.

GROUP TASK Activity
List the tasks performed by Fred during his work with Pet Buddies so far. Identify the skills Fred possesses to complete these tasks.

REQUIREMENTS PROTOTYPES

Requirements Prototype
A working model of an information system, built in order to understand the requirements of the system.

Requirements prototypes model the software parts of the system with which the users will interact. The model is composed of screen mock-ups and perhaps sample reports. A requirements prototype accurately simulates the look and behaviour of the final application with minimal effort. A typical requirements prototype is in effect a simulation of the user interface. It includes all the screens, menus and screen elements together with the ability for users to enter sample data and even view sample reports. Users, and in particular participants, use the requirements prototype as they simulate the tasks they will perform with the real system. Requirements prototypes do not contain any real processing – for instance records are not really added, edited or even validated. The aim is to confirm, clarify and better understand the requirements.

Often a sequence of requirements prototypes is produced, each new prototype being a refinement of the previous version in response to feedback. This repetitive process continues until both the developers and the participants are satisfied that all requirements are understood. The final prototype can be used exclusively to refine the requirements or it can be used as the basis for development of the real system.

The visual nature of prototypes makes them valuable tools for confirming understanding and sparking new ideas compared to more traditional lists of requirements. Participants are able to experience the proposed system; they can easily comprehend how the new system will operate. Members of the development team can sit down with participants to observe and discuss the detail of the system. For instance, simple things, like adding a keyboard shortcut or moving a seldom used field lower on the screen are easily identified. It's rather difficult to think of such detail without such simulations.

Although requirements prototypes are particularly well suited to gathering and clarifying user interface requirements they can also assist with understanding and generating ideas for other system requirements. For instance, a participant who works on accounts may notice there is no function for identifying invoices that have only been partly paid. A salesman after viewing a prototype screen displaying recent sales leads might suggest the system send new leads to salesmen as text messages.

Requirements prototypes can also be designed for distribution to a broad audience. Software tools are available that can create standalone requirements prototypes that include the ability for users to add comments or even make changes to screen layouts. These comments and edits are then returned electronically to the development team for further analysis and possible inclusion in the final system.

There are specialised software applications for creating requirements prototypes. Many of these specialised products are part of larger requirements definition applications. For smaller projects, requirements prototypes can be generated using standard application packages. For example, Microsoft Access is able to create forms and reports that can be used for this purpose. In addition, many modern programming language environments provide a similar facility without requiring any programming expertise. When the requirements prototypes are created using the same software that will be used for developing the real system then the prototype can actually evolve into the new system. Most specialised requirements definition packages are also able to export prototypes for use within many popular software development products.

Consider the following extract on the iRise software product suite:

iRise – The Power of Simulation

Simulation: iRise simulations look and behave exactly like the final business application, eliminating confusion and getting everyone on the same page.

Business Analysis: BAs use iRise to quickly assemble a visual blueprint of business applications – before coding. Iterative stakeholder review & approval is accelerated with the iRise collaborative platform.

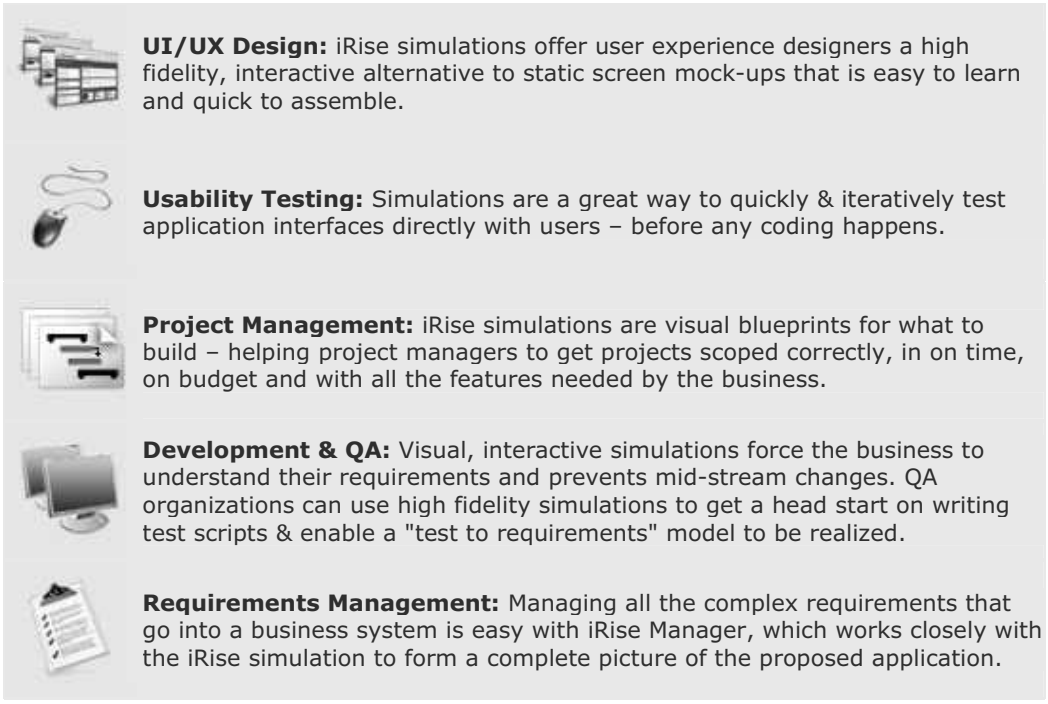


Fig 1.11
Extract of an overview of the iRise product suite

GROUP TASK Discussion
Identify the development personnel and development tasks mentioned in the above extract. Discuss how simulations (or requirements prototypes) assist these people to perform their tasks.

DEFINE THE REQUIREMENTS FOR A NEW SYSTEM

The previous activities aimed to provide sufficient information to enable the creation of a complete set of requirements for the new system. These requirements are expressed within a formal 'Requirements Report'; this report is the most significant deliverable from the first stage of the SDLC.

A Requirements Report can be considered as a 'black box' – it specifies the inputs and the outputs together with their relationships to each other. However it makes no attempt to solve the problem. In fact when creating the Requirements Report the systems analyst must be careful to avoid references and inferences that imply a particular solution. For example, "The system shall operate continuously should a storage device fail" is a better requirement than "The system shall include a RAID device where hard disks can be hot swapped". The second version specifies a particular solution and effectively rules out other possible solutions. Furthermore, the second version is likely to make little sense to the client.

The Requirements Report should be expressed in such a way that it is understandable to the client and also useful as a technical specification for the new system's developers. In most instances these two parties have a very different view of the system, hence it is often appropriate for two different versions of the requirements report to be produced. Each version contains the same content organised into a form that meets the specific needs of each party. In essence the Requirements Report forms a communication interface between the client and the system's technical developers. Ensuring each party understands the Requirements Report is absolutely essential as all subsequent stages of the SDLC rely on its content.

The process of preparing a Requirements Report for a project is known as 'Requirements Analysis', which is itself a complete discipline. There are university courses, technical books and dedicated requirements analysis professionals. Many versions of the SDLC include requirements analysis as a distinct stage. In IPT we can only hope to scratch the surface of the requirements analysis process. To do this we limit our discussion to:

- a description of how the Requirements Report is used during the remaining stages of the SDLC, and
- the content of a typical Requirements Report when using the traditional system development approach.

Consider the following:

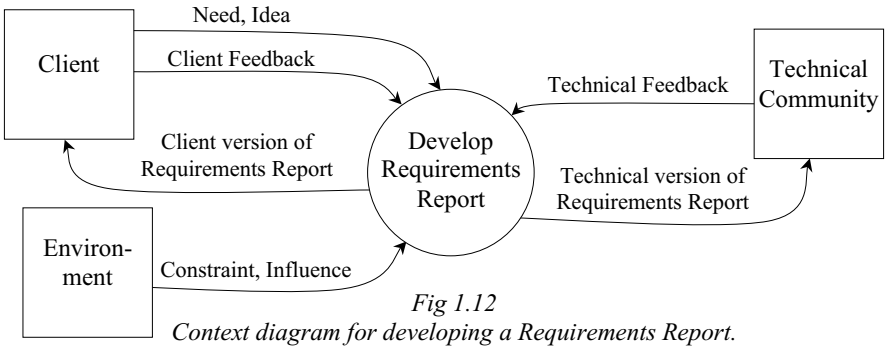


Fig 1.12
Context diagram for developing a Requirements Report.

The context diagram in *Fig 1.12* is a modified version of a similar diagram included within the *IEEE Guide for Developing Systems Requirements Specifications (IEEE Std 1233, 1998 Edition).* The diagram indicates that developing a Requirements Report involves feedback from both the client and the technical community – possibly numerous times. The client is the organisation, or their representative, who approves the requirements. The technical community includes all the development personnel who will eventually design, build and test the new system.

The diagram includes the environment as an entity that influences and places constraints on the requirements. In the IEEE 1233 standard the environmental influences include political, market, cultural and organisational influences.

GROUP TASK Discussion
Describe the flow of data modelled on the context diagram in *Fig 1.12*. Can you explain why no data flows to the environment?

GROUP TASK Discussion
Compare and contrast the client's view of the Requirements Report with the technical community's view of the Requirements Report.

How Requirements Reports are used during the SDLC

When planning, the Requirements Report is used to determine possible solution options and their feasibility. Different solutions can be compared fairly, as they all aim to achieve identical requirements. The Requirements Report is a 'blue print' of what the system will do, as such it forms the basis of the contract between the client and the system's development team. The contract is a formal legal agreement, signed by both the client and the system's developers. The system is complete once all requirements have been met, and hence the contract is also complete.

During the planning stage a particular solution and system development approach is chosen. Once this has occurred the Requirements Report can be updated to include specific detail about the selected solution. For instance, details of the subtasks, timing of tasks, participants, information technology and data/information can be identified and documented within the report.

During the design of the solution, the overriding aim is to achieve all of the requirements specified in the Requirements Report. Commonly the design process involves the creation of various subsystems. Each subsystem aims to meet specific requirements, however these requirements may well originate from different areas of the Requirements Report. For example, requirements concerning the storage and retrieval of data are likely to be present throughout many areas of the Requirements Report, yet the system's designers may choose to meet these requirements within a single subsystem – perhaps using a database management system and its associated hardware. At all times the Requirements Report remains the common ground, it describes unambiguously what the system will do, whilst the designers determine the detail of how it will be done.

When implementing new systems it is necessary to decide on a method for converting from the old to the new system. As the Requirements Report describes what the new system does then it also determines which (and when) existing systems and subsystems can be removed. Furthermore the conversion requires participants to be trained on the new system. The Requirements Report highlights areas of participant interaction that training should address.

Testing and evaluation of the new system is all about checking that each requirement has been met. Clearly the Requirements Report is central to this process. Tests are designed to specifically verify that each requirement has been met. Once all tests are successful then the client, and the developers, can be confident the system will meet its purpose.

Once the new system is operational it must continue to be maintained. Requirements change and new requirements will emerge over time. The Requirements Report must evolve to accommodate such modifications to the system. Furthermore, it forms the basis for ensuring new modifications do not replicate or affect the achievement of existing requirements.

The content of a typical Requirements Report when using the traditional system development approach

Clearly the most important content within a Requirements Report is the system requirements themselves, however other details are needed to introduce and support the formal requirements.

In this section we examine some general areas for inclusion within a typical traditional Requirements Report. One possible outline is reproduced in *Fig 1.13*. This sample is intended to cover most aspects of most new information systems in a logical manner. Remember there is an infinite variety of possible information systems, hence this outline will need to be adjusted, or perhaps significantly changed, to suit each new system's specific needs.

The outline shown in *Fig 1.13* implies a printed report will be produced. Be aware that the

Table of contents
Glossary

1. **Introduction**
   - 1a System purpose.
   - 1b The needs of the users.
   - 1c System scope.
2. **General system description**
   - 2a System context
   - 2b Major system requirements
   - 2c Participant characteristics
3. **System requirements**
   - 3a Physical
   - 3b Performance
   - 3c Security
   - 3d Data/Information
   - 3e System operations

Fig 1.13
Sample Requirements Report outline

organisation and format of the final requirements report can take many forms. It may indeed be a printed text document, or it could be a hypertext document that includes the final requirements prototype, or it could be a series of linked interactive diagrams that enable the requirements to be viewed from different perspectives. The method of organising and formatting the report should be chosen to effectively and efficiently communicate the requirements of the particular information system to the particular client and system developers.

Let us briefly consider the content under each heading contained in the sample outline shown in *Fig 1.13*:

1. **Introduction**
   - **1a System purpose**
     Identifies the overall aims and objectives of the system. Often the identified needs of the client are also included. The purpose is the reason the system is being developed.
   - **1b The needs of the users**
     The final set of user needs that will be addressed by the new system. This list may not include all the needs identified when surveying/interviewing users. Rather it includes just the user needs that the client has agreed the new system should address.
   - **1c System scope**
     An explanation of what the system will and will not do. All major functionality that will be included in the new system is explained. Perhaps more importantly, any functionality that could possibly be interpreted as being part of the new system but is actually not going to be part of the system should be specifically excluded. In essence the boundaries of the system are defined – what is part of the system and what is not.
2. **General system description**
   - **2a System context**
     An overview of all the data/information that enters and leaves the system, including its source and destination. Commonly a context diagram is used together with a written description.
   - **2b Major system requirements**
     A description of the major capabilities of the new system. The description may include diagrams as well as written descriptions.
   - **2c Participant characteristics**
     Each different type of participant is identified and the nature of their use of the system described.
3. **System requirements**
   - **3a Physical**
     This section includes any requirements that specify aspects of the system's physical equipment and the physical environment in which it will operate. This includes requirements in regard to the construction, weight, dimensions, quality, future expansion and life expectancy of the hardware. In regard to the physical environment, typical requirements will deal with temperature, humidity, motion, noise and electromagnetic interference levels. If the equipment will be outside then requirements in regard to rain and wind conditions should be included.
   - **3b Performance**
     This section includes requirements that relate to the ability of the system to complete its processes correctly and efficiently. It includes requirements in

regard to the time taken by the system to complete tasks, the accuracy of the information produced and the frequency with which tasks occur.

3c Security

All requirements that deal with access to the system and privacy of data/information within the system are included in this section. This includes requirements that address both accidental and intentional security breaches. It should also include requirements in regard to protecting against loss of data, such as backup and recovery.

3d Data/Information

This section includes requirements that address the data and information needs of the system. This includes requirements specifying what data is kept and what information is produced. Requirements relating to the organisation and storage of data can also be included.

3e System operations

This section addresses requirements relating to the system during its operation. This includes human factors such as requirements in regard to the user interface within software and the ergonomic design of equipment, including both hardware and software. It also includes requirements that support the system's continued operation such as regular preventative maintenance, reliability and also repair times should a fault occur.

GROUP TASK Discussion

Identify and discuss reasons why the System Scope may have been included within the Requirements Report outline in *Fig 1.13*.

Notice that the outline above does not group requirements that address information technology separately to those that address information processes. For example, under the heading 'Performance' the time taken to complete tasks is mentioned. A typical requirement might state "The system shall complete task A in less than B microseconds". Such a requirement is likely to have consequences in regard to the selection of a suitable CPU and also in regard to the efficient design of the information processes used to complete task A. One possible solution may rely heavily on a fast CPU whilst another relies on a more efficient use of information processes. If the requirement was listed under the heading 'Information Technology' then the second, and perhaps better solution is unlikely to emerge. Similarly if the requirement was listed under the heading 'Information Processes' then the first solution is less likely to be considered. Remember the aim at all times is to specify what the system must do without indicating or even implying a specific solution – the sample outline discussed above assists in this regard.

GROUP TASK Discussion

"The details of the information processes occurring within an information system are essentially the solution to the problem, hence such details should never form part of a Requirements Report".
Do you agree? Discuss.

The Requirements Report outline described above is particularly suitable for systems developed using the traditional approach. Many of the alternative approaches, in particular prototyping and agile approaches, allow new requirements to emerge and existing requirements to change as the system is being designed. When such approaches are used the Requirements Report must also be allowed to evolve and change to encompass modifications and additions. The use of software for managing requirement changes is recommended when such systems are developed by a team.

Suitable procedures need to be in place to ensure all team members are kept up to date with changes as soon as they occur. Such procedures would be documented within the communication management plan. Various changes to other parts of the project plan will no doubt be needed, for instance updates to the schedule and budget.

Consider the following:

Fig 1.14 shows a screenshot from 'Objectiver', a requirements engineering software application written by the Belgium company CEDITI. Objectiver is able to produce both printed and interactive HTML requirements reports that include diagrams and/or textual information. Different views of the same requirements can be produced to suit different audiences.

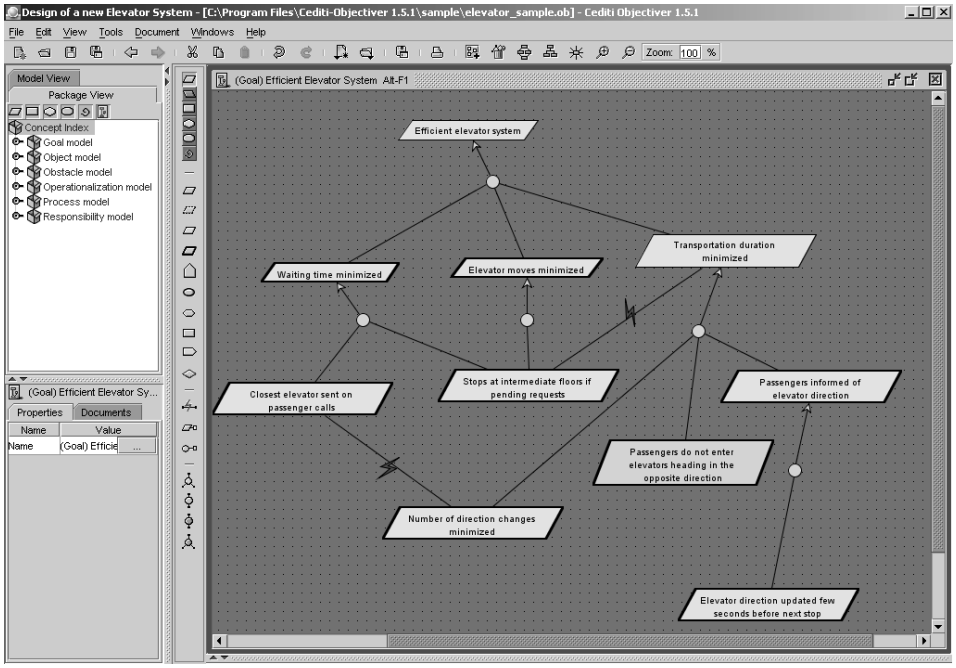


Fig 1.14
Screenshot from Objectiver, a requirements engineering software application produced by the Belgium company CEDITI.

Objectiver is based on a goal-oriented methodology called KAOS. The highest or top-level goals are essentially the aims and objectives that must be met to achieve the system's purpose. Each goal is progressively refined into a verifiable set of requirements. The HTML reports produced by Objectiver allow the progress of the requirements analysis process to be easily shared with all interested parties. Furthermore any alterations to the requirements that occur throughout the SDLC can easily be distributed to all parties involved in the system's development.

GROUP TASK Discussion
Identify advantages of using a software application such as Objectiver compared to using a word processor to prepare a Requirements Report.

GROUP TASK Research
Research, using the Internet or otherwise, at least one other example of a requirements engineering software application. Briefly describe its major features. Share your findings with other members of your class.

Consider Pet Buddies Pty. Ltd.

Selected sections of the final Requirements Report developed by Fred for Pet Buddies are reproduced below in *Fig 1.15* and *Fig 1.16*.

1. Introduction

Pet Buddies provides professional confidential expert home care services to breeders and keepers of birds, reptiles, fish, dogs and cats. Many of their customers are professional large-scale breeders who maintain extensive animal collections. The value of their customer's collections range from $5000 up to $10 million, the average value being approximately $40,000.

1a. System Purpose

The purpose of this system is to:

- automate the generation and distribution of activity reports.
- personalise contact between customers and experts during home care services.
- improve the accuracy of quotations for home care services.

1b. Pet Buddies' Customers Needs

Pet Buddies' Customers need:

- reassurance that all specified activities are being completed.
- feedback on problems encountered during home care services.
- to be confident in the ability of the expert performing their home care service.
- to be confident that details of their animal collection and its location remain confidential.

1c. System Scope

The system will:

- collect sufficient data to enable accurate quotations to be produced.
- collect data required to generate the activity reports.
- generate activity reports at the correct times.
- facilitate the display of activity reports to customers.
- ensure customer data is secure.

The system will NOT:

- create or generate quotations.
- include or provide functionality in regard to invoicing or any other financial functions of the business.
- perform any marketing functions.

Fig 1.15
Pet Buddies Requirements Report Introduction

GROUP TASK Discussion

It is clear from the above introduction that the proposed system addresses just some of Pet Buddies' information system needs. Suggest and discuss possible reasons why this decision may have been made.

GROUP TASK Discussion

Presumably much of the existing system will remain in operation. Identify and describe possible consequences for the new system in terms of its development and also in terms of its operation.

3. System Requirements

3a. Physical

The system shall:

3a.1. use mobile devices weighing less than 5kg.
3a.2. use mobile devices that operate for at least 9 hours without accessing mains power.
3a.3. include hardware components that are replaceable within 24 hours.
3a.4. include hardware components that regulate their own temperature without the need for external cooling.
3a.5. include components with a minimum life expectancy of greater than 2 years.
3a.6. use computer communication hardware compatible with Pet Buddies existing gigabit Ethernet LAN.

3b. Performance

The system shall:

3b.1. provide activity reports to customers within 60 minutes of the necessary data being received by the system.
3b.2. enable experts to submit data for activity reports from any location, including whilst on the customers premises.
3b.3. include the facility for Pet Buddies management to at their discretion check and/or edit the content of any activity report prior to its release to a customer.
3b.4. include the facility for Pet Buddies management to specify that all activity reports from a particular expert or to a particular customer must be approved by Pet Buddies management before release to customers.
3b.5. alert Pet Buddies staff immediately an activity report becomes overdue.
3b.6. provide the facility for customers to provide feedback on the content of activity reports at any time, including immediately after receiving an activity report.
3b.7. alert Pet Buddies management immediately customer feedback specified in 3b.6 is received.
3b.8. include the facility for the system to collect and store all quotation data directly from experts within 60 minutes of the expert determining such data.
3b.9. alert Pet Buddies management immediately quotation data specified in 3b.8 is received.
3b.10. reuse the collected quotation data to generate outlines for use during the production of activity reports.
3b.11. collect data from experts on the total time taken to complete each home care service.
3b.12. generate statistical reports on demand that compare the actual time taken to perform each home care service with the estimated time on the quotation. Reports can be generated for individual customers, individual experts, individual animal types and/or within specified date ranges.

Fig 1.16
Section 3a and 3b of Pet Buddies Requirements Report.

GROUP TASK Discussion
How does each of the above requirements assist in the achievement of the system's purpose? Discuss.

GROUP TASK Discussion
The security and data/information sections of the Requirements Report have not been reproduced above. Develop a list of possible requirements that these two sections of the report would likely include.

Consider Pet Buddies Pty. Ltd.

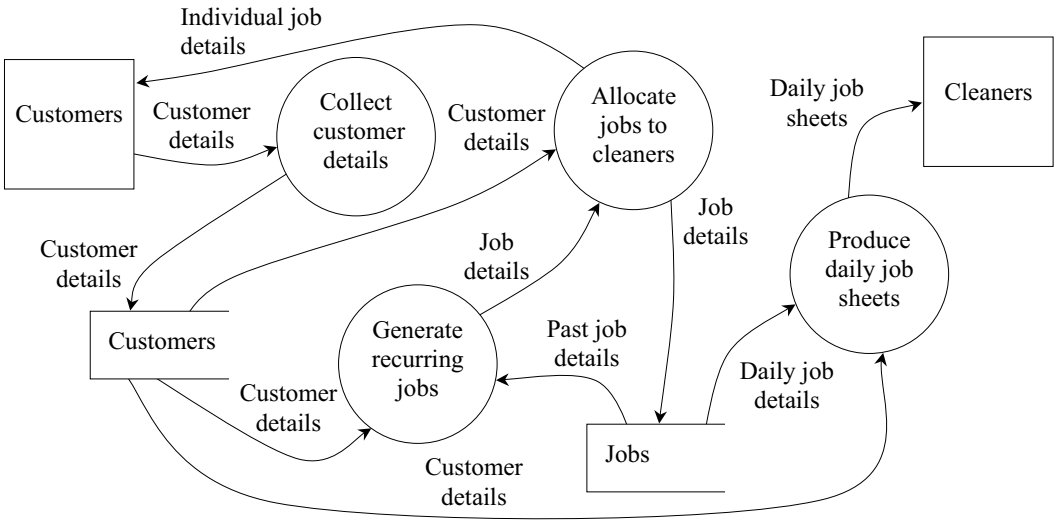
Fred intends to submit the Requirements Report to various businesses to obtain ideas, and quotations, in regard to possible solutions. Fred advises Iris and Tom that before this occurs they need to determine some idea of a budget and also some idea of when the system should be operational. This information is required to enable Fred to explore possible solution options that meet the requirements, including budget and time constraints.

After discussion, Iris and Tom inform Fred that the budget should be set based on the principle that development costs will be recovered within 2 years of the system becoming operational. In essence the cost of the new system should be covered by increased company profits within 2 years. Fred, although he agrees, points out various other considerations. For example, he points out that Iris and Tom will have more time for leisure and/or business development and marketing activities. He also mentions the likely increase in capitol value of the business due to a lowered reliance on their personal skills and knowledge – in essence the business will be more self-sufficient as an independent entity.

GROUP TASK Discussion
Is it always necessary for the budget and the date of system completion to be known prior to considering possible solution options? Discuss.

HSC style question:

A cleaning business currently uses a manual system to collect customer information and allocate jobs to each of its cleaners. They are investigating the possibility of computerising their existing system. The data flow diagram below models the existing manual information system. Currently each process is completed manually by one of the system's participants.



(a) Two different symbols on the data flow diagram refer to customers. Compare and contrast the use of these two symbols using specific examples from the data flow diagram.

(b) Cleaning jobs are allocated on a priority basis. All customers are allocated a certain priority, higher priority customers having their job completed first. Recurring jobs are allocated a particular time and all other jobs must be allocated around these times.

Using the data flow diagram together with the above information describe the likely contents of the data flows labelled 'Customer details' and 'Job details'.

(c) Propose suitable techniques that could be used to identify problems present within the existing manual system.

Suggested Solution

(a) The customer's entity refers to the actual human customers who are the source of the customer details used during the collection process. The customer's data store is a file that contains details of each of the business's customers. Both deal with customer data, but one is the source of this data whilst the other is a storage area for the data – probably a filing cabinet.

(b) The 'Customer Details' data flow would contain a customer's name, address, phone number, how long the job will take, any unusual aspects to the job, preferred day of the week and/or time, and also whether it is a recurring job. If it is a recurring job then the frequency and priority of the job would be included.

The 'Job Details' data flow passes data regarding each individual cleaning job that is assigned to a cleaner. This would include the date, time and duration of the job together with the customer's contact details and the cleaner who has been assigned the job.

(c) A simple customer satisfaction survey form could be created and distributed to existing customers. Perhaps the cleaners could leave the survey after they complete each job. The survey would ask customers to comment on both negative and positive aspects of the cleaning business – including questions about their experiences booking jobs and also whether their job was completed at a convenient time. Each cleaner could also be surveyed to obtain information about any problems with regards to their daily job sheets.

Once the surveys have been completed the results will need to be analysed to identify significant problems. This list of problems could then be distributed to each of the participants so they are able to express any ideas they have in regard to possible solutions. In additions the participants can also be asked about any other problems they perceive. Interviews with participants could take place so that their ideas and possible solutions can be explored in more detail.

In the new computerised system most of the information processes will be automated. Hence a requirements prototype would be a valuable aid for ensuring all of the current manual processes are addressed and also for introducing the general nature of the proposed system to the participants.

Comments

- In an HSC or Trial HSC examination part (a) would likely attract 2 marks, part (b) would attract 3 marks and part (c) would attract approximately 4 marks.
- In part (b) it is important to notice that the Customer Details includes details of recurring jobs in addition to name, address and phone numbers.
- A variety of different suitable techniques could have been proposed in part (c).

SET 1B

1. The person who determines requirements and designs new information systems is best described as a:
   (A) Project manager.
   (B) Participant.
   (C) System analyst.
   (D) Engineer.

2. Feedback from users should occur during which stages of the SDLC?
   (A) Understanding the problem and planning stages.
   (B) Designing and implementing stages.
   (C) Testing, evaluation and maintaining stage.
   (D) All stages of the SDLC.

3. Which type of information is more likely to be obtained from interviews compared to surveys?
   (A) New ideas and needs.
   (B) Details of existing issues.
   (C) Current procedures for completing tasks.
   (D) Responses from many users.

4. Tools for diagrammatically representing existing systems include:
   (A) requirements reports and requirements prototypes.
   (B) interviews/surveys of users and participants.
   (C) application packages and requirements definition packages.
   (D) context and data flow diagrams.

5. During testing and evaluation the requirements report is used to:
   (A) determine the most suitable method for converting from the old to the new system.
   (B) design the information processes that will form part of the new system.
   (C) determine the feasibility of possible solution options.
   (D) verify all requirements have been met.

6. An explanation of what the system will and will not do helps to define the:
   (A) needs of users.
   (B) system scope.
   (C) system purpose.
   (D) characteristics of participants.

7. In IPT, which of the following lists of SDLC stages is in the correct sequence?
   (A) Understanding the problem, planning, designing, implementing, testing, evaluation and maintaining.
   (B) Understanding the problem, designing, planning, implementing, testing, evaluation and maintaining.
   (C) Understanding the problem, implementing, designing, planning, testing, evaluation and maintaining.
   (D) Planning, understanding the problem, designing, implementing, testing, evaluation and maintaining.

8. A simulation of a new system built to understand the system's requirements is known as a:
   (A) Requirements Report.
   (B) Requirements Prototype.
   (C) Requirements Model.
   (D) Evolutionary Prototype.

9. Features, properties or behaviours a system must have to achieve its purpose are called:
   (A) requirements.
   (B) needs.
   (C) decisions.
   (D) processes.

10. When using a traditional system development approach the main deliverable from the "Understanding the problem" stage is the:
    (A) Interview and surveys.
    (B) Feasibility study.
    (C) Operational system.
    (D) Requirements report.

11. Define each of the following terms.
    (a) survey (b) interview (c) requirement (d) system purpose

12. Describe the content of a typical requirements report.

13. Explain how the requirements report is used during the system development lifecycle.

14. Assess the value of requirements prototypes compared to surveying and interviewing users and participants.

15. Explain why it is necessary to analyse the operation of existing systems when developing new systems.

PLANNING

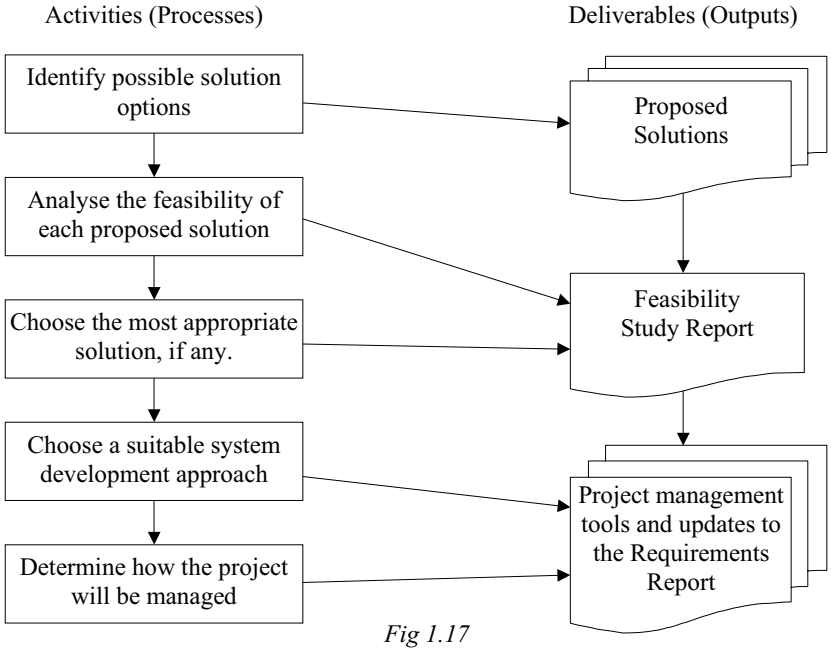


Fig 1.17
Activities performed and deliverables produced during the 'Planning' stage of the SDLC.

In this, the second stage of the system development cycle, the aim is to decide which possible solution, if any, should be developed and then decide how it should be developed and managed. In other words the feasibility of developing the new system is analysed to create the Feasibility Study Report. Assuming an appropriate solution is found then a system development approach can be determined that is suited to developing that solution. Finally project management tools are used to document the detail of how the project will be managed and the Requirements Report is updated to include and reflect details of the chosen solution and system development approach.

FEASIBILITY STUDY

Feasible
Capable of being achieved using the available resources and meeting the identified requirements.

So what is a feasibility study? Consider making some large purchase – say a new car, a new computer or some new piece of furniture. Prior to making such a purchase you ask yourself various questions. What kind do I want? What features do I want? Will it do what I need it to do? What will it cost and can I afford it? Will it require maintenance and what will that cost? And finally should I actually buy it? In essence you are performing an informal mini-feasibility study. Asking and answering similar questions is the essence of all feasibility studies. The ultimate aim is to determine the feasibility of each possible solution and then recommend the most suitable solution. Remember it is possible, and reasonably common for no feasible solution to be recommended, meaning the existing system will remain.

The feasibility of each possible solution must be assessed fairly – the Requirements Report plays a major role in this regard. Without a common set of requirements it would be difficult to make a fair comparison between different solution options. This presents a new problem – if a number of solutions are able to meet the requirements then on what basis can a decision be made? The 'Feasibility Study' is also concerned with addressing criteria upon which the answer to this question is based.

Feasibility studies generally examine each possible solution option in terms of the following four feasibility criteria:

GROUP TASK Discussion
A solution that meets each of the requirements within the requirements report must be the preferred solution. Do you agree? Discuss.

- technical feasibility
- economic feasibility
- schedule feasibility
- operational feasibility

Let us examine each of these areas and consider questions that should be addressed under each area as part of a feasibility study.

Technical Feasibility

The technical feasibility of a solution is concerned with the availability of the required information technology, its ability to operate with other technology and the technical expertise of participants and users to effectively use the new technology. For example a new off-the-shelf state-of-the-art software application may, according to its specifications, meet the system's requirements, however without a large customer base there are likely to be concerns in regard to continuing support and upgrades. Furthermore few people will be trained in the use of the application. This means it will be difficult to replace trained personnel during the system's future operation.

Questions used to determine a solution's technical feasibility include:

- Do we currently possess the necessary technology?
- Is the technology readily available?
- How widely used is the technology?
- Are existing users of the technology happy with its quality and performance?
- Will the technology continue to be upgraded and supported in the future?
- Will the technology operate with other existing and possible future new or emerging technologies?

GROUP TASK Discussion
Identify from whom and how answers to each of the above questions could be obtained.

GROUP TASK Discussion
How could the answers to the above questions be compiled in order to compare the technical feasibility of different solution options? Discuss.

Economic Feasibility

The economic feasibility of each solution option is determined by performing a "Cost-benefit analysis". This involves calculating all the costs involved in the development and implementation of each solution option. On the surface it would appear that the least expensive option to develop and implement would be the most economically feasible, however this is not always the case. There are various other factors that contribute to the economic feasibility of a solution and should be considered as part of a cost-benefit analysis. Let us consider such factors and then discuss issues that should be considered when analysing the economic feasibility of a solution.

- **Factors affecting a solution's economic feasibility**

Development costs
- Cost of the development team
- Systems analyst and other consultancy fees
- Software costs to purchase or build the software
- Hardware costs to purchase, lease and/or assembly the hardware
- Infrastructure costs such as new buildings, communication links and power.
- Installation of the system
- Training participants and users
- Converting from the old system to the new system

Ongoing operational costs
- Hardware maintenance and repair costs
- Software licences and upgrade
- Maintenance of infrastructure that supports the system
- Salary/wages for participants
- Support costs for users, including ongoing training
- Consumables such as toner cartridges and paper

Tangible benefits (that can relatively easily be assigned a dollar value)
- increased sales
- cost reductions
- increased efficiency
- increased profit on sales
- more effective use of staff time

Intangible benefits (that are difficult to assign a dollar value)
- increased flexibility of the system
- higher quality products or services
- improved customer satisfaction
- better staff morale

GROUP TASK Discussion
Explain how a dollar value could be determined for each of the tangible benefits list above.

GROUP TASK Discussion
Discuss possible techniques for determining a dollar value for the intangible benefits listed above.

- **Issues to consider during a cost-benefit analysis**

Cost-benefit analysis, as the name implies, compares all the costs with all the benefits in an attempt to determine the total return on the money invested into the new system. One would imagine that if the benefits, in dollar terms, exceed the costs then the system is economically feasible – unfortunately things are not quite so simple! Cost-benefit analysis aims to determine the real benefits of each solution option. The techniques used are the same as those used by economists to analyse investments.

Issues to consider include:
- The money spent on the new system could have been invested elsewhere; hence the benefits of the new system must also exceed the benefits that would have been realised without the new system. In accounting terms the Net Present Value (NPV) is determined. A positive NPV indicates a good investment, and the largest

NPV indicates the best investment. Negative NPV values indicate investments that should not be developed further.

- Comparing the percentage profitability of each solution option rather than just the absolute profit. This is known as return on investment (ROI) analysis. ROI describes the percentage increase of an investment over time.
- When will the new system have paid for itself? This is known as the 'break-even point' – the point in time where the new system has been paid for and it begins to make a profit. For example, in *Fig 1.18* solution option A has a break-even point of 2 years whilst solution option B has a break-even point of 3.5 years. The period of time prior to the break-even point is called the payback period.

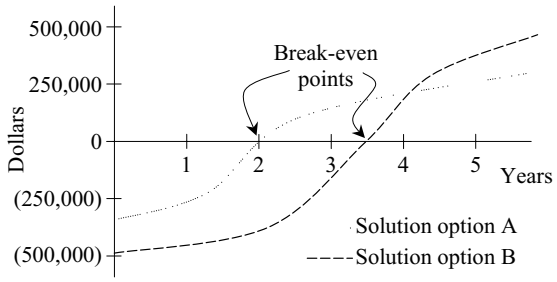


Fig 1.18
Break-even analysis is used to determine when each solution option becomes profitable.

Solutions with a high NPV, high ROI and short payback period will be the most economically feasible. Unfortunately all these measures are based on future predictions, hence they can never be determined with complete accuracy. Furthermore, different clients will have different needs that will affect the relative importance of each measure when determining the economic feasibility of solutions to particular problems.

GROUP TASK Discussion
The most profitable solution is not always the most economically feasible solution. Do you agree? Justify your answer using examples.

Schedule Feasibility

Schedule feasibility is largely about whether the solution can be completed on time. The project plan, and in particular the Gantt chart, will specify the deadlines for completion of each development task. Schedule feasibility aims to determine if such deadlines can be met. It should also examine the consequences should some tasks and even the entire project fail to meet its specified deadlines.

Questions used to determine a solution's schedule feasibility include:

- How long will it take to obtain the required information technology?
- If new personnel need to be employed then how long will that take?
- How long will it take to retrain existing team members?
- Will retraining affect the ability of staff to complete existing tasks on time?
- Are the deadlines mandatory or are they desirable?
- If the project runs over time what are the consequences?
- Is it possible to install an incomplete solution should deadlines not be met?
- How can development of the solution be monitored to verify deadlines are indeed being met?

GROUP TASK Discussion
Identify from whom and how answers to each of the above questions could be obtained.

Operational Feasibility

Operational feasibility aims to evaluate whether each solution option will work in practice rather than whether it can work. It considers support for the new system from management and existing employees. In essence a solution option is likely to be operationally feasible if it meets the needs of the participants and users of the system.

Questions used to determine a solution's operational feasibility include:

- Do existing staff support the solution option?
- Do management support the solution option?
- Does the nature of the solution 'fit in' or conflict with the nature of other systems that will remain in place?
- Will the nature of work change for participants?
- Are participants open to change or resistant to change?
- How do the end-users feel about the delivery of information from the new system?
- Do participants already possess the technical expertise?
- Do users already possess the technical skills to use the technology?
- Is training and support available and will it remain available?

GROUP TASK Discussion
Identify from whom and how answers to each of the above questions could be obtained.

GROUP TASK Discussion
How could the answers to the above questions be compiled in order to compare the operational feasibility of different solution options? Discuss.

Consider Pet Buddies Pty. Ltd.

Fred has now researched possible solutions and has determined two solution options. A brief outline of each option in regard to the production of activity reports is reproduced below:

Pet Buddies' solution option A

1. Each expert is provided with a personal digital assistant (PDA) device. The expert enters activity report data into their PDA using the device's handwriting recognition capabilities.
2. Each expert then connects their PDA to the Internet via their mobile phone and emails the text data to a dedicated email address at Pet Buddies.
3. Software at Pet Buddies receives the message, notifies Iris and Tom and stores the data in a database linked to the customer's name.
4. The message generated for Iris and Tom provides them with an option to view and edit the report. In all cases they must indicate their approval before the report is made available to the customer.
5. To retrieve activity reports the customer phones Pet Buddies and is connected to a computerised voice mail system. The voice mail system collects the customer's ID number and then gives the customer the option of listening to activity reports or having them faxed.

6. If the customer chooses to listen then the data is retrieved from the database and read over the phone using TTS software, otherwise the data is formatted into an activity report, which is subsequently faxed to the customer's fax number.

Pet Buddies' solution option B

1. A voice mail software application is installed at Pet Buddies. This application interfaces with the existing customer database and provides a separate password protected mailbox for each customer's activity reports. It also includes mailboxes for each expert that store the initial activity report data prior to it being checked.
2. Whilst onsite experts ring Pet Buddies voice mail system using their mobile phone. The system establishes their identify and also the customer's identity.
3. The voice mail system then uses TTS to ask the expert to comment on each area needed to complete the particular customer's activity report. The expert's responses are digitally recorded along with the synthesised questions.
4. A message is generated for Iris and Tom that provides them with the option to view and edit the report. In all cases they must indicate their approval before the report is made available to the customer.
5. To retrieve activity reports the customer phones Pet Buddies and is connected to the voice mail system. The voice mail system collects the customer's ID number and then gives the customer the option of listening to activity reports or having them faxed.
6. If the customer chooses to listen then the data is retrieved from the database and read over the phone, otherwise the data is sent to a speech recognition engine where it is converted to text. The text is then formatted into an activity report, which is subsequently faxed to the customer's fax number.

GROUP TASK Discussion
Compare each of the above solution options to the system requirements in *Fig 1.16* on page 42. In regard to the activity reports, do you think both options are capable of meeting all of these requirements? Discuss.

GROUP TASK Discussion
Identify the essential differences between the two solution options in terms of:
- the tasks performed by the experts,
- the tasks performed by the software at Pet Buddies, and
- the different types of media used by the systems.

Fred is currently conducting a feasibility study in order to determine which solution, if any, should be developed. Below is a brief summary of his initial research and thoughts grouped according to the four feasibility criteria described above:

Technical feasibility

Option A	Option B
– PDAs must be purchased for each expert. – Experts have various different models of mobile phone, hence different interface cables are needed. + Most mobile phones contain inbuilt modem functionality. + Free suitable software for the PDAs is readily available.	+ All experts currently have a mobile phone. + Experts do not require additional information technology. + Minimal training is needed for experts. + Voice mail software is readily available and has a wide market. – Speech recognition not 100% accurate using telephone quality audio recordings.

– Mobile phone coverage is limited in some areas serviced by Pet Buddies. + Millions of users worldwide use PDAs in conjunction with mobile phones for email. + Lower spec computer as only text files are stored. TTS occurs in real time.	– Custom software is needed to automate the data transfer to the speech recognition engine and then back to the voice mail system. – More powerful computer and much larger storage needed for audio files.

Economic feasibility

Option A	Option B
– Significant costs involved in the purchase of PDAs for each expert. – Cost of interface cables for each expert. – Pet Buddies responsible for maintenance costs in regard to PDAs. + As only text data is being emailed connection charges are low for each text file sent. + TTS software is inexpensive yet accurate. – Synthesised speech not so acceptable to customers. + Faxed reports more accurate. – Training of experts will be more costly. + Low spec computer will cost less.	– Mobile call charges are high, particularly during peak periods in the middle of the day. – Custom software will be costly to develop. – A high quality (and expensive) speech recognition engine is needed. + Spoken activity reports will be higher quality. + Spoken activity reports use experts voice so more personal and acceptable to customers. – Faxed reports less accurate. – Edited voice reports will be obvious and less acceptable to customers as the voice will be different. – Higher spec computer will cost more.

Schedule feasibility

Option A	Option B
– Experts require more training. + All information technology is readily available. – Correct operation of TTS software is critical to improving the efficiency of the system as most customers require voice reports.	– Custom software will take significant time to develop and implement. + Speech recognition and custom software can be added later. This would require fax reports to be manually typed as per existing system.

Operational feasibility

Option A	Option B
+ No restriction on the number of experts submitting reports at any one time. – It is likely experts will be less supportive due to their increased tasks. – Significant changes to expert's work. – Few of the experts have experience using PDAs.	– Number of experts submitting reports at one time is limited to the number of telephone lines into the voice mail system. – Editing voice versions of reports will require more work by Iris and Tom. + Minor changes to expert's work.

GROUP TASK Discussion
Consider the economic feasibility points above. Categorise each point according to the Factors affecting a solution's economic feasibility on p48.

GROUP TASK Discussion
Based on the above information which option do you think is the most suitable? Discuss.

CHOOSING A SYSTEM DEVELOPMENT APPROACH

There are numerous system development approaches that can be used in isolation, combined and/or integrated to form a system development approach appropriate for developing almost any project. The particular nature of the development team and the individual characteristics of each project determines which system development approach should be selected.

In this section we describe the defining characteristics of a variety of different approaches. Be aware that it is unusual for a single approach to be used in isolation; rather for most projects different approaches are combined and integrated to create an appropriate system development approach for developing each particular system.

We consider characteristics of the following system development approaches:

- Traditional
- Outsourcing
- Prototyping
- Customisation
- Participant development
- Agile methods

Traditional

The traditional or structured approach to system development involves very formal step-by-step stages. Each stage must be completed before progressing to the next stage. As we discussed earlier in this chapter, the traditional approach produces detailed deliverables from each stage that become the essential inputs necessary to begin the next stage. For example when Understanding the problem all requirements must be precisely determined and documented. The deliverable being the final requirements report. This report is required to assess the feasibility of possible solutions in the Planning stage. Unlike the traditional approach, other system development approaches accept and allow for requirement changes.

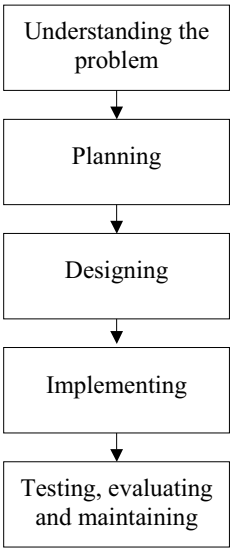


Fig 1.19
Traditional system development approach.

As each stage is completed deliverables feed down to the next stage and also into all subsequent stages, for this reason the traditional approach is also known as the waterfall approach. For example the requirements report is used by all the stages that follow, similarly for the project plan, system models and various other deliverables. In addition, waterfalls flow downhill not uphill. In relation to the traditional approach this means there is no returning to a previous stage and there are also few opportunities for users and others to provide ongoing feedback. Unfortunately this means errors or omissions can feed through the system development cycle without detection. For instance, omitting a requirement within the requirements report is difficult to detect until the system is operational.

For most systems the cost to correct such issues increases exponentially as development progresses. In general a problem or oversight in the first "Understanding the problem" stage will cost five times more to correct if not detected until the "Planning" stage. It will cost ten times more to correct when not detected until the "Designing" stage, forty times more when detected in the "Implementation" stage and the cost of correction can be hundred of times more expensive once the system is operational.

Despite these concerns, the traditional approach remains well suited to the development of many types of information systems. For instance, most large critical systems and also most new hardware products are developed using this approach. The performance and reliability of these systems is vital and furthermore the requirements for these systems can be determined in advance.

Consider the following information systems:

- Upgrades to the infrastructure that connect banks together within the EFT system.
- A new model of mobile phone is to be developed. It is expected that in excess of 100,000 units will be manufactured.
- A computer controlled water jet cutting machine. The machine can cut intricate parts from plastic and sheet metal material based on information within CAD files.

GROUP TASK Discussion
Identify aspects of the above systems that make them suitable for development using the traditional system development approach.

Outsourcing

Outsourcing of development tasks involves using another company to develop parts of the system or even the complete system. It is often more cost effective to outsource specialised tasks to an experienced company rather than employ new staff or train existing staff. This is particularly the case when the information system is being developed in-house or when aspects of the system require highly specialised skills that are unlikely to be required once the system is operational.

For many new information systems the entire project is outsourced to a professional development or consultancy company. In many cases this company will, in turn, further outsource specialised aspects of the system's development. For instance, in most industries there are specialist IT consultancy companies. These IT consultants have worked with a large number of businesses within the industry and have extensive experience with all the available IT options. The consultant performs all the systems analysis tasks, including preparing a feasibility study. They then liaise with suppliers and development companies during the design and implementation phases. Often the extra cost involved to hire such consultants is more than returned through higher quality systems that better meet requirements.

Contracting and outsourcing, although similar in some respects, do have some fundamental differences. When an outside organisation is contracted they perform their tasks under the direct management and control of the contracting organisation. Outsourcing is different; it involves passing control for the entire process over to the outsourced company. When development tasks are outsourced the requirements and a time for completion are negotiated in advance – the project management and development approaches are determined and controlled by the outsourced company. For example, software development is often outsourced to offshore companies. The offshore company receives detailed requirements, however they design the software and also project manage its development.

GROUP TASK Discussion
Currently many products, including IT hardware, are manufactured in China and many software applications are developed in India. Identify and discuss reasons why such offshore outsourcing is now common.

Prototyping

Earlier in this chapter we discussed requirements prototypes, whose main aim is to verify and determine the requirements for a new system. The prototyping approach extends the use of such requirements prototypes such that they evolve to a point where they actually become the final solution or they become sufficiently detailed that they can be used to present the concept for full scale development. Furthermore, concept prototypes, as they are accurate simulations of the final system, become an essential part of the requirements for the new system.

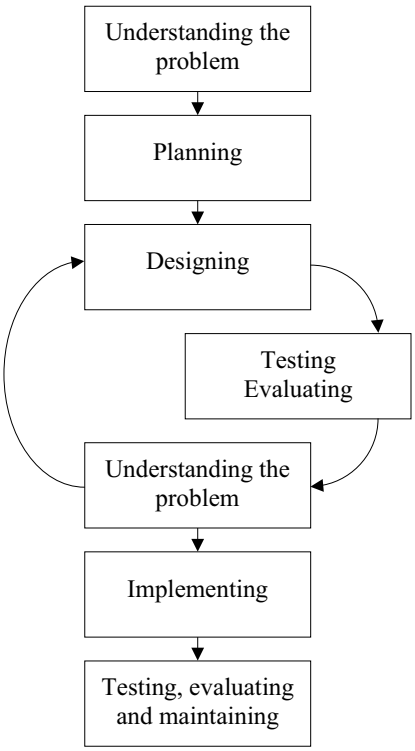


Fig 1.20
Prototyping system development approach.

The diagram in *Fig 1.20* describes the phases occurring when a prototype evolves into the final solution. Notice the loop containing "Designing", "Testing, Evaluating" and "Understanding the problem". Each iteration through this loop produces an enhanced prototype that meets more of the systems requirements. Indeed new or modified system requirements are determined during each "Understanding the problem" phase. After many iterations the prototype reaches a stage where the problem is sufficiently well understood, which means it successfully meets its requirements and is now ready for implementation.

Prototyping acknowledges that many system requirements cannot be determined precisely until development is underway. During each "Understanding the problem" phases users, participants and other stakeholders are able to view the prototype and suggest modifications and additions. Therefore as the prototype evolves, so too does the system's requirements. Clearly this is an enormous advantage in terms of the system meeting the needs of those for whom it is designed. However, it can also lead to "blow outs" in the scope of projects. Users will think of new functionality as they view a working prototype that they would not initially have considered. Project management techniques are required to ensure such issues do not cripple the project. In particular management strategies are needed to ensure the project remains within budget and time constraints. It is often wise to prioritise requirements, such that necessary requirements are met prior to less critical requirements. If time and/or money run low then the system can still be implemented that meets all necessary requirements.

The prototyping approach is particularly well suited to the development of the software components of information systems. Ongoing feedback from users and participants can be incorporated into the solution or the concept prototype during each iteration. If the prototype will evolve into the final solution then the tools used to design and create the software must be able to accommodate "on the fly" changes and must also be appropriate for final implementation. For large and/or critical systems the performance, reliability and quality requirements mean this is often not possible. However for smaller less critical systems rapid application development (RAD) tools, such as visual programming environments, or even customised versions of standard applications are quite able to produce software of sufficient quality and performance.

GROUP TASK Discussion
Distinguish between Requirements Prototypes, Concept Prototypes and Prototypes that evolve into the final solution. Brainstorm and discuss examples of information systems where each type is appropriate.

Customisation

For many new information systems it is economically unviable to develop a completely new system. Instead an existing system is customised to suit the specific needs and requirements of the new system. In reality most business systems are customised versions of existing systems. For example, virtually all Hotels across the globe use one of only a handful of commercially available software and hardware systems. One of these systems is selected and customised to suit each hotel's specific requirements. For example a small hotel likely has a single restaurant and a single bar, whilst larger hotels contain many restaurants and bars.

Customisation may involve alterations to system settings within the hardware and software or it may involve underlying customisation of the actual hardware or software itself. For instance an "off the shelf" server could be customised by adding extra RAM or installing a RAID storage device. Standard applications, such as word processors, spreadsheets and databases, can be customised to perform new functions. Existing software applications can also have their source code modified to implement custom features. Often mass produced information technology is able to meet the large majority of the system's requirements. Tweaking and modifying such products is generally much more cost effective compared to developing from scratch.

Consider the following systems:

- A school has analysed various commercially available software solutions for producing reports to parents. One software package almost meets their needs however in its present state it is not able to produce summary tables specifying the number of students in each performance band for each course.
- A department store has decided to invest in a particular point of sale (POS) system. This system includes terminals where the keyboards are integrated with the cash drawers. For most departments this is fine, however various food departments have requested waterproof keyboards with larger keys.
- A warehouse is developing a new automated vehicle picking system based on commercially available automated forklifts. The computer controlled forklift vehicles retrieve pallets of product from storage and deliver them to the existing computer controlled conveyor and packing system. The software that controls the new forklift system is unable to interface with the existing conveyor and packing system.
- A courier company currently uses a two-way radio system to communicate with their drivers. They have decided to introduce a new commercially available information system for allocating courier jobs to drivers. The new system sends messages to driver's mobile phones but not two-way radios.

GROUP TASK Discussion
Identify aspects of the development of the above systems that could be customised to meet the system's requirements. Discuss other development strategies that could be used instead of customisation.

Participant Development

The participant development approach simply means that the same people who will use and operate the final system develop the system. As the users and participants are the people who largely determine the requirements there is little need to consult widely. Although this will no doubt speed up development considerably there are of course numerous disadvantages that can have the opposite effect. Firstly the user must have sufficient skills to be able to create the system and secondly they must understand the extent of their skills. Sometimes a little technical knowledge can be worse than no knowledge at all. With most information systems the extent of technical know-how required is not obvious until well into the design stage. All too often it's the small detail that takes time, skill and experience to complete. In general, user developed systems will be of lower quality than those developed professionally.

So what types of project are suited to user development? Systems that will only be used by the developer/user and perhaps a few other people are often suitable candidates. There is no need for detailed documentation – the developer is always on hand to answer questions and even make modifications. If the system can be developed using common software applications that include reusable and quality components then the project has a higher chance of success. For instance a spreadsheet program could be used to create a template for a teacher's mark book. The developer/user requires skills with regard to designing formulas, however more advanced features such as securing the resulting spreadsheet files or validating input can be left out. The solution will meet the user/developers requirements but is unlikely to be suitable for commercial distribution. Such detail and quality issues are a feature of most user-developed systems. They perform the processing they must perform with no extra bells and whistles.

End user or participant development has many advantages for small business and home users who would not otherwise be able to afford a professional solution. They are able to automate functions themselves and are then able to modify the solution as new requirements emerge.

Consider the following systems:

- Thomas operates a used car yard. Currently he completes all paper work manually, however this is becoming unmanageable as the business grows. He has decided records of each vehicle in his yard together with payroll functions need to be automated. Thomas already lists each of his vehicles on a number of websites; therefore having each vehicle's details in electronic form will greatly simplify uploading this data to the web.
- Bethany operates a home business selling products using eBay. She imports product in bulk lots from various overseas suppliers and lists them individually on eBay. Bethany already uses an open source software product to list items and automatically create and send invoices to customers. She wishes to track stock levels of each product from the time she orders them from her supplier, however she would like her stock control system to interface with the her existing invoicing system. Each time an invoice is generated the stock level of each product sold should reduce automatically.
- Stuart and Jennifer operate a water carting business. They have a number of contracts with local councils, whereby they supply water on a "as required" basis. They would like to track just how many loads of water are actually delivered so they can determine the actual costs associated with servicing each contract.

GROUP TASK Discussion
Outline the skills required to develop Thomas, Bethany, and Stuart and Jennifer's systems. For each system, do you think participant development is a suitable system development approach? Discuss.

Agile Methods

Agile development methods have emerged in response to the "adhoc" reality of many software development projects. They place emphasis on the team developing the system rather than following predefined structured development processes. Agile methods remove the need for detailed requirements and complex design documentation. Rather they encourage cooperation and team work. Agile methods are particularly well suited to web-based software development and other software applications that are modified regularly such that they evolve over time. In general, agile methods are for developing software rather than total information systems.

Typically quite small teams of developers are used. It would be unusual for an agile team to have more than about half a dozen members. It is preferable for one team member to be a knowledgeable and experienced user or participant. Small teams are better able to share ideas and work on solutions together. Larger teams tend to break into smaller groups – for agile methods to be a success everyone must be an equal member with a clear shared purpose.

Let us work through the activities described in *Fig 1.21*. Initially the general nature of the problem is determined and the development team is formed. The team first meet to create a basic plan and a general design for the software – only minimal detail at this stage, just enough to get started. The basic idea is to only plan, design and document details as they're actually needed. Often a simple whiteboard is used to sketch out the general design. The team then gets straight to the task of creating an initial solution. As this occurs they informally consult and negotiate with each other. The user team member is always present to answer questions, make suggestions and generally ensure the solution will be workable in practice.

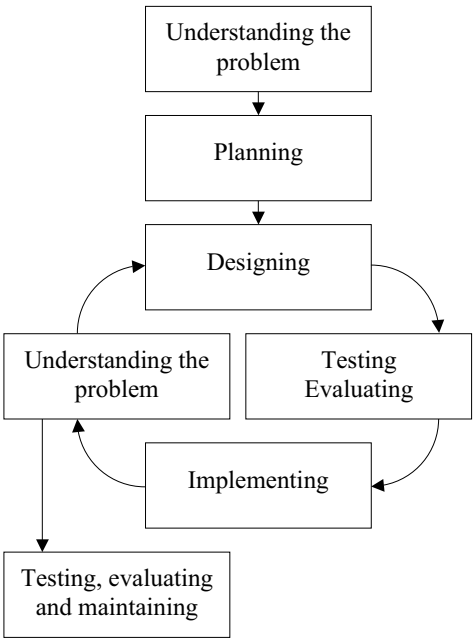


Fig 1.21
Agile system development approach.

Once an initial, yet simplified, solution is produced it is immediately tested, evaluated and then implemented. This means the solution is actually being used by real users and participants – usually the client but it could be a sample of users or even globally to all users via the web. The users see exactly what has been achieved, can provide feedback and make suggestions about further additions. In effect we have entered a new mini "Understanding the problem" phase. The team again meets informally to discuss the next part of the design. The design incorporates feedback from users together their own ideas. They then go straight to work coding this next part of the design. The solution is again thoroughly tested and evaluated before being implemented. This process repeats many times with each iteration implementing further functionality and detail. Typically a single iteration takes weeks or even just days. Each design meeting is short, maybe just an hour or so, whilst coding and testing consumes the majority of the development time.

When developing software it's all the minute details that combine to form the total solution. Agile methods are a response to the reality that intricate details are difficult to specify accurately in advance. Each part of a software solution relies heavily on many other related parts. Until the related parts exist, it is wasteful to continue designing. Much of the design will prove unworkable and will need to be redesigned or significantly altered. Compare this to the traditional approach where specific and intricate detail is created well in advance.

One significant issue with agile methods is how to construct agreements when outsourcing the development. Traditionally a strict set of detailed requirements, together with the total cost and time for completion is negotiated. When using agile methods no detailed requirements exist – they emerge during development. A common solution to this dilemma is to fix the budget and time and allow the requirements to change. Once the budget and time is exhausted then the current solution becomes the final solution. To enter into such agreements requires significant trust to be established between the client and developer. The client stands to gain, as they are heavily involved throughout the development process and hence are more likely to receive a final product that better meets their actual and current requirements.

Consider the following situations:

- Google, Yahoo and other search engine companies continually update their systems. This includes both the software and also the data and its underlying organisation.
- Currently most operating systems, and in particular Microsoft Windows, are regularly updated via automatic download to add new functionality and also to overcome security flaws.
- Large businesses commonly employ their own teams of information system developers. These teams are continually working to fulfil new and changing user requirements.
- Small businesses and even individuals regularly modify their websites. Once the new site has been uploaded to their web server it is immediately operational for all end-users.
- A company has decided to create a new information system. They already have a team of developers, however the existing team is comprised of members with different specific skills and no agile development experience.

GROUP TASK Discussion

Critically analyse each of the above situations in terms of its suitability for development using agile methods. Identify any issues that should be addressed if agile development is to be a success.

DETERMINE HOW THE PROJECT WILL BE MANAGED AND UPDATE THE REQUIREMENTS REPORT

Once a particular solution has been identified and a suitable development approach has been determined then sufficient information is available to determine suitable techniques and strategies for managing the project's development. Furthermore, the Requirements Report can be updated with specifics in regard to the chosen solution and to reflect the selected system development approach. We discussed various project management tools in detail earlier in this chapter, namely:

- Gantt charts for scheduling of tasks.
- Journals and diaries for recording the completion of tasks and other details.

- Funding management plan for allocating money to tasks.
- Communication management plans to specify how all stakeholders will communicate with each other during the development of the new system.

How these tools are used will depend on the system development approach and also on the development needs of the specific solution. The chosen solution allows the project manager to identify and take account of the new system's participants, information technology, data/information and of course the needs and requirements of its users as they formulate a strategy for managing the project. In addition to creating project management tools, the Requirements Report is also updated appropriately to document these areas based on the specifics of the system being developed.

Areas likely to affect project management decisions and that are commonly documented within the Requirements Report include:

- Participants should be identified, and in particular mechanisms for obtaining their feedback, should be considered. This is not an issue for participant developed solutions. If using agile methods then consider including a knowledgeable system participant on the team. For other development approaches regular sessions, meetings or other forms of communication should be planned and documented in advance to ensure ongoing, regular communication takes place.
- Information technology, which includes the hardware and software for the new system should be identified. This must be purchased or its development planned. In most information systems hardware is purchased with minor modifications made to suit the particular solution. How such purchases are to be made must be specified. Perhaps a number of quotes must be obtained followed by further negotiation with suppliers. Developing the software is often the most significant development task. If it is to be outsourced then agreements will need to be negotiated – funding management plans should specify when, how and under what terms such agreements are to be made. Clearly software developed using other approaches will require extensive planning, management and probably the addition of new requirements throughout the system development lifecycle.
- Data is the input into the new system and the information is the output. It is likely that some data will be sourced from other systems, whilst other data is entered directly into the system by users. In either case sample or existing real data will be needed for testing during the design and subsequent development stages. How and when this data is to be obtained should be documented. The information the system produces is what ultimately meets the system's purpose. Samples that identify the precise nature of this information are a valuable resource. The ability of the system to produce particular information is often the primary means for verifying that system requirements have indeed been met. Many development tasks aim to produce specific information. In essence each piece of information helps determine the development subtasks, in terms of project management it also determines how each task should be scheduled and costed. Furthermore successful completion of subtasks is clearly signalled when the information can be accurately produced.
- Meeting the needs and requirements of users is the aim of all successful information systems. If a traditional system development approach is being used then these requirements will have already been established. For other approaches, in particular the iterative prototyping and agile methods, user needs and requirements emerge and change development progresses. When using such iterative approaches project management techniques and the Requirements Report must be flexible, whilst maintaining control of cost and time constraints. Ongoing, regular and meaningful communication between developers and users is essential.

Details of the specific communication strategies and techniques are specified within the communication management plan. Journals and diaries are used to document each communication and the funding management plan will detail the mechanisms for reallocating money as requirements emerge and change.

Consider Pet Buddies Pty. Ltd.

Fred's feasibility report strongly recommends solution option B (refer page 50 – 51) and Iris and Tom agree. Fred will negotiate the purchase of all required hardware and also the voice mail software. He will also upgrade and modify Pet Buddies existing database to suit the requirements of the new system. Development of the speech recognition software will be outsourced to a specialist software development company. Fred feels a traditional approach should be used by the outsourced specialist, as the software does not interact directly with users; rather it obtains all input from the audio files in the database and outputs text files back to the database. Fred now has sufficient information to create a workable schedule including each of the project's subtasks.

GROUP TASK Discussion
With reference to Pet Buddies Solution Option B (page 51), identify the major tasks Fred needs to include on a Gantt chart for the project. Discuss a suitable sequence for completing these tasks.

GROUP TASK Discussion
Is a single system development approach appropriate for developing the new Pet Buddies systems? Discuss various alternatives.

GROUP TASK Discussion
Why would Fred choose to outsource development of the speech recognition software? Do you agree that a traditional development approach is suitable for developing this software? Discuss.

HSC style question:

The Australian federal government is considering implementing a new system for doctor's patients to claim their Medicare rebates for doctors who do not bulk bill their patients.

The existing system works in the following way:

- a patient goes to their doctor for a consultation
- the patient receives an account (bill) from their doctor
- the patient pays the doctor's account and receives a receipt
- the patient takes (or posts) their receipt to a Medicare office
- the receipt is processed by the Medicare office
- the patient receives a rebate (partial re-imbursement) from Medicare.

The new system would amend the current system so that the doctor's surgery would be connected via a Wide Area Network to the Medicare office and as a result the processing of the account would occur at the surgery directly following the payment of the account. Patients would receive their rebate by direct deposit from Medicare into their bank account immediately after the account has been paid.

(a) Describe THREE specific issues that should be considered when assessing the feasibility of the new system.

(b) Assuming the new system is to be developed, recommend and justify a suitable system development approach.

Suggested Solution

(a) No doubt there are a large variety of different billing software packages used by different doctors and some doctors may still use manual systems. How will the new system interface with such a broad range of systems? Is it technically feasible for such a large and diverse range of systems to be accommodated?

The new system removes work from Medicare offices and also from the end-user patients. Essentially this work is transferred to the Doctor's surgery staff (and also the new software). There are no direct advantages for the Doctor's surgeries and hence they are unlikely to embrace the new system. This could result in operational problems, as the primary participants will be resistant to changes brought about by implementing the new system.

Each Doctor's surgery throughout the country will require a secure communication link and associated communication hardware. Purchasing and installing this equipment will be costly. However perhaps more significant will be the ongoing maintenance of the network and hardware. Although Medicare offices will require less staff to process rebates, more technical staff will need to be employed. Such issues will affect the economic feasibility of the new system.

(b) The communication network software and hardware would be best developed using a traditional structured approach. The hardware at each Medicare office and at each Doctor's surgery can be largely of the same design. Because there are no doubt thousands of Doctor's surgeries and hundreds of Medicare offices it is worth the effort to ensure the system is as reliable and secure as is possible. Furthermore the requirements for the network information technology can be specified in advance and only limited technical user interaction is required.

The software to interface with the new system and the account systems used by Doctor's surgeries could be developed using a prototyping approach. Each completed prototype can be sent for testing and feedback to sample Doctor's surgeries and also to software companies that develop software for Doctor's surgeries – in effect these are the actual people most affected. In this way the prototypes can be modified so they evolve in response to feedback and the software companies can modify and also verify that their products will operate with the new Medicare system.

Comments

- In an HSC or Trial HSC Examination both parts (a) and (b) would likely attract 3 to 4 marks each.
- In part (a) there are numerous other issues that could be discussed. For instance, ongoing training and support for new surgeries and surgeries that change their billing systems. The system requires patients to have a bank account and to be willing to have the account details within the system – some patients may have privacy concerns. Under the previous system patients could visit Medicare to obtain their rebate prior to paying the bill, under the new system patients must pay the account first, which requires them to have sufficient funds available.
- In part (b) a number of different system development approaches could legitimately be recommended and justified. It is likely that better responses would combine a number of development approaches to form a system development approach tailored to the development needs of this specific system.

SET 1C

1. Cost-benefit analysis is part of assessing each solution's:
   (A) technical feasibility.
   (B) economic feasibility.
   (C) schedule feasibility.
   (D) operational feasibility.

2. The ability of participants to effectively use new information technology is part of assessing each solution's:
   (A) technical feasibility.
   (B) economic feasibility.
   (C) schedule feasibility.
   (D) operational feasibility.

3. Determining whether a solution can be developed within the available time is part of assessing each solution's:
   (A) technical feasibility.
   (B) economic feasibility.
   (C) schedule feasibility.
   (D) operational feasibility.

4. Support from users and participants for each solution is considered when assessing each solution's:
   (A) technical feasibility.
   (B) economic feasibility.
   (C) schedule feasibility.
   (D) operational feasibility.

5. Altering an existing solution occurs when using which development approach?
   (A) Agile.
   (B) Outsourcing.
   (C) Prototyping.
   (D) Customisation.

6. Using outside specialists to develop all or part of the solution is known as:
   (A) Customisation.
   (B) Prototyping.
   (C) Outsourcing.
   (D) Agile methods.

7. System development methods that acknowledge the changing nature of requirements during development include:
   (A) prototyping and customisation.
   (B) prototyping and agile methods.
   (C) traditional and agile methods.
   (D) outsourcing and customisation.

8. Which approach does NOT require detailed user documentation to be produced?
   (A) Traditional approach.
   (B) Prototyping approach.
   (C) Participant development approach.
   (D) Agile approach.

9. Planning and designing just before the solution is created is a characteristic of:
   (A) agile methods.
   (B) traditional system development.
   (C) customisation.
   (D) outsourcing.

10. Each stage of the SDLC is completed in sequence when using which system development approach?
    (A) Traditional.
    (B) Prototyping.
    (C) Outsourcing.
    (D) Participant development.

11. Define each of the following.
    (a) feasible (b) deadline (c) payback period (d) NPV

12. Outline factors affecting a solution's:
    (a) economic feasibility (c) operational feasibility
    (b) technical feasibility (d) schedule feasibility

13. List characteristics of each of the following development methods.
    (a) Traditional (c) Prototyping (e) Participant development
    (b) Outsourcing (d) Customisation (f) Agile methods

14. Contrast the traditional system development approach with:
    (a) prototyping (b) agile methods

15. During the planning stage the feasibility study is completed, then the most appropriate solution selected, followed by determining a suitable system development approach and finally planning how the project will be managed and updating the Requirements Report.
    Discuss reasons why these activities are performed in this particular sequence.

DESIGNING

This third stage of the system development lifecycle (SDLC) is where the actual solution is designed and built. This includes describing the information processes and specifying the system resources required to perform these processes. The resources used by the new information system include the participants, data/information and information technology (see *Fig 1.22*). Information technology includes all the hardware and software resources used by the system's information processes. Some new information systems may require completely new hardware and software, whilst others may utilise existing hardware and software to perform new information processes – in fact any combination of new and existing information technology is possible, it depends on the requirements of the new system and the needs of its information processes.

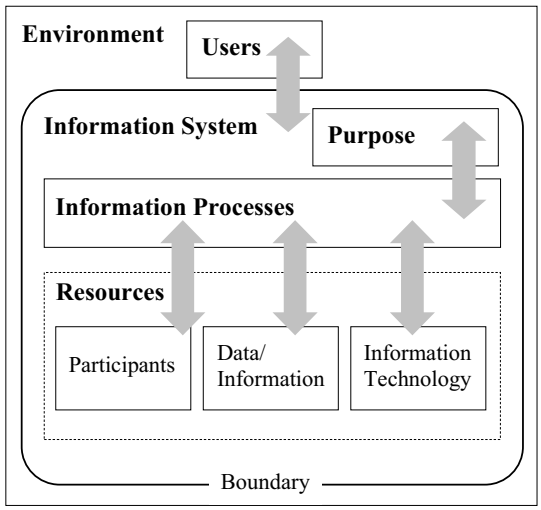


Fig 1.22
Diagrammatic representation of an information system.

The design process will differ according to the system development approach used. However for all approaches designing involves identifying and describing the detail of the new system's information processes. System models are created, using tools such as context diagrams, data flow diagrams, decision trees and tables and also storyboards. During the modelling process, the data and information used and produced by the system is determined and clearly defined within a data dictionary. Once the processing and data/information is understood the particular information technology that will perform these processes can be accurately determined. Depending on the individual system and the selected development approach, it may be necessary to have new software developed, existing software modified or specific hardware components assembled. Furthermore, specifications and suppliers for required outside communication lines, network cabling, furniture, off-the-shelf software and standard hardware are determined in preparation for negotiating their purchase and/or installation. Agreements with regard to outsourced development should be finalised early so that their development can progress. Hardware or software that will be customised will need to be purchased in advance. Throughout the entire design process consultation with both users and participants should be ongoing. It is essential that the needs and concerns of all people affected by the final operational system remain central to the design process.

GROUP TASK Discussion
Discuss techniques appropriate to different system development approaches that ensure user and participant's needs and concerns are not overlooked during the design stage.

GROUP TASK Discussion
Precisely when detailed system models are required varies depending on the system development approach. Discuss such differences with particular reference to the traditional, prototyping and agile approaches.

SYSTEM DESIGN TOOLS FOR UNDERSTANDING, EXPLAINING AND DOCUMENTING THE OPERATION OF THE SYSTEM

The vital link between all the system's resources is the information processes, which will operate within the new system. Describing the detail of such processes is critical to all aspects of the design – including hardware purchases. As a consequence detailed models of the solution should be produced. In this course we examine a variety of design tools, namely context diagrams, data flow diagrams, decision trees and tables, data dictionaries and storyboards. It is vital to understand how to create, read and use the tools, as they will be utilised numerous times throughout the remainder of this course. In this section we introduce each tool with emphasis on their use as tools to assist in the design of new systems. In future chapters they will also be used to assist in understanding and explaining the operation of numerous existing systems.

Context Diagrams

Context diagrams represent the entire system as a single process. They do not attempt to describe the information processes within the system; rather they identify the data entering and the information leaving the system together with its source and its destination (sink). The sources and sinks are called "external entities". As is implied by the word external, these entities are present within the system's environment. Context diagrams are really top-level data flow diagrams and are often known as level 0 data flow diagrams.

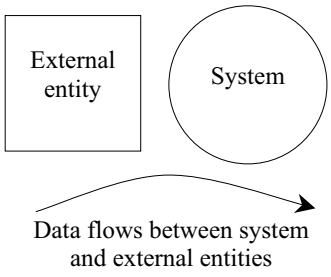


Fig 1.23
Symbols used on context diagrams.

Squares are used to represent each of the external entities. Common examples of external entities include users, other organisations and other systems. These entities are not part of the system being described, as they do not perform any of the system's information processes. Rather the system acquires (collects or receives) data from each source entity and/or the system supplies (displays or transmits) information to each sink entity. The entire system is represented using a circle, with labelled data flow arrows used to describe the data and its direction of flow between the system and it's external entities. Data flows from each source into the system, and data (information) flows from the system to each sink.

Each data flow label should clearly identify the nature of the data using simple clear words. Remember each data flow describes data not a process, for example if a user enters a password then an appropriate data flow label would be "User password", not "Enter password". Furthermore in this example each user is the source of a single password, so "User password" is a more appropriate label than "User passwords". If many data items flow together then a plural label would be more appropriate, however in most systems this is a rare occurrence.

The systems participants require special consideration as they are part of the system – participants are a special class of user who carry out the information processes within the system. As participants are part of the system they are not automatically included as external entities. It is only when the participants also supply the system with data or receive information from the system that they become external entities – in essence they are also acting as more general users. For instance, within the new Pet Buddies system Iris and Tom are clearly participants, they initiate and perform many information processes. However Iris and Tom view the draft activity reports, make edits to these reports and approve each activity report. It is often helpful to try to separate data and processes within your mind. The system displays (process) each

draft activity report (data) to Iris and Tom hence they are a sink. The system collects (process) edited activity data (data) and approval for activity reports (data) from Iris and Tom hence they are also a data source. All data entering the system and all data (information) leaving the system must be included on the context diagram. All processes performed by the system are part of the single system circle and are not detailed on the context diagram.

So how does a context diagram assist the design process? Context diagrams indicate where the new system interfaces with its environment. They define the data and information that passes through each interface and in which direction it travels. Descriptions of this data and information is further detailed within a data dictionary. Ultimately the data entering the system from all its sources must be sufficient to create all the information leaving the system to its sinks.

Consider Pet Buddies Pty. Ltd.

Recall that solution option B (refer page 51) has been accepted. Fred is now commencing work on the design of the new activity report creation system. He has developed the context diagram reproduced in *Fig 1.24* below.

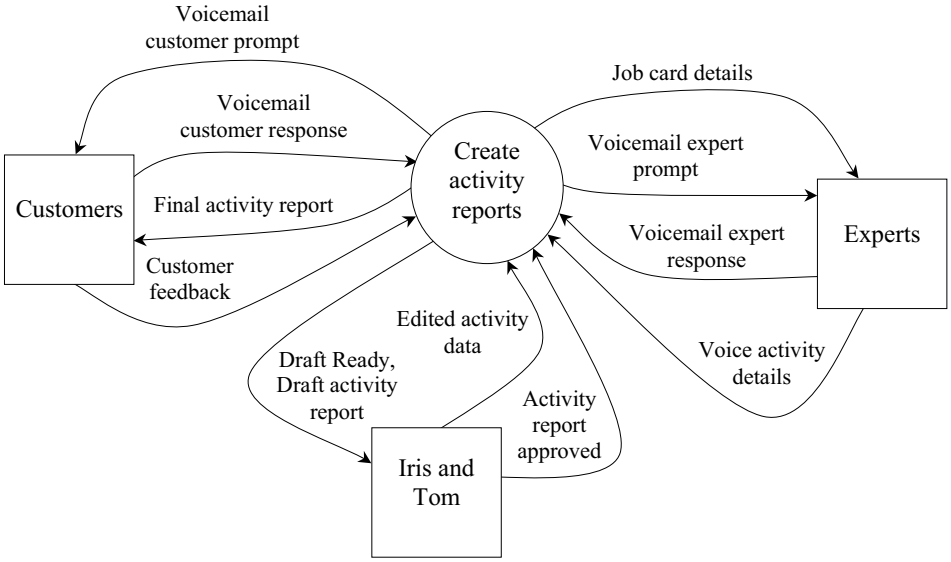


Fig 1.24
Context diagram for Pet Buddies new information system.

GROUP TASK Discussion
Analyse the above *Fig 1.24* context diagram in relation to the Option B solution outline on page 51.

Data Dictionaries

Data dictionaries are used to detail each of the data items used by the system. They are tables where each row describes a particular data item and each column describes an attribute or detail of the data item. Clearly the name or identifier given to the data item must be included, together with a variety of other details such as its data type, storage size, description and so on.

Data dictionaries are often associated solely with the design of databases where they are used to document details of each field. Commonly such details include at least the field name, data type, data format, field size, description and perhaps an example.

However data dictionaries are also used in conjunction with many design tools. For instance a data dictionary can be used to specify details of each data flow used on context and data flow diagrams. The details specified for each data item should be selected to suit the purpose for which the data dictionary is created. Context diagrams describe an overall view of the system and hence specifying the data type, a description and perhaps an example will likely suffice. When designing a database much more detailed specifications are needed, including the previously mentioned details and possibly other additional detail such as data validation, default value, whether it is key field and so on. Software developers also use data dictionaries to document all the variables and data structures within their code.

Consider Pet Buddies Pty. Ltd.

Fred has created the following data dictionary to document his context diagram.

Data Flow Name	Media/Data type	Description
Job card details	Hardcopy text	A printed report containing the customer's details and the activities to be completed by the expert during each home care visit.
Voicemail expert prompt	Analog Audio	Synthesised voice used to prompt expert for input.
Voicemail expert response	Numeric	Response from expert entered using telephone keypad.
Voice activity details	Analog Audio	Analog voice recording via expert's telephone.
Draft ready	Boolean	Used to alert Iris and Tom that a draft activity report is waiting for editing and approval.
Draft activity report	Digital Audio	Digital recording of a total activity report prior to its approval.
Edited activity data	Digital Audio	Voice recording from Iris or Tom to replace portions of the draft activity report.
Activity report approved	Boolean	Approval for activity report to be made available to the customer.
Voicemail customer prompt	Analog Audio	Synthesised voice used to prompt customer for input.
Voicemail customer response	Numeric	Response from customer entered using telephone keypad.
Final activity report	Analog Audio, or Facsimile	The final activity report received by the customer. Could be over the telephone or could be a faxed version created by the speech recognition engine and associated software.
Customer feedback	Analog Audio	Analog voice recording via customer's telephone.

Fig 1.25
Data dictionary accompanying Pet Buddies' context diagram.

GROUP TASK Discussion
With reference to the *Fig 1.24* context diagram, identify the outputs from the new Pet Buddies system. Analyse each output to determine the inputs that are processed by the system to produce each of these outputs.

GROUP TASK Discussion
Describe the nature of the interfaces between the system and each of the three external entities. Refer to both the context diagram (*Fig 1.24*) and the data dictionary (*Fig 1.25*) to justify your responses.

Data Flow Diagrams (DFDs)

DFDs do not attempt to describe the step-by-step logic of individual processes within a system. Rather they describe the movement and changes in data between processes. As all processes alter data then the data leaving or output from a process must be different in some way to the data that entered or was input into that process. This is what all processes do; they alter data in some way. The aim of DFDs is to represent systems by describing the changes in data as it passes through processes. For example a process that adds up numbers receives various numbers as its input and outputs their sum. On DFDs there is no attempt to describe how the numbers are summed. Rather the emphasis is on where the numbers come from and where the sum is headed.

To represent the data moving between processes we use labelled data flow arrows. The label describes the data and the direction of the arrow describes the movement. Processes are represented using circles. The label within the circle describes the process. As processes change data the labels used should imply some action – verbs should be used, such as create, update, collect, to emphasise that some action is performed.

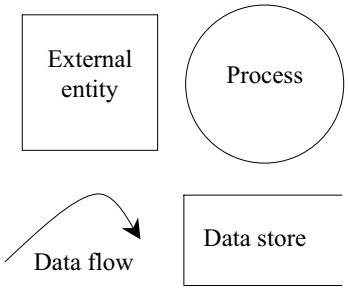


Fig 1.26
Symbols used on data flow diagrams.

The final symbol used on DFDs represents data stores. A data store is where data is maintained prior to and after it has been processed. In most cases a data store will be a file or database stored on a secondary storage device, however it could also be some form of non-computer storage such as a file within a filing cabinet. An open rectangle together with a descriptive label is used to represent data stores. Data stores allow the system to pause or halt between processes and they also allow processes to occur in different sequences and at different times. In effect processes are freed to execute independently of each other. Consider a typical process that collects data from a user and stores it within a data store. This single process can execute many times simultaneously whilst at other times it sits idle. The data is maintained within the data store where it can be retrieved and used by other processes when and as they require.

Context diagrams are top-level data flow diagrams –also called level 0 DFDs. They specify all external entities with the complete system represented as a single process. A level 1 data flow diagram expands this single process into multiple processes. A series of Level 2 DFDs are drawn to expand each level 1 DFD process into further processes. Level 3 DFDs similarly expand each level 2 process and so on. A series of progressively more and more detailed DFDs refine the system into its component sub-processes. Eventually the lowest level DFDs will contain processes that can be solved independently. Breaking down a system's processes into smaller and smaller sub-processes is known as 'top-down design'. The component sub-processes can be solved and even tested independent of other processes. Once all the sub-processes are solved and working as expected they combine to form the complete solution.

On some level 1 and lower-level DFDs the external entities are included, whilst on others they are not. If a context diagram has already been produced or external entities have been included on a higher-level DFD then it is common practice to omit the external entities from the derived lower-level DFDs. A similar practice is also true for data stores, however in the interest of improved clarity it is more common to reproduce data stores on lower-level DFDs. To improve clarity it is also permissible to include the same external entity or data store multiple times within the same DFD.

For instance in *Fig 1.27* below the "Widget Sales Team" entity is included twice simply to improve readability. This DFD could easily be reformatted using a single "Widget Sales Team" entity with both data flows attached.

On most DFDs the processes are numbered in addition to their labels. Consider the example level 1 DFD in *Fig 1.27* – it contains the three processes, 1. Filter sales records, 2. Calculate widget statistics and 3. Produce widget sales graphs. Three level 2 DFDs would then be produced – one for each process in the level 1 DFD. *Fig 1.28* shows an expansion of process 2. Calculate widget statistics into a level 2 DFD containing four processes. These four processes are numbered from 2.1 to 2.4 – the 2 indicating their connection to process 2 on the level 1 DFD. If process 2.1 required further expansion into a level 3 DFD then its processes would be numbered 2.1.1, 2.1.2, 2.1.3 and so on.

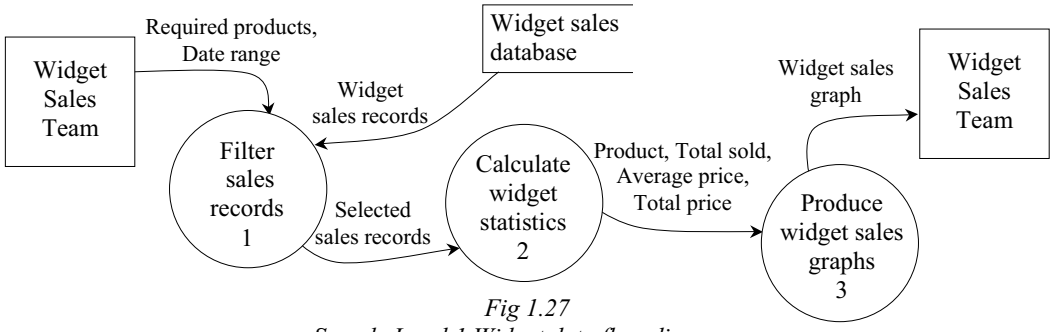


Fig 1.27
Sample Level 1 Widget data flow diagram.

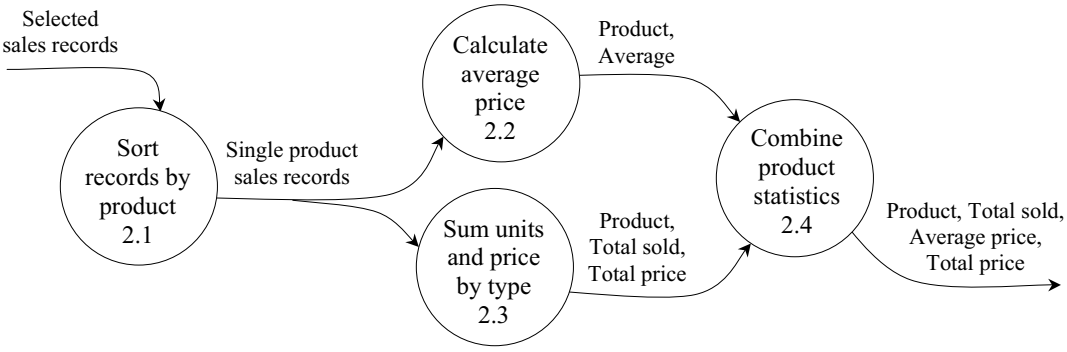


Fig 1.28
2. Calculate widget statistics DFD.

Consider the following DFD summary points:

- All processes must have a different set of inputs and outputs.
- All lower-level DFDs must have identical inputs and outputs as the higher-level process they expand.
- External entities and data stores can be reproduced on lower-level DFDs.
- External entities must be present on context diagrams (level 0 DFDs) but are optional on lower-level DFDs.
- A single output data flow can be the input to multiple other processes.
- Labels for processes should include verbs that describe the action taking place.

GROUP TASK Activity
Identify examples within *Fig 1.27* and *Fig 1.28* above that illustrate each of the above dot points.

Consider Pet Buddies Pty. Ltd.

Fred has further refined the context diagram in *Fig 1.24* into the more detailed level 1 DFD reproduced below in *Fig 1.29*. Within the DFD Fred has deliberately split the system into four independent processes. Once operational each of these processes can occur at different times or they could occur at the same time. For instance, process 1 outputs "Draft ready", which is used to alert Iris and Tom via a message displayed on their screens that an activity report is awaiting approval, however there is no requirement that they respond to this message and complete process 2 immediately.

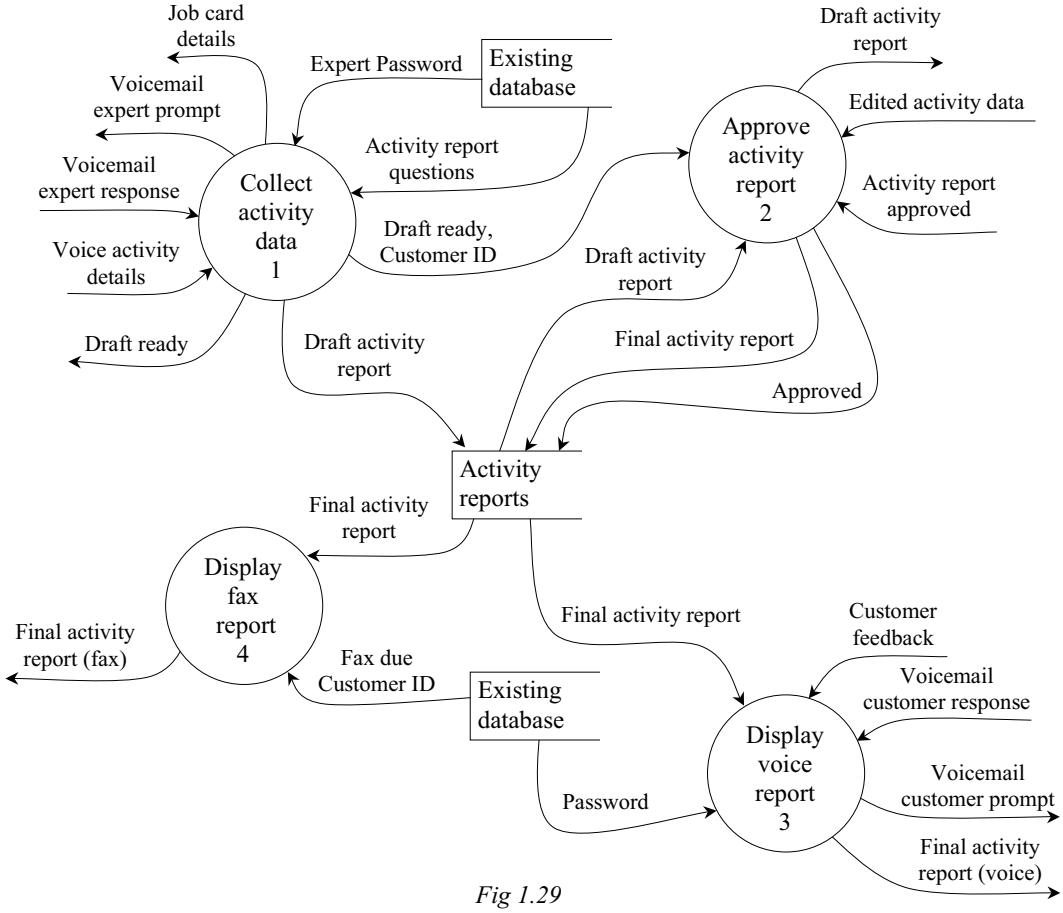


Fig 1.29
Pet Buddies level 1 DFD.

Process 1 and 3 will be performed by the voicemail application. Essentially process 1 involves the expert recording their responses to each question in the activity report. The resulting audio files are stored in the customer's mailbox within the activity reports data store. At this stage they are marked as drafts – process 2 approves these drafts. In process 3 customers essentially access their mailbox and listen to their messages, which are the final voice activity reports. Process 4 periodically checks for any fax reports that are due. When such a report is identified it is converted to text and faxed to the customer.

As the voice mail software operates using multiple phone lines, it is possible for multiple experts and customers to be using the system at the same time. That is, both process 1 and 2 can be executing simultaneously multiple times.

Process 4 requires the digital audio files to be converted into text and then faxed. The development of the software for performing this process is to be outsourced to a

specialist software developer – the software developers will work through their own version of the software development cycle. Process 4 can be performed manually without affecting the operation of the other processes. This means its completion will not affect the scheduled implementation date.

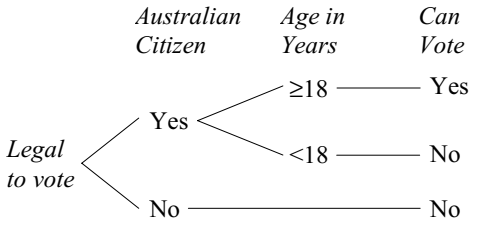
GROUP TASK Activity
The data dictionary for the DFD in *Fig 1.29* has not been reproduced. Split your class into 4 groups. Each group is to produce a data dictionary for 1 of the 4 processes.

GROUP TASK Discussion
There are two unique data stores on the DFD in *Fig 1.29* – the existing database is included twice merely for clarity. Describe the data held in each of these data stores.

GROUP TASK Discussion
Identify and discuss aspects of the above design that ensure the system is 'human centred' rather than 'machine centred'.

Decision Trees and Decision Tables

Decisions are made when one alternative is chosen from a range of possible alternatives. In terms of information systems, each of the alternatives results in some action or process being performed. Decision trees and decision tables are tools for documenting the logic upon which decisions are made. They describe a strict set of rules where each rule leads to a particular decision alternative or action. Each rule is composed of one or more conditions that must be satisfied for the rule to be true. For example, if you are an Australian citizen and you are 18 years or older then you can vote. This rule contains two conditions namely; "Australian citizen" and "18 years or older" and the single action "can vote". In this rule both conditions must be True for the action to take place. We could produce further rules for when either or both conditions are false, which would result in the action "can NOT vote".



Conditions	Rules			
Australian Citizen	×	✓	×	✓
18 years or older	×	×	✓	✓
Actions				
Can vote	×	×	×	✓
Can NOT vote	✓	✓	✓	×

Fig 1.30
Sample decision tree (left) and decision table (right).

Decision trees represent the rules, conditions and actions as a diagram whilst decision tables use a two-dimensional table. In a decision tree each unique left to right sequence of branches represents a complete rule. Each rule results in some action – the actions are listed at the right hand side of the branches. In *Fig 1.30* there are three branch sequences and therefore three rules. Notice in this example there is no need to evaluate the "Age in Years" when "Australian Citizen" is already known to be False; hence three rather than four rules are required. This is possible because decision trees document the particular sequence in which conditions are evaluated – this is not true for decision tables.

Within a decision tree each condition is split into two parts, the first being a variable written above the tree and the second being the value (or range of values) for that variable within the tree. In our voting decision tree we have two variables, "Australian Citizen" and "Age in years". Australian Citizen is a Boolean variable and hence only two conditions are possible – "Australian Citizen = Yes" and "Australian Citizen = No". Age in years is an integer variable, so it's possible to have any number of associated conditions – in our example there are just two, namely, "Age in years ≥ 18" and "Age in Years < 18". Branches are followed to the right only when a condition is true. For negative conditions this can cause some confusion, for instance when it is true that "Australian Citizen = No" you follow the branch to the right.

Decision tables represent rules vertically within columns, within the *Fig 1.30* decision table there are four rule columns representing four rules. In general, the number of rules in a decision table is a function of the number of conditions. In our *Fig 1.30* example there are two conditions which each can be true or false, hence a total of 2^2=4 rules appear in the decision table. If we had three conditions then 2^3=8 rules would be required to cover all possibilities. Conditions and actions are represented within rows. Ticks and crosses (or Yes/No, True/False) are used to indicate the result of conditions and the resulting actions. There are decision situations where the result of some conditions has no effect on some actions. In these cases it is common practice to leave the square blank to indicate either true or false is acceptable – when there are many conditions this practice significantly reduces the number of rule columns.

Consider the following:

The following Australian Taxation Office tables detail individual income tax rates for the 2007 and 2008 financial years.

Residents

Taxable Income	*Tax on this income*
$0 – $6,000	Nil
$6,001 – $30,000	15c for each $1 over $6,000
$30,001 – $75,000	$3,600 plus 30c for each $1 over $30,000
$75,001 – $150,000	$17,100 plus 40c for each $1 over $75,000
$150,001 and over	$47,100 plus 45c for each $1 over $150,000

Non-residents

Taxable Income	*Tax on this income*
$0 – $30,000	29c for each $1
$30,001 – $75,000	$8,700 plus 30c for each $1 over $25,000
$75,001 – $150,000	$22,200 plus 40c for each $1 over $75,000
$150,001 and over	$52,200 plus 45c for each $1 over $150,000

GROUP TASK Discussion
Identify examples of conditions and actions present in the above tables. Discuss logical and efficient rules for determining how a person's tax should be calculated.

GROUP TASK Discussion
Is a decision tree or a decision table the best tool for describing this decision situation? Discuss. Create the model using your preferred tool.

Storyboards

Storyboards are tools for designing the user interface within software. They document the layout of elements on individual screens and also the connections between screens. Storyboards are often hand drawn sketches of each screen that shows the placement of each screen element. Each sketch includes comments to document specific details. Software applications are available for creating storyboards, however often simple pencil and paper storyboards are easier to construct and alter. Storyboard software has the advantage that colour and some interaction can be included that makes them suitable for creating basic requirements prototypes.

Storyboards can also include a diagram (navigation map) that shows the navigational links between screens – this is particularly valuable for hypermedia software, such as websites and multimedia systems. We shall consider storyboards that use different navigation structures suited to hypermedia in chapter 2 and again within the multimedia option in chapter 6. In this section we focus on general user interface design issues that should be considered during the design of screens.

User Interface
Part of a software application that displays information for the user. The user interface provides the means by which users interact with software.

The user interface is the most obvious element within most software applications and also forms the basis upon which prototypes are built. The user interface is more than just the placement of components on the screens; rather it provides the total interaction between the user and the software application. User-friendly interfaces are easy to understand. They include standard screen elements that perform in predictable ways. They guide and assist participants as they enter data and initiate information processes.

There are numerous design factors that influence the efficiency and accuracy of user interfaces. Indeed the study of user interface design is itself a complete discipline; nevertheless let us consider some basic principles that should be considered when designing quality user interfaces.

- Know who the users are. What are their goals, skills, experience and needs? Answers to these questions are required before an accurate assessment of the user interface can be made. For example a data entry screen that will be used every day by data entry operators will be quite different to one used infrequently by unskilled users. It will require keyboard shortcuts, consideration of paper forms from which data is entered, quick response times, and so on.
- Consistency with known software and also consistency within the application. Users expect certain components to operate in similar ways and to be located in similar locations. For example the file menu is located in the top left hand corner of the screen, placing it elsewhere would be inconsistent and confusing. Radio buttons permit just one item to be selected; allowing users to make more than one selection is inconsistent. Consistency allows users to utilise their existing skills with confidence when learning new software applications.
- Components on data entry screens should be readable. This includes the words used as well as the logical placement and grouping of components. The interface should include blank areas (white space) to visually imply grouping and to rest the eye. Colour and graphics should be used with caution and only when they convey information more efficiently than other means.
- Clearly show what functions are available. Users like to explore the user interface; this is how most people learn new applications, therefore functions should not be

hidden too deeply. If a particular function is not relevant then it is better for it to be dulled than for it to be hidden, this allows users to absorb all possibilities. At the same time the user interface should not be overly complex. For instance if many data items need to be collected, consider splitting the data into logical groups with each group on its own screen.

- Every action by a user should cause a reaction in the user interface. This is called feedback; without feedback that something is occurring, or has occurred, users will either feel insecure or will reinitiate the task in the belief that nothing has happened. Feedback can be provided in simple ways; such as the cursor moving to the next field, a command button depressing or the mouse pointer changing. Tasks that take some time to complete should provide more obvious feedback indicating the likely time for the task to complete.
- User actions that perform potentially dangerous changes should provide a way out. Many modern software applications include an 'undo' function, whilst others provide warning messages prior to such dangerous tasks commencing. In either case the user is given a method to reverse their action.

Consider the following

MockupScreens is a software application for creating storyboards that could be used as simple requirements prototypes. The software includes a variety of standard screen elements together with the facility to easily add example data to each element. The resulting screens can be exported as HTML files for distribution to users. No user interaction with the final storyboard is possible; rather the screens are presented as a simple sequential slide show.

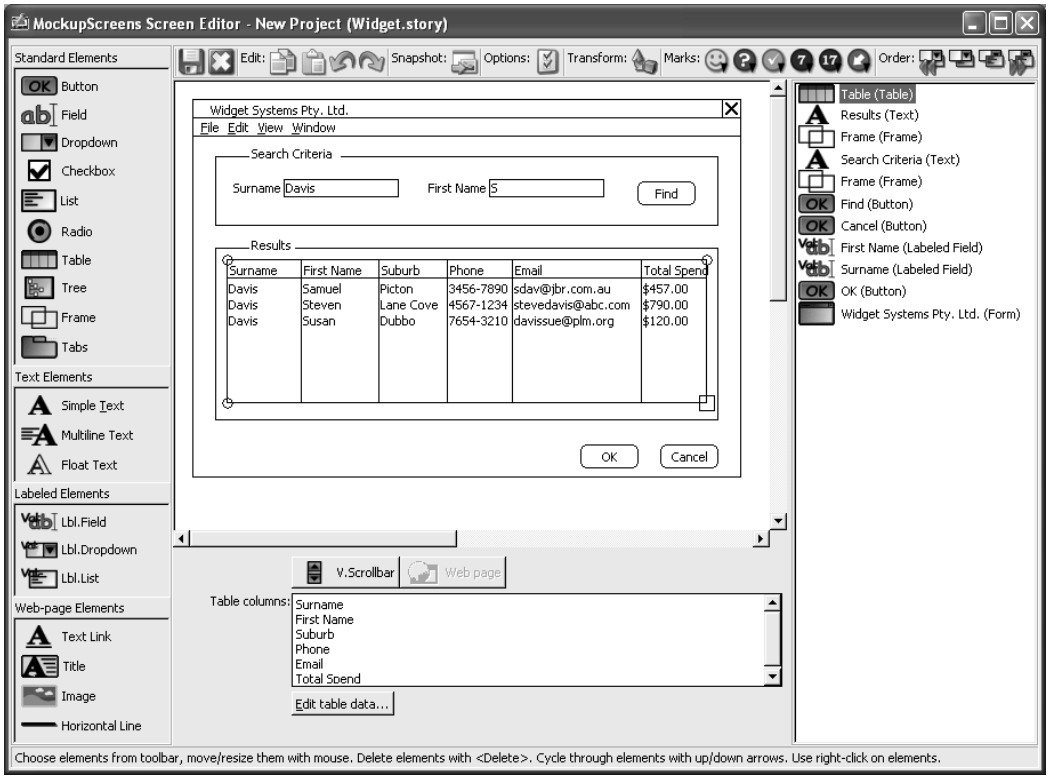


Fig 1.31
Designing a screen layout within MockupScreens.

GROUP TASK Research
Research, using the Internet or otherwise, examples of screen design tools that assist in the creation of storyboards. Assess the usefulness of these tools for creating requirements prototypes.

GROUP TASK Discussion
Critically evaluate the simple screen being designed within MockupScreens in *Fig 1.31* in terms of the dot points on the previous page.

DESIGNING THE INFORMATION TECHNOLOGY

The use of design tools to create system models was about designing the information processes and defining the detail of the data and information used and produced by the system. The information processes, data/information and also the information technology all work together with the participants to achieve the system's purpose. The data is processed into information using the system's information technology – the hardware and software. Hence the hardware and software must be chosen to maximise the efficiency of the information processing.

The ability of the hardware and software to perform the system's information processes is of course essential, however there are various other factors that also require consideration. Many of these factors are likely to be specified within the Requirements Report. Some possible questions that address these factors include:

- Is it maintainable? Are there regular upgrades and will these upgrades continue?
- Are spare parts available for hardware now and in the future?
- Is the software user friendly and easy to learn? Does it use appropriate terminology?
- Does the user interface behave similarly to other applications known to the users? Can the user interface be modified to suit the user's needs?
- Is the software human centred rather than machine centred? Does it enable tasks to be performed according to the user's preferences not the machines?
- Will it meet future needs and expansion? How easily can it be modified or expanded?
- Can it be customised to meet new and emerging requirements?
- Is there a large customer base for the technology? Are their many users proficient in the use of the technology? Do they recommend the technology or do they have complaints?
- Is the technology mature? That is, is it stable with few errors or is it experimental?
- Are the human interface devices ergonomically sound?
- Is the furniture and environment in which users will work ergonomically sound?

The nature of the system and its requirements will determine which of the above questions are relevant. For example, when designing a website the hardware and the furniture used by end-users is part of the system environment – it is beyond the scope (or control) of the system. It is important to confirm answers to each question using sources other than the manufacturer or distributor. Existing customers who have used the technology for some time are often in the best position to confirm claims made by manufacturers and distributors.

GROUP TASK Discussion
Both hardware and software should be designed for ease of maintenance. Focus on a particular hardware device or a particular software application and identify design features that simplify its ability to be maintained.

Consider Pet Buddies Pty. Ltd.

Fred has identified the hardware and software for the new system. A brief discussion of each component, including his recommendations follows:

Hardware

- Analog PCI telephony board capable of managing up to 8 telephone lines. 4 telephone lines will initially be connected. Fred has researched possible cards and found the Talk Voice 8LV board by CallURL to be the most reliable and most highly recommended device (see *Fig 1.32*).
- RAID storage utilising at least 4 HDDs and using both striping and mirroring. Total capacity greater than 1Tb. Must include the ability to 'hot swap' HDDs. A number of recommended units are available, including the SOHORAID SR6500 from Silicon Memory (see *Fig 1.33*)
- Intel Pentium based computer. At least 2Gb of RAM. As Pet Buddies already has existing Dell systems, a Dell machine is recommended. Various performance options are possible. The computer must contain a serial ATA interface to connect the RAID device.
- A Gigabit Ethernet NIC must be supplied with the computer.

Software

- Windows operating system to match existing LAN. Currently Windows XP Professional. This is more cost effective if included with the purchase of the server.
- RAID software is included with the RAID hardware.
- Voice mail software is the most critical software component. The software should include:
  - capacity for up to 1500 mailboxes.
  - ability to accept calls from at least 8 lines simultaneously.
  - outgoing messages and menu items for each mailbox that can be customised.
  - ability for messages to be created by voice synthesis of text retrieved using SQL commands.
  - ability to connect to an existing Microsoft SQL Server database.
  - messages stored as individual .WAV files.
  - ability to edit .WAV file messages.

  TeleSound Pty. Ltd. markets IVR (Interactive Voice Response) Phone Assistant, a product that is widely used, developed in Australia and meets or exceeds each of the above dot points. Furthermore, TeleSound is able to provide a technician to setup the software and make any modifications to ensure it meets Pet Buddies' requirements.
- Custom software for performing speech recognition and subsequent faxing. This is process 4 on the DFD in *Fig 1.29* and will be outsourced to a specialist software developer.

GROUP TASK Discussion
Explain why each of the above components are required for Pet Buddies' new information system.

www.CallURL.com

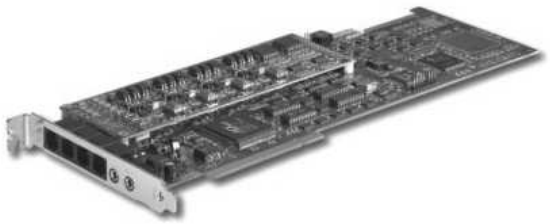
Talk Voice 4/8LV

4- and 8-Port DSP-based Analog Interface Voice Board

VOICE

VoIP

TAPI



Features

- Four/eight independent voice processing ports supporting low-to medium-density voice systems in a single PC
- Two extra audio jacks allow voice monitoring and auxiliary audio input
- μ-law voice coding with selectable data rates 32 Kb/s or 64 Kb/s on a channel-by-channel basis
- Configure multiple boards in a single PC for easy and cost-effective system expansion, capable of scaling from 4 to 80 ports
- Caller ID detection, supports both FSK as well as DTMF
- Full duplex voice channels for VoIP applications
- Bundled with TAPI and wave drivers for Windows 2000 and Windows XP
- Optional CT32 voice and fax development kit for Windows 2000 and Windows XP
- Support LC Drop/Polarity Reversal detection
- Detailed call progress monitoring & analysis

Applications

- Interactive Voice Response (IVR)
- Automatic Call Distribution (ACD)
- Unified Messaging System (UMS)
- Unified Communication (UC)
- TAPI-compliant applications

Technical Description

Configuration

The Talk Voice 4/8LV provides four/eight telephone line interface circuits for direct connection to analog loop start lines. Multiple Talk Voice boards can be installed in a single PC chassis enabling port expansion up to 80 ports per system.

Network Interface

The Talk Voice 4/8LV supports powerful call progress analysis, provides parameters programmable for country-specific regulation ,enhanced DTMF detection & cut-through, and PABX-specific tones detection. Quality voice recording in ADPCM, PCM or WAV formats.

Open Architecture

The Talk Voice 4/8LV boards share a common driver architecture with other Talk Voice boards, offering users and developers both, maximum flexibility and scalability. All Talk Voice products support standard TAPI and wave drivers for Windows 2000 and Windows XPA CT32 voice and fax development kit is provided for application developers.

Fig 1.32
Brochure describing the recommended analog PCI telephony board.

GROUP TASK Discussion
Why do you think Fred has recommended a specialist telephony board, when most dial-up modems include voice capabilities? Discuss.

SOHORAID SR6500

New generation of EzyRaid,
with 5 bay RAID 0, 1, 3, 5.

Welcome to the new generation of EzyRaid. The SR6500 is the ultimate in cost effective high performance RAID, when the integrity and security of your data is your highest priority. With RAID levels 0, 1, 3 or 5, the SR6500 can meet the needs of any user who may require simple data protection or for content creation professionals who have large capacity requirements as well as the need for speed.

True Plug'n'play
Super speeds, and true plug and play performance under Win98SE/ME/2000/XP or Mac OS 9.2 or above.

Lots of space
Five bays allow for a storage capacity of over 1 terabyte.

Workspace Efficient
The SR6500 is neat and tidy with dimensions of 16.2cm x 27.0cm x 22.8cm.

Totally Flexible RAID options: Interfaces
Three interfaces maximise flexibility with FireWire 800 (1394b), USB 2.0, and Serial ATA ports.

- FireWire 800 Transfer rate up to 88MB/sec
- USB 2.0 Transfer rate up to 56MB/sec
- Serial ATA Transfer rate up to 150MB/sec (1.5Gbps)

Disk Bus Interface

- Ultra ATA/133 compatible
- 5 channels, operating in parallel
- 5 hot-swap drive bays (not supported under SATA)
- ATA 6: Supports big drives 300GB+ (normal IDE internal hard drives)

Auto Configure
Auto Rebuild Function: Easy rebuilding without interrupting the system operation.
Hot Swap Function: In the event of a hard drive failure

Fig 1.33
Description of the SOHORAID SR6500 from Silicon Memory.

GROUP TASK Discussion
Identify features of the above RAID storage device that improve the maintainability of the device.

BUILDING/CREATING THE SYSTEM

The detail of how the system will actually be created and built is determined by the chosen system development approach and the specifics of each individual system. Clearly we cannot hope to cover such a wide range of possible strategies, tools and techniques. Rather we will consider two examples – refining existing prototypes and guided processes within application packages. The first is relevant when a prototyping approach is used and the second is particularly relevant to end-user or participant developed systems.

Refining Existing Prototypes

Existing prototypes can be refined such that they evolve into the final software solution. There are a variety of different software applications that allow the user interface and even the underlying processes to be easily modified to fulfil new and modified requirements. In addition to the mechanics of actually building each prototype these applications should also assist the developers to maintain different versions of the solution and provide mechanisms for them to easily revert to previous versions as required. Often changes made to one prototype will be rejected or perhaps a prior version of some aspect of the solution will eventually become part of the preferred solution. The ability to accurately track such changes avoids the need to duplicate work that has already been done.

Ongoing user feedback is critical to the success of the prototyping approach. When prototypes are distributed widely or distance is significant then electronic feedback will be needed. Mechanisms for the provision of such feedback should be simple and can be included within the prototypes. For instance, the ability to add comments alongside screens. The ability to provide feedback as the prototype is actually being used is a distinct advantage. Furthermore the context of the feedback is available to assist developers understand the specific detail of each feedback comment.

Many prototyping tools also include functions that simplify the creation of prototypes so they can easily be delivered to and installed by users. For instance, a website that will connect to an operational database can include a simulated sample database built using a file-based database management system which avoids the need to install database server software.

Consider the following prototyping situations:

- A company's website is being modified to include a shopping cart so that users can purchase and pay for goods online. The system will store data within a database running on a database server. The system produces various reports, including delivery dockets and invoices, for use by the system's participants.
- A factory's assembly line is automating some additional functions that are currently performed manually. One new function involves various computer controlled sensors and robotic arms removing components from the assembly line that include faults. In the final system this function will be duplicated and performed by ten different work units.
- A new model motorcycle has just been released. To assist mechanics throughout the world an expert system is being developed. This system simulates the knowledge of an expert mechanic who has detailed knowledge about the new motorcycle. The system asks the mechanic a series of logical questions in an attempt to diagnose problems they encounter with particular motorcycles.

- A school has outsourced the development of a new pastoral care system to a software developer who is using a prototyping approach. The system will link to the school's existing student and timetable database. Essentially the new system adds and subtracts points to and from students based on their actions. When students reach different point levels they are given punishments or rewards. A web interface is included so that teachers can use the system over the school's intranet.

GROUP TASK Discussion
For each of the above situations, discuss issues associated with the refinement and distribution of prototypes to participants. Also describe possible strategies for obtaining ongoing feedback from users.

Guided Processes in Application Packages

Many software applications include Wizards or assistants that guide the user as they perform common tasks. Initiating such a guided process is often the first step when developing a system using the end-user or participant development approach. Furthermore specific wizards can be initiated to build specific parts of the solution. For instance, many software applications include a Wizard or assistant to simplify connecting to and querying an existing database or other data source. *Fig 1.34* shows a screen from Microsoft Excel's Wizard that automates the connection and querying of external databases. Other guided processes automate the creation of common solutions. For example, *Fig 1.35* is a screen from an Access Database Wizard that automates the creation of a resource-tracking database. Once the basic structure of the solution has been created the developer/user can make modifications to meet their specific requirements.

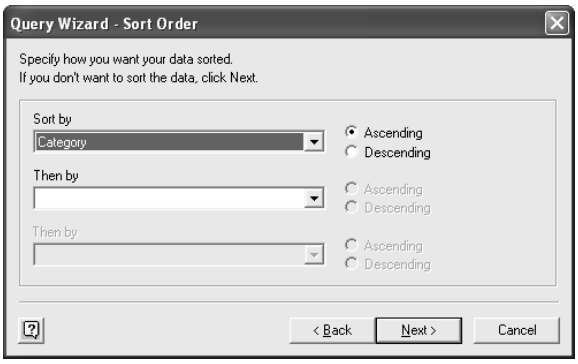


Fig 1.34
Connecting and querying a database within Microsoft's Excel spreadsheet.

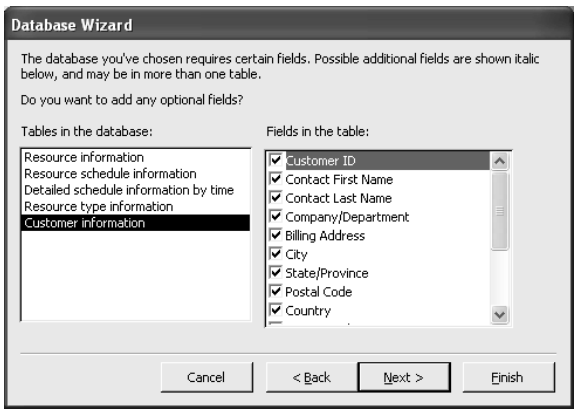


Fig 1.35
Screen from a MS-Access Wizard for creating resource-tracking databases.

Consider guided processes for accomplishing the following:

- Connecting a word processor file to a database to perform a mail merge.
- Extracting data from a database for analysis within a spreadsheet.
- Creating a chart within a spreadsheet.
- Building the structure of a database.
- Creating a data entry form based on an existing database.
- Recording a sequence of keystrokes as a macro that can be played back.

GROUP TASK Discussion
Guided processes are often composed of a sequence of steps. Outline typical steps required to complete each of the above guided processes.

HSC style question:

A bank's loan approval system uses the following decision table as the basis for deciding on the type of loan granted to home buyers. Each home buyer submits their income and the purchase price, whilst the bank's existing system provides the current total in the home buyer's accounts together with their repayment history.

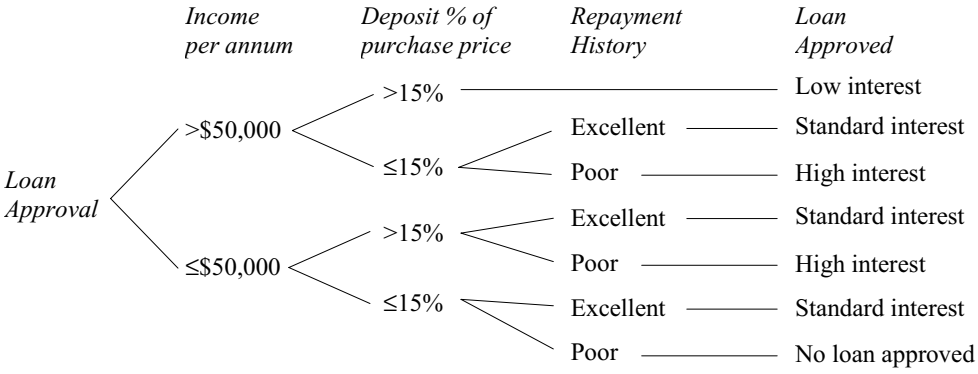
Conditions	*Rules*							
Income >$50,000 per annum	✓	✓	✓	✓	×	×	×	×
Deposit >15% of purchase price	✓	✓	×	×	✓	✓	×	×
Excellent repayment history	✓	×	✓	×	✓	×	✓	×
Actions								
Approve low interest loan	✓	✓	×	×	×	×	×	×
Approve standard loan	×	×	✓	×	✓	×	✓	×
Approve high interest loan	×	×	×	✓	×	✓	×	×

(a) Construct a suitable decision tree for this decision.

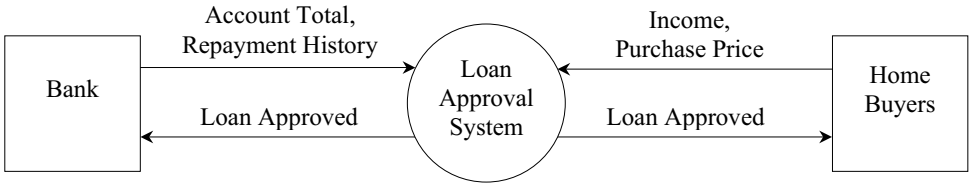
(b) Construct a context diagram for the bank's loan approval system.

Suggested Solution

(a)



(b)



Comments

- In an HSC or Trial HSC examination each part would likely attract 3 marks.
- In part (a) 8 rules have been reduced to 7. Using different sequences of conditions will yield slightly different rules. Is there a solution using less than 7 rules?
- In part (b) it is reasonable to assume both Bank and Home Buyer are informed of the Loan Approved.

SET 1D

1. Which of the following lists includes the resources used to perform the system's information processes?
   (A) Context diagrams, data flow diagrams and data dictionaries.
   (B) Participants, information technology and data/information.
   (C) External entities, processes and data flows.
   (D) Hardware and software.

2. Data flows on context diagrams always:
   (A) flow from a process into another process.
   (B) flow from an external entity into the system.
   (C) describe the processes occurring to transform data into information.
   (D) describe data moving to and from the system and its external entities.

3. A data flow diagram contains four processes that are numbered 4.2.1, 4.2.2, 4.2.3 and 4.2.4. What level data flow diagram is this an example?
   (A) 1
   (B) 2
   (C) 3
   (D) 4

4. What is the best reason why the outputs from a process must be different to the inputs into the process?
   (A) All data flows must have different labels.
   (B) All processes alter data in some way.
   (C) To simplify the construction of data dictionaries.
   (D) This is a requirement when constructing data flow diagrams.

5. Which tool would be most useful when designing the user interface?
   (A) Context diagram
   (B) Data dictionary
   (C) Decision tree or table
   (D) Storyboard

6. Which of the following best defines a sink?
   (A) An external entity that is not part of a system but supplies data to a system.
   (B) People who receive information from the system.
   (C) A process that gets input from the system but does not supply data to the system.
   (D) An entity that is external to the system which receives information from the system.

7. A table describing details of each data item processed by a system is known as a:
   (A) context diagram.
   (B) data dictionary.
   (C) data flow diagram.
   (D) decision tree.

8. Within a system, which of the following allows processing to pause?
   (A) External entities
   (B) Data flows
   (C) Processes
   (D) Data stores

9. In a decision table, rules are represented:
   (A) by each horizontal row.
   (B) by each vertical column.
   (C) as a sequence of conditions.
   (D) as sets of actions.

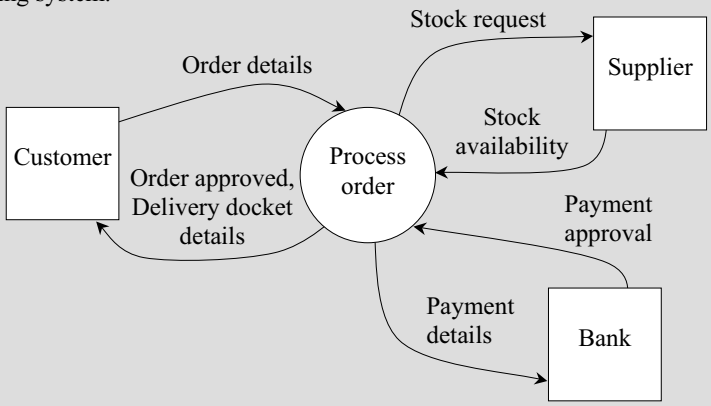
10. A decision is made based on whether an account is overdue, if the total owing on the account is greater than $1000 and whether the customer is "Trusted". Which of the following is TRUE when constructing a decision tree for this decision?
    (A) Exactly 8 unique branch sequences are required.
    (B) At least 8 unique branch sequences are required.
    (C) 4 unique branch sequences are required.
    (D) A maximum of 8 unique branch sequences are required.

11. Define each the following and describe how they are included when constructing context and/or data flow diagrams.
    (a) External entities
    (b) Processes
    (c) Data flows
    (d) Data stores

12. Identify and describe factors that should be considered when choosing or designing information technology that affect the ability of the hardware or software to be maintained.

13. Construct a context diagram for the following systems.
    (a) A handheld GPS system gets location data from satellites and the final destination from the user. The system then directs the user to their destination.
    (b) A booking system is being developed for an upcoming conference. The system receives online bookings from conference delegates, sends payment details to PayPal for processing and approval, and then sends each delegate an email to confirm details of each booking has been made and payment has been completed.

14. Consider the following context diagram that models the flow of data to and from a company's ordering system.



    To process an order the order details are used to determine the total cost of the order using data from the company's product orders database. This database is also used to determine if the warehouse already holds sufficient stock of each product. If new stock needs to be ordered then a stock request is sent to the appropriate supplier who returns details in regard to availability of the product. Assuming all products are available the system sends the payment details to the bank for processing and approval. Orders are only approved and stored in the orders database if all products are available and payment has been approved. When all products are present in the warehouse the order is delivered together with a delivery docket.
    (a) Expand the context diagram into a level 1 data flow diagram.
    (b) Create a data dictionary for your level 1 data flow diagram.
    (b) Construct a decision table to model the decision to approve or not approve each order.

15. A salesman is developing a customer database to store details of each of their potential and actual customers. When a customer phones the salesman first wishes to check if they are already in the database. This involves searching on the customer's name, phone number and also on their address. If any of these details match then the existing record is updated as needed. If no match is found then a new record is created. Each record includes the customer's surname, first name, phone number, email address and postal address.
    (a) Design a screen or screens for this system using a storyboard. If your design includes more than one screen ensure you include the navigational links between the screens.
    (b) Construct a decision tree to model the decision resulting in actions to either add a new record or update an existing record.
    (c) Create a data dictionary for the customer database.

IMPLEMENTING

This fourth stage of the system development lifecycle is where the new system is installed and commences operation. The old system ceases operation and is replaced with the new system. There are various different methods for performing this conversion. However, all these conversion methods require a similar set of tasks to be documented and then completed prior to the system commencing operation. The details are specified within an implementation plan. Typical implementation steps include:

1. Installing network cabling and outside communication lines.
2. Acquiring and installing new hardware and software.
3. Configuring the new hardware.
4. Installing and configuring the software.
5. Converting data from the old system to the new.
6. Training the users and participants.

GROUP TASK Discussion
Do the 6 steps above need to be completed in the precise order they are listed? Justify and explain your answer.

In this section we first consider the content of a typical implementation plan, we then consider four common methods of implementing or converting from an old system to a new system. Finally we discuss techniques for training users and participants to operate and understand the new system.

IMPLEMENTATION PLAN

Many people and organisations are involved in the implementation of most new information systems. For example organisations that supply and deliver the hardware, technicians who install communication and other hardware and the people who install, configure and test the operation of the software. There also trainers who teach the participants to use the new system and also the participants themselves. All these people must be organised so they complete their tasks in the correct sequence and at the correct time. For this to occur requires planning.

A typical implementation plan should consider and document in advance solutions to the following questions:

- How and when the participants are to be trained to operate the new system. Will there be formal training sessions in advance of the system being installed? Will the training be onsite or offsite? Will specialist trainers be employed or will members of the development team perform this function? Will an operational manual be produced that details specific procedures participants should follow? How will other work be completed whilst participants are being trained?
- The method of converting from the old system to the new system. Is it acceptable for no system to operate during installation? Should or can both old and new systems remain operational until the operation of the new system is ensured? What happens if something goes wrong during conversion? What conversion tasks need to be completed and in what order? How will conversion affect other systems that are operating? Can conversion occur outside normal working/office hours?
- How the system will be tested. Is sample data available for onsite testing? When and which parts of the system will be ready for testing? Consider testing each system component independently as it is installed, then test the larger system as components are connected. Schedule and plan for testing throughout installation –

both hardware and software testing. Consider creating a backup plan in the event some components fail.

- Conversion of data for the new system. Often data within the existing system will need to be converted to operate with the new system. Are automated processes available to simplify such data conversion? How long will data conversion take? How accurately can the data be converted? Will the existing system remain operational? Does the new system access and process the same data as the existing system? If so will the old processes affect the new, or the new processes affect the old? What happens to data that is processed whilst data conversion takes place?

The implementation plan should address the above issues. Think of the implementation plan as a project plan that identifies the tasks, people, processes, timing and also cost of the system's implementation.

GROUP TASK Discussion

Consider the implementation of an information system into a new fast food outlet. The system includes a LAN with six point of sale terminals and five other computers and printers. The system uses proprietary software used by all stores within the fast food chain. Discuss the implementation plan for this system with reference to the above points.

METHODS OF CONVERSION

There are a number of methods of introducing a new system and each of these methods suits different circumstances. Usually implementation of a new system includes converting from an old system to the new system.

We consider the following four methods of conversion:

- Direct conversion
- Parallel conversion
- Phased conversion
- Pilot conversion

Direct Conversion

This method involves the old system being completely dropped and the new system being completely implemented at a single point in time. The old system is no longer available. As a consequence, you must be absolutely sure that the new system will operate correctly and meet all of its requirements. Furthermore full and complete testing at the time of installation is needed to confirm that all components are indeed operating as expected. It is particularly important to anticipate and plan for possible faults – perhaps ensuring replacements are readily available or having duplicates on hand for any critical components.

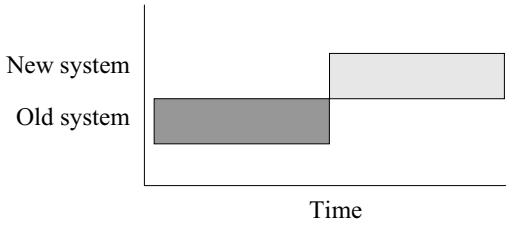


Fig 1.36
Direct conversion method of implementation.

The direct conversion method is used when it is not feasible to continue operating two systems together, for example it may be impractical for large amounts of data to be entered into two systems. Any data to be used in the new system must be converted and imported from the old system. Often neither system operates whilst this conversion takes place – a suitable quiet time should be chosen or perhaps temporary manual processes can be used. Participants must be fully trained in the operation of the new system before the conversion takes place.

Parallel Conversion

The parallel method of conversion involves operating both the old and new systems together for a period of time. This allows any major problems with the new system to be encountered and corrected without the loss of data. Parallel conversion also means users of the system have time to familiarise themselves fully with the operation of the new system. In essence, the old system remains operational as a backup for the new system. Once the new system has been fully tested and is found to be meeting requirements then operation of the old system can cease. The parallel method often involves double the workload for participants as all tasks must be performed using both the old and the new systems.

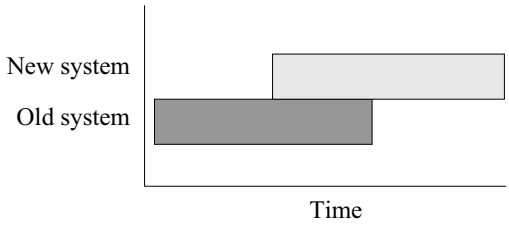


Fig 1.37
Parallel conversion method of implementation.

Parallel conversion is especially useful when the processing is of a crucial nature. That is, dire consequences would result if the new system were to fail. By continuing operation of the old system, the crucial nature of the data is protected.

Phased Conversion

The phased method of converting from an old system to a new system involves a gradual introduction of the new system whilst the old system is progressively discarded. This can be achieved by introducing new parts of the new product one at a time while the older parts being replaced are removed.

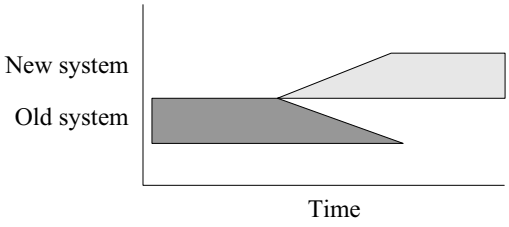


Fig 1.38
Phased conversion method of implementation.

Often phased conversion is used because the system, as a whole, is still under development. When agile methods are used to develop the software a phased conversion is often appropriate. Completed sub-systems are released to customers as they become available. Phased conversion can also mean, for large organisations, that the conversion process is more manageable. Parts of the total system are introduced systematically across the organisation, each part replacing a component of the old system. Over time the complete system will be converted.

Pilot Conversion

With the Pilot method of conversion the new system is installed for a small number of users. These users learn, use and evaluate the new system. Once the new system is deemed to be performing satisfactorily then the system is installed and used by all. This method is particularly useful for systems with a large number of users as it ensures the system is able to operate and perform correctly in a real operational setting. The pilot method also allows a base of users to learn the new system. These users can then assist with the training of others during the systems full implementation. The pilot conversion method can be used as the final acceptance testing of the product. Both the developers and the customer are able to ensure the system meets requirements in an operational environment.

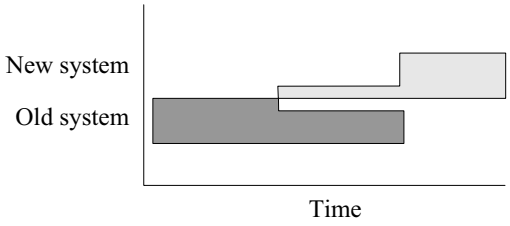


Fig 1.39
Pilot conversion method of implementation.

Consider the following scenarios:

1. A large restaurant is implementing a new information system. There are essentially four sub-systems that interface together to operate the functions of the restaurant – point of sale, accounting, wages and ordering/stocktaking.
2. Chemsoft is a company that specialises in information systems to support the operations of pharmacies. They currently have around 4000 chemists using their system. Chemsoft constantly works on upgrading their software to include new functions and correct bugs. As each upgrade is completed, it needs to be distributed to each chemist for installation. In general upgrades are produced and need to be distributed approximately 3 times per year.
3. A bank is introducing a new Automatic Teller Machine into many of its suburban branches. This new ATM includes a colour touch screen together with various enhanced security features. The software that controls the ATM has been thoroughly tested.
4. Five new computer-controlled life support systems have been purchased by a hospital for use in their intensive care unit. The systems have been used successfully in hundreds of hospitals across the world. The life support systems monitor a patient's temperature, blood pressure and various other vital signs. When an irregularity is detected, the medical staff are alerted electronically. However, the medical staff at the hospital are sceptical, they wish to continue manually monitoring each patient's vital signs and recording them on a paper chart on the end of each patient's bed.
5. Digital mobile phone networks are now the only type of mobile network available in Australia. Digital mobile networks were introduced in Australia in the early 1990s, however the old analog mobile networks were only taken out of service in the late 1990s. Both systems operated together for some 5-10 years.

GROUP TASK Discussion
Identify and justify a suitable method of converting the old system to the new system for each of the above scenarios. Note that it is possible for any combination of conversion techniques to be used.

IMPLEMENTING TRAINING FOR PARTICIPANTS AND USERS

Successful training requires motivated learners. Even the best trainers, using fantastic training techniques and materials will fail if the learners are simply not motivated. For example, nearly all of us complete subjects at school that we are not really enthused about. As a consequence learning in these subjects is an effort. In contrast, even the most unmotivated student is able to learn incredible amounts of information about their favourite hobby or sport. When people are motivated about a subject they actively seek out information, often without prompting. This is not to say that the training methods used are insignificant, rather the point is that motivated learners are vital if the training methods are to be a success.

GROUP TASK Discussion
Choose a subject where some of the class is motivated to learn whilst others are not. Identify reasons for each individual's level of motivation. (Don't choose IPT, as no doubt everyone is highly motivated!)

In regard to new information systems, the learners are the participants and the users. These people are likely to be motivated learners when they:

- are open to change.
- understand how the new system will meet their needs.
- have provided input that has been acted upon during the development of the system.
- have an overall view of the larger system and how their particular tasks will assist in achieving the system's purpose.

These characteristics are achieved through continuous two-way communication throughout the SDLC. For example, if a user has provided an idea during the development process then they should receive feedback regardless of whether the idea has been implemented or not. Indeed feedback on ideas that have not been included is particularly important. Most people will accept rejection if they can see their ideas were considered and that there is a logical reason their ideas were not included.

Let us assume the participants and users are on the whole motivated. We still need to implement some formal training to enable them to commence operating the new system. Some possible training techniques include:

- **Traditional group training sessions**

The trainer can be a member of the system development team or an outsourced specialist trainer. If the software has been purchased with little modification then an outsourced training specialist is likely to provide a better service due to their intimate knowledge of the software. If the software has been customised then a member of the development team is perhaps a better choice. In either case the training can be performed onsite or at a separate premises. Onsite group training can often lead to problems as apparently urgent, but unrelated matters, often interrupt the sessions. Off site training allows participants to focus more fully on the training.

- **Peer training**

One or more users undergo intensive training in regard to the operation and skills needed by the new system. These users are also trained in regard to how to train others to use the system. The trained users are then used to train their peers. Peer training is often a one-to-one process. The trained user is essentially an onsite expert who works alongside and assists other users as they learn the skills to operate the new system. This technique allows users to learn skills, as they are required over time.

- **Online training such as tutorials and help systems**

Online tutorials and help systems allow users to learn new skills at their own pace and as they are needed. It is common for larger systems to be provided with a complete tutorial system. Such systems include sample files and databases that can be manipulated and changed without fear of altering or deleting the real data. Many help systems are now context sensitive. This means they display information relevant to the task being completed.

- **Operation manuals**

Printed operation manuals contain procedural information similar to many online tutorial and help systems. However, operation manuals describe step-by-step instructions specific to the new system. For instance, detailed instructions on how to perform backups, how to add a new customer account or what to do if a product is returned. Such processes likely include both manual and computer-based tasks that differ according to the policies of the organisation. We discuss operation manuals in more detail in the Testing, evaluating and maintaining section later in this chapter.

Consider Pet Buddies Pty. Ltd.

Pet Buddies new system is about to be implemented. Fred, Iris and Tom are discussing the most appropriate method of conversion. The following comments are made during their conversation:

Fred *The speech recognition and faxing software is still not complete. The software developer needs another 3 weeks to complete her work. I think we can go ahead regardless.*

Iris *Some of the experts are over 60 years old. I think it will take them some time to feel comfortable talking to a computer. Also, some customers have expressed their concern in regard to the security of the new system.*

Tom *Do we really need to collect all the activity reports using the new system straight away? We can easily continue using the manual system and just mark reports as done on the computer system.*

Fred *You're going to lose two of your voice telephone lines, so you can't have too many experts continuing to use the old system for long. Also it will be difficult to inform customers. Some will dial the old number and others will need to call the new voicemail number.*

Tom *Iris and I are still unclear about why we need the new RAID device. Our existing server is secure, we're not sure why we can't simply added extra storage.*

Fred *It's about fault tolerance and performance. Each hardware system operates independently. If one fails then the other can continue. Furthermore the amount of audio data stored is enormous compared to your existing database. There is no need for the audio data to be totally secure, it will not contain any personal customer information.*

Iris *I'm nervous about understanding how to use the voicemail software. I'd like someone from Telesound to come out and do some intensive training with us.*

Fred *A technician is coming out to configure the voicemail software a few days before the system goes live. They have requested we all be present to answer any questions they may have. In the afternoon the technician will provide us with a hands-on training session. We can always book further training, if needed.*

Tom *We'll have to inform our customers of the changes. We'll create a brochure that includes a step-by-step explanation of the voicemail operation. The experts can give out the brochure when they're doing each quotation. In this way customers can ask questions face-to-face.*

GROUP TASK Discussion
Recommend a suitable method for converting form Pet Buddies old system to their new system. Use evidence from the above conversation to justify your recommendation.

GROUP TASK Discussion
Explain how Iris and Tom, the experts and Pet Buddies' customers can best be trained to use the new system.

TESTING, EVALUATING AND MAINTAINING

Testing, evaluating and maintaining is the fifth and final stage of the software development lifecycle (SDLC). Unlike the previous stages of the SDLC, aspects of this final stage continue throughout the life of the system.

Tasks included in the testing, evaluating and maintaining stage include:

- testing to ensure the system meets requirements,
- trialling and using the operation manual,
- ongoing evaluation to monitor performance,
- ongoing evaluation to review the effect on users, participants and people within the environment,
- maintaining the system to ensure it continues to meet requirements, and
- modifying parts of the system where problems are identified.

GROUP TASK Discussion
Testing and evaluation occurs throughout all stages of the SDLC. Identify examples of testing and evaluation used during each preceding stage.

TESTING TO ENSURE THE SYSTEM MEETS REQUIREMENTS

Acceptance Tests
Formal tests conducted to verify whether or not a system meets its requirements. Acceptance testing enables the client to determine whether or not to accept the new system.

The testing, evaluating and maintaining stage commences with formal testing of the operational system to ensure it meets the requirements specified in the Requirements Report – this is known as acceptance testing. Once the tests confirm the requirements have been met the system is signed off as complete. The client and the system developers usually agree to use the results of the acceptance tests as the basis for determining completion of the new system. If the tests are successful then the client makes their final payment and the development team's job is complete.

For large-scale information systems acceptance testing is best performed by an outside specialist testing organisation. Even for smaller systems it is preferable for acceptance tests to be performed by people who were not involved in the system's development. People involved in the system development process are likely to be biased. They have designed and implemented the new system, so clearly they will feel the requirements have been met. Furthermore they will, unsurprisingly, view their particular solution as superior to other possibilities.

Although using outside testers are preferable, it is not unusual for the client to perform their own acceptance tests prior to finally accepting and signing off the new system. This is understandable, given that all systems are ultimately developed to meet the needs of clients. Unfortunately disagreement between the clients view of an acceptable system can differ from the views of the developers. It is preferable to agree on the precise nature of the testing and who will perform the tests early in the SDLC – in terms of the traditional development approach this should occur during the creation of the Requirements Report. This can easily become a significant problem with less structured development approaches.

The system is tested and evaluated using a variety of different tests and test data including volume data, simulated data and live data. Such tests ensure the system will meet all system requirements when operational.

Volume Data

Many systems are required to process large amounts of data. Volume data is test data designed to ensure the system performs within its requirements when processes are subjected to large volumes of data. For example, queries within a database application may return their results quickly when the database contains a few hundred records, however how will it perform when each query must examine millions of records? Volume test data aims to answer such questions.

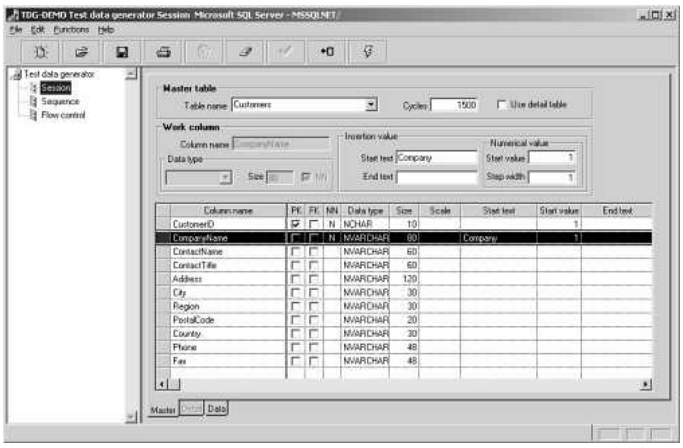


Fig 1.40
Screen shot form TDG (Test Data Generator)
by IGS-EDV Systems Germany.

How can such large amounts of data be obtained? Perhaps the existing system already contains suitable data, if not then software tools are available that will automatically generate large amounts of data with specific characteristics. For example, TDG (Test Data Generator) by IGS-EDV systems of Germany is able to read the definition of databases and create large quantities of compatible test data automatically (see *Fig 1.40*). Volume testing measures response times as well as ensuring the system continues to operate and process data when presented with large amounts of data.

Simulated Data

Simulated test data aims to test the performance of systems under simulated operational conditions, such as when many users, connections or different processes are all occurring in different combinations and at the same time. Clearly it is impractical to enrol hundreds of users to log into a system and all perform different tasks. Instead software is used to simulate this situation. Such simulated testing aims to evaluate the system performance under a variety of different scenarios. For example under anticipated maximum loads, when part of the system fails, when exceptional loads are applied, when users don't respond to prompts or simply cancel or close windows during operations, when the network cannot support the number of requests, etc.

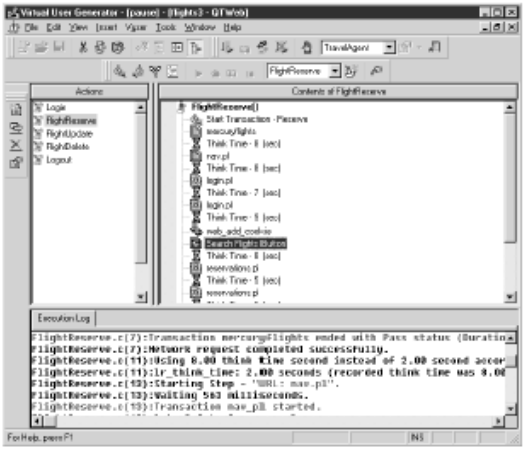


Fig 1.41
Screen shot from Mercury Interactive's
LoadRunner testing software product.

Various companies specialise in the provision of simulated tests and there are also software tools available to perform such tests. One example is Mercury Interactive's LoadRunner, a software tool that simulates many users performing a range of processes and produces information on average response time together with specific details of each problem encountered whilst performing such processes.

Live Data

Live data, as the name implies, is the actual data that is processed by the operational system. Live testing takes place once the system has been installed to ensure it is operating as expected. Testing with live data ensures the system operates under real conditions. Other types of test data are formulated in advance by the developers and hence can only hope to include data and tests that the developers anticipate may cause problems.

Live tests confirm all parts of the installed system are working as expected and meeting the system requirements. For most systems it is impractical to build and test the complete system in advance. Rather such testing occurs onsite once the system is actually installed. Different communication links, computers, operating system settings and various other different hardware and software combinations are likely to be present within the final operational environment. Furthermore newly installed hardware and software must also be tested. Commonly live tests are the final step prior to the completed system being accepted by the client.

GROUP TASK Discussion
Brainstorm issues that may be uncovered by live tests, that cannot be detected by tests conducted prior to the system being installed.

Consider Pet Buddies Pty. Ltd.

Fred, Iris and Tom agreed to test and verify each requirement within the Requirement Report themselves once the system was operational. They are currently working through the list of requirements (refer *Fig 1.16* on page 42) and testing each as they go. Unfortunately they did not specify the precise tests that would be used to verify each requirement. Nevertheless they do agree that most of the requirements have been met. There are just a few requirements whose verification is causing problems. Two examples follow:

3b.4 [The system shall] include the facility for Pet Buddies management to specify that all activity reports from a particular expert or to a particular customer must be approved by Pet Buddies management before release to customers.

Iris and Tom feel this requirement has not been addressed at all. Fred's view is that requirement 3b.3 encompasses this requirement. 3b.3 specifies that any activity report can be checked and/or edited. This means the need to specify particular activity reports is redundant.

3b.11 [The system shall] collect data from experts on the total time taken to complete each home care service.

The new system collects data on the total time each expert spends at each customer's premises. Iris and Tom argue that the phrase 'each home care service' means each particular activity performed by the expert. Fred argues that the current implementation is correct.

GROUP TASK Discussion
Debate each side of the above two arguments. Explain how the disagreements could be resolved and suggest how such problems could be avoided in the first place.

TRIALLING AND USING THE OPERATION MANUAL

The operation manual describes the procedures participants follow as they use the new system. Once the new system is operational, the participants start using the operation manual as they perform their work. During this initial trial period the operation manual will likely require modification to reflect the policies of the organisation and the realities of the system's operation.

Operation manuals include step-by-step descriptions of each task and decision that should be accomplished to perform specific system processes. Operation manuals that specify these procedures are an example of 'procedural documentation'. Procedural documentation in the form of an operation manual is created for a specific information system – usually in written form, either as a printed manual or its electronic equivalent. The specific procedures used commonly include both manual and computer-based tasks. Procedural documentation is also included as part of the online help system within many software applications. Such online help provides step-by-step help specific to common tasks performed by the software package rather than by an entire information system.

Procedure
The series of steps required to complete a process successfully.

Operation manuals are not static documents; they should be continually updated to reflect changes in the information system and changes in the organisation's policies. For example, a company may introduce a new policy requiring direct phone contact with all customers who have outstanding accounts. Previously an overdue account was faxed or mailed. To implement this new policy requires changes to the operation manual with regard to procedures participants follow when chasing overdue accounts.

As operation manuals are intended for participants they should be structured in terms of the processes or tasks performed by these people. Each task should have a clearly defined purpose. For example if a participant commonly needs to generate and fax statements to individual clients then the procedure necessary to perform this task should be included within the operation manual for the system. Such a task is likely to involve initiating a number of system processes, many of which are also used to perform other tasks. Hence operation manuals are not simply a description of each isolated process but rather a description of how these processes are used to perform particular tasks.

For each identified task the operation manual should include:

- What the task is and why it is required. In essence a general statement describing the overall process and its purpose. For example a particular task may be "How to generate and fax individual client statements". This task is required because individual clients regularly request statements at different times.
- How the task relates to other tasks within the system. For example commonly orders do not appear on a client's account until the goods have been despatched. A user preparing client statements must be aware of this.
- Who is responsible for the task and who performs the task? Each task is assigned to a particular participant or group of participants. For example performing backups may be the responsibility of the system administrator, however another participant performs the actual task.
- When the task is to be completed. Many tasks must be completed at a particular time or under particular circumstances. For example overdue accounts maybe

generated every 30 days, or virus software should be installed prior to new computers being added to the network.

- How to complete the task. This section describes the steps the user must perform to complete the task. In most cases this is the major part of each entry.

Consider the following page from an operation manual:

Accounts: Creating a new customer account

Related tasks:	**Officer responsible:**	**Frequency:**
Creating an order. Updating credit limits.	Accounts Manager.	As required.

Task notes:

Potential new customers are frequently indicated when no account number is present on a purchase order received via fax, email or mail. A new customer account must be created for all new customers. Cash customers are assigned a zero credit limit, which causes the system to demand prepayment of orders prior to goods being dispatched. Often cash clients are unaware that an account is maintained in their name and hence do not quote their account number. Credit is only made available to customers once supplier references have been confirmed or a history of past cash orders is present.

Procedure:

1. Determine that the order is in fact from a new customer.
   A. Enter the customer details via the *new account* option on the accounts menu. This process will create a new account number for the new customer.
   B. Select *find matches* on the new accounts screen. This function looks for similar customer details based on phone, fax and address details.
   C. If a match is found then contact the existing customer to resolve the issue. If no clear resolution is determined then the matter is referred to the accounts manager.
   D. If no match then write down the account number and *save* the record. (Credit limit must be 0).
2. Contact the new customer by phone.
   A. Inform customer that the order has been received.
   B. Determine if a credit account is required.
   C. If no credit required then redirect call to an orders clerk. Supply the order clerk with the new account number prior to connecting the customer. End of procedure.
   D. If credit is required then go to step 3.
3. Initiate credit account application.
   A. Explain requirements for opening a credit account as listed on the *Credit Account Application*.
   B. Write account number on *Credit Account Application* and forward to customer.
   C. Inform client that current order cannot be processed without either prepayment or waiting for credit approval.
   D. If prepayment is desired for current order then redirect call to an orders clerk. Supply the order clerk with the new account number prior to connecting the customer.
   E. If waiting for credit approval is desired then write the account number and date on the original order together with the words "Awaiting credit approval". When, and if, the application is approved the order is forwarded to an orders clerk.
   F. When the completed Credit Account Application is received follow the procedures described in **Accounts:Updating Credit Limits.**

GROUP TASK Discussion

Why is it desirable to have step-by-step descriptions like the one above? Discuss.

GROUP TASK Activity

Identify procedural aspects of help systems present in a variety of software applications. Is it necessary for organisations to develop their own procedural operation manuals if they are using these applications as part of their information systems? Discuss.

ONGOING EVALUATION TO MONITOR PERFORMANCE

Evaluation
The process of examining a system to determine the extent to which it is meeting its requirements.

There are two essentials factors to consider in regard to monitoring the performance of a system. Performance can be monitored from a technical viewpoint – is the system continuing to achieve its requirements? Or the system's performance can be monitored from a financial viewpoint – is the system resulting in improved profits? Each of these factors requires ongoing examination to determine the extent to which the system is meeting expectations. This is the process of evaluation.

Technical performance monitoring

Technical performance monitoring aims to evaluate the continuing achievement of the systems evolving requirements. Notice we say 'evolving' requirements. Some old requirements may go down in priority over time or even become irrelevant. Other totally new requirements will emerge and existing requirements will change. This is the nature of virtually all information systems – they change over time. Ongoing evaluation of technical performance aims to verify that requirements continue to be met and identify any changes that may require modifications to the system.

Consider the following:

Some common issues uncovered when performing ongoing technical system evaluation relate to the following factors:

- As the amount of data in the system grows, storage and retrieval processes slow. For instance, when we first purchase a new computer it seems even large video files can be accessed almost instantly, over time the hard drive fills and access slows markedly.
- As the number of transactions increase, response speeds decrease. For example, making a withdrawal from a bank is fast at 10am in the morning, however at 4pm on a Friday afternoon transactions are intolerably slow.
- As users gain more experience their tolerance of poor performance and usability issues decreases. In other words 'familiarity breeds contempt'. For example, a user interface that generates a simple warning message after each new record is added may be acceptable and even useful to new or irregular users. When entering large quantities of data, experienced users will find responding to such messages hundreds of times a day very irritating.

GROUP TASK Discussion
One example is given for each of the above dot point. Identify and describe further examples of each dot point.

Financial performance monitoring

During the 'Planning' stage of the SDLC a feasibility study was undertaken. This study included analysis of the system's economic feasibility. Financial performance monitoring is largely about evaluating the accuracy of the real economic situation against the economic predictions made in the feasibility study. The aim being to evaluate the extent to which the new system is achieving its economic goals.

Data collected during the evaluation should therefore be sufficient to produce accurate comparisons with the expected results within the feasibility study. Consider the graph in *Fig 1.42*, it shows the results of the original break-even analysis compared to the actual situation for a particular project. A simple analysis of this graph indicates that the project ran slightly over budget when it first became operational some 4½ years ago. Despite this the system managed to reach its break-even point a month prior to expectations. Furthermore, according to the graph, the system has failed to realise its expected economic potential over the last 12 months. Although all of the preceding comments are true of the graph, they are not necessarily true of the system. Perhaps a new competitor entered the market a year ago? Maybe 2 years ago there was a major recession? Environmental factors such as these should be considered when performing financial performance monitoring on an information system.

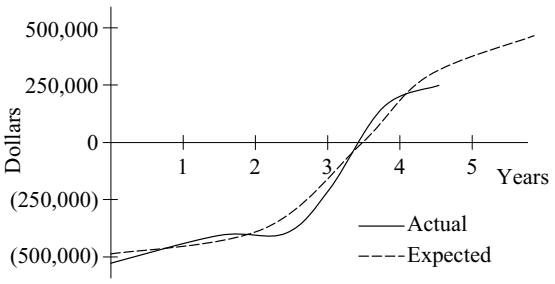


Fig 1.42
Business performance monitoring evaluates actual compared to expected performance.

ONGOING EVALUATION TO REVIEW THE EFFECT ON USERS, PARTICIPANTS AND PEOPLE WITHIN THE ENVIRONMENT

Have you ever participated in market research, been interviewed about a product or service, or completed a survey? If so then it is likely you were part of ongoing user evaluation. Similar techniques can be used to assess the effect of information systems on users, participants and people in the environment. People are the most critical elements of an information system. If they are positive about the system then it is more than likely to be a success, however the opposite is also true.

Following is a brief discussion of some of the effects of information systems on people. All these items are worth considering when creating evaluation tools.

Decreased privacy including perceptions of decreased privacy

Consequences of the *Privacy Act 1988* mean that information systems that contain personal information must legally be able to:

- explain why personal information is being collected and how it will be used
- provide individuals with access to their records
- correct inaccurate information
- divulge details of other organisations that may be provided with information from the system
- describe to individuals the purpose of holding the information
- describe the information held and how it is managed

Changes in the type and nature of employment

New systems will and do alter the work performed by particpants and others who use or are affected by the system. Whenever such change occurs there is potential for both negative and postive effects. New tasks commonly require more advanced skills in regard to using technology rather than skills that substitute for technology. For example, a clerk no longer needs to manually search through filing cabinets, rather they need to be able to use software to query a database. As the search now takes seconds rather than hours, it is likely the clerk will now perform many new and varied tasks or perhaps their work hours have been reduced.

Health and safety concerns

All workers are exposed to potential health and safety problems whilst undertaking their work. Employers are responsible for ensuring these risks are minimised. In NSW the *Occupational Health and Safety Act 2000*, together with the *Occupational Health and Safety Regulation 2001* are the legal documents outlining the rights and responsibilities of employers and employees in regard to occupational health and safety. Workcover NSW administers this act in NSW to ensure and monitor compliance. Employers must setup a procedure for identifying and acting on occupational health and safety (OHS) issues. This requirement is often fulfilled by appointing either an OHS representative or by forming an OHS committee.

Ergonomics is the study of the relationship between human workers and their work environment, it is not just about the design and placement of furniture, rather it is about anything and everything that affects the work experience. This includes physical, emotional and psychological aspects of work.

Ergonomics
The study of the relationship between human workers and their work environment.

Most participants in information systems primarily work in offices at computer workstations. Some broad ergonomic issues relevant to this type of work environment include:

- Furniture and computer hardware design and placement should be appropriate to the task. This includes desks, chairs, keyboards, monitors, pointing devices, etc.
- Artificial lighting should appropriately light the work area. Outside and overhead lighting should not cause glare.
- Noise levels generated by equipment, but also from other workers, to be at reasonable levels. Research shows that conversations from fellow workers are a major distraction to most workers.
- Work routine should include a variety of tasks designed to minimise boredom and discomfort. Working continuously on the same task is the greatest cause of repetitive strain injury (RSI).
- Software design should be intuitive and provide shortcuts for experienced users. The user should drive the software, the software should not drive the user. Training should be thorough and ongoing.
- Procedures for reporting potential OHS problems should be in place and understood by all employees.

Be aware that lack of job satisfaction has been shown to be closely linked to poor ergonomics. Health and safety is not just about minimising and dealing with injuries, rather it concerns the total work experience.

Little or no sense of accomplishment

All people need to feel a sense of accomplishment. There should be a well-defined purpose to every task they perform. Also, each task should have a distinct start and end point. For example, it is most demoralising to work within a system where a single task is continuous, extra work is always present and no end is ever in sight. Unfortunately many existing information systems include such monotonous tasks. Altering the work routine to include a variety of tasks and assigning responsibility for task completion can often assist. Evaluation should identify such occurrences so that modifications can be made.

Deskilling

Deskilling occurs when the information system performs processes that were once performed by participants. For example, when desktop publishing software revolutionised the printing industry the "type setting" trade changed almost overnight. All the existing type setting skills required to manually set lead type were no longer needed. These workers had to either leave the industry or retrain to use the new software. Deskilling can also occur when an information system restricts participants to particular tasks and excludes them from others.

Loss of social contact

Loss of social contact is becoming a common issue. Efficient communication systems allow more and more people to work from home. There is no doubt that this has many advantages, however people are social creatures and they need to develop and maintain relationships with each other. Loss of social contact can also occur when an information system requires participants, particularly those involved in data entry, to spend long periods of time at a computer.

GROUP TASK Discussion

List and describe evaluation techniques that could be used to identify the effects of a new system on users, participants and people within the new system's environment.

MAINTAINING THE SYSTEM TO ENSURE IT CONTINUES TO MEET REQUIREMENTS

Information systems require regular maintenance if they are to continue to meet their requirements. In this regard information systems are just like any other system. For example, a car requires regular servicing if it is to continue to function correctly. However even cars that have been serviced according to the manufacturer's specifications do break down. It is the same with information systems. Therefore maintaining an information system involves:

1. regular maintenance, and
2. repairs when faults occur.

Let us briefly consider typical maintenance tasks performed during the operation of an information system.

- Maintaining a hardware and software inventory. An inventory is a detailed list of all the hardware, software and any other equipment used by the system. It should include where each item is located, when it was purchased and how much it cost.
- Perform backups of the system's data and ensure these backup copies are secured in a safe location. Restore data from backups should a fault occur.
- Protect against viruses by ensuring virus protection software is used and updated. If a virus is detected then initiate processes to remove the virus and protect the rest of the system from infection.
- Ensure illegal software is not installed and that all required software is correctly licensed. Should unlicensed or illegal software be found it should be removed.
- Maintain hardware by carrying out all recommended cleaning and other maintenance tasks.
- Ensure stock of all required consumables is at hand. Consumables include printer toner cartridges, disks, recordable CDs and tapes.
- Install and configure replacement or additional hardware and software.

- Setup network access for new users. This is includes assigning data access rights together with installing the hardware.
- Monitoring the use of peripheral devices.
- Purchasing and replacing faulty hardware components as problems occur.
- Ensuring new users receive training in regard to the operation of the new system.

Consider Pet Buddies Pty. Ltd.

Pet Buddies LAN now connects a total of six computers. They also have a tape backup unit, DVD burner, colour laser printer and an inkjet printer. In regard to software Pet Buddies has their voice mail software, the new custom speech recognition software, SQL server and various other standard applications. Currently a single copy of a virus protection application is installed on the machine that provides Internet access via a cable modem.

GROUP TASK Discussion
Using the above dot points as a guide, identify and describe some of the maintenance tasks that Pet Buddies should perform.

MODIFYING PARTS OF THE SYSTEM WHERE PROBLEMS ARE IDENTIFIED

Problems identified during any of the above tasks will require modifications to the system. In addition, new requirements will emerge over the life of the system that will require modifications to be made. For each modification the system development lifecycle (SDLC) commences again. Even if the modification is relatively minor each stage of the SDLC should be completed. This is necessary to ensure the modification works correctly with all parts of the existing system and also to ensure all documentation is updated so it continues to reflect the current operational system.

Consider Pet Buddies Pty. Ltd.

After six months of operation a formal review of Pet Buddies system is undertaken. Questionnaires are distributed to experts and customers. Various issues are identified and then prioritised. The three most critical issues are listed below:

- Faxed activity reports are often poorly worded to the extent they are virtually unreadable.
- In the evening experts are often unable to reach Pet Buddies to submit voicemail activity reports.
- Customers who already know their expert would prefer to contact them directly rather than obtain activity reports from the Pet Buddies system.

GROUP TASK Discussion
Critically analyse the development of the Pet Buddies system to determine why these issues were not foreseen and resolved earlier.

GROUP TASK Discussion
Propose suitable modifications to the system that would help resolve each of the above issues.

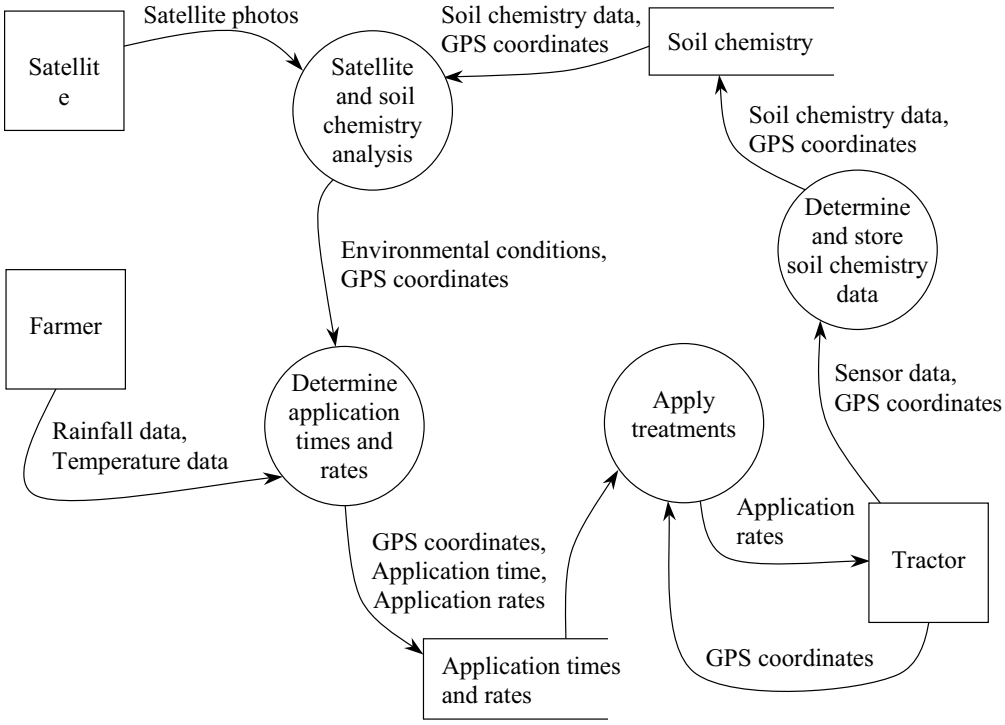
HSC style question:

A farmer has recently read an article on a relatively new farming technique known as "Precision Agriculture". The article claims that Precision Agriculture increases yield and significantly reduces fertilizer, insecticide and other treatment costs.

According to the article Precision Agriculture involves the detailed computer analysis of satellite photographs and soil chemistry data (from actual field tests) to determine differences in environmental conditions within precise areas of each field – some implementations analysed conditions for individual areas measuring less than a square metre. This information, together with historical rainfall and temperature data for the property (which is routinely collected by farmers on a daily basis), is used to accurately determine the optimum time and application rate of fertilizer, insecticide and/or other treatment for each specific area of each field.

During treatment of a field GPS technology is used to determine the tractor's precise location. The location is fed into an onboard computer, which causes the correct rate of each treatment to be applied to each specific area of the field. Sensors attached to the tractor collect soil chemistry data during the application of treatments – this data is then available when formulating future treatment plans.

The following data flow diagram is an attempt to describe this system:



(a) Identify and briefly describe each of the inputs into this system.

(b) Identify the information technology present on the tractor.

(c) Explain why files are required to store the Soil chemistry data and Application times and rates data within the above system.

(d) Assume the farmer has decided to implement "Precision Agriculture". Propose and justify a suitable method of conversion.

Suggested Solution

(a) There are five inputs into the system, namely:
- Satellite photos – bitmap images that are of sufficiently high resolution that areas of less than 1 square metre can be analysed with accuracy.
- Rainfall data – dates and rainfall for each day.
- Temperature data – dates and temperature readings for each day.
- GPS coordinates – numeric data specifying the current location of the tractor.
- Sensor data – numeric data describing the soil chemistry at the tractor's current location.

(b) The tractor contains the following information technology:
- A GPS transmitter/receiver to determine its current location.
- Sensors that are able to detect differences in soil chemistry.
- Actuators to adjust the rate of each treatment applied.
- An on board computer and software to perform both the Apply treatments process and the Determine and store chemistry data process.
- A hard disk or other secondary storage device that holds both the Application times and rates data store and the Soil chemistry data store.

(c) The soil chemistry data is collected at a completely different time to when it is used to generate the environmental conditions. This means it must be stored during the intervening period of time. Also the Soil Chemistry data is collected during the operation of the tractor, hence a data store is needed so that the data is maintained for later copying to the farmer's computer. The Application times and rates data is generated by the farmer's computer, but is used during the tractor's operation. Using a file means that the system can halt whilst the data is transferred to the tractor.

(d) A two stage phased strategy for conversion could be used. Firstly the parts of the system that do not require the tractor could be implemented. These processes are software based and hence the cost would be minimal compared to the large capital required to purchase the specialised tractor hardware. A sample of the application times and rates output from the system can then be analysed on site by the farmer using his experience and a hand held GPS device. If the farmer agrees with the data then the final more expensive phase can be implemented.

Comments

- In an HSC or Trial HSC examination each part would likely attract 3 or 4 marks. Hence this would be a significant question worth a total of 12 to 16 marks.
- In part (a) the inputs to the system are all data flows commencing from an external entity.
- In part (b) and also in part (c) it is possible to assume a wireless link exists between the tractor and another computer. If this were true then the data stores would be on the other computer and the tractor would require wireless communication devices and related software. This would also be reflected in answers to part (c).
- In part (d) a number of different conversion methods could be proposed and justified. For instance direct conversion could be used, with justification based on the fact that the system has already been implemented on other farms. Parallel conversion could also be argued whereby the farmer uses the new system on some paddocks and his old system for others. This would allow him to assess the advantages of the new system for his particular property. Marks would be awarded for a logically justified conversion strategy.

SET 1E

1. Which document details training, testing and conversion of the existing system and data to the new system?
   (A) Project plan.
   (B) Implementation plan.
   (C) Requirements report
   (D) Operation manual

2. Both old and new systems operate together for some time when which method of conversion is used?
   (A) Parallel
   (B) Direct
   (C) Phased
   (D) Pilot

3. Parts of a new system are introduced over time when which method of conversion is used?
   (A) Parallel
   (B) Direct
   (C) Phased
   (D) Pilot

4. Training participants to use the new system should occur during which stage of the system development lifecycle?
   (A) Planning
   (B) Design
   (C) Implementation.
   (D) Testing, evaluating and maintaining

5. Which of the following best describes "acceptance testing"?
   (A) Tests conducted to ensure the system meets requirements so the client will accept the new system as complete.
   (B) Formal tests to ensure the new system interfaces correctly with other existing systems.
   (C) A series of predetermined tests that are formally undertaken to monitor the ongoing performance of the system.
   (D) Ongoing evaluation to monitor the financial benefits of a new system.

6. Testing to verify that the system meets requirements when subjected large amounts of data is known as:
   (A) acceptance testing.
   (B) volume testing.
   (C) simulated testing.
   (D) live testing.

7. Which of the following best describes the use of sample files as participants learn to perform the new system's processes?
   (A) Peer training
   (B) Context sensitive help
   (C) Online tutorial
   (D) Procedural help

8. Testing to ensure the system performs when many different processes are occurring together is best achieved using:
   (A) volume tests
   (B) simulated tests
   (C) live tests
   (D) acceptance tests

9. Which document describes participant procedures for completing tasks specific to the new information system?
   (A) System models
   (B) Implementation plan
   (C) Requirements report
   (D) Operation manual

10. Which term describes the ongoing assessment of a system to monitor the extent to which it continues to meet requirements?
    (A) Maintenance
    (B) Testing
    (C) Evaluation
    (D) Ergonomics

11. Describe the typical content of each of the following documents.
    (a) Implementation plan
    (b) Operation manual

12. Distinguish between volume data, simulated data and live data.

13. Describe each of the following methods of conversion and provide an example situation where each would be suitable.
    (a) Parallel conversion
    (b) Direct conversion
    (c) Phased conversion
    (d) Pilot conversion

14. Describe different techniques for training participants to use a new system.

15. Research and develop procedural documentation suitable for inclusion in an operation manual for each of the following tasks.
    (a) The steps performed when a new student enrols at your school.
    (b) The steps performed by a user as they list their first item on eBay.

CHAPTER 1 REVIEW

1. Management of projects is documented using:
   (A) Requirements reports
   (B) Operation manuals
   (C) Implementation plans
   (D) Project management tools

2. The benefits, risks and costs of possible solutions are assessed when:
   (A) analysing the existing system.
   (B) conducting a feasibility study.
   (C) creating system models.
   (D) interviewing and/or surveying users and participants.

3. A team can best be described as:
   (A) a group of people who work together.
   (B) people with a similar set of skills and training who all work on a project.
   (C) a mixture of skills, personality and behaviour types.
   (D) people with complimentary personality and behaviours who are committed to a common goal.

4. According to Tuckman's four stages of team development, when is conflict most likely to occur?
   (A) Forming
   (B) Storming
   (C) Norming
   (D) Performing

5. Which of the following development methods iteratively produces regular operational systems with progressively more functionality?
   (A) Agile methods
   (B) Traditional methods
   (C) Prototyping methods
   (D) Customisation

6. Where would team members document details of development tasks as they are completed?
   (A) Journal
   (B) Operation manual
   (C) Gantt chart
   (D) Communication management plan.

7. All context diagrams must contain which of the following?
   (A) A single external entity and one or more processes.
   (B) A single process and one or more external entities.
   (C) One or more external entities and one or more processes.
   (D) A single external entity and a single process.

8. Responding with words related to the speaker's message is an essential part of:
   (A) conflict resolution.
   (B) active listening.
   (C) negotiation.
   (D) project management.

9. Which is the most significant deliverable from the designing stage?
   (A) Requirements report
   (B) Gantt chart
   (C) System models
   (D) The new system

10. Details with regard to the operation of the existing system are most likely to be obtained from:
    (A) end-users
    (B) participants
    (C) the project manager
    (D) the development team

11. Describe the content of each of the following documents.
    (a) Funding management plan
    (b) Communication management plan
    (c) Feasibility study report
    (d) Requirements Report
    (e) Implementation plan

12. Describe the communication skills required to successfully manage the development of new information systems, including:
    (a) active listening skills
    (b) conflict resolution skills
    (c) negotiation skills
    (d) interview skills
    (e) team building skills

13. Summarise the essential features of each of the following system development approaches.
    (a) Traditional approach
    (b) Outsourcing
    (c) Prototyping
    (d) Customisation
    (e) Participant development
    (f) Agile methods

14. Recount the sequence of activities occurring during each of the following stages of the SDLC as a system is developed using the traditional system development approach.
    (a) Understanding the problem
    (b) Planning
    (c) Designing
    (d) Implementing
    (e) Testing, evaluating and maintaining

15. Create summaries describing points relevant to the production of each of the following system design tools.
    (a) Context diagrams
    (b) Data flow diagrams
    (c) Decision trees
    (d) Decision tables
    (e) Data dictionaries
    (f) Storyboards

In this chapter you will learn to:

- identify the type and purpose of a given information system
- represent an information system using a systems representation tool
  - identify the purpose, information processes, information technology and participants within a given system
  - represent diagrammatically the flow of information within an information system
- identify participants, data/information and information technology for the given examples of database information systems
- describe the relationships between participants, data/information and information technology for the given examples of database information systems
- choose between a computer based or non-computer based method to organise data, given a particular set of circumstances
- identify situations where one type of database is more appropriate than another
- represent an existing relational database in a schematic diagram
- create a schematic diagram for a scenario where the data is to be organised into a relational database
- modify an existing schema to meet a change in user requirements
- choose and justify the most appropriate type of database, flat-file or relational, to organise a given set of data
- create a simple relational database from a schematic diagram and data dictionary
- populate a relational database with data
- describe the similarities and differences between flat-file and relational databases
- create a data dictionary for a given set of data
- create documentation, including data modelling, to indicate how a relational database has been used to organise data
- demonstrate an awareness of issues of privacy, security and accuracy in handling data
- compare and contrast hypermedia and databases for organising data
- design and develop a storyboard to represent a set of data items and links between them
- construct a hypertext document from a storyboard
- use software that links data, such as:
  - HTML editors
  - web page creation software
- search a database using relational and logical operators
- output sorted data from a database
- generate reports from a database
- construct an SQL query to select data from a given database, matching given criteria
- calculate the storage requirements for a given number of records (given a data dictionary for a database)
- summarise, extrapolate and report on data retrieved from the Internet
- use search engines to locate data on the World Wide Web
- describe the principles of the operation of a search engine
- design and create screens for interacting with selected parts of a database and justify their appropriateness
- design and generate reports from a database
- identify and apply issues of ownership, accuracy, data quality, security and privacy of information, data matching
- discuss issues of access to and control of information
- validate information retrieved from the Internet

Which will make you more able to:

- apply and explain an understanding of the nature and function of information technologies to a specific practical situation
- explain and justify the way in which information systems relate to information processes in a specific context
- analyse and describe a system in terms of the information processes involved
- develop solutions for an identified need which address all of the information processes
- evaluate and discuss the effect of information systems on the individual, society and the environment
- demonstrate and explain ethical practice in the use of information systems, technologies and processes
- propose and justify ways in which information systems will meet emerging needs
- justify the selection and use of appropriate resources and tools to effectively develop and manage projects
- assess the ethical implications of selecting and using specific resources and tools, recommends and justifies the choices
- analyse situations, identify needs, propose and then develop solutions
- select, justify and apply methodical approaches to planning, designing or implementing solutions
- implement effective management techniques
- use methods to thoroughly document the development of individual or team projects.

In this chapter you will learn about:

Information systems

- the characteristics of an information system, namely:
  - the organisation of data into information
  - the analysing of information to give knowledge
- the different types of and purposes for information systems, including systems used to:
  - process transactions
  - provide users with information about an organisation
  - help decision-making
  - manage information used within an organisation

Database information systems

- school databases holding information on teachers, subjects, classrooms and students
- the Roads and Traffic Authority holding information on automobiles and holders of drivers licences
- video stores holding information on borrowers and videos

Organisation

- non-computer methods of organising including:
  - telephone books
  - card based applications
- computer based methods of organising, including:
  - flat-file systems
  - database management systems
  - hypermedia
- the advantages and disadvantages of computer based and non-computer based organisation methods
- the logical organisation of flat-file databases, including:
  - files
  - records
  - fields, key fields
  - characters
- the logical organisation of relational databases, including:
  - schemas as consisting of
    - entities
    - attributes
    - relationships including one to one, one to many and many to many
  - tables as the implementation of entities consisting of attributes, records
  - linking tables using primary and foreign keys
  - user views for different purposes
- data modelling tools for organising databases, including:
  - data dictionaries to describe the characteristics of data including:
    - field name
    - data type
    - data format
    - field size
    - description
    - example
  - schematic diagrams that show the relationships between entities
  - normalising data to reduce data redundancy
- the logical organisation of hypermedia, including:
  - nodes and links
  - uniform resource locators
  - metadata such as HTML tags
- tools for organising hypermedia, including:
  - story boards to represent data organised using hyperlinks
  - software that allows text, graphics and sounds to be hyper linked

Storage and retrieval

- database management systems (DBMS) including:
  - the role of a DBMS in handling access to a database
  - the independence of data from the DBMS
- direct and sequential access of data
- on-line and off-line storage
- centralised and distributed databases
- storage media including:
  - hard discs
  - CD-ROMS
  - cartridge and tape
- encryption and decryption
- backup and security procedures
- tools for database storage and retrieval, including:
  - extracting relevant information through searching and sorting a database
  - selecting data from a relational database using query by example (QBE) and Structured Query Languages (SQL) commands, including:
    - SELECT
    - FROM
    - WHERE
    - ORDER BY
- tools for hypermedia search and retrieval, including:
  - free text searching
  - operation of a search engine
    - indexing and search robots
    - metadata
- reporting on data found in hypermedia systems

Other information processes for database information systems

- displaying
  - reporting on relevant information held in a database
  - constructing different views of a database for different purposes

Issues related to information systems and databases

- acknowledgment of data sources
- the freedom of information act
- privacy principles
- quality of data
- accuracy of data and the reliability of data sources
- access to data, ownership and control of data
- data matching to cross link data across multiple data bases
- current and emerging trends in the organisation, processing, storage and retrieval of data such as
  - data warehousing and data mining
  - Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP)

2

INFORMATION SYSTEMS AND DATABASES

The aim of all information systems is to produce information from data for use by the system's end-users. The end-users analyse this information to gain knowledge. It is only when knowledge has been gained that the system's purpose can be achieved. To produce such information requires all the information processes, however two of the processes are of particular significance – the data needs to be appropriately organised and it must be able to be stored and retrieved efficiently. Hence in this topic we emphasise both these information processes as they occur within databases and also hypermedia.

Information
Information is the meaning that a human assigns to data. Knowledge is acquired when information is received.

Databases contain the raw data used by the majority of information systems. In this course there are four option topics of which you will study two, namely:

- Transaction processing systems,
- Decision support systems,
- Automated manufacturing systems, and
- Multimedia systems.

Purpose
The aim or objective of the system and the reason the system exists. The purpose fulfils the needs of those for whom the system is created.

Common examples of all these systems include some form of database as the data store for the data they process. For example, transactions are sets of operations that must all occur if the overall transaction is to be completed successfully. Each operation commonly alters, deletes or adds data within one or more databases. An expert system is a type of decision support system that contains a database of facts. This database is interrogated to infer likely conclusions.

Consider the following:

1. The operation of an EFTPOS machine.
2. A GPS navigation system in a car as it directs the driver.
3. The operation of a search engine on the Internet.
4. A computer controlled lathe machining a specific engine component.

GROUP TASK Discussion
For each of the above processes, briefly describe the data used and the information produced.

GROUP TASK Discussion
Identify any databases that are likely to be used during each of the above processes.

We shall examine in this chapter:

- Examples of database information systems.
- We then examine in detail three commonly used methods for organising and storing and retrieving data within information systems:
  1. Flat-file databases (including non-computer examples),
  2. Relational databases, and
  3. Hypermedia or hypertext.
- Finally we consider issues related to the use of information systems and databases.

EXAMPLES OF DATABASE INFORMATION SYSTEMS

In this section we examine three examples of database information systems:

- A school timetable system holding information on teachers, subjects, classrooms and students.
- The Roads and Traffic Authority holding information on vehicles and holders of driver's licences.
- Video stores holding information on borrowers and videos.

For each example we identify the system's environment/boundaries, purpose, participants, data/information, information technology and information processes. We describe the flow of data/information through each system using data flow diagrams. Our aim is to gain an overall view of each system's components and how they work together to achieve the system's purpose.

SCHOOL TIMETABLE SYSTEM

Environment/Boundaries

In this example we consider a school's timetable system as a complete system, however in our particular example it is actually a subsystem within the larger school administration system. School admin systems perform many functions, one of these functions being the maintenance of the school's timetable. The larger administration system forms part of the environment within which the timetable system operates – an entity on the context diagram for the school timetable system.

The larger school administration system provides and obtains data via an interface crossing a boundary to the timetable system – represented by data flows in both directions on the school timetable context diagram (*Fig 2.1)*. For example teacher and student names move from the larger system to the timetable system. The teachers and students personal details, including their names, are maintained somewhere else within the larger system. Note that individual student and teacher timetables can be edited or even removed from within the timetable system, however personal student and teacher details cannot be removed from within the timetable system. The timetable system also provides data to other parts of the larger administration system via queries. For example information on each student's subjects is output from the timetable system to the larger system to enable subject fees to be charged, Board of Studies reports to be prepared, student reports to be produced, etc.

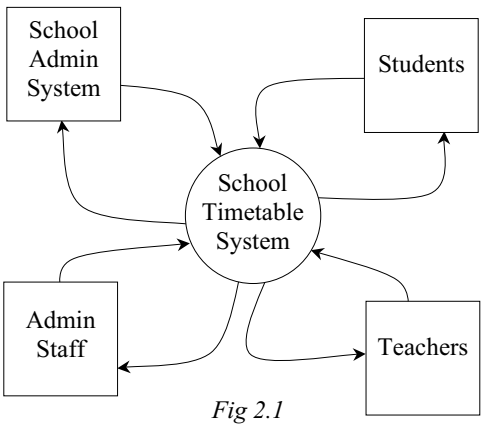


Fig 2.1
Context diagram for a school timetable system (without data flows labeled).

The actual teachers and students are also present within the timetable system's environment. Both teachers and students provide data to the system – teachers indicate classes they wish to teach and students provide subject selections. Conversely both receive their personal timetables from the system. Hence the teachers and students form external entities on the school timetable context diagram (see *Fig 2.1*).

The final entity is the administration staff. This includes office staff, the deputy, the principal and others who may need to locate particular teachers and students during the school day. Note that these people are also likely to be participants and also users within the system.

Environment
The circumstances and conditions that surround an information system.
Everything that influences or is influenced by the system.

The context diagram in *Fig 2.1* above graphically describes the environment in terms of data/information flowing into and out of the school timetable system. However, the environment includes more than just the entities shown on a context diagram – it includes everything that influences or is influenced by the system. The environment includes physical components that affect the system such as the network connections along which data moves and the power supply to the hardware. It is likely that the timetable system operates and shares hardware, and some software that is part of the larger admin system – if this is the case then this information technology is also part of the timetable system's environment.

GROUP TASK Discussion
Why do you think personal student details are maintained outside the timetable system? Discuss.

Purpose

The purpose fulfils the needs of those for whom the system is created. A school's timetable must therefore fulfil the primary needs for teachers and students to know where to go and what to do at all times. Other people within the school, such as admin staff on behalf of parents, need to be able to locate individual teachers or students at any time. Furthermore the larger school admin system needs various different forms of information from the timetable systems to achieve its purpose.

The purpose of a school timetable system is therefore to:

- provide accurate details to each teacher and student with regard to where and what they should be doing throughout each school day.
- enable the location of any teacher or student to be accurately determined at any time throughout each school day.
- provide flexible retrieval methods so timetable data in various forms can be provided to the school's administration system.

Notice that the purpose is not to ensure students and teachers are in the correct place at the correct time; rather its task is to provide the information to enable this to occur. Clearly an information system cannot hope to force students to be in class, on time, every time!

GROUP TASK Discussion
In reality, is there really a difference between needs and the system's purpose? Discuss.

Data/information

In our timetable example we have already mentioned much of the data/information entering and leaving the school timetable system. The following table summarises the data/information mentioned throughout our discussion so far:

Data/Information	External Entity	Source	OR Sink
Teacher Names Student Names	School Admin System	✓	
Subject Selections	Students	✓	
Student Timetables	Students		✓
Class Selections	Teachers	✓	
Teacher Timetables	Teachers		✓
Teacher Name Student Name Update Details	Admin Staff	✓	
Teacher Location Student Location	Admin Staff		✓
Timetable Query	School Admin System	✓	
Query Results	School Admin System		✓

The details from the above table form the basis for labelling each of the data flows on the context diagram (see *Fig 2.2*). Notice that data flow arrows pointing to an external entity indicate sinks, whilst arrows from an external entity and towards the school timetable system indicate sources of data. In this example all the external entities are both sources and sinks – they both provide data to and receive data from the system.

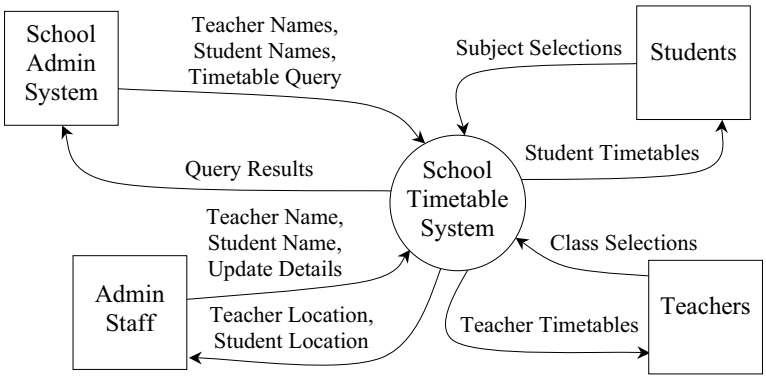


Fig 2.2
Context diagram for a school timetable system.

Consider the following:

To produce information requires data to be analysed and processed. Hence an examination of the final information output from a system is critical when identifying the data that must enter the system. Note that if we were developing a new system a series of verifiable requirements would be created that aim to ensure the system's purpose is realised – many of these requirements would specify the precise nature of the information produced by the system.

GROUP TASK Activity
Examine your own personal school timetable and discuss the data required by your school's timetable system to produce your timetable.

Participants

Participants are those people who perform or initiate the information processes – therefore they are part of the information system. Within our timetable system the primary participants are the administration staff, including those teachers who create and update the timetable. For example office staff probably perform most of the bulk data entry of student subject selections. The teachers who create the timetable analyse the number of students selecting each course to decide on the number of classes that will operate. They also analyse the different combinations of subject selections to best place each class so that the maximum number of students and teacher selections are satisfied. In most timetable systems these processes are accomplished using a combination of manual and computer based processes.

Participants
People who carry out or initiate information processes within an information system. An integral part of the system during information processing.

Consider the following:

Users are not the same as participants, however users can be participants and participants can be users – somewhat confusing! A user is someone who provides data to the system and/or receives information from the system but they need not be part of the system. In general, users who are not participants are indirect users.

GROUP TASK Discussion
In some school timetable systems the students are both users and participants, whilst in most schools students are indirect users but not participants. Identify and describe possible differences in these systems that make this possible.

Information technology

Much of the information technology used within this particular school timetable system is common to the larger school administration system. The following table details the general nature of the hardware and software used:

Information Technology
The hardware and software used by an information system to carry out its information processes.

Description	Purpose	Part of larger Admin System	Hardware	Software
File server with RAID1	Physical data storage	✓	✓	×
SQL Server DBMS	Provide access/security of data	✓	×	✓
Personal computers	Execute software that queries the timetable database	✓	✓	×
Laser Printers	Fast printing of student and teacher timetables	✓	✓	×
LAN	Provide connectivity between server and personal computers	✓	✓	✓
Timechart	Dedicated software application for constructing the timetable	×	×	✓
SAS Timetable module	Application which performs all timetable processes during the school year.	✓	×	✓

Information Processes

The school timetable system is composed of five processes:

1. The creation of the timetable, which includes the collection of subject selections from students and class selections from teachers. This process results in the initial timetable that is used at the start of the school year.
2. Generating student timetables, which involves querying the timetable database and formatting then printing all individual student timetables.
3. Generating teacher timetables, which involves querying the timetable database and formatting then printing all individual teacher timetables.
4. Locating teachers or students includes collecting the student or teachers name and then querying the timetable to determine their location at the current time.
5. Executing SQL (Structured Query Language) statements of various types on the timetable database. The resulting data (if any) from the query being returned to the querying process. This process is used by each of the other processes apart from during the creation of the initial timetable.

Information Processes
What needs to be done to transform the data into useful information. These actions coordinate and direct the system's resources to achieve the system's purpose.

The data flow diagram in *Fig 2.3* is a decomposition of the context diagram to describe these five processes.

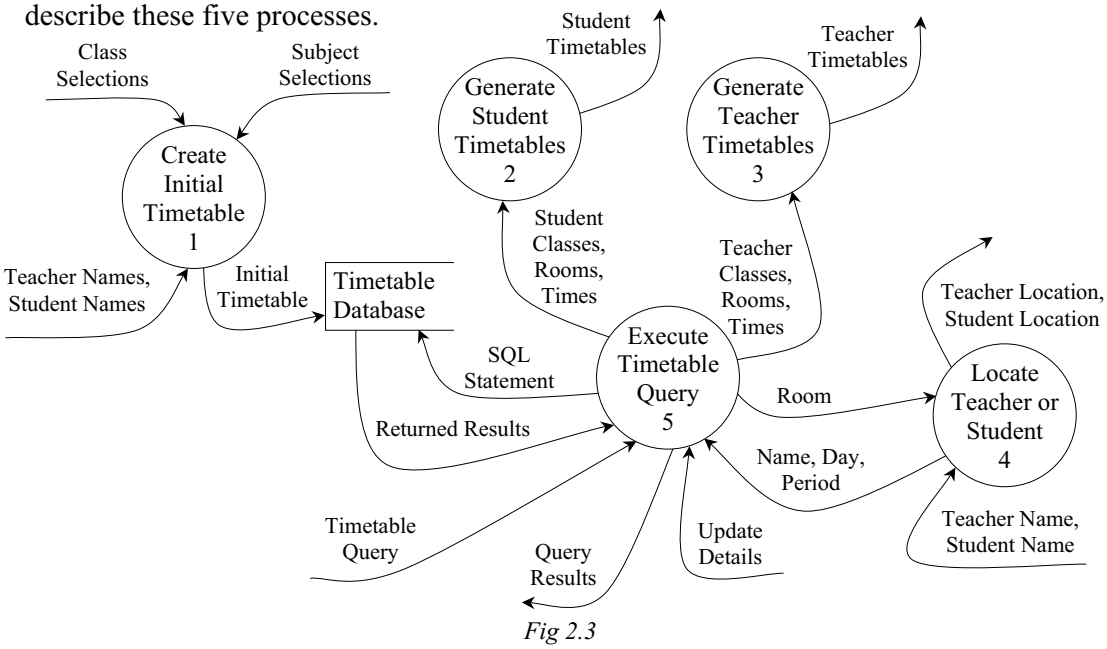


Fig 2.3
Level 1 DFD for a school timetable system.

GROUP TASK Discussion
Three significant software tools are itemised in the table on the previous page. Decide which software tool is most likely to accomplish each process on the above DFD.

GROUP TASK Activity
Choose process 2, 3 ,4 or 5 on the above DFD. Decompose this process further based on your school's timetable system.

THE ROADS AND TRAFFIC AUTHORITY HOLDING INFORMATION ON VEHICLES AND HOLDERS OF DRIVER'S LICENCES

Environment/Boundaries

The Roads and Traffic Authority (RTA) is a NSW statutory authority responsible for managing the road network to ensure efficient traffic flows and improved road safety. This includes building new roads and improving and maintaining existing roads. The RTA is also responsible for testing and licensing drivers and registering and inspecting vehicles – it is this area of responsibility that we shall consider.

In 2005 the RTA operated 131 motor registries, a customer call centre located in Newcastle and approximately 80-90 other centres that are either mobile or operate as agencies within regional areas. In NSW there are more than 4.5 million licensed drivers and a similar number of vehicles requiring yearly registration. The RTA operates a system called DRIVES (Driver Vehicle System) that processes all registration and licence transactions.

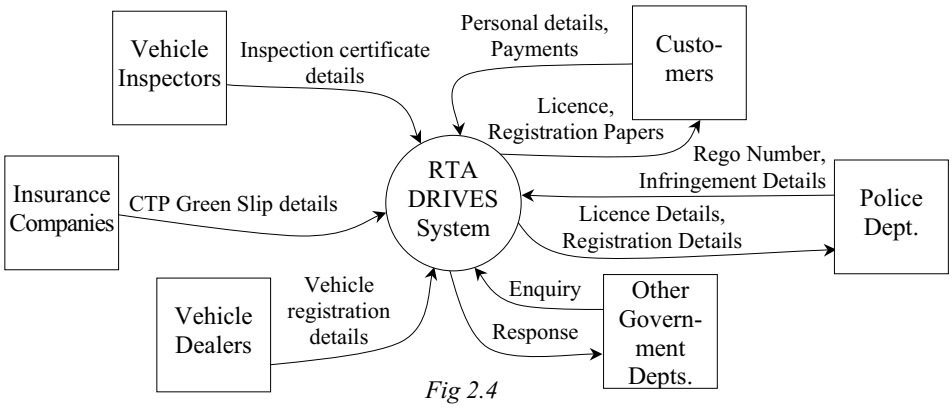


Fig 2.4
Context diagram for RTA DRIVES system.

The context diagram in *Fig 2.4* above describes the significant external entities that either provide or obtain data from DRIVES. For example customers provide their personal details in the form of various "proof of identity" documents when they apply for a driver's licence. Other areas of government, including other sections of the RTA, are able to access DRIVES – for instance statistics on the number of vehicles registered in particular suburbs assists when planning upgrades to the road system or other infrastructure.

GROUP TASK Activity
Classify each of the external entities on the context diagram in *Fig 2.4* as either a source and/or a sink.

Purpose

The purpose of DRIVES includes:

- Maintaining accurate records of all licence and vehicle registrations within NSW.
- Assigning demerit points to licence holders as a consequence of infringements.
- Ensuring the privacy of customer's personal details.
- Providing information to other government departments.

GROUP TASK Activity
Briefly discuss how each data flow on the context diagram in *Fig 2.4* assists DRIVES to achieve its purpose.

Data/Information

Details of each data flow on the DRIVES context diagram in *Fig 2.4* follow:

Data/Information	*Detailed Description*
Personal Details	Name, address, photograph and proof of identity documents.
Payments	Credit card numbers, details for EFTPOS transaction, cash.
Licence	NSW photo licence.
Registration Papers	NSW vehicle registration papers.
Rego Number	Licence plate number used as a unique identifier to determine the vehicle's registered owner.
Infringement Details	Type of infringement and date/time together with the driver's licence number and other personal details. Also includes the vehicle's details.
Enquiry	Various authorised queries for information from the DRIVES database.
Response	Information returned from DRIVES in response to an enquiry.
Vehicle Registration Details	Vehicle dealers submit personal details of each car purchaser together with the vehicle's details for each car sold.
CTP Green Slip Details	Insurance companies inform the RTA directly each time a Green Slip is issued.
Inspection Certificate Details	Pink slip and blue slip details either on paper certificates or transmitted electronically to RTA.

Participants

Most of the information processing within DRIVES is performed by RTA staff, hence these are the most significant participants within the system. The system also allows many of the other users to enter data directly into the system – when this occurs then those people are also participants.

Examples where people other than RTA staff are participants include:

- myRTA website which allows customers to perform a range of transactions online including renewing their registration, changing address and checking their demerit points.
- Dealer online (DOL) system that enables motor vehicle dealers to register vehicles and transfer registrations using the Internet.
- E-safety check system which allows registered vehicle inspectors to electronically transmit pink slip details to the RTA.
- Employees of CTP Green Slip insurers transmit details of each paid Green Slip directly to the RTA system.

Information Technology

The NSW RTA has outsourced responsibility and provision of its data management technology to Fujitsu since 1997. Fujitsu manages the entire NSW RTA information technology environment, which includes DRIVES. The main data centre is currently located in Ultimo (an inner Sydney suburb) where both application software and data is hosted on Sun Fire™ E6900 servers – two of these servers hosting DRIVES. Together these servers support approximately 5500 client computers in some 220 locations throughout the state. The current contract with Fujitsu includes detailed specifications including reliability, response times and recovery times. The E6900 servers assist in this regard as they include inbuilt redundancy for most of their components.

Currently (2007) the client computers used within registry offices are largely Apple G4 iMacs – these were selected because of their ergonomic design and their ability to integrate easily within the Unix-based network. The DRIVES software is a custom application that processes licence and registration data held in an Oracle database accessed via the Sun E6900 servers. Each motor registry workstation includes the iMac computer, a printer, EFTPOS terminal and access to at least one digital camera. The DRIVES software is an integrated application capable of processing EFTPOS transactions, capturing photos and producing licences and of course accessing the main Oracle database.

GROUP TASK Research
Determine the basic specifications of the Sun Fire™ E6900 server and Oracle's database system.

GROUP TASK Discussion
Brainstorm possible reasons why the RTA has outsourced responsibility and provision of its data management technology.

Information Processes

Some of the information processes performed by DRIVES include:

- Renewing vehicle registrations. This includes generating and posting renewal notices, receiving pink and green slip details, processing payments and approving renewals.
- Editing registration details. Includes change of ownership and/or address, collecting stamp duty payments, verifying personal details and creating registration records for new vehicles.
- Issuing new and renewed licences. Includes testing, processing payments, taking photos, verifying personal details and producing photo licences.
- Retrieving and transmitting details of the registered owner of vehicles to police.
- Issuing licence suspension notices when twelve or more demerit points are accumulated within a period of 3 years.

GROUP TASK Discussion
Brainstorm a list of other information processes performed by DRIVES.

Consider the following decision table for renewing vehicle registrations:

Conditions	**Rules**							
Current CTP Green Slip	×	✓	×	×	✓	✓	×	✓
Pink Slip Passed	×	×	✓	×	✓	×	✓	✓
Payment Approved	×	×	×	✓	×	✓	✓	✓
Actions								
Registration Renewed	×	×	×	×	×	×	×	✓
Registration NOT renewed	✓	✓	✓	✓	✓	✓	✓	×

GROUP TASK Activity
Convert the above decision table into an equivalent decision tree.

VIDEO STORES HOLDING INFORMATION ON BORROWERS AND VIDEOS

HSC style question

A small video store records details of its customers and the videos and DVDs they have borrowed using vStore – a software application connected to a database. The store has a single personal computer attached to a cash drawer, bar code scanner and printer. The owner of the store uses the computer to generate various financial and statistical reports from the database. The sales staff use the computer when enrolling new members, processing sales and entering returned movies.

The customers are provided with a membership card that includes a barcode representing their membership number. Similarly each video and DVD has a sticker with a unique barcode. A separate EFTPOS machine is used to process all non-cash payments.

(a) Identify each of the following components in the context of the above information system.
- Purpose
- Participants
- Data/information
- Information technology

(b) Draw a data flow diagram to describe the information system, including the following:
- external entities
- information processes mentioned above
- data flows

Suggested Solution

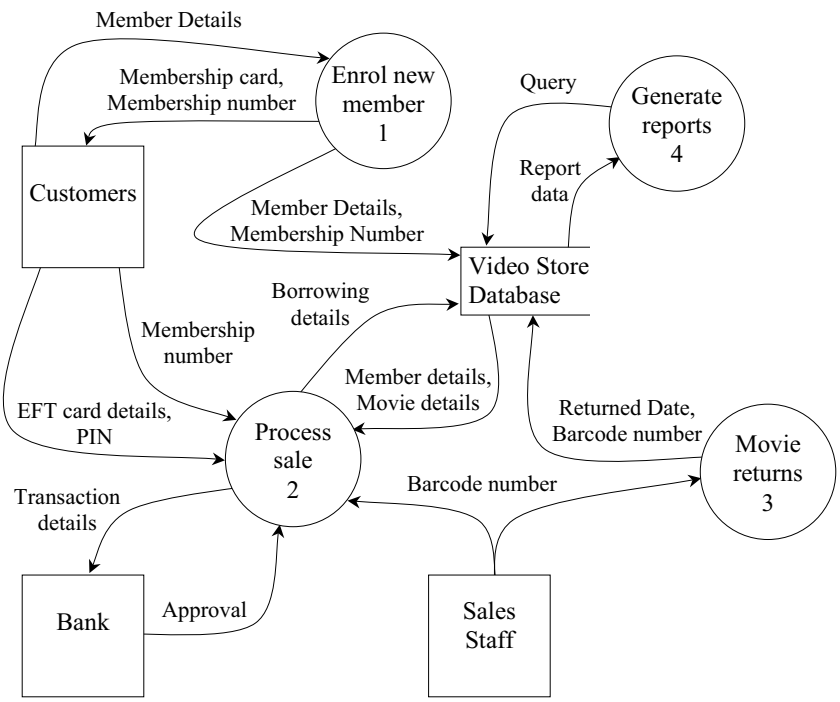
(a) Purpose
- To maintain accurate records of members and the videos and DVDs they borrow and subsequently return including payments made.
- Produce financial and statistical reports for the owner.

Participants
- Owner when generating reports.
- Sales staff enrolling new members, processing sales and entering returned movies.

Data/information
- Customer details including their membership number.
- Details of each video including a unique number/barcode for each.
- Borrowing details including membership number, date borrowed, date for return, unique number for each video and DVD borrowed and payment.
- Financial and statistical reports.
- EFTPOS details including details from customer's EFT cards and approval from bank.

Information Technology
- vStore software, PC, cash drawer, bar code scanner and printer.
- EFTPOS machine, including its connection to bank.

(b)



Comments

- Customers are not participants as they do not directly interact with the information system. The customers are indirect users who provide data to and receive information from the system; hence they are included as an external entity.
- Participants are not included as external entities to the system unless they provide data to the system or receive data from the system. Participants are an integral part of the system as they initiate and perform the systems information processes. These actions occur within the boundaries of the system. In the Video Store question, enrolling new members is performed by the sales staff as part of their role as a participant within the system. The sales staff also scan the actual videos and DVDs to input Barcode numbers into the system, hence in this context the sales staff are included as an external entity.
- The Video returns process uses the barcode number on each video together with the current date to execute an update query that adds the returned date to the record that holds the borrowing details for the video.
- There are other processes that occur within a real video store information system, for example chasing late returns, charging overdue fees, linking family members to memberships, etc. Such processes need not be included as they are not mentioned in the initial scenario.

GROUP TASK Discussion
Do you think the above suggested solution should receive full marks? Justify your response.

GROUP TASK Activity
Create a data dictionary to describe the data noted on each data flow in the above data flow diagram. Include the data type and a brief description.

SET 2A

1. The system's purpose:
   (A) fulfils the needs of those for whom the system is created.
   (B) is the reason the system exists.
   (C) is the aim or objective of the system.
   (D) All of the above.

2. On DFDs, all processes must include:
   (A) a data flow directly to and from an external entity.
   (B) a different data flow entering the process to the one leaving the process.
   (C) at least one data flow which may be either entering or leaving the process.
   (D) a data flow that either enters or leaves a data store.

3. The environment in which a system operates is best described as:
   (A) the hardware and software outside the system.
   (B) the hardware and software within the system.
   (C) all the information technology and processes contained within the system.
   (D) everything that surrounds yet influences or is influenced by the system.

4. Within a school's timetable system an example of knowledge would be:
   (A) student's subject selections.
   (B) student and teacher timetables.
   (C) office staff being able to find any student at any time.
   (D) All of the above.

5. Indirect users are:
   (A) not usually participants.
   (B) usually participants.
   (C) people within the system.
   (D) people who initiate information processes.

6. Which of the following is NOT an example of information technology?
   (A) DBMS server software.
   (B) RAID storage system.
   (C) Executing queries.
   (D) Personal computer.

7. On context diagrams an interface always exists between:
   (A) external entities and data flows.
   (B) external entities and data stores.
   (C) external entities and the system.
   (D) data flows and the system.

8. Examples of unique identifiers within the RTA's information system include:
   (A) driver's licence numbers.
   (B) credit card numbers.
   (C) registration plate numbers.
   (D) All of the above.

9. Within the RTA system described in the text which of the following is true?
   (A) The Sun Fire servers are hardware and the Oracle system is software.
   (B) The Sun Fire servers are software and the Oracle system is hardware.
   (C) The Sun Fire servers and the Oracle system is software.
   (D) The Sun Fire servers and the Oracle system is hardware.

10. Which of the following best describes a transaction?
    (A) A single process on a DFD occurring using actual data to produce particular information.
    (B) Collection of a series of related data items and their subsequent storage.
    (C) The processing of a sale.
    (D) A set of operations that must all occur successfully. If any one operation fails then all other operations are reversed.

11. Define the following terms and provide an example of each:
    (a) participants (b) data (c) information (d) information technology

12. Construct a context diagram to model the system used at an ATM machine.

13. Decompose the *Create Initial Timetable 1* process on the school timetable DFD in *Fig 2.3* into a level 2 DFD. This process collects student subject selections and teacher class selections. It then calculates the number of classes in each subject that should run in the school. Next each class is scheduled, roomed and assigned a teacher; finally students are placed into classes.

14. Describe the sequence of steps performed to renew an individual vehicle's registration from the point of view of the vehicle's owner. Commence from the time the renewal certificate is received until the renewal is complete. Detail various techniques for acquiring green and pink slips, together with the final payment to the RTA.

15. Many schools charge fees based on the subjects each student studies. Consider the generation and processing of subject fees as a separate information system. Identify the participants, data/information and information processes within this system.

ORGANISATION METHODS

Organising
The information process that determines the format in which data will be arranged and represented in preparation for other information processes.

Organising is the information process that prepares data for use by other information processes. It is the information process that determines how the data will be arranged and represented. The aim is to organise the data in such a way that it simplifies other information processes.

For virtually all databases the method chosen to organise the data for storage is critical if the data is to be processed efficiently. This is particularly true for large commercial and government databases that are accessed by many users. The method used is determined as part of the information system's initial development and is difficult to alter significantly once the system is operational. Hence designing the most appropriate method of organisation for a database is vital and becomes more and more so as the quantity of data and number of users increases.

Flat file databases are the simplest form of database. Most non-computer databases are examples of flat-files, for example telephone books, appointment diaries and even filing cabinets. This explains why flat-files were the first to be computerised - they were essentially a direct implementation of existing non-computer databases. Flat-files still remain popular within a variety of simple applications.

Relational databases are used extensively as the data stores for all types of applications. All three of the examples studied in the previous section of this chapter utilised a relational database accessed using a database management system (DBMS) – the school timetable used Microsoft's SQL (pronounced *sequel*) Server and the RTA system used Oracle. Much of our work in the remainder of this chapter involves the theory, design and implementation of relational databases.

Hypertext/hypermedia is based on the connection of related data using hyperlinks. The World Wide Web can be considered as one large hypermedia data store. Web pages are linked together as the author of the page sees fit. Similarly users are free to follow hyperlinks in any direction available. There is very little formal structure, however this does not mean there is no formal method of organisation. There are many rules and protocols to follow if it is all to operate seamlessly.

Consider the following:

Herman Hollerith developed the idea that all United States citizens could be represented by a string of exactly 80 letters and digits. These 80 characters included data that represented each resident's age, address, state and so on. Spaces were added where needed so each field occupied the same number of characters. Hollerith sold his idea, together with his machine and punched cards to the US Census Bureau where it was used to store and tabulate data for the 1890 US census.

Fig 2.5
Herman Hollerith's tabulating machine used during the 1890 US Census.

GROUP TASK Research
Research how Hollerith's machine stored data and tabulated results, and the effect this had on the time required to analyse the 1890 census.

ORGANISATION OF FLAT-FILE DATABASES

A flat-file database is organised as a two dimensional table of data items, hence it can be displayed as a simple table such as the names and date of birth data shown below in *Fig 2.6*. Each row includes all the data about a single individual item and is known as a record or tuple. All records in a table are composed of the same set of attributes – the columns in the table.

Flat-file Database
A single table of data stored as a single file. All rows (records) are composed of the same sequence of fields (attributes).

In *Fig 2.6* each record contains all the data about an individual entity and each individual entity has three attributes – Surname, FirstName and DateOfBirth. Each attribute describes a particular aspect of each individual entity. A particular attribute of a particular record is known as a field, however the distinction between an attribute and a field is seldom observed, rather the terms attribute and field are used interchangeably.

Surname	FirstName	DateOfBirth
Nerk	Fred	15/7/1975
Lamb	Mary	2/4/1955
Jones	John	3/9/2001
Wilson	Julie	28/2/1994
Matthews	Wilbur	19/12/1988

Fig 2.6
Example of a flat-file database.

Each attribute of a particular record is only ever relevant to that record. In *Fig 2.6* it would not make sense to sort on Surname without also rearranging the other fields so they remain with their correct related surname. The underlying organisation of all databases enforce this rule. In reality each record is itself of a particular data type that is composed of individual fields. Indeed all databases process records as complete units of data. This is the most significant difference between the organisation of a spreadsheet and a database. On a spreadsheet each cell is an individual data item that is processed individually. Hence when using a spreadsheet it is possible to sort single columns whilst adjoining columns remain unaltered – this is not possible nor is it desirable within a database.

So far we have discussed the arrangement or structure of a flat-file database, that is, a two-dimensional table containing records that are composed of fields. Consider now how this data is represented or coded. Different fields can contain different types of data, for example in *Fig 2.6* the Surname and FirstName fields contain text and the DateOfBirth field holds dates. Text is composed of a sequence of characters, where each character is represented by an integer using a coding system such as ASCII. Dates are represented as real numbers (usually double precision floating-point) where the whole number portion represents the number of days since some particular date (often 30/12/1899 or 04/04/1904) and the fractional part represents the portion of the day that has elapsed.

The data types used are determined by the software application or database management system (DBMS) used to access the database. Many software applications include the ability to read and write flat-file structures as an integral part of the application. For example, many computer games use a flat-file structure to store player details such as player names, high scores and levels reached. The programmer determines the data types used within such software. Software applications that utilise large amounts of data use a dedicated DBMS, usually a relational DBMS (RDBMS). However, even RDBMSs can be used at a simple level to create and access simple flat-file databases – relational databases, as we shall later learn, are composed of multiple two-dimensional tables. The DBMS includes a collection of available data types and the data type of each field is described within a data dictionary.

Choosing appropriate field data types

The table in *Fig 2.7* describes the general field data types available within most database management systems. Note that different DBMSs use different names for specifying each of these types and most include many different versions of each type. For example, in MySQL there are five differently sized standard integer data types, namely, TINYINT, SMALLINT, MEDIUMINT, INT and BIGINT represented by 1 byte, 2 bytes, 3 bytes, 4 bytes and 8 bytes respectively. In Microsoft Access there are three standard integer types, Byte which predictably uses 1 byte, Integer using 2 bytes, and Long Integer using 4 bytes of storage.

Data Type	Description
Integer	Exact whole numbers (usually both negative and positive).
Decimal/ Fixed-Point	Exact fixed decimal point numbers with limited precision. A scaled version of an integer data type.
Real/Float	Approximate fractional numbers with a very large range.
Money/ Currency	Exact, essentially an integer scaled to have four decimal places and optimised for financial accuracy.
Boolean/ Bit	Yes/No or True/False data.
Date/Time	A number representing the days since (or prior to) a specific date.
Text/Char	String data represented as a sequence of individual characters.
Binary/ BLOB	Raw binary data. Used for storing images, audio or other non-numeric or text data.

Fig 2.7
Summary of common field data types.

So how do you decide upon the most appropriate data type to assign to each field?

Let us first exclude the binary or BLOB (Binary Large Object) data type from our discussion. Selecting a Binary data type is straightforward as, for our purposes, it is only ever used to store image, audio, video and various other data created by other software applications.

Now consider whether a text or numeric data type is needed. Note that apart from Text and Binary all the data types in *Fig 2.7* are classed as being numeric. Some points to consider when making the decision between text and numeric include:

- Do you wish to perform arithmetic operations (addition, subtraction, multiplication, etc.) on the data? If so then a numeric data type should be used. Consider dates. It is common to subtract dates to determine the number of days in between, for example, to calculate someone's age or the number of days an item is overdue. If a date is represented as text then such calculations become extremely difficult.
- Some data is composed of just digits, yet performing mathematical operations does not make sense. For example phone numbers and postcodes are composed of digits yet adding, subtracting or multiplying phone numbers or postcodes is unheard of. In these cases a text data type is a better choice. Furthermore significant leading zeros (0s) appear in both phone numbers and postcodes, for example mobile phone numbers and Northern Territory postcodes.
- The data type assigned to a field determines how the data will be stored and processed, not how it will be formatted for display. Dates and times stored using the DBMS's Date/Time data types can easily be formatted in many ways for display. For example, May 28 2006, 28/5/2006 and Sunday 28th May 2006 are formatted differently but are all the same date hence they should be represented the same. Also Boolean/Bit fields can be formatted for display as Yes/No, Black/White, Orange/Apple, or any other pair of text values.
- Do you wish to sort alphabetically or numerically? Numbers entered into a text field will sort alphabetically. For example, the list 1, 10, 103, 12, 2, 21, 245, 5 is sorted alphabetically, which obviously does not give us the intuitively correct order. Furthermore most date formats sort incorrectly if entered into text fields.

If a text data type is required then consider the following points:

- Is there a limit to the number of characters that will be entered? If so then specify the smallest text data type that will accommodate the data. For example, in Microsoft Access, Text has a maximum length of 255 characters whilst memo fields hold an almost unlimited amount of data but cannot be indexed, searched or sorted efficiently.
- Most DBMSs allow the maximum size or length to be specified. This restricts the number of characters a user can enter (which is a simple form of data validation) and reduces the amount of storage space used. Text data types for most common DBMSs only store the characters actually entered even if the allowable number of characters is greater.
- Unicode or ASCII? Unicode is an extension of the ASCII character set to include many foreign language characters and a variety of other special symbols. Unicode requires 2 bytes to represent each character (2^{16} = 65536 different characters). For most text fields a 1-byte character set is sufficient (2^8 = 256 different characters) – essentially ASCII. In general, unless the field will be storing foreign language characters use a 1-byte text data type and use half the storage space. For example, in MySQL and SQL Server the varchar data type uses 1 byte per character whilst nvarchar uses 2 bytes per character.

If a numeric data type is required then consider the following points:

- Will the values stored always be integers (whole numbers)? If so use an integer data type. Integers use the least amount of storage and are processed faster than the other numeric data types. Furthermore in many larger DBMS products it is possible to specify signed or unsigned integers. For example, a signed 1-byte integer has a range of –128 to 127 whilst an unsigned version has a range from 0 to 255.
- What range of values is required? For integers choose the data type that includes the required range but uses the smallest amount of storage. When real numbers (those that include fractions) are required then the precision of the data type should be considered (see next points).
- Are the values currency or money values? If available use the DBMSs currency data type. This data type has been optimised to ensure the highest level of accuracy for financial calculations. Furthermore currency data types include "Bankers Rounding" (see discussion on next page).
- How precisely or exactly must the numbers be represented? Integer data types store the number entered exactly and should always be used for whole numbers. Both fixed-point and floating-point data types are available for real numbers - floating-point is by far the most common and is almost certainly your best choice. Neither floating nor fixed-point represent all possible real numbers exactly. In general single precision floating-point is accurate to around 7 significant figures and double precision floating-point is accurate to about 15 significant figures regardless of the position of the decimal point – more than enough accuracy for most purposes. Fixed-point representations are scaled versions of integers and hence they represent numbers exactly but with limited precision. For example in Microsoft Access a Decimal data type with precision 4 and scale 2 has a range from –99.99 to 99.99, that is four digits in total with two to the right of the decimal point. Every number within the range that has up to two decimal points is represented exactly, however numbers with more than 2 decimal points simply cannot be represented at all. Fixed-point data types are reserved for specialised applications where a precise range and precision is required.

Consider the following:

Most currency or money data types are a modified form of fixed-point representation. For example in Microsoft Access a Decimal with precision 19 and scale set to 4 has an identical range and can represent exactly the same numbers as the Currency data type. So what is the difference? Currency data types use a system of rounding known as "Bankers Rounding". With Bankers Rounding, values below 0.5 go down and values above 0.5 go up as normal. However, values of exactly 0.5 go to the nearest even number. So 14.5 will be rounded down to 14 and a value of 13.5 will be rounded up to 14. In the case of the Currency data type in Microsoft Access values entered (or more likely calculated) as $1.00135 and $1.00145 are both stored as $1.0014. This occurs to ensure overall fairness in rounding – when working with millions of transactions and billions of dollars it becomes significant.

GROUP TASK Practical Activity
Create an experiment to confirm that "Bankers Rounding" is used for the currency data type within a DBMS with which you are familiar.

GROUP TASK Discussion
Between 0 and 1 there are nine 0.1 steps. Consider how each of these is normally rounded. Discuss your results in terms of the fairness of the "Bankers Rounding" system.

HSC style question

Major changes are planned for a real estate agent's information system. The diagram below represents example data from the rental table in the existing database.

Renter Code	Telephone Number	Postcode	Rent	Occupation Date	Under lease
458703	9123 4567	2056	$230.00	3/12/2005	Y
594223	9567 4321	2057	$395.00	4/3/1999	N
934882	02 4632 2345	2570	$410.58	31/10/2001	N
239922	4589 7654	2690	$195.00	4/3/2006	Y
345533	4322 8933	2856	$240.00	16/7/2006	Y

(a) Construct a data dictionary to describe the data stored in the rental table.

Include the following columns in your data dictionary:

- field name,
- data type,
- storage size, and
- description.

(b) Justify your choice of data types and storage sizes.

(c) Calculate the approximate storage required if the rental table contained 1000 records.

Suggested Solution

(a)

Field Name	Data Type	Storage Size	Description
Renter Code	Integer	3 bytes	Unique code identifying the renter.
Telephone Number	Text	10 bytes	Renters contact telephone number.
Postcode	Text	4 bytes	Renters postcode.
Rent	Money	8 bytes	Weekly rent chanrged.
Occupation Date	Date	8 bytes	Date renter moved in.
Under Lease	Boolean	1 bit	Y means a lease is still current.

(b) Considering each field in turn:

- Renter Code in each example is an integer containing 6 digits. To obtain a range with 6 digits requires a 3 byte integer as the range is then greater than 1 million.
- Telephone Number is text as no maths is done on phone numbers and leading zeros are significant. 10 bytes correspond to the 10 characters needed for the longest phone number in the example.
- Postcode is text as no maths is done and leading zeros are possible (e.g. NT postcodes). Each character requires 1 byte of storage, hence 4 bytes are needed.
- Rent is an amount of money hence the DBMSs specific currency/money data type should be used. In most database such types require 8 bytes of storage.
- Occupation Date is clearly a date and should be stored as such so that maths can be performed and the format adjusted to suit different needs. Date formats commonly require 8 bytes (as they use double precision floating-point).
- Under Lease is a Yes/No field hence just a single bit is needed to store either a 1 or a 0.

(c) 1 record = (3 + 10 + 4 + 8 + 8) bytes + 1 bit
= 33 bytes + 1 bit

1000 records = 33000 bytes + 1000 bits
= 33 KB (Approximately)

(Note: there are 1024 bytes per KB hence the extra 1000 bits are included within the extra 24 * 33 bytes).

Comments

- The storage sizes will depend on the databases with which you have had experience, however they should not vary significantly from those in the suggested solution above.
- Justifications of storage size should address the length of text fields and the range for numeric fields.
- It is reasonable to assume all text fields require 1 byte per character. It is uncommon for Unicode 2 byte per character text to be used.
- For questions like part (c), first calculate the storage required for a single record and then multiple by the total number of records.
- At the time this book was printed no calculators were permitted in IPT HSC examinations. As a consequence approximations could be asked, the question may requires only simple arithmetic or you could be asked to "show how you would calculate...", which expects full working to be shown without the need to actually calculate the final answer.

NON-COMPUTER EXAMPLES OF FLAT-FILES

Most non-computer databases are really flat-file structures that are permanently ordered according to one or more fields. It is unusual for them to be physically stored as a two-dimensional table. Rather it is more common for each record to be stored individually on a piece of paper, card or within an individual file. Separating records in this way makes it far simpler to add and delete records without destroying their order – a new record can be inserted between two existing records or it can be physically removed.

Consider the following examples of non-computer databases.

- Many small business offices maintain a filing cabinet that contains a folder for each customer. The folders are physically ordered alphabetically by the most commonly accessed field – usually surname or company name. Each folder includes various documents that contain individual data items describing different aspects of each customer.
- Telephone books use enormous amounts of paper, yet virtually every household and business throughout the world receives a new telephone book, or set of telephone books each year. In Australia two sets of telephone books are distributed; the White Pages, which is arranged alphabetically by surname, and the Yellow Pages, which is arranged into business categories and then alphabetically within each category.
- Card catalogues were until recently used in libraries. The books are physically arranged on the shelves by their call numbers with at least two separate card catalogues being maintained. One catalogue was sorted by title and the other by author; when a new book was added to the collection a new card was added to each card catalogue.
- Salesmen commonly maintain a card system to track their sales leads (potential customers). Each card contains the details of each lead and all the cards are stored within a box on their desk (see *Fig 2.8*).

Fig 2.8
Typical card based system

- Many reference books are organised similarly to flat-files. For example recipe books, encyclopaedias and even computer programming language reference texts.

GROUP TASK Discussion
Each of the above examples is organised similarly to a flat-file database. For each of the above examples, explain the organisation of the data in terms of records and fields.

GROUP TASK Discussion
Computerised versions of each of the above examples are freely available. List possible reasons people still prefer the non-computerised versions.

SET 2B

1. The organising information process:
   (A) transforms data into information.
   (B) represents data on physical storage media.
   (C) arranges and represents data in a form suited to further processes.
   (D) only occurs during the design of information systems.

2. Rows in a flat-file database are also known as:
   (A) attributes or fields.
   (B) fields or tuples.
   (C) records or tuples.
   (D) attributes or records.

3. Columns in a flat-file database are also known as:
   (A) attributes or fields.
   (B) fields or tuples.
   (C) records or tuples.
   (D) attributes or records.

4. Sorting an individual column in a table without affecting the order of other columns is possible when using a:
   (A) flat-file database.
   (B) spreadsheet.
   (C) DBMS.
   (D) RDBMS.

5. The most suitable data type for storing post codes is:
   (A) Integer.
   (B) Fixed-point decimal.
   (C) Text.
   (D) Boolean.

6. Which of the following has the least amount of formal organisation?
   (A) flat-file database
   (B) relational database
   (C) spreadsheet
   (D) hypermedia

7. HSC marks in 2 Unit courses are whole numbers within the range 0 to 100. The best data type for storing these marks would be:
   (A) 4 byte integer.
   (B) 3 byte text.
   (C) double precision floating-point.
   (D) 1 byte integer.

8. In regard to floating and fixed-point representations, which of the following is FALSE?
   (A) Fixed-point has a much smaller range than floating-point.
   (B) Fixed-point is exact for the numbers it can represent.
   (C) Floating-point represents many numbers approximately.
   (D) Floating-point data types are really scaled integers.

9. A flat-file contains 300 tuples. There are 5 attributes and each attribute holds integers in the range 0 to 65535. What is the approximate size of this file?
   (A) 3KB
   (B) 3Kb
   (C) 600B
   (D) 1500B

10. Which of the following is true in regard to the data type used for dates?
    (A) A text data type should be used so they can be entered in the desired format.
    (B) A number data type should be used so they can be sorted and processed numerically.
    (C) Using a text data type means the format can be more easily changed to suit the system's requirements.
    (D) Dates should be stored as three separate integer fields – one each for day, month and year.

11. Define each of the following terms:
    (a) Flat-file database. (b) Data type. (c) Record (d) Attribute

12. Explain why phone numbers and postcodes are commonly represented using text data types.

13. Compare and contrast:
    (a) Fixed-point and Integer data types. (b) Floating and fixed-point data types.

14. Design and create a flat-file database to store the details of each of your HSC assessment tasks.

15. "Many people still continue to use paper-based flat-file systems despite owning computers and flat-file software." Explain reasons this is so. Include examples as part of your explanation.

RELATIONAL DATABASES

In simple terms a relational database is a collection of two-dimensional tables, where the organisation of each table is almost identical to a simple flat file database. All information processes within a relational database system are performed on tables. This is what a relational database management system (RDBMS) does; it performs information processing on the tables within relational databases. This includes processes performed on the data as well as processes that create and modify the design of the tables. Currently the large majority of computer-based databases conform to the relational model, however other database models exist, such as the hierarchical and network models.

GROUP TASK Research
Research, using the Internet or otherwise, the general method of organisation used within hierarchical and network database models.

So what is it about relational databases that make them such a popular choice? Clearly all databases are designed to store data, however the main problem with data is that it keeps changing over time – new records are added, existing records are changed and deleted and even changes to the underlying structure are made. Relational databases include mechanisms built into their basic design to make such processes as painless as possible. As we study the logical organisation of relational databases we shall introduce many of these mechanisms. At times our discussion will become quite theoretical; remain focussed and keep asking, "how does this assist the processing of the data?"

Before we commence our discussion on the logical organisation of relational databases we need a general understanding of the role DBMS software performs within information systems. We have already mentioned some examples of DBMSs, namely Microsoft Access, MySQL, Oracle and SQL Server – these are all relational DBMSs (RDBMSs) but there are many others. It is likely that you interact with one or more RDBMSs every day of your life without even being aware. For example a RDBMS is operating when using an ATM, chatting on the Internet, using a search engine or looking up references in the school library.

GROUP TASK Activity
Brainstorm a list of activities you have performed this week that are likely to have included interaction with a relational database.

RDBMSs commonly operate between software applications and the actual relational database (see *Fig 2.9*). A command is created by the software application and passed to the RDBMS, the RDBMS checks the user has permission, performs the processes required to carry out the command on the database and sends a response back to the software application. The response maybe as simple as an acknowledgement that the command was executed or it may be a series of records retrieved from the database. In most modern RDBMSs the commands are issued in the form of SQL (pronounced *sequel*) statements.

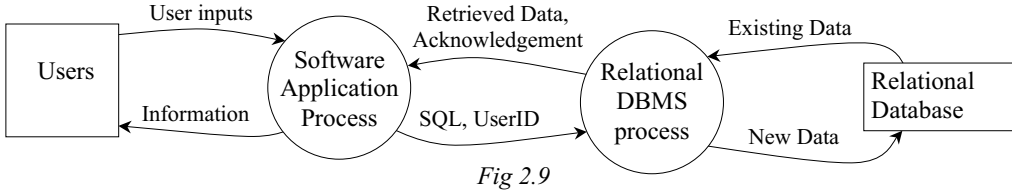


Fig 2.9
RDBMS operate between software applications and relational databases.

Throughout the discussion that follows we will use Microsoft Access as an example. Access is a true relational DBMS however it also includes the ability to design and execute data entry forms and hardcopy reports (amongst other things). By default Access creates a single file that includes the tables, queries, forms and reports. This file can be shared across a network, however each user executes their own copy of the Access DBMS. This is not the case with server based DBMSs such as MySQL, SQL server or Oracle. Server based DBMSs execute on a server and each user executes their own client software application. The client application creates all the pretty stuff – the data entry screens and the nicely formatted reports. The database server looks after the data access and maintains the actual database.

The client software application, and the DBMS controlling the database are independent of each other. In many cases a variety of different client software applications access data from the same database via the DBMS. The DBMS controls access to the data and also organises the data into a form suited to each client software application. In effect the use of a DBMS means the organisation of the database is independent of the organisation required by different client applications.

THE LOGICAL ORGANISATION OF RELATIONAL DATABASES

Let us consider the key concepts in regard to the organisation of relational databases as a series of points – we consider tables, primary keys, relationships and the concept of referential integrity. During our discussion we will illustrate each point using examples from a library database created with Microsoft Access.

Tables

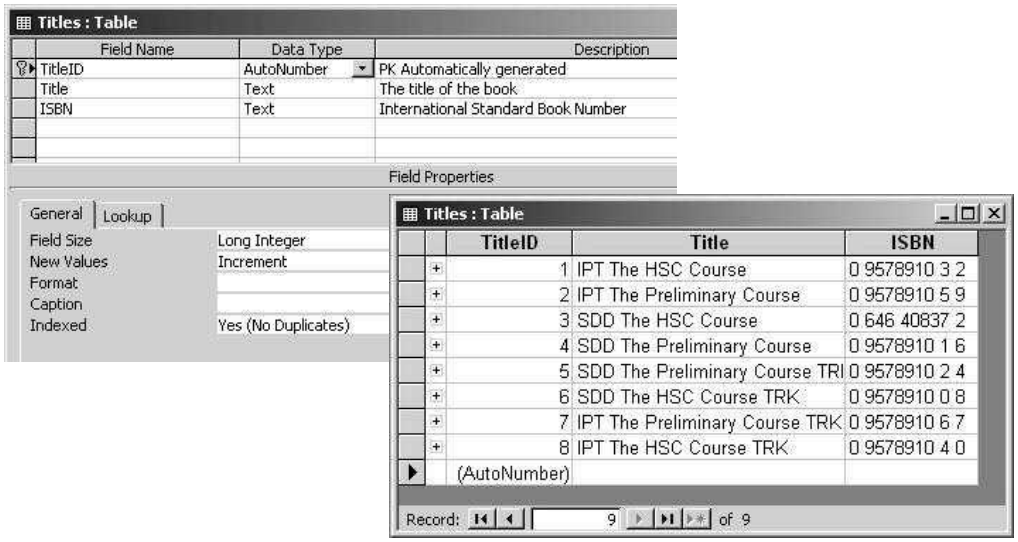


Fig 2.10
Data dictionary and some example data in the Titles table of the library database.

The basic building block of all relational databases is the table. Some key concepts involving tables include:

- Each table is a set of rows (records) and columns (fields). There is no predefined order to the rows or the columns, that is, theoretically the data does not reside or appear in any defined order. In *Fig 2.10* above the Titles table contains 8 records and 3 fields. The order of the records and the order in which the fields appear is not significant. Notice that none of the field names contain spaces; this simplifies the writing of SQL statements.
- A single table is very much like an individual flat-file database. It is composed of records, which are composed of fields. Each record in a table has the same set of

fields. Each of the 8 records in *Fig 2.10* has the same set of three fields, namely TitleID, Title and ISBN.

- Records are also known as tuples and fields are also known as attributes. In *Fig 2.10* each of the 8 records is also called a tuple. Each tuple has a TitleID, Title and ISBN attribute.
- Tables are also known as entities or relations. The term entity is used as each row in each table describes all the data about a particular individual entity. The word relation means an association between two things, in this case there are two-dimensions – the rows and the columns.

  The Titles table is also an entity or a relation. Each row in the Titles table describes all the data about a particular title. Significantly the table does not contain data about each title's author as many authors write many books – including the author in the titles table would introduce redundant data.
- Each record within a table is unique, that is, there are never two records where the contents of all fields are identical. It is not possible for more than one record in the Titles table to have the same TitleID hence all records are unique.

Primary Keys

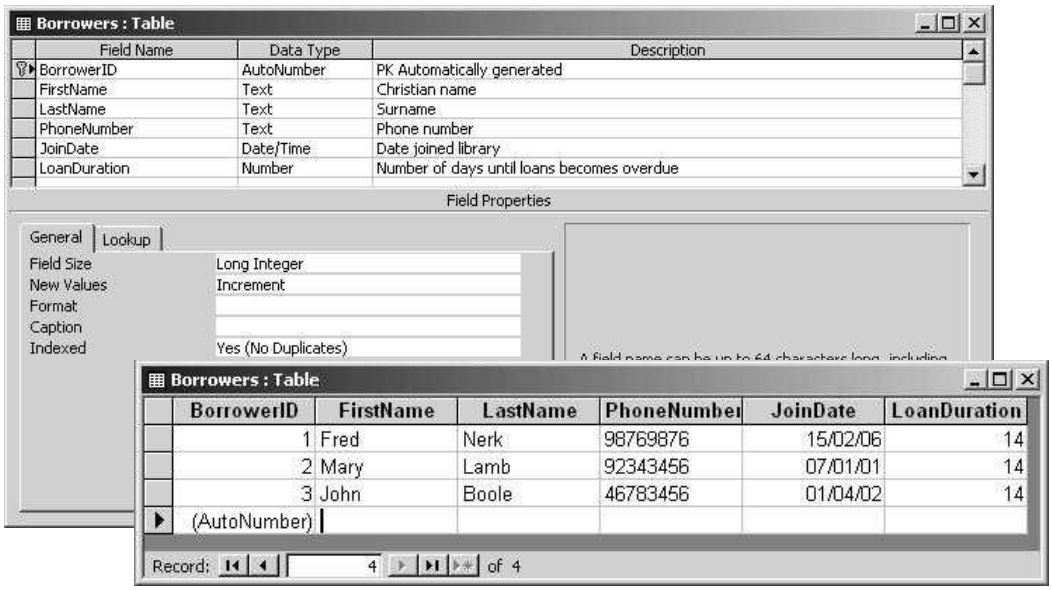


Fig 2.11
Data dictionary and some example data in the Borrowers table of the library database.

Every table within a relational database must have a primary key (PK) – a field or combination of fields that uniquely identifies each record. Key concepts in regard to primary keys include:

- Any single field or combination of fields that uniquely identifies a record is called a candidate key. One candidate key is selected as the primary key (PK) for the table.

  In the Titles table described in *Fig 2.10* TitleID and ISBN are candidate keys. In this table TitleID has been defined as the PK – in Access this is indicated by the key symbol to the left of the field name. Title may appear to be a candidate key however it is possible, and actually quite likely, that two different books will have the same title.

  In the Borrowers table described in *Fig 2.11* BorrowerID is a candidate key and so too is a combination of FirstName, LastName combined with either PhoneNumber and/or JoinDate. In this case BorrowerID has been selected as the PK. FirstName

and LastName are never good choices for a PK as it is not uncommon for two people to have the same name.

- It is usually more convenient to use a single integer field as the primary key. Often the primary key is a new field created specifically for this purpose. Commonly an integer data type is used together with a DBMS feature that automatically generates unique numbers.

  TitleID in the Titles table is an autonumber PK field and so too is BorrowerID in the Borrowers table. Both these PK fields increment automatically for each new record. An alternative is to generate the unique integer values randomly. This random strategy is used to generate the LoanID PK within the Loans table, which explains the somewhat odd looking values for LoanID within the example data in *Fig 2.12*.

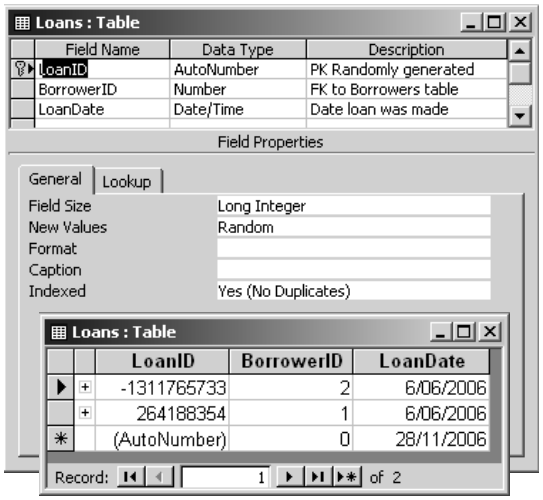


Fig 2.12
Data Dictionary and example data in the Loans table of the library database.

- When more than one field is used as the primary key it is called a composite key.

  In the BookLoans table described in *Fig 2.13* both the LoanID and the BookID fields combine to form the primary key so they form a composite key. The BookLoans table will make more sense once we have discussed relationships and viewed the entire schema for the Library database.

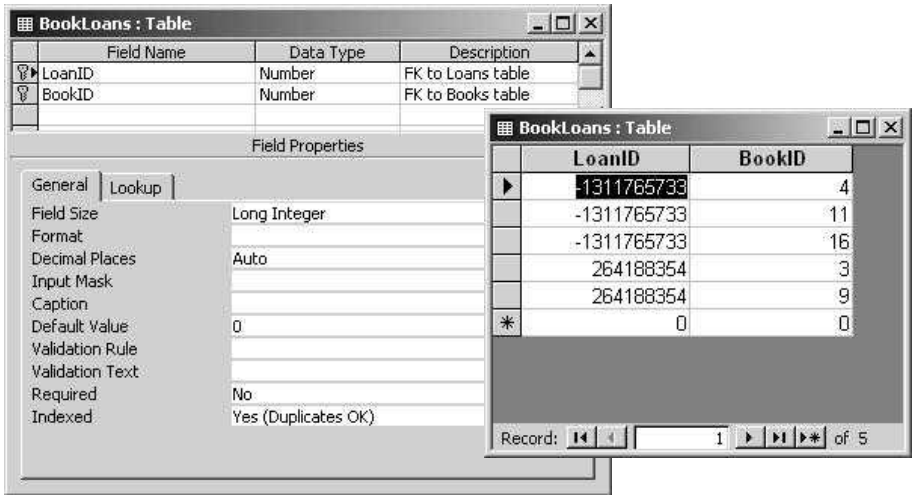


Fig 2.13
Data dictionary and some example data in the BookLoans table of the library database.

Consider the following:

The following Access SQL statement when executed creates the basic structure of the Borrowers table described previously in *Fig 2.11.* This is an example of a Microsoft Access DDL (Data Definition Language) SQL statement. In Access the simplest method of entering and executing DDL SQL is via the SQL view of a query.

```
CREATE TABLE Borrowers (
    BorrowerID COUNTER PRIMARY KEY,
    FirstName Text(30),
    LastName Text(50),
    PhoneNumber Text(10),
    JoinDate Date,
    LoanDuration Byte)
```

GROUP TASK Practical Activity
Use a DBMS to execute the above DDL SQL. If not using MS Access then some adjustments to the data type specifications will be required.

GROUP TASK Discussion
In Access, tables are normally created using the built-in user interface. Describe real scenarios where DDL SQL statements would be used.

Relationships

Tables are linked together via relationships. A relationship creates a join between the primary key in one table and a foreign key in another. Each of the tables together with their relationships to each other is modelled using a schema – *Fig 2.14* shows the initial (not complete) schema for our library database.

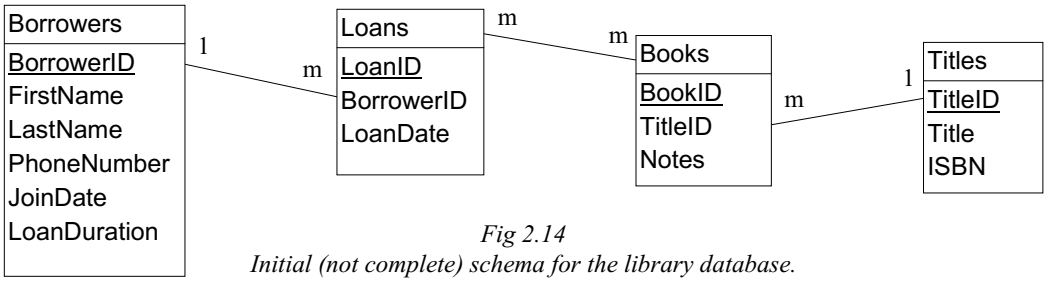


Fig 2.14
Initial (not complete) schema for the library database.

Some key concepts in regard to relationships include:

- Database schemas or schematic diagrams are a technique for modelling the relationships within a relational database. Schemas include each entity (table) together with its attributes (field names). The primary key field (or fields) is underlined. Lines between attributes represent the relationships or joins between tables. Each relationship line is labelled to indicate the nature of the join.

  In *Fig 2.14* the schema includes four tables and three relationships or joins. Schemas are commonly called Entity Relationship Diagrams (ERDs). Schemas and ERDs are not strictly the same thing; for our purposes the distinction is not important, it is sufficient to consider a schema as a type of ERD. There are various other techniques for constructing ERDs, however the information being modelled is essentially the same.

GROUP TASK Research
Using the Internet, or otherwise, find at least one example of an ERD that uses a different technique to database schemas.

- Foreign keys (FK) are fields that contain data that must match data from the primary key (PK) of another table. Hence the data type of a foreign key must always match the data type of the related table's primary key.

  In the above schema the BorrowerID field within the Loans table is a foreign key (FK) as it forms one side of the join to the PK of the Borrowers table.

- By far the most common type of relationship is one to many (1:m). This means that for each record in the primary key's table there can exist multiple records in the foreign key's table. There are two 1:m relationships present in *Fig 2.14*.

  The join between the Borrowers table and the Loans table means that an individual borrower can have many loans, however each loan can only have a single borrower. For example Fred can visit the library and borrow books on say Monday and then again on say Friday. However, Mary and Fred cannot borrow books together – each loan must be recorded against a single borrower.

  The join between the Books table and the Titles table means there can be many books that are the same title. Note that in our example database a record in the Title table describes a particular published title – the library may have no copies of the title or it may have one or more copies. The Books table includes a record for each copy of a book the library actually owns. If the library has 10 copies of a particular title then there will be 10 records in the Books table – each of these 10 records are related to the same single record in the Titles table.
- One to one (1:1) relationships are seldom required, however there are some situations where they are included to improve performance and reduce storage. A one to one join means that at most one record from table A is associated with one record from table B.

  When a one to one relationship is detected then it is always possible to include all the attributes from both tables into a single table. For example employee's names could be held in one table and their date of birth in another with a 1:1 relationship joining the two tables. Both tables can be combined into a single table that includes attributes for employee names and date of birth.

  There are real situations where 1:1 relationships should remain. Consider the partial schema in *Fig 2.15*. In this database some employees have an office whilst others do not, however an individual office can only ever be occupied by one and only one employee.

  Lets say some company has 100 employees and 20 offices. There will therefore be 100 records in the Employees table and just 20 records in the EmployeeOffices table. If the attributes from the EmployeeOffices table where included in the Employees table then 80 employee records would contain NULL entries within the new attributes. Furthermore, and more significantly, reassigning employees to offices would be more difficult. The OfficeName and OfficeLocation data needs to be removed from the existing employee assigned the office and then this data must be re-entered within the new employee's record. The structure in *Fig 2.15* means the EmployeeID in the EmployeeOffices table is simply edited to reflect the newly assigned employee.

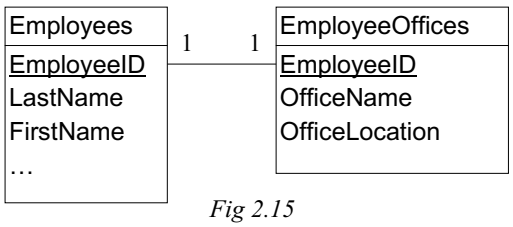


Fig 2.15
Example of a one to one relationship.

- Many to many (m:m) relationships must be resolved by creating a join table with two 1:m relationships. The new join table must contain foreign keys to both the primary key fields within the original two tables. Together these fields form the primary key (actually a composite key) within the new table.

  In the initial schema for the library database (*Fig 2.14*) a many to many relationship exists between the Loans table and the Books table. This m:m join means that many books can be associated with each loan and also each book can form part of many loans. In theory this sounds fine, indeed this is exactly what we

wish to occur, unfortunately this initial structure cannot be implemented directly (this is why in *Fig 2.14* the join merely points to each table's name).

Let us consider this problem more deeply. Each book can form part of many loans – surely this is a 1:m relationship so we could add a BookID (FK) to the Loans table. Also each loan can contain many books – another 1:m relationship so we could add LoanID (FK) to the Books table. These possibilities are shown in *Fig 2.16* but they are incorrect – they don't work in practice. They also mean redundant (duplicate) data is stored because data in regard to each loan and book combination is stored twice. This will cause significant problems, as both sets of data must always be updated together.

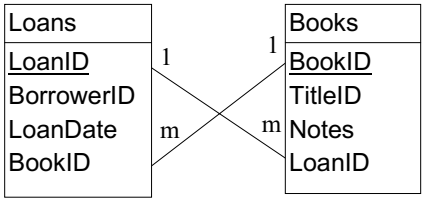


Fig 2.16
Incorrect m:m solution.

Furthermore, consider what would happen when many books are borrowed as part of a single loan. Say the LoanID is 10 and the BookIDs are 101,102,103 and 104. We require 4 loan records to store the 4 BookIDs – if this occured then the uniqueness of the LoanID primary key would be violated. In addition the LoanDate would be stored 4 times – another example of redundant data. Maybe we could add four BookIDs to the Loans table? But what if someone borrows 10 books or 15? This is not a good idea and it also destroys the ability to efficiently access the data. So adding the BookID (FK) to the Loans table is out of the question.

Now consider storing the LoanID as a FK in the Books table. This would correctly identify the Loan and thence the Borrower of each book whilst it is out of the library. When a book is in the library the LoanID (FK) could be set to NULL. When a book is loaned again then the Book record is updated so the LoanID (FK) matches the LoanID in the Loan table. So what's the problem? We are not maintaining any records of previous loans for any books. In essence we have lost the "each book can form part of many loans" part of the m:m join. We can't query to find out the most popular or unpopular books, the average number of books borrowed or any other information that requires historical data on books borrowed.

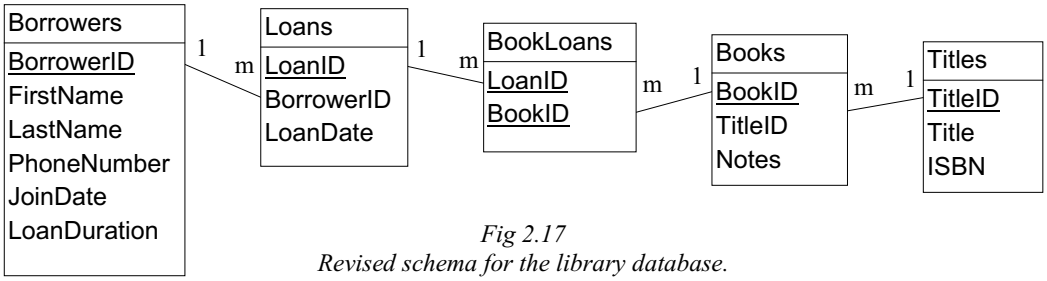


Fig 2.17
Revised schema for the library database.

The solution is to create a new table connected to each of the existing tables using 1:m joins. This new table includes foreign keys to each of the primary keys in the existing tables. In our example we create a table called BookLoans that includes LoanID and BookID attributes. These two attributes combine to form the primary key of the new table. The revised schematic diagram is shown above in *Fig 2.17*.

GROUP TASK Discussion
Identify and describe all the records involved in the processing of a single loan where say 4 books are borrowed.

- Recursive relationships are permitted. This occurs when an attribute of a table is joined to the primary key within the same table.

  For example in *Fig 2.18* each employee is assigned a single manager, however each manager is also an employee. Notice that a single employee can manage many employees, therefore a 1:m join exists between EmployeeID (PK) and ManagerID (FK) attributes.

  As an aside, notice that for recursive relationships the FK has a different name to the PK – clearly necessary, they're in the same table! In all our other examples we have used the same field names for either side of each join. This is not a necessity, rather it simply makes the design clearer when designing queries and SQL statements.

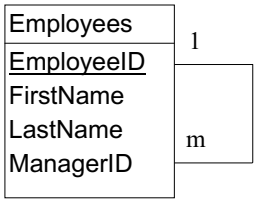
  

  Fig 2.18
  Recursive relationship.

  To create a recursive relationship in the Microsoft Access Relationships window it is necessary to add the table to the window twice and create the join between the original and the copy.

Referential Integrity

- Referential integrity is ensured when each foreign key always matches a related primary key. The only exception is NULL values if they are permitted in the foreign key. RDBMS include mechanisms that enforce referential integrity.

  In our library example every Book record must always have a TitleID (FK) that matches one of the TitleIDs of a record in the Title table. In other words we wish to enforce referential integrity and we do not want to allow NULL values in the FK. By default RDBMSs enforce referential integrity, however it is necessary to specify that NULL values are not permitted. Note that a NULL value means no data at all – an empty string is not a null and neither is a zero value. In Microsoft Access setting the *required* property of a field to Yes prevents NULL values (see *Fig 2.19*).

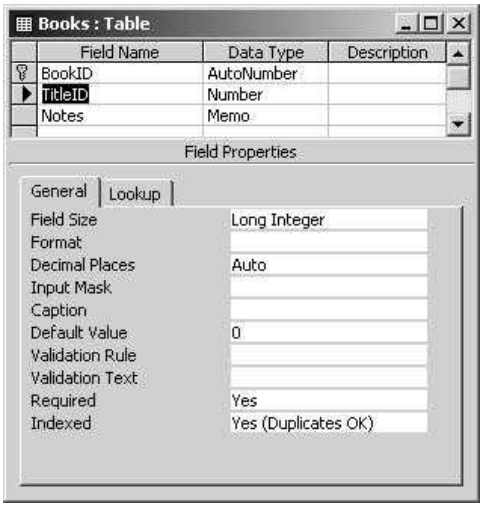
  

  Fig 2.19
  NULL values are not permitted when the required property is set to Yes.

  Referential integrity is enforced by default within relational databases. If this were not the case then over time records with foreign key values would exist in the database with no associated primary key in the parent table. It is difficult to imagine a situation where referential integrity should not be enforced.

- Two issues result from enforced referential integrity that need to be resolved – what to do if a primary key is updated (changed) and what to do if a primary key record is deleted when related records exist?

  The update problem occurs when the primary key on the 1 or parent side of a relationship is altered. Without enforced referential integrity this would result in foreign key values that are orphaned (i.e. they have no parent record). Enforced referential integrity solves this problem in two possible ways. Either the change to the primary key is simply not allowed (generating an error message) or the foreign keys of all related records are automatically updated to match the new primary key value. In MS-Access these two solutions are implemented by either not selecting

or selecting the "Cascade Update Related Fields" option within the Edit Relationships dialogue (refer *Fig 2.20*).

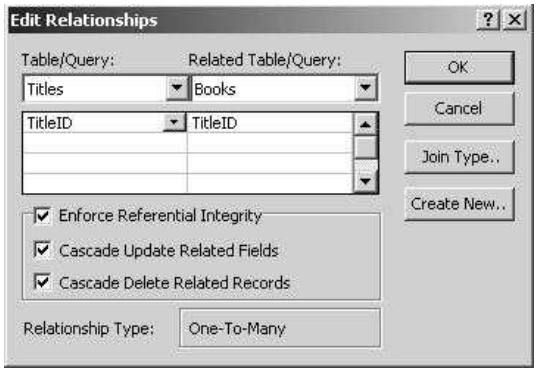


Fig 2.20
Edit Relationships dialogue in Microsoft Access.

The delete problem occurs when a record on the parent side of a relationship is deleted. Without referential integrity enforced this would result in orphaned records. Normally referential integrity is enforced and as for the update problem there are two possible strategies. Either don't delete the record or also delete all the related records. In either case it would be wise to inform the user. MS-Access implements these two strategies by either not selecting or selecting "Cascade Delete Related Records" in the Edit Relationships dialogue (refer *Fig 2.20*).

Consider the following sample data in the library database:

Titles : Table

	TitleID	Title	ISBN
+	1	IPT The HSC Course	0 9578910 3 2
+	2	IPT The Preliminary Course	0 9578910 5 9
+	3	SDD The HSC Course	0 646 40837 2
+	4	SDD The Preliminary Course	0 9578910 1 6
+	5	SDD The Preliminary Course TRI	0 9578910 2 4
+	6	SDD The HSC Course TRK	0 9578910 0 8
+	7	IPT The Preliminary Course TRK	0 9578910 6 7
+	8	IPT The HSC Course TRK	0 9578910 4 0
*	(AutoNumber)		

Record: 1 of 8

BookLoans : Table

LoanID	BookID
-1311765733	4
-1311765733	16
-1311765733	22
264188354	3
264188354	9
0	0

Record: 1 of 5

Loans : Table

	LoanID	BorrowerID	LoanDate
+	-1311765733	2	06/06/06
+	264188354	1	06/06/06
*	(AutoNumber)	0	08/06/06

Record: 1 of 2

Books : Table

	BookID	TitleID	Notes
+	1	1	
+	2	1	
+	3	1	
+	4	1	
+	5	1	
+	6	2	
+	7	2	
+	8	3	
+	9	3	
+	10	3	
+	11	3	
+	12	3	
+	13	3	
+	14	3	
+	15	3	
+	16	3	
+	17	3	
+	18	3	
+	19	4	
+	20	4	
+	21	2	
+	22	2	
+	23	1	
+	24	1	
+	25	1	
+	26	2	
+	27	3	
*	(AutoNumber)	0	

Record: 1 of 27

Borrowers : Table

	BorrowerID	FirstName	LastName	PhoneNumber	JoinDate	LoanDuration
+	1	Fred	Nerk	98769876	15/02/06	14
+	2	Mary	Lamb	92343456	07/01/01	14
+	3	John	Boole	46783456	01/04/02	14
*	(AutoNumber)					

Record: 1 of 3

Fig 2.21
Sample data within each table of the library example database.

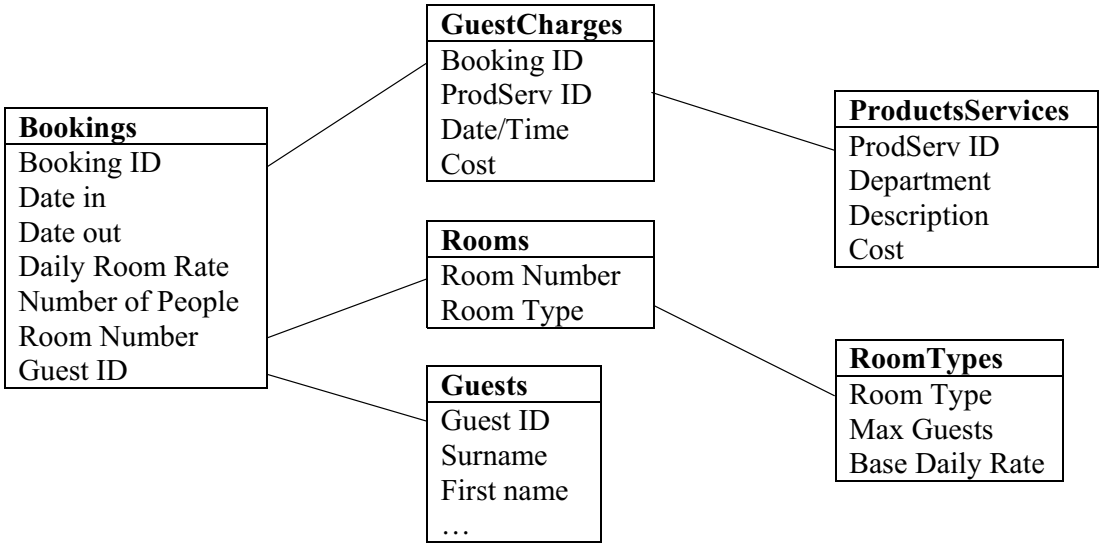
GROUP TASK Activity
Analyse the above sample data to determine the title of each of the books Fred Nerk has borrowed.

GROUP TASK Discussion
If Fred Nerk's borrower record was deleted then which other records would also be removed if "Cascade Delete Related Records" is selected for all relationships.

GROUP TASK Discussion
Currently we are not recording when books have been returned. Identify possible alterations to include such data. Justify the best alternative.

HSC style question

The Grandview Hotel utilises a database to store and process all data required during a guest's stay at the hotel. The schema for the parts of the hotel database required to produce each guest's final account is shown below:



(a) (i) The above schema does not indicate the nature of each of the Relationships. Add this information to the above schema.

(ii) A Cost field is included in both the GuestCharges and ProductsServices tables. Using examples, explain why both these fields are needed.

(iii) Explain how the total cost of a guest's visit can be calculated at the conclusion of their stay.

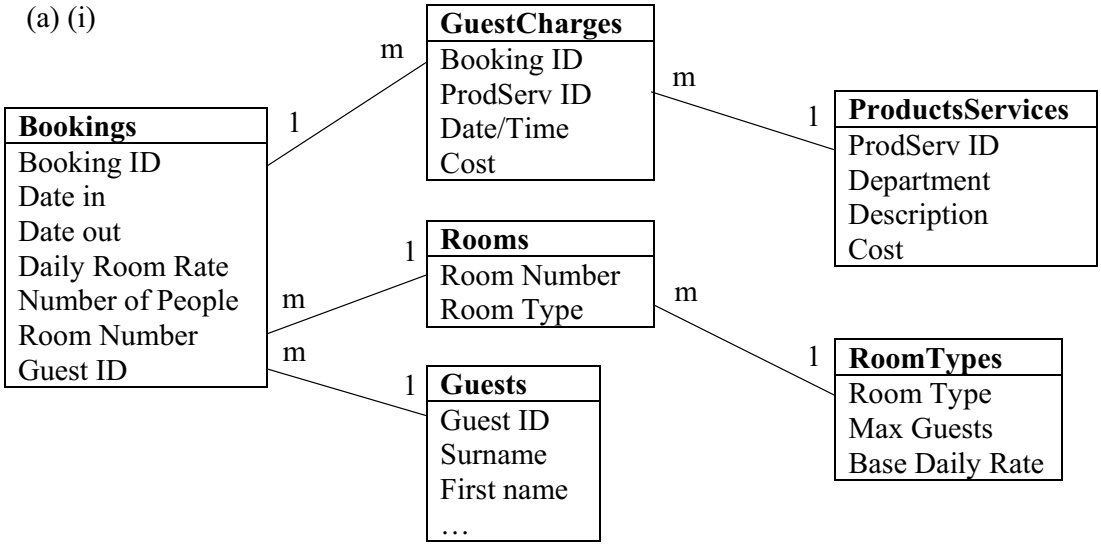
(b) An exclusive 'Grand Members Club' is being introduced to encourage frequent guests to increase their spending whilst at the hotel and also to increase their visits to the Hotel.

The 'Grand Members Club' will offer a fast check-in service, as well as a variety of different discounts on the hotel's other products and services. A club newsletter will be distributed to members each month detailing the different percentage discounts being offered for different products and services.

Propose and justify suitable modifications to the database schema so that the appropriate discounts can be applied when a member's final account is being generated.

Suggested Solution

(a) (i)



(a) (ii) The cost field in the ProductServices table is the current default cost for that product or service. This is likely to change over time as prices rise. When this default cost is altered it would be wrong for past guest's charges to also change. Also some products and services may have their cost modified for a particular guest. For example in a restaurant a guest may wish to order extra chips which incurs an extra $2 charge. Having the cost field in the GuestCharges table allows such changes to be made without affecting other guest's charges or the normal cost for the item.

(a) (iii) The total number of days is calculated using the Date in and Date out fields. This total is multiplied by the Daily Room Rate (not the Base Daily Rate) to produce the total room cost.

The total guest charges are calculated by adding the Cost field of all GuestCharges records that match the Booking ID for the Guest's current visit.

The sum of the total room cost and total guest charges is the total cost of the guest's visit.

(b) Modifications could include:

- An additional field added to the Guest table to indicate membership of the club. This field could be a simple Boolean type or it could contain some sort of membership number.
- A field added to the ProductsServices table called ClubDiscount. This field would contain a real number from 0 to 1 (0 being the default) indicating the percentage discount on this item for members.
- When the final account is being produced a check should be made to see if the Guest is a club member. If so then each GuestCharge.Cost for the current booking must be reduced by the discount within the corresponding ProductsServices.Discount field.

SET 2C

1. Examples of RDBMS include:
   (A) Microsoft Access, SQL Server.
   (B) MySQL, Oracle.
   (C) Oracle, Microsoft Access.
   (D) All of the above.

2. Within relational databases a set of rows that all have the same attributes is called a:
   (A) primary key.
   (B) table.
   (C) relationship.
   (D) record.

3. A candidate key is:
   (A) one or more fields that uniquely identify each row.
   (B) the same as a primary key.
   (C) one or more fields within a table.
   (D) a record that could be used as the primary key.

4. If the primary key in one table is joined to the primary key in another table, which type of relationship would be formed?
   (A) 1:m
   (B) m:m
   (C) 1:1
   (D) Two 1:m relationships.

5. Alternative names for tables, records and fields respectively are:
   (A) files, entities and attributes.
   (B) entities, attributes and tuples.
   (C) relations, tuples and attributes.
   (D) tuples, relationships, keys.

6. A composite key is an example of a:
   (A) foreign key.
   (B) primary key.
   (C) relationship.
   (D) candidate key.

7. With regard to relationships, which of the following is true?
   (A) They join a primary key in one table to a candidate key in another table.
   (B) They join a foreign key in one table to a candidate key in another table.
   (C) They join a primary key in one table to a composite key in another table.
   (D) They join a primary key in one table to a foreign key in another table.

8. In relational databases, how is a many to many join created?
   (A) Join the primary key in one table to the foreign key in the other table and vice versa.
   (B) Join the primary keys in each table together.
   (C) Create a new table containing foreign keys to each existing table. The foreign keys link back to the primary keys in the existing table.
   (D) Any of the above is possible depending on the requirements of the database system.

9. The mechanism within RDBMSs that ensures each FK matches a PK is called:
   (A) referential integrity.
   (B) a database schema.
   (C) a recursive relationship.
   (D) a one to many relationship.

10. Which type of relationship is most commonly used in relational databases?
    (A) one to one.
    (B) one to many.
    (C) many to many.
    (D) Each of the above relationships is equally likely.

11. Define each of the following terms and provide an example of each.
    (a) Table (b) Primary key (c) Foreign key (d) Relationship

12. Compare and contrast the organisation of flat-file databases with the organisation of relational databases.

13. Identify problems that can occur and strategies for resolving these problems for each of the following.
    (a) A record on the ‘one’ side of a ‘one to many’ relationship is deleted.
    (b) The value of a primary key is altered on the ‘one’ side of a ‘one to many’ relationship.

14. Describe the components of a database schema.

15. Consider the Grandview Hotel HSC Style Question.
    (a) Create a data dictionary for each table. Include columns for the field name, data type, field size, description and an example of a typical data item.
    (b) Create the tables and relationships using a RDBMS.
    (c) Add records to populate the tables with some data. Include at least 10 records in the Bookings table.

NORMALISING DATABASES

Normalisation
The process of normalising the design of a database to exclude redundant data. Progressively decomposing the design into a sequence of normal forms.

Redundant Data
Unnecessary duplicate data. Reducing or preferably eliminating data redundancy is the aim of normalisation.

Normalising is the process of designing an efficient database schema for the logical organisation of data within a relational database. It involves splitting the data into tables linked by relationships. The overall aim of normalisation is to remove the possibility of redundant data (unnecessary and duplicate data).

Redundant data wastes storage space and creates maintenance problems. If duplicate data exists in different locations then when this data requires alteration the changes must be made numerous times. If all copies are not altered then data integrity problems will emerge. Imagine a large products table contains the supplier address along with each product's details. When a supplier changes their address it needs to be updated for every product they supply. A properly normalised database eliminates such problems altogether.

The normalisation process is theoretically performed by progressively decomposing the design into a sequence of normal forms, where each normal form is a rule with which the database must comply. Technically there are some eight recognised normal forms. In reality experienced database designers achieve normalised databases using a more intuitive process and then analyse their final design against at least the first three normal form rules. Real world information systems must operate within the limits of the information technology. In many instances progressing past or even fully to the conclusion of the third normal form can have a negative effect on performance – the outcome being too many tables and relationships resulting in overly complex queries.

GROUP TASK Research
Using the Internet, or otherwise, create a list and brief description of each of the normal forms that exist in addition to 1NF, 2NF and 3NF.

We restrict our discussion to a relatively non-technical structured process for working through the first 3 normal forms – referred to as 1NF, 2NF and 3NF. As an example to illustrate the normalising process we shall develop a normalised schema for a simple invoicing database used by a small business.

Consider the sample tax invoice in *Fig 2.22* at right – analysing sample information greatly assists when identifying data that needs to be stored. Notice the GST and Inc-Tax Cost columns and also each of the Totals are calculated. Calculated data should not be stored within a database. It is in effect redundant, therefore it is better to recalculate directly from the source data each time it is required.

TAX INVOICE

To: Fred Smith
12 Gouldian St
Whatsville 2345

09-Jun-06
Invoice Number: 4942
Order Number: H3411

Product	Unit Cost	Units	Ex-Tax Cost	GST	Inc-Tax Cost
Grommet	$23.50	1	$23.50	$2.35	$25.85
Trinket	$3.65	5	$18.25	$1.83	$20.08
Wigwam	$18.00	2	$36.00	$3.60	$39.60
		Totals:	$77.75	$7.77	$85.52

Fig 2.22
Sample Tax Invoice for a small business.

As a starting point we represent the data implied by the sample invoice within a single table – we use Microsoft Access, however any RDBMS could be used. *Fig 2.23* shows this table together with some fictitious data. Notice this table has lots of fields and hence is very wide. In terms of efficient DBMS processing, fields are expensive whereas records are cheap – normalising ultimately results in narrow tables and more efficient processing.

InitialTable : Table

FullName	Address	Town	InvNum	OrderNum	InvDate	Product1	UnitCost1	Units1	Product2	UnitCost2	Units2	Product3	UnitCost3	Units3
Fred Smith	12 Gouldian St	Whatsville 2345	4942	H3411	09/06/06	Grommett	$23.50	1	Trinkett	$3.65	5	Wigwam	$18.00	2
Mary Dougall	16 Klyma Drive	Frogtown 3456	4943	39/231	09/06/06	Grommett	$23.50	4						
Sue Quizzy	19 Groads Rd	Glymer 4567	4944	00234	10/06/06	Trinkett	$3.65	10	Wigwam	$18.00	3			
Jack Spratt	19 Griswold St	Fatsville 5678	4945	CA239	13/06/06	Wigwam	$18.00	1	Grommett	$23.50	2			
Fred Smith	12 Gouldian St	Whatsville 2345	4946	H3416	13/06/06	Trinkett	$23.50	3						

Record: 5 of 5

Fig 2.23
Initial flat-file table for invoicing database.

First Normal Form (1NF)

First normal form deals with the removal of repeating attributes across horizontal rows and ensures each field holds single data items.

To achieve first normal form the following is required:

1. Each field stores single data items.
2. There are no multiple data items within individual fields and no fields are repeated.

To meet the above requirements the following processes are commonly performed:

- Splitting fields into smaller units of data (to achieve 1).
- Deleting repeated fields (to achieve 2).
- Creating new records for each multiple data item and each repeated field (to achieve 2).

In our invoicing database FullName should be split into FirstName and LastName – this simplifies sorts on customer names. The Town field contains both town and postcode; this field should also be split. We could also split the address field however for our purpose this is not necessary. We know that all information processing we will later perform uses the entire contents of this field; therefore within our system Address already holds single data items.

In our initial table in *Fig 2.23* Product, UnitCost and Units are repeated multiple times – an example of repeating fields. We simply delete the repeat fields and then create additional records to contain the data that has been removed. This change when implemented using the sample data from *Fig 2.23* requires 9 records. The table is now in first normal form (1NF), however it contains more redundant data than the original! Fortunately this will be corrected as we work on achieving second normal form.

1NF Version 1 : Table

FirstName	LastName	Address	Town	Postcode	InvNum	OrderNum	InvDate	Product	UnitCost	Units
Fred	Smith	12 Gouldian St	Whatsville	2345	4942	H3411	09/06/06	Wigwam	$18.00	2
Fred	Smith	12 Gouldian St	Whatsville	2345	4942	H3411	09/06/06	Trinkett	$3.65	5
Fred	Smith	12 Gouldian St	Whatsville	2345	4942	H3411	09/06/06	Grommett	$23.50	1
Mary	Dougall	16 Klyma Drive	Frogtown	3456	4943	39/231	09/06/06	Grommett	$23.50	4
Sue	Quizzy	19 Groads Rd	Glymer	4567	4944	00234	10/06/06	Wigwam	$18.00	3
Sue	Quizzy	19 Groads Rd	Glymer	4567	4944	00234	10/06/06	Trinkett	$3.65	10
Jack	Spratt	19 Griswold St	Fatsville	5678	4945	CA239	13/06/06	Gromett	$23.50	2
Jack	Spratt	19 Griswold St	Fatsville	5678	4945	CA239	13/06/06	Wigwam	$18.00	1
Fred	Smith	12 Gouldian St	Whatsville	2345	4946	H3416	13/06/06	Trinkett	$23.50	3

Record: 9 of 9

Fig 2.24
Invoicing database sample data in 1NF.

Our invoicing example did not include multiple data items within individual fields. This occurs when lists of data items are entered into one field usually with separating commas. For example storing Fishing, Surfing, Rugby all in one Hobby field. The 1NF solution is to create new copies of the record each with a different Hobby. This solution is similar to how we solved the repeating Product, UnitCost and Units fields problem in our invoicing database.

GROUP TASK Discussion
Identify possible candidate keys for the 1NF version of the invoicing database (*Fig 2.24*). Which candidate key do you think makes the best primary key? Discuss.

Second Normal Form (2NF)

Second normal form removes redundant data within vertical columns or fields.

To achieve second normal form the following is required:

1. All tables must be in first normal form.
2. Every non-key attribute is functionally dependent on the table's primary key.

Clearly we need to understand what in the world the term "functionally dependent" means? In mathematics a relation links two sets of numbers – usually x and y. A relation $y=f(x)$ is a function if every value of x results in exactly one y value. However the reverse is not true, that is, each value of y can result in zero, one or more values of x. Consider $y=x^2$; a simple parabola. It doesn't matter what value of x you choose you'll always get exactly one answer. However, if you put in a y value you can get zero, one or more than one solutions for x. For example, if y equals 4 then $y=x^2$ has two solutions for x, namely –2 and 2.

Functional dependency operates similarly to mathematical functions. In a table there is a unique primary key for every record – think of the values in the primary key field as the x values in a math function. Now consider some other non-key attribute of the record – these are the y values. In each record the primary key value x identifies exactly one non-key attribute y. However, each non-key attribute y may appear alongside any number of primary keys x. If this is true then y is said to be functionally dependent on x. This can be written as $x \rightarrow y$ which is read as x identifies y.

To meet point 2 above and hence fulfil the requirements of 2NF the following processes are commonly performed:

- Determine functional dependencies. Consider columns that contain redundant data. Look for redundant horizontal sets of data items. The attributes of these duplicates are likely to be functionally dependent on the same primary key –check this is the case for all possible data.
- Determine a primary key (which may require creating one) for each set of functionally dependent attributes determined above.
- Create tables containing each set of functionally dependent attributes including the primary key.
- Move each set of functionally dependent data including the PK into the new tables.
- Apart from the determined primary key columns (which become the foreign keys), delete all other moved attributes from the original table.

Let us work through this process with our 1NF invoicing database shown in *Fig 2.24*. The FirstName, LastName, Address, Town and Postcode columns are always the same for each customer. As this will always be true (that is, it is true for all possible customers not just our sample) then we have FirstName, LastName, Address, Town

and Postcode are functionally dependent on some primary key. There is no obvious existing candidate key so we'll create one in our 1NF table called CustomerID.

We now observe that UnitCost and Product are redundant. Also they are both the same horizontally. For example the Wigwam Product always has a UnitCost of $18.00. This is true for all Products in our sample and indeed for all possible Products. Hence, UnitCost is functionally dependent on Product. We could use the Product field as our new primary key, however in reality it is likely that the name of products will change over time. Hence we decide to create a ProductID primary key in our 1NF table. This means both Product and UnitCost attributes are functionally dependent on ProductID.

2NF Version 1 : Table

CustomerID	FirstName	LastName	Address	Town	Postcode	InvNum	OrderNum	InvDate	ProductID	Product	UnitCost	Units
1	Fred	Smith	12 Gouldian St	Whatsville	2345	4942	H3411	09/06/06	3	Wigwam	$18.00	2
1	Fred	Smith	12 Gouldian St	Whatsville	2345	4942	H3411	09/06/06	2	Trinkett	$3.65	5
1	Fred	Smith	12 Gouldian St	Whatsville	2345	4942	H3411	09/06/06	1	Grommett	$23.50	1
2	Mary	Dougall	16 Klyma Drive	Frogtown	3456	4943	39/231	09/06/06	1	Grommett	$23.50	4
3	Sue	Quizzy	19 Groads Rd	Glymer	4567	4944	00234	10/06/06	3	Wigwam	$18.00	3
3	Sue	Quizzy	19 Groads Rd	Glymer	4567	4944	00234	10/06/06	2	Trinkett	$3.65	10
4	Jack	Spratt	19 Griswold St	Fatsville	5678	4945	CA239	13/06/06	1	Gromett	$23.50	2
4	Jack	Spratt	19 Griswold St	Fatsville	5678	4945	CA239	13/06/06	3	Wigwam	$18.00	1
1	Fred	Smith	12 Gouldian St	Whatsville	2345	4946	H3416	13/06/06	2	Trinkett	$23.50	3
0									0			

Record: 9 of 9

Fig 2.25
Invoicing sample database in partial 2NF.

The additional CustomerID and ProductID fields are added to the table (see *Fig 2.25*). Although we suspect further functional dependencies exist they are difficult to see with all the distracting customer and product attributes so we create our Customers and Products tables, move the data into the new tables and finally delete the functionally dependent customer and product attributes from the main table (see *Fig 2.26*). This results in a clearer view of the data in the main table.

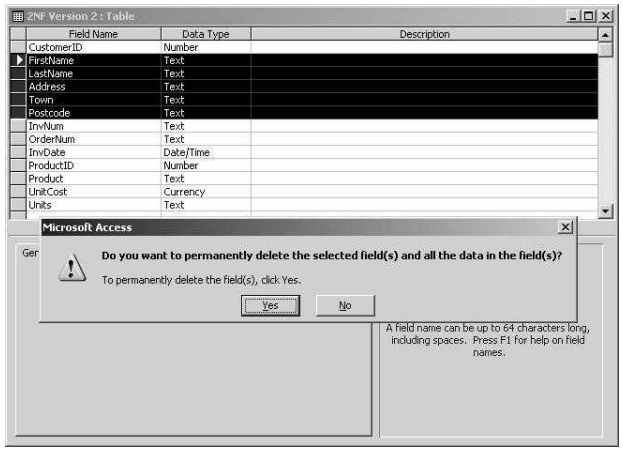


Fig 2.26
Deleting functionally dependent customer attributes from the main table.

The current state of our database schema is reproduced in *Fig 2.27* below. Notice that CustomerID is the PK in the new customer table, so each customer only appears once. Similarly in the Products table each product appears once only. We have also selected a composite key for the main table composed of InvNum and ProductID.

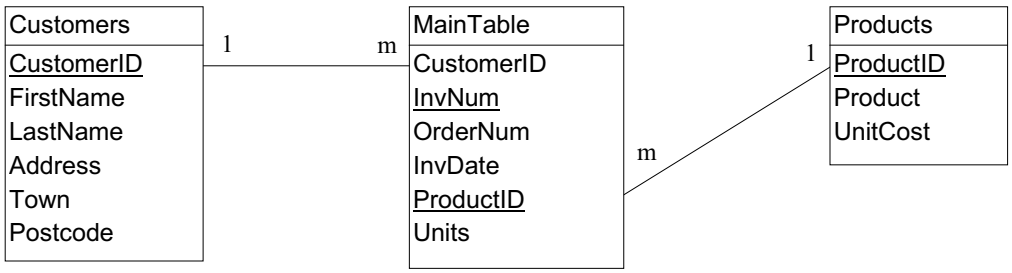


Fig 2.27
Invoicing database incomplete 2NF schema.

Consider the new version of the main table reproduced in *Fig 2.28*. There are still remaining functional dependencies – try to find them before reading on!

2NF Version 2 : Table

CustomerID	InvNum	OrderNum	InvDate	ProductID	Units
1	4942	H3411	09/06/06	3	2
1	4942	H3411	09/06/06	2	5
1	4942	H3411	09/06/06	1	1
2	4943	39/231	09/06/06	1	4
3	4944	00234	10/06/06	3	3
3	4944	00234	10/06/06	2	10
4	4945	CA239	13/06/06	1	2
4	4945	CA239	13/06/06	3	1
1	4946	H3416	13/06/06	2	3

Record: 9 of 9

Fig 2.28
Improved but incomplete 2NF main table.

Consider the redundant data within the InvNum column. For each InvNum the OrderNum, InvDate and CustomerID attributes contain the same data. It appears that InvNum identifies CustomerID, OrderNum and InvDate. This means the three attributes CustomerID, OrderNum, and InvDate are functionally dependent on the primary key InvNum. Note that both InvNum and OrderNum are possible candidate keys worth considering, we choose InvNum. OrderNum is rejected as it is supplied by the customers (presumably on their purchase orders) hence it is possible for two (or more) customers to submit purchase orders with identical order numbers; this would violate the uniqueness of the PK.

In this case we do not need to create a new primary key as InvNum already exists. Notice that the four attributes in question, namely InvNum (PK), OrderNum, InvDate and CustomerID are all attributes of a particular invoice, hence we logically name our new table Invoices. We create the Invoices table, move the data in and then delete the CustomerID, OrderNum and InvDate attributes from the main table.

Now examine the remaining main table – it now contains just the InvNum, ProductID and Units attributes (consider just these three columns in *Fig 2.28*). Each unique InvNum, ProductID combination determines exactly one value in the Units column. Therefore the Units attribute is functionally dependent on both InvNum and ProductID. ProductID and InvNum together must form the composite key. This table is therefore in 2NF – so at last we're finished!

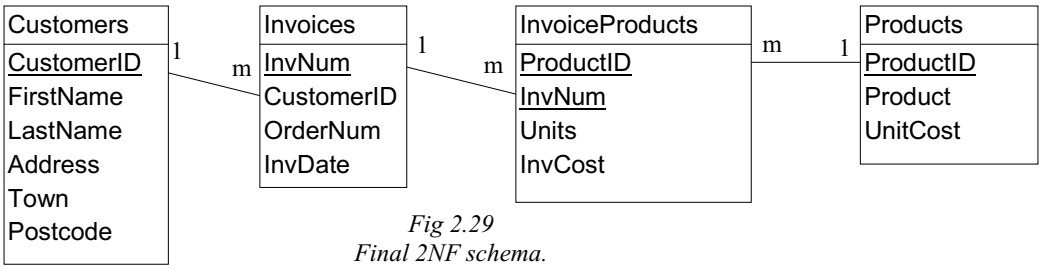


Fig 2.29
Final 2NF schema.

Our revised final 2NF schema together with our sample data within the main table (renamed as InvoiceProducts) is reproduced in *Fig 2.29* and *Fig 2.30* respectively. Two additional alterations have been made to the schema.

InvoiceProducts : Table

InvNum	ProductID	Units	InvCost
4942	1	1	$23.50
4942	2	5	$3.65
4942	3	2	$18.00
4943	1	4	$23.50
4944	2	10	$3.65
4944	3	3	$18.00
4945	1	2	$23.50
4945	3	1	$18.00
4946	2	3	$3.65

Record: 9 of 9

Fig 2.30
Final main table in 2NF renamed InvoiceProducts.

The main table has been renamed InvoiceProducts. This name change makes sense given the position of the table within the schema and furthermore each record within this table describes a product present on an individual invoice.

The second alteration is the addition of the InvCost attribute to the InvoiceProducts table. This addition appears to violate 2NF. It is a necessary addition to

meet the requirements of most invoicing systems. Suppose the UnitCost is changed for a product due to a price increase. If the UnitCost is held just once (within the Products table) then the cost of that product on all existing invoices will also change and be incorrect. Therefore the UnitCost should be included in both the Products and the InvoiceProducts tables. As invoices are entered the current UnitCost from the Products table is used as the default value for the InvoiceProducts InvCost field. In reality InvCost is functionally dependent on the composite key ProductID and InvNum.

GROUP TASK Activity
Draw a grid for each table in the final 2NF schema shown in *Fig 2.29*. Include all the data within the initial table shown in *Fig 2.23*.

GROUP TASK Practical Activity
Create the initial *Fig 2.23* table within a RDBMS. Work through the 1NF and 2NF normalisation processes described on the previous pages. Try to only use copy and paste to move data into new tables as they are created.

Third Normal Form (3NF)

Third normal form removes further redundant data within vertical columns or fields.

To achieve third normal form the following is required:

1. All tables must be in second normal form.
2. Every non-key attribute is functionally dependent only on the table's primary key and not on any other attributes of the table.

To achieve third normal form the database must be in second normal form and then the following is performed:

- In each table look for non-key attributes that are functionally dependent on another non-key attribute. (Note that all attributes will already be functionally dependent on the primary key as the tables are in 2NF).
- Determine a primary key (which may require creating one) for each set of functionally dependent attributes determined above.
- Create tables containing each set of functionally dependent attributes including the primary key.
- Move each set of functionally dependent data including the PK into the new tables.
- Apart from the determined primary key columns (which become the foreign keys), delete all other moved attributes from the original table.

In practice it can prove counter productive to pursue 3NF to its logical conclusion. 3NF often involves removing attributes that seldom change. Such processes increase the complexity of queries yet in many cases have a minimal effect in terms of both storage and improved data integrity.

Consider our Customers table; the Town determines the Postcode as each Town has exactly one Postcode (although a single postcode can relate to multiple towns). That is the Postcode is functionally dependent on the Town. Hence to achieve 3NF in the Customers table we must create a separate linked table of Towns with Town and Postcode attributes together with a primary key. For large commercial and government databases where thousands or even millions of addresses are held the effort is worthwhile. In our small business system such detail is not justified as Towns and Postcodes rarely change.

Situations where 3NF is worth pursuing involve data that changes fairly regularly or where there are only a small number of possible data item combinations. Consider a

typical Students table within a school. Say this table contains attributes for StudentID (PK), FirstName, LastName, YearLevel and YearAdvisor. All non-key attributes are functionally dependent on the StudentID (PK) as the Students table is in 2NF. However, YearLevel identifies exactly one Year Advisor (a reasonable assumption in most schools) That is, the YearAdvisor attribute is functionally dependent on the YearLevel attribute. Furthermore it is likely that the YearAdvisor for each YearLevel will change at least every year, also within most high schools there are only six year levels and six year advisors – not very much data at all. In this case it makes sense to create a YearAdvisor table containing just the composite key composed of YearLevel and TeacherID. The YearLevel attribute being a FK back to the Students table and the TeacherID attribute being a FK to the Teachers table (see *Fig 2.31*).

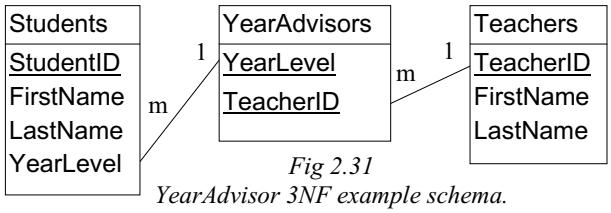


Fig 2.31
YearAdvisor 3NF example schema.

GROUP TASK Discussion
The YearAdvisor schema in *Fig 2.31* allows a single teacher to be year advisor for more than one year level. Suggest changes to the schema so that a teacher can be a year advisor for at most one year level.

Consider the following:

The normalisation process aims to remove the possibility of redundant data. But why is reducing data redundancy so important? To answer this question we need to consider the types of information processes that are performed on databases and then consider why duplicate data (or redundant data) is a problem for each of these processes. We'll use our initial non-normalised invoice database (reproduced in *Fig 2.32* below) to illustrate each problem.

InitialTable : Table

FullName	Address	Town	InvNum	OrderNum	InvDate	Product1	UnitCost1	Units1	Product2	UnitCost2	Units2	Product3	UnitCost3	Units3
Fred Smith	12 Gouldian St	Whatsville 2345	4942	H3411	09/06/06	Grommett	$23.50	1	Trinkett	$3.65	5	Wigwam	$18.00	2
Mary Dougall	16 Klyma Drive	Frogtown 3456	4943	39/231	09/06/06	Grommett	$23.50	4						
Sue Quizzy	19 Groads Rd	Glymer 4567	4944	00234	10/06/06	Trinkett	$3.65	10	Wigwam	$18.00	3			
Jack Spratt	19 Griswold St	Fatsville 5678	4945	CA239	13/06/06	Wigwam	$18.00	1	Grommett	$23.50	2			
Fred Smith	12 Gouldian St	Whatsville 2345	4946	H3416	13/06/06	Trinkett	$23.50	3						

Record: 5 of 5

Fig 2.32
Initial flat-file table for invoicing database.

- Collecting is the information process that gathers information. Within databases collecting adds or inserts new records into the database. In SQL the INSERT keyword is used to create new records as part of the collecting information process. Database problems occurring as data is collected are known as INSERT anomalies.

  There are two types of common INSERT anomalies:

  - Extra data needs to be inserted along with the data you actually wish to insert. Imagine the business has a new product they wish to add to the database. This cannot be done until a customer orders that product. Similarly a new customer's details cannot be added until they have placed an order.

- Data that already exists must be re-entered along with the new data. When a customer reorders their details must be re-entered. Similarly each time a product is ordered its name and cost must be re-entered.

- Processing information processes manipulate data by editing and updating it; in essence the data is changed. This includes modifying or updating data, such as changing an address, and it also includes deleting data, such as removing a product from an invoice. In database terms these processes are known as UPDATE and DELETE processes (these are SQL terms).

  These problems are known as UPDATE anomalies and DELETE anomalies.

  - DELETE anomalies occur when deleting a record also removes data not intended for deletion. Say you wish to delete a particular invoice. If this is the only invoice for that customer then the customer's details are also lost.
  - UPDATE anomalies occur when changing a specific data item requires the same change in many places. Say a customer's address changes, this change must be made to every invoice that relates to that customer.

GROUP TASK Discussion
Consider the final normalised invoice database (refer *Fig 2.29*). Discuss how each of the INSERT, DELETE and UPDATE anomaly problems mentioned above has been resolved.

- Analysing information processes transform data into information. As information is for users then it must subsequently be displayed. Within databases many analysing processes involve sorting and/or searching the data.

  Say in our initial invoicing database we require a list of customers who have ordered a particular product. This is difficult as it involves searching three fields – Product1, Product2 and Product3. Furthermore if the product has been misspelled somewhere then it will be missed completely.

  What about a simple alphabetical list of products the business sells? This is also difficult as the products are in different fields. Even if we succeed any misspelled products will appear multiple times.

GROUP TASK Discussion
Explain how the search and the sort mentioned above have been simplified within the final normalised invoice database.

- Storing and retrieving information processes save, reload and maintain data. Transmitting and receiving transfers data within the system. These processes move data to and from each of the other information processes.

  Clearly storing the same data many times requires extra storage space. However in most cases this is not the most significant problem – secondary storage is pretty cheap these days. The speed at which the data is moved is much more critical. DBMSs deal at the record level – they save, reload, transmit and receive complete records not individual fields. Obviously moving longer records around is going to take longer than moving shorter records. Compare moving the initial set of records in *Fig 2.32* with moving the final set of records in the InvoiceProducts table.

GROUP TASK Activity
Calculate the approximate storage size of each record in the initial table (*Fig 2.32*) compared with the storage size of a record in the final InvoiceProducts table (*Fig 2.30*).

HSC style question

Louise works for a large department store. She is responsible for maintaining records in regard to the loan of items to various departments. Louise currently stores this data in a single Loans table linked to an Employee table she obtained from the IT department.

- The Employee table includes EmployeeID, LastName and FirstName attributes.
- Each Department has a single supervisor who borrows items on behalf of employees in their department.
- Each employee works within a single department.
- Each item has a label attached with a unique item number.

Some of the records in Louise's Loans table are reproduced below.

Item Name.	*Item Number*	*SupervisorID*	*EmployeeID*	*Department name*	*Date borrowed*	*Date Returned*
Cash register	2341	JWA	FNE	Ladies Wear	15/2/05	
Stocktake scanner	6634	MRO	MDA	Electronics	10/5/06	9/6/06
Stocktake scanner	4511	SMI	MDA	Mens Wear	10/4/05	17/4/05
Laptop Computer	2433	SMI	SMI	Mens Wear	11/8/05	12/2/06
Stocktake scanner	6634	JWA	FNE	Ladies Wear	12/6/06	
Laptop Computer	1866	JWA	SDA	Ladies wear	18/3/05	21/9/06
Laptop Computer	2433	SMI	SMI	Mens Wear	18/5/06	
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

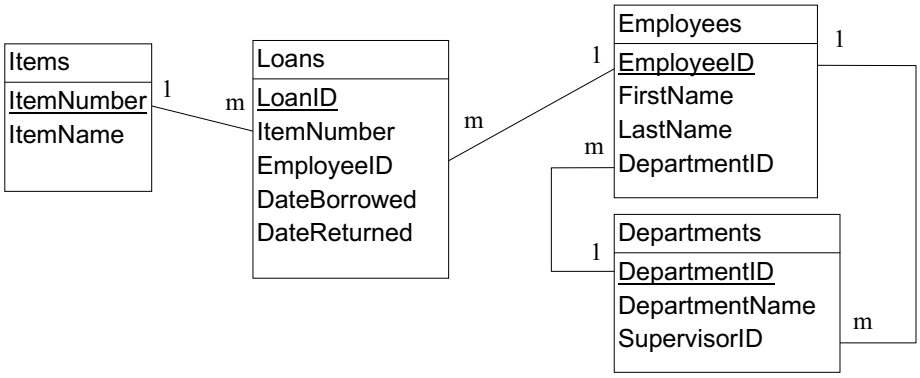
(a) With reference to the above sample Loans table, identify an example of data redundancy and describe problems that could arise as a consequence.

(b) Normalise this relational database into four tables (including the Employee table). Indicate all necessary relationships, primary keys and foreign keys.

Suggested Solution

(a) Department name and SupervisorID are duplicated. Neither of these attributes are needed as SupervisorID and DepartmentName are functionally dependent on EmployeeID. Including them in the table means Louise must re-enter both department name and SupervisorID each time an item is borrowed. Also if a department's supervisor changes then the SupervisorID must be altered in every record that relates to that department.

(b)



Comments

- For (a) the Item Name attribute also contains redundant data, as Item Name is functionally dependent on Item Number. Louise must enter both the Item Number and Item Name for each new loan. Also it is not possible to maintain a record of items that have never been loaned.

GROUP TASK Discussion
The Item Name field contains the same data for different Item Numbers. For example 2433 and 1866 are both named "Laptop Computer". How could this redundancy be removed? Is it worth removing? Discuss.

- The question asks for foreign keys to be indicated. The 1:m relationship lines pointing to the foreign keys should be sufficient indication, but it would be prudent to physically label each of the relevant fields – perhaps using FK to label foreign keys and also labelling each primary key using PK.
- In the Loans table a combination of ItemNumber and DateBorrowed is a candidate key if DateBorrowed includes the time in sufficient detail. Clearly a single item cannot be loaned to more than one employee at the same time.
- Note that the Date Returned attribute cannot be considered as part of the PK as it is NULL whilst an item is on loan. Primary keys and components of composite keys can never be NULL.
- Notice that two relationships link the Employees and Departments tables. That is, each employee is linked to a department and each department has a supervisor who is also an employee. When trying to make sense of such schemas try to consider each relationship in isolation.
- Within the suggested answer schema it is possible for a supervisor to supervise many departments. This is okay in terms of the question – the question specifies that each department has one supervisor, not the opposite. However this may not be desirable in reality. Most DBMSs include a unique property for each field – setting this property for the SupervisorID would solve this problem.

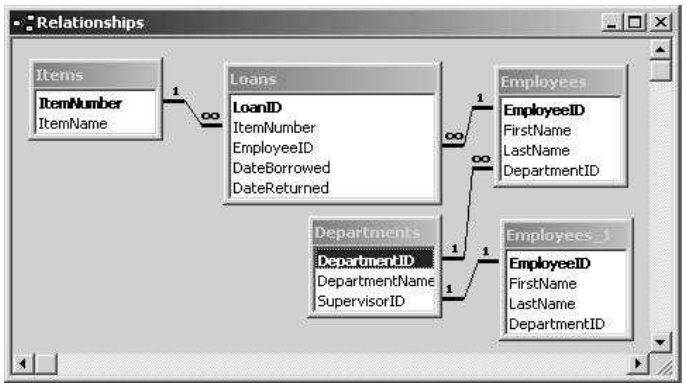


Fig 2.33
MS-Access Relationships for Loans database schema.

- To create the schema for the suggested answer within MS Access's relationship window requires the Employee table to be included twice as shown above in *Fig 2.33* above. Note that a 1:1 join between SupervisorID and EmployeeID is shown, indicating the unique property has been set for the SupervisorID field.

GROUP TASK Practical Activity
Create the Loans database using a RDBMS such as MS-Access. Enter the sample data from the question into the database.

SET 2D

1. In general, normalising a flat-file database results in:
   (A) many tables.
   (B) reduced data redundancy.
   (C) no INSERT, DELETE or UPDATE anomalies.
   (D) All of the above.
2. A table normalised into 1NF commonly:
   (A) includes more attributes.
   (B) contains more records.
   (C) contains less records.
   (D) has no redundant data.
3. For a table to be in 2NF it must be in 1NF and also:
   (A) All non-key attributes must be candidate keys.
   (B) All non-key attributes must be functionally dependent on the primary key.
   (C) The primary key must be functionally dependent on all other attributes.
   (D) There must be one and only one candidate key that is the primary key.
4. To alter a product name requires the name to be changed in 5 different places. This is an example of a:
   (A) DELETE anomaly.
   (B) INSERT anomaly.
   (C) UPDATE anomaly.
   (D) CREATE anomaly.
5. A school database's Students table contains the name and address details of each student. However there are many brothers and sisters in the school who live at the same address. Splitting the address details into their own table would occur when normalising the Students table into:
   (A) 1NF.
   (B) 2NF.
   (C) 3NF.
   (D) 4NF.
6. In a normalised table, the attribute *p* is functionally dependent on the attribute *q*. Which of the following is true?
   (A) There can be repeating values in the *p* column.
   (B) The *q* column is a unique identifier.
   (C) Each value for *q* identifies a single value of *p*.
   (D) All of the above.
7. A table contains data about products and customers. Splitting this table into two would occur when normalising the table into:
   (A) 1NF.
   (B) 2NF.
   (C) 3NF.
   (D) 4NF.
8. A field in a database contains lists of items. This would be corrected when normalising the database in to:
   (A) 1NF.
   (B) 2NF.
   (C) 3NF.
   (D) 4NF.
9. A table is in 3NF when it is in 2NF and:
   (A) all fields (apart from the PK) are functionally dependent on only the PK.
   (B) no records have the same data contained within the same attribute.
   (C) every attribute (including the primary key) is a candidate key.
   (D) a primary key uniquely identifies every record.
10. During normalisation it is first noticed that each time a particular value in attribute *p* occurs attribute *q* has the same value. Which normal form is being considered?
    (A) 1NF.
    (B) 2NF.
    (C) 3NF.
    (D) 4NF.
11. Define each of the following terms.
    (a) Normalisation (b) Functionally dependent (c) Redundant data
12. Consider the Library database schema shown in *Fig 2.17*.
    (a) List 5 examples of functional dependencies present in this schema.
    (b) Is each table in the library database in 3NF? Justify your response.
13. Identify and describe problems that are solved by normalising a database.
14. Create a step-by-step summary describing how a table is normalised into 3NF.
15. Normalise the following *Vehicles* table into 2NF (assume there are many more records).

Rego	*Description*	*Cylinders /Capacity*	*Year*	*CTP Insurer*	*Owner*	*Address*
QZN-712	Ford Festiva	4/1300cc	1993	AAMI	Melissa Davis	15 Kiama St Wallytown 2345
NPO-933	Holden Commodore	6/3800cc	2004	NRMA	Martin Wilson	6 Juniper Rd Elberton 3409

HYPERTEXT/HYPERMEDIA

Hypertext is a term used to describe bodies of text that are linked in a non-sequential manner. The related term, hypermedia, is an extension of hypertext to include links to a variety of different media types including image, sound, and video. In everyday usage, particularly in regard to the World Wide Web, the word hypertext has taken on the same meaning as hypermedia; in our discussions we shall just use the term hypertext. Be aware that when we discuss links to other documents, these other documents are not necessarily text; they could be images, audio, video or any mix of media types.

Hypertext
Bodies of text that are linked in a non-sequential manner. Each block of text contains links to other blocks of text.

Hypermedia
An extension of hypertext to include non-sequential links with other media types, such as image, audio and video.

Today most people associate the term hypertext with the World Wide Web (WWW), however the WWW commenced operation in the early 1990s – in 1992 there were just 50 web sites. In reality hypertext in various forms has been around since the late 1960s. Computerised versions of dictionaries and encyclopaedias used hypertext so readers could quickly navigate to specific words or topics. Apple Computer released HyperCard in 1987, a hypertext program included with the Macintosh. Hypercard allowed users to create multi-linked databases. Each card was similar to a record in a database table, with the addition that fields could contain links to other cards. Many computer games use hypermedia concepts to guide the user through a storyline. The storyline changes each time the game is played or a different choice or action is performed.

Despite the unstructured nature of hypertext, it actually reflects the operation of the human mind more closely than other methods of data organisation. The human mind operates largely on associations; we read a passage of text and our mind generates various related associations based on past experiences. Our thoughts move continually from one association to another; hypertext is an attempt to better reflect this behaviour. It enables us to explore associations by following links.

Consider the following:

Theodor Holm (Ted) Nelson was the first to use the term “Hypertext”. The following extracts are taken from his 1965 paper titled “"A File Structure for the Complex, the Changing, and the Indeterminate."

Under the heading “Discrete Hypertexts.” Nelson writes:

> *“‘Hypertext’ means forms of writing which branch or perform on request; they are best presented on computer display screens… Discrete, or chunk style, hypertexts consist of separate pieces of text connected by links.”*

In this next extract Nelson discusses a further form of hypertext he calls “stretchtext”:

> *“This form of hypertext is easy to use without getting lost… There are a screen and two throttles. The first throttle moves the text forward and backward, up and down on the screen. The second throttle causes changes in the writing itself: throttling toward you causes the text to become longer by minute degrees.”*

GROUP TASK Discussion
With reference to the above extracts, do you think Ted Nelson's vision for hypertext has largely been realised? Discuss.

THE LOGICAL ORGANISATION OF HYPERTEXT/HYPERMEDIA

The organisation of hypertext is based on links (often called hyperlinks) and nodes. A set of nodes and their various links form a web – the World Wide Web being the most obvious and largest example.

In general usage the term node means a point where links are connected. In a computer network a node is any device connected (linked) to the network. Hypertext nodes are also connected (via links) to each other; each node is part of a hypertext network known as a web. In hypertext terms a node is usually some block or unit of information – perhaps a web page, a simple block of text, a video sequence or some richer information that combines many media types. The user follows a link embedded within a node and is taken to another node; this new node may also contain links to further nodes.

Navigation between nodes within even moderately sized webs can theoretically take place in many complex and unstructured ways. The WWW is an extreme case where the number of possible navigation paths is virtually infinite. When designing a web it is desirable to logically structure the possible navigation paths or at least indicate some common paths through the web. Storyboards are a tool designed for such a purpose. In this section we examine storyboards and we then consider Hypertext Markup Language (HTML) – the hypertext language of the WWW.

GROUP TASK Discussion
There are many other applications where hypertext is used apart from the WWW. List examples of such applications. In each example describe a typical node and link.

STORYBOARDS

Storyboarding is a technique that was first used for the creation of video information, including film, television and animation. These storyboards show a hand drawn sketch of each scene together with a hand written description.

Video data by its very nature is linear, that is scenes are arranged into a strict sequence that tells a story (see *Fig 2.34*). However hypertext screen displays are different, they provide the ability for users to navigate in a variety of different ways. As a consequence, storyboards created for computer-based screen display are typically composed of two primary elements – the individual screen layouts with descriptions, together with a navigation map illustrating the links between these screens.

The individual screen layouts should clearly show the placement of navigational items, titles, headings and content. It is useful to indicate which items exist on multiple pages – such as contact details and menus. Notes that describe elements or actions that are not obvious should be made. Each layout should not just

Fig 2.34
Video storyboards are always linear.

include the functional elements, it should also adequately show the look and feel of the page. Commonly a theme for the overall design is used – this can be detailed separately to each of the individual screen designs. Often each screen is hand drawn on separate pieces of paper. Once these layouts are complete they can be arranged in various combinations to assist when finalising the structure of the navigation map.

A navigation map describes the organisation of a hypertext web. It is composed of a sketch that includes each node or screen within the web, together with arrows indicating links between nodes.

There are four commonly used navigation structures: linear, hierarchical, non-linear and composite (see *Fig 2.35*). The nature of the information largely determines the selection of a particular structure. For example a research project has a very different natural structure compared to an online supermarket. There are two somewhat conflicting aims when designing a navigation structure. Firstly the structure must convey the information to users in the manner intended by the author, and secondly the users should be able to locate information without being forced to wade through irrelevant information. The structure should offer the user sufficient flexibility to navigate easily to information they require. Designers of hypertext must balance the achievement of these aims as they choose the most effective navigation structure.

The linear structure forces the user through a particular sequence of nodes. This structure is particularly useful for training where the content of each node requires knowledge obtained from previous nodes. For example PowerPoint presentations are almost always linear. Linear navigation is also used on commercial sites where data is sequentially collected from users to process a transaction. For example, making a purchase online requires customers to progress through the same sequence of screens each time they make a purchase.

Hierarchical structures are common as they are simple for users to visualise. As a user drills down the tree they are presented with more and more detailed information. Most large commercial and government web sites use this structure. It is particularly suited to information that falls into categories and sub-categories. Once in a particular category, users are not overwhelmed by information from other categories. To navigate to some

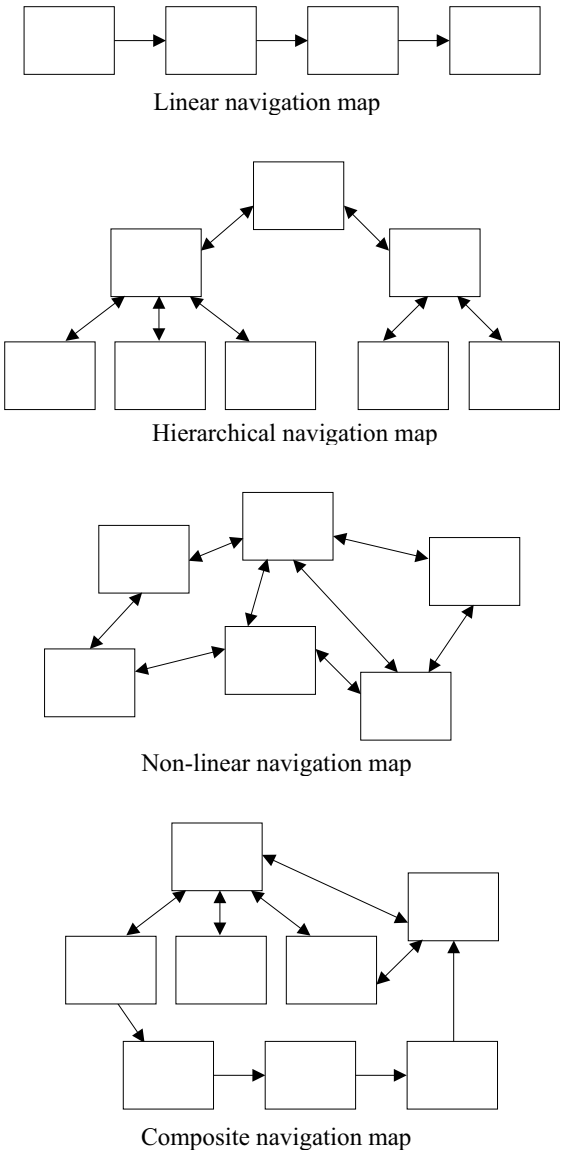


Fig 2.35
Common navigation structures used on storyboards.

other category they must move back up the hierarchy and then select a different downward path.

Non-linear or unstructured navigation is difficult for users to visualise. It allows maximum flexibility of design, but it is easy for users to get lost in a maze of screens. If a non-linear structure is used then in most cases some form of map should also be provided for users. Games are one area where non-linear structures are used to great advantage. Within games the experience is enhanced when knowledge of what comes next is unknown.

Composite structures combine aspects of each of the other structures. In reality most hypertext webs use a composite structure. This makes sense given that most webs include instructional nodes that form a sequence, together with informational nodes that have some form of inherent classification.

Consider the following screen layouts:

Janine is designing a website for Angelo's Italian Restaurant – she is currently working on a storyboard. Two of her initial screen layouts are reproduced in *Fig 2.36* below.

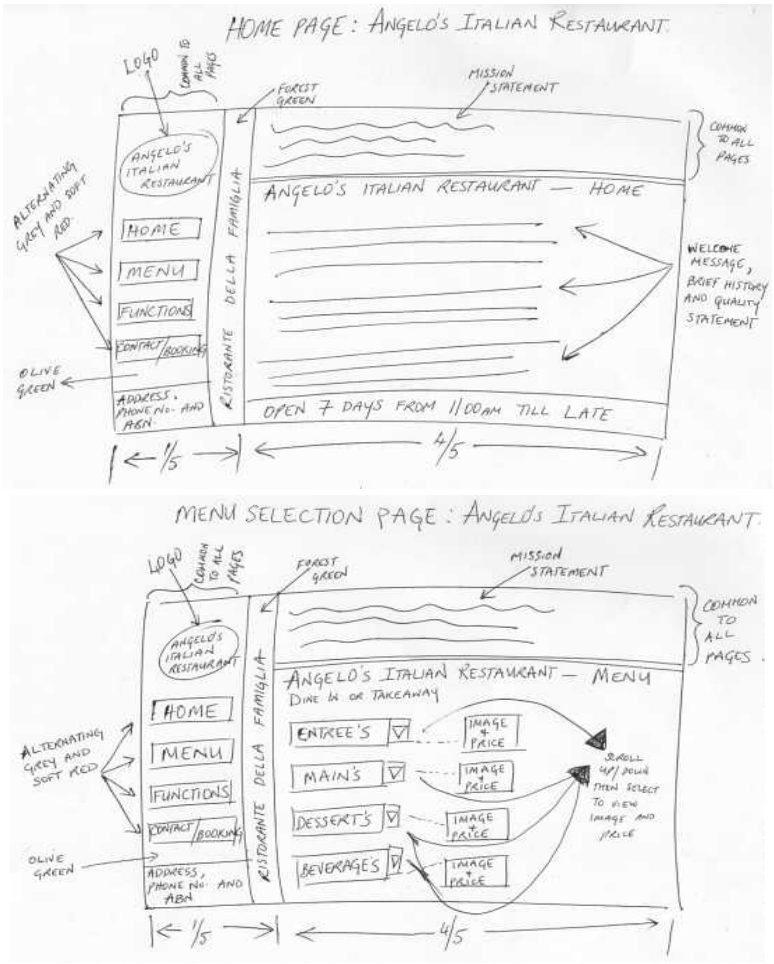


Fig 2.36
Screen layouts for Angelo's Home Page (top) and Menu Page (bottom).

GROUP TASK Activity
Complete Janine's storyboard by creating layouts for the "Functions" and "Contact/Booking" web pages, together with a suitable navigation map.

HYPERTEXT MARKUP LANGUAGE (HTML)

Documents accessed via the World Wide Web (WWW) make extensive use of hyperlinks; these documents are primarily based on HTML. HTML is the primary method of organising hypertext for use on the WWW. In general, each document is an HTML file that is displayed as a web page within the user's browser. Clicking on a link within a web page can take you to a bookmark within the current page or to another page stored on virtually any computer throughout the world. From the user's point of view, the web page is just retrieved and displayed in their web browser; the physical location of the page is irrelevant.

Let us consider the organisation of a typical HTML file. All HTML files are really simple text files, that is, a sequential list of characters. Hence, HTML files can be created and edited using any simple text editor. *Fig 2.37* shows Microsoft's home page together with its source HTML file shown within a text editor, in this case notepad. Various software applications, collectively called HTML or web page editors, are available to assist when creating HTML files; text editors are the simplest.

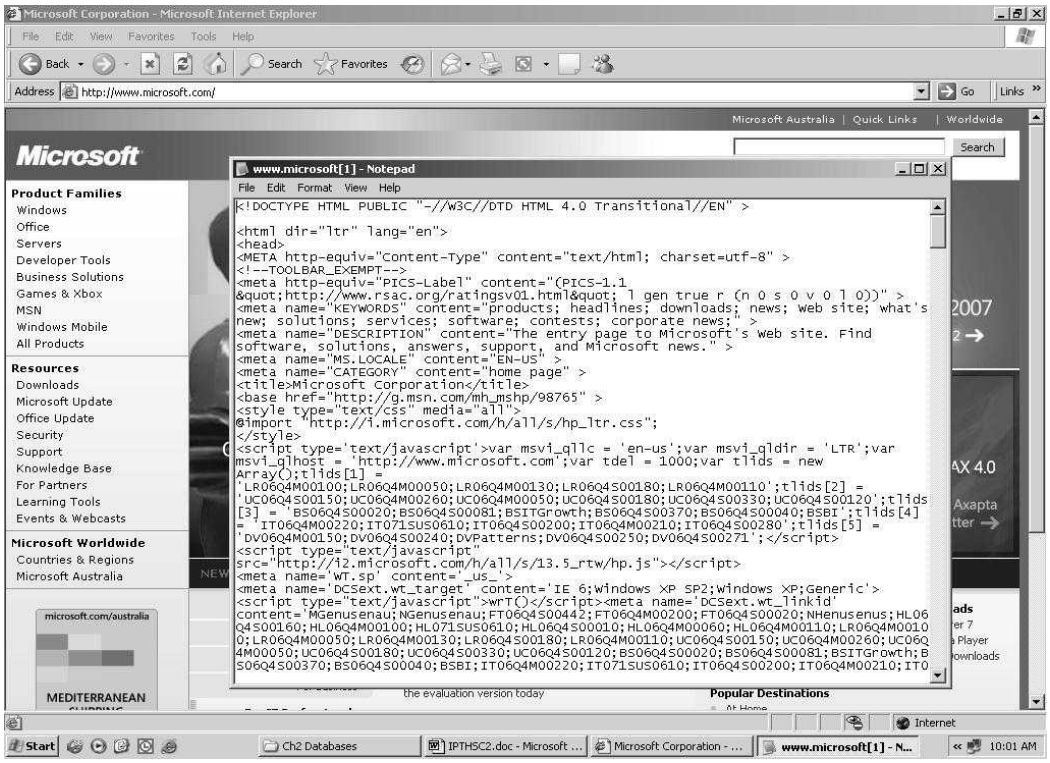


Fig 2.37
Microsoft© home page and source HTML code within notepad.

In the past web designers required extensive technical knowledge in regard to the details of HTML, this is no longer the case. Today most web designers are visual design professionals; they use dedicated web page creation software such as Dreamweaver, where the focus is directed towards the artistic layout of the pages. These software packages remove the need for designers to understand the intricate technical detail of HTML; rather they work in a WYSIWYG ("what you see is what you get") environment. In essence web page creation software automates the generation of the final HTML files in much the same way that desktop publishing software automates the production of final hardcopy. Nevertheless it is still worthwhile having a basic knowledge of HTML. Many designers use sophisticated web page creation software for much of the design, and then they edit the underlying HTML to include specific fine detail within their pages.

HTML uses tags to specify formatting, hyperlinks and numerous other functions – some common examples are included in *Fig 2.38* below. All tags are enclosed within angled brackets < >; these brackets indicate to the web browser that the text enclosed is an instruction rather than text for display. In most cases, pairs of tags are required; a start tag and an end tag. The function specified by the start tag is applied to the text contained between the tags. For example, in *Fig 2.37* above, the `<title>` and `</title>` tags surround the page title; the text between these two tags, namely "Microsoft Corporation" is displayed in the title bar of the browser. In this case the browser has also appended its name to the title – "Microsoft Internet Explorer".

Basic Tags	
<html></html>	Creates an HTML document
<head></head>	Sets off the title and other information that isn't displayed on the Web page itself
<body></body>	Sets off the visible portion of the document
Header Tags	
<title></title>	Puts the name of the document in the title bar
Body Attributes	
<body bgcolor=?>	Sets the background color, using name or hex value
<body text=?>	Sets the text color, using name or hex value
<body link=?>	Sets the color of links, using name or hex value
<body vlink=?>	Sets the color of followed links, using name or hex value
<body alink=?>	Sets the color of links on click
Text Tags	
<pre></pre>	Creates preformatted text
<hl></hl>	Creates the largest headline
	Creates the smallest headline
	Creates bold text
	Creates italic text
	Emphasizes a word (with italic or bold)
	Sets size of font, from 1 to 7
 	Sets font color, using name or hex value
Anchor tags (Links)	
 	Creates a hyperlink
 	Creates a mailto link
 	Creates a target location within a document
 	Links to that target location from elsewhere in the document

Formatting	
	Creates a new paragraph
	Aligns a paragraph to the left, right, or center
 	Inserts a line break
<blockquote></block quote>	Indents text from both sides
	Creates a numbered list
	Precedes each list item, and adds a number
	Creates a bulleted list
Image Elements	
<img src="name">	Adds an image
<img src="name" align=?>	Aligns an image: left, right, center; bottom, top, middle
<img src="name" border=?>	Sets size of border around an image
<hr>	Inserts a horizontal rule
<hr size=?>	Sets size (height) of rule
<hr width=?>	Sets width of rule, in percentage or absolute value
Tables	
<table></table>	Creates a table
<tr></tr>	Sets off each row in a table
<td></td>	Sets off each cell in a row
<th></th>	Sets off the table header
Table Attributes	
<table border=#>	Sets width of border around table cells
<table cellspacing=#>	Sets amount of space between table cells
<table cellpadding=#>	Sets amount of space between a cell's border and its contents
<table width=# or %>	Sets width of table — in pixels or as a percentage of document width
<tr align=?> or <td align=?>	Sets alignment for cell(s) (left, center, or right)
<tr valign=?> or <td valign=?>	Sets vertical alignment for cell(s) (top, middle, or bottom)
<td colspan=#>	Sets number of columns a cell should span
<td rowspan=#>	Sets number of rows a cell should span (default=1)

Fig 2.38
Some common HTML tags.

HTML tags are an example of metadata. Metadata is data that defines or describes other data. Within a relational database both data dictionaries and schematic diagrams are examples of metadata – both these tools define the data within the database. There are countless other examples of metadata, HTML tags and storyboards included. There are literally hundreds of possible HTML tags available to web designers – just some of them are shown above in *Fig 2.38*. Note that HTML tags can be entered in either upper or lower case. For our purpose we restrict our discussion to two common examples; meta tags that describe the data within a page and anchor tags used to link pages. We then consider the organisation of uniform resource locators (URLs) used within links.

Metadata
Data that defines or describes other data.

META tag

The `META` tag is a special HTML tag that is used to store information that describes the data within a Web page rather than defining how it should be displayed. `META` tags provide information including what program was used to create the page, a description of the page, and keywords relevant to the page. Many search engines display the page title and then the description from the `META` tags for each page they find. The `META name="keywords"` option was designed to assist search engines. When early search engines performed a full text search to identify keywords within pages they often identified words that were not necessarily relevant. The `META name="keywords"` option was introduced so web page designers could specify their own keywords directly. Unfortunately the `keywords` option has been misused by designers in an attempt to attract extra traffic to their web site. Today search engines use much more sophisticated techniques for identifying keywords and hence few search engines today utilise this `keyword` information.

```
<HEAD>
<TITLE>The world according to Zorp</TITLE>
<META name="description" content="Zorp describes his view on the
world. A fascinating insight into the mind of Zorp.">
<META name="keywords" content="zorp, world view, insightful">
</HEAD>
```

Fig 2.39
The HTML META tag is used to describe the data within a web page.

`META` tags, when used, are included between the `<HEAD>` and `</HEAD>` tags. For example, the web page in *Fig 2.39* would be described within most search engines as "The world according to Zorp." followed by the description "Zorp describes his view on the world. A fascinating insight into the mind of Zorp." If the search engine uses the `keywords` option then anyone using the words "zorp", "world view" or "insightful" in their search would find this page.

GROUP TASK Research
There are many other examples of metadata. Research and describe at least two other examples of metadata.

Anchor tags

Anchor tags are used to specify all the links within and between web pages. It is this tag that single handedly connects all web pages together to form the largest web of all; the World Wide Web. Every time a user clicks on a link within a browser they are

activating an anchor tag. This includes links to external web pages, navigational elements within individual webs and even links that open images, audio, video and any other type of media file.

There are various options available when using the anchor tag. We restrict our discussion to common examples that deal with the nature of the link itself rather than options in regard to how the link will be formatted on the page or how it will be performed.

- `<A HREF="http://www.pedc.com.au/">PEC Website</A>`

Creates a link to the server that hosts the `www.pedc.com.au` website. `HREF` is short for hypertext reference. The text between the tags (`PEC Website` in the example) forms the link displayed on the page. By default most browsers display this text blue and underlined. Clicking on the link will cause the file `index.htm` (or `index.html`) to be retrieved from the website, interpreted as HTML by the browser and then displayed within the browser.

- `<A HREF="mailto:info@pedc.com.au">information</A>`

Creates a link to the email address `info@pedc.com.au`. In this example the word `information` is displayed in blue and underlined. Clicking on the link causes the user's email program to open with a new message addressed to `info@pedc.com.au`.

- `<A NAME="menu"></A>`

This example creates a bookmark within the web page that may be linked to. In this example the bookmark is called `menu`.

- `<A HREF="#menu">jump to the menu</A>`

Creates a link to the bookmark named `menu` within the current page. When the user clicks on the text `jump to the menu` the browser adjusts the window so the location of the `menu` bookmark is in view.

- `<A HREF="http://www.pedc.com.au/IPT.htm#menu">IPT Menu</A>`

Creates a link to the `menu` bookmark within the file `IPT.htm` that is located on the `www.pedc.com.au` website.

- `<A HREF="images/weblogo.gif">Logo</A>`

Creates a link to the GIF image file `weblogo.gif` located within the `images` directory, which is within the same directory as the web page containing the link. The text `Logo` forms the link, which when clicked retrieves and displays the image.

- `<A HREF="http://www.pedc.com.au/"><IMG SRC="weblogo.gif"></A>`

Creates a link to the `www.pedc.com.au` website, similar to the first example. However instead of the link being displayed as text, the image `weblogo.gif` is displayed and can be clicked. The tag `IMG SRC` is short for image source.

GROUP TASK Practical Activity

Create simple HTML pages using a text editor such as notepad. Include links to each page and back again, and then a link to different websites, individual web pages and also to at least one email address. Test your links by opening the file within a browser.

Uniform resource locators (URLs)

Uniform Resource Locators or URLS are used to identify individual files and resources on the Internet, including the WWW. When using a browser we see the URL of the current web page shown in the address bar at the top of the screen. Entering the URL of a web page into the address bar causes the page to be retrieved and displayed within the browser.

URLs are not only used to access HTML files within web browsers, they are used to uniquely identify and retrieve all types of resources present on the Internet. Most browsers are able to control the transfer of HTML and other files, however they include the ability to redirect requests for other resources to the appropriate client application. For example, `news:microsoft.public.access` when entered into a browser starts the default newsreader and initiates a connection to the newsgroup called `microsoft.public.access`. Similarly `mailto:info@pedc.com.au` when entered into the address bar will execute the default email client with a new message to `info@pedc.com.au`.

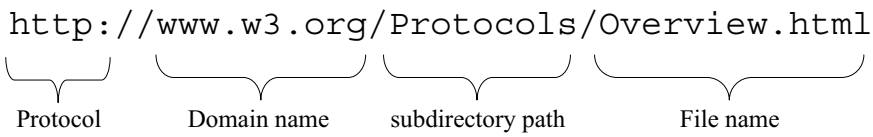


Fig 2.40
Components of a typical URL

Let us consider each of the components of the "typical URL" shown in *Fig 2.40* above. Our discussion is restricted to URLs used to locate web pages and download files within browsers.

- **Protocol**

The protocol identifies the format and method of transmission to be used. A colon follows the abbreviated protocol name. The most common protocol used on the Internet is http (hypertext transfer protocol), this is the protocol used to transfer HTML pages between web servers and web browsers. Most browsers support a secure version of http called Secure Sockets Layer (SSL) or https; https encrypts data during transfer and is commonly used to transfer sensitive data such as bank and other financial transactions.

File transfer protocol (ftp) is used for transferring files of any type. When a file is downloaded directly to a local hard disk or uploaded to a website the transfer is usually accomplished using the ftp protocol. The ftp protocol is supported within most browsers (particularly for downloads). Uploading of website files is usually performed using either dedicated ftp applications or with utilities included within web creation applications.

- **Domain name**

This is the name for the website on the Internet - often called the host name. The domain name is preceded by two forward slashes (//). The domain name is used to locate the computer (web server) that hosts the domain's website.

Every domain name must be unique and is always associated with a unique IP (Internet Protocol) address. The IP address is composed of a set of 4 numbers – each number within the range 0 to 255. For example the IP address for the domain name `www.pedc.com.au` is `203.57.144.42` – not very easy to remember, hence the need for English like domain names. It is possible to enter the IP address directly into the browser in place of the domain name.

Browsers and other Internet software applications communicate with a Domain Name Service server (DNS server) to resolve each domain name into its associated IP address. The IP address is used to locate the correct server as each packet of data is transferred across the Internet. The Windows operating system includes a DNS client called nslookup, which can be executed from a command prompt. For example, typing `nslookup www.pedc.com.au` returns the IP address `203.57.144.42`.

GROUP TASK Practical Activity
Use nslookup, or a similar DNS client, to determine the IP address of some of the domains you have visited lately. Confirm the IP addresses are correct by entering them into a browser's address bar.

Domain names are composed of elements intended for human readers. In general website domain names should commence with www followed by a word or words that describe the company or organisation who owns the domain. The top level of the domain name is the last part that follows the final full stop. There are two types of top level domain names:

1. Generic top level domains (gTLDs). These include `.net`, `.com`, `.org`, `.biz`, `.info` and `.name`. For example `www.microsoft.com` includes the gTLD of `.com`. In the past these domains indicated sites within the USA – this is no longer enforced.
2. Country Code top level domains (ccTLDs). These identify the country of origin for the domain using a 2 letter code. Examples include `.au` for Australia, `.uk` for the United Kingdom, `.nz` for New Zealand, `.us` for the USA, etc. The policy for these names is set by a domain name authority in each country. Each country controls the rules for the allocation of second level domains and hence differences between countries are common. For example, in Australia commercial sites commonly use the `.com.au` second level domain whilst in New Zealand commercial sites use `.co.nz`.

GROUP TASK Research
Create a list of common second level domains used within Australia. Find the equivalent second level domains used in say, New Zealand.

- **Subdirectory path and filename**

Following the domain name is the directory structure that leads to the individual file. The subdirectory path may include many nested directories. In *Fig 2.40* the HTML file named `Overview.html` is located within the `Protocols` directory within the `www.w3.org` directory on the web server that hosts the web site.

It is also possible that a query appears after the filename – the query must be preceded by a question mark to separate it from the file name. For example the URL `http://www.google.com.au/search?q=suzuki+hayabusa` initiates a Google search using the words "Suzuki" and "hayabusa".

HSC style question

Margo is working on a hypertext web presentation describing how to bake a cake. She has created a sequence of four HTML web pages named `cake1.htm`, `cake2.htm`, `cake3.htm` and `cake4.htm`.

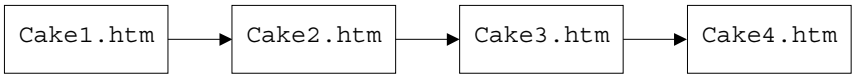
Each of these files is within a single directory on her local hard disk. Within this directory is a subdirectory named `pics`, which contains all the images used on Margo's web pages.

(a) Construct a simplified storyboard that includes a box to represent each web page together with the links between them. Briefly justify your choice of links.

(b) Margo has an image called `candle.gif` that is to be used on each page as a "clickable" image link to the next page. Describe the HTML code required to implement these links.

(c) Margo's web site will be uploaded to the subdirectory `margo` within the `www.cooking.net.au` domain. Identify and describe the URL required to view Margo's cake presentation.

Suggested solution

(a) 

A linear navigation structure is suitable as making a cake is a sequential process where each step needs to be completed prior to the next step commencing. Margo developed her web pages in a strict sequence, so a linear structure encourages users to follow Margo's intended order.

(b) `<A HREF="cake2.htm"><IMG SRC="pics/candle.gif"></A>`

The example above links to the second of the cake web pages (`cake2.htm`). This version would be used on the `cake1.htm` page to link to the `cake2.htm` page. The `cake2.htm` reference would need to be altered appropriately to link to other pages.

The `candle.gif` image is within the `pics` directory, which is a subdirectory within the directory containing the web pages. The source of the candle image must contain the relative path from the web page location to the image, hence the `pics` directory and the file name is needed.

(c) URL is `http://www.cooking.net.au/margo/cake1.htm`

`http:` (Hypertext Transfer Protocol) is the protocol used to transfer HTML files across the Internet. Usually between browsers and web servers.

`www.cooking.net.au` is the domain name. In this case within the top-level domain `.au` which indicates an Australian web site. The domain name is associated with a unique IP address that enables the computer hosting the web site to be located.

`/margo/cake1.htm` is the path and filename of the first web page on Margo's site. This is used to specify the location of the file on the server hosting the `www.cooking.net.au` domain.

Comments

- In part (b) a description of the link is required rather than the actual HTML code. The IPT syllabus does not specify the particular HTML tags that should be known. However, in this question knowing and including the HTML code certainly makes the description simpler.
- In many references the domain name part of a URL is referred to as the name of a computer. This is not strictly true as in most cases a single computer (web server) hosts web sites for many different domains.
- In part (b) the suggested answer refers to a relative path. When a relative path is specified it means the path to the directory containing the relative reference is added to the start of the relative path. For example the relative path `pics/candle.gif` is used within the HTML file `cake1.htm`. The path to the directory that contains the file `cake1.htm` is `www.cooking.net.au/margo/`. This path is added to the relative path resulting in the full path `www.cooking.net.au/margo/pics/candle.gif`

SET 2E

1. In general, most hypertext documents are linked together:
   (A) sequentially.
   (B) non-sequentially.
   (C) randomly.
   (D) using hypermedia.

2. The term 'hypertext' was first used:
   (A) by Apple computer within their Hypercard software.
   (B) when the WWW was created.
   (C) to describe the thought processes of the human mind.
   (D) by Ted Nelson in the early 1960s.

3. A single path through a series of nodes indicates a:
   (A) linear system of navigation.
   (B) hierarchical system of navigation.
   (C) non-linear system of navigation.
   (D) composite system of navigation.

4. The HTML tag
   <A HREF="http://www.eckie.com/pic.jpg">
   www.eckie.com </A>:
   (A) displays an image that links to the www.eckie.com website.
   (B) displays a small version of pic.jpg and links to the full size version.
   (C) displays www.eckie.com which links to the image pic.jpg on the site.
   (D) causes the image pic.jpg to be displayed.

5. Storyboards for designing hypertext displays are composed of:
   (A) nodes and links.
   (B) screen layouts and descriptions.
   (C) a navigation map.
   (D) Both B and C.

6. Metadata is used to:
   (A) describe and define data.
   (B) enter and display data.
   (C) provide search engines with information about an HTML page.
   (D) summarise the content of a web page.

7. Hypertext is thought to better reflect the human mind because:
   (A) the human mind has no structure, thoughts occur randomly.
   (B) it largely operates on associations (links) just like the human mind..
   (C) our minds do not follow logical patterns.
   (D) All of the above.

8. The HTML anchor tag is used to:
   (A) link to email addresses.
   (B) link to images.
   (C) link to other web pages.
   (D) All of the above.

9. The domain name within a URL is:
   (A) the name of a computer.
   (B) the same as an IP address.
   (C) only used by DNS servers.
   (D) the name of an Internet website.

10. In the domain www.hello.com.au:
    (A) .au is the top level domain and .com.au is the second level domain.
    (B) .au is the Australian domain and .com.au is the top level domain.
    (C) com.au is the top level domain and hello.com.au is the second level domain.
    (D) .au is the top level domain and hello.com.au is the second level domain

11. Define the following terms.
    (a) hypertext (b) hypermedia (c) storyboard (d) HTML

12. Explain how links are implemented within HTML web pages. Use examples to assist your explanation.

13. Compare and contrast storyboards used for movie production with storyboards used during the design of hypertext websites.

14. Define the term metadata and explain how metadata is specified and used within Internet webpages.

15. Collect together a sequence of five images. Create five HTML pages that each contains one of these images. Each image should link to the next image's page. The last image's page is to link back to the first image's page.

STORAGE AND RETRIEVAL

Storage and retrieval of data occurs within all information systems, however it is particularly critical in regard to maintaining large data stores. Examples of large data stores include relational databases accessed using a DBMS and also web pages and other online files accessed via the Internet. The performance of such data centred systems is dependent on the efficiency and security of storage and retrieval information processes. This efficiency is determined by a combination of both the hardware that physically maintains and moves the data and also the software that controls and directs this hardware.

Storing and retrieving is a two-part process; storing saves data or information and retrieving reloads data or information. Storing and retrieving supports all other information processes; it provides a mechanism for maintaining data and information prior to and after other information processes. Within large online data stores retrieving occurs just prior to the transmission of data, similarly storing occurs just after data has been received. For such large data stores database management systems (DBMSs) running on dedicated servers are used.

Data Independence
The separation of data and its management from the software applications that process the data.

DBMSs separate data and its management, including its storage and retrieval, from the software applications that process the data. The separation of data and processing is known as "data independence"; it provides the ability of data and its organisation to be altered without affecting or changing the software applications that process the data. For example, adding new fields to a table or altering the data type or length of a field is performed using the DBMS. The DBMS not only supplies data to the software applications, it also supplies details of how the data is organised. This means software applications that process this data do not need to be altered, rather they are able to detect the change and adapt accordingly.

Data independence also makes it easier for different software applications to process the same data and also for particular software applications to process data from a variety of differently organised databases. In addition, a DBMS performs all the storing and retrieving processes – software applications need not concern themselves with the detail of how the data is physically stored, retrieved or even secured.

The opposite of data independence is data dependence, this occurs when software defines the organisation of its own data. Often the software hides how it organises data within the application itself. This makes sharing data with other applications difficult, as they have no mechanism for determining the hidden detail of the data organisation. Data dependence can also occur when specific data values are "hard coded" within software. For instance, a software application may be hard coded with 10% as the value for GST. If the GST rate changes then the software itself must be altered – a much more significant modification compared to simply editing a single data value within a database.

When storing and retrieving processes are viewed at their most basic level, it is true to say that the actual data is not changed, rather the physical method of representing the data changes. For example, when saving data on a hard disk the storing information process physically represents the received data using magnetic fields; when this data is later reloaded the retrieval process converts these magnetic fields into varying electrical signals suitable for transmission. In this section we take a somewhat broader view of storing and retrieving to encompass a variety of other sub-processes.

Consider the DFD (Data Flow Diagram) in *Fig 2.41*, which has been reproduced from our earlier introduction to relational databases. Within this data flow diagram it is clear that the processes performed by the RDBMS essentially involve the storage and retrieval of data – existing data is retrieved and new data is stored. A similar DFD could be drawn for an email, web or file server.

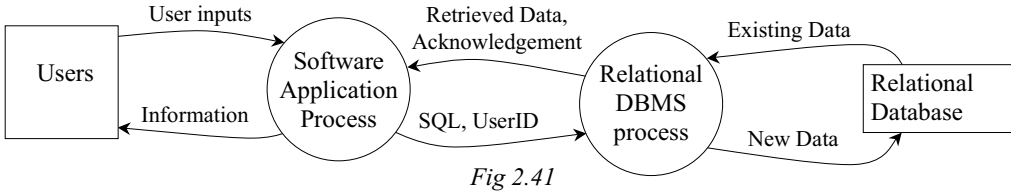


Fig 2.41

RDBMS operate between software applications and relational databases.

In this section our aim is to understand what goes on within the server process of such DFDs. For example, in *Fig 2.41* we cannot see the detail of what goes on within the "Relational DBMS process." We expand this server process into its sub-processes to produce a lower level DFD. As much of the work in this chapter deals specifically with relational databases let us concentrate on this particular type of server for a moment – *Fig 2.42* is a lower level DFD for the "Relational DBMS process" within the *Fig 2.41* DFD.

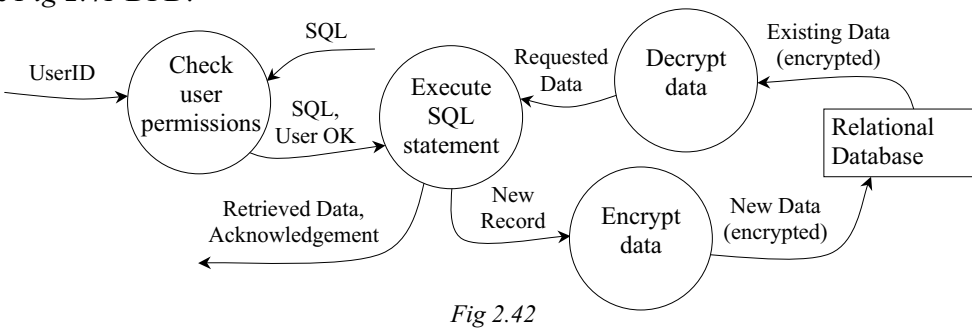


Fig 2.42

DFD describing the sub-processes performed by a RDBMS.

GROUP TASK Activity

Construct a set of DFDs, similar to *Fig 2.41* and *Fig 2.42* above, that describe the data flows and processes performed by a web server.

In *Fig 2.42* ensuring the security of the data figures prominently, namely checking user permissions, encrypting data and decrypting data. Hence we investigate different techniques used to secure data. The "Execute SQL statement" process is where the real work is done – new records are created and stored, existing records are altered, deleted or simply retrieved. Therefore we investigate SQL statements and other tools and query techniques used to both search and sort data. One area not highlighted on the DFD is the hardware used to physically perform these processes – clearly some understanding of the storage hardware is needed.

We therefore consider the following:

- Types of storage hardware including on-line and off-line storage, direct access storage media, namely hard disks and optical disks, as well as tape media used for sequential storage. We examine how the data is physically stored as well as how such devices operate as they store and retrieve data. We also consider the operation of RAID systems and tape libraries used within larger systems.
- Various techniques used to secure data including backup and recovery, user names and passwords, encryption and decryption and also specific techniques used by DBMSs.

- Searching and sorting, including database queries (in particular SQL) and tools used to search hypertext (in particular search engines). We also consider distributed databases where data is stored at different locations yet can be searched as a single entity.

Consider the following:

In our discussion above we continually referred to servers, in particular DBMS servers. A server performs centralised processing for clients – this is known as client-server processing . For example in the DFD shown in *Fig 2.41* the software applications are the clients who are sending requests in the form of SQL statements to the RDBMS server. The server executes the request and returns a response – in our database example the response is either the retrieved data or a simple acknowledgement that the request was performed. The interactions between a browser and a web server are performed in a similar manner.

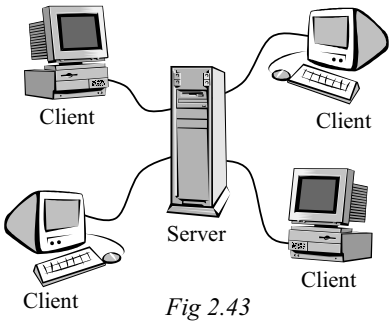


Fig 2.43
Each server provides resources to multiple clients.

These interactions between client and server occur over a network – could be a LAN or even the Internet. Furthermore there are usually many clients making requests to each server. The whole client-server model requires a reliable connection to the server. If the server is offline then the clients cannot continue processing. As a consequence different techniques have been implemented that allow local client processing to continue despite the server being offline.

GROUP TASK Research
Caching of web pages and replication of databases are two techniques that allow client processing to continue offline. Research and describe the fundamentals principles of these two techniques.

STORAGE HARDWARE

During the preliminary course we examined the detailed characteristics and operation of a variety of storage hardware (refer to chapter 6 of the related Preliminary text). Hence we now restrict our treatment to a brief review of this material.

Direct and sequential access

Direct access refers to the ability to go to any data item in any order. Once the location of the required data is known then that data can be read or written directly without accessing or affecting any other data. Often the term random access is used because the data can be accessed in any order, however in reality accessing any data item at random is virtually unheard of.

Sequential access means the data must be stored and retrieved in a linear sequence. For example, in *Fig 2.44* the sixth data item is needed so the preceding five data items must first be accessed. In terms of hardware devices, tape drives are the only widely used sequential storage devices. The time taken to locate data makes sequential storage unsuitable for most applications apart from backup.

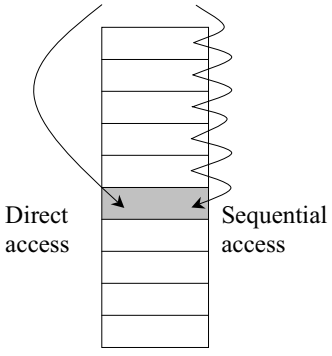


Fig 2.44
Direct access versus sequential access.

On-line and Off-line Storage

Off-line storage refers to data stored such that it cannot be accessed until the storage media is mounted into a drive. Common examples of off-line storage include magnetic tape, optical media such as CDs and DVDs and other portable drives such as thumb or USB storage devices. In terms of large information systems off-line storage is used to maintain backup copies of the on-line data.

On-line storage is available immediately to connected computers. It includes hard disks within a single computer and also storage devices accessed via a network or even over the Internet. On-line storage is usually in the form of hard disk drives, however tape libraries, CD and DVD juke boxes can be used to provide on-line access to tape and optical media. Conversely, systems also exist where hard disk drives are used for off-line backup.

On-line storage over the Internet is becoming common. In this case a third party organisation provides secure, yet flexible, backup and restore services. Many of these services allow backup copies of individual files to be opened and saved on-line across the globe.

GROUP TASK Research
Research examples of organisations that offer on-line Internet storage. Outline the services and security offered by these organisations.

Magnetic Storage

Magnetic storage is currently the most popular method for maintaining large quantities of data. It provides large storage capacity and, in the case of hard disks, it allows for direct access at high speed for both storing and retrieving processes. Optical storage, at the current time, is unable to compete in terms of times required for storing processes.

Digital data is composed of a sequence of binary digits, zeros and ones. These zeros and ones are spaced along the surface of the magnetic medium so they pass under the read/write head at equal time intervals. High magnetic forces are present where the direction of the magnetic field changes; these points are really magnetic poles – indicated by N or S in *Fig 2.45*. It is the strength of the magnetic force that determines a one or a zero, not the direction of the magnetic force. Low magnetic forces occur between two poles and represent zeros. High magnetic forces are present at the poles and represent ones.

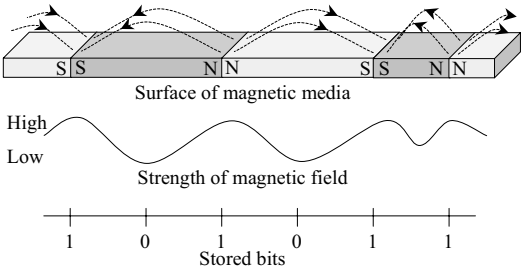


Fig 2.45
Microscopic detail of magnetic storage medium.

Magnetic data is written on to hard magnetic material using tiny electromagnets. These electromagnets form the write heads for both hard disks and tape drives. Essentially an electromagnet is comprised of a copper coil of wire wrapped around soft magnetic material (see *Fig 2.46*). The soft magnetic material is in the shape of a loop that is not quite joined; this tiny gap in the loop is where the magnetic field is produced and the writing takes place.

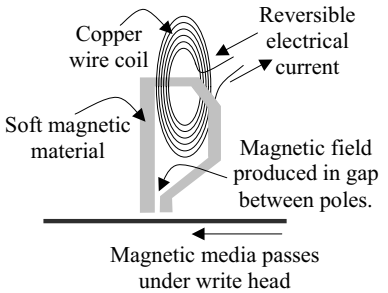


Fig 2.46
Detail of magnetic write head.

Magneto resistant (MR) materials conduct electricity in the presence of stronger magnetic fields. They form the basis of most modern read heads (see *Fig 2.47*). When stronger magnetic forces are detected, representing a 1, the current flow through the MR material increases and hence the voltage increases; similarly when the force is weaker the current and voltage decreases. These voltage fluctuations reflect the original binary data and are suitable for further processing by the computer.

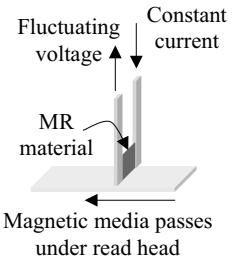


Fig 2.47
Detail of an MR read head.

- **Hard Disks**

Hard disks store data magnetically on precision aluminium or glass platters. The platters have a layer of hard magnetic material (primarily composed of iron oxide) into which the magnetic data is stored. Each platter is double sided, so two read/write heads are required for each platter contained within the drive's casing. The casing is sealed to protect the platters and heads from dust and humidity.

Data is arranged on each platter into tracks and sectors. The tracks are laid down as a series of concentric circles. At the time of writing a typical platter contains more than ten thousand tracks with each track split into two hundred to five hundred sectors. The diagram in *Fig 2.48* implies an equal number of sectors per track; on old hard disks this was true however on modern hard disks this is not the case, rather the number of sectors increases as the radius of the tracks increase. Each sector stores the same amount of data, in most cases 512 bytes. The read/write heads store and retrieve data from complete sectors.

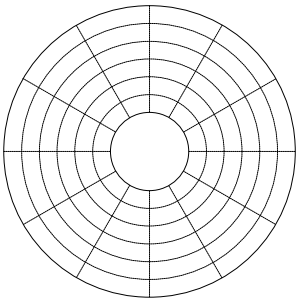


Fig 2.48
Each disk platter is arranged into tracks and sectors.

Each read/write head is attached to a head arm with all the head arms attached to a single pivot point, consequently all the read/write heads move together. This means just a single read/write head on a single platter is actually operational at any instant. Each read/write head is extremely small, so small it is difficult to see with the naked eye. Air pressure created by the spinning platters causes the sliders to float a few nanometers (billionths of a metre) above the surface of the disk.

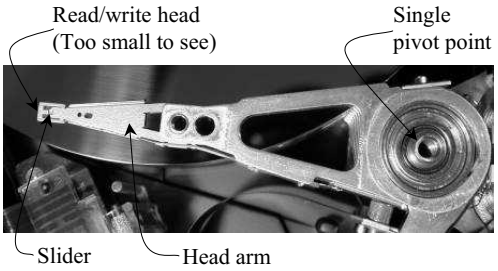


Fig 2.49
Expanded view of a head arm assembly.

- **RAID (Redundant Array of Independent Disks)**

RAID utilises multiple hard disk drives together with a RAID controller. The RAID controller manages the data flowing between the hard disks and the attached computer; the attached computer just sees the RAID device as a normal single hard disk. The RAID controller can be a dedicated hardware device or it can be software running on a computer. In most cases the computer attached to the RAID device is a server on a network. Simple RAID systems contain just two hard disks whilst large systems may contain many hundreds of disks.

RAID is based on two basic processes, striping and mirroring. Striping splits the data into chunks and stores chunks equally across a number of hard disks. During a typical storing or retrieving process a number of different hard drives are writing/reading

different chunks of data simultaneously (see *Fig 2.50*). As the relatively slow physical processes within each drive occur in parallel, a significant improvement in data access times is achieved.

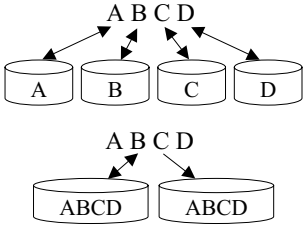


Fig 2.50
Striping (top) and mirroring (bottom) processes are the basis of RAID systems.

Mirroring involves writing the same data to more than one hard disk at the same time. *Fig 2.50* shows the simplest example of mirroring using just two hard disks where both disks contain identical data. When identical copies of data are present on different hard disks the system is said to have 100% data redundancy. Should one disk fail then no data is lost, furthermore the system can continue to operate without rebuilding any data. Hence mirroring makes it possible to swap complete hard disks without halting the system; this is known as 'hot swapping'. Many larger RAID systems also include various other redundant components, such as power supplies; these components can also be 'hot swapped'. Data redundancy and the ability to 'hot swap' components improve the system's fault tolerance.

GROUP TASK Discussion
Data redundancy in RAID systems is a good thing, however data redundancy within relational databases is a bad thing. Discuss reasons for this apparent contradiction.

GROUP TASK Research
There are various different RAID levels – RAID 0, RAID 1, RAID 5 and RAID 0+1 are commonly used examples. Research and describe how each of these RAID levels implements striping and mirroring.

- **Cartridge and Tape**

Magnetic tape has been used consistently for data storage since the early 1950s. At this early stage magnetic tape was the principal secondary storage technology; hard disk technologies first appeared in the late 1950s. Today magnetic tape is contained within cassettes or cartridges. Such cartridges range in size from roughly the size of a matchbox to the size of a standard VHS tape. Tape remains the most convenient and cost effective media for backup of large quantities of data. A single inexpensive magnetic tape can store the complete contents of virtually any hard disk; currently magnetic tapes (and tape drives) are available that can store more than 500GB of data at only a few cents per gigabyte. The ability to backup the entire contents of a hard disk using just one tape far outweighs the disadvantages of sequential access; in any case both backup and restore procedures are essentially sequential processes.

Fig 2.51
Examples magnetic tape cartridges.

There are two different technologies currently used to store data on magnetic tape, helical and linear. Helical tape drives use technology originally developed for video and audio tapes; in fact the majority of the components, often including the actual tape cartridges, are borrowed directly from camcorders. Linear tape technologies were designed specifically for archiving data; hence in terms of data storage most linear systems perform their task more efficiently than helical systems.

- **Tape libraries**

Have you ever made a complete backup copy of a hard disk? It involves manually swapping media and a good deal of time; these are major disincentives. Now imagine performing the same process for all the data held by a large organisation; hundreds or even thousands of tapes need to be swapped taking days or even weeks to complete. Clearly the backup process needs to be automated; this is the purpose of tape libraries.

Fig 2.52
Sony TSL-400C tape library.

Various different size tape library devices are available to suit the demands of different information systems. The smallest, such as Sony's TSL-SA400C in *Fig 2.52*, hold just four tapes and use a single drive; these devices provide capacities suited to most small businesses. Larger devices hold hundreds or even thousands of tapes and contain many drives. Large government departments and organisations link multiple tape library devices together; such systems hold hundreds of thousands of tapes and many thousands of tape drives. Backup processes on such large systems continue 24 hours a day, seven days a week.

Large tape libraries, such as StorageTek's SL8500 shown in *Fig 2.53*, include a robotic system to move tapes between the storage racks and the tape drives. The actual tape drives are just standard single tape drives whose operation has been automated. The robots select individual tapes from racks and place them individually into each drive just like a human hand would. The use of standard tape drives allows faulty drives to be replaced whilst the system continues operating – the remaining drives simply take up the slack. Other components are also duplicated – such as the robotics, power supplies and even the circuit boards controlling the system – the aim being to improve fault tolerance.

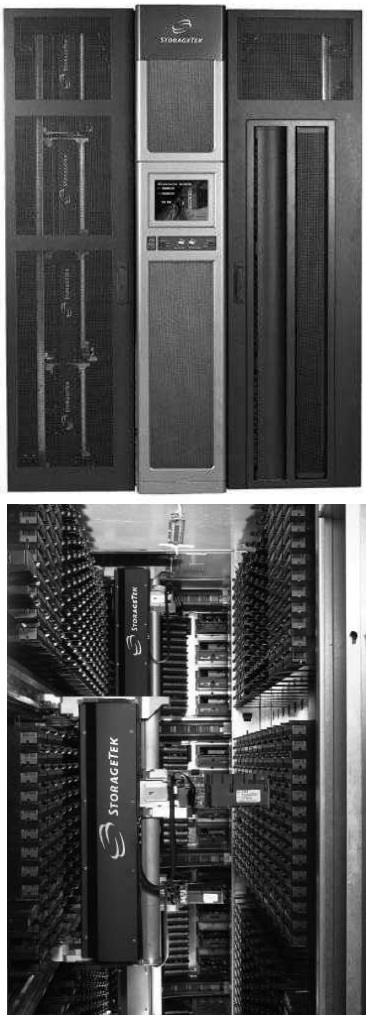


Fig 2.53
Exterior and interior of StorageTek's StreamLine™ SL8500 tape library.

GROUP TASK Discussion
Redundant (duplicate) components are common within many devices present in server-based systems. Explain how these redundant components improve the fault tolerance of such systems.

GROUP TASK Research
Research, using the Internet or otherwise, the storage capacity and data access rates for a single tape drive. How do these statistics compare with similar statistics for single hard disks?

Optical storage

Optical storage processes are based on reflection of light; either the light reflects well or it reflects poorly back to the drive's sensor. It is the transition from good reflection to poor reflection or vice versa, that is used to represent a binary one (1); when reflection is constant a zero (0) is represented. This is similar to magnetic retrieval where a change in direction of the magnetic force represents a binary one and no change represents a zero.

As the data is so tightly packed on both compact disks (CDs) and digital versatile disks (DVDs) it is essential that the light used for optical storage processes be as consistent and as highly focussed as is possible; lasers provide such light. Essentially a laser produces an intense parallel beam of light; accurately focussing this light produces just what is needed for optical storage and retrieval processes. Relatively weak lasers are used during the retrieval of data and much higher-powered lasers when storing data. Higher-powered lasers produce the heat necessary to alter the material used during the CD or DVD burning process.

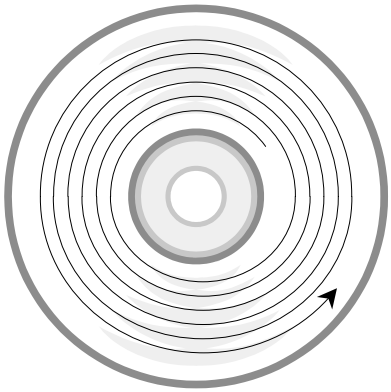


Fig 2.54
CDs and DVDs contain spiral tracks.

CDs contain a single spiral track that commences at the inner portion of the disk and spirals outward toward the edge of the disk (see *Fig 2.54*). This single track is able to store up to 680 megabytes of data. DVDs contain similar but much more densely packed tracks, each track can store up to 4.7 gigabytes of data. Furthermore, DVDs may be double sided and they may also be dual layered. Therefore a double sided, dual layer DVD would contain a total of four spiral tracks; in total up to 17 gigabytes of data can be stored.

Each spiral track, whether on a CD or a DVD, is composed of a sequence of pits and lands. On commercially produced disks the pits really are physical indentations within the upper side of the disk. *Fig 2.55* depicts the underside of a disk; this is the side read by the laser, and hence the pits appear as raised bumps above the surrounding surface. On writeable media the pits are in fact not pits at all; rather they are areas that reflect light differently. The essential point is that pits reflect virtually no light back to the sensor whilst lands reflect most of the light back to the sensor.

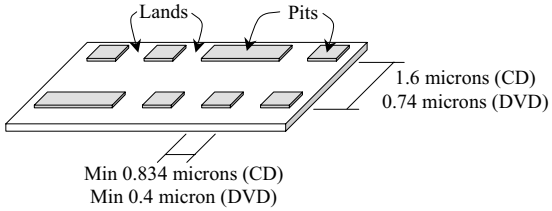


Fig 2.55
Magnified view of the underside of an optical disk.

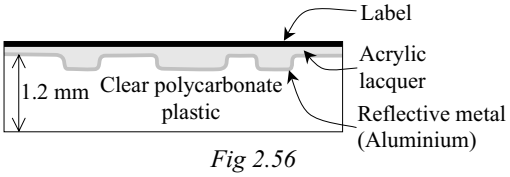


Fig 2.56
Cross section of a typical commercially produced CD or single sided single layer DVD.

Both CD and DVD media are approximately 1.2mm thick and are primarily clear polycarbonate plastic. On commercially produced disks the pits are stamped into the top surface of the plastic, which is then covered by a fine layer of reflective metal (commonly aluminium), followed by a protective acrylic lacquer and finally some sort of printed label. On recordable and rewriteable media a further layer is added between the polycarbonate and the reflective layer; it is this layer whose reflective properties can be altered to store data using a higher powered laser. Double layer DVDs contain two data layers where the outside layer is semi reflective; this allows light to pass through to the lower layer. The laser is accurately focussed onto the layer being read.

SECURING DATA

Data security is about achieving two somewhat distinct aims. Firstly it aims to prevent data being lost or corrupted; this ensures the system remains operational – or at least can be put back into an operational state. Secondly it aims to prevent unauthorised access to data; this includes restricting access completely to outsiders and it also includes assigning specific levels of access to participants within the system.

The table shown in *Fig 2.57* lists common techniques for securing data aligned with the above two aims. No single technique is sufficient on its own; rather a combination of many techniques should be used. Different information systems will require a different balance of data security techniques. The choice of techniques is largely determined by the sensitivity of the data and how critical the data is to the organisation's continued operation. Consideration should be given to the potential repercussions should the data be lost completely, corrupted and/or accessed by others.

	Protects against:	
Techniques	Data loss	Unauthorised access
Backup and recovery	✓	×
Physical security measures	✓	✓
Usernames and passwords	✓	✓
Encryption and decryption	×	✓
Restricting access using DBMS views	×	✓
Record locks in DBMSs	✓	×
RAID (Mirroring only)	✓	×

Fig 2.57
Data security techniques.

Backup and Recovery

Making a backup of data is the process of storing or copying the data to another permanent storage device, commonly recordable CD/DVD, magnetic tape or a second hard disk. In the classroom you may well use a USB thumb drive as your preferred backup device. Recovery of data is the opposite process where the data is retrieved or restored from the backup copy and placed back into the system.

Backup
To copy files to a separate secondary storage device as a precaution in case the first device fails or data is lost.

The aim of creating backups is to prevent data loss in the unfortunate event that the original data is damaged or lost. Such damage most often results from hard disk failures; in fact it is inevitable that all hard disks will eventually fail. Some other reasons for data loss or damage includes software faults, theft, fire, viruses, intentional malicious damage, insufficient or inappropriate validation that accepts unreasonable data, and even intentional changes that are later found to be incorrect. For backup copies to most effectively guard against such occurrences regular backups are required and these backup copies should be kept in a fireproof safe or at a separate physical location.

Even the most reliable computer will eventually break down and the consequences of such breakdowns can be devastating if no backups have been made. Consider a small business with some 100 clients; a total loss of data means loss of all client records, orders and invoices, together with any correspondence and marketing materials. Even if much of this information is maintained in paper-based storage the cost of recovering from such a loss is enormous in comparison to the minor costs involved to maintain regular backups. Now extrapolate this impact to a large corporate organisation and imagine the effect if all their data is lost.

There are two types of backup that are commonly used; full backups and partial backups. A full backup includes all files whereas a partial backup includes only those files that have been created or altered. Most operating systems include an archive bit stored with each file to simplify partial backups; each time a file is created or altered the archive bit is set to true. Backup and recovery utilities examine this bit to determine files to be included in a partial backup. Partial backups only include files where the archive bit is set to true.

Incremental and differential backups are two common backup strategies that include partial backups. Both strategies require a full backup to be made at regular intervals; commonly once a week, such as each Friday. Each full backup sets all archive bits to false. On other days a partial backup is performed. Incremental backup strategies set the archive bit on each successfully copied file to false during each partial backup, whilst differential backup strategies do not. Therefore each partial backup made using an incremental strategy contains only files that were created or changed since the last partial backup. If a failure occurs then a sequence of backup copies must be restored in the order they were originally made commencing with the last full backup. On the other hand, each partial backup made using a differential strategy will contain all files that have been created or changed since the last full backup was made. If a failure occurs then the last full backup is restored followed by the most recent partial backup.

The frequency at which backups are made depends on how critical the data is to the organisation and how frequently the data changes. Usually a full backup is made at least once a week with partial backups being made daily. A further safeguard against data loss is to rotate the media used for backups; commonly three complete sets are used. This means that should one set of backups also be corrupted then the previous set can be used for data recovery.

GROUP TASK Research
Research the backup strategy used at your school or work. Analyse this strategy to determine the maximum data loss possible if all data in the operational system is lost.

Physical Security Measures

Physically securing the room in which servers and other system critical devices are located is an obvious technique for reducing data loss and unauthorised access. For large systems all hardware critical to the system's operation is held within a locked climate controlled room of substantial construction (see *Fig 2.58*). Only persons who need to use the room are given access.

Fig 2.58
Secure climate controlled server room.

Access controls of various types are implemented to prevent unauthorised persons entering the room. Such persons include relatively innocuous people such as interested employees simply wishing to have a look, all the way through to terrorists who may wish to bomb and completely destroy the facility. Clearly the level of security should reflect the nature of the perceived threats. For example a local ISP would include secure locks on an otherwise normal room whilst a government's military computer facility would be housed within a solid bomb proof concrete bunker

style room. Locks on doors can be controlled by keys, passwords, smart cards or in many cases biometric readers such as fingerprint and iris scanners.

GROUP TASK Discussion
In high security systems even the nature of the physical security is a secret. Why is this? Discuss.

Climate control systems within such facilities monitor and adjust both temperature and humidity. Components expand and contract as temperature changes – particularly precision metallic parts. Maintaining a constant operating temperature minimises such effects and increases the life of components. Moisture is the enemy of all electrical and mechanical parts; hence maintaining low humidity levels prolongs the life of components and increases the system's reliability.

Usernames and Passwords

Passwords can be used to secure individual files, directories or even entire storage devices. A combination of user names and passwords are used by operating systems, network software and various other multi-user applications to confirm the identity of users. Once the user has been verified the system assigns permissions based on their user name – typically create, read, write and delete access to particular directories and software applications are assigned to the user. If the files are accessed over a network then these permissions are set by the network administrator – we discuss these tasks in some detail within the Communication Systems topic. Users can set passwords for individual files from within the file's related software application.

Data secured by passwords is only secure whilst the passwords remain secret. There are numerous techniques and also software applications available for working out passwords. Furthermore, remembering many different passwords is difficult, hence people tend to either use the same password for multiple systems or they write down their passwords. There have been cases where the user names and passwords for entire systems have been typed into totally unsecured text files, which are easily accessible to intending hackers. The next two security techniques, namely encryption/decryption and the use of database views also require the use of user names and passwords.

GROUP TASK Discussion
Many online systems specify the minimum length for passwords and they do not allow certain passwords, such as words or all digits. Other systems ask for passwords to be re-entered at regular intervals. Identify types of security threats such techniques would protect against and also threats such techniques would not protect against.

Encryption and decryption

The science of developing and analysing encryption and decryption technologies is called cryptography. The military have used cryptography to secure messages for hundreds of years. In fact many of the techniques and strategies now widely used evolved from these military applications. Cryptography has now become a major industry due to the widespread need to secure sensitive digital data.

Encryption
The process of making data unreadable by those who do not possess the decryption key.

Decryption
The process of decoding encrypted data using a key. The opposite of encryption.

Encryption alters raw data in such a way that the resulting data is virtually impossible to read. Therefore should unauthorised access occur the infiltrator just sees a meaningless jumble of nonsense. Of course, this would be a pointless exercise if authorised persons cannot reverse the process and decrypt the data. To enable decryption, secret information, called keys, are used. The key contains sufficient information to encrypt and/or decrypt data to the required level of security. Some systems use a single key for both encryption and decryption whilst others use a different key for each process.

Single key encryption is commonly called symmetrical or secret key encryption. The same key is used to decrypt the data as was used for encryption. Such systems are commonly used to encrypt data held on secondary storage devices. Software on the device itself, or at least the attached computer, does all the encrypting and decrypting. As a consequence it is not necessary for the secret key to be shared, although it must be securely protected. If the user or computer decrypting the data is different from the one who encrypted the data then the secret key must be shared with both parties. A secure encryption technique is needed to communicate the secret key. Solving issues such as this is the job of cryptographers; one solution is the use of systems that use two keys.

Two key systems utilise a public key for encryption and a private key for decryption; they are known as asymmetrical or public key systems. Each user of the system has a public key and a private key. The public key can be distributed freely to anybody or any computer, however the private key must never be divulged. Let us consider a typical transfer of data, say from Fred to Jane (see *Fig 2.59*). Jane has her own personal public and private key, as does Fred. Fred first sends a plain message to Jane requesting her public key. Jane responds by sending Fred a copy of her public key; Fred uses this key to encrypt the message. He then sends the encrypted message to Jane. Jane receives the message and decrypts it using her private key. The message is secure during the transfer as only Jane's private key is able to decrypt the message, and Jane is the only one who has this key. It doesn't matter if Jane's public key is intercepted during the transfer as it can only be used for encrypting messages, not decrypting them. Our example used two people, but in reality the transfer is more likely between two computers.

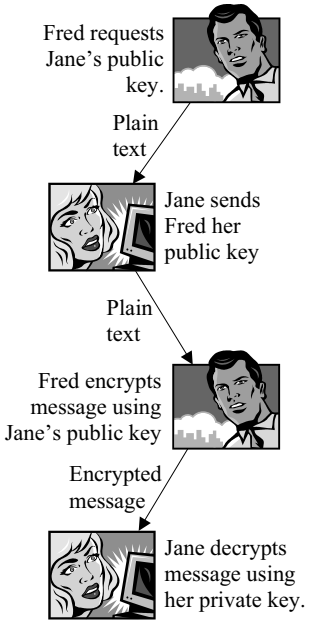


Fig 2.59
Typical transfer using a public or two key system.

GROUP TASK Discussion
Secure Sockets Layer (SSL) is a protocol included within all current browsers. The SSL protocol is being used when a URL commences with HTTPS: and a small lock appears in the browser's status bar. Create a list of examples where HTTPS: is used.

Consider the following:

It is common for systems that store highly sensitive data to use a combination of encryption techniques. In many organisations users carry flash memory-based smart cards containing their private keys. These cards must be inserted into a reader before

any data can be decrypted and viewed. On file servers, data is encrypted using a different technique, often involving further levels of encryption.

The data stored on many file servers is encrypted, and the key for decrypting this data is itself held on a removable flash device attached to the file server. During retrieval the file server uses the key on its flash device to decrypt the data, then prior to transmission the data is encrypted using the public key of the current user. Once the user receives the data it is decrypted using the private key on their smart card.

However what if a user's smart card is stolen? Surely the thief then has complete access. To counteract this possibility a password can be used to confirm the user's identity corresponds with the owner of the smart card. But passwords can be guessed, or users can divulge their password. Such problems can be overcome using biometric data such as fingerprints to replace passwords, the biometric data being used to confirm the identity of the user.

Even more elaborate schemes can be used. Some storage systems use a different key to encrypt every file. They then encrypt each of these individual keys using the key on the server's flash card. Such systems allow the key on the flash card to be changed at any time without the need to decrypt and then encrypt all the data on the entire storage device. Similarly the use of smart cards for users means their public and private keys can easily be altered at any time.

GROUP TASK Discussion
Are such detailed encryption techniques really necessary? What types of data are so important that they need this level of security? Identify and discuss examples of data where such encryption is necessary.

Restricting Access using DBMS Views (or User Views)

Restricting access within databases commonly involves restricting access to particular views of the data based on usernames and the client applications being executed by these users. A view or user view is essentially the resulting data from a SELECT query – SELECT queries can be used to restrict the fields and records retrieved from one or more tables. We examine SELECT queries in some detail later in this chapter – indeed this section on views will become clearer once we complete this work. The difference between views and select queries is the way they are treated by the DBMS. A view is optimised by the DBMS to improve performance and its details are stored as an integral part of the database. Queries are constructed and executed as required – usually at the request of client software applications.

View (or User View)
The restricted portion of a database made available to a user. Views select particular data but have no affect on the underlying organisation of the database.

When a DBMS view is created it behaves and can be manipulated just like a real table – views are also known as virtual tables. The view itself does not contain any actual data, rather it specifies the organisation of parts of the real database's organisation. The actual organisation of a database is described within its schema, hence views are sometimes known as sub-schemas because they include parts of the complete database schema. When setting user permissions each user is given access to particular views rather than particular tables and fields. This technique provides the flexibility to include the current user within the view specification, for example when accessing banking details over the Internet you are accessing a view of the data that selects records that match your username. Users and client software applications use views in

the same way as they use tables; in fact from the perspective of users and client applications views are effectively identical to real database tables.

Views are not merely created to assist data security; they also improve data independence by providing a simplified view of the data suited to the needs of particular client software applications. Most large databases are accessed by a number of different applications, each application is written with the expectation that the data will be available in its preferred format. DBMS views allow the data within a single database to be manipulated by different software applications in a format that suits that application. For example in a Hotel system one software application is used at the front desk to check guests in and out, and another is used behind the scenes to create financial reports. Each of these applications uses a different view of the same data.

Each user is assigned a set of permissions for each view of the data they require to perform their processes. For example, an order entry clerk may be able to read customer details but not change them, yet they may be able to both add and edit invoices. The order entry clerk would be assigned read permission for the customer details view of the data and create, read and write (and probably delete) permissions to the invoice view of the data. Each of these views would exclude fields not required by the data entry clerk to complete their work.

Usually users are required to enter a user name and password each time they use a particular database, however larger DBMS systems utilise the network user name to verify the identity of the current user. In either case the identity of the user is determined and their data access rights assigned accordingly.

GROUP TASK Discussion
User views of the data are used when creating data entry screens (forms) and also when creating reports to output information. Discuss reasons why a view would be preferred over accessing the actual tables directly.

Record locks in DBMSs

DBMS software retrieves records rather than files, as a consequence editing can also be controlled based on records rather than complete files. Imagine two users have retrieved the same record, if both users subsequently make changes to this record then which version of the record should be stored? The DBMS must implement a strategy whereby records can be locked; commonly DBMSs provide two different strategies; pessimistic locking and optimistic locking.

Pessimistic locking, as the name suggests, is somewhat negative. The first user to start editing the record effectively locks the record and hence subsequent users must wait for the updated record to be stored before they can commence editing; often a visual aid is used to inform the user. Microsoft Access uses the symbol "⊘" to indicate that another user is currently editing the record (see *Fig 2.60*). A pessimistic strategy requires the DBMS to be informed and lock the record whenever a user commences editing any record. Such a strategy, although the most common, adds considerably to the amount of processing required of the DBMS.

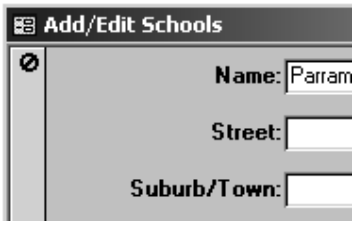


Fig 2.60
Microsoft Access displays a ⊘ symbol when pessimistic locking is active and another user is editing a record.

Optimistic locking is a much more positive strategy. It is based on the assumption that conflicts will rarely occur. Such a strategy does not require the DBMS to be informed as editing commences, rather the DBMS checks for record changes prior to storing each record. If another user has made a change to a record then there are two possible

options, either the currently stored record can be overwritten or the current users changes can be discarded. Commonly the user is given the task of making this decision via a warning message. *Fig 2.61* shows the default message generated by Microsoft Access. In either case all but one user is destined to lose their changes. Clearly an optimistic locking strategy can have dangerous consequences in terms of maintaining data integrity. For instance say user A is updating a customer's phone number and whilst this is occurring User B begins updating the same customer's address. Using an optimistic locking strategy one of the two changes will definitely be lost. This cannot occur when pessimistic locking is implemented.

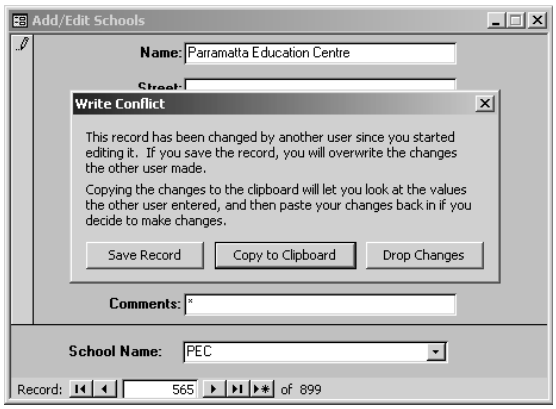


Fig 2.61
Microsoft Access provides 3 options in response to write conflicts when optimistic locking is enabled.

GROUP TASK Practical Activity
Have two users attempt to edit the same record simultaneously within a database. Determine the locking strategy being used.

Consider the following unfortunate situations:

- A hard disk drive fails within a server that manages data critical to the day-to-day operations of a small business. The IT manager discovers to his dismay that his only tape backup will not restore correctly.
- The building that houses an Internet service provider (ISP) is completely destroyed by fire. The ISP maintained duplicate servers on their site that each included mirrored RAID storage. Unfortunately all hardware and data it contained has been irreparably damaged.
- You have been working on a large assignment on your computer over a number of weeks. You have just spent an hour or so making changes suggested by some of your friends. You have been regularly saving your work. During the next day at school you realise that the changes suggested by your friends were incorrect. Unfortunately you are unable to easily reverse all the changes you made last night.
- An executive strongly suspects that members of the IT department are reading her emails. She is unable to prove her suspicions, however it seems IT staff are aware of many new company initiatives that have only ever been described within private email messages.
- A company's database server that contains confidential data including credit card numbers is stolen. During the weeks that follow the robbery, many customers report fraudulent purchases against their credit card accounts.

GROUP TASK Discussion
Propose suitable security strategies and procedures that would have prevented each of the above unfortunate situations.

SET 2F

1. Storing and retrieving information processes:
   (A) alter the actual data within the system.
   (B) are used to maintain data in support of other information processes.
   (C) represent data as a sequence of high and low voltages.
   (D) encrypts and decrypts data.

2. Which of the following is an example of a sequential storage device?
   (A) hard disk.
   (B) optical disk.
   (C) RAID.
   (D) tape.

3. On magnetic media binary ones are represented:
   (A) where the magnetic forces are low.
   (B) where the direction of the magnetic force changes.
   (C) where the direction of the magnetic force is constant.
   (D) between the north and south poles.

4. Which of the following is true of magneto resistant (MR) materials?
   (A) Current increases through MR material in the presence of higher magnetic fields.
   (B) MR material is used within the read heads of magnetic storage devices.
   (C) The voltage through the MR material changes in proportion to the stored magnetic field.
   (D) All of the above.

5. On modern hard disks, which of the following is FALSE?
   (A) All tracks contain the same number of sectors.
   (B) Each sector stores the same amount of data.
   (C) Complete sectors of data are read and written.
   (D) Each platter has its own read/write head.

6. Within many RAID systems the same data is written to different disks. This is known as:
   (A) mirroring.
   (B) striping.
   (C) hot swapping.
   (D) fault tolerance.

7. On optical storage, how are binary ones represented?
   (A) Each pit represents a binary one.
   (B) Each land represents a binary one.
   (C) Continuous pits or lands represent binary ones.
   (D) The transition from pit to land and land to pit represents binary ones.

8. Which of the following is true for single or secret key encryption?
   (A) Two different keys are used.
   (B) The key must be known by both sender and receiver.
   (C) Only the receiver knows the decryption key.
   (D) Commonly used to secure the initial data transferred between two parties.

9. How does creating views of a database help secure data?
   (A) Views only allow users to see one record at a time.
   (B) A view presents the data in a form suited to the requirements of client software applications.
   (C) Users are unable to access data not included in their assigned views of the data.
   (D) Both B and C.

10. In a backup system that uses archive bits, what must occur after a full backup?
    (A) All archive bits are set to true.
    (B) The archive bits for new and altered files are set to true.
    (C) All archive bits are set to false.
    (D) The archive bits for existing files are set to true.

11. Describe the following processes and provide an example of each.
    (a) Backup (b) Recovery (c) Encryption (d) Decryption

12. Identify and describe techniques used by a DBMS to secure data.

13. Explain how data is physically stored on:
    (a) Hard disks (b) Magnetic tape (c) CD-ROM

14. With regard to data security, compare and contrast RAID systems with tape libraries.

15. Many people now use the Internet to perform many bank transactions. Identify and describe likely techniques used to secure data during these transactions.

OVERVIEW OF SEARCHING, SELECTING AND SORTING

Searching, selecting and sorting are really analysing information processes; they take data and transform it into information. In most cases the information is displayed immediately after these processes are completed. For example within databases records are selected and sorted and then the resulting information is immediately displayed on forms and reports. Note that in the context of databases (and also search engines) these analysis processes determine the data that is retrieved, therefore searching, selecting and sorting are also sub-processes that occur within retrieving information processes.

Search
To look through a collection of data in order to locate required data.

Sort
To arrange a collection of items in some specified order.

Both searching and selecting are processes that combine to identify the data to be retrieved. Commonly the term 'searching' is used to describe the process of actually looking through the data – comparing each data item against some specified criteria. Selecting then takes over and highlights or lists each of the found items. Within databases only the records that precisely match the search criteria are selected and retrieved. Most search engines are less pedantic; they ignore some common words and try to correct minor spelling errors.

Sorting arranges data into alphabetical or numerical order. When data is sorted, it becomes easier for people to understand and use – that is, the data is transformed into information. Furthermore sorting is used to arrange data into categories – the data in each category has one or more attributes in common. For example, sorting high school students by their school year results in a list of all year 7 students, followed by all year 8 students, and so on up to all of year 12 – the students are categorised by school year.

Digital data is always represented as binary numbers; therefore sorting digital data is always a numerical process– even for text. Alphabetical sorts use the binary number codes that represent each character to determine the sort order. For example in the ASCII system A is represented by 1000001_2 or 65, B is represented by 66, C by 67 and so on. Alphabetical sorts compare the numerical value representing characters from left to right; if two characters are found to have the same value then the next corresponding characters are considered. An ascending alphabetical sort, as expected, places "Balloon" before "Barrow" as the ASCII value or number representing "l" comes before the number representing "r".

Problems occur when numbers are incorrectly represented as text. For example if the following data is defined as text, sorting -500, -5.6, -0.001, 2, 12 and 100 into ascending alphabetical order produces the result -0.001, 100, 12, 2, -5.6 and –500 in most databases. This is unlikely to be the required result. Essentially an alphabetical sort has been performed rather than a numerical sort, but the hyphens (which look like negative signs) have been ignored. This occurs because most databases use an alphabetical sort technique known as "word sort". Word sort ignores hyphens and apostrophes and rates other punctuation before normal digits and letters.

Numerical sorts consider the total numeric value of the data item; hence an ascending numerical sort, as one would expect, arranges the data from smallest negative value to highest positive value. For example, -500, -5.6, -0.001, 2, 12 and 100 is the result when the same data items are defined as numeric and sorted into ascending numerical order. Predictably, a descending numerical sort results in this list being reversed.

GROUP TASK Practical Activity
Experiment with suitable sample data in a database to determine if "word sort" is being used by the DBMS.

TOOLS FOR DATABASE SEARCHING AND RETRIEVAL

In a database table the order in which records are physically stored is not significant; conceptually records exist in no particular order. When a search process is initiated either each record must be examined in turn or the records must be arranged in order so the search can occur more efficiently. When sorts are applied to large tables with many thousands or even millions of records either technique is a potentially lengthy process. The answer is to use indexes.

Indexes within database tables are similar to those in the back of a book. Think about the index in a book; it provides an alphabetical listing, where each entry points to a specific page. Indexes within database tables operate in a similar way; they describe a particular record order without actually ordering the records. The index is in order hence it can be used to quickly search through the data. The required records can then be retrieved. For indexes to perform they must remain up to date – inaccuracies can occur each time a new record is added or data in an indexed field is edited. In smaller systems the index is updated immediately new data is entered or existing data is edited. Within larger systems this can take some time, so indexes are rebuilt at a later time – commonly late at night when the system is relatively idle. Indexes should only be specified for key fields and other specific fields that are used within common searches and sorts.

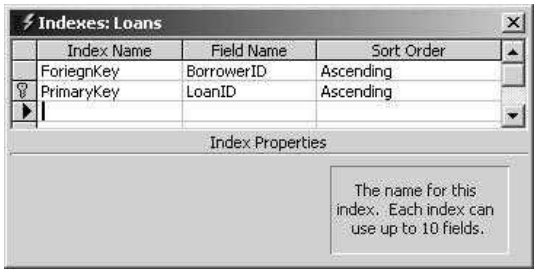


Fig 2.62
Defining indexes in Microsoft Access.

GROUP TASK Discussion
DBMSs automatically create an index for each primary key. In most DBMSs these PK indexes cannot be removed. Why is this? Discuss.

The remainder of our discussion on searching and sorting databases examines the syntax required to specify different types of searches and sorts – primarily using SQL We first examine examples where the source of the data is a single table (or it could be a simple flat-file). We then consider searches across multiple tables in relational databases. Throughout our discussion we shall use our Library and Invoicing databases created using Microsoft Access earlier in this chapter.

Searching and Sorting Single Tables (including Flat-Files)

In SQL (Structured Query Language) searching and sorting is performed using the SELECT statement. The general syntax of the SELECT statement is described in *Fig 2.63*. This is by no means a thorough definition of the SELECT statement, however it is sufficient for our current purpose. Following the SELECT keyword is the list of attributes or fields that will be retrieved – replacing this list with an asterisk "*" causes all attributes to be retrieved. The FROM keyword is used to specify the tables from which the data will be retrieved – currently we're interested in single tables so just one name

```
SELECT (attributes to retrieve)
FROM (list of table names)
WHERE (search criteria)
ORDER BY (list of attributes)
```

Fig 2.63
SQL SELECT statement general syntax

will be used here. The WHERE keyword specifies the search criteria, for example WHERE LastName="Nerk". The ORDER BY clause specifies how the retrieved records should be sorted, for example, ORDER BY LastName, FirstName.

Let us consider our Borrowers table from the Library database we created earlier in this chapter. Assume the table holds the 10 records shown in *Fig 2.64* below:

Borrowers : Table

BorrowerID	FirstName	LastName	PhoneNumber	JoinDate	LoanDuration
1	Fred	Nerk	98769876	15/02/06	14
2	Mary	Lamb	92343456	07/01/01	14
3	John	Boole	46783456	01/04/02	14
4	Jack	Spratt	34567890	02/05/04	21
5	Michael	Nerk	90123456	01/08/02	14
6	Martin	Lamb	33344567	03/12/03	14
7	Susan	Nerk	67845599	03/12/02	21
8	Kylie	Nerk	98769876	05/11/04	14
9	Harold	Nugent	45536589	02/03/04	14
10	Lachlan	Adams	23345667	23/05/05	14
(AutoNumber)					

Record: 10 of 10

Fig 2.64
Sample records in the Borrowers table

Say, we wish to find all the borrowers whose last name is Nerk and sort these records into alphabetical order based on their last name and then their first name. In Microsoft Access this process is simplified using the included query design grid graphical user interface (GUI) – *Fig 2.65* shows this GUI with the specifications of our query. Using a GUI such as the one supplied with Access greatly simplifies writing SQL (and is also a great way to learn SQL). In Access the equivalent SQL is displayed or viewed via the SQL view window. SQL view can be used to enter the SQL statements directly. In this case the equivalent SQL is:

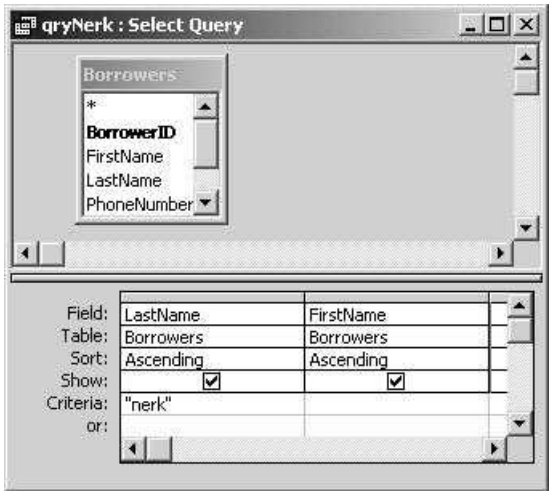


Fig 2.65
Access GUI for creating SELECT queries.

```
SELECT Borrowers.LastName, Borrowers.FirstName
FROM Borrowers
WHERE Borrowers.LastName="Nerk"
ORDER BY Borrowers.LastName, Borrowers.FirstName
```

qryNerk : Select Query

LastName	FirstName
Nerk	Fred
Nerk	Kylie
Nerk	Michael
Nerk	Susan

Record: 4 of 4

Fig 2.66
Data sheet view in Access shows the retrieved records.

In SQL starting a new line for each clause in the SELECT statement is not necessary, rather it simply makes the SQL easier for us humans to read – the DBMS does not care. Notice that both the table name and the attribute name are included; this is needed to avoid confusion should two or more tables include the same attribute name – when all attribute names are different use of the table name is optional. The search data "Nerk" is enclosed within double quotes (single quotes are also okay) – this is the standard way of

differentiating particular text from an attribute name. When numeric attributes are specified then specific data values must be numbers. As numbers are not legitimate attribute names then no delimiting quotes are required. The data retrieved by this query is reproduced in *Fig 2.66* – Microsoft Access calls this datasheet view.

Let us focus on the search criteria following the WHERE keyword. The search criteria is constructed using various relational and logical operators. Common examples of these operators are shown in *Fig 2.67*. Rather than explain the detail of each operator let us consider example queries that use these operators within their WHERE clause. We shall base each of our examples on the following SELECT query applied to the sample data shown in *Fig 2.64*. Note when no WHERE clause is included at all, the query returns 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 – all the BorrowerIDs. The records returned and brief comments accompany each example.

```
SELECT Borrowers.BorrowerID
FROM Borrowers
WHERE (search criteria)
ORDER BY Borrowers.BorrowerID
```

Relational Operators	
English meaning	**SQL**
CONTAINS	LIKE
DOES NOT CONTAIN	NOT LIKE
EQUALS	=
NOT EQUAL TO	<>
GREATER THAN	>
GREATER THAN OR EQUAL TO	>=
LESS THAN	<
LESS THAN OR EQUAL TO	<=
Logical Operators	
True when both expressions True	AND
True when at least one expression True	OR
Opposite	NOT

Fig 2.67
Common relational and logical operators.

Consider the above SQL when the search criteria is:

- LastName=”Nerk”
  Returns 1, 5, 7, 8. All records where LastName exactly equals Nerk. Although note that many DBMSs by default are not case sensistive, that is, nerk and nERk would also match.
- LastName>”Nerk”
  Returns 4, 9. Only last names that are alphabetically after Nerk are returned.
- LastName LIKE ”n*”
  Returns 1, 5, 7, 8, 9. The asterisk is a wild card that represents zero or more characters, hence all last names commencing with an ‘n’ are returned.
- LastName LIKE ”*m*”
  Returns 2, 6, 10. All last names that contain an ‘m’.
- NOT(LastName LIKE ”*m*”)
  Returns 1, 3, 4, 5, 7, 8, 9. Opposite of the previous example, that is all last names that do not contain an ‘m’.
- LastName LIKE “???a*”
  Returns 4. The question mark is a wild card that represents any single character, hence all last names where the fourth character is an ‘a’ are returned.
- LoanDuration=21
  Returns 4, 7. Record where loan duration equals 21. Quotes are not required around numeric values.
- LoanDuration/7>2
  Returns 4, 7. Arithmetic operators can be used within search criteria, in this case the loan duration divided by 7 must be greater than 2 for the record to match.
- Month(JoinDate)>6
  Returns 5, 6, 7, 8. The month function returns a number from 1 to 12. All records where the join date was in the second half of the year. Specialised functions exist for dates and times, including Year, Month, Day, Hour, Minute, Second and also WeekDay. WeekDay returns a number from 1 to 7 representing the day of the week.

- LastName ="Nerk" AND FirstName LIKE "*a*"

  Returns 5, 7. The last name must be 'Nerk' and the first name must contain an 'a'.
- LastName ="Nerk" OR FirstName LIKE "*a*"

  Returns 1, 2, 4, 5, 6, 7, 8, 9, 10. The last name could be 'Nerk' or the first name could contain an 'a' or both could be true.
- NOT(LastName ="Nerk" OR FirstName LIKE "*a*")

  Returns 3. The opposite of the previous search criteria.
- LastName <>"Nerk" AND FirstName NOT(LIKE "*a*")

  Returns 3. Also the opposite of the example two above. Notice that each operator has been reversed rather than reversing the entire expression, however the effect is exactly the same. Note that AND is the reverse of OR and viceversa.
- LastName <=FirstName

  Returns 2, 3, 6, 7, 10. It is possible to compare attributes within each record as part of the search criteria. In this case records where the last name is alphabetically less than or equal to the first name are returned.

GROUP TASK Practical Activity
Use a DBMS to create the borrowers table and enter the data in *Fig 2.64*. Try each of the above queries- if using Access many are simpler to enter in SQL view rather than using the GUI. Check your results with those above.

The following examples focus on the ORDER BY clause. Each example is based on the following SELECT query using our borrower's table – SELECT * causes all attributes to be returned. Simply replace the ORDER BY clause in each case.

```
SELECT *
FROM Borrowers
ORDER BY (list of attributes)
```

The resulting order of the BorrowerID field and some comments accompany each of the examples – in reality all fields are returned:

- ORDER BY LastName, FirstName

  Order is 10, 3, 6, 2, 1, 8, 5, 7, 9, 4. Ascending alphabetical sort on last name. If last name is the same then for each matching group the first name is sorted into ascending alphabetical order. Alphabetical sort is used because the data type of both the LastName and the FirstName fields is text.
- ORDER BY LastName DESC, FirstName DESC

  Order is 4, 9, 7, 5, 8, 1, 2, 6, 3, 10. Records are in the opposite order to the previous example. Descending alphabetical sort on last name, then descending alphabetical sort on first name within matching last names.
- ORDER BY LastName, FirstName DESC

  Order is 10, 3, 2, 6, 7, 5, 8, 1, 9, 4. Ascending alphabetical sort on last name, then descending alphabetical sort on first name within matching last names.
- ORDER BY JoinDate

  Order is 2, 3, 5, 7, 6, 9, 4, 8, 10, 1. Ascending numerical sort on join date. Remember that date fields are really numbers where the integer part represents the number of days, hence sorting numerically arranges dates into chronological order.
- ORDER BY Month(JoinDate), Day(JoinDate)

  Order is 2, 1, 9, 3, 4, 10, 5, 8, 7, 6. Ascending numerical sort on the month number of each join date, and then on the day number of each join date. This example shows how functions can be used within the ORDER BY clause.

- ORDER BY 30-LoanDuration, BorrowerID

  Order is 4, 7, 1, 2, 3, 5, 6, 8, 9, 10. Arithmetic operators can be used within the ORDER BY clause. In our sample data loan duration is either 21 or 14 days, therefore 30-LoanDuration equals 9 or 16 – the 9s sort first. BorrowerID is included simple to ensure a unique order for the result.
- ORDER BY WeekDayName(WeekDay(JoinDate)), BorrowerID

  Order is 5, 2, 4, 8, 1, 6, 3, 10, 7, 9. Ascending alphabetical sort on the weekday name, that is, Sunday, Monday, Tuesday, etc. sorted alphabetically, then sorted on BorrowerID. Clearly this is very difficult to determine by simply inspecting the sample records. The WeekDayName function returns text, hence the sort is performed alphabetically. The WeekDayName function is not standard SQL so it may not work in all DBMSs – however a similar function will be available.

In the last three examples we used arithmetic operators and functions within the ORDER BY clause. These operators and functions can also be used within the SELECT clause. For example the result of the following SQL statement is reproduced in *Fig 2.68*.

```
SELECT BorrowerID, WeekDayName(WeekDay(JoinDate)) As DayName
FROM Borrowers
ORDER BY WeekDayName(WeekDay(JoinDate)), BorrowerID
```

Notice that the result of the function call is given the name DayName - any name could have been used. This ability is particularly useful when creating views (note that currently MS-Access does not support the creation of views using the standard GUI).

GROUP TASK Activity
Try each of the above queries – if using MS-Access it is simpler to use SQL view. Check your results with those above.

GROUP TASK Activity
DBMS views are composed of SQL SELECT queries. Discuss how such views of the data can be used to restrict user's data access.

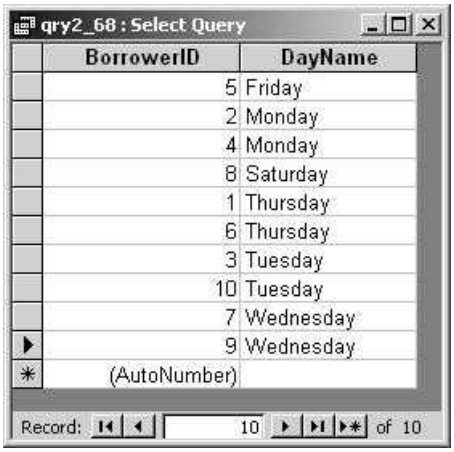


Fig 2.68
Query results in Microsoft Access.

Query by Example (QBE)

QBE (Query by Example)
A visual method for specifying a query. Often search criteria are entered into fields within what appears to be an empty record.

Query by example (QBE) is a visual technique for specifying a database query. Commonly search criteria are entered into what appears to be an empty record. IBM originally developed QBE for use on mainframe computers; it is now included in a variety of different ways within many software applications that are used to search databases. QBE is used to simplify how end-users specify search criteria. The users do not need to understand the details of SQL. For instance, a modified QBE system is the basis of most advanced search screens used by Internet search engines. No doubt most users of search engines are completely unaware that SQL even exists.

In its original and simplest form, QBE displays what looks like an empty record and the user enters example data or conditions into each field. The query engine then creates the corresponding SQL statements, performs the query using the current records and displays the results. In Microsoft Access this functionality is known as filtering. *Fig 2.69* shows this QBE facility being used to filter records from the Borrowers table in our library database where the LastName field equals nerk. Notice the "Or" tab at the bottom of each screen, which allows more complex criteria to be specified.

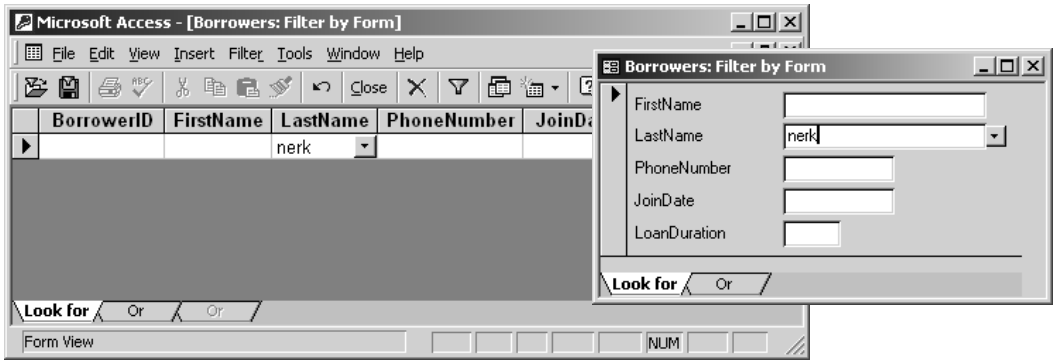


Fig 2.69
In Microsoft Access simple QBE is implemented using filters.

Sophisticated "QBE like" utilities are also available for use with most modern DBMSs. These visual utilities are either add-ons or are an integrated part of the DBMS software. They all aim to simplify the design of complex queries. The query design grid included within Microsoft Access is one example (refer *Fig 2.65*). As we shall see in the next section the Access query design grid greatly simplifies the design of complex SQL queries that include multiple tables and relationships. Although such "QBE like" facilities are an excellent aid when commencing the design of a new query, it is often necessary (or simpler) to edit the underlying SQL to meet unusual requirements.

GROUP TASK Practical Activity
Use the borrowers table from above to trial the QBE features available with the DBMS you are using. If using Microsoft Access first create a simple form based on the Borrowers table and then use the filter functionality to select records that match different criteria.

Searching and Sorting Multiple Tables

The real advantages of relational databases are not realised until the related data from multiple tables can be retrieved. For this to occur, each component table and the relationships that join the tables need to be specified.

At first glance it seems reasonable that simply listing each of the table names within the FROM clause will enable us to correctly retrieve data. Unfortunately this does not produce the desired result. In this section our main task is to understand how the joins (relationships) between tables are specified in SQL. To assist our discussion let us return to our invoicing relational database – the same one we normalised earlier in this chapter. We will use the same sample data (with two new customers) as we used earlier, however to save flicking back through the text the final schema together with the sample contents of each table are reproduced on the next page in *Fig 2.70*.

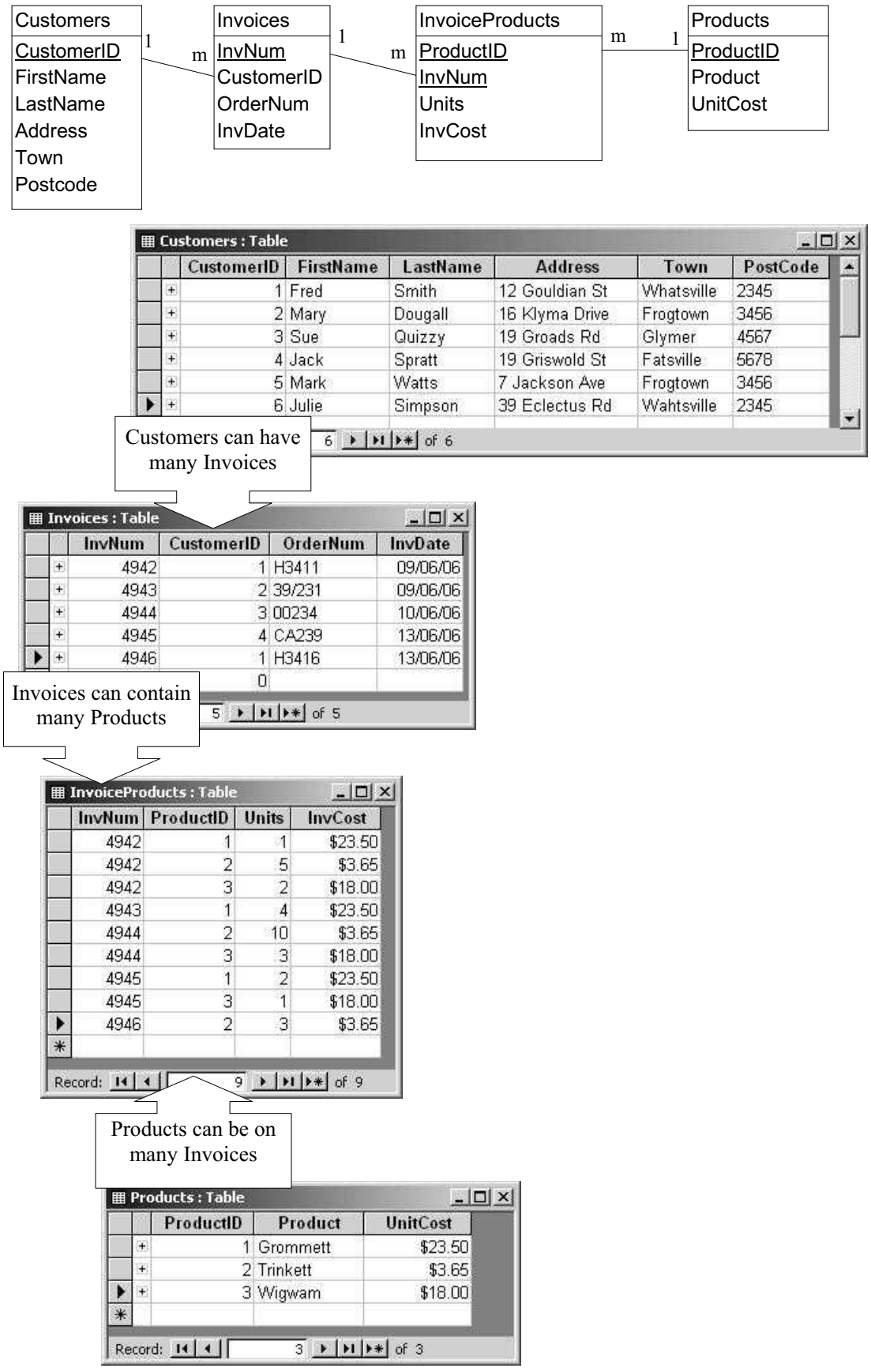


Fig 2.70
Final schema and sample data in the invoicing relational database.

Firstly let us examine what happens if we simply include multiple table names in the FROM clause of a SELECT query. In our invoice sample, say we wish to generate a list of all customers with each of their related invoice numbers; we could try:

```
SELECT Customers.LastName, Customers.FirstName,
Invoices.InvNum
FROM Customers, Invoices
```

This returns the results shown in *Fig 2.71* (in no particular order). Clearly this is not what is required – there is a row for every customer to invoice combination. As there are 6 customer records and 5 invoice records the DBMS logically links all customers and all invoices, so 6×5=30 rows are returned. In essence we have not specified that our primary and foreign keys must match.

A sensible second attempt is to include the search criteria specifying that the primary key field must match the foreign key field. Our revised query is now:

```
SELECT Customers.LastName, Customers.FirstName,
Invoices.InvNum
FROM Customers, Invoices
WHERE Customers.CustomerID=Invoices.CustomerID
```

The results from this query are shown in *Fig 2.72*. This time 5 rows are returned – 1 row for each invoice, which is probably what we actually need. Notice that Fred Smith appears twice as he has two invoices within the database, however Mark Watts and Julie Simpson do not appear, as they have no invoices.

Until 1992 this was precisely how the SQL standard specified such queries should be written and currently (2007) this is all that is required in an HSC response. In 1992 a new SQL standard was released; this standard specified how joins should be written as part of the FROM clause. This makes more sense as relationships (or joins) are not search criteria at all, rather they are structural parts of a database.

Qry2_71 : Select Query

LastName	FirstName	InvNum
Smith	Fred	4942
Smith	Fred	4943
Smith	Fred	4944
Smith	Fred	4945
Smith	Fred	4946
Dougall	Mary	4942
Dougall	Mary	4943
Dougall	Mary	4944
Dougall	Mary	4945
Dougall	Mary	4946
Quizzy	Sue	4942
Quizzy	Sue	4943
Quizzy	Sue	4944
Quizzy	Sue	4945
Quizzy	Sue	4946
Spratt	Jack	4942
Spratt	Jack	4943
Spratt	Jack	4944
Spratt	Jack	4945
Spratt	Jack	4946
Watts	Mark	4942
Watts	Mark	4943
Watts	Mark	4944
Watts	Mark	4945
Watts	Mark	4946
Simpson	Julie	4942
Simpson	Julie	4943
Simpson	Julie	4944
Simpson	Julie	4945
Simpson	Julie	4946

Record: 30 of 30

Fig 2.71
Results of multiple table query with no join specification.

Qry2_72 : Select Query

LastName	FirstName	InvNum
Smith	Fred	4942
Smith	Fred	4946
Dougall	Mary	4943
Quizzy	Sue	4944
Spratt	Jack	4945

Record: 5 of 5

Fig 2.72
Results when join specified in WHERE clause.

The identical rows shown in *Fig 2.72* are returned from the following SQL statement:

```
SELECT Customers.LastName, Customers.FirstName, Invoices.InvNum
FROM Customers INNER JOIN Invoices ON Customers.CustomerID = Invoices.CustomerID
```

GROUP TASK Practical Activity
Try each of the three SQL statements on this page using a DBMS. (Assuming you have previously constructed the Invoicing database, don't forget to add the 2 new customers shown in *Fig 2.70*).

The SQL keywords INNER JOIN indicate that only the records that match on each side of the join should be retrieved. In particular INNER means include only those records that form part of a match between the specified fields; records that don't match a record on the other side are not included at all. The order in which the tables appear on either side of the INNER JOIN keywords is irrelevant, however this order is significant for outer joins.

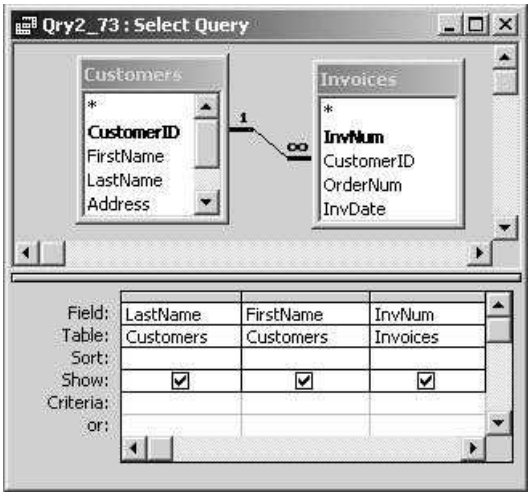


Fig 2.73
Query design grid in MS-Access

Most DBMSs include some form of graphical user interface to simplify the construction of SQL joins. These tools automatically create joins based on the relationships defined within the database. *Fig 2.73* shows the previous query as it is represented on the MS-Access query design GUI. Often it is easier to create the joins using the GUI and then modify the SQL to suit individual requirements – such as adding additional search criteria or sorts.

Outer joins are used when we wish to include all records on one or both sides of a join. A left outer join includes all records from the table on the left hand side of the join – in SQL the keywords LEFT OUTER JOIN or simply LEFT JOIN are used. A right outer join includes all the records on the right hand side of the join – the SQL keywords RIGHT OUTER JOIN or RIGHT JOIN are used.

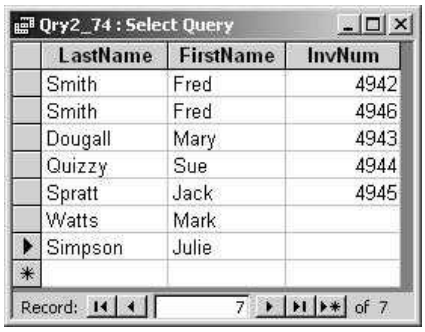


Fig 2.74
Customers LEFT JOIN Invoices example query results.

Full outer joins include all records on both sides and use the SQL keywords FULL JOIN. In reality full joins are rarely if ever used when the database has been normalised – the SQL standard includes full outer joins however they are not currently supported in MS-Access or mySQL. Larger commercial DBMS server software does support full joins, for example SQL server and Oracle. It is possible to simulate full joins using a combination of inner and outer joins and then combining the results using the UNION keyword.

Let us consider some examples using our sample invoices databases:

- SELECT Customers.LastName, Customers.FirstName, Invoices.InvNum
  FROM Customers LEFT JOIN Invoices ON Customers.CustomerID = Invoices.CustomerID
  This SQL statement will include all customers and only the invoices that match. Note that in our example database all invoices must match a customer but customers who have no related invoices are possible. The result is shown in *Fig 2.74*.
- SELECT Customers.LastName, Customers.FirstName, Invoices.InvNum
  FROM Invoices RIGHT JOIN Customers ON Customers.CustomerID = Invoices.CustomerID
  This example is equivalent to the previous example – note that the table names have been reversed together with the join type.

- SELECT Customers.LastName, Customers.FirstName, Invoices.InvNum
  FROM Customers LEFT JOIN Invoices ON Customers.CustomerID = Invoices.CustomerID
  WHERE Invoices.CustomerID is NULL

  This example is similar to the previous examples with the addition of a search criteria. In this case only customers that do not have a related invoice are returned – in this example, just Mark Watts and Julie Simpson. In effect the query searches for all customers who have never placed an order.

We now consider queries that include aggregate functions. Imagine in our invoicing database we wish to determine the number of each product that has been sold. Let us progressively create this query.

- SELECT Products.Product, InvoiceProducts.Units
  FROM Products INNER JOIN InvoiceProducts ON
  Products.ProductID = InvoiceProducts.ProductID
  ORDER BY Products.Product;

  The records retrieved are shown in *Fig 2.75*. Clearly we need to sum the Units column for each product.

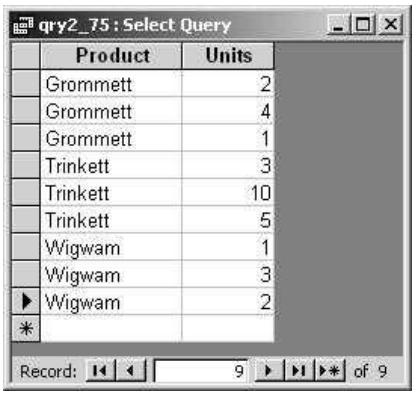


Fig 2.75
Calculating number of each product sold – first attempt.

- SELECT Products.Product, Sum(InvoiceProducts.Units) AS TotalUnitsSold
  FROM Products INNER JOIN InvoiceProducts ON Products.ProductID = InvoiceProducts.ProductID
  GROUP BY Products.Product
  ORDER BY Products.Product

  The records retrieved are shown in *Fig 2.76*. The Sum function is used to add up all the field values within each group specified by the GROUP BY clause. The resulting TotalUnitsSold column is known as a calculated or derived field – such calculated data is not stored.

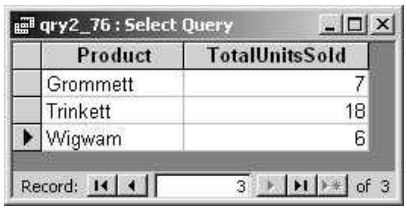


Fig 2.76
Calculating number of each product sold.

To return the number of invoices containing each product simply replace the Sum function with the Count function – the Count function returns the number of records in each group. Similarly the average (Avg) function returns the average of the field values in each group. There are many other functions available – some of the more common ones include Min, Max, First and Last.

Consider the following (extension work):

In real world systems it is common to include multiple joins and also other queries (subqueries) within a single query. It is unlikely you would be asked such questions in the HSC, however if your project includes a relational database then it is likely you will need to construct such complex queries at some stage.

Let us construct a query to determine the most valuable customers within our sample invoicing database. For our purpose we shall determine the value of a customer based on the total cost of all their invoices. In reality this is not a fair indicator – new customers will be distinctly disadvantaged.

In general terms, we need to calculate the total cost of all invoices for each customer with the results sorted in descending total cost order. The first step is to calculate the cost of each product on each invoice for each customer. The following query accomplishes this task:

```
SELECT Invoices.CustomerID, Sum([InvCost]*[Units]) AS TotalCost
FROM Invoices INNER JOIN InvoiceProducts ON Invoices.InvNum = InvoiceProducts.InvNum
GROUP BY Invoices.CustomerID
```

Notice in this query we have not included the customer table and we have grouped on CustomerID. It may be tempting to include the customer names and group on these names; this is not a good idea as it is possible for many customers to have the same name whilst CustomerID is guaranteed to be unique.

Consider the expression Sum([InvCost]*[Units]) AS TotalCost. This expression first multiplies the invoice cost (InvCost) by the number of units ordered (Units) for every record in the table InvoiceProducts – in effect we are calculating the cost of each invoice's product. As we are grouping on CustomerID then the Sum function causes these product costs to be added together for each different CustomerID. Finally we name this new column TotalCost. The results, using our sample data, are shown in *Fig 2.77* – note that we save this query with the name qryCusTotCost.

qryCusTotCost : Select Query

CustomerID	TotalCost
1	$88.70
2	$94.00
3	$90.50
4	$65.00

Record: 4 of 4

Fig 2.77
Results from qryCusTotCost

We now need to sort these results into descending order and include each customer's name. Rather than modify our query we'll create a new query that includes a join between qryCustomerTotalCost and the Customers table:

```
SELECT [FirstName] & " " & [LastName] AS CustomerName, qryCusTotCost.TotalCost
FROM Customers INNER JOIN qryCusTotCost ON Customers.CustomerID = qryCusTotCost.CustomerID
ORDER BY qryCusTotCost.TotalCost DESC
```

The above query returns the desired results (see *Fig 2.78*). Notice we have combined (concatenated) the first and last name of each customer together with a space between.

qryMostValuedCus : Select Query

CustomerName	TotalCost
Mary Dougall	$94.00
Sue Quizzy	$90.50
Fred Smith	$88.70
Jack Spratt	$65.00

Record: 4 of 4

Fig 2.78
Final most valuable customer results

Exactly the same result can be obtained in many different ways, including using a single query. This query returns exactly the same results as the previous one:

```
SELECT [FirstName] & " " & [LastName] AS CustomerName, Sum([InvCost]*[Units]) AS TotalCost
FROM (Customers INNER JOIN Invoices ON Customers.CustomerID = Invoices.CustomerID)
        INNER JOIN InvoiceProducts ON Invoices.InvNum = InvoiceProducts.InvNum
GROUP BY Customers.CustomerID, [FirstName] & " " & [LastName]
ORDER BY Sum([InvCost]*[Units]) DESC;
```

Notice that brackets have been used to indicate the order in which the joins should be made. In our example, and in most situations, the join order is not significant. For example our FROM clause above could have been:

```
FROM Customers INNER JOIN (Invoices INNER JOIN InvoiceProducts ON Invoices.InvNum =
        InvoiceProducts.InvNum) ON Customers.CustomerID = Invoices.CustomerID
```

In the original the join between the Customers and InvoiceProducts tables is performed first, whereas in the second version the join between the Invoices and InvoiceProducts table is performed first.

GROUP TASK Practical Activity
Reproduce each of the above SQL statements using a DBMS.
(Assuming you have previously constructed the Invoicing database).

HSC style question

Mia is working on a personal address book system to store details of her personal contacts. The data dictionary for her *Contacts* table is reproduced below:

Field name	*Field type*	*Description*
ContactID	Integer	Primary key
LastName	Text	Last name of contact
FirstName	Text	First name of contact
DOB	Date	Date of birth
StreetAddress	Text	Street address e.g. 110 Harold Avenue
Town	Text	Town or suburb
Postcode	Text	Postcode e.g. 2066

(a) Mia wishes to create a list in alphabetical order by last name of all her contacts that have a birthday during May. The list should include first name, last name and date of birth. Construct an SQL query to retrieve this information.

(b) Mia has created a separate table for phone numbers; some sample data within this *PhoneNumbers* table is reproduced below:

PhoneID	*ContactID*	*NumberType*	*PhoneNumber*
3455	2455	Mobile	0455 678 906
3456	1034	Mobile	0434 123 456
3457	2455	Home	02 9657 1234
3459	2455	Work	02 8899 0033
3460	3115	Mobile	0422 345 678
3461	3115	Home	02 9456 7890
3463	1066	Home	02 9543 4321

(i) Explain why Mia has created a separate table to store phone numbers.

(ii) Construct an SQL query to return a list of first and last names for all the contact's that have a mobile phone number within Mia's database.

Suggested Solution

(a)
```
SELECT FirstName, LastName, DOB
FROM Contacts
WHERE Month(DOB)=5
ORDER BY LastName
```

(b) (i) Using a separate phone number table means each contact can have any number of phone numbers stored. For example a contact could have a mobile, home, work, holiday, second mobile, or any other type of phone number. If the phone numbers were in the Contacts table then a new field for each number would need to be present – even if only one contact had this type of number. Furthermore it is easier to create queries that retrieve all of a contacts phone numbers.

(ii)
```
SELECT FirstName, LastName
FROM Contacts INNER JOIN PhoneNumbers ON Contacts.ContactID=PhoneNumbers.ContactID
WHERE NumberType="Mobile"
```

SET 2G

1. Sorting 3, 31, 26, 4, 2 into descending alphabetical order results in:
   (A) 2, 3, 4, 26, 31.
   (B) 2, 26, 3, 31, 4.
   (C) 31, 26, 4, 3, 2.
   (D) 4, 31, 3, 26, 2.
2. Sorting Bull, Cow, Cat, Ant, Car, Bike into ascending alphabetical order results in:
   (A) Ant, Bike, Bull, Cow, Car, Cat.
   (B) Cow, Cat, Car, Bull, Bike, Ant.
   (C) Ant, Bull, Bike, Cow, Car, Cat.
   (D) Ant, Bike, Bull, Car, Cat, Cow.
3. Which list contains only logical operators?
   (A) <, >=<, =, like.
   (B) AND, OR, NOT.
   (C) LIKE, AND, OR.
   (D) CONTAINS, EQUALS, NOT, LIKE.
4. SELECT Car FROM Vehicles ORDER BY Make will display:
   (A) The Car attribute in Make order.
   (B) The Car and Make attributes in Make order.
   (C) The Make attribute in Car order.
   (D) Car attributes that include data in the Make attribute.
5. SELECT * FROM Manufacturers will return:
   (A) All records in all tables.
   (B) All attribute names in the Manufacturers table.
   (C) All fields and records in the Manufacturers table.
   (D) The primary key for each record in the Manufacturers table.
6. With regard to indexes within databases, which of the following is true?
   (A) Records are stored in index order.
   (B) The data must always be retrieved in index order.
   (C) Indexes are always updated as soon as an indexed field is updated.
   (D) Indexes detail the sort order without actually sorting the records.
7. Which expression selects all records where the Suburb field ends in 'town'?
   (A) Suburb LIKE "*town*"
   (B) Suburb LIKE "town*"
   (C) Suburb LIKE "?town"
   (D) Suburb LIKE "*town"
8. Table A contains 5 records and table B contains 15 records. A one to many relationship exists from table A to table B. The SQL SELECT * FROM A,B will return:
   (A) 15 records.
   (B) 5 records.
   (C) 20 records.
   (D) 75 records.
9. Table A contains 5 records and table B contains 15 records. An enforced one to many relationship exists from table A to table B. What is the maximum number of records returned by an outer join?
   (A) 5 records.
   (B) 15 records.
   (C) 19 records.
   (D) 20 records.
10. A table Alphabet contains a single attribute called Letter. Each of the 26 records holds a different letter. The table Alphabet2 is the same as the Alphabet table except one record is missing. Which SELECT statement returns the missing letter?
    (A)
    ```
    SELECT Alphabet.Letter
    FROM Alphabet INNER JOIN Alphabet2 ON
    Alphabet.Letter = Alphabet2.Letter
    WHERE Alphabet2.Letter Is Null
    ```
    (B)
    ```
    SELECT Alphabet.Letter
    FROM Alphabet RIGHT JOIN Alphabet2
    WHERE Alphabet2.Letter Is Null
    ```
    (C)
    ```
    SELECT Alphabet.Letter
    FROM Alphabet LEFT JOIN Alphabet2 ON
    Alphabet.Letter = Alphabet2.Letter
    WHERE Alphabet2.Letter Is Null
    ```
    (D)
    ```
    SELECT Alphabet.Letter
    FROM Alphabet, Alphabet2
    WHERE Alphabet2.Letter Is Null
    ```
11. Explain the purpose of each of the following SQL keywords.
    (a) SELECT (b) FROM (c) WHERE (d) ORDER BY (e) INNER JOIN
12. Consider the invoicing relational database schema (*Fig 2.69*). Construct SQL statements to perform each of the following.
    (a) Return all customers who live in NSW sorted on their last name.
    (b) Return the number of different products for each unique order number.
    (c) Return the date of each customer's most recent invoice.
    (d) Return the data required to construct the original invoice in *Fig 2.22* (page 139). Some totals on the *Fig 2.22* example just don't add up, can you explain why?
13. Explain why indexes are created automatically for primary keys in most RDBMS systems.
14. With the aid of examples, explain the difference between inner and outer joins.
15. Consider Mia's personal address book system from the HSC Style Question on the previous page. Construct an SQL statement to return the names of all her contacts that do not have a mobile phone number within Mia's database.

CENTRALISED AND DISTRIBUTED DATABASES

It is rare today for a large organisation to store all its data within a single centralised database at a single location. In reality they maintain many databases in many different remote locations on many different computers. The data at all remote sites is connected in such a way that it appears to users as one large single database. Such databases are called distributed databases and the software that controls the sharing of data is called a Distributed Database Management System (DDBMS).

Centralised Database
A single database under the control of a single DBMS. All users and client applications connect directly to the DBMS.

Distributed Database
A set of connected databases stored on multiple computers that appears to users as a single database.

The overriding aim of all distributed database systems is to ensure correct and complete data within all databases is available to all sites at all times – by no means a simple task! A centralised database does not need to be concerned with such issues as the whole database is under the control of a single DBMS and every user accesses data via this DBMS. This works well on a single site, however performance problems begin to emerge as user numbers increase. Even the largest and most powerful server machines are unable to cope with the demands of many thousands of concurrent users. For instance consider the number of users accessing bank databases, popular online sites such as hotmail or government department databases. The problem is further exacerbated when many users are connecting remotely. Communication lines must be able to cope with the demand without response times slowing or the lines failing.

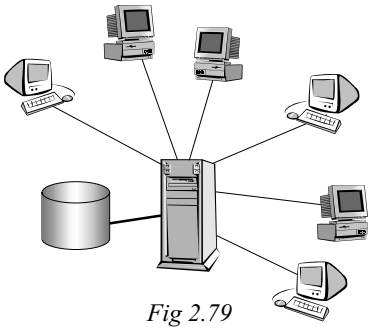


Fig 2.79
Centralised databases are accessed through a single DBMS server

An early solution to the problems encountered with centralised databases was to create distributed systems where one database was designated as the central database server. Each of the other database servers would regularly upload their changes to the central server. It was either left up to the central server to deal with any data integrity issues or the system was designed such that particular records could only be stored within one of the smaller databases. There are still many organisations that maintain such systems. Today there are various sophisticated techniques that mean distributed databases can be implemented such that users have access and can make alterations to any data on any of the individual distributed databases. These solutions aim to make the entire distributed database appear to be a single centralised database from the user's perspective.

Consider *Fig 2.80* on the next page, this diagram describes a typical distributed database system. This system includes four databases where each database is accessed via a local DDBMS server. Each of these servers provide the usual DBMS services to their local clients. In addition each server communicates with all other servers to share data. The structure of each database together with the sharing strategy determines how this data sharing occurs. Individual users, and the client software applications they run, are unaware of the distributed database – they operate as if the complete database exists at their site.

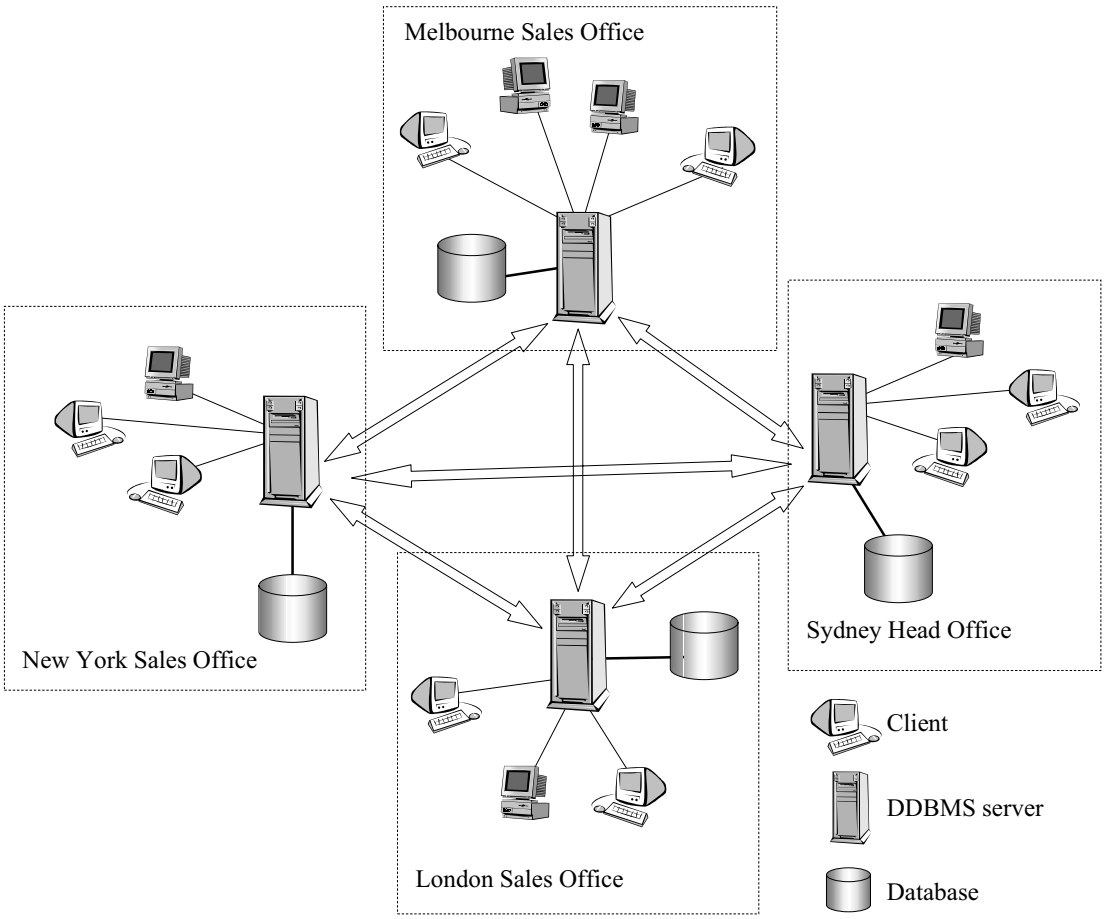


Fig 2.80
Major components within a typical distributed database system.

Types of Distributed Database Systems

There are many different types of distributed database systems – indeed almost every implementation is different. The choice depends on the nature of the data and the individual system's requirements. Let us consider three common structures and strategies - be aware that in reality a combination of these techniques is used.

- **Fragmentation**

Different parts of the database are stored at different locations. Using this system individual data items are physically stored once only at one single location. To execute a query that includes data from a remote location always requires the data to be physically retrieved from the remote server. For this to work correctly requires a fast and reliable connection between all DDBMS servers.

Horizontal fragmentation stores different records of the same table at different locations. Consider the system described in *Fig 2.80* and *Fig 2.81*; each sales office database contains a sales table containing records of sales made in that office. Say Fred at head office in Sydney executes a query to calculate total sales for the entire company. The Sydney head office DDBMS server examines Fred's query and splits it into three sub-queries, one for each sale's office. Each sub-query is the same – an SQL statement to calculate the total sales. The head office DDBMS is acting as a client to each of the three remote sale's servers. It must wait for the results of each query to be returned. Once all results arrive the head office DDBMS compiles the results – in this example it simply adds up the three totals – and sends this result to Fred's client application.

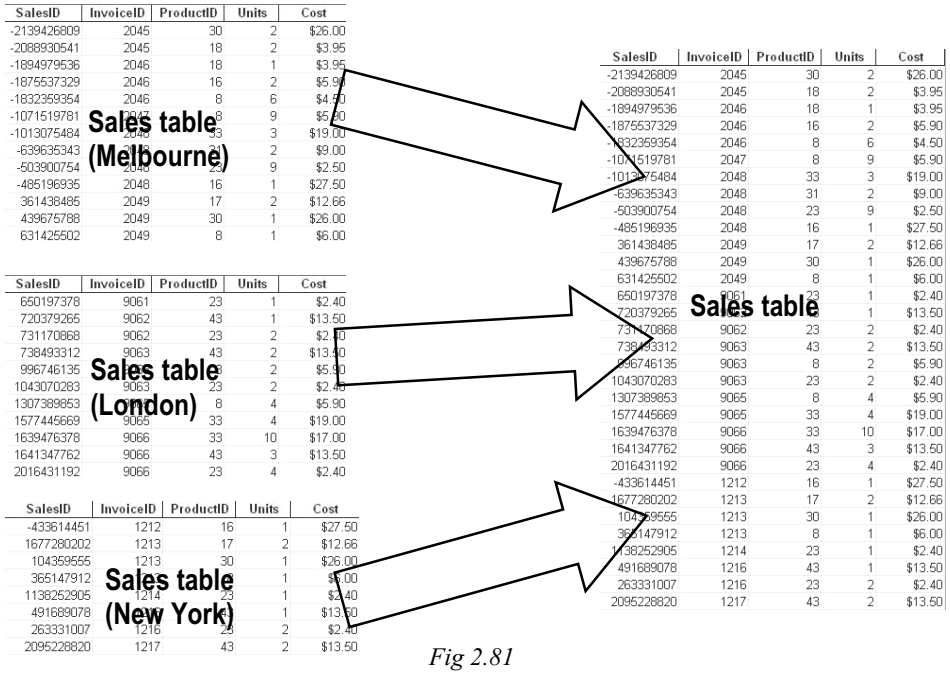


Fig 2.81
Horizontal fragmentation stores records in different locations.

Vertical fragmentation stores different attributes at different locations. Consider *Fig 2.80* and *Fig 2.82*; now imagine the payroll system for the company is located at head office in Sydney. The payroll data is structured such that the Employees table contains a number of attributes (fields) used solely for payroll data – say annual salary, bank account details and sick days remaining. The remaining employee attributes, such as names, addresses and phone numbers are stored in the Employees table at each individual office. Note that each individual data item is stored once within the distributed database. Furthermore the primary key is stored at all locations.

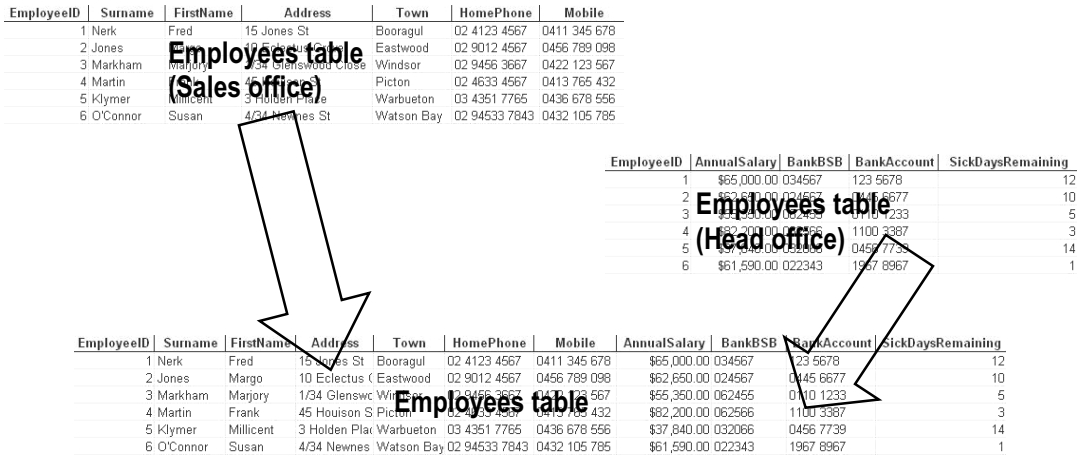


Fig 2.82
Vertical fragmentation stores attributes in different locations

Say Madge is the sales manager at the company's New York office. She wishes to know how many remaining sick days each of her sales staff has. The client software on Madge's computer creates a SELECT query based entirely on the Employees table, For example:

```
SELECT Employees.Name, Employees.SickDaysRemaining
FROM Employees
WHERE Location="New York"
```

This query is sent to Madge's local DDBMS server in New York. This server examines the SQL and realises it doesn't have the SickDaysRemaining field in its local database, rather the field is held at head office. The New York DDBMS splits the query into two sub-queries.

```
SELECT Name, EmployeeID                SELECT SickDaysRemaining, EmployeeID
FROM Employees                         FROM Employees
WHERE Location="New York"              WHERE EmployeeID IN (list of EmployeeIDs from first query)
```

The first sub-query gets all the employee names (and primary keys) from the local database. The second sub-query gets the SickDaysRemaining based on the primary keys found by the first sub-query. The second sub-query is sent to the head office DDBMS, where it is executed. The results are returned to the New York server where they are combined with the employee names and sent on to Madge's computer. In this example the New York server acts as a client to the head office server. In essence all DDBMS servers can also behave as clients.

- **Downloading**

In this type of distributed database each server downloads copies of data as it is required from remote databases and stores the data within its local database. This system is suited to data that rarely changes yet fast data access times are required.

The DNS (Domain Name Service) system uses this 'downloading' strategy. The DNS system is composed of thousands of DNS servers – each server is linked to a database of domain names and related IP addresses. When a web server performs a DNS lookup it is simply requesting the IP address for a specific domain name. If the DNS server cannot find the domain name within its local database then it performs a DNS lookup on another DNS server. If this DNS server doesn't know the domain name it performs yet another DNS lookup at another DNS server. This sequence continues until a DNS server can complete the request. When this occurs the found IP address filters back through each DNS server in the chain – each DNS server stores the IP address for the domain name in its local database and passes the result onto the requesting server. Eventually the web server receives the IP address. As a consequence commonly accessed domains and their related IP addresses are held within most DNS databases. The IP addresses of new domains propagate through the DNS system at a rate proportional to the number of users wishing to access the domain's web site.

Let us consider a possible 'downloading' strategy within our *Fig 2.80* system. Say each sales office database contains its own customer's table. This table holds the name, address, phone number and so on for all of the local office's customers. Now imagine a salesman gets an order from a new customer say Mitch – is Mitch really a new customer or is he an existing customer who has previously ordered through another office? This question needs to be resolved. If a downloading strategy is used the following would occur. First the local DDBMS checks its own database – the customer Mitch is not found here. The DDBMS then sends a request (again an SQL SELECT query) to all the remote

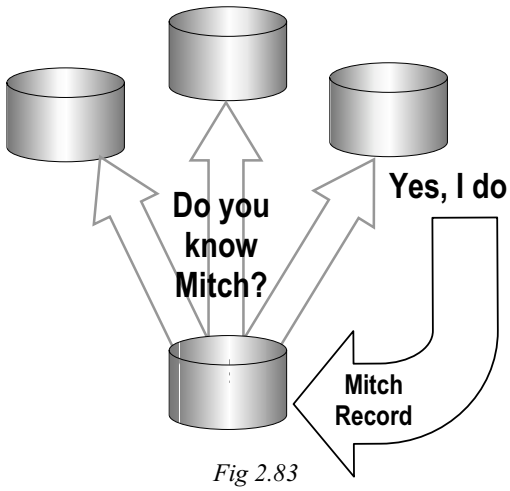


Fig 2.83
Downloading involves checking if new records already exist at remote locations.

DDBMS servers (see *Fig 2.83*). If one of these remote servers returns a matching record then the customer Mitch exists elsewhere and the returned record is stored to the local database. If no record is returned from any remote servers then the new Mitch customer record is created in the local database.

- **Replication**

In this type of distributed database the aim is for all local databases to hold copies of all the data all the time – in reality each local database holds MOST of the data MOST of the time! One database is designated as the master and all other databases are known as replicants. Each replicant is synchronised with the master at regular intervals. The synchronising process copies all altered records and new records in both directions. That is, the replicant receives new and altered records from the master and the master receives new and altered records from the replicant – this process is known as replication. Over time each replicant receives all the changes and additions made in the master and in all other replicants. The time interval between replication events determines the accuracy of all the copies. In most systems replication takes place each night, however the timing is adjusted to suit the needs of the individual system.

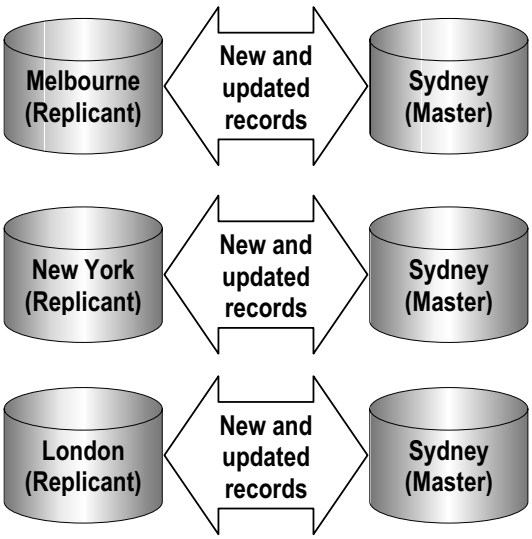


Fig 2.84
Each replicant is synchronised in turn with the master.

Replication is suited to database systems where the same records are rarely altered or added at similar times but at different sites. Furthermore replication does not rely on a stable connection between remote servers. For example replication commonly occurs over a standard broadband Internet connection.

Imagine the system in *Fig 2.80* uses replication. Say the head office database is the master and each of the sales office databases are replicants. Now say Bob, a lowly salesman in Melbourne, notices that the name of their best selling 'Wooble' product is misspelt as 'Wouble', Bob makes the change. That is, the Name attribute for the 'Wooble' product within the Products table is updated within the local Melbourne office's database. Now suppose Madge and her New York sales team pronounce 'Wooble' as 'Wowble' – different to Aussie Bob, so Madge updates the spelling in her local New York database to 'Wowble'. We now have the same record updated differently in two different locations with two different 'Wooble' spellings. This is called an update conflict. How can replication resolve this conflict? There are various strategies for resolving update conflicts and most systems will use different strategies based on the particular table where the conflict occurs.

The simplest strategy for resolving update conflicts is to simply use the most recently updated version. This strategy requires a time stamp to be stored for every change made. In our example if Bob made his change after Madge then Bob's change would ultimately appear in all copies of the database.

Another popular strategy is to prioritise sites and/or users. In our example the head office database may have priority over the sales offices. This would not resolve the conflict. However Madge is a sales manager and Bob a lowly salesman, hence Madge

would likely have priority over Bob. In this case Madge's 'Wowble' spelling will, after replication, appear in all copies of the database. But what if Bob was also a sales manager with equal priority to Madge? In this case the conflict cannot be resolved automatically. Therefore the conflict is logged for manual resolution – in large systems a replication manager is appointed whose main task is resolving update conflicts.

GROUP TASK Activity
Microsoft Access includes the ability to perform replication. If you have access to Access, examine the help system and investigate the tools and strategies used for replication within Access.

GROUP TASK Discussion
For most large commercial and government organisations fragmentation is the most commonly used technique. Why do you think this is? Discuss.

Consider the following:

Currently 'Grids' are all the rage. A 'Grid' not only shares data across multiple databases but it also shares processing and server resources. Grids aim to maximise the use of data and processing resources. Large companies have data processing facilities all over the world. As night falls on different continents, so too does the work load at each of these facilities. Grid computing or 'The Grid' enables these idle resources to be automatically shared on a global basis. The night time example is an extreme case – the processing load would be constantly changing to balance the work load between all machines within the grid.

GROUP TASK Research
Research, using the Internet or otherwise, an example of a large organisation that uses Grid technology.

TOOLS FOR HYPERMEDIA SEARCHING AND RETRIEVAL

Hypermedia (and hypertext), as we learnt earlier in this chapter, is by its very nature unstructured. There is no real predefined organisation, all we know is that nodes are connected to each other via links. The term 'web' is a somewhat apt description of hypermedia – links go all over the place, connecting nodes in all directions, much like a spider's web. This lack of defined structure makes searching for and then locating particular information a difficult task. In this section we concentrate on the operation of search engines – the primary tool used to search the World Wide Web (WWW).

Search Engine
A program that builds an index of website content. Users can search the indexed content to locate relevant website content.

Many of the technologies used within search engines are also used to search other types of hypermedia. For instance natural language searching allows search criteria to be entered in normal spoken English – usually questions. Many help systems within software applications include such capabilities. For example, Microsoft Word includes an answer wizard within its help system; questions are entered just like spoken English and the system responds with links to possible relevant topics.

Operation of Search Engines

Search engines enable web users to locate web pages about specified topics. The user enters a search criteria (which may be in the form of natural language) and the search engine returns a list of matches. Traditionally there are two distinct techniques used to locate web pages; the first is used by web directories and the second is used by search engines.

Web directories are created by humans based on submissions from users. Commonly the creator of a new website submits details of their site to the web directory, the details are checked by an employee who then adds the site to the appropriate directory. Search engines use automated software tools called search robots. These robots crawl the web and automatically locate and index individual web pages. Some search engines examine keywords specified within HTML metatags, whilst most determine their own keywords as they examine the content on each page.

Today most leading sites that include web directories also include search engines (for example Yahoo); hence the term 'search engine' encompasses both types. In any case both types of search engine must maintain an enormous store of data describing millions of web pages.

There is one further type of search engine, sometimes called a 'meta-search engine', These sites submit the search criteria entered to other search engines and simply compile the results - MetaCrawler is one example of such a site.

GROUP TASK Discussion
Metadata is data that describes or defines other data. Describe similarities between metadata and meta-search engines.

GROUP TASK Practical Activity
Visit a number of popular search engines and determine whether they include search engines, web directories and/or meta-search engines.

We now restrict our discussion to the operation of search engines. That is, search engines that crawl the web to compile and rank their indexed content. The general processes performed by search engines include:

- Crawling the web to locate and retrieve web pages.
- Indexing and ranking each web page found.
- Analysing search criteria entered by users.
- Retrieving suitably ranked web page results.

The context diagram in *Fig 2.85* describes the data flowing into and out of a typical search engine. The World Wide Web provides web pages, millions or even billions of them, to the system. The system is continually processing web pages 24 hours a day 7 days a week. Users enter search criteria and the system generates and displays ranked results specific to their search criteria. This is a fairly trivial view of the system's operation. Let us expand this context diagram into a more detailed level 1 data flow diagram (see *Fig 2.86* below).

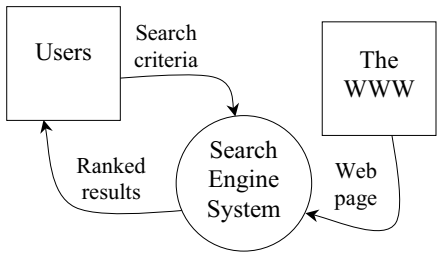


Fig 2.85
Context diagram for a search engine.

In *Fig 2.86* we clearly see the two primary processes performed by search engines. They create the searchable databases and they process user searches. Let us consider

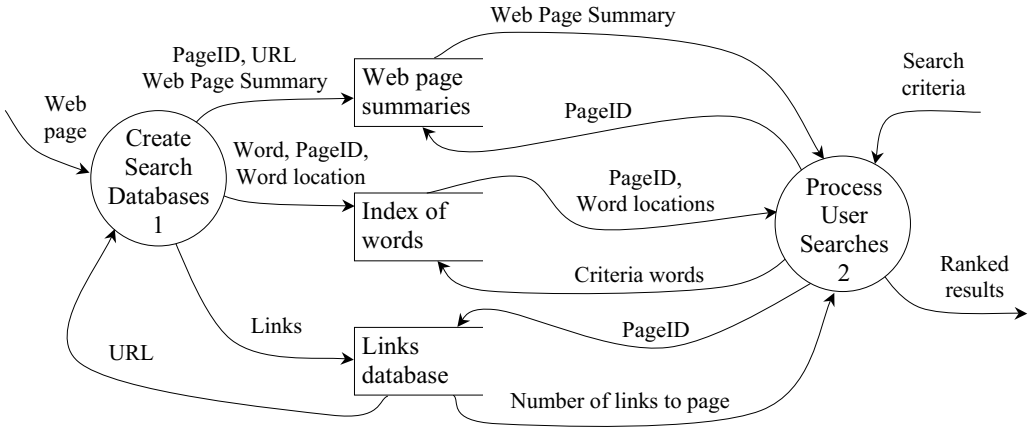


Fig 2.86
Search engine Level 1 DFD.

the general operations occurring during each of these processes. In reality the process is far more complex than we can hope to describe and furthermore large commercial search engine companies closely guard the technical details of their operation. In fact, it is the sophistication of these operations that distinguish between different search engines.

- **Create Search Databases**

Most search engines use hundreds or even thousands of small computers (often simple personal computers) to crawl the web. These computers run software known as search robots, spiders, web crawlers or even simply bots - Google calls their search robot Googlebot. These search robots retrieve web pages from web servers, extract any hyperlinks to other web pages and pass the complete web page onto the indexer. Thousands of these individual search robots work continuously 24 hours a day; essentially a distributed network of computers all operating in parallel to complete one task – indexing the World Wide Web.

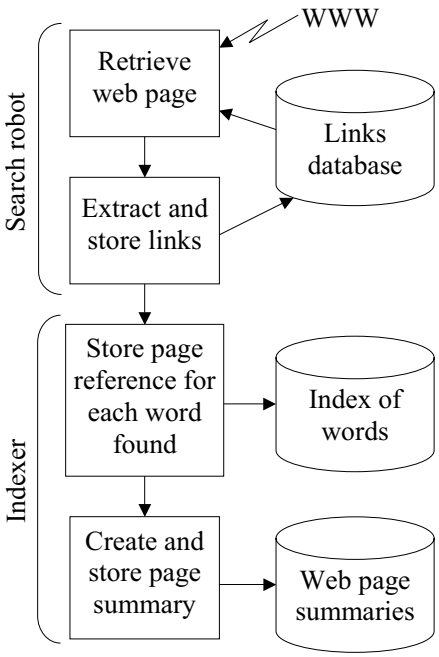


Fig 2.87
Flowchart describing the indexing of a single web page.

In the past many search robots used to follow each link they found in turn from each parent web page. This is the origin of the name spider and crawler – they followed their nose through the web based on where the web led them. Today this strategy is rarely used; rather the hyperlinks found on each page are stored in a links database – essentially an enormous list of URLs.

The URLs within the links database is not a unique list – some URLs will appear once whilst others will appear many thousands of times. The number of times a URL appears indicates the number of times that page has been linked to from other pages the robot has examined.

The links database is continuously being compared to the web page summaries database. If a URL is found that does not have a corresponding web page summary then that URL is added towards the front of the queue for the search robots to retrieve. All URLs are eventually sent again to the search robots – this ensures the web summaries remain reasonably up to date. Those URLs that appear more often are sent to the robots more often. The theory being that these URLs are probably more popular and hence it is important they remain as up to date as is possible.

Let us return to the indexer (refer *Fig 2.87*). The job of the indexer is to create an index of words upon which user searches are based. This index is just like the index in the back of a book, it contains a list of words in alphabetical order together with a reference linking the word to every web page that contains the word (on our level 1 DFD in *Fig 2.86* we have named this reference PageID). In reality the precise location of each word on each page is also stored.

The indexer receives entire web pages from the search robot. Firstly the indexer assigns the web page a unique PageID. It then works freely and sequentially through the page examining each word. For each word found, apart from 'stop words', the indexer stores the PageID and the precise location of the word alongside the word within the index database. Stop words are common words that are irrelevant in terms of narrowing searches; examples include the, is, and, or, how, why and single digits and letters. This search process is known as "free text searching". Finally the indexer creates a web page summary and stores it and the PageID within the Web Page Summaries database. The web page summary commonly includes the page title from within any HTML Title tags, the data specified within any description HTML META tags and the first sentence or two from the body text.

- **Process user searches**

The flowchart shown in *Fig 2.88* describes the general sequence performed to process a single search. First the user enters their search criteria via the search engine's web page. The criteria can include a simple list of relevant words, words not to be included, logical expressions, phrases and even free text such as questions and sentences. The search criteria are transmitted to the web server hosting the search engine's web page.

Search criteria for most search engines will not include words preceded by a minus sign, for example '-frog' would exclude web pages containing the word frog. Words to include can be preceded by a plus '+' sign, however

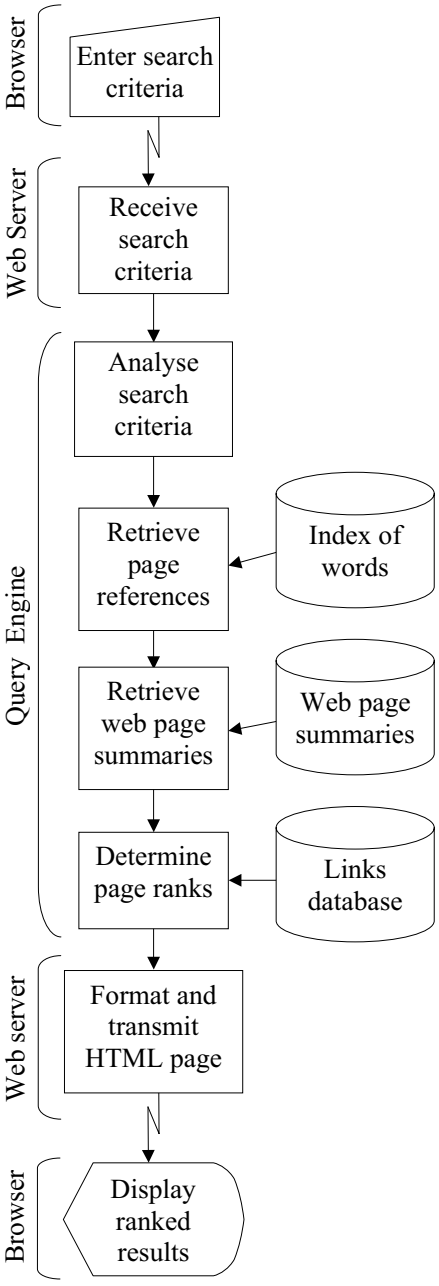


Fig 2.88
Flowchart describing the processing of a user's search.

generally the plus sign is optional. The logical operator OR can be used however AND is optional – lists of words are assumed to be separated by AND operators. Each search engine includes many other options, usually accessible via an advanced search screen.

The web server passes the search criteria to the query engine. The query engine is software whose overall task is to produce a set of ranked web pages. First the query engine analyses the search criteria and transforms it into a logical expression. If free text was entered then stop words are removed and in some engines synonyms are included for significant words. For example a search criteria that includes the word 'good' may also search for pages containing the words 'decent' and 'nice'.

The query engine now performs the search. This involves looking up each word in the index database and retrieving the associated page references. References for words that are excluded are removed. Each unique reference indicates a specific web page summary – these summaries are now retrieved.

Determining how the pages will be ranked is an involved and complex task. In our simplified example the number of times each link appears in the links database is used – those with many links being ranked higher. In reality many more factors are considered - for example, how often the page has been accessed by users, the position and size font of the search words within the page and how close together the search words appear on the page. The detail of how rankings are precisely determined is kept secret and no doubt is significantly different for competing search engines. Google is reported to use more than one hundred different factors when determining their rankings.

Consider the following:

Search Engines

Source: Consult-X.com (Modified)

Tips & Hints to increase your site's ranking

WARNING: Do not 'spam' the engines. It is easy to do so even unwittingly.

Register with relevant search engines - this is best done manually, but automated software submission may also be used. Although engines should spider all sites eventually, this is often a very hit or miss affair.

Keyword Placement - keywords or phrases that users are expected to enter into search engines must be appropriately seeded in the relevant pages. Choice of keywords is critical. Optimisation is a black art that requires keywords be used in the right locations (e.g. Titles, Headers, body copy, URLs, anchor text, etc.) with the right frequency (e.g. not more than 7 times per page), and density (relative to the total volume of words on a page).

Meta Tags - Not all search engines recognise meta tags, but some of the most important do. Title tags and Keyword meta tags may be used in the ranking process by some engines, but they must relate to the page.

Link Popularity - used by many engines, notably Google, as an important criterion for relevance and authoritativeness. Sites with few external links pointing to them may find it difficult to achieve high rankings.

Freshness - web sites that are not updated for long periods of time are considered 'stale' and therefore less important. Ensure you make frequent changes that are noticeable to the engines.

Structure - avoid frames, and more than two or three levels of directories; some engines just won't follow the links.

GROUP TASK Discussion

There are many companies whose sole task is to improve the rankings of their customer's web sites. Why is a high ranking so valued? Discuss.

GROUP TASK Discussion

Explain why only tips and hints are available rather than guaranteed techniques for improving search engine rankings.

SET 2H

1. All distributed databases:
   (A) are shared between multiple users.
   (B) are stored on multiple computers.
   (C) include each data item once only.
   (D) contain multiple copies of data.

2. Storing different records from the same table in different locations is known as:
   (A) horizontal fragmentation.
   (B) vertical fragmentation.
   (C) downloading.
   (D) replication.

3. Each database holds an almost complete copy of all records when the distributed database is based upon which strategy?
   (A) fragmentation
   (B) downloading
   (C) replication
   (D) Both B and C.

4. The domain name service uses which distributed database strategy?
   (A) horizontal fragmentation.
   (B) vertical fragmentation.
   (C) downloading.
   (D) replication.

5. What type of search engine sends search criteria to other search engines?
   (A) Meta-search engine.
   (B) Web crawler based search engine.
   (C) Web directories.
   (D) All search engines do this.

6. Criteria used by search engines to rank pages includes:
   (A) The number of links to the page from other pages.
   (B) The proximity of search words on each page.
   (C) The font size of search words on matching pages.
   (D) All of the above.

7. Which of the following best describes the actions of search robots?
   (A) Extracting and storing links from retrieved web pages.
   (B) Creating an index of all words on a retrieved web page.
   (C) Analysing web pages to create and store page summaries.
   (D) Retrieve ranked lists of links based on user inputs.

8. Which of the following best describes 'free text searching'?
   (A) No charge is made to execute the search.
   (B) A system used to locate specific words within a text document.
   (C) A formal system of specifying search criteria using symbols and logical operators.
   (D) The process performed by search engines as they examine all text on each page of a web site.

9. A distributed database contains a master *m* and two replicants *p* and *q*. A record is added to *p*. What needs to occur for this record to appear in *q*?
   (A) Replication between *m* and *p*.
   (B) Replication between *m* and *q*.
   (C) A and then B.
   (D) B and then A.

10. *p*, *q* and *r* are DNS servers. A web browser performs a DNS lookup at *p*. *p* does not know the IP address so it performs a DNS lookup at *q*. *q* knows the IP address. Eventually the web browser displays the requested web page. Who definitely knows the IP address for the displayed web page?
    (A) The browser and DNS server *q*.
    (B) The browser and DNS servers *p* and *q*.
    (C) The browser and all three DNS servers.
    (D) The browser and DNS server *p*.

11. Define each of the following terms.
    (a) Distributed database (b) Search engine (c) Search robot (d) Indexer

12. Assume a simple distributed database contains a single table. Explain how a new record added at one location is made available to users at all locations for each of the following strategies.
    (a) Fragmentation (b) Downloading (c) Replication

13. Imagine you have just created a new web site. Within days the web site appears within the results returned by many search engines, yet you have not submitted the site to any of these search engines. Explain how this can occur.

14. Many search engines rank pages based on the number of links to the page. Explain how search engines determine the number of links to a page.

15. With regard to distributed databases, compare and contrast fragmentation and replication strategies.

COLLECTING AND DISPLAYING FOR DATABASE SYSTEMS

Collecting is the information process that gathers data from the system's users. This data is processed by the system and the resulting information is returned to users via the displaying information process. Collecting and displaying are the only information processes that directly interact with users. Indeed many users choose particular software applications based largely on the look and feel of user interfaces. The design of collection and display screens is critical – after all these screens are the user's only window into the system.

Collecting
The information process that gathers data from the environment. It includes knowing what data is required, from where it will come and how it will be gathered.

Displaying
The information process that outputs information from an information system.

Collecting is an input process and displaying is an output process, between these two processes various other information processes occur to transform the data into information. The context diagram in *Fig 2.89* appears to simply describe this traditional Input Process Output (IPO) model of data processing. However there is more going on here than it may first appear – in reality collecting and displaying support each other. Consider a data entry screen; clearly the display is being used to control the collecting process. Similarly printed and on-screen reports are displayed in response to user inputs – data is collected to determine the parameters or search criteria upon which the report is based.

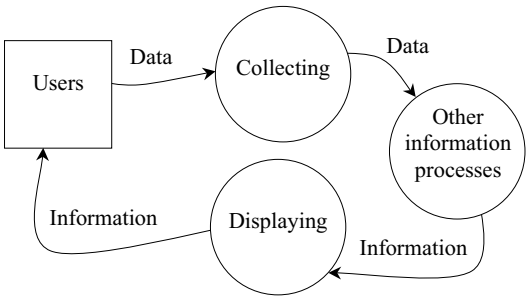


Fig 2.89
Collecting and displaying interact with the system's users.

Within DBMS server based systems data entry screens (forms) and reports are produced within the client software applications. The client software application generates SQL statements and sends them to the DBMS server. Behind forms `INSERT` and `UPDATE` queries are created to add new records or alter existing records. Within reports `SELECT` queries are used to retrieve information. Most forms also display existing records – the source of this data is also a `SELECT` query.

GROUP TASK Discussion
What are the advantages of using a database server to execute queries? Why not simply execute all queries within the client applications?

In larger database systems different views of the data are made available to different users. The underlying record source for both forms and reports is built using SQL statements based on defined views of the database rather than the raw tables. For example within a hospital the views available to the administration staff would not include notes made by doctors. On the other hand the doctors views would include such notes, however they would not include patient's financial records.

GROUP TASK Discussion
Consider a school's information system. Discuss data that should be excluded from student views of the data.

SCREEN AND REPORT DESIGN PRINCIPLES

Great data entry screens or forms can be used with minimal training. The user is able to infer how to use the screen based on past experience and clues inherent in the user interface. This is a clear indicator of a truly user-friendly screen. But how is this achieved? In this section, we examine a series of design principles that will help you to design and create intuitive data entry screens.

Report design requires an assessment of the nature of the information. How is it categorised and sorted? What user needs does it hope to address? And importantly who are the intended users? Are they internal employees or members of the general public? The aim is to produce a format that best communicates the information to the intended audience. Every report is different; hence in this section we consider general principles that apply in most cases.

In this section we examine a number of general principles that should be considered when developing data entry screens and reports.

Screen and report design principles:

- Consistency of design
- Grouping of information
- Use of white space
- Judicious use of colour and graphics
- Legibility of text

Additional screen design principles:

- Data validation
- Effective prompts

Consistency of design

Perhaps the most important aspect of screen and report design is consistency. For example, labels describing data should be placed in consistent places on all screens. All reports should have a consistent feel – consider using the same typeface in various sizes and styles. Use of a consistent colour scheme and layout also assists in this regard.

Consistency of screen design allows users to transfer skills from other applications and other data entry screens within your software. A mental model of the software application's operation is built up in the user's mind; consistency reinforces this model.

Some points vital to the development of consistent screens include:

- Setting standards and sticking to them. Set design standards based on industry standards and then develop extra standards specific to your application. Most operating system developers have sets of design standards; these standards will often define the majority of your screen design needs.
- Explain the rules. Develop a simple set of rules that apply to your entire application. In this way, you will need to explain your rules only once. This makes more sense than explaining the details of how to use each individual screen within your application.
- Use interface elements correctly. Know when to use screen elements and how to use them correctly. For example, command buttons, check boxes, radio buttons, list boxes, text boxes, menus, etc. Follow recognised standards for each of these elements. Never alter the operation of standard screen elements to perform unexpected functions.

Each of the other principles examined in this section should be used consistently. Readability is greatly increased when screens and reports make consistent use of white space, colour, graphics, grouping and text. When using screens users should be able to predict the result of their actions. When generating reports, users should be able to pretty much predict what the report will look like based on prior experience generating your other reports.

GROUP TASK Practical Activity
Examine a number of applications or websites in regard to the consistency of their screens. List any items of inconsistency that you find.

GROUP TASK Activity
Large organisations, such as banks and phone companies post out printed reports of various types to their customers. Examine examples of reports from one organisation and comment on their consistency.

Grouping of information

It makes sense to logically group related items and data together. On data entry screens the actual screen elements can be logically grouped. On reports the data itself is grouped into logical categories. In either case grouping allows users to, at a glance, internalise the overall content of the screen or report. They can then focus on the required elements more efficiently.

Grouping is emphasised using borders, lines, different fonts, colours or even page breaks. Often a label is used to describe each group. The label should concisely communicate the nature and purpose of the elements within each group. When multiple records are displayed labels normally appear at the top of each column. When single records are displayed the labels appear to the left of text and to the right of check boxes and radio buttons.

The query upon which data entry screens and reports are based is usually sorted. The sort order is often the basis used to group multiple records. On data entry screens that display a single record the sort order determines which record appears next as the user navigates through the records. On multiple record layouts the sort order determines where each group starts and ends. For example, queries for producing invoices are generally based on a query that sorts by customer, then invoice number and then products ordered. The customer, including their name and address details, is placed in the header, the invoice number determines page breaks and the products ordered are listed in rows. An example of such an invoice was used as the stimulus for our discussion on normalising databases (see *Fig 2.22* earlier in this chapter).

Consider the following:

Parramatta Education Centre, the publisher of this text, also writes software used to operate schools and small businesses. Most of these applications use a DBMS. *Fig 2.90* on the next page shows three screens from a school's reporting package. We shall refer back to these screens throughout this section of the text.

This first data entry screen is used to set-up each course prior to reports being written by teachers. Essentially this screen collects data for a single record in the Courses table of the reporting database. The second screen collects teacher details. The third screen is used by teachers to enter marks, comments and other data as they write each student report. Various methods of grouping are used on all these screens.

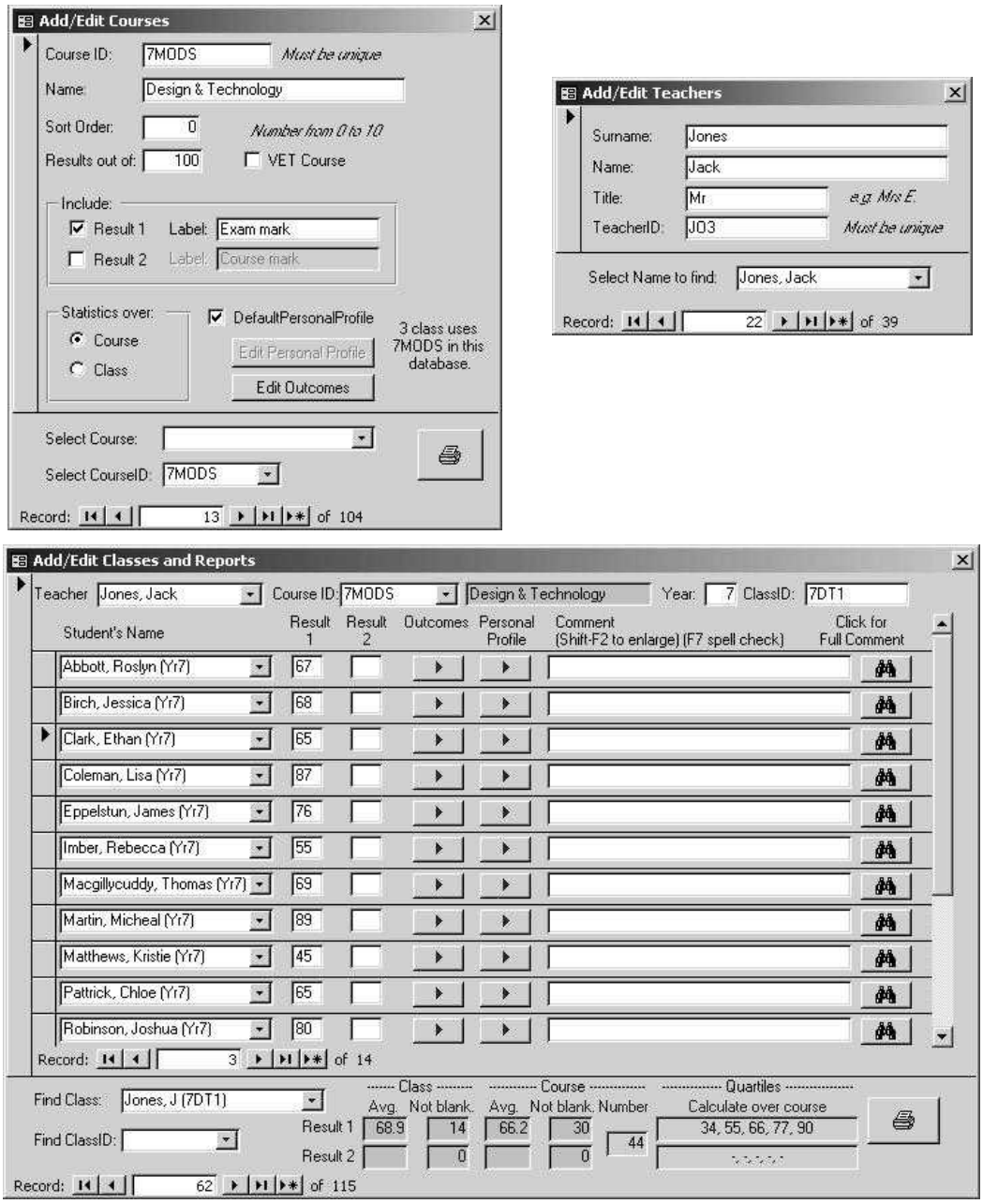


Fig 2.90
Data entry screens from a school's reporting application.

GROUP TASK Discussion
Identify the likely attributes present in the Courses table based on evidence from the "Add/Edit Courses" data entry screen above.

GROUP TASK Discussion
Do you think the three screens above have a consistent design? Justify your response using examples.

GROUP TASK Discussion
Discuss how grouping of screen elements is used on each of the above screens.

Consider the following:

The screen in *Fig 2.91* below is the results of an SQL query. This query will be used as the data source for a report that prints out class lists for teachers.

qryClassLists : Select Query

Teachers.Surname	Title	Courses.Name	ClassID	Students.Surname	Students.Name
Hundrill	Mrs	Design & Technology	9DT4	Lennane	Aron
Hundrill	Mrs	Design & Technology	9DT4	Liebich	Scott
Hundrill	Mrs	Design & Technology	9DT4	Logan	Jonathon
Hundrill	Mrs	Design & Technology	9DT4	Matthews	Glenn
Hundrill	Mrs	Design & Technology	9DT4	McFarlane	Luke
Hundrill	Mrs	Design & Technology	9DT4	O'Brien	Clint
Hundrill	Mrs	Design & Technology	9DT4	Walker	Jonathan
Hundrill	Mrs	Design & Technology	9DT4	Walmsley	Brett
Hundrill	Mrs	Design & Technology	9DT4	Wood	Christopher
Jones	Mr	Design & Technology	7DT1	Abbott	Roslyn
Jones	Mr	Design & Technology	7DT1	Birch	Jessica
Jones	Mr	Design & Technology	7DT1	Clark	Ethan
Jones	Mr	Design & Technology	7DT1	Coleman	Lisa
Jones	Mr	Design & Technology	7DT1	Eppelstun	James
Jones	Mr	Design & Technology	7DT1	Imber	Rebecca
Jones	Mr	Design & Technology	7DT1	Macgillycuddy	Thomas
Jones	Mr	Design & Technology	7DT1	Martin	Micheal
Jones	Mr	Design & Technology	7DT1	Matthews	Kristie
Jones	Mr	Design & Technology	7DT1	Pattrick	Chloe
Jones	Mr	Design & Technology	7DT1	Robinson	Joshua
Jones	Mr	Design & Technology	7DT1	Smith	Cherie
Jones	Mr	Design & Technology	7DT1	Stonestreet	Justan
Jones	Mr	Design & Technology	7DT1	Troy	Adam
Jones	Mr	Design & Technology	7DT2	Bowerman	Nathan
Jones	Mr	Design & Technology	7DT2	Burn	Tegan
Jones	Mr	Design & Technology	7DT2	Clapham	Julia
Jones	Mr	Design & Technology	7DT2	Clark	Erin
Jones	Mr	Design & Technology	7DT2	Cupitt	Craig
Jones	Mr	Design & Technology	7DT2	Draffin	Sian

Record: 731 of 1679

Fig 2.91
Results of a query used to create class lists for a school.

FORUM

GROUP TASK Discussion
Examine the above query results and discuss how this information could be suitably grouped and laid out so that each teacher receives an appropriately formatted list for each of their classes.

Use of white space

We see white space all the time, yet often we are unaware of it. It is the portion of any written or artistic works that isn't filled. White space is the area on screens and reports that is not used. However, white space is a vital component of all user interfaces and printed documents. White space need not be white. In the case of graphical user interfaces it is often grey and on command-based interfaces it is commonly black

White space should be a neutral colour that does not attract or draw the user's attention. White space breaks up the screen into sections. It draws the user's eye to important elements and highlights these elements. Consider the two magazine advertisements in *Fig 2.92*. Clearly the one on the right uses large areas of white

space to focus the reader's eyes on the shirt box and hence the Cold Power logo. In comparison, the advertisement on the left does not clearly direct the reader to one particular element of the design. Although these are extreme examples, they do clearly illustrate the importance of white space.

Fig 2.92
The use of white space in advertising.

Judicious use of colour and graphics

Effective screens and reports use colour and graphics to achieve some specific purpose. It is common practice to use blue text for hyperlinks and red to signal some potential danger or problem – perhaps an account that has a negative balance. Extensive use of coloured text can be extremely distracting for users. Although the colours chosen may look appealing to the designer, it is likely that many users will not agree. Different display devices show colours differently, combinations that are readable on one display may be next to impossible to read on other displays. Use coloured text sparingly and only where it has some purpose or message to deliver.

Graphics used as icons should deliver a clear message to the user as to their purpose. Effective icons communicate their purpose more quickly than the equivalent text. However the reverse is also true, poor icons confuse users. If the meaning of an icon is not immediately clear then text should be used.

The use of colour and graphics for purely aesthetic purposes should generally be avoided. Aesthetic elements that are attractive at first sight quickly become dreary distractions for frequent users of database software. If used, then an option should be included to turn these elements off or modify them to suit the user's preferences.

Consider the following:

The icons used on the toolbar in *Fig 2.93* are supposed to improve the readability of this portion of the screen. Unfortunately, the purpose of many of the icons used is unclear.

GROUP TASK Discussion
Examine each icon on the toolbar in *Fig 2.93*. Discuss each icon's possible purpose. Do you think the purpose of each icon is clear? Discuss.

Fig 2.93
Sample icons

Legibility of text

Legibility of text refers to the user's ability to make out each word and or character on the screen or report. Primarily, the font used and how it is justified or aligned on the screen, influences legibility. Different fonts and methods of justification are suited to different uses. We need to understand how legibility is affected by our choice and use of fonts and justification.

10 point Bold Arial
12 point Italic Arial
14 point normal Arial

Fig 2.94
Three different fonts that all use the Arial typeface.

A font is a complete set of characters that are of the same design. Each font is a particular example of a typeface. For example, Arial is a typeface and 10 point bold Arial is a font. Therefore, each font possesses a series of

properties; it uses a typeface e.g. Arial, a typestyle e.g. italic or bold and a size e.g. 10 point. In most cases one or at most two typefaces should be used. To highlight particular elements on the screen or report use a different size or typestyle but keep the typeface the same.

Fonts are classified as either serif or sans serif fonts. Serifs are the little ticks or blobs attached to each end of the curves and lines that make up each character. Sans serif means, in French, no serifs. Hence sans serif fonts have no serifs and serif fonts do - *Fig 2.95* shows an example of a serif and sans serif font.

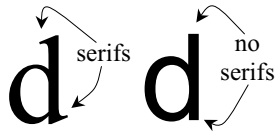


Fig 2.95
A serif font on the left and a sans serif font on the right.

So how do fonts affect legibility and as a consequence the readability of text? It depends on where the font is to be used. Reports that will be printed generally use serif fonts for the main body of the document and sans serif fonts for headings. Research has shown that serifs help the reader to more efficiently make out the shape of each character. They also assist in keeping the eye tracking across each line. The smaller the font then the more significant this becomes.

What about screens? The resolution of a printed document is quite different to that of a computer monitor. Most books are printed using a resolution of at least 1200 dots per inch; monitors have a resolution of around 70 dots per inch. Command based systems often have a far lower resolution than this. Unfortunately this low resolution can often blur the serifs resulting in lowered legibility. Simple sans serif fonts are not affected to the same degree. *Fig 2.96* shows an enlarged view of the letter 'd' as viewed on a typical monitor. The simpler shape of the sans serif version on the right makes it more legible. Remember, it is the simpler curves and lines that make the font more readable. A fancy sans serif font may be less legible than a basic serif font.

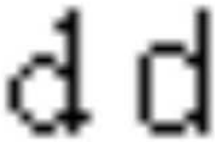


Fig 2.96
Serifs are blurred on computer monitors.

Justification is how text is aligned to the margins. Left justified text is tight against the left margin. Conversely right justified text is tight against the right margin. Full justification means the text is spread evenly between both margins. Centred means the text is equidistant from both margins. Many word processors use the term alignment in preference to the term justification.

Screens and reports rarely use multiple lines of text. However, they often contain lists of labels and data. In general all lists of text should be left justified. It is easier to absorb a list of items when each commences at the same point. Centred text should only be used for headings or for specialised screen elements such as the text on command buttons. Right justification is used for numbers. This ensures the decimal points line up directly underneath each other.

Consider the following:

A data entry screen is required to gather and display client information. The fields to be included on this screen include: surname, first name, sex, street address, suburb, postcode, phone number, fax number, mobile number and email address. Two screen designs that formed part of the first and second prototypes have been created and are shown on the next page in *Fig 2.97*

GROUP TASK Discussion
Comment on the fonts, justification and grouping used on the screens in *Fig 2.97*.

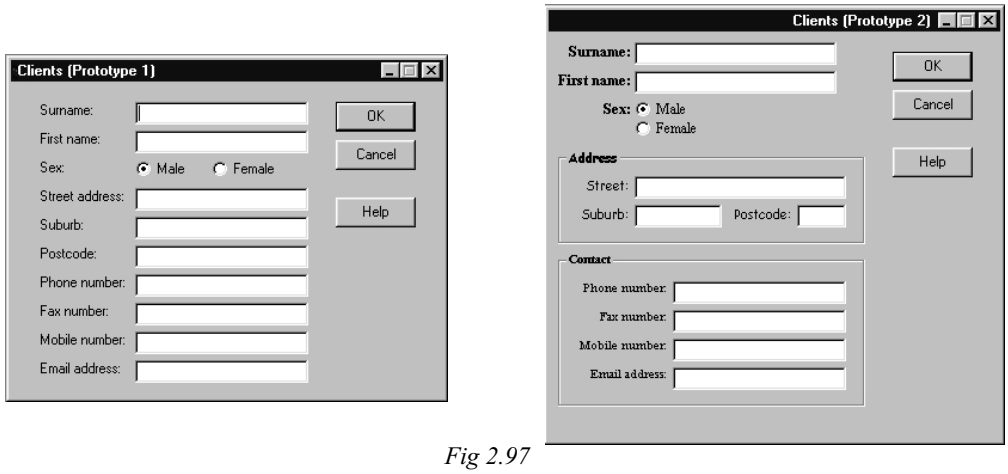


Fig 2.97
Two screen designs for collecting and displaying client information.

Data Validation

Data Validation
A check, at the time of data collection, to ensure the data is reasonable and meets certain criteria.

Data validation is a check during data collection to ensure that reasonable data is entered. For example, when entering the cost of a product, data validation criteria would likely ensure a positive number is entered. Data validation certainly helps to improve the integrity or correctness of data, however it cannot ensure the data is actually correct. The real cost of a product may be $5.50, however if $5.60 is entered then no data validation is going to identify the error – data verification procedures are needed to correct these types of errors.

For DBMS based information systems data validation can be implemented in two locations; within the client software applications or it can be specified and enforced within the database itself. In most cases a combination of both these locations is used. There are two competing aims that need to be balanced – the data must be kept clean within the database and secondly each data item should logically be validated immediately a user has entered it.

DBMS software operates on databases at the record level. New records are inserted and existing records are updated as complete units of data. This means the DBMS software can only perform validation once it has received a complete record. If data validation only occurs when a record is sent to the DBMS then validation errors will only occur once the user has entered an entire record of data.

Consider a typical online purchase. The user selects a product, enters their name, address, phone number, email address, credit card number, expiry data, etc, etc... they then click the submit button. The record is sent to the DBMS, the DBMS tries to write it to the database, the database finds the email address doesn't comply with its validation rules. Eventually the user is presented with a message stating the email address error – this occurs some time after they originally entered the email address. Now consider the same scenario when the client application performs the validation. Immediately after the email address has been entered the client software application performs its validation. The user is immediately notified - clearly a much more user-friendly solution. So what's the problem, why not simply validate within the client applications?

The problem occurs because most large databases are accessed from many different client software applications. The administrators of such databases cannot rely on each

client application to validate all fields correctly; hence they include validation within the database itself. Furthermore as there is just one central database any changes to data validation rules need only be made once – within the database.

Within MS-Access validation rules can be set for each attribute of each table For example in *Fig 2.98* the Result1 field will only allow numbers from 0 to 100 (or Nulls). If numbers outside this range are entered then a message is displayed containing the validation text "Number from 0 to 100 expected." Access also allows individual screen elements to perform further validation. When a screen element is connected or bound to a field then both validation rules are checked before a record is stored

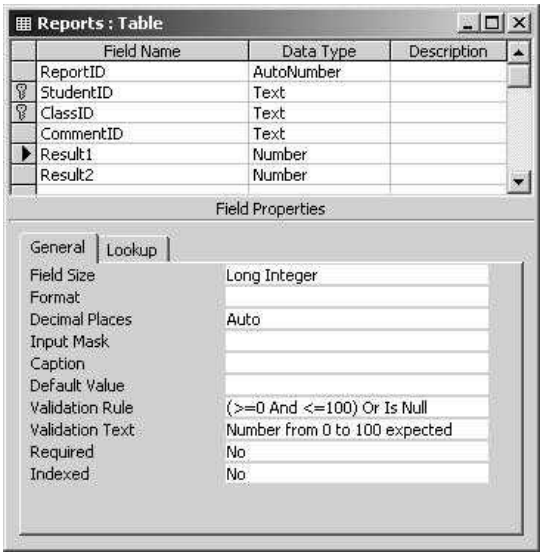


Fig 2.98
Setting validation rules in MS-Access at the table level.

Data entry screens often use self-validating components that ensure only valid data can be entered. For example, sets of radio buttons restrict the range of data that can physically be entered to one of the available choices, hence radio buttons are said to be self-validating. Other self-validating screen elements include list boxes and combination boxes where the possible inputs can be restricted to those present within the list.

GROUP TASK Activity
Examine data entry screens from software applications installed on your school or home computer. Describe the different types of data validation used for each element on these screens.

Effective prompts

A prompt is a reminder or a cue as to what is required. For example, most stage productions have prompters standing in the wings. Their job is to prompt the actors using cues should any of them forget their lines. Prompts on user interfaces perform a similar task for users rather than actors. They must be concise yet they should accurately communicate their message.

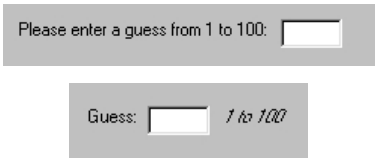


Fig 2.99
Prompts should accurately and concisely communicate a single simple message.

A prompt is not the place to teach users about the details of the screen. Neither is it necessary or desirable to embellish the wording of prompts. Consider the two prompts in *Fig 2.99*. The top one sounds nice and fuzzy and friendly when first read. What if you had to read it hundreds of times during each day? Suddenly the fuzzy and friendly becomes irritating and annoying. Obviously you need to enter a guess, if you didn't then why is there a text box containing a flashing cursor. The bottom prompt in *Fig 2.99* communicates the same message and is far less likely to irritate.

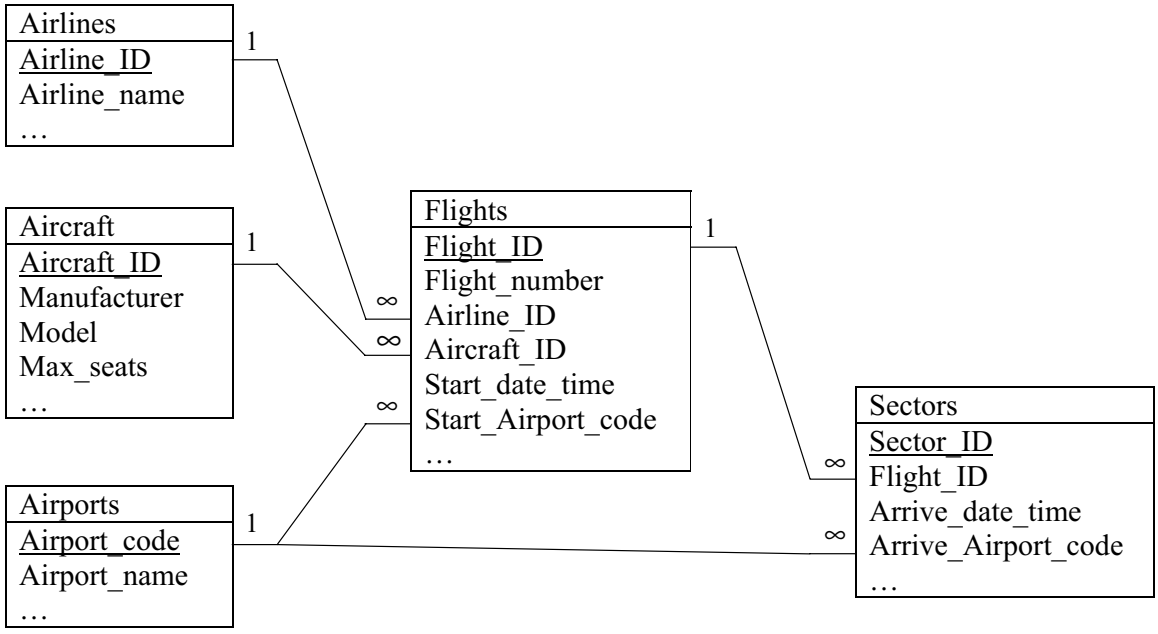
Prompts are the main method of communicating with users. Most screen elements contain or are linked to prompts. Without them it would be impossible for users to understand and use software. It is well worth spending time considering the words used to ensure they correctly communicate the desired message.

Some general guidelines for developing prompts are:

- Use verbs or doing words if choosing a screen element activates some process or action. Often verbs are used on menus e.g. file, edit, print or format. The implication is that choosing one of these items will lead to some action being performed.
- If the prompt is merely used to gather data of some sort then the prompt should be a noun that describes the data e.g. Surname rather than Enter surname.
- It is common practice to use an ellipsis (…) after prompts that open new windows. The ellipsis also signifies that the item does not directly activate a process.

HSC style question

The following schema describes the organisation of an aircraft flight information database.



(a) Identify an example of an entity, attribute and relationship within the above schema.

(b) A particular Qantas Boeing 747 flight leaves Sydney and flies via Auckland to Hawaii. Describe the records required in each table to represent this flight.

(c) Construct an SQL statement to retrieve all flights commencing from Los Angeles airport (Airport code LAX) on the 23/08/2004. The results are to be sorted on departure time and only the flight number and departure date/time should be displayed.

(d) Design a screen that could be used to enter and edit flights, including flight sectors.

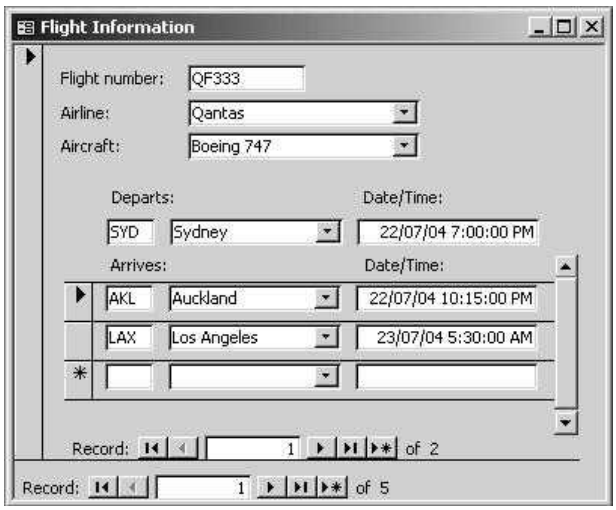
Suggested Solution

(a) Entity: Airlines table, Attribute: Airline_ID within Airlines table, Relationship: Airline.Airline_ID is a primary key with a 1 to many join to the foriegn key Flights.Airline_ID.

(b) A single record exists in the Flights table. This record links the flight to a single record in each of the Airlines, Aircraft and Airports tables. The Flights record contains the primary key Flight_ID, which is used to link to each sector record for the flight. In this example the Flights record contains the Airport code for Sydney in Start_Airport_code. There would be two records for this flight within the Sectors table. One would have an Arrive_Airport_code for Auckland and the other would contain the code for Hawaii. Both these records relate to the corresponding record in the Airports table. In total, 8 records contribute to represent the flight.

(c)
```
SELECT Flights.Flight_number, Flights.Start_date_time
FROM Flights
WHERE Start_date_time>='23/8/2004' And Start_date_time<'24/8/2004' AND Start_Airport_Code="lax"
ORDER BY Flights.Start_date_time
```

(d)



Comments

- In part (a) any entity, attribute or relationship could have been identified.
- The single quote delimiters surrounding the dates in part (c) are different for individual DBMSs. The delimiters are not important in terms of marks awarded.
- Obviously the screen design in part (d) was produced using Microsoft Access. Clearly a hand drawn sketch that includes all significant features would be produced in an exam.
- Required points for the screen design to attract full marks include:
  - Includes all necessary fields.
  - Does not include key fields that should be hidden from users.
  - Appropriate descriptive labels (NOT field names).
  - Uses appropriate screen elements (such as combo boxes).
  - Provides navigation elements.
  - Layout of elements makes logical sense.

SET 2I

1. With regard to collecting and displaying information processes, which of the following is true?
   (A) Collecting is an input process and displaying is an output processes.
   (B) Both collecting and displaying are output processes.
   (C) Both collecting and displaying are input processes.
   (D) Collecting turns data into information for displaying.
2. The data displayed on most data entry screens and reports is retrieved using SQL:
   (A) SELECT statements.
   (B) UPDATE statements.
   (C) DELETE statements.
   (D) INSERT statements.
3. Which of the following is the most significant reason for using a consistent design for all data entry screens?
   (A) The screens look better and hence higher sales for the product are more likely.
   (B) The screens can be copied and pasted, thus time spent creating the screens is reduced.
   (C) Users are able to transfer skills from other screens and even other products.
   (D) The data entry screens can then be used to access data from many different DBMS servers.
4. Which of the following is NOT true in regard to the use of white space on data entry screens and reports?
   (A) It is always white.
   (B) It rests the user's eye.
   (C) It breaks the screen or report into logical sections.
   (D) It draws the user eye to important elements.
5. Sans serif fonts should generally be used for:
   (A) computer monitors.
   (B) large headings.
   (C) low resolution output.
   (D) All of the above.
6. Within DBMS server based systems new records are added using SQL:
   (A) SELECT statements.
   (B) UPDATE statements.
   (C) DELETE statements.
   (D) INSERT statements.
7. On a web page a message is displayed immediately after each incorrect input. Where is the data validation most likely to be occurring?
   (A) At the web server.
   (B) At the DBMS server.
   (C) Within the browser.
   (D) At the ISP.
8. Which of the following is definitely NOT a self-validating screen element?
   (A) radio buttons
   (B) check box
   (C) text box
   (D) list box
9. On data entry screens the prompts or labels identifying each data item to enter should:
   (A) be phrased as questions.
   (B) include information about the range of data items expected.
   (C) include a typical example of the data to be entered.
   (D) be simple nouns.
10. 12 point bold times new roman is an example of a:
    (A) typeface
    (B) font
    (C) typestyle
    (D) font size
11. Explain how each of the following principles affect the design of data entry screens and reports.
    (a) Consistency (b) Grouping (c) White space (d) Legibility
12. Describe the connections between data entry screens, SQL statements, DBMS servers and databases.
13. Consider the aircraft flight information HSC style question on the previous pages. Design reports to display the following information.
    (a) The movements of a particular aircraft over the next week.
    (b) A list of all flights arriving at a particular airport on a particular day.
14. Consider the invoicing database from *Fig 2.70* on page 185. Design data entry screens to collect and/or edit the following data.
    (a) Customer details.
    (b) Invoice details, including products and quantities ordered.
15. When entering data into a database validation often occurs twice on the same data. Explain how and why this occurs.

ISSUES RELATED TO INFORMATION SYSTEMS AND DATABASES

In this section we first consider ethical issues relevant to the use of databases and also hypertext. These issues largely revolve around how data is collected and from whom, how the data is kept accurate and who controls and is permitted to view the data. We consider:

- Acknowledgement of data sources
- Access, ownership and control of data
- Accuracy and reliability of data

Finally we briefly examine current and emerging trends, in particular data warehousing, data mining, online analytical processing (OLAP) and online transaction processing (OLTP).

ACKNOWLEDGEMENT OF DATA SOURCES

The most significant source of data for most information systems is ultimately people – often these people indirectly supply data to the system without even being aware. Consider school assignments – you read or examine a number of primary and secondary data sources - these sources were written by people. You consider the ideas within these data sources and then formulate your own ideas. The final assignment includes your original ideas developed using the ideas of others. Likely your school requires you to acknowledge any outside data sources, particularly if you use direct quotes. However complying with your school's regulations should not be the overriding reason for acknowledging sources, rather it is simply the right thing to do. The same applies to data within databases and hypertext.

GROUP TASK Discussion

How would you feel if you found copies of an assignment you had written were being circulated on the Internet without your permission? Would your feelings change if you found some copies acknowledged someone else as the author? Discuss.

Databases and hypertext webs include large quantities of data collected from a diverse set of data sources. Often tracking the primary source can be difficult – the data has been obtained from a secondary source that obtained the data from another secondary source and so on. If the data contains private information about individuals then privacy laws may govern its use. If the organisation is included within freedom of information legislation then tracking the source of data may well be a legal necessity. If the data contains original ideas or artistic works then the laws of copyright may apply. Even if such laws do not apply, acknowledging the source of data is still ethically the right thing to do.

In the preliminary course (Chapter 1 of the related text), we examined the *Copyright Act 1968* and its implications when using or copying software applications and databases of information. We found the laws governing copyright do not apply to the actual information within a database but rather to the work and expense used to gather the information together. This means copyright is breached when an existing database is copied without permission and acknowledgement.

There are many reasons for acknowledging the source of data used within databases. Some of these reasons include:

- Justification of outputs. For example, the results from surveys will only be accepted if the source of the data can be shown to be accurate. Describing and acknowledging the data source assists in this process.

- Providing a mechanism for tracking and auditing data. If the source of data is unknown then it is difficult to track and determine the accuracy of the data. For example, audits of financial transactions must be able to determine the precise source of each transaction to check its authenticity.
- Requirements of the source organisation. Many sources require, or at least request, that they be acknowledged when others use their data.

GROUP TASK Discussion
Why do you think copyright laws in Australia, and many other countries, cover the work and expense used to gather information rather than covering the information itself? Discuss using examples.

ACCESS, OWNERSHIP AND CONTROL OF DATA

Many databases store private information on individuals and confidential information with regard to the operation of organisations. Access to such data clearly needs to be controlled. We examined various techniques for securing data earlier in this chapter – these techniques are a means to control the data and ensure only authorised persons have access. Control is clearly needed if data access is to be restricted, however there is a downside. Control of any asset gives the controller power. When the asset is private or confidential information then the techniques used to control the data can have the effect of making the data unavailable to those the information describes. Individuals feel controlled and powerless, they are 'small fry' compared to the size of large commercial and government organisations. To redress this imbalance most governments around the world have created laws to ensure data on individuals and organisations is made available to those the data describes, but is not made available to unauthorised persons.

GROUP TASK Discussion
Many organisations own private and confidential data about you. You don't own this data, they do! How can this be ethical? Discuss.

The appropriate use of information systems is often detailed as a policy statement for users within the organisation. In legal terms, these policies must meet the requirements of the relevant freedom of information and privacy legislation. Such policies outline inappropriate activities together with the consequences should a user violate any of the conditions. Typically such a policy statement would include the following activities as inappropriate usage:

- Unauthorised access, alteration or destruction of another user's data, programs, electronic mail or voice mail.
- Attempts to obtain unauthorised access to either local or remote computer systems or networks.
- Attempts to circumvent established security procedures or to obtain access privileges to which the user is not entitled.
- Attempts to modify computer systems or software in any unauthorised manner.
- Unauthorised use of computing resources for private purposes.
- Transmitting unsolicited material such as repetitive mass mailings, advertising or chain messages.
- Release of confidential information.
- Unauthorised release of information.

GROUP TASK Research
Use the Internet, or otherwise, to locate examples of policy statements that deal with appropriate information use.

In Australia the federal *Freedom of Information Act 1982* and in NSW the *New South Wales Freedom of Information Act 1989* are the legal documents specifying the laws in regard to an individual's access to information. Currently these acts only apply to government departments and their related statutory authorities – they do not apply to commercial organisations. In other countries freedom of information law covers organisations of all types.

The laws in regard to restricting access to private and confidential data are specified within the federal *Privacy Act 1988 (Cth)* and in NSW within the *Privacy and Personal Information Protection Act 1998* and also within the *Health Records and Information Privacy Act 2002*. Other states in Austalia have similar legislative Acts to NSW. Within this section of the text will restrict our treatment to a brief examination of the nature of the *New South Wales Freedom of Information Act 1989* and the federal *Privacy Act 1988 (Cth)*.

Freedom of Information (FOI) Acts

In New South Wales there are two legal documents to consider; the federal *Freedom of Information Act 1982* and the *New South Wales Freedom of Information Act 1989*. The following questions and answers are an extract reproduced from the NSW Premier's Department website at http://www.premiers.nsw.gov.au - they relate to the *New South Wales Freedom of Information Act 1989*.

- **What is Freedom of Information?**

  In New South Wales, the *Freedom of Information Act 1989* gives you the legal right to:
  - Obtain access to information held as records by State Government Agencies, a Government Minister, local government and other public bodies;
  - Request amendments to records of a personal nature that are inaccurate; and
  - Appeal against a decision not to grant access to information or to amend personal records.

- **What sort of information can I ask for?**

  You can ask for any kind of personal or non-personal information. Personal information includes your public education and school records, health, welfare and superannuation records, and examination and training records. Non-personal information includes government policy documents, research materials, instruction and procedure manuals, and market research and product testing records. Information can be in the form of certificates, files, computer printouts, maps, films, photographs, tape recordings and video recordings.

- **What agencies and other public bodies can give me this information?**

  Agencies and public bodies that must give you information under FOI include:
  - Government departments and authorities
  - State boards and committees
  - Government Ministers
  - Local and municipal councils
  - Universities
  - Public hospitals
  - Regulatory bodies eg the Harness Racing Authority

- **Is any Information not available?**

  Under the Freedom of Information Act, wherever possible, agencies are required to make information available. You may be denied right of access to information only where, for example, there is a legitimate need for confidentiality or where another person's privacy may be invaded. This information is called "exempt" and includes:
  - State Government cabinet and executive council documents (with the exception of those that are factual or statistical and do not disclose cabinet or executive council deliberations or decisions);
  - Documents which are exempt under Commonwealth or other States FOI legislation;
  - Documents concerning law enforcement and public safety;
  - Documents subject to legal professional privilege; and
  - Documents subject to secrecy provisions in other legislation.

  Other information which may be exempt includes documents affecting:
  - Personal affairs of another person;
  - Business affairs of another person or business; and
  - The economy of the State.

  The Premier of NSW, as the Minister responsible for FOI, has the right to issue a Ministerial Certificate stating that a specific document is exempt and restricted.

- **Can I correct inaccurate documents about me?**

  Yes, if you believe any information about you is incomplete, incorrect, misleading, or out-of-date, you have the right to request that it is corrected.

GROUP TASK Discussion
Currently both federal and state freedom of information law does not cover private organisations. Do you think they should? And if they did, how could such laws be enforced? Discuss.

Privacy Principles

Privacy is about protecting an individual's personal information. Personal information is any information that allows others to identify you. Privacy is a fundamental principle of our society, we have the right to know who holds our personal information and that they will keep this information confidential. We need to feel confident that our personal information will not be collected, disclosed or otherwise used without our knowledge or permission.

Personal information is required, quite legitimately by many organisations when carrying out their various functions. This creates a problem; how do we ensure this information is used only for its intended task and how do we know what these intended tasks are? Laws are needed that require organisations to provide individuals with answers to these questions. In this way individuals can protect their privacy.

In NSW, privacy is legally protected via the federal *Privacy Act 1988 (Cth)*, the NSW *Privacy and Personal Information Protection Act 1998* and also the *Health Records and Information Privacy Act 2002*. We shall concentrate on the federal privacy legislation. This legislation contains ten National Privacy Principles (NPPs), that set standards that organisations are required to meet when dealing with personal information; the text in *Fig 2.100* on the next page briefly explains each of these principles.

Consequences of the *Privacy Act 1988 (Cth)* mean that information systems that contain personal information must legally be able to:

- explain why personal information is being collected and how it will be used.
- provide individuals with access to their records.
- correct inaccurate information.
- divulge details of other organisations that may be provided with information from the system.
- describe to individuals the purpose of holding the information.
- describe the information held and how it is managed.

What are the ten National Privacy Principles?

The following briefly explains what the NPPs mean for you.

NPP1: Collection - describes what an organisation should do when collecting your personal information.

NPP2: Use and Disclosure - outlines how organisations can use and disclose your personal information.

NPP3: Data Quality & NPP4: Data Security - set the standards that organisations must meet for the accuracy, currency, completeness and security of your personal information.

NPP5: Openness - requires organisations to be open about how they handle your personal information.

NPP6: Access & Correction - gives you a general right of access to your own personal information, and the right to have that information corrected, if it is inaccurate, incomplete or out of date.

NPP7: Identifiers - says that generally, Commonwealth government identifiers (such as the Medicare number or the Veterans Affairs number) can only be used for the purposes for which they were issued.

NPP8: Anonymity - where possible, requires organisations to provide the opportunity for you to interact with them without identifying yourself.

NPP9: Transborder Data Flows - outlines privacy protections that apply to the transfer of your personal information out of Australia.

NPP10: Sensitive Information - requires your consent when an organisation collects sensitive information about you such as health information, or information about your racial or ethnic background, or criminal record. Sensitive information is a subset of personal information and special protection applies to this information.

Fig 2.100
The ten 'National Privacy Principles' briefly described from the Office of the Federal Privacy Commissioner's website at http://www.privacy.gov.au

GROUP TASK Discussion

NPP7 specifically mentions Medicare numbers, however there are many other numbers that uniquely identify individuals. Brainstorm a list of such identifiers. Discuss where these identifiers are stored and how they could potentially be linked together.

ACCURACY AND RELIABILITY OF DATA

The accuracy or correctness of data is vital if the information generated is to be current and valid. For example every time someone moves home their address must be updated in numerous databases. When moving most people inform banks, telephone companies, the RTA and other organisations they deal with fairly frequently. However it is rare to move house and not receive mail addressed to the previous tenant. Updating incorrect information does not only help individuals, it is also vital for the organisations that hold the information.

All organisations, in particular commercial organisations, benefit greatly when their data is accurate and reliable. This includes not just personal details on customers and contacts, but all data. The term 'data integrity' is used to describe the correctness and quality of data.

Data Integrity
A measure of how correct and accurately data reflects its source. The quality of the data.

Within databases data integrity is improved using data validation and data verification checks. We discussed data validation earlier in this chapter – it ensures the data entered is reasonable but it in no way determines if it is correct. Data verification is an ongoing process that aims to ensure data is correct and remains correct over time.

Data Verification
A check to ensure the data collected and stored matches and continues to match the source of the data.

Data verification is a much more difficult task than data validation. For example, the computer can quite easily check that a phone number contains the correct number of digits however verifying that these digits are indeed the person's phone number is a different and somewhat more difficult task. Furthermore, people change their phone numbers.

Data verification includes all the procedures that are used to verify the correctness of the data within an information system. In regard to data entry, data verification is often implemented as a procedure whereby the data entry operator must manually compare the source data to the data just entered. For example, when taking a credit card order over the phone, the operator verifies the credit card number entered by reading it back to the customer.

What about verifying data over time? This is a much more involved and potentially expensive task that rarely detects all inaccuracies. Consider customer mailing addresses; potential integrity problems can be determined based on marketing mail outs that are undeliverable or are returned. However this does not provide the new address. Furthermore marketing mail outs are likely to be thrown out by new tenants rather than returned to the sender. Data verification aims to achieve perfect data integrity; in reality this is rarely achieved or even possible.

GROUP TASK Discussion
Does your school perform data verification checks to ensure your personal details are correct? Describe the procedure used (or that could be used).

So far our discussion on accuracy has centred on the accuracy of data within databases. Accuracy and reliability relates to verifying the correctness of all types of data and information. This is of particular importance when information is sourced from the Internet. Often conflicting opinions or even conflicting statements of fact exist. How does one validate and verify which information is of sufficient quality to be trusted? A checklist to assist is reproduced below from the preliminary course text. This list is based on the five criteria traditionally used to assess print media – they are just as valid as an aid for assessing the quality of information on the Internet.

1. Accuracy
   - Is the information well written and edited?
   - Have sources upon which the information is based been acknowledged?
2. Authority
   - Who wrote or is responsible for the information?

- Are the author's qualifications clearly stated?
- Is a phone number and address for the author or their company included?

3. Objectivity
   - Is the information free of advertising?
   - Is the information trying to alter or sway your opinion?
   - On commercial sites, is the information biased towards the company's products?
4. Currency
   - Is the information up-to-date?
   - Is it clear when the information was published?
5. Coverage
   - Is the information complete or is it still under construction?
   - What topics are covered and are they explored in depth?
   - Is this the entire work or is there a more detailed version?

HSC style question

Surroundpix is a business that creates virtual tours on behalf of real estate agents and their clients. A screen shot of a final virtual tour web page is reproduced below.

The process of creating a virtual tour includes:

I Taking photographs.
   - Still images are taken by a Surroundpix photographer with a high quality 6 mega pixel camera.
   - Each SurroundPix sequence requires 12 web-quality still images to be taken by rotating a wide angled lens camera 30 degrees between each frame.

II Image processing.
   - All images are digitally processed to enhance their appearance.
   - Still images are stored in two formats – one suitable for commercial printing and another version suitable for use on the web.

- Surroundpix files are created using proprietary software that stitches the 12 images together into a complete 360-degree continuous image.

III Website production.

- Surroundpix files and still images (in both print and web format) are added to the database located on Surroundpix's web server together with other text data about the property.
- An email is sent to the real estate agent that includes two URLs. One URL enables download of the high resolution images and the other is a direct link to the completed virtual tour.

Consider the Surroundpix files and still images. Discuss relevant issues regarding:

(a) ownership of the data

(b) control of the data

(c) accuracy of the data.

Suggested Solution

(a) Ownership – The property is owned by the vendor, however the photographs were taken by an employee of Surroundpix. Also it is likely that the real estate agent has paid Surroundpix meaning that they too may have a claim in regard to ownership. There needs to be a clear contract between the three parties that specifies who retains ownership of the images.

(b) Control – Surroundpix has control of the Surroundpix files as they require their application to be viewed. The still images are available for use by the agent (and vendor) hence they have control over how these images are used.

(c) Accuracy – Surroundpix enhances all the images, which may result in them not reflecting the true state of the property. Also the photographer uses a wide angled lens which means the Surroundpix files will show a slightly distorted view of the property. As a consequence potential buyers may be mislead.

CURRENT AND EMERGING TRENDS

Data warehouses

A data warehouse is a database that includes copies of data from each of an organisation's operational databases. The data warehouse is used by various other systems as they analyse the activities of the organisation. The aim is to provide evidence to assist decision makers improve the organisation's performance.

Typically a data warehouse will include financial, sales, marketing, staffing and customer data gathered over an extended period of time. The data from each of these source databases is uploaded to the data warehouse at regular intervals – commonly weekly or monthly. Hence data warehouses have the added benefit of providing a backup and archival role.

Data Warehouse
A large separate combined copy of different databases used by an organisation. It includes historical data, which is used to analyse the activities of the organisation.

Generally data warehouses, once loaded with data, are read only. As the data does not change many of the problems present in operational systems simply are not present. For example data redundancy is not an issue, as records are never updated. Backup is simple as new records are uploaded on mass rather than appearing over time. The DBMS software used to access data warehouses never needs to monitor who is accessing records, as they are never changed.

The downside is the enormous size of data warehouses compared to operational systems. Typically a data warehouse is much larger than the total size of operational databases – often 300 or 400 times the size. Operational systems are purged of old data on a regular basis – this old data is food for the data warehouse.

At the time of writing (2007) Wal-Mart – an American retail chain – claims to have the largest data warehouses in the world. Wal-Mart uses their data warehouse for many significant business decisions such as identifying the most profitable locations for new stores, determining under performing products and stores and optimising the amount of stock of each product held at each store. The Wal-Mart data warehouse contains a record of every scanned item purchased in any of its thousands of stores.

GROUP TASK Research
Use the Internet to identify an organisation that maintains a data warehouse. Describe the data held in this data warehouse.

Data mining

The aim of data mining is to discover new knowledge through the exploration of data. Data mining is a form of detailed data analysis used on large databases – usually data warehouses. Data mining is performed by software that uses various different strategies in an attempt to uncover patterns that are non-obvious within the data. The uncovered patterns are usually predictive, that is they predict some future behaviour based on past trends.

Data Mining
The process of discovering non-obvious patterns within large collections of data.

The phrase 'non-obvious pattern' is critical to an understanding of data mining – it is what separates data mining from simply querying a database or performing statistical analysis. Obvious patterns are those that someone thinks up – much like a scientific theory. The theory is postulated and then evidence is gathered in an attempt to support or disprove the theory. If enough evidence supports the theory then it becomes accepted knowledge. This is not how data mining works, indeed data mining uses quite the opposite strategy – it searches for evidence that leads to a theory. The evidence is the non-obvious patterns.

So what is a non-obvious pattern? It is a pattern detected automatically by the data mining software. The data mining software explores the data searching for relationships that were never planned. These relationships are rarely definite nor do they necessarily apply to all the data, rather they tend to be general trends known as patterns. For example, data mining software may detect a pattern indicating that men tend to buy luxury food items early in the week if they are also purchasing nappies. The pattern detected is the evidence and what it indicates is the theory. The management of grocery stores could exploit this knowledge by placing luxury food items between the baby products and the checkout or they could choose not to discount luxury food items early in the week. Some patterns or trends uncovered using data mining may be completely coincidental or they may have no real world significance. It is up to those who make decisions to assess the relevance of such patterns before making critical business decisions based on them.

Data mining software uses complex statistical and artificial intelligence techniques together with a lot of hard work. Many of the techniques used are not new but the hard work is. The hard work is the processing of enormous amounts of data. This has only become feasible due to the increased performance of computer technology. If your class studies the decision support systems option, you will study some of the strategies used by data mining systems.

GROUP TASK Research
Use the Internet to identify an organisation that uses data mining. Describe examples of knowledge acquired from their data mining system.

Online Analytical Processing (OLAP)

Online analytical processing is a technique for providing business decision makers with statistical evidence, largely based on past trends, upon which they can make intelligent decisions. This is also a primary aim of data mining, however OLAP aims to provide this critical information visually, online, as needed and as quickly as possible. Furthermore the analysis can be performed by the decision maker from the comfort of his or her own office. To do this OLAP tools must optimise the organisation of large data stores so that decision makers can simply and quickly get answers to assist them to make decisions.

OLAP tools organise and combine multiple databases into multidimensional structures known as data cubes. The word cube implies three dimensions, however many more dimensions are possible and are routinely used. So what do we mean by dimensions? A relational database is composed of tables where each table has just two dimensions – records and fields. An OLAP tool organises the data into many more dimensions. Users are able to focus on a single attribute and then drill down through further dimensions to uncover new relationships and details. For example, the data cube in *Fig 2.101* includes three dimensions, namely products, customers and dates. When performing analysis, one could commence by studying a chart displaying a product or group of products. It is then possible to click on one product to drill down and explore characteristics of customers who purchased the product. Perhaps displaying graphs concerning their age or statistics with regard to their location. Drilling down further would uncover trends in regard to when these customers purchased the product and even other products they purchased at the same time. The multidimensional organisation of the data allows such visual aids to be generated in seconds. Examples of OLAP are studied in the decision support systems option topic (page 472).

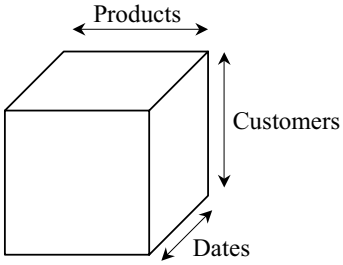


Fig 2.101
Example OLAP data cube

Online Transaction Processing (OLTP)

Databases that allow transactions to be processed immediately by remote users are known as online transaction processing systems. A transaction is a sequence of operations that must all complete successfully or all fail. Completing an online purchase over the Internet is a common example of a transaction performed by an OLTP system. Such financial transactions must occur in close to real time. Funds are moved from one bank account to another and then the purchase is confirmed. Both purchaser and seller must be confident that all operations within the transaction have completed successfully.

OLTP commonly involves many different database systems communicating to complete a single transaction. For instance, in our Internet purchasing example, there will be at least two banks, in addition to the sellers system and probably some other system, such as PayPal, that manages the overall approval process. Many OLTP systems include transaction monitoring software, whose central task is to direct, manage and control operations as they are performed by the various systems involved in each transaction. We examine OLTP in more detail within the transaction processing systems option.

SET 2J

1. Distributing original material without permission from the author is an infringement of:
   (A) Copyright Law.
   (B) FOI Law.
   (C) Privacy Law.
   (D) Criminal Law.

2. Your local ISP office is broken into and the web server hosting your web site is stolen. Which law would most likely to be used as the basis for charging the thief?
   (A) Copyright Law.
   (B) FOI Law.
   (C) Privacy Law.
   (D) Criminal Law.

3. The ten NPPs are contained within the:
   (A) Privacy and Personal Protection Act 1998.
   (B) Copyright Act 1968.
   (C) Freedom of Information Act 1982.
   (D) Privacy Act 1988 (Cth).

4. The integrity of data is a measure of its:
   (A) accuracy.
   (B) quality.
   (C) correctness.
   (D) All of the above.

5. Which of the following is NOT a breach of Privacy law?
   (A) Storing and reusing personal information contained on a sales order on future sales orders.
   (B) Selling private information without consent.
   (C) Using private information for unintended purposes without permission.
   (D) Refusing to explain why personal information is held.

6. A mail out that includes a change of address form on the back is an example of a:
   (A) data validation check.
   (B) data integrity check.
   (C) data verification check.
   (D) data entry check.

7. You discover a private health company is using your Medicare number to link your medical and financial records. The company is breaching:
   (A) Copyright Law.
   (B) FOI Law.
   (C) Privacy Law.
   (D) Criminal Law.

8. Which laws require government organisations to make available certain personal information such as health and superannuation records?
   (A) Copyright Laws.
   (B) FOI Laws.
   (C) Privacy Laws.
   (D) Criminal Laws.

9. Which of the following best describes the data verification process?
   (A) Checks to ensure data is within a suitable range.
   (B) The correctness and quality of the data.
   (C) Updating records to reflect changes over time.
   (D) Checks to ensure the data in the system matches and continues to match its source.

10. Government policy documents are made available to the general public as a consequence of:
    (A) Copyright Law.
    (B) FOI Law.
    (C) Privacy Law.
    (D) Criminal Law.

11. Define each of the following terms:
    (a) Data integrity
    (b) Data verification
    (c) Data warehouse
    (d) Data mining
    (e) OLAP
    (f) OLTP

12. Compare and contrast Freedom of Information Law with Privacy Law.

13. Virtually all businesses maintain a database containing personal information on each of their customers. Outline what this information is and how it can legally be used.

14. Explain how the accuracy and reliability of information obtained via the Internet can be assessed.

15. (a) Create a list detailing the main features of most data warehouses.
    (b) Compare creating and executing SQL queries with the techniques of data mining.
    (c) Explain the purpose of OLAP.
    (d) List and briefly describe OLTP systems you have used in the last month.

CHAPTER 2 REVIEW

1. What is the main purpose of all information systems?
   (A) To process data into information.
   (B) To meet the needs for whom the system is created.
   (C) To store data securely and efficiently.
   (D) To manage the system's resources so that the integrity of the data is maximised.

2. A complete database is composed of a single 2 dimensional table of data. This is an example of a:
   (A) flat-file database.
   (B) relational database.
   (C) hierarchical database.
   (D) network database.

3. With regard to field data types, which of the following is FALSE?
   (A) The data type determines the storage size.
   (B) Each attribute has a single data type.
   (C) Each record in a table has the same data type.
   (D) The data type determines how the field is formatted for display.

4. With regard to candidate keys, which of the following is FALSE?
   (A) They uniquely identify each record.
   (B) One is selected as the primary key.
   (C) Candidate keys are foreign keys.
   (D) Secondary keys are candidate keys.

5. Seven records in one table are linked to five records in another table. In a relational database how can this be achieved?
   (A) The primary key in each table is included as a foreign key in the other table.
   (B) A new table is created that includes all the records from both tables.
   (C) A new table is created that includes the primary key field from both tables as foreign keys.
   (D) The data from each table is included in the other table.

6. The process that ensures each foreign key always matches a primary key is called:
   (A) data validation.
   (B) referential integrity.
   (C) normalisation.
   (D) functional dependency.

7. New tables are not created when normalising to which normal form?
   (A) 1NF
   (B) 2NF
   (C) 3NF
   (D) All of the above.

8. Which of the following HTML tags links to the image lake.jpg on the www.pedc.com.au website?
   (A) <A HREF="http://www.pedc.com.au /lake.jpg">Lake</A>
   (B) <A HREF="lake.jpg"> www.pedc.com.au</A>
   (C) <A HREF="pedc.com.au/lake.jpg"> Lake</A>
   (D) <A HREF=Lake>http://www.pedc.com.au /lake.jpg</A>

9. Which of the following is TRUE of an incremental backup?
   (A) All the files altered or created since the last full backup are copied.
   (B) Only files with their archive bit set to false are copied.
   (C) Only files altered or created since the last incremental or full backup are copied.
   (D) Every file is backed up.

10. Which of the following is TRUE for a database that is horizontally fragmented?
    (A) It is a distributed database.
    (B) It is a relational database.
    (C) Different records are held at different locations.
    (D) All of the above.

11. Describe the organisation of data within:
    (a) Flat-file databases (b) Relational databases (c) Hypertext webs.

12. Distinguish between text and numeric data types in terms of:
    (a) sorting (b) representation (c) storage size

13. Compare and contrast each of the following:
    (a) Data validation with data verification. (b) Data warehouses with data mining.

14. Discuss reasons why relational databases are normalised.

15. Describe the principles of the operation of a search engine.

In this chapter you will learn to:

- use applications to create and transmit messages
- establish a communications link and describe the steps that take place in its establishment
- identify and describe specified protocols at different stages of the communication
- identify client processing and server processing
- describe the advantages and disadvantages of client – server architecture
- use a communication system to transmit and receive audio, video and text data
- for given examples, identify the participants, information/data, information technology, need and purpose
- for given examples explain how data is transmitted and received
- for given examples, identify the advantages and disadvantages of the system
- compare and contrast traditional communication systems with current electronic methods
- represent a communication system diagrammatically
- predict developments in communication systems based on current trends
- simulate activities involved with communication in areas such as
  - e commerce
  - EFTPOS
  - Internet banking
- for a given scenario, choose and justify the most appropriate transmission media
- diagrammatically represent the topology
- describe the location and role of hardware components on the network
- compare the functions of different hardware components
- identify the main characteristics of network operating software
- compare and contrast the Internet, intranets and extranets
- distinguish between data in analog and digital form
- justify the need to encode and decode data
- identify where in a communication system signal conversion takes place
- describe the structure of a data packet
- describe methods to check the accuracy of data being transmitted
- detail the network management software in a given network
- describe the role of the network administrator and conduct network administration tasks
- demonstrate logon and logoff procedures, and justify their use
- adopt procedures to manage electronic mail
- describe and justify the need for ethical behaviour when using the Internet
- discuss the social and ethical issues that have arisen from use of the Internet, including:
  - the availability of material normally restricted
  - electronic commerce
  - domination of content and control of access to the Internet
  - the changing nature of social interactions
- identify the issues associated with the use of communication systems including:
  - teleconferencing systems
  - messaging systems
  - e commerce
  - EFTPOS
  - electronic banking
- design and implement a communication system to meet an individual need
- predict developments in communication systems based on current trends

Which will make you more able to:

- apply and explain an understanding of the nature and function of information technologies to a specific practical situation
- explain and justify the way in which information systems relate to information processes in a specific context
- analyse and describe a system in terms of the information processes involved
- develop solutions for an identified need which address all of the information processes
- evaluate and discuss the effect of information systems on the individual, society and the environment
- demonstrate and explain ethical practice in the use of information systems, technologies and processes
- propose and justify ways in which information systems will meet emerging needs
- justify the selection and use of appropriate resources and tools to effectively develop and manage projects
- assess the ethical implications of selecting and using specific resources and tools, recommends and justifies the choices
- analyse situations, identify a need and develop solutions
- select and apply a methodical approach to planning, designing or implementing a solution
- implement effective management techniques
- use methods to thoroughly document the development of individual or team projects

In this chapter you will learn about:

Characteristics of communication systems

- communication systems as being those systems which enable users to send and receive data and information
- the framework in which communication systems function, demonstrated by the *Fig 3.1* model
- the functions performed within the communication systems in passing messages between source and destination, including:
  - message creation
  - organisation of packets at the interface between source and transmitter
  - signal generation by the transmitter
  - transmission
  - synchronising the exchange
  - addressing and routing
  - error detection and correction
  - security and management
- the roles of protocols in communication
  - handshaking and its importance in a communications link
  - functions performed by protocols at different levels
- the client - server model
  - the role of the client and the server
  - thin clients and fat clients
  - examples of clients such as web browsers and mail clients
  - examples of servers such as print servers, mail servers and web server

Examples of communication systems

- teleconferencing systems
- messaging systems, including email, voice mail and voice over Internet protocol (VOIP)
- other systems dependent on communication technology such as:
  - e commerce
  - EFTPOS
  - electronic banking

Transmitting and receiving in communication systems

- transmission media including:
  - wired transmission, including twisted pair, coaxial cable and optic fibre
  - wireless transmission, including microwave, satellite, radio and infrared
- characteristics of media in terms of speed, capacity, cost and security
- communication protocols, including:
  - application level protocols, including http, smtp and SSL
  - communication control and addressing level protocols, including TCP and IP
  - transmission level protocols, including Ethernet and Token ring
- strategies for error detection and error correction
- network topologies, including star, bus, ring, hybrid and wireless networks
- the functions performed by hardware components in communication systems including
  - hubs and switches
  - routers
  - modems
  - bridges and gateways
  - network interface cards (NIC)
  - mobile phones
  - cable
  - wireless access points
  - Bluetooth devices
- characteristics of network operating software
- the similarities and differences between the Internet, intranets and extranets

Other information processes in communication systems

- collecting, such as
  - the phone as the collection device with voice mail
  - EFTPOS terminal as a collection device for electronic banking
- processing, including
  - encoding and decoding analog and digital signals
  - formation of data packets
  - routing
  - encryption and decryption
  - error checking
    - parity bit check
    - check sum
    - cycle redundancy check
- displaying, such as
  - the phone as the display device with voice mail
  - EFTPOS terminal as a display device for electronic banking

Managing communication systems

- network administration tasks, such as:
  - adding/removing users
  - assigning users to printers
  - giving users file access rights
  - installation of software and sharing with users
  - client installation and protocol assignment
  - logon and logoff procedures
  - network based applications

Issues related to communication systems

- security
- globalisation
- changing nature of work
- interpersonal relationships
- e crime
- legal
- virtual communities
- current and emerging trends in communications, including
  - blogs
  - wikis
  - RSS feeds
  - podcasts
  - online radio, TV and video on demand
  - 3G technologies for mobile communications

3

COMMUNICATION SYSTEMS

Communication systems enable people and systems to share and exchange data and information electronically. This communication occurs between transmitting and receiving hardware and software over a network – each device on a network is called a node. Consider the diagram in *Fig 3.1*. As each message leaves its source it is encoded into a form suitable for transmission along the communication medium, which could be a wired or wireless connection. During its travels, the message may follow a variety of different paths through many different networks and connection devices. Different types of connection device use different strategies to determine which path each message will follow – switches decide based on the MAC address, whilst routers use the IP address, for example. Eventually the message arrives at the receiver, who decodes the message as it arrives at its destination. The network could be a local area network (LAN), a wide area network (WAN), it could be the Internet, an intranet, extranet or any combination of network types.

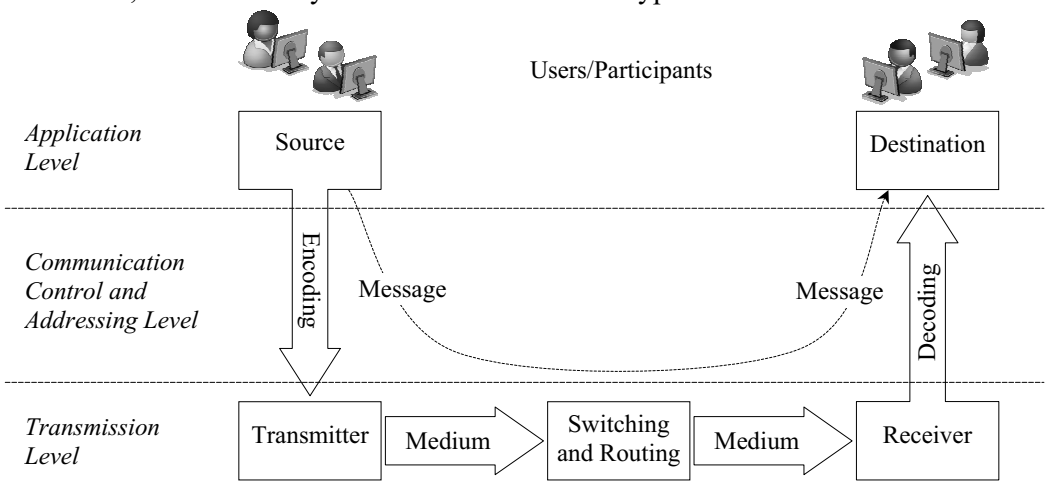


Fig 3.1
Communication system framework from NSW Board of Studies IPT syllabus (modified).

For communication to be successful requires components to agree on a set of rules known as protocols. Establishing and agreeing on which set of protocols will be used and the specific detail of each protocol must occur before any data can be transmitted or received – this process is known as handshaking. Protocols are classified according to the level or layer in which they operate. In the IPT course we classify protocols into three levels, namely; Application Level, Communication Control and Addressing Level, and Transmission Level (refer *Fig 3.1*). As messages pass through the interface between sender and transmitter they are encoded, meaning they descend the stack of protocols and are finally transmitted – each message is progressively encoded using the protocol (or protocols) operating at each level. Conversely, as messages are received they pass through the interface between receiver and destination – the original message is decoded by each protocol in turn as it ascends through each level of the protocol stack.

In the IPT syllabus three levels of protocols are defined; this framework provides a simplified view of the more detailed OSI (Open Systems Interconnection) model. The OSI model defines seven layers, where each layer can be further expanded into sub-

layers. Layers specified within the OSI model are combined to form the levels of the IPT model as shown in *Fig 3.2*. In IPT the OSI Presentation and Application layers (layer 6 and 7) are combined to form the IPT Application Level. OSI layers 3, 4 and 5, the network, transport and session layers are combined to form the IPT Communication Control and Addressing Level. Finally, protocols operating within the Physical and Data link layers (layer 1 and 2) of the OSI model are included in the IPT Transmission level. Throughout this chapter we focus on the IPT version with reference to the OSI model when appropriate.

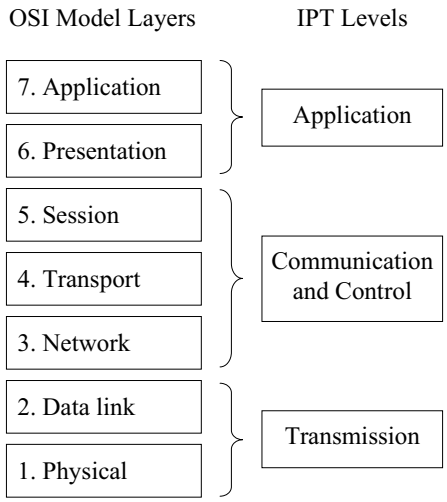


Fig 3.2
Comparison of the seven layers of the OSI model with the three levels used in IPT.

Note that in most cases communication occurs in both directions, even when the actual message only travels in one direction. The receiver transmits data back to the transmitter including data to acknowledge receipt, request more data or to ask for the data to be resent should it not be received correctly. The details of such exchanges are specific to the particular protocol being used.

In this chapter we consider:

- Characteristics of communication systems, including an overview of each protocol level based on the OSI model, details of how messages pass from source to destination, examples of protocols operating at each level, measurements of transmission speed and common error checking methods.
- Examples of communication systems including teleconferencing, messaging systems and financial systems.
- Network communication concepts including client-server architecture, network physical and logical topologies and methods for encoding and decoding digital and analog data.
- Network hardware including transmission media, network hardware devices such as hubs, switches and routers, and also servers such as file, print, email and web servers.
- Software to control networks including network operating software, network administration tasks and other network-based applications.
- Finally we consider issues related to communication systems and current and emerging trends in communication.

Consider the following examples of communication:

1. A conversation with a young child.
2. Sending a birthday card to your grandmother.
3. Watching television.
4. Ordering a meal in a restaurant.

GROUP TASK Discussion
For each example, identify the source, destination and medium over which messages are sent. Describe suitable communication rules (protocols).

CHARACTERISTICS OF COMMUNICATION SYSTEMS

Before we examine the details of particular examples of communication systems it is worthwhile understanding some communication concepts and terminology common to most communication system. The knowledge gained in this section underpins much of the work covered in the remainder of this chapter.

OVERVIEW OF PROTOCOL LEVELS

Software is used to control and direct the operation of hardware. The transmitter and the receiver must agree on how the hardware will be used to transfer messages. This is not a simple matter, a large variety of applications transfer data using a wide variety of operating systems, protocols, devices and transmission media. In 1978 a set of standards was first developed by the International Standards Organisation (ISO) to address such issues. These standards are known as the *Seven-Layer Model for Open Systems Interconnection* or more simply as the *OSI Model*. This seven-layer model has been largely accepted and used by network engineers when creating all types of transmission hardware and software.

The hardware actually used for transmission resides within the IPT Transmission Level, which includes the physical layer of the OSI Model. The physical layer includes NICs, hubs and the various different types of physical and wireless transmission media. These components actually move the data from the transmitter to the receiver. How they do this is determined by the higher software layers. Each layer performs its functions with data from the layer above during transmitting and the layer below during receiving.

The seven layers of the OSI model are referred to as the OSI stack. Each packet of data must descend the stack, be transmitted and then ascend the stack on the receiving computer. A brief explanation of the general tasks performed at each of the OSI layers and IPT levels follows. To avoid confusion between IPT levels and OSI layers we will always refer to the IPT syllabus levels as "IPT… Level" and OSI layers as "OSI… Layer".

IPT Presentation Level

7. OSI Application Layer – The actual data to be transmitted is created by a software application, this data is organised in a format understood by the application that will receive the data.
6. OSI Presentation Layer – The data is reorganised into a form suitable for subsequent transmission. For example, compressing an image and then representing it as a sequence of ASCII characters suited to the operating system. The presentation layer is commonly part of the application or is executed directly by the application and is often related to the requirements of the operating system. Protocols operating at this level include HTTP, DNS, FTP, SMTP, POP, IMAP and SSL.

IPT Communication Control and Addressing Level

5 OSI Session Layer – This is where communication with the network is established, commences and is maintained. It determines when a communication session is started with a remote computer and also when it ends. For example, when performing an Internet banking transaction it is the session layer that ensures communication continues until the entire transaction is completed. Layer 5 also includes security to ensure a user has the appropriate access rights.

4. OSI Transport Layer – The transport layer manages the correct transmission of each packet of data. This layer ensures that packets failing to reach their

destination are retransmitted. For example, TCP (Transport Control Protocol) operates within layer 4. TCP is used on TCP/IP networks, such as the Internet, to ensure the correct delivery of each data packet actually occurs.

3. OSI Network Layer – This is where packets are directed to their destination. IP (Internet Protocol) operates here – its job is to address and forward packets to their destination. There is no attempt to check each packet actually arrives. Routers also operate at this layer by directing packets along the best path based on their IP address. Routers often have their software stored in flash memory and can be configured remotely from an attached computer.

IPT Transmission Level

MAC Address
Media Access Controller Address hardwired into each device. A hardware address that uniquely identifies each node on a network.

2. OSI Data Link Layer – This layer defines how the transmission media is actually shared. Device drivers that control the physical transmission hardware operate at this layer. They determine the final size of transmitted packets, the speed of transfer, and various other physical characteristics of the transfer. Switches and the Ethernet protocol operate at this level, directing messages based on their destination MAC (Media Access Controller) address. Other data link protocols include Token Ring, SONET and FDDI.
1. OSI Physical Layer – This layer performs the actual physical transfer, hence it is composed solely of hardware. It converts the bits in each message into the signals that are transmitted down the transmission media. The transmission media could be twisted pair within a LAN, copper telephone cable in an ADSL connection, coaxial cable, optical fibre or even a wireless connection.

OVERVIEW OF HOW MESSAGES ARE PASSED BETWEEN SOURCE AND DESTINATION

In this section we explain the general processes occurring from when a message is first created at the source until it arrives at its final destination. Most of the points made here will be expanded and elaborated upon throughout the remainder of this chapter. The intention of this overview is to explain how all the different processes and information technology we will study fit together to form a logical operational communication system. It may be worthwhile rereading this overview as you work through this chapter to help explain where each new area of study fits within the overall communication process.

Message creation

The message is compiled at the source in preparation for sending. This takes place using some type of software application and perhaps involves the collection of message data from one of the system's users or participants.

Some examples of message creation include:

- A user writing an email using an email client such as Outlook.
- A web server retrieving requested HTML files from secondary storage in preparation for transmission to a web browser.
- A DBMS server extracting records from a database for transmission to a client application.
- Speaking during a VOIP (Voice Over Internet Protocol) phone conversation.
- Pressing the delete key to remove a file stored on a file server.

GROUP TASK Discussion
Brainstorm other examples where messages are created in preparation for transmission. In each case identify the software used to create the message.

Organisation of packets at the interface between source and transmitter

In general, when a message is being prepared for transmission it descends the stack of protocols from the Application Level down to where it is ready for physical transmission by the hardware operating at the Transmission Level. Each protocol wraps the data packet (or frame or segment – different names are used depending on the particular protocol) from the layer above with its own header and trailer. The header and trailer contain data relevant to the protocol operating at that layer. The protocol operating within the next lower layer considers each entire packet from the prior layer to be data and adds its own header and trailer (refer *Fig 3.3*). Hence the protocols within each layer are applied independently of the protocols operating in other layers. Some protocols include the address of the receiver within the header and many include some form of error detection code within their header or trailer.

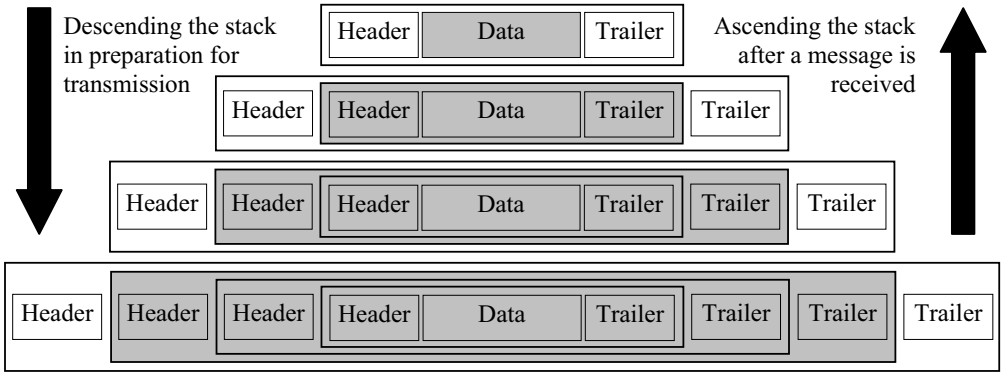


Fig 3.3
Descending and ascending the stack occurs during transmitting and receiving respectively.

Fig 3.3 implies each layer is creating a single data packet from the packet passed from the preceding layer. This need not be the case; usually multiple packets are created based on the requirements of the individual protocol being applied.

Let us work through a typical example. The software application, perhaps after direction from a user, first initiates the processes required to prepare the message for transmission. Essentially commands that include the message are issued to the protocol operating at the Application Level. For instance, to send an email message the email client software issues SMTP commands that include the recipient's email address and the content of the email message. To request a web page a web browser issues an HTTP command that includes the URL of the requested page. At this level we still have a single complete message. Furthermore the Application Level protocol is part of the software application; hence at this stage all processing has been performed by the same software that created the message.

Next the message is passed on to the Communication Control and Addressing Level. Commonly two or more protocols are involved, for example TCP in the OSI Transport Layer and then IP within the OSI Network Layer. Protocols operating at this level operate under the control of the operating system. They are not part of individual software applications, rather they are installed and managed by the operating system. The Communication Control and Addressing Level ensures packets reach their destination correctly. They include error checks, flow control and also the

source and destination address. Imagine the data packet has been passed to TCP. If the packet is longer than 536 bytes then TCP splits it into segments. The header within each segment includes a checksum and also information used by IP. TCP creates a connection between the source and destination that is used to control the flow and correct delivery of all segments within the total message. As each TCP segment is produced it is passed on to IP – TCP requires that IP be used. IP is the protocol that routes data across the network to its destination. IP packets are known as datagrams. During transmission routers determine where to send each datagram based on the destination IP address. The final Communication Control and Addressing protocol passes each packet to the Transmission Level protocol(s) that operates in conjunction with the physical transmission hardware.

At the receiving end the processes described above are essentially reversed – each protocol strips off its header and trailer, performs any error checks, and passes the data packet up to the next protocol. The specifics of different protocols are described in detail later in this chapter.

Consider the following:

TCP/IP is actually a collection of many protocols operating above layer 2 of the OSI model. As TCP/IP does not include data link (layer 2) and physical (layer 1) protocols it is able to operate across almost any type of communication hardware. This is the central reason why TCP/IP is so suited to the transfer of data and information over the Internet.

The suite of TCP/IP protocols does not precisely mirror the seven layers of the OSI model. Commonly layers 5, 6 and 7 are combined in TCP/IP references and are collectively called the application layer. Layer 4 remains as the transport layer and layer 3 is renamed as the Internet layer.

GROUP TASK Discussion
Explain why Transmission Level protocols (layer 1 and 2 of the OSI model) do not form part of the TCP/IP protocols. How does this assist TCP/IP to operate across almost any network?

GROUP TASK Research
Using the Internet, or otherwise, determine TCP/IP protocols operating within the application, transport and Internet layers mentioned above.

Signal generation by the transmitter

The transmitter is the physical hardware that generates or encodes the data onto the medium creating a signal. In most cases transmitters and receivers are contained within the same hardware device – receivers decode the signal on the medium. This hardware is controlled by protocols operating at the Transmission Level. The main task of the transmitter is to represent individual bits or patterns of bits as a wave – this wave is the signal that is actually transmitted through the medium. For instance, on copper wires bits are represented by altering voltage, on optical fibres light waves are altered, and for wireless mediums radio waves, infrared waves or microwaves are altered. In all cases characteristics of some type of wave is altered by the transmitter. The rules of the Transmission Level protocol determine precisely which characteristics are altered. Some rules determine how each pattern of bits is encoded, others determine the speed of transmission and others are used to control and

synchronise the exchange. Examples of devices that include a transmitter (and also a receiver) include NICs, switches, routers, ADSL and cable modems, and even mobile phones and Bluetooth devices.

Transmission

Transmission occurs as the signal travels or propagates through the medium. Each bit or often pattern of bits moves from transmitter to receiver as a particular waveform. The transmitter creates each waveform and maintains it on the medium for a small period of time. Consider a Transmission protocol transmitting at 5Msym/s. This means the transmitter generates 5 million distinct symbols (wave forms representing bit patterns) every second. And it also means each distinct symbol is maintained on the medium by the transmitter for a period of one five millionth of a second. If each symbol represents 8-bits (1-byte) of data then one megabyte of data could potentially be transferred in one fifth of a second – as 1 million bytes requires 1 million symbols, and 5 million symbols can be transferred in one second. One fifth of a second is the time required for the physical transmission of one megabyte of binary data if the transmission occurs as a continuous stream of symbols and the transmitter and receiver are physically close together. In reality, data is split into packets, which are not sent continuously, errors occur that need to be corrected and some mediums exist over enormous distances – such as up to satellites or across oceans. Furthermore some protocols wait for acknowledgement from the receiver before they send the next data packet. This in itself has the potential to double transmission times – flow control is used by protocols to help overcome this problem.

Synchronising the exchange

To accurately decode the signal requires the receiver to sample the incoming signal using precisely the same timing used by the transmitter during encoding. This synchronising process ensures each symbol or waveform is detected by the receiver. If both transmitter and receiver use a common clock then transmission can take place in the knowledge that sampling is almost perfectly synchronised with transmitting. This is the most obvious method of achieving synchronous communication, for example the system clock is used during synchronous communication between components on the motherboard. Unfortunately, the use of a common clock is rarely a practical possibility when communication occurs outside of a single computer. As a consequence, other techniques must be used in an attempt to bring the receiver into synch with the transmitter.

Today synchronous transmission systems have almost completely replaced older asynchronous links, which transferred individual bytes separately using start and stop bits. Synchronous communication does not transfer bytes individually; rather it transfers larger data packets usually called frames. Frames vary in size depending upon the individual implementation. 10baseT Ethernet networks use a frame size of up to 1500 bytes and frame sizes in excess of 4000 bytes are common on high-speed dedicated links.

There are two elements commonly used to assist the synchronising process. A preamble can be included at the start of each frame whose purpose is initial synchronisation of the receive and transmit clocks. The second element is included or embedded within the data and is used to ensure synchronisation is maintained throughout transmission of each frame. Let us consider each of these elements.

Firstly each frame commences with a preamble. The Ethernet Transmission Level protocol uses an 8 bytes (64 bits) long preamble, which is simply a sequence of alternating 1s and 0s that end with a terminating pattern (commonly 1 1) called a frame delimiter. The receiver uses the preamble to adjust its clock to the correct phase

as the transmitting clock (see *Fig 3.4*). A frame delimiter is needed at the end of the preamble because the receiver may lose some bits during clock adjustment so these delimiting bits act as a flag indicating the start of the actual data.

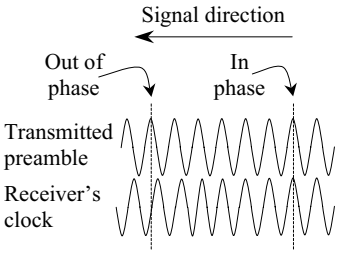


Fig 3.4
The preamble is used to synchronise the phase of the receiver's clock to match the transmitter's clock.

The preamble is followed by the data that needs to be received. The representation of the bits within the signal provides the second element used to maintain synchronisation. Commonly bits are represented not as high or low signals but using the transitions between these states. An example of such a system is Manchester Encoding used within 10baseT Ethernet networks. Using this system a low to high transition represents a 1 and a high to low transition represents a 0. As the clocks are initially synchronised then the location of the transitions representing the bits is known. The receiver detects each transition, if they are slightly out of synch then the receiving clock adjusts accordingly, hence Manchester Encoding is an example of a self-clocking code. As can be seen in *Fig 3.5*, two frequencies are needed to implement such a system; a base frequency and a frequency that is precisely double the base frequency. Data is transmitted at the same rate as the base frequency. For example 10baseT Ethernet transfers data at 10 megabits per second and therefore a base frequency of 10 mega hertz is used.

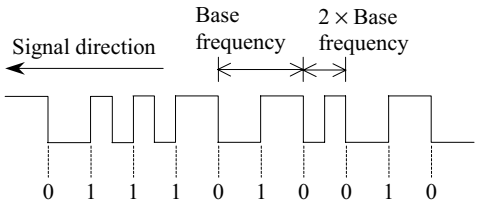


Fig 3.5
Manchester encoding uses the transitions between high and low to represent bits.

Other Transmission Level protocols use similar synchronisation strategies. For instance ADSL connections transmit superframes that contain many data frames. The header of the superframe contains synchronisation data much like the preamble of an Ethernet frame. Each data frame begins at equal and precisely spaced intervals.

Addressing and routing

During transmission data packets may pass through many different and varied links – particularly when the communication is over the Internet. Furthermore it is likely that packets forming part of a single file will travel over quite different paths from the transmitter to the receiver. Each new communication link will have its own protocol or set of protocols and hence each packet must ascend the protocol stack until it reaches the addressing or routing protocol and then descend the protocol stack as it is prepared for transmission down the next path.

Ethernet and other Transmission Level protocols use the receiver's MAC address to determine the path leading to the receiver. For instance an Ethernet switch maintains a table of all the MAC addresses of attached devices. Frames can therefore be directed down the precise connection that leads to the receiver. Most routers use the IP address within IP datagrams together with their own routing table to determine the next hop in a datagrams travels. The routing table is continually being updated to reflect the current state of attached networks and surrounding routers. Routers can therefore divert datagrams around faulty or poorly performing network connections.

GROUP TASK Research
Determine the protocols operating on either your own or your friend's home network. Explain how a message is sent using these protocols.

Error detection and correction

As messages descend the stack prior to transmission many protocols calculate checksums or CRC (Cyclic Redundancy Check) values and include them within their headers or footers. Once the message has been received it ascends the protocol stack, where each protocol examines its own received headers and trailers. If error detection is used by the protocol then the error check calculation is again performed to ensure the result matches the received checksum or CRC value. Whenever an error is detected virtually all protocols discard the entire packet and the sender will need to resend the packet to correct the problem. In general, CRCs are used within hardware operating within the Transmission Level, whilst checksums are used within many higher level protocols.

Clearly some strategy is needed so the sender can determine that an error was detected by the receiver and within which data packet the error occurred. Some protocols acknowledge only correct packets. This strategy is used by TCP and requires the sender to maintain a list of transmitted packets; as each acknowledgement arrives the associated packet is removed from the list. Packets remaining on the list for some specified period of time are resent. Within other protocols, such as Ethernet the receiver specifically requests packets to be resent each time an error is detected. There are specialised protocols that include self-correcting error detection codes – in this case some errors can be corrected at the destination without the need to resend the packet. Other protocols, such as IP, simply discard the message without any attempt to notify the sender.

GROUP TASK Discussion
Specialist systems, such as space probes, don't both with error correction; rather they send the whole message multiple times. Why is this strategy inappropriate for most communication systems? Discuss.

Security and management

Many protocols restrict messages based on user names and passwords, and others go a step further by encrypting messages during transmission. For example, POP (Post Office Protocol) operates on most mail servers. Top retrieve email messages from a POP server the user must first be authenticated – meaning a correct user name and password combination must be included. In this case the user name also identifies the mail box from which email messages are retrieved. SSL (or https) uses a public key encryption and decryption system to secure critical data transfers such as financial transactions. We explained encryption and decryption strategies in some detail within Chapter 2 and we will describe their implementation within the SSL protocol later in this chapter when we examine electronic banking.

GROUP TASK Discussion
Review the explanation of encryption and decryption in Chapter 2. Is encryption only used to secure messages during transmission? Discuss.

PROTOCOLS

Protocol
A formal set of rules and procedures that must be observed for two devices to transfer data efficiently and successfully.

There are literally thousands of different protocols that exist. Each protocol is designed to specify a particular set of rules and accomplish particular tasks. For example Ethernet is the most widespread Transmission Level protocol for the transfer of data between nodes on local

area networks, however Ethernet is not suitable for communication over wide area networks (WANs) carrying enormous amounts of data over long distances. Commonly such networks use protocols such as ATM (Asynchronous Transfer Mode) or SONET (Synchronous Optical Network) – ATM is used on most ADSL connections and SONET for connections between network access points (NAPs) that connect different cities and even continents. Ethernet, ATM and SONET all operate at the Transmission Level (OSI layer 1 and 2).

Before two devices can communicate they must first agree on the protocol or series of protocols they will utilise. This process is known as 'handshaking'. Handshaking commences when one device asks to communicate with another; the devices then exchange messages until they have agreed upon the rules that will be used. Depending on the protocol being used handshaking may occur just after the devices are powered up or it may occur prior to each communication session occurring.

Handshaking
The process of negotiating and establishing the rules of communication between two or more devices.

In IPT we study three common examples of Application Level protocols, namely http, smtp and SSL – we examine HTTP in this section, smtp later as we discuss email and SSL during our discussion on electronic banking. Two Communication Control and Adressing protocols are required, namely TCP and IP. We describe each of these in this section and as they are common to most of today's networks we expand on this discussion throughout the text. At the Transmission Level we need to cover Ethernet and also the token ring protocol. We deal with Ethernet in this section and token ring later in the chapter as we discuss the operation of ring topologies.

HTTP, TCP, IP and usually Ethernet all contribute during the transfer of web pages – these four protocols are described in this section.

Hypertext Transfer Protocol (HTTP)

HTTP operates within the IPT Application Level and within layer 6 of the OSI model. HTTP is the primary protocol used by web browsers to communicate and retrieve web pages from web servers. A client-server connection is used where the browser is the client and the web server is the server. There are three primary HTTP commands (or methods) used by browsers – GET, HEAD and POST.

The HTTP GET method retrieves entire documents – the documents retrieved could be HTML files, image files, video files or any other type of file. The browser requests a document from a particular web server using a GET command together with the URL (Universal Resource Locator) of the document. The web server responds by transmitting the document to the browser. The header, which precedes the file data, indicates the nature of the data in the file – the browser reads this header data to determine how it should display the data in the file that follows. For example if it is an HTML file then the browser will interpret and display the file based on its HTML tags.

The HTTP HEAD method retrieves just the header information for the file. This is commonly used to check if the file has been updated since the browser last retrieved the file. If the file has not been updated then there is no need to retrieve the entire file, rather the existing version held in the browser's cache can be displayed.

The HTTP POST method is used to send data from the browser to a web server. Commonly the POST method is used to send all the data input by users within web-based forms. For example many web sites require users to create an account. The users details are sent back to the web server using the HTTP POST method.

Consider the following:

Using a Telnet client it is possible to execute HTTP methods (or commands) directly. The following steps outline how to accomplish this task using a machine running current versions of Microsoft's Windows operating system.

1. Start a DOS command prompt by entering cmd at the run command located on the start menu.
2. From the command prompt start Telnet with a connection to the required domain on port 80. Port 80 is the standard HTTP port on most web servers. For example telnet www.microsoft.com 80 will initiate a connection to Microsoft.com.
3. Turn on local echo so you can see what you are typing. First type Ctrl+], then type set localecho and press enter. Press enter again on a blank line.
4. Type your HTTP GET or HEAD command, including the host name and then hit enter twice. For example GET /index.htm HTTP/1.1 then press enter, now type Host: www.microsoft.com and press enter twice. For GET commands the server will respond by sending the HTTP header followed by the document. For HEAD commands the server responds with just the HTTP header for the file. An example is shown below in *Fig 3.3*.

Fig 3.6
Screen dump of a Telnet session showing the HTTP HEAD method and the results for the file index.htm on the www.pedc.com.au domain.

GROUP TASK Practical Activity
Locate a simple web page using a web browser. Now use Telnet to retrieve the page using an HTTP GET command and then retrieve just the header using an HTTP HEAD command.

GROUP TASK Discussion
Discuss possible uses for the information contained within the HTTP headers returned by web servers.

Transmission Control Protocol (TCP)

TCP operates within the Communication Control and Addressing Level (Transport layer 4 of the OSI model). TCP, together with IP, are the protocols responsible for the transmission of most data across the Internet. The primary responsibility of transport layer protocols such as TCP is ensuring messages are actually delivered correctly.

Unlike most protocols that operate completely independently of their neighbouring protocols, TCP requires IP to be operating. TCP considers elements of the IP header – the reverse is not true, IP can operate without TCP, however for almost all implementations both TCP and IP are operating. This is why both TCP and IP are commonly referred to as TCP/IP.

In TCP terminology each packet is called a segment, where a segment includes a string of bytes forming part of the data to be sent. TCP includes checks for errors within each segment and also uses a system known as "sliding windows" to control the flow of data and ensure every byte of data is acknowledged once it has been successfully received. TCP is often called a "connection oriented" and "byte oriented" protocol as it maintains information about individual bytes transferred within a particular communication session.

Each TCP segment includes a header that includes the sequence of bytes contained within the segment and a checksum – we discuss the detail of checksums later in this section. The checksum is produced prior to the segment being sent. Upon arrival of each segment the checksum is recalculated to ensure it matches the checksum within the header. If it matches then the bytes received within the segment are acknowledged.

By default TCP segments contain a total of 576 bytes. This total includes 20 bytes for the TCP header and 20 bytes for the IP header, leaving 536 bytes for data. The sender in a TCP session continues sending segments of data up to the limit (window size) specified within acknowledgements from the receiver. Conceptually as subsequent segments are sent and received the window slides progressively along the length of the total message data, hence the name "sliding window". This flow control mechanism allows the receiver to adjust the rate of data it receives.

Consider the following:

Fig 3.7 below is a simplified conceptual view of the TCP sliding windows system at a particular point in time during a TCP communication session. In this diagram the "…cat sat on the mat…" text forms the complete message to be sent using multiple segments. Some data has been sent by the sender and acknowledged as correct by the receiver, some data has been sent but not yet acknowledged.

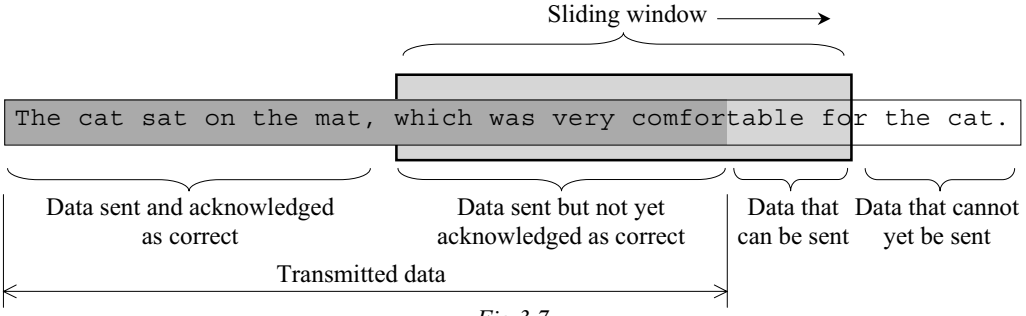


Fig 3.7
TCP uses a system known as "sliding windows" for flow control.

As the sender receives acknowledgements for transmitted segments the sliding window moves to the right. This movement enlarges the width of the "data that can be sent" region, hence the sender transmits more segments. Should segments fail to reach the receiver, contain errors or become delayed by network congestion then the window slides more slowly. When segments arrive quickly and without error the window slides more rapidly.

The receiver can adjust the width of the sliding window as part of their acknowledgement messages. A smaller window size slows the transmission whilst larger windows speed up the transmission.

GROUP TASK Discussion
Many other protocols wait for acknowledgement from the receiver before sending the next packet of data. Such systems are known as PAR or Positive Acknowledgement with Retransmission.
Discuss advantages of the sliding windows system over PAR systems, particularly with regard to communication over the Internet.

Internet Protocol (IP)

IP is the workhorse of the Internet. It is the protocol that causes data packets (called datagrams) to move from sender to receiver. The Internet Protocol operates at the OSI Network layer 3, which is called the Internet layer in references that specifically discuss TCP/IP. IP has been designed so it will operate with all types of networks and hardware. It was originally created so the different network systems used by the United States Army, Air Force and Navy could exchange and share data.

IP does not guarantee datagrams will reach their destination and it makes no attempt to acknowledge datagrams that have been received. Rather IP simply fires off each datagram one after the other. For these reasons IP is known as a connectionless protocol – as far as IP is concerned each datagram has no connection or relationship to any other datagram (unless fragmentation of a single datagram occurs). In essence IP cannot be relied upon to successfully transmit datagrams. At first this may seem to be a significant shortcoming of IP, however in reality it makes sense. For some data, such as streamed video, the speed of delivery is more important than its accuracy. Losing a single frame in a video sequence is unlikely to be even noticed; hence the significant overhead required for error checking is not needed. The only error check within an IP datagram is a checksum of the bytes within the header – no error checking is performed on the data. Note that TCP provides error checking in layer 4 and is used for data that must be delivered accurately. On the other hand the User Datagram Protocol (UDP) can be used in OSI layer 4 when speed is a higher priority than accuracy. Furthermore layer 2 data link protocols generally include robust error checks.

Where IP excels is in its ability to reroute messages over the most efficient path to their destination – using routers, which in turn utilise yet another protocol in the TCP/IP suite, ARP (Address Resolution Protocol), to determine the next hop for each datagram. Should a portion of the network fail then messages are automatically rerouted around the problem area. This was a requirement for the original designers of IP who needed to ensure communication between US defence sites would not be disrupted should individual sites be damaged during conflict. We discuss the operation of routers in more detail later in this chapter; at this stage we introduce IP addresses together with their underlying structure.

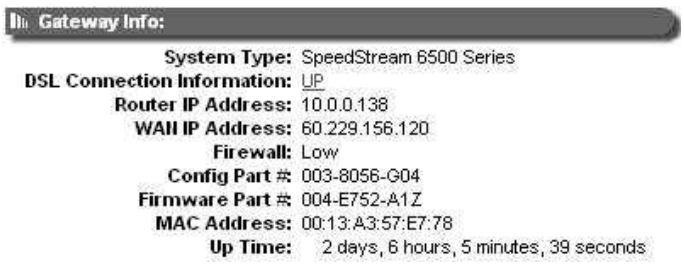


Fig 3.8
Router information showing the internal LAN IP address and the external WAN or Internet IP address.

Each IP address is composed of four bytes (a total of 32 bits). Every device on the Internet (or on any IP network) must have at least one unique IP address. Routers, and some other devices, require more than one IP address – one IP address for each network they are connected to. In *Fig 3.8* on the previous page the router's LAN IP address is 10.0.0.138 and its IP address on the Internet is 60.229.156.120. The header of every IP datagram includes the sender's IP address and the destination's IP address. Routers examine the destination IP address in the header of each IP datagram to determine which network connection they should use to retransmit the datagram.

Often IP addresses are expressed as dotted decimals, for example 140.123.54.67. Each of the four decimal numbers represents 8 bits; the IP address 140.123.54.67 is equivalent to the 32-bit IP address 10001100 01111011 00110110 01000011. Every IP address is composed of a network ID and a host ID. The network ID is a particular number of bits starting from the left hand side of the binary IP address, the remaining bits form the host ID. For example the IP address expressed as 140.123.54.67/24 means that the first 24 bits form the network ID and the remaining 8 bits form the host ID.

Network IDs form a hierarchical structure that splits larger networks into sub-networks, sub-sub networks, sub-sub-sub networks, etc. Sub networks lower in the hierarchy have longer network IDs, that is more bits in each IP address are used for the network ID, whilst sub networks higher in the hierarchy have shorter network IDs. It is the network ID that is used by IP (and routers) to determine the path a datagram takes to its destination. It is not until an IP datagram arrives at the router attached to the network matching the full destination network ID that the host ID part of the IP address is even considered. At this final delivery stage the host ID determines the individual destination device that receives the IP datagram.

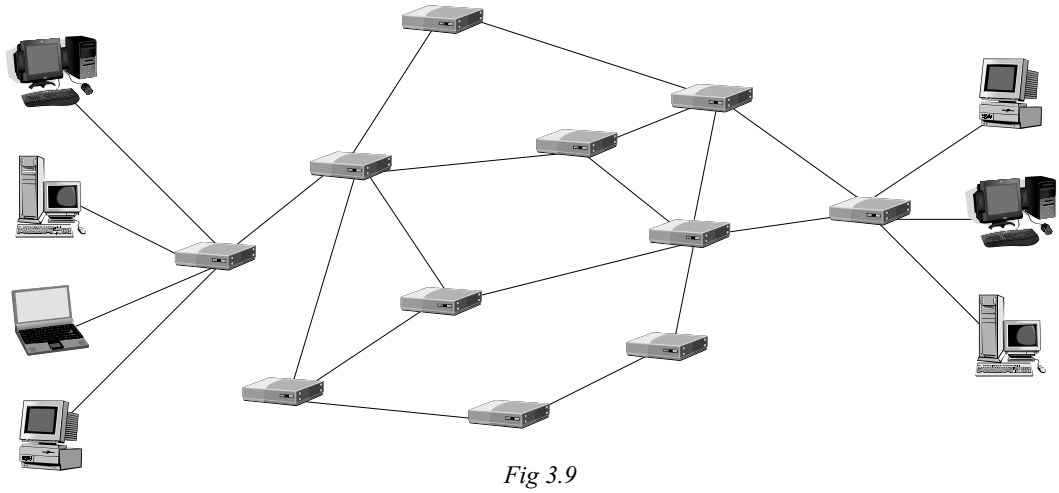


Fig 3.9
Each line between routers represents a possible network hop for IP datagrams and can potentially utilise a different data link protocol and different physical hardware.

During the transmission of an IP datagram across the Internet it is likely to pass through many varied network hops as it moves from router to router (see *Fig 3.9*). Each network hop potentially uses different hardware and a different protocol at the Transmission Level. The size of the frames physically transmitted differs depending on the OSI layer 2 protocol and also the hardware used at the physical layer. As a consequence the Internet protocol includes a mechanism known as 'fragmentation' to split complete datagrams into a series of smaller datagrams suited to the protocol operating at the OSI data link layer 2 of the current network hop.

The smaller IP datagrams created during fragmentation are not recombined until they reach their final destination. This means the size of fragments received is determined by the network hop with the smallest maximum frame size – known as the MTU or maximum transmission unit. It is preferable to avoid fragmentation and in most cases it is unnecessary as most OSI layer 2 data link protocols have MTU values significantly greater than TCP's default 576 byte segment size, for example Ethernet frames have an MTU of 1500 bytes.

Consider the following:

The header of each IP datagram is at least 20 bytes long and includes a 1-byte time to live (TTL) field. Each router encountered during the datagram's journey reduces the value of this field by one. If the TTL field is zero then the router discards the datagram. In fact any errors found within a datagram cause it to simply be discarded – no attempt is made to notify either the sender or the receiver.

GROUP TASK Discussion

Identify possible transmission problems where the TTL field will reduce to 0 and cause the datagram to be discarded. How will the sender and receiver become aware that an IP datagram has been discarded?

Ethernet

Ethernet operates at the IPT Transmission Level including OSI data link layer 2 and also at the OSI physical layer 1. Because Ethernet operates at the physical level it must be built into the various hardware devices used to transmit and receive. The term 'Ether' was proposed by the original Ethernet inventors Robert Metcalf and David Boggs to indicate that Ethernet can be applied to any medium – copper wire, optical fibre and even wireless mediums.

The original format and design details of Ethernet where first developed by Xerox in 1972 at their Palo Alto Research Centre in California. Digital, Intel and Xerox further developed the Ethernet standard in partnership and its current form is known as Ethernet II (DIX). The IEEE 802.3 committee formalised a slightly different Ethernet standard known as Ethernet 802.3. The differences between these two is not significant at our level of treatment.

Preamble (8 bytes)	Destination MAC Address (6 bytes)	Source MAC Address (6 bytes)	Type (2 bytes)	Data (46-1500 bytes)	CRC (4 bytes)

Fig 3.10
Ethernet II (DIX) frame format.

Ethernet packets are known as frames – *Fig 3.10* describes the format of an Ethernet II (DIX) frame. Packets of data from the Communication Control and Addressing Level form the data within each Ethernet frame. The length of the data must be between 46 and 1500 bytes. If the data is a default TCP/IP datagram then the TCP segment requires 576 bytes with an additional 20 bytes added for the IP header, therefore most IP datagrams require approximately 596 bytes – well below the 1500 MTU of Ethernet frames. The type field indicates the higher-layer protocol being used. In Ethernet 802.3 frames the type field is replaced by a field indicating the length of the data portion of the frame.

The preamble is a sequence of alternating zeros and ones and is used to synchronise the phase of the sender and receiver's clocks. In general, the ones and zeros within each frame are physically represented as transitions from high to low and low to high respectively. For these transitions to be accurately identified by the receiver requires the sender and receivers clock to be initially in phase with each other.

The MAC (Media Access Controller) address of both the sender and the receiver is included in the frame header. Every node on an Ethernet network must have its own unique 6-byte MAC address. For example the network interface card (NIC) on the computer I am currently using has the hexadecimal MAC address 00-00-E2-66-E3-CC as shown in *Fig 3.11*. Each node examines the destination MAC address of every Ethernet frame sent over their segment, if it matches their own MAC address then they accept the frame. If it does not match then the frame is simply ignored. Note that a node is any device attached to the network that is able to send and/or receive frames. For example *Fig 3.8* includes the MAC address 00-13-A3-57-E7-78 for the SpeedStream router.

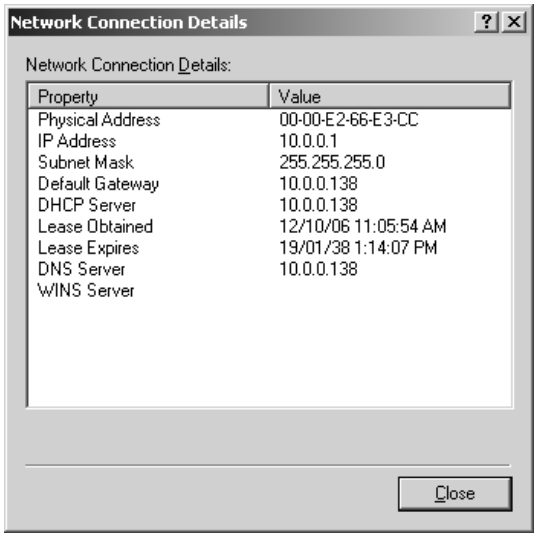


Fig 3.11
In Windows XP the physical address is equivalent to the MAC address of a computer's NIC.

The final 4-byte CRC of each Ethernet frame is used for error checking. Cyclic redundancy checks (CRCs) are a more accurate error checking technique than checksums. We examine CRCs in more detail later in this chapter. In general the sender calculates the CRC based on the contents of the frame. The receiver performs the same calculation and only accepts the frame if the two CRCs match. If the CRCs do not match then the receiver informs the sender so that the frame can be resent.

Using Ethernet it is possible for two nodes to transmit a frame at the same time. If these nodes share the same physical transmission line (i.e. are on the same segment) then a data collision will occur and both frames will be corrupted. Ethernet uses a system called Carrier Sense Multiple Access with Collision Detection (CSMA/CD) to deal with such collisions. Modern Ethernet networks prevent collisions altogether through the use of switches where just two nodes (including the switch) exist on each segment. We examine the operation of CSMA/CD and switches later in this chapter when we consider network topologies and network hardware.

Consider the following:

There are many different Ethernet standards that specify the speed of transmission together with details of the transmission medium used. For example 1000Base-T transfers data at up to 1000 megabits per second (1000Mbps) over twisted pair (Cat 5) cable. 1000Mb is equivalent to 1Gb, hence 1000Base-T is known as Gigabit Ethernet.

GROUP TASK Research
Using the Internet or otherwise identify the different Ethernet standards commonly used today.

SET 3A

1. During transmission data is represented using a:
   (A) transmitter
   (B) medium
   (C) message
   (D) wave

2. The MAC address is primarily used at which of the following layers of the OSI model?
   (A) network
   (B) transport
   (C) data link
   (D) presentation and application

3. Establishing and negotiating the rules for communication is the process known as:
   (A) handshaking.
   (B) protocol assignment.
   (C) sliding windows.
   (D) routing.

4. Which of the following is TRUE for all IP addresses?
   (A) They are transmitted as dotted decimals.
   (B) They always correspond to a unique domain name.
   (C) They are assigned by hardware manufacturers and cannot readily be changed.
   (D) They include a network ID and a host ID.

5. Why would an HTTP HEAD method be used?
   (A) To upload a new version of a file to a web server.
   (B) To determine if the user is permitted to download an HTML file.
   (C) To test the speed of a TCP/IP connection prior to download.
   (D) To determine if a file has been altered compared to the local cached version.

6. The system known as "sliding windows" is used to:
   (A) ensure TCP segments are acknowledged prior to further segments being sent.
   (B) monitor and record the destination of files sent from a web server.
   (C) adjust the speed of transmission during TCP sessions.
   (D) equitably share the bandwidth of communication channels.

7. In terms of the protocol stack, what occurs at the interface between source and transmitter?
   (A) Messages ascend the stack.
   (B) Messages descend the stack.
   (C) Messages are stripped of their headers and trailers.
   (D) Each protocol is influenced by the protocols operating at adjoining layers.

8. Data collisions, if possible, are detected by protocols operating at which layers of the OSI model?
   (A) Layers 1 and 2.
   (B) Layers 2 and 3.
   (C) Layers 3 and 4.
   (D) Layers 4 and 5.

9. As messages move across the Internet the protocols that change for each network hop would most likely operate at which level?
   (A) Transmission Level
   (B) Communication Control and Addressing Level
   (C) Application Level
   (D) Addressing and Routing Level

10. Which list includes only protocols that perform error checking?
    (A) TCP, IP.
    (B) Ethernet, TCP.
    (C) HTTP, UDP.
    (D) Ethernet, IP.

11. Define each of the following terms.
    (a) Protocol
    (b) Handshaking
    (c) IP Address
    (d) MAC Address

12. Explain what occurs as a message ascends the protocol stack.

13. IP does not guarantee delivery of datagrams. Is this a problem? Discuss.

14. TCP uses a flow control system known as "sliding windows". Outline the "sliding windows" process.

15. A particular router has a single MAC address but has many IP addresses. Why is this? Explain.

MEASUREMENTS OF SPEED

Bits per second (bps), baud rate and bandwidth are all measures commonly used to describe the speed of communication. Unfortunately many references use these terms incorrectly. The most common error is to use all three terms interchangeably to mean bits per second. In this section we consider the technical meaning of each of these measures, together with their relationship to each other.

Bits per second

Bits per second (bps)
The number of bits transferred each second. The speed of binary data transmission.

Bits per second is the rate at which binary digital data is transferred. For instance a speed of 2400bps, means 2400 binary digits can be transferred each second. Notice bps means bits per second not bytes per second. If a measure refers to bytes a capital B should be used, and if it refers to bits then a lower case b should be used; for example kB means kilobyte and kb means kilobit, similarly MB means megabyte whilst Mb means megabit. It is customary to refer to bits when describing transmission speeds.

Consider an Ethernet network based on the Fast Ethernet 100Base-T standard. This network is able to transfer data at a maximum speed of approximately 100Mbps. Now imagine we wish to transfer a 15MB video from one machine to another. 15MB = 15 × 8 Mb = 120Mb, therefore the transfer should take approximately 1.2 seconds. In reality the transfer will take significantly longer due to the overheads required to create the frames at the source and decode the frames at the destination. Also the headers and trailers added by each communication protocol involved have not been included in our calculation, yet they too must be transferred.

Baud rate

Baud (or baud rate)
The number of signal events occurring each second along a communication channel. Equivalent to the number of symbols per second.

Baud rate is a measure of the number of distinct signal events occurring each second along a communication channel. A signal event being a change in the transmission signal used to represent the data. Technically each of these signal events is called a baud, however commonly the term baud is used as a shortened form of the term baud rate.

Most modern communication systems represent multiple bits using a single signal event. For example, a connection could represent 2 bits within each baud by transmitting say +12 volts to represent the bits 11, +6 volts for 10, -6 volts for 01, and –12 volts for 00. If this connection were operating at 1200 baud then 2400bps could be transmitted. This example is trivial, in reality various complex systems are used where up to 4, 6, 8 or more bits are represented by each baud. In these situations different waveforms or symbols are needed to represent each bit pattern. The number of different symbols required doubles for each extra bit represented, for example to represent 4 bits requires $2^4 = 16$ different symbols whilst 5 bits requires

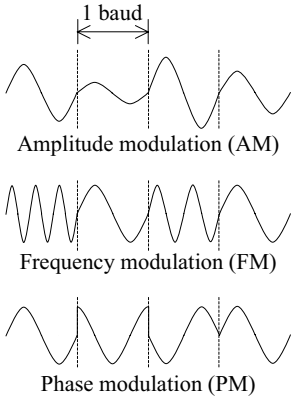


Fig 3.12
Examples of amplitude, frequency and phase modulation.

$2 \times 16 = 32$ different symbols. Altering or modulating the amplitude, frequency and/or phase of the signal produces these different symbols; *Fig 3.12* shows these modulation techniques separately. As most high-speed data communication is restricted to a particular range of frequencies, most encoding systems use a combination of amplitude and phase modulation.

Today few modern communication devices use the term Baud, rather they use the related measure symbols per second (sym/s). In most cases the speed is such that ksym/s and Msym/s is generally quoted within the specifications of these devices. Each symbol is transmitted as a distinct signal event; therefore the symbol rate is an equivalent measure to the Baud rate.

To calculate the time required to transfer a message of a certain size requires more than just the symbol or baud rate of the communication channel – it also requires the number of symbols represented by each distinct baud or signal event. For example a communication channel that uses 64QAM represents 6 bits within each symbol – 2^6=64 different symbols. If this channel is able to transfer 5Msym/s then it is able to communicate at a speed of 5×6 Mbps = 30Mbps. Say we wish to transfer a 15MB video over this communication channel, 15MB = 15×8 Mb = 120Mb. Therefore the minimum time for the transfer will be 120 / 30 = 4 seconds. Again significant overheads for transmitting and receiving processes, together with the various headers and trailers would increase this time significantly.

Consider the following

The time taken for each individual symbol to travel (or propagate) along the medium from the transmitter to the receiver can also affect transmission times. In regard to the transmission of individual data packets this is relatively insignificant. It only becomes significant over longer distances, particularly when each data packet must be acknowledged before the next one can be sent.

The speed at which waves propagate from transmitter to receiver approaches the speed of light – the speed of light (3×10^8m/s) is only achieved as waves travel through a vacuum. In copper wire and other mediums speeds of around 2×10^8m/s are more realistic. In any case the speed of the wave is incredibly fast. At a speed of 2×10^8m/s, travelling the 20,000km around to the other side of the Earth takes one tenth of a second.

GROUP TASK Activity
Calculate the minimum transmission time required to transfer a 1kB packet at 10Mbps to a satellite located 40,000 metres above the Earth.

GROUP TASK Discussion
Why is the speed of wave propagation particularly significant over longer distances when each data packet must be acknowledged before the next one can be sent? Discuss.

GROUP TASK Discussion
As CPU speeds increase and motherboards transfer data faster, will the speed of wave propagation within and between motherboard components become significant? Discuss.

Bandwidth

The term bandwidth is often used incorrectly, people make statements such as "video requires much more bandwidth than text" or "my bandwidth decreases as more people use the Internet". Statements such as these are incorrect; they are using bandwidth when they really mean speed or bps. Bandwidth is not a measure of speed at all; rather it is the range of frequencies used by a transmission channel. Presumably misunderstandings have occurred because the theoretical maximum speed does increase as the bandwidth of a channel increases. However, it is simply impossible for the bandwidth of most channels to change during transmission. Each channel is assigned a particular range of frequencies when it is first setup; unless you run a high-speed Internet company or are creating your own hardware transmitters and receivers, then altering bandwidth is really beyond your control.

Bandwidth
The difference between the highest and lowest frequencies in a transmission channel. Hence bandwidth is expressed in hertz (Hz), usually kilohertz (kHz) or megahertz (MHz).

So what is bandwidth? It is the difference between the highest and the lowest frequencies used by a transmission channel. Frequency is measured in hertz (Hz), meaning cycles per second. Each cycle is a complete wavelength of an electromagnetic wave, so 20Hz means 20 complete wavelengths occur every second. As frequency is expressed in hertz then so to is bandwidth. For example, standard telephone equipment used for voice operates within a frequency range from about 200Hz to 3400Hz, so the available bandwidth is approximately 3200Hz. As high-speed connections routinely use bandwidths larger than 1,000Hz or even 1,000,000Hz, bandwidth is usually expressed using kilohertz (kHz) or megahertz (MHz). For example 3200Hz would be expressed as 3.2kHz.

All signals need to be modulated in such a way that they remain within their allocated bandwidth. This places restrictions on the degree of frequency modulation that can be used. As a consequence most modulation systems rely on amplitude and phase modulation. For example, most current connections to the Internet use Quadrature Amplitude Modulation (QAM), this system represents different bit patterns by altering only the amplitude and phase of the wave. 16QAM uses 16 different symbols to represent 4 bits/symbol, 64QAM uses 64 different symbols to represent 6 bits/symbol and 256QAM uses 256 different symbols representing 8 bits/symbol.

Amplitude, phase and frequency are related; altering one has an effect on each of the others. Increasing the available frequency range (bandwidth) results in a corresponding increase in the total number of unique amplitude and phase change combinations (symbols) that can accurately be represented and detected. In general, it is true that the speed of data transfer increases as the bandwidth is increased.

It is difficult to discuss bandwidth without mentioning the related term 'broadband'. Broadband, is a shortened form of the words broad and bandwidth. As is the case with numerous computer related terms there are various accepted meanings. In common usage broadband simply refers to a communication channel with a large bandwidth. However, the term is also used in reference to a physical transmission medium that carries more than one channel. In essence, the total bandwidth is split into separate channels that each use a distinct range of frequencies. Using either meaning, most long distance Internet connections and both ADSL (Asymmetrical Digital Subscriber Line) and cable are examples of broadband technologies. They all deliver high data rates (theoretically in excess of 5Mbps) by splitting the total bandwidth into separate communication channels. The opposite of broadband is baseband. Baseband

connections include Ethernet, 56kbps modem links and 128kbps ISDN links where a single communication channel is used. The term "narrowband" refers to a single channel that occupies a small bandwidth, such as traditional voice telephone lines.

Consider the following:

A 2MB file is to be transferred over the following communication channels:

1. A 56kbps dial-up modem link.
2. A 10Base-T Ethernet connection.
2. A 100Base-T Fast Ethernet connection.
3. A 1000Base-T Gigabit Ethernet connection.
4. A 640ksym/s cable modem channel that uses 16QAM.
5. An ADSL channel operating at 1.5Mbps.
6. A DSL channel that uses 64QAM and a symbol rate of 4Msym/s.

GROUP TASK Activity
Calculate the minimum time taken to transfer the 2MB file over each of the above communication channels.

GROUP TASK Discussion
Identify and discuss reasons why it is unlikely that the minimum times calculated above would be realised in reality.

GROUP TASK Practical Research Activity
Investigate the specifications of communication channels used at your school and at home. Determine the minimum time to transfer a 2MB file. Now actually transfer a 2MB file and determine if your calculations are close to the actual time taken.

ERROR CHECKING METHODS

In our previous section on protocols we learnt that TCP includes a checksum within the header of each segment, IP includes a checksum of just the header fields and Ethernet frames contain a 32-bit CRC (Cyclic Redundancy Check). In this section we consider the detail of checksums and CRCs, however we commence by examining simple parity checks.

Note that when an error is detected the receiver can respond in various ways depending on the rules of the particular protocol. The receiver may simply drop the data packet, such as occurs with IP. They may only acknowledge correct packets as occurs with TCP or they may specifically request that packets containing errors be resent, as occurs with Ethernet.

Parity bit check

Early modems transmitted and received each character separately as its 7-bit ASCII code. In essence each packet of data contained just 7-bits, for such small packets a simple error checking technique known as parity was all that was used. Furthermore much of the data was text, so the occasional incorrect character was not a significant problem. Today data is transmitted by all modems in much larger packets that utilise more sophisticated error check techniques such as checksums and CRCs. Nevertheless most serial ports (and dial-up modems) still provide the ability to use parity checks to allow compatibility with older (and much slower) methods of communication (see *Fig 3.13* on the next page).

Parity bits are still used internally by components on the motherboard. For example many types of RAM chip include parity bits for each byte of storage and the PCI bus uses a modification of the parity system to detect errors within addresses and commands communicated between the PCI controller and attached devices on the motherboard.

Parity bits are single bits appended either before or after the data so that the total number of ones is either odd or even. During handshaking the sender and receiver decide on whether odd or even parity will be used. Parity bits can be created for any length message, however their use is generally restricted to individual characters or bytes of data.

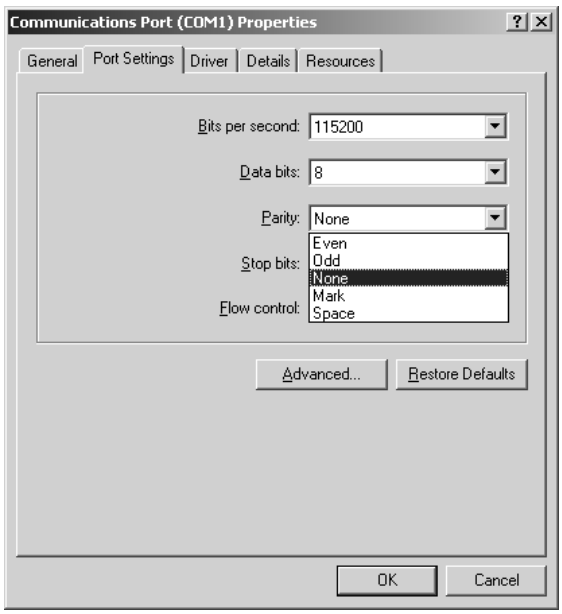


Fig 3.13
Serial or COM port settings include a Parity option within Windows XP.

You may have noticed that in *Fig 3.13* there are five parity options in the drop down box – even, odd, none, mark and space. Odd and even are the only two options that provide error checking. None means no parity bit is included in the transmission, mark means a 1 is always transmitted as the parity bit and space means a 0 is always transmitted. The mark and space options provide compatibility with some specialised devices that connect via a serial port, for example a device may specify 8M1 as its required port setting, this means 8 data bits, mark parity (i.e. always 1) and 1 stop bit.

Consider the transmission of the word "ARK" using odd parity, where the parity bit is appended to the end of the character bits (refer *Fig 3.14*). The ASCII code for A is 65, which is 1000001 in binary. There are two 1s hence to make the total number of bits odd requires the parity bit to be set to 1. The letter A is therefore transmitted as 10000011 – note that the total number of 1s is now the odd number 3. Similarly the letter R is transmitted as 10100100 and the letter K is transmitted as 10010111. If even parity had been used rather than odd parity then each parity bit would be reversed to make the total number of 1s an even number.

Char	ASCII code		Odd Parity Bit
	Dec	Binary	
A	65	1000001	1
R	82	1010010	0
K	75	1001011	1

Fig 3.14
The word 'ARK' using odd parity.

Consider what occurs if bits are corrupted (reversed) during transmission. If any single bit (including the parity bit) is corrupted then the receiver will detect an error. Indeed an error is detected whenever an odd number of bits are corrupted. However whenever an even number of bits are reversed no error is detected at all. The total number of ones remains an odd number when using odd parity (or an even number if using even parity). This is a significant problem with parity checks when the communication is over external media that is influenced by environmental interference; hence parity checks are unsuitable for detecting network transmission errors. However within components and between components on the motherboard by far the most common error is a simple reversal of a single bit; in these cases a simple parity check will detect the large majority of errors.

GROUP TASK Discussion

As most computers boot they perform a RAM parity test. This test writes a byte to each memory location and then reads the byte together with the parity bit. Memory locations that do not pass are simply not used. Discuss how this system differs from parity checks used during the transfer of data.

Checksums

Checksums, as the name suggests, are calculated by summing or adding up. The simplest checksums simply add all the bytes as if they were integers within the message. The resulting sum is then sent along with the message. The receiver also calculates the sum of the bytes and compares their result with the received checksum.

To reduce the size of the checksum only a portion of the least significant bits (right hand bits) are usually sent. For example an 8-bit checksum sends only the 8 least significant bits. To simplify the math we can simulate this process in decimal. Say we wish to transmit the following five numbers 130, 203, 97, 38 and 181. The sum of these numbers is 649. Now 8-bits is the size of our checksum and 8-bits can represent numbers from 0 to 255, therefore we require the remainder after division by 256. With our example we calculate 649 ÷ 256 = 2 with a remainder of 137. We send the checksum 137 along with our data. *Fig 3.15* shows this calculation in both decimal and binary. Notice that in binary there is no need to perform any division, rather we can simply discard the excess bits.

Decimal	*8-bit Binary*
+ 130	+ 10000010
203	11001011
97	01100001
38	00100110
181	10110101
649	1010001001
137	10001001
Checksum is remainder after dividing by 256	Checksum is the 8 least significant bits

Fig 3.15
Initial calculation of an 8-bit checksum.

The above example (in *Fig 3.15*) is trivial – the checksum is just 1-byte (8-bits) long and the data itself is just five bytes long. In reality checksums are usually 2-bytes (16-bits) long or even 4-bytes (32-bits) long. IP headers include a 16-bit checksum over their header fields and TCP includes a 16-bit checksum over the complete segment (packet). Clearly both IP and TCP generate checksums over much larger amounts of data. To simplify our discussion we will continue using our 8-bit example using just 5 bytes of data.

There are two significant problems with our initial checksum calculation. Firstly if the data being sent contains all zeros then the checksum will also be zero. Errors can occur in either software or hardware that cause empty packets (all zeros) to be sent and our initial checksum would not detect such problems. To solve this issue the calculated checksum is simply reversed – all zeros become ones and all ones become zeros. Technically this transformation finds the ones complement of the checksum. Our "all zeros" problem is now solved, as an actual real packet of zeros will have a checksum that is a sequence of ones rather than a sequence of zeros. This transformation is performed for virtually all checksums, including IP and TCP checksums. Modifying our initial example from *Fig 3.15* above the checksum sent becomes 01110110 rather than 10001001 as shown in *Fig 3.16*. A bonus side effect of this ones complement transformation simplifies the work required by the receiver. The receiver now simply adds up all the data including the checksum and the result must always be a sequence of ones.

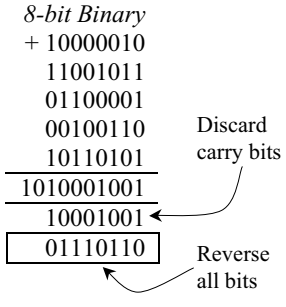


Fig 3.16
Modified calculation of an 8-bit checksum.

GROUP TASK Activity
Confirm the receiver will calculate a sum of 11111111 using our *Fig 3.16* example and a calculator in binary mode – add all five data bytes and the checksum 01110110. Create and test some other examples.

The second significant problem is not unlike the parity problem, where reversing an even number of bits caused the data to be received without an error being detected. In the case of checksums this problem is less severe as it only occurs as a result of corruption of the most significant bits (MSBs or left hand side bits) in the data bytes. For example in *Fig 3.16* if the MSB of the first two data bytes are reversed such that zeros rather than ones are received the addition performed by the receiver still results in 11111111 and the data is accepted by the receiver despite the errors.

GROUP TASK Activity
Confirm with a calculator that the checksum is unchanged when an even number of MSBs are reversed. Try altering an odd number of MSBs and altering an even and odd number of various other bits. Confirm that for these cases the checksum does indeed change.

To understand the solution to this problem consider the sum of the data bytes prior to discarding the carry bits. In our *Fig 3.16* example the uncorrupted data bytes sum to 10 1000 1001 and when the MSB of the first two data bytes is corrupted the sum is 1 1000 1001. Note that the carry is different – originally the excess carry bits were 10, whilst the corrupted sum has a carry of just 1. If we can include the carry as part of our checksum then the problem will be solved – currently we are simply discarding the carry. At first glance we may be tempted to simply extend the length of the checksum to include the carry bits. This possibility is ruled out, as with larger, more realistically sized data packets the carry is potentially as long as the original checksum. This additional overhead would slow transmission significantly – the length of all checksums would need to be doubled. A better solution is to simply add the carry to the sum. Technically this process is identical to ones complement addition. *Fig 3.17* shows the complete process of creating an 8-bit checksum. Note that the carry bits, 10 in this case are added to the sum prior to reversal. At the receiving end the data and checksum must sum to 11111111 for the packet to be accepted.

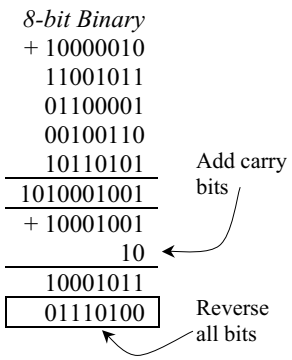


Fig 3.17
Final calculation of an 8-bit checksum.

The ability of checksums to detect errors is far better than simple parity checks, however some errors are still possible. Determining the precise theoretical accuracy of a checksum requires consideration of the length of the data packet together with the length of the checksum. Furthermore not all types of errors are equally likely on all communication links. For these reasons it is not a simple process to determine the actual percentage of errors that will be detected. Nevertheless we can calculate a reasonable prediction of accuracy based solely on the length of the checksum.

GROUP TASK Discussion Activity
Identify examples of corruption that could occur that will not be detected by a checksum. Use a calculator to confirm that the checksum is the same for both the original and corrupted versions of the data.

To simplify our discussion consider an 8-bit checksum. There are exactly $2^8 = 256$ different possible checksums that can be generated and sent. Every possible message packet results in one of these possible 256 checksums. The only times when the receiver will NOT detect an error is when the message packet is corrupted in such a way that it still produces the same checksum as the original message produces. If all possible corruptions of message packets are equally likely (which in reality is not true) then the probability that a message will be corrupted in such a way that its checksum remains the same must be 1 in 256. Therefore for an 8-bit checksum the probability that an error is detected must be $1 - \frac{1}{256}$ or approximately 99.6% of the time. For checksums of any length n we can generalise our formula such that the probability of an error being detected is approximately $1 - \frac{1}{2^n}$.

Applying our general formula to the more common 16 and 32 bit checksums we expect to detect errors approximately 99.9985% of the time with a 16-bit checksum and 99.999999977% of the time with a 32-bit checksum. This means the 16-bit checksum used by IP datagram headers and TCP segments will, based on our theory, fail to detect just one or two errors in every one hundred thousand transmissions. In reality checksums are not quite this accurate as all errors are not equally likely. Cyclic redundancy (CRC) checks are an attempt to deal with this issue. Remember that further error checks exist within other OSI layers; hence even errors that pass through a protocol within one layer undetected are likely to be detected by protocols operating within other OSI layers.

GROUP TASK Discussion
IP datagrams include a 16-bit checksum calculated using just their header whilst TCP segments include a 16-bit checksum calculated using the entire message. Discuss possible reasons for this difference and describe likely differences between the accuracy of IP and TCP checksums.

Cyclic Redundancy Check (CRC)

Cyclic Redundancy Checks or CRCs form part of many Transmission Level (OSI layer 1 and 2) protocols including Ethernet, ATM and SONET. The calculation of CRC values is generally built into and performed by the hardware. Note that most secondary storage devices perform CRC checks as data is accessed from the drive – this includes hard disks, CD/DVD drives and also tape drives. CRCs are a significantly stronger technique for detecting errors than checksums and are far superior to simple parity checks. It is the method of calculating CRC values that is different to checksums rather than the way they are used during transmission. Both checksums and CRC values are calculated and included within the header or trailer of each message packet by the sender. The receiver calculates the CRC value and compares it to the CRC value within the received message packet.

CRC values are calculated using division whilst checksums use addition. In simple terms, to calculate a CRC we consider the entire message to be a complete number. This number is then divided by another predetermined number (called a generator polynomial). The remainder from this division becomes the CRC value.

Lets us perform a CRC calculation using our example 5-byte message from *Fig 3.16*. To further simplify our working we'll calculate an 8-bit CRC value using the generator polynomial 110010001 (which is equivalent to 401 in decimal). I just made up this example 9-bit generator polynomial – the only requirement at this stage of our discussion being that it contain one more bit than the size of the CRC value we wish

to generate. Later we shall discuss the significance of the generator polynomial, and also why it is called a generator polynomial.

The five bytes to be transmitted are 10000010, 11001011, 01100001, 00100110 and 10110101. We consider this to be one single complete binary number. This binary number is equivalent to the decimal number 561,757,890,229. Dividing by 401 we get 1,400,892,494 remainder 135. Now the remainder 135 in binary is 10000111, we could use this CRC value and send it with the message and it would have most of the benefits of a real CRC value. Unfortunately such long divisions are laborious and for computers they require many machine instructions. Many of these machine instructions are unnecessary in terms of achieving the purpose of a strong error checking technique. It is critical that the calculation is as efficient as possible when you consider that every frame of data sent using Ethernet (and other low level protocols) requires the CRC calculation to be performed by both the sender and the receiver.

In reality CRC values are calculated using a simpler long division based on polynomial division. This technique does not require us to worry about carries at all when performing the required subtractions – mathematically each binary number represents the coefficients of a polynomial and we perform the subtractions using modulo 2 arithmetic. For our level of treatment we need not concern ourselves with polynomials however it does explain the use of the term 'generator polynomial'. Modulo 2 arithmetic is really easy; addition and subtraction are the same and there are only two possible answers to any addition – either 0 or 1. If we're adding an even number of 1s then the answer is 0 and adding an odd number of 1s results in 1. To calculate CRCs we really only need to know that 0+0=0, 0+1=1, 1+0=1 and 1+1=0. These results are simple to implement using hardware as a single logic gate called an XOR gate performs precisely this process. An example calculation using this system and performed using our data from *Fig 3.16* and the generator polynomial 110010001 is reproduced in *Fig 3.18* on the next page. It is worthwhile examining this example to understand the process more thoroughly, however in IPT it is highly unlikely that you would be asked to perform such a calculation in an examination.

GROUP TASK Activity
Perform the division 1100110011001100 divided by 10011 using the system described above and within *Fig 3.18*. The final remainder (or CRC value) calculated should be 101.

CRCs are stronger than checksums because they are able to detect many of the more common types of transmission errors. For example, checksums are unable to detect errors where 2 bits within one column of the addition have been corrupted - CRCs detect all such errors. Furthermore CRCs will detect all error bursts that are less than or equal to the length of the generated CRC value. For example a 32-bit CRC detects all errors where the number of bits counting from the first corrupted bit to the last corrupted bit is less than or equal to 32. This is due to the way remainders after division change compared to how sums after addition change. In practice corruption of bits during transmission tends to occur more often in bursts – it is rare for the corrupted bits to be distributed throughout the entire message packet.

The specific types of error detected by CRCs changes when different generator polynomials are used. The mathematics required to explain the effect of different generator polynomials is well beyond what is required in IPT. Nevertheless there are standard generator polynomials that have been shown to detect the largest range of likely transmission errors that occur in most communication systems.

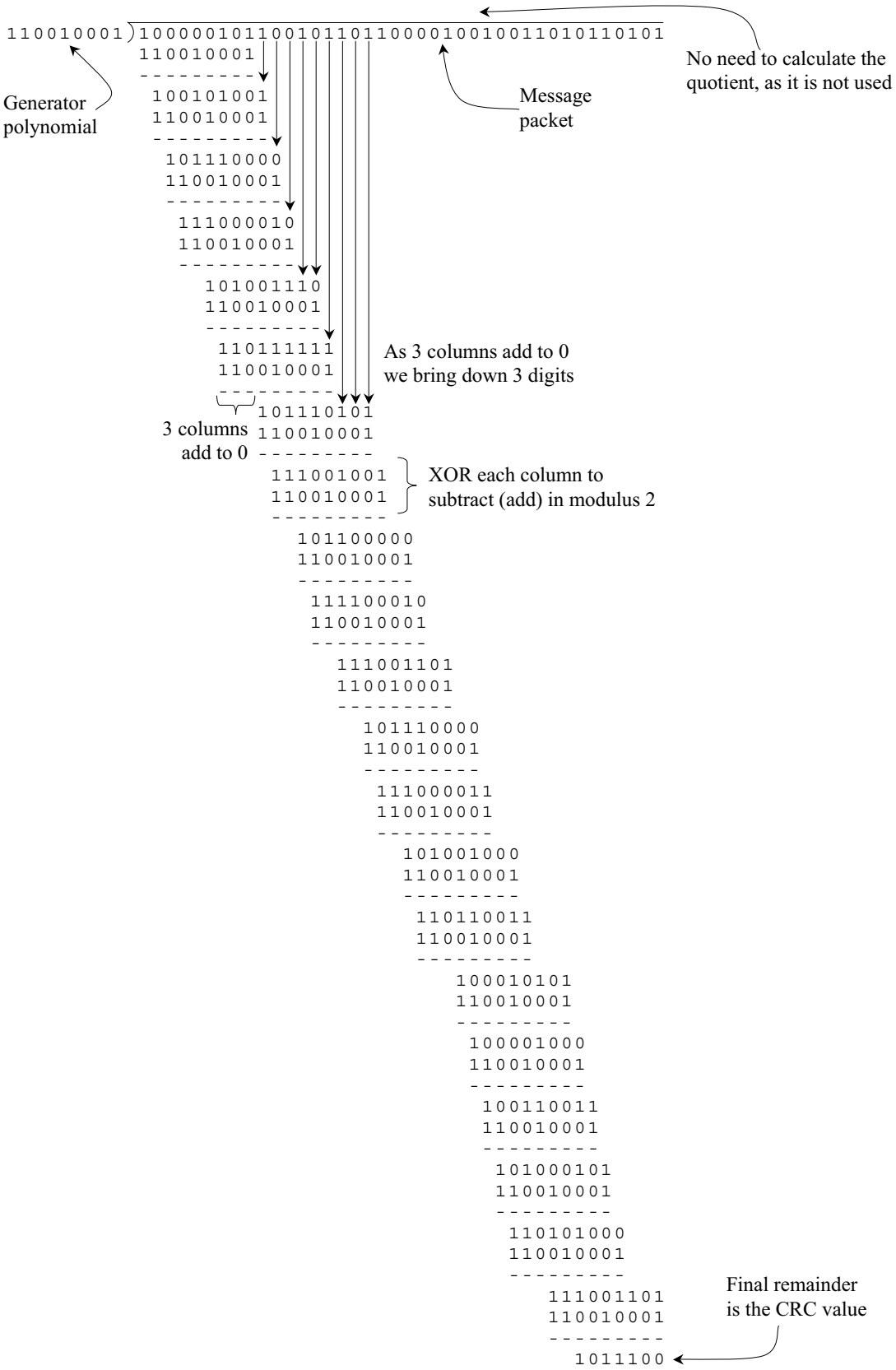


Fig 3.18
CRC calculation example.

There are some common CRC standards and generator polynomials that are each used by many protocols – CRC-16-X25, CRC-16-BYSNCH and CRC-32. The generator polynomials together with example protocols that use the standard are reproduced in *Fig 3.19*. Ethernet uses the CRC-32 standard whilst fax machines and many other telephone line devices use the CRC-16-X25 version within the X.25-CCITT protocol. Many high-speed long-distance protocols such as SONET use 64-bit or even 128-bit CRCs. All CRCs are calculated using essentially the same 'division like' process as that described above. However there are slight differences in the way they are implemented. For example when using CRC-32 the final CRC value is reversed prior to sending.

	CRC-16-X25	CRC-16-BYSNCH	CRC-32
Width	16 bits	16 bits	32 bits
Generator Polynomial	1 0001 0000 0010 0001 69,665 (Decimal)	1 1000 0000 0000 0101 98,309 (Decimal)	1 0000 0100 1100 0001 0001 1101 1011 0111 4,374,732,215 (Decimal)
Example Protocols	ITU-TSS, X.25-CCITT, V.41, XModem, IMB SDLC, PPP	IBM BISYNCH LHA, PKPAK, ZOO	Ethernet, ATM, FDDI, PPP, PKZip,

Fig 3.19
Common CRC standards.

In general, CRCs detect more errors than a checksum of the same length. Determining the actual probability of a particular CRC detecting errors is a difficult task. For our purposes it is sufficient to state that we expect them to detect more errors than our probability calculations for checksums. That is, when using a 16-bit CRC we expect better than 99.9985% of errors to be detected and when using a 32-bit CRC we expect more than 99.999999977% of errors to be detected.

GROUP TASK Discussion
Propose possible reasons why transmission errors that occur within message packets are more likely to occur in bursts rather than being more evenly distributed throughout the packets.

Hamming Distances and Error Correction (Extension)

The number of changes between two patterns is known as the 'hamming distance' For example, the hamming distance between the word 'sock' and the word 'silk' is 2, as the two letters 'o' and 'c' have changed to the letters 'i' and 'l' respectively. Similarly, the patterns of bits 10011100 and 10101101 have a hamming distance of 3 – as the third, fourth and last bits have changed. If the bit patterns are message packets that both result in the same checksum or CRC value then corruption such that one bit pattern becomes the other will not be detected.

Computer engineers design error checks that aim to maximise the minimum hamming distance between messages that result in the same check value. The theory being that corruption of a small number of bits is much more likely than corruption of a larger number of bits. For example, if the minimum hamming distance for a particular error checking technique to produce the same check value is 8 then all errors where less than 8 bits are corrupted will be detected.

This 'hamming distance' information is used by some error checking techniques to not only detect errors but to also correct errors without the need for the message packet to be resent. Consider our example error checking technique where all errors with less than 8 bits corrupted are detected. Say an error is detected within a received

message. We know the check value hence the correct message packet must be one that produces this check value – this limits the set of possible correct message packets significantly. We can then select from this set any message packets that are close (in terms of hamming distance) to the corrupted message received. In our example we would choose packets where the hamming distance between the packet and the corrupted packet is less than 8. If the smallest hamming distance calculated is the same for more than one possible packet then we cannot correct the error. On the other hand if just one possible message packet is closest then we can reasonably assume that this is the correct packet.

GROUP TASK Discussion

The length of an error burst is the number of bits between the first and last corrupted bit. For example an error burst may be 8 bits long yet the hamming distance is just 2. In terms of error checking, compare and contrast the significance of the length of an error burst with its hamming distance.

HSC style question:

Web browsers are applications that retrieve web pages from web servers, they then format and display the retrieved web pages.

(a) Identify and briefly describe TWO communication protocols in use during the retrieval of a web page from a web server.

(b) Identify and describe the operation of TWO error checking methods used during the transmission of a web page from a web server to a web browser.

Suggested solution

(a) There are many protocols involved in the transfer of files from a web server to a web browser. However in all cases the protocols will include HTTP and IP.

HTTP or Hypertext Transfer protocol operates within the application layer and is used by the web browser to request a particular file from the web server using an HTTP GET command that includes the URL of the requested file. In most cases the file will be an HTML document. The web server responds by sending the file back to the web browser. The web browser examines the header that precedes the file to determine its type and how it should be formatted and displayed.

All data is transmitted across the Internet using IP (Internet Protocol). IP is an OSI model layer 3 protocol whose main task is to deliver IP datagrams to their destination. IP does not include any mechanism for acknowledgement of messages – in fact there is no guarantee that IP datagrams will reach their destination. Datagrams sent using IP are directed through many network hops by routers. The router uses the destination IP address within the header of each datagram to determine the next hop for the message. Each header also contains a TTL (time to live field) that is decremented for each network hop the datagram passes. If this TTL field becomes zero then the datagram is discarded.

(b) Most web servers exist on Ethernet networks as do most machines running web browsers, therefore the message commences and ends its journey as a sequence of layer 1 and 2 Ethernet protocol frames – Ethernet includes CRC-32 error

checking. Also IP is used to transmit datagrams within layer 3 across the Internet and includes a 16-bit checksum of each datagram's header.

IP 16-bit checksums are calculated by summing each double byte (16-bits) within the header of the IP datagram. This total is likely to contain carries in excess of the 16-bit checksum. These carry bits are added back into the checksum. It is the reverse of this result that is sent as the checksum. The receiving device (which may be a router somewhere on the Internet) adds the header and checksum and discards datagrams where the result is not a string of ones.

The CRC-32 system used by Ethernet is a much stronger error checking method than the above 16-bit checksum. In the case of Ethernet the CRC value is calculated over the whole message frame. Cyclic Redundancy Checks (CRCs) are calculated using a special type of division based on polynomial division and modulus 2 arithmetic. The message data is considered to be a long binary number, this number is then divided by a predetermined binary number known as the generator polynomial. It is the remainder after this modified division process that is sent as the CRC check value. Using Ethernet the sender specifically requests corrupted packets to be resent.

Comments

- In an HSC or Trial examination this question would likely be worth six marks – three marks for each part.
- Many other protocols could have been identified and described in part (a).
- The description of the error checking methods should address the specific implementation used by the protocol rather than just the general operation of the error checking method.
- It would be risky to discuss parity checks for part (b) unless justification is included that retrieval of the file from hard disk is part of the transfer.

SET 3B

1. The number of signal events occurring each second is known as the:
   (A) bits per second.
   (B) bandwidth.
   (C) Baud rate.
   (D) modulation scheme.
2. A communication channel modulates waves using 256 QAM and transmits 8 million symbols each second. Approximately how long will it take to transfer 10MB?
   (A) 64 seconds
   (B) 8 seconds
   (C) 0.125 seconds
   (D) 1.25 seconds
3. Which of the following includes only baseband communication links?
   (A) Ethernet, ISDN
   (B) ADSL, ISDN
   (C) Ethernet, ADSL
   (D) Cable, ADSL
4. Which of the following is TRUE in terms of 8-bit checksums?
   (A) Approximately 99.6% of errors are detected.
   (B) Approximately 99.6% of data packets will be received correctly.
   (C) Approximately 99.6% of packets will not be corrupted during transmission.
   (D) Approximately 99.6% of detected errors can be corrected.
5. Protocols that include checksums include:
   (A) Ethernet and SONET.
   (B) TCP and IP.
   (C) ATM and IP.
   (D) TCP and Ethernet.
6. A parity bit is added to each byte of data sent. If all data bits are reversed what will occur?
   (A) The error will always be detected.
   (B) No error will ever be detected.
   (C) Some errors will be detected.
   (D) Most errors will be detected.
7. 7-bit ASCII data is sent one character at a time using odd parity. The received data contains errors. Which of the following is most likely?
   (A) An odd number of bits in some bytes were corrupted.
   (B) The parity bit in some bytes was corrupted.
   (C) An even number of bits in some bytes were corrupted.
   (D) The receiver has different port settings to the sender.
8. The range of frequencies a transmission channel occupies is know as its:
   (A) symbol rate
   (B) Baud
   (C) speed
   (D) bandwidth.
9. The most significant advantage of CRCs compared to checksums is:
   (A) CRCs are used by lower OSI layer protocols than checksums.
   (B) CRCs are better at detecting commonly occurring types of transmission errors.
   (C) Division is a more reliable operation than addition.
   (D) CRCs are usually implemented within the hardware while checksums are implemented within software.
10. When using parity bits, checksums and CRCs, what must occur for an error to go undetected?
    (A) The message must be corrupted such that the parity bit, checksum or CRC is unaltered.
    (B) An even number of bits within the message must be corrupted.
    (C) The error must be the result of hardware errors rather than software or interference errors.
    (D) The message must be corrupted in such a way that is becomes some other legitimate message.
11. Define each of the following terms.
    (a) Bits per second (b) Baud rate (c) Bandwidth (d) Broadband (e) Baseband
12. The word "CAR" is sent using 7-bit ASCII and even parity. The following data is received: 10000111, 10000011 and 10101001.
    (a) Comment on errors detected and undetected.
    (b) Explain how detected error(s) could be corrected.
13. Calculate an 8-bit checksum for the following 6 bytes of data using the calculation method described in *Fig 3.17*. 00001111, 11110000, 10101010, 01010101, 11001100, 00110011.
14. Compare and contrast checksums and CRCs in terms of their:
    (a) method of calculation. (b) ability to detect errors.
15. For each of the following protocols, outline the method of error detection and method of error correction used.
    (a) TCP (b) IP (c) Ethernet

EXAMPLES OF COMMUNICATION SYSTEMS

In this section we consider three broad types of communication system. Firstly, teleconferencing systems where real time audio, video and/or other data is shared between participants. Secondly, messaging systems including traditional phone and fax as well as voice mail and email systems. Finally, we examine electronic commerce systems, specifically EFTPOS and Internet or electronic banking.

Before we commence, we first describe relevant characteristics of three large wide area networks used to transmit data within the above communication systems – namely, the Internet, the public switched telephone network (PSTN) and intranets and extranets.

Internet

The Internet is a worldwide packet switched public network based on the Internet Protocol where all data moves between nodes within IP datagrams. As we learnt previously, there is no guarantee that IP datagrams will reach their destination. Furthermore the Internet is "connectionless" meaning there is no connection maintained between the sender and receiver – in effect each IP datagram is on its own and may follow a different path to its destination. As a consequence IP datagrams can arrive out of sequence or not arrive at all. These issues are insignificant when the communication between participants is asynchronous. However these are significant issues when real time (synchronous) communication is required. Synchronous in this context refers to the ability of participants to hold a real time conversation whilst asynchronous refers to systems where there is (or can be) a pause between sending and receiving processes. The Internet was designed for asynchronous rather than synchronous transfers.

GROUP TASK Discussion
Consider different forms of communication between groups of people. Classify each as either asynchronous or synchronous.

Public Switched Telephone Network (PSTN)

The PSTN is the network that carries traditional telephone calls throughout the world – it is also known as the "Plain Old Telephone Service" or POTS. The PSTN differs from the Internet because it creates and maintains an individual circuit between the participants during each conversation. When a phone call begins a single direct connection is created between the two telephones. This connection or circuit is used for the duration of the call – hence the PSTN is known as a "connection-based" or "circuit switched" network. This connection-based system was designed for real time synchronous voice communication using telephones as the collection and display devices.

The significant infrastructure of the PSTN has been in place for many years and is owned and maintained by governments and large telecommunication companies. Somewhat confusingly, much of the data transferred over the Internet actually travels across the PSTN. Most Internet service providers, rather than installing their own dedicated lines, lease connections on the PSTN. This means many connectionless IP datagrams actually travel along network hops alongside connection-based data.

GROUP TASK Research
ISDN is a set of layer 1, 2 ,3 protocols that transfers data over the PSTN In Australia ISDN was once popular for business communication. Research and explain why ISDN is no longer popular.

Intranet and Extranet

An intranet is a private network maintained by a company or government organisation and is based on the Internet protocol (IP). Many intranets include leased high-speed lines to connect their local area networks (LANs) into a private wide area network (WAN). The leased lines are dedicated to traffic on a specific private intranet. Such leased lines mean that the amount of data transferred is under the direct control of the intranet owner. This control becomes significant when real time synchronous applications are used. Some intranets connect LANs using the public Internet where all messages are encrypted during transmission to ensure privacy is maintained.

Extranets are an extension of an intranet to allow access to customers and other users outside the organisation. The interface between the extranet and the intranet must be secure – commonly firewalls, user names and passwords and also encryption is used. Extranets allow companies to share their services with other companies. For instance a large bank may provide online banking services to other smaller banks via its extranet.

Both intranets and extranets can also include virtual private networks (VPNs). VPNs use the infrastructure of the public Internet to provide secure and private connections to a company's internal network. A VPN allows employees to securely communicate with their company's network using any Internet connection. VPNs include tunnelling Transmission protocols, which not only encrypt and secure messages but also encrypt all internal network addresses. Examples of tunnelling protocols include Microsoft's Point to Point Tunnelling Protocol (PPTP), Cisco's Layer 2 Forwarding protocol (L2F) and the Layer 2 Tunnelling Protocol (L2TP) which is a standard that aims to combine the benefits and functions within both PPTP and L2F.

GROUP TASK Research
Explain why organisations may choose to set up an intranet in preference to simply using the public and less expensive infrastructure of the Internet.

TELECONFERENCING

Teleconference
A multi-location, multi-person conference where audio, video and/or other data is communicated in real time to all participants.

The term "teleconference" encompasses a wide variety of different real-time conference systems. From a simple three-way call using standard telephones to systems that share audio, video and other types of data between tens or even hundreds of participants. The essential feature of all teleconferencing systems is synchronous communication between many people in many different locations. Commonly many participants are present at one location whilst single participants are present at other locations. For example teleconferencing is routinely used for meetings between an organisation's head office and its branch offices. There are many participants present at head office and other participants at each branch office.

Historically the term "teleconference" referred to multi-person multi-location conferences sharing just audio over the PSTN - this audio only meaning is still used by many. Today such conferences routinely include video and various other types of data in addition to audio. Many references recommend using more descriptive terms, such as videoconference to describe systems that include video or e-conference when many data types are shared. In our discussion we shall use the more general meaning of teleconferencing that includes the real-time sharing of a variety of different data types.

We cannot hope to describe all the possible types of teleconferencing systems available. Rather we examine two particular examples of teleconferencing that utilise different information technology to achieve their purpose, namely:

1. Business meeting system, sharing audio over the PSTN.
2. Distance education system, sharing audio, video and other data using both the PSTN and the Internet.

For each teleconferencing system we identify the environment and boundaries, purpose, data/information, participants and information technology. We then discuss the information processes, in particular the essential transmitting and receiving processes used by the system. Finally we consider the advantages and disadvantages of teleconferencing within the context of the particular system.

GROUP TASK Research
Using the Internet, or otherwise, create a list of specific examples where teleconferencing is used.

1. BUSINESS MEETING SYSTEM, SHARING AUDIO OVER THE PSTN.

Environment/Boundaries

In this example we consider a medium sized business that has a head office in Sydney and five branch offices in country towns throughout NSW. At some stage during each Tuesday a teleconference is scheduled between the general manager, the four division managers and each of the branch mangers. The general manager and the division managers have offices within head office. Each of the division managers takes turns to chair and manage the weekly meeting.

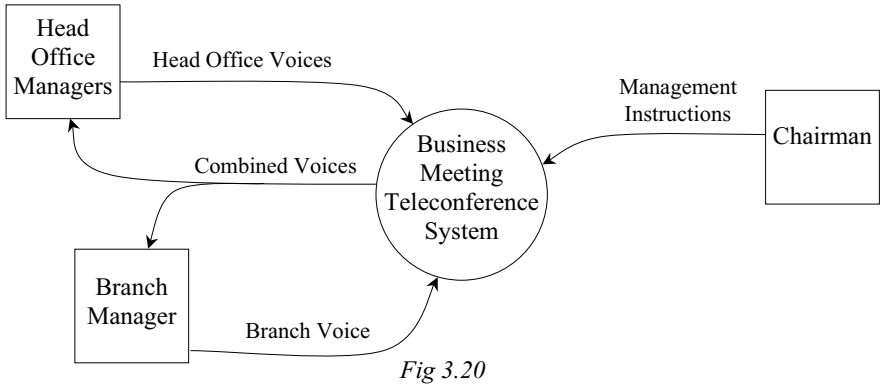


Fig 3.20
Initial context diagram for a business meeting teleconference system.

An initial context diagram describing this teleconferencing system is reproduced in *Fig 3.20* – the data flows and labels at this stage are incomplete. On this diagram just one of the branch managers is shown, in reality there are five branch managers. It makes sense to include the chairman as a separate entity as the inputs into the system from the chairman are different to their contributions as a member of the head office managers.

GROUP TASK Discussion
How does the initial context diagram in *Fig 3.20* assist to define the boundaries of the teleconferencing system?

Purpose

The needs that the weekly management meetings aim to fulfil include:

- Efficiently disseminating information to all managers throughout the organisation.
- Improving the efficiency of decision-making processes by managers – particularly with regard to including branch managers in the decision making process.
- Encouraging the sharing of ideas and strategies between members of the management team.
- Sharing of staff issues occurring at the local level with a view to more amicably and consistently resolving such issues across the entire organisation.
- Maintaining and enhancing interpersonal relationships between members of the management team.
- Inclusion of all managers, even if this means rescheduling the meeting at late notice.

Taking these needs and other more general business needs into account, the purpose of this business teleconferencing system is to:

- Provide the ability for all managers to contribute equally at weekly management meetings.
- Enable managers at remote locations to participate in all meetings without the need to travel.
- Output audio of sufficient quality such that all voices can be understood at all locations, including when multiple people are speaking at the same or different locations.
- Reduce costs through a reduction in the number of face-to-face management meetings required throughout the year.
- Be simple to setup, such that meetings can be rescheduled at late notice with minimal effort.
- Include only reliable, commonly available, well-tested technologies that provide a high quality of service without the need for onsite technical expertise during use.

GROUP TASK Discussion
Discuss how each of the above purpose statements assist in fulfilling one or more of the needs on the previous page.

Data/Information

The following table summarises the data/information used by the teleconference system. The table includes the audio input to and output from the system together with data required to access and manage the setup and operation of the system.

In this example system the meeting agenda and the minutes produced after the meeting are not included. Such data and information is outside the boundaries of the system that were defined on the initial context diagram.

Data/Information	*Data type*	*External Entity*	*Source OR Sink*	
Head Office Voices	Audio	Head Office Managers	✓	
Branch Voice	Audio	Branch Manager	✓	
Combined Voices	Audio	Head Office Managers /Branch Manager		✓
Management Commands	Numeric	Chairman	✓	
Start Date/Time	Numeric	Chairman	✓	

Host PIN	Numeric	Chairman	✓	
Guest PIN	Numeric	Branch Manager	✓	
Dial in Number	Numeric	Chairman /Branch Manager	✓	
Simulated Voice Response	Audio	Chairman /Branch Manager		✓

The details from the above table form the basis for completing the data flows on the initial context diagram – the final version is reproduced in *Fig 3.21*. Note that the chairman has the responsibility for setting up the technology including when the conference will take place prior to each conference. All non-audio inputs are numeric as they are entered via a telephone keypad.

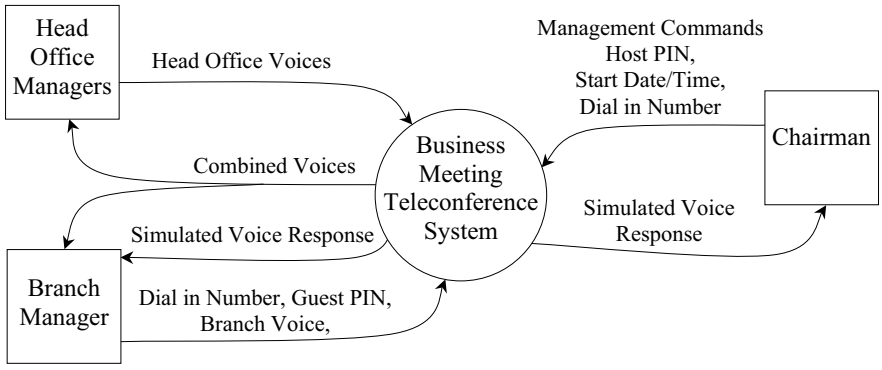


Fig 3.21
Final context diagram for a business meeting teleconference system.

Participants

The general manager and the four division managers at head office, one of which acts as the chairman. The five branch managers located in different country towns throughout the NSW.

Information Technology

- Standard telephones used by each branch manager to dial into the system, enter their Guest PIN and also to speak and listen during the conference.
- Polycom Sound Station 2W™ Wireless Conference phone used at head office (see *Fig 3.22*). The Polycom Sound Station 2W™ includes three high quality microphones to collect head office participant's voices. It also includes a high quality speaker for displaying audio from branch managers. The conference phone is full-duplex to allow branch voices to be heard whilst head office participants are speaking.

Fig 3.22
Polycom Sound Station 2W conference phone.

- Teleconferencing server controlling a PABX (Private Automatic Branch Exchange) that connects the PSTN circuits originating from head office with each of the PSTN circuits originating from the branches (see *Fig 3.23*). This server is maintained by a teleconferencing company who charges for its service on a per minute per connection basis for each conference.
- PSTN used to transmit and receive all data. The data is in analog form at each branch, at head office and also as it enters the PABX at the Teleconferencing Company.

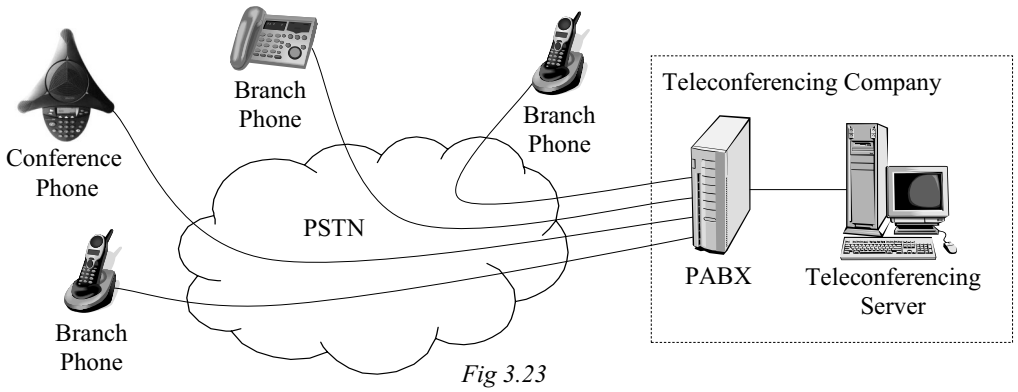


Fig 3.23

Network diagram including significant hardware within the business meeting system, sharing audio over the PSTN.

Information Processes

The following processes occur during a typical teleconference:

- Step 1. Setup by chairman

Prior to the teleconference the chairman rings the phone number of the teleconferencing server (Dial in Number). The chairman enters the Host PIN and is then prompted by the server to configure the conference. The server uses simulated voice prompts and the chairman responds by entering numbers through their phone keypad. The configuration includes the date and time of the conference together with the creation of a Guest PIN. The chairman provides the time and Guest PIN to each of the branch manager participants.

- Step 2. Participants enter conference

Just prior to the scheduled start time the chairman dials the teleconferencing server and enters the Host PIN using the conference phone. They follow the voice prompts to commence the conference. To join the conference each branch manager participant dials the "Dial in Number" and enters the Guest PIN. The teleconferencing server directs the PABX to connect the telephone line from each branch manager participant to the head office line. Once all branch managers have dialled in the conference can commence. The company pays a per minute charge for each connection used during a teleconference.

- Step 3. Conference takes place

During the teleconference all participants' voices are transmitted and received along the same single circuit. As is the case with any standard phone call, each local telephone only displays remote voices (and other audio). Prior to display local audio is filtered from the signal by the local phone.

- Step 4. Conference ends

The conference ends automatically when the conference phone hangs up. This occurs as soon as the teleconferencing server detects that the phone line that commenced the conference has been disconnected. The teleconferencing server then calculates the charge for the conference based on the total conference time and the number of participants.

GROUP TASK Activity

Create a step-by-step description of the steps required to setup and run one of the business teleconferences.

Advantages/Disadvantages

Advantages include:

- Reduction in costs associated with travel and accommodation. Furthermore branch managers are not absent from their offices as often and unproductive travel time can be used more productively.
- No additional hardware or software required apart from the conference phone at head office. There is no need for onsite technical help as the technical side of the conference has been outsourced to the teleconferencing company.
- Simple to setup and schedule conferences as required. Face to face meetings must be scheduled well in advance, whilst teleconferences can occur when and as required. This allows urgent decisions and issues to be resolved and information to be disseminated more efficiently.
- More regular communication between the complete management team results in better informed decisions and improved communication of these decisions. Furthermore issues occurring at the local level are better understood by head office, hence more appropriate solutions result.

Disadvantages include:

- Face to face communication includes body language and facial expressions – such communication is totally lost using a voice only system.
- Branch managers are not physically present, whilst division managers and the general manager are. This reduces the ability of branch managers to develop close inter-personal relationships with other members of management.
- It is difficult to maintain concentration during extended phone calls. From the branch manager perspective each teleconference is essentially an extended phone call.

GROUP TASK Discussion
The business described above has outsourced the technical side of its teleconferencing. Identify advantages and disadvantages of outsourcing in this situation.

2. DISTANCE EDUCATION SYSTEM, SHARING AUDIO, VIDEO AND OTHER DATA USING BOTH THE PSTN AND THE INTERNET.

Environment/Boundaries

In this example we consider a teleconferencing (or web conferencing) system used by ABC University. The system transmits audio over the PSTN using a system similar to the previous business meeting system. The system also transmits and receives live video and other digital data using IP over the Internet. Various University courses use the system so that students at remote sites can both observe and contribute to live presentations as they occur in front of local students.

The presenter and the local students are present within a purpose built teleconferencing room at ABC University. Each remote student connects to the conference via a standard telephone line for audio content and via a web browser running on a personal computer with a broadband Internet connection for video and other data.

Purpose

Students at ABC University are able to complete many degrees as either full-time on-campus students or as part-time off-campus students. The teleconferencing system aims to provide the off-campus students equal access to live presentations without the need for lecturers to duplicate or significantly modify their presentations.

The purpose of this teleconferencing system is to:

- Enable remote off-campus students to be equal participants in live presentations.
- Remove the need for lecturers to prepare different material for on and off campus students.
- Allow individual remote students to connect to teleconferences using their existing hardware and broadband Internet connections.
- Allow presenters to seamlessly operate the technology with minimal disruption to the local student's view of the presentation.

Data/Information

Data/Information	*Data type*	*Description*
Participant Audio	Audio	Audio from the teleconferencing room and remote students is added to a shared PSTN circuit.
Combined Audio	Audio	Mixed audio from all sites is present on the shared PSTN circuit.
Participant Video	Video	Video from the teleconferencing room and each remote student is transmitted using IP and the Internet to a remote chat and video conferencing server.
Video Stream	Video	Video from the chat and video server is transmitted using IP to participant's web browsers. A separate stream is used for each connection and is tailored to suit the actual speed of the individual connection..
Application Data	Various	Includes data to enable the sharing of documents, virtual whiteboard, desktops and other types of digital data. This includes the ability to concurrently edit the virtual whiteboard and single documents.
Chat Data	Text	The system includes an instant messenger chat feature. Chat data can be broadcast to all participants or between specific individuals. All chat data passes through the Chat and Video Conferencing Server.
Conference IP Address	Numeric	The IP address of the conference management server used by all participants to connect to the system.
Participant IP Address	Numeric	The IP address of each computer participating in the conference.
Dial in Number	Numeric	Used to connect voice via the PSTN to the remote telephone conferencing server.
Student PIN	Numeric	Used by students to verify their identity as they initiate telephone and web sessions.
Presenter PIN	Numeric	Used by the presenter to verify their identity as they initiate telephone and web sessions.

Participants

- Lecturers who present material from the purpose built teleconferencing room.
- Full-time students who are present within the teleconferencing room.
- Part-time students who connect to the teleconference presentation from their own home or office.

Information Technology

Fig 3.24
Purpose built audio/video/web teleconferencing room.

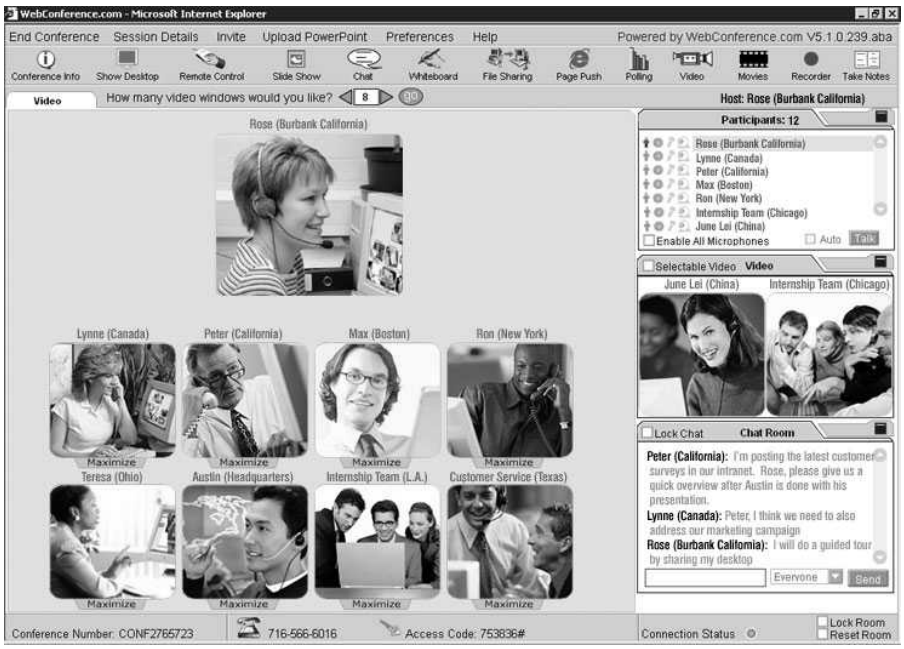


Fig 3.25
WebConference.com™ software within Internet Explorer.

Teleconferencing room:

- Personal computer with web browser, WebConference.com™ software and high-speed Internet connection.
- Three large monitors – one for displaying video of participants, another for other application data. The third monitor is used to display data to the presenter so they do not need to turn away from their audience.
- DLP data projector used by the presenter to display any data source to the local students using a remote control.
- Document camera for collecting images and video of paper documents as well as 3D objects.
- Video camera with pan, tilt and focussing functions as well as the ability to follow the current speaker's voice.

- DVD and video player – the output can replace the normal video camera.
- High quality microphones throughout the room. The main presenter wears a lapel microphone. The microphone system includes echo cancelling so that audio from the speakers is not retransmitted.
- High quality speaker system optimised for voice frequency output.

Remote Students:

- Personal computer with web browser connected to a broadband Internet connection.
- WebConference.com™ software which is downloaded and run automatically within the student's browser – an example screenshot is reproduced above in *Fig 3.25*.
- Web camera for collecting local video.
- Standard telephone, however a headset is recommended.

Teleconferencing Service Provider (in this example WebConference.com™):

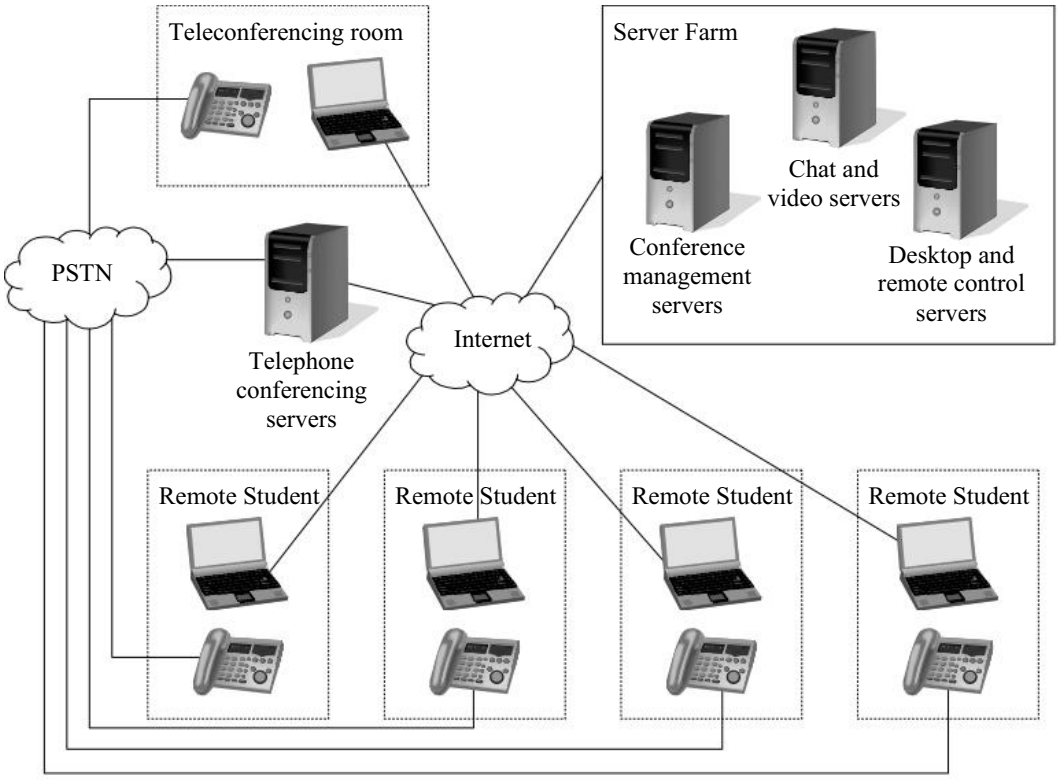


Fig 3.26
Network diagram including significant hardware within the WebConference.com™ system, sharing audio over the PSTN and IP data over the Internet.

- Multiple server farms (see *Fig 3.26*) that include collections of the following servers in a variety of different locations throughout the world.
- Conferencing management server used to control the setup and running of each conference. This includes directing connections to other servers and other server farms before and during the conference to ensure a continuous high quality of service.
- Chat and video server receives video and chat data from all participants and transmits this data out as required. The server creates and transmits suitable streams of video data to each participant's web browser based on the current speed of each participant's Internet connection.

- Desktop and remote control server used to receive and transmit application data. For example the presenter may share an open Word document on their local machine such that remote students can edit the document synchronously.
- Telephone conferencing server used to connect all PSTN lines from all participants to form a single shared circuit.

Information Processes

Some general collecting and displaying information processes occurring include:

- Collecting – audio using telephone and conference room microphones, video using cameras, text using keyboard, images using document camera.
- Displaying – audio using speakers in conference room and speaker in remote student's phones, video and other data types are displayed on monitors and using the DLP data projector.

Let us consider how video is transmitted and received in some detail. The data flow diagram in *Fig 3.27* describes this process for a single stream travelling from the teleconferencing room to a single remote student – clearly there are potentially numerous other streams travelling in all directions between all participants. The points that follow elaborate on the DFD:

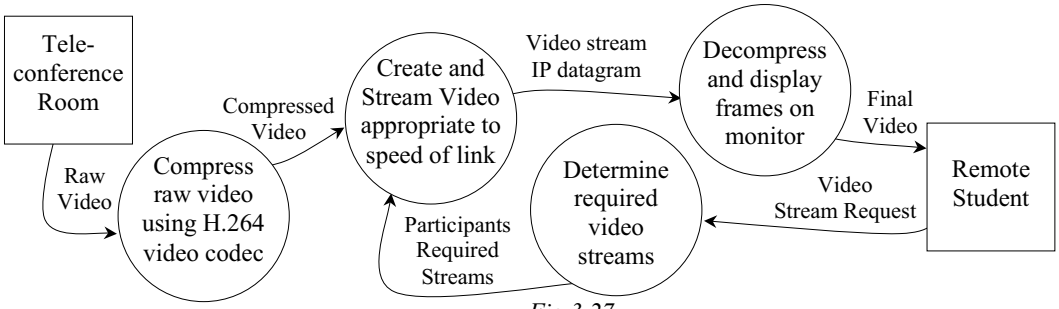


Fig 3.27
DFD describing the transmission of a single video stream.

- Raw video is collected as a sequence of images called frames by the video camera. For many applications the video camera includes a microphone and hence sound samples are also collected – within this example system no audio is collected by the video cameras. The raw frames from the conference room are collected at a far higher resolution than those collected from each remote student's web camera.
- The raw video frames are fed in real time through a software-based codec. In this example the MPEG-4 part 10 or H.264 codec is used. A codec is used to compress and decompress data prior to and after transmission. The codec compresses the video using an efficient block-based compression technique. We discussed block-based coding in some detail on page 59 and 60 of the related IPT Preliminary Text.
- The compressed video data is transmitted via the Internet to the Chat and Video server. This server determines which streams of video data each participant requires and prepares to transmit just those streams to the participant's web browser. For example typically a remote student will view video from the teleconference room and perhaps streams from two or three other remote students.
- Each chat and video server includes streaming video server software. This software is able to determine the optimum transmission speed for each participant's Internet link. The job of the streaming server is to adjust the resolution and frame rate of each video stream to maximise the quality of the video transmitted to each participant. For example a slower link will receive smaller and fewer frames than a faster link. Furthermore the quality of the video can be altered by the streaming server in real time should the speed of a link change during the conference.

- The stream of video is ultimately transmitted as a sequence of IP datagrams. Higher resolutions and frame rates require more IP datagrams per second than lower resolutions and frame rates.
- As the stream of IP datagrams are received the same H.264 codec is used by the receiver's computer to decompress the video. Finally the decompressed frames are displayed on the receiver's monitor.

Advantages/Disadvantages

For this example we restrict our advantages/disadvantages to those concerned with technical aspects of the system.

Technical advantages include:

- Remote students do not require any specialised or dedicated information technology apart from the free and automatically installed WebConference.com™ software operating within their browser.
- Video streams are automatically adjusted to suit the speed of each participant's Internet connection. This means lower speed connections receive a continuous video experience, albeit at reduced resolution and frame rates.
- The quality of audio is not affected by poor or congested Internet connections. The PSTN provides an audio signal of equal quality to all remote participants. Even if a student's Internet connection is lost the audio is still active.
- The system includes redundant servers and server farms so that failure of a single server or connection to a single server farm does not disrupt conferences.

Technical disadvantages include:

- Some remote students will experience poor quality video due to slower Internet connections. Most remote students are likely to receive video of somewhat lower quality compared to those students present within the teleconferencing room.
- Most remote students connect from their home. Therefore their home telephone is tied up for the duration of each conference.

GROUP TASK Discussion
Identify and describe more general advantages and disadvantages of the above system for each of the system's participants.

Consider the following:

During a conference the same video stream originating from the teleconferencing room is being sent multiple times as a separate stream to each remote student. This system is an example of a multipoint Unicast transfer. There are currently two types of multipoint transfer that can be used over an IP network – Unicast and Multicast. Unicast is a point-to-point system where each IP datagram travels to exactly one recipient – this is the normal method currently used to transfer virtually all IP datagrams across the Internet. Multicast is a one-to-many system where a single IP datagram is sent to many recipients.

The multicast system requires a multicast destination IP address within the IP datagram. During transmission of a multicast IP datagram each router examines the multicast destination address and may then decide to forward the datagram along more than one connection. The multicast system has the potential to significantly improve the speed of transfer for streamed video (and also audio) over the Internet. Although many current routers include support for the required multicast protocols there are many that do not and there are many other routers where multicasting is turned off – multicast IP datagrams arriving at such routers are simply discarded.

GROUP TASK Research
Using the Internet, or otherwise, identify and briefly describe the protocols used by routers to route multicast IP datagrams.

GROUP TASK Discussion
Explain how multicasting can significantly speed up the transfer of streamed audio and video.

HSC style question:

A company has won a contract to supply security infrastructure and personnel for the 2008 Beijing Olympics. The company has offices in Sydney, London, New York and now Beijing. Each week the senior management at all offices participate in a teleconference over the Internet that includes both audio and video.

(a) Compare and contrast the use of teleconferencing with traditional telephone and face-to-face communication in this situation.

(b) Identify and briefly describe the information technology required by this teleconferencing system.

(c) Describe how data is transmitted and received between offices during one of the weekly teleconferences.

Suggested solution

(a) Both teleconferencing and traditional methods allow people from different offices in different parts of the world to communicate effectively. This teleconferencing system includes video in addition to audio. Multiple participants can hear and see the other participants of the conference. For this company the participants are located in different offices across the world. Therefore the system requires high speed Internet links to transmit the video and audio data. The quality of the video and audio is dependent on these public Internet links.

Face-to-face communication can only occur between people in the same location. This means face-to-face meetings would need to be scheduled at one of the offices (Sydney, London, New York or Beijing) and there would be large expenses and work time lost in getting people from the other offices in for the meeting. Furthermore it would be impractical for such face-to-face meetings to occur on a regular basis.

Traditional telephone is audio communication between two people over the PSTN – or three people, if a three-way conference call is possible. The participants can only hear the other person's voice, there are no visuals and so body language plays no part in the conversation, hence business and personal relationships are harder to build. This teleconferencing system assists in this regard as it includes video and it supports synchronous communication between many more participants.

In this example the audio is transmitted over the Internet. Due to the packet-switched nature of IP transmissions the audio will be of lower quality than is possible using a normal circuit-switched telephone line. Also the company does not control the Internet, hence transmission speeds between participants will vary which will affect the quality of both the audio and video.

The significant advantage of teleconferencing for an international company is that none of their workers need to leave their home country to participate in the conference. The use of teleconferencing reduces expenses (no plane and accommodation costs) and maintains productivity (no wasted hours on plane trips). It also allows the company to have frequent meetings at short notice and at relatively minimal cost.

(b) The hardware required by each participant includes a video and audio capture device at each location. This is likely to be a simple webcam with microphone. Each location must also have a screen in which to display the images from each location as well as speakers to play the audio. Inside the computer there needs to be a sound and video card.

A high-speed network link to the Internet is needed so that the data (video and audio) can be transmitted and received in nearly real time. Faster links resulting in high resolution and smoother video together audio that is in sync with the video. This means that each office will require a fast broadband Internet connection.

Software is required that captures the video and audio and streams across the Internet to the teleconferencing server. In this case the video and audio would be combined (multiplexed) and sent together as a continuous stream of IP datagrams.

A teleconferencing server is needed with multiple high-speed Internet links. It receives the streams from each participant and sends out an individual video/audio stream to each participant. Multicasting is unlikely to be possible as the transmission is over the public Internet.

(c) At a teleconference each participant's analog data is captured as digital video frames and digital sound samples. This data is then multiplexed and compressed together using a codec such MPEG 4. The data is then streamed over the Internet to the teleconferencing server as a sequence of IP datagrams.

The teleconferencing server receives the video/audio streams from each participant. It also determines the particular streams requested by each participant and the current speed of their individual transmission links. The server then produces a suitable stream for each participant that will maximise the quality of his or her received video and audio. The stream sent is altered during the conference in response to changing transmission speeds.

At each participant location the received data is decompressed and then broken down into the audio and video components. Finally the audio samples are converted to analog and output through the speakers. As this occurs the video frames are displayed in sequence on the participant's screen.

Comments

- In an HSC or Trial examination this question would likely be worth nine marks – three marks for each part.
- A multicast system could be described, however at the time of writing there were few Internet connections that support IP multicasting between different countries.
- Presently most business teleconferencing systems use the PSTN for audio. In this case the question states that the Internet is used for both video and audio.
- A conference phone could be used at each office as it is likely that more than one participant is present at some locations.

SET 3C

1. During a telephone call over the PSTN, which of the following is TRUE?
   (A) Data can travel over a variety of different routes during a conversation.
   (B) A single connection is maintained for the duration of the call.
   (C) The data is split into packets that travel independently of each other.
   (D) The same circuit may be shared with IP and other voice data.
2. Which of the following terms best describes a private WAN connecting a company's various offices?
   (A) Intranet
   (B) Extranet
   (C) Internet
   (D) PSTN
3. The PSTN is currently used for audio in many teleconferences because:
   (A) voice quality is better on a connectionless network.
   (B) currently multicasting is not widely implemented on the Internet.
   (C) circuit switched networks provide higher levels of security.
   (D) voice quality is better on a connection-based network.
4. When participants are widely dispersed, which of the following is an advantage of teleconferencing systems compared to face-to-face meetings?
   (A) Ability to develop personal relationships is enhanced.
   (B) Specialised information technology is required.
   (C) Significant savings in terms of money and time.
   (D) All of the above.
5. Which of the following is TRUE for PSTN based audio conferences?
   (A) Each participant has a different circuit.
   (B) Audio from each participant is transferred as a sequence of packets.
   (C) All participants share a single circuit.
   (D) Each participant must use a dedicated conference phone.
6. The purpose of a streaming video server is:
   (A) to adjust the quality of the video stream sent to each participant based on their transmission speed.
   (B) to transmit identical streams of video to all conference participants.
   (C) to ensure a continuous connection between all participants is maintained.
   (D) to connect and disconnect participants as they enter and leave the conference.
7. With regard to the video received during a videoconference, which of the following is TRUE?
   (A) All participants in a video conference must receive video of identical quality.
   (B) The quality can never exceed that of the collected video.
   (C) The codec used by the sender can be different to the codec used by the receivers.
   (D) Video quality decreases as transmission rates increase.
8. When IP multicast is used, which of the following occurs?
   (A) Each participant receives the same stream.
   (B) Each participant receives their own stream.
   (C) A dedicated streaming server is definitely required.
   (D) Video cannot be sent from multiple locations.
9. Teleconferencing can best be described as:
   (A) synchronous and simplex.
   (B) asynchronous and full duplex.
   (C) asynchronous and simplex.
   (D) synchronous and full duplex.
10. Which list contains devices used to collect data during teleconferences?
    (A) Phone, monitor, keyboard, mouse.
    (B) Speakers, monitors, headsets, projectors.
    (C) Phone, video camera, document camera, keyboard, mouse.
    (D) Video camera, document camera, speakers, scanners.
11. Define each of the following terms:
    (a) Internet
    (b) PSTN
    (c) Intranet
    (d) Extranet
    (e) Teleconference
12. Compare and contrast IP unicasting with IP multicasting with regard to their use in teleconferencing systems over an intranet and over the Internet.
13. Explain the differences between packet switched connectionless networks and circuit switched connection-based networks.
14. Outline the processes performed by teleconferencing servers when:
    (a) sharing audio over the PSTN.
    (b) sharing video over the Internet.
15. Compare and contrast teleconferencing systems with face-to-face meetings.

MESSAGING SYSTEMS

In this section we first consider the basic operation of traditional phone and fax systems operating over the PSTN. We then consider enhancements to the traditional phone system to include voice mail and information services. We then consider VoIP, a system for making phone calls using the Internet. Finally we examine the characteristics of email and how it is transmitted and received.

1. TRADITIONAL PHONE AND FAX

Telephones

Telephones and the PSTN network connecting homes and organisations operate using similar principles as the original system first implemented over 100 years ago. Essentially all telephones have a microphone, a speaker, some sort of bell and a simple switch to connect the phone to the telephone network. A 100-year-old phone will still operate on most of today's phone lines. The only significant difference being the signals used to dial numbers – older phones use pulse dialling whereas current phones use tone dialling. When pulse dialling, the phone switch is rapidly disconnected and connected the same number of times as the number being dialled – techniques included tapping the hook the required number of times or rotating a dial. Tone dialling transmits different frequencies to represent each number.

Fig 3.28
Rotary dial telephone in common use from 1940-1990.

In many older homes the copper wires connecting the phone to the PSTN network have been in place for many more years than originally intended, it is what happens once the wires reach the local telephone exchange that has changed. In the past, actual mechanical switches were used to connect the copper wire from your home phone directly with the copper wires connected to the phone being called. Circuit switching creates a direct connection or circuit between the two phones. In the days of manual switchboards, operators would manually connect the wires running from your home with the wires running to the person's phone you wished to call. Although manual switching has now been completely replaced by electronic switching, the PSTN circuit switched network operates using this very same connection-based principle, that is, a direct connection is setup and maintained whilst each conversation takes place.

During a typical conversation we spend less than half the time listening, less than half the time speaking and the remaining time in relative silence. This is not such a concern between a phone and its local exchange, however over longer distances the inefficiencies are significant. Today, apart from the connection between telephones and their local exchange, the remainder of the PSTN is essentially digital. Digital networks make much more efficient use of the lines. By digitising the analog voice signals it becomes possible to compress the bits and also to combine (multiplex) many conversations on a single physical connection. This means many conversations share the same line simultaneously. Various different modulation schemes are used depending on the range of frequencies used and the physical attributes of the cable. For example time division multiplexing (TDM), used on tier 1 (T1) lines, samples each voice 8000 times per second and each of these samples is coded into 7-bits. A total of 24 voice channels are combined onto a single copper circuit. Most medium to large organisations do away with analog lines altogether, rather they have one or more T1 lines that directly enters their premises.

It is the digital nature of most of the PSTN that has allowed most phone companies to provide their customers with additional features, such as call waiting, caller id, three-

way calls, call diversion and voice mail. The processing required to implement these features occurs at the telephone exchange – the customer sends commands to access and control the feature using tones generated by their phone's keypad. Furthermore much of the PSTN's digital infrastructure is used to transmit IP data across the Internet.

GROUP TASK Discussion
Explain the difference between analog and digital voice signals. Why do you think analog signals are still used between most phones and their local telephone exchange? Discuss.

Facsimile (Fax)

Alexander Bain first patented the basic principle of the facsimile, or fax machine, in 1843. Incredibly this is some 33 years before the telephone was invented. It was some twenty years later that the first operational fax machines and transmissions commenced. Initially it seems odd that fax pre-dates telephones, however in fact it makes sense. At this time the telegraph system using Morse code was in operation. Morse code was transmitted by opening and closing a circuit, which is similar to the binary ones and zeros used by today's fax machines.

It wasn't until the late 1960s that fax machines became commercially viable; these machines adhered to the CCITT Group 1 standard, which used analog signals and took some 6 minutes to send each page. The message was sent as a series of tones, one for white and another for black, these tones were then converted to an image using heat sensitive paper. By the late 1970s the fax machine had become a standard inclusion in most offices. A new Group 2 standard was introduced; these Group 2 machines generated digital signals and used light sensors to read images on plain paper originals. Soon after machines were developed that used inkjet and laser printer technologies to print directly onto plain paper. The Group 3 standard was introduced in 1983; it contained various different resolutions together with methods of compressing the digital data.

Fig 3.29
Fax machines are standard items in almost all offices.

Today computers are routinely used to produce, send and receive faxes; in fact most dial-up modems have built in fax capabilities. There are even Internet sites that allow a single fax to be broadcast to many thousands of fax machines simultaneously. It is common today for a single device to integrate scanning, faxing and printing.

GROUP TASK Discussion
Brainstorm specific examples where fax has been used. For each example, discuss reasons why fax has been used in preference to phone, email or other messaging systems.

2. VOICE MAIL AND PHONE INFORMATION SERVICES

Voice mail, in its simplest form, is much like a digital version of a traditional answering machine. Calls that are not answered after a predefined number of rings are diverted to the voice mail system. The voice mail system answers the call and plays a pre-recorded outgoing message (OGM). The OGM welcomes the caller and provides instruction on how to leave a message – for residential phones the OGM may be as simple as "Hi, you've reached Sam, please leave a message after the tone and I'll get back to you ASAP." The voice mail system then digitally records the users voice and

stores it within the customers voice mailbox. At some later time the customer rings the voice mail system, verifies their identity using a numeric password and listens to the voice messages held in their voice mailbox. During message retrieval the customer uses their phone keypad to enter commands that control the voice mail system. No doubt we are all familiar with such systems.

GROUP TASK Activity
Create a DFD to describe the data flows, external entities and basic processes in the simple voice mail system described above. Include just two processes – "Leave Message" and "Retrieve Messages".

The familiar voice mail system described above is normally a service provided by the customer's local telephone service provider – Telstra, Optus, Orange, etc. The servers used to process messages are located and owned by the telephone company. More sophisticated voice mail systems are used by business and government organisations. These organisations maintain their own systems. Such systems include a multitude of features designed to meet the needs of the individual organisation and its customers. They do a lot more than maintaining voice mail for many users. Commonly such systems integrate with other messaging systems such as email and fax, and they provide automated information services and call forwarding functionality to customers. For our purposes we more accurately describe such systems as Phone Information Services.

The majority of phone information systems include a hierarchical audio menu whereby customers navigate down through the hierarchy of menus to locate information or be directed to specific personnel. The available options at each level of the hierarchy are read out as an OGM, the customer responds using their phone's keypad or using voice commands to progress to the next level.

Some of the features present within Phone Information Services include:

- Voice mail management for many users. Customers enter the extension number of the required person and if not answered the system records the message to the person's mailbox.
- Support for multiple incoming and outgoing lines of different types. Today large organisations will have many digital T1 lines connected directly to the PSTN and also VoIP (voice over IP) lines connected to the Internet via broadband connection.
- Fax on demand where customers navigate a menu system to locate and request particular documents to be faxed back.
- Call attendant functions where the menu system filters callers through to the correct department based on the caller's selections. Some systems can forward calls to other external lines.
- Text to speech (TTS) capabilities that allow text to be read to users over the phone. For example, TTS can be used to read emails and other text documents or more simply it is often used to read numbers and currency amounts back to customers to verify their data entry.
- Call logging to databases. For example records commonly include the caller id, time and length of call. This data is analysed to provide management information to the organisation.
- Provision of information to customers. The OGMs include information rather than just details of how to navigate the menu system. For example, in Australia numbers with the prefix 1900 provide such information on a user pays basis.

- Automated ordering systems that allow customers to order and pay for products without the need for a human operator. Often includes collecting and verifying credit card payments.
- Automated surveys where answers to questions are stored within a linked database. Some commercial surveys use the 1900 system or the SMS system where the user is charged on their telephone bill for their contribution. The telephone company forwards the funds to the survey provider.
- Integration of voice mail with other messaging systems. For example voice mail messages are converted to email messages and appear in the recipients email inbox. The email can include the voice message as an audio attachment or the audio can be converted to text using voice recognition.

GROUP TASK Discussion
Brainstorm a list of phone information services members of your class have used. Identify and briefly describe features within these services.

GROUP TASK Research
Currently VoIP is becoming a popular alternative to standard PSTN lines. It is likely that by the time you read this it will be a routine method for making phone calls. Research VoIP and describe its essential differences compared to traditional telephone lines.

Consider the following:

ISO/IEC 13714 is the international standard for interactive voice response (IVR) systems. Recommendations within this standard include how each key on a standard telephone keypad should be used when designing menus for IVR systems. These recommendations include:

- # key – used to delimit data input or to stop recording and move to the next step. It can also be used as a decimal point. The preferred name for # is "hash".
- * key – used to stop the current action and return the caller back to the previous step. Often this means the last OGM will replay. When entering data the * key should clear the current entry. The preferred name for * is "star".
- 0 key – if possible the 0 key should be used to transfer the call to an operator or to provide help on the current feature or action. The preferred name for 0 is "zero".
- 9 key – used to hang-up the call where this is a suitable option.
- Yes/No responses – the 1 key should be used for Yes and the 2 key used for No.
- Alpha to numeric conversions – America and the rest of the world use slightly different mappings. To ensure IVR systems work on both systems the following mappings should be used:

1 – QZ	4 – GHI	7 – PQRS
2 – ABC	5 – JKL	8 – TUV
3 – DEF	6 – MNO	9 – WXYZ

  Note that 1 and 7 map to Q and that 1 and 9 map to Z.
- OGMs should refer to numbers on the telephone keypad not letters.
- OGMs should be phrased with the function first followed by the key to press. For example "To pay an invoice press 2".
- Menu OGMs should be in ascending numerical order with no gaps in numbering.

- Commonly used functions should be listed first. For example pressing 1 causes the most commonly used function to activate.
- In general menus should be limited to 4 commands (excluding help, operator transfer, back and hang-up commands).

GROUP TASK Discussion
In your experience, have these recommendations been implemented within phone information services you have used? Discuss reasons for the existence of the ISO/IEC 13714 standard.

Storyboarding and simulating an example IPT Phone Information Service

In this example we shall develop a phone information system to provide basic information about the IPT HSC course and each of its component topics. In addition the system will be able to record student's questions into a voice mailbox corresponding to the topic.

Consider the essentially hierarchical storyboard reproduced in *Fig 3.30*. Each rectangle on this storyboard corresponds to an OGM (outgoing message) – some OGMs are menus, others simply provide information and some do both. Think of an OGM as the audio version of a screen on a normal storyboard – both screens and OGMs display data. The lines between each OGM rectangle include the key used to navigate from OGM to OGM. Notice that a line exists from each topic to the voice mailboxes. In the final system a separate mailbox will be maintained for each topic. Each mailbox is linked to the email address of an expert on that topic. When a question is left by a student caller it is immediately emailed to the corresponding topic expert. The email includes the phone number (CallerID) of the student caller together with an audio file attachment and the topic name.

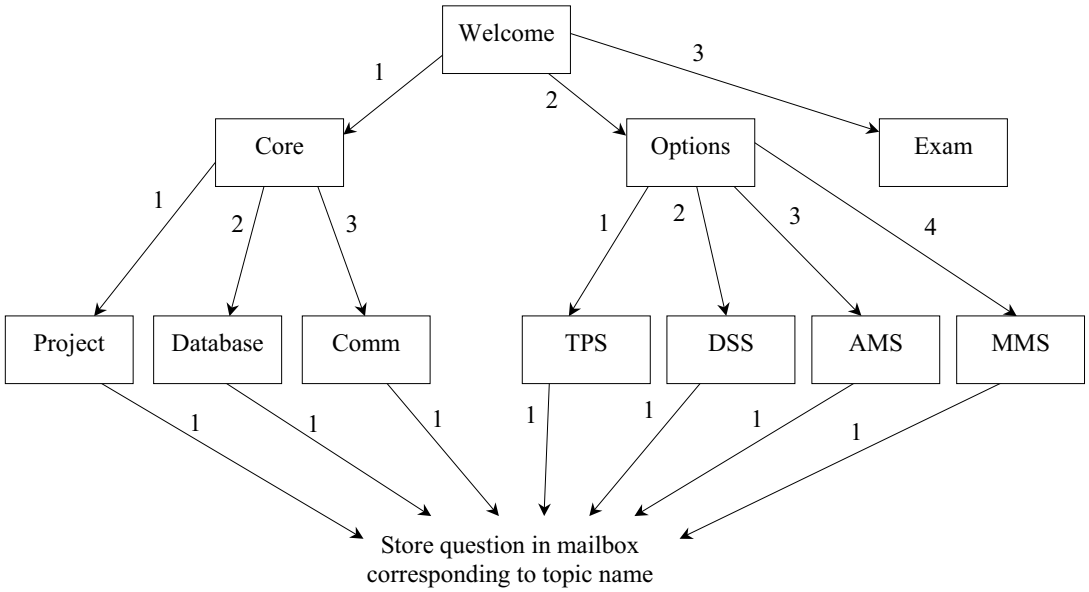


Fig 3.30
Storyboard showing the links between OGMs for the example IPT Phone Information Service.

When we create a storyboard for a user interface we also create designs for each individual screen. When designing OGMs we need simply design the text that will be spoken (or synthesised) for each OGM. The table in *Fig 3.31* details the text of each OGM together with actions performed in response to user key presses.

OGM Name	*Text*	*Action*
Welcome	Welcome to the IPT HSC command centre. We provide general information and answers to specific questions on all topics. For core topics please press 1. For option topics please press 2. For HSC examination details press 3.	1- go to Core OGM 2- go to Options OGM 3- go to Exam OGM *- repeat Welcome OGM 9- end call.
Core	There are 3 core topics each worth 20 percent. For project work press 1. For Information systems and databases press 2. For Communication Systems press 3.	1- go to Project OGM 2- go to Database OGM 3- go to Comm OGM *- go to Welcome OGM 9- end call.
Options	There are 4 options of which 2 must be completed. For Transaction Processing Systems press 1. For Decision Support Systems press 2. For Automated Manufacturing Systems press 3. For Multimedia systems press 4.	1- go to TPS OGM 2- go to DSS OGM 3- go to AMS OGM 4- go to MMS OGM *- go to Welcome OGM 9- end call.
Exam	The IPT HSC Examination is a 3 hour exam that contains 3 sections. Section 1 is worth 20 marks and is composed of 20 multiple choice questions based on the 3 core topics. Section 2 is worth 40 marks and is composed of 4 free response questions based on the 3 core topics. Section 3 is worth 40 marks and contains one 20 mark question for each option topic. You must complete 2 questions. To return to the previous menu press the star key.	*- go to Welcome OGM 9- end call.
Project	Project work involves planning, designing and implementing an information system that has a specific purpose. To leave a question about project work press 1.	1- leave message in Project mail box *- go to Core OGM 9- end call.
Database	Information systems and databases emphasises the organising, storing and retrieving processes within database systems and hypermedia. To leave a question about information systems and databases press 1.	1- leave message in Database mail box *- go to Core OGM 9- end call.
Comm	Communication systems support people by enabling the exchange of data and information electronically. This topic emphasises the transmitting and receiving processes. To leave a question about communication systems press 1.	1- leave message in Comm mail box *- go to Core OGM 9- end call.
TPS	Transaction processing systems meet record keeping and event tracking needs of organisations. To leave a question about transaction processing systems press 1. To go back to the previous menu press the star key.	1- leave message in TPS mail box *- go to Core OGM 9- end call.
DSS	Decision support systems use models, analytical tools, databases and automated processes to assist decision making. To leave a question about decision support systems press 1. To go back to the previous menu press the star key.	1- leave message in DSS mail box *- go to Core OGM 9- end call.
AMS	Automated manufacturing systems gather data through sensors, process this data and send signals to actuators that perform some mechanical task. To leave a question about automated manufacturing systems press 1. To go back to the previous menu press the star key.	1- leave message in AMS mail box *- go to Core OGM 9- end call.
MMS	Multimedia systems combine different types of data. To leave a question about multimedia systems press 1. To go back to the previous menu press the star key.	1- leave message in MMS mail box *- go to Core OGM 9- end call.

Fig 3.31
Details of each OGM in the example IPT HSC Phone Information system.

To implement this IPT phone information system requires either VoIP or traditional phone lines. Analog PSTN lines connect to a computer via voice modems or a purpose built telephony board. Digital lines such as ISDN or T1 still require modems to convert the digital data to and from the computer. Many current ISDN and T1 modems support both circuit switched PSTN lines and also IP Internet data – including VoIP. In each case the software controlling the processing receives digital audio data from callers via the modem and sends digital audio data to callers via the modem. We restrict our discussion of the information technology to an example software application called "IVM Answering Attendant" that is written and distributed by NCH Swift Sound. At the time of writing a shareware version of this product was available for evaluation purposes.

IVM Answering Attendant includes a call test simulator – a screen shot is reproduced in *Fig 3.32*. This simulator plays OGMs through the computer's speakers. The computer's microphone and the onscreen phone keypad are used to record voices and enter commands. This feature is used to test the OGMs and actions during the design of the solution.

Fig 3.32
Call Test Simulator within IVM Answering Attendant.

Each OGM is created and edited using the OGM Manager within the software. The text for each OGM can be entered and then converted to audio using a TTS (text to speech) engine or it can be recorded directly using a microphone. The properties window for each OGM includes a Key Response tab (see *Fig 3.33*) where the actions to perform in response to user key presses are specified. For example *Fig 3.33* shows the Core OGM where the response to pressing 2 is being specified – go to Database OGM.

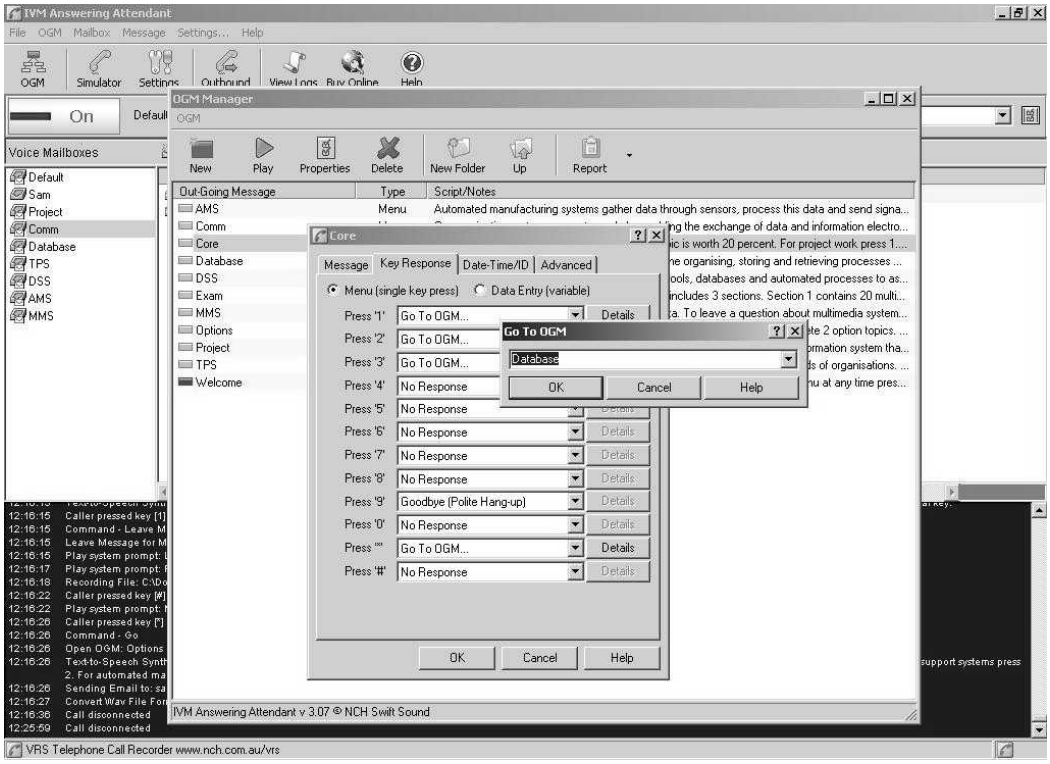


Fig 3.33
Specifying key response actions to OGMs within IVM Answering Attendant ©NCH Swift Sound.

Each mailbox (in our example system there are seven) includes various delivery options as shown in *Fig 3.34*. For our IPT Phone Information system each voice mail message is emailed to the topic expert as an audio file attachment. The text of the message includes the mailbox name together with the CallerID (phone number) of the student.

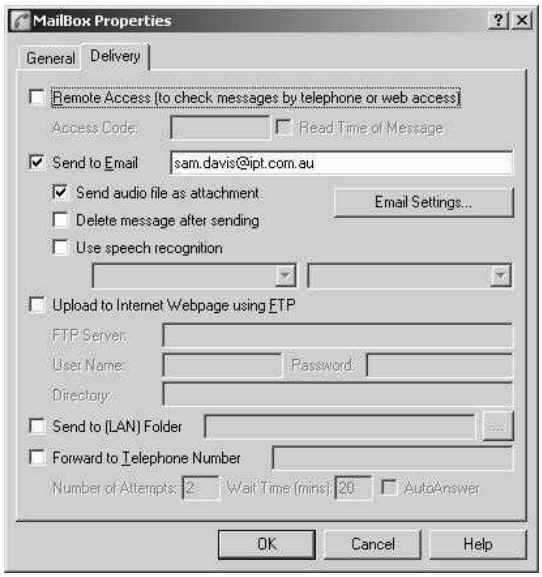


Fig 3.34
Mailbox delivery options in IVM Answering Attendant ©NCH Swift Sound.

Speech recognition is also possible, however currently most speech recognition engines are only accurate when they have been trained to a specific user's voice. As a consequence speech recognition is only a viable option when single words from a specific set of possible words are used. For example Yes/No questions or perhaps a suburb name – such data can be validated using TTS to read back the callers input. For most voice mail systems the first Remote Access option is selected. This allows users to retrieve their messages from any telephone (or over the web) by entering their access code.

GROUP TASK Discussion
Brainstorm a list of example phone information systems where each of the options shown in *Fig 3.34* would likely be used.

GROUP TASK Research
Using the Internet, or otherwise, find examples of voice modems, telephony boards, ISDN modems, T1 modems and VoIP modems. Briefly describe the functions performed by each type of modem.

GROUP TASK Discussion
Identify the participants, information/data, information technology and information processes within the above IPT Phone Information System.

3. VOICE OVER INTERNET PROTOCOL (VoIP)

Voice over Internet Protocol, as the name suggests, transfers voice calls over the public Internet. VoIP is also known as IP Telephony, Voice over broadband and Internet telephony. All these names indicate some of the basics of VoIP – a broadband Internet connection is used to transfer telephone calls using IP. However it is possible to transfer voice over the Internet using any Internet connection combined with a microphone, speakers and one of the many free instant messaging applications, such as MS Messenger, for example. So how is VoIP different? VoIP goes one step further and provides an interface to the PSTN. This is the defining feature of VoIP – it allows VoIP calls to be made to any normal telephone across the globe. Furthermore calls are significantly cheaper as the public Internet carries the data for free regardless of distance. If both ends of the call are using VoIP then the commercial PSTN is not used at all – such calls are often free, apart from the cost of the Internet connection.

VoIP is not a single protocol rather it is suite of protocols. For instance, audio codecs are included to digitise and compress the analog voice data, and then decompress and convert it back to analog at the receiving end. Once the data has been converted from analog to digital it passes through a stack of protocols – commonly RTP (Real Time Protocol) and UDP (User Datagram Protocol) at the OSI Transport Layer 4 and then IP at the OSI network Layer 3. RTP is used to control streaming of data packets, including maintaining a constant speed and also keeping packets in the correct sequence. UDP is used rather than TCP as UDP fires off packets more rapidly without the overhead of error checking and flow control.

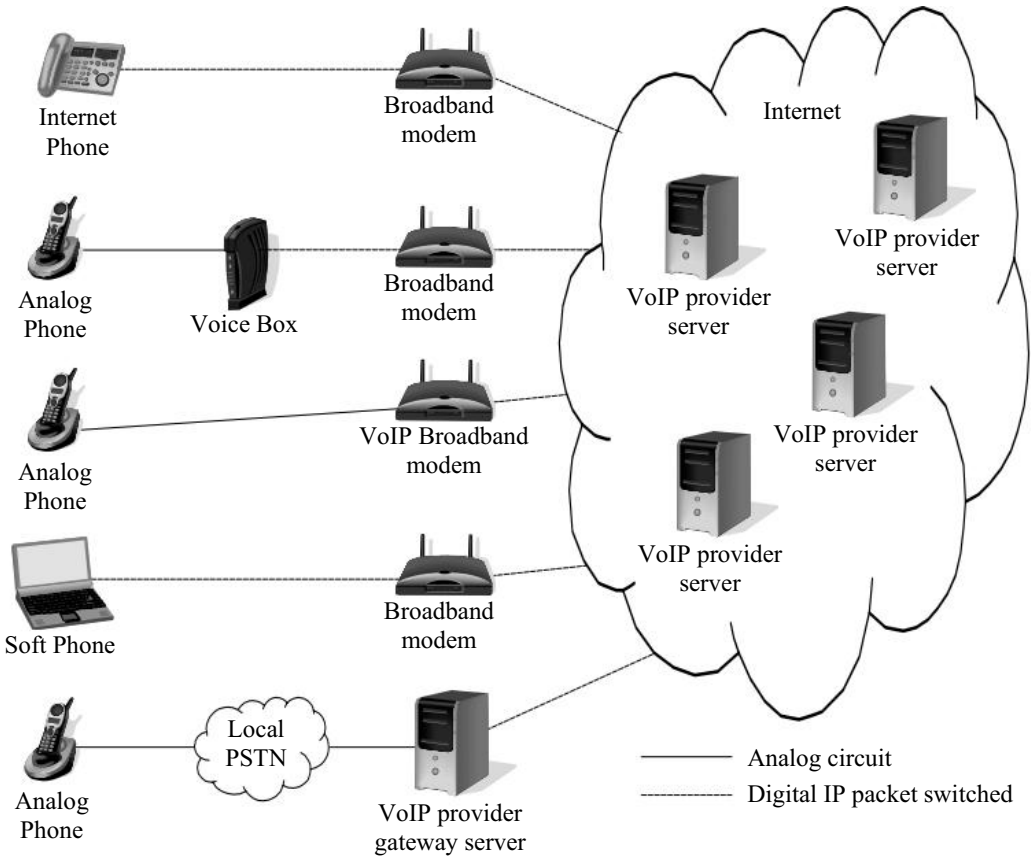


Fig 3.35
VoIP network diagram including different hardware combinations used to connect VoIP users .

There are various hardware combinations that are all commonly used to connect VoIP users – five possibilities are shown in *Fig 3.35*. The VoIP provider maintains one or more servers whose central task is to translate normal telephone numbers into IP addresses. VoIP providers also maintain gateway servers which convert analog phone calls to IP packets and viceversa – a gateway is a devcie that connects two different networks.

Users who sign up with a VoIP provider commonly connect using their existing broadband modem and Internet connection. Broadband modems are also available with built-in support for VoIP, in this case a standard analog telephone is simply plugged into the modem. Other possibilties include soft phones, where a VoIP software application operates on an existing Internet connected computer. Voice boxes are also available that connect existing analog handsets to existing broadband modems.

Now consider users who don't have an account with a VoIP provider, rather they have a traditional PSTN phone line. VoIP providers must maintain a network that allows

their customers to connect to phones on the PSTN. To implement this functionality and still save money on long distance calls requires VoIP gateway servers to be installed in locations throughout the world. Clearly it would not be economically viable for each VoIP provider to install gateway servers in every country. Therefore VoIP providers share their gateway servers with other international VoIP providers. Each local VoIP provider enters into an agreement with their local PSTN phone company. The local PSTN then creates a circuit between the PSTN users and the local VoIP gateway server for the duration of each VoIP call. The VoIP gateway server manages the packet switched side of the connection and the conversion of data between the Internet and the local PSTN.

Advantages of VoIP compared to traditional PSTN includes:

- Low cost long distance calls.
- No added cabling is required to add extra VoIP lines.
- Additional digital services, such as voice mail, conference calls and video calls are much simpler to add as the data is digital.
- VoIP calls can originate from any location with an Internet connection. For example, a user with an Australian VoIP account can use their account from any country just like they do at home.

Disadvantages of VoIP compared to traditional PSTN includes:

- IP and the Internet form a packet switched network, which was not designed for continuous delivery of real time data. If congestion occurs then some packets will be delayed or lost causing poor quality audio. The PSTN maintains a complete circuit for the duration of each call, hence such problems are rarely encountered.
- Emergency VoIP calls cannot be made when there is a power failure. PSTN lines are powered by the local telephone line and hence they continue to operate even when the power to the home or business is cut off.
- Broadband Internet connections are unreliable in terms of Quality of Service (QoS), compared to the PSTN. Most countries have laws that require PSTN lines to be available and that specify how quickly faults must be repaired. Currently no such laws exist for Internet connections.

GROUP TASK Research
VoIP providers also offer VoIP lines to users who do not have an Internet connection. Research to determine why some of these systems allow VoIP calls to be made but not received.

4. ELECTRONIC MAIL

In this section we describe the characteristics and organisation of email messages. This includes the components or fields within an email message as well as how the message data and any attachments are encoded. We also identify and briefly discuss the application/presentation layer protocols used to transmit and receive email messages across the Internet – all email is ultimately transmitted as ASCII text.

During transmission all email messages are composed of two broad components, an envelope and a contents component. The envelope contains the information required to transfer the message to its destination – much like a paper envelope. The envelope data is examined and used by SMTP (Simple Mail Transfer Protocol) servers to relay email messages to other SMTP servers and finally to their destination. The contents component contains various headers together with the actual message. SMTP examines and adds to these headers, however it does not alter the actual message.

Email Contents Component

The contents component contains the actual message data together with various header fields used to specify the sender, receiver, date/time, subject and also the relationship of the message to other related messages. RFC2822 "Internet Message Format" is the current standard that specifies how the content of all email messages are organised. From a user's perspective creating an email involves specifying header fields for the recipients (receivers) and a subject – as well as entering the body of the message. The email client application adds the sender's address, date/time and various other headers. Examples of the more common header fields are shown in *Fig 3.36*. This screen includes four header fields, namely To:, Cc:, Bcc: and Subject:. The To:, Cc: and Bcc: fields are known as destination fields as they are used to specify the recipients of the email. Each of these fields can contain multiple email addresses separated by commas. Note that in *Fig 3.36* MS-Outlook has converted the commas to semi-colons – RFC2822 specifies commas as the separators, presumably MS-Outlook would use commas when the message is actually sent.

Fig 3.36
Email created in the email client MS-Outlook.

The content of all email messages are composed of a sequence of header fields followed by lines of text that form the body of the message. All data being represented as a sequence of ASCII characters. Each header field is composed of a field name followed by a colon ':', the field data and finally a carriage return line feed combination (often referred to as 'CRLF' meaning the ASCII character 13 followed by the ASCII character 10). For example in *Fig 3.36* the To: field is actually sent as `To: fred@thisdomain.com`, the field name being `To` and the field data being `fred@thisdomain.com`.

RFC2822 specifies all the possible header fields. They are broadly grouped into seven categories as destination address fields, originator fields, identification fields, informational fields, resent fields, trace fields and optional fields. We shall consider the first four in some detail and then briefly describe the purpose of the final three categories. Finally we describe MIME – the standard for coding non-text email data.

- **Destination Address Fields**

Destination address header fields include To:, Cc: and Bcc:. The To: field contains the addresses of the primary recipients of the message. These are the people who the message is directly written to. Cc is short for carbon copy; these recipients receive a copy however the message is not directed at them. The blind carbon copy (Bcc:) header field is for recipients who also receive the message but their addresses are not to be revealed to any other recipients. In *Fig 3.36* the message is sent to a total of five recipients. However when the message is sent to the To; and Cc: recipients the Bcc: header field is completely removed.

There are two possibilities that arise when the Bcc: field contains a list of recipient email addresses. Remember the email client must ensure that the individual email addresses of Bcc: recipients are not sent to any other recipients. One solution is to

alter the Bcc: header prior to sending each message so it contains just the individual recipient's address. This solution requires the message to be sent multiple times – once for each of the Bcc: recipients and one time for all the To: and Cc: recipients. Other email clients remove the Bcc: field completely for all recipients – in this case the message is sent just once to all recipients, including the Bcc: recipients. Note it is the envelope that actually determines who is sent a copy of the message – the header fields within the contents are used to determine who these recipients should be. At first the second option appears to be the most satisfactory, however it has security implications. When a Bcc: recipient receives such an email their email address is not shown at all (as it was removed by the sender). As a consequence they may not realise the message was sent confidentially and they may unknowingly reply to one or more of the To: or Cc: recipients. These reply recipients will then be aware that the Bcc: recipient had received the original message.

- **Originator Fields**

Originator fields include Date:, From:, Sender: and Reply-To:. All email messages must contain at least a Date: and From: originator field – the other two fields are used as required.

The Date: field must always be included and is used to specify the date and time that the user indicated that the message was complete and ready to send. Commonly this is the time that the user pressed the send or submit button within the email client application. In many cases the message is not actually sent by SMTP until some later time, for example the user may not currently be connected to the Internet.

It is possible for a message to be sent from more than one person. When this is the case the From: field contains multiple email addresses and the Sender: field is used to specify the single email address that actually sent the message. For example senior management may formulate an email message that is actually sent by a secretary. In this case the From: field contains each of the manager's email addresses whilst the Sender: field would contain the secretary's email address.

The Reply-To: field is optionally used to specify one or more email addresses where replies should be sent. If no Reply-To: field exists then the address or addresses in the From: field are used for replies.

- **Identification Fields**

Identification field headers are used to identify individual messages and to allow email applications to maintain links between a thread of messages. They are designed for machines to read rather than humans. There are three possible identification fields - Message-ID:, In-Reply-To: and References:. Each of these fields contains unique identifiers for individual email messages. Message-ID: should exist within all messages, whilst the other two fields should be included within replies.

The unique identifier used as the field data for the Message-ID: field must be globally unique. That is, no two messages travelling over the Internet can ever have the same Message-ID:. In most cases this uniqueness is achieved by using the domain name (or IP address) on the right hand side of an @ symbol with a unique code for that domain on the left hand side. Some systems use the date and time or the user's mail box in combination with some other unique code on the left hand side.

When a user replies to a message an In-Reply-To: field is created that contains the original message's Message-ID. Furthermore the original message's Message-ID: is also appended to the References: field. This means messages that form part of a conversation include a References: header field that lists all the Message-IDs of the previous related messages. Email applications use this information to display the thread of all related messages.

- **Informational Fields**

Informational fields include the familiar Subject: header together with Comment: and Keywords: header fields. All three of these header fields are for human readers and are optional, however it is desirable to include a Subject: field in all messages.

The Subject: field is used to briefly identify the topic of the message, however it may contain any unstructured text. When replying to messages the string "Re:" is appended to the start of the existing subject field data. The Comment: field is designed for additional comments about the message. The Keywords: field contains a comma separated list of important words or phrases that maybe of relevance to the receiver.

- **Resent, Trace and Optional Fields**

Resent header fields are added to the start of a message each time that an existing message is resubmitted by a user for transmission. The resent fields include Resent-From:, Resent-To:, Resent-Message-ID: and all other corresponding originator and destination fields. The resent headers are for information only – the data in the original message's originator and destination fields are used by email client applications when replies are created.

Trace fields are added by the various SMTP servers who deliver messages across the Internet. They describe the path the message has taken from sender to receiver. These trace header fields are added to the start of each message by each SMTP server. The purpose of such trace headers is to enable technical staff to determine the path taken by each message should delivery problems occur. Most email clients and the majority of SMTP servers provide a command so that such headers can be viewed. For example in current versions of MS-Outlook the Internet Headers for a message can be viewed via the View-Options menu item.

Optional header fields are added to provide additional functionality such as virus checking and for specifying MIME (Multipurpose Internet Mail Extensions) headers. MIME headers are used to specify the details of non-text formatted messages and attachments. Often such header names commence with the string "X-", although this is not strictly necessary.

Consider the following

RFC stands for "Request For Comment", RFCs are initially working documents produced by members of the "Internet Society". The "Internet Society" is a global non-profit organisation that produces and maintains open standards for most of the protocols used over the Internet. Once an RFC has been widely circulated and edited it becomes a standard.

RFC2821 specifies SMTP details (the envelope) and RFC2822 specifies the content of emails. A further series of standards (RFC2046-2049) specify how attachments should be encoded using MIME (Multipurpose Internet Mail Extensions). MIME encoded attachments form part of the content of an email message.

GROUP TASK Research
Using the Internet, or otherwise, identify different Internet standards that are specified using RFCs. Explain why the RFC system is well suited to the creation of Internet standards.

- **MIME (Multipurpose Internet Mail Extensions)**

MIME is the protocol used to code non-textual data and attachments into ASCII so that it can be transmitted within email messages. MIME is used to code HTML email messages, image files, video files and any other type of file that is attached and transmitted by email. Furthermore MIME allows for the transmission of many foreign language characters that cannot be represented using the 127 7-bit ASCII characters.

In all cases the entire message data, including attachments, is included within the content component of the email. The SMTP servers that deliver the email treat the entire message as simple ASCII text (a sequence of 7-bit binary ASCII codes). The receiving email client reads the MIME headers and formats the message accordingly. If an attachment is detected then the original file is recreated. For example the typical headers shown in *Fig 3.37* specify that the body of the message is to be interpreted as HTML and that it is encoded as 7-bit ASCII.

```
Message-ID: <38944.1161439.JavaMail.webadm@nus090pc>
Date: Tue, 24 Oct 2006 16:09:11 +1000 (EST)
To: sam.davis@pedc.com.au
Subject: Telstra Bill - Arrival Notification
Mime-Version: 1.0
Content-Type: text/html
Content-Transfer-Encoding: 7bit
```

Fig 3.37
Example mail headers including MIME headers.

Let us briefly describe how MIME encodes binary data so that it is represented as sequences of 7-bit ASCII codes. The primary MIME technique for encoding binary data into character data is called base64. In this system there are just 65 possible characters that correspond to all the bit patterns possible with 6 binary digits, plus an extra character '=' that is used as padding. The encoding system used is reproduced in *Fig 3.38*. For example say a single 24-bit pixel within an image is represented in binary as 11100101 01110101 01010110 – each byte represents the intensity of red, green and blue respectively. To encode this pixel using Base64 we first split it into

Binary	*Dec*	*Char*
000000	0	A
000001	1	B
000010	2	C
000011	3	D
000100	4	E
000101	5	F
000110	6	G
000111	7	H
001000	8	I
001001	9	J
001010	10	K
001011	11	L
001100	12	M
001101	13	N
001110	14	O
001111	15	P
010000	16	Q

Binary	*Dec*	*Char*
010001	17	R
010010	18	S
010011	19	T
010100	20	U
010101	21	V
010110	22	W
010111	23	X
011000	24	Y
011001	25	Z
011010	26	a
011011	27	b
011100	28	c
011101	29	d
011110	30	e
011111	31	f
100000	32	g
100001	33	h

Binary	*Dec*	*Char*
100010	34	i
100011	35	j
100100	36	k
100101	37	l
100110	38	m
100111	39	n
101000	40	o
101001	41	p
101010	42	q
101011	43	r
101100	44	s
101101	45	t
101110	46	u
101111	47	v
110000	48	w
110001	49	x
110010	50	y

Binary	*Dec*	*Char*
110011	51	z
110100	52	0
110101	53	1
110110	54	2
110111	55	3
111000	56	4
111001	57	5
111010	58	6
111011	59	7
111100	60	8
111101	61	9
111110	62	+
111111	63	/

(pad) =

Fig 3.38
MIME base64 encoding table.

four 6-bit sequences – 111001 010111 010101 010110. We then use our table in *Fig 3.38* to encode each 6-bit sequence as the corresponding character; hence our sound sample is sent within an email as 5XVW. This encoding system works fine when the total number of bits is an exact multiple of 6, in fact the MIME standard insists that the total number of bits be made to be an exact multiple of 24. When this is not the case the pad character '=' is used. For example, to encode the 16-bit pattern 00110001 01111001 we split it into 6-bit sections resulting in 001100 010111 1001. Note that we have two lots of 6-bits and one with just 4-bits. The 4-bits are extended to 6 by

simply adding two more zeros. We now have 001100 010111 100100 which encodes to MXk, however we have just 18-bits not the required multiple of 24-bits, hence we add the pad character, so our data is sent in an email as MXk=.

Clearly most files sent as attachments are significantly longer than our above examples. When the file reaches its destination the reverse process takes place to decode the data. Base64 deliberately uses only characters that are available universally – there are no strange punctuation or non-printable characters. This means the text can be transformed and represented using many different coding systems during its transmission without the risk of corruption. The receiving machine needs only to know the details in *Fig 3.38* to successfully decode the data – the actual characters received can be represented using any character coding system known to the receiver.

GROUP TASK Discussion

Why do you think groups of 6 bits have been chosen to represent single characters in MIME? Why not use 7-bits? Discuss.

Transmitting and Receiving Email Messages

Email uses two different Application Level protocols; SMTP and either POP or IMAP. Email client applications, such as Microsoft Outlook, must be able to communicate using these protocols. SMTP (Simple Mail Transfer Protocol) is used to send email messages from an email SMTP client application to an SMTP server. Emails are received by an email client application from a POP (Post Office Protocol) server or IMAP (Internet Message Access Protocol) server. *Fig 3.39* shows these server settings for a particular email account within Microsoft Outlook.

Sending an email using the account in *Fig 3.39* involves the email SMTP client, in this case Microsoft Outlook, establishing an SMTP connection to the SMTP server called smtp.mydomain.com.au. The email is then transferred to this server. If the user wishes to download their email then Microsoft Outlook establishes a POP connection with pop.mydomain.com.au, logs into the server using the account name and password, and finally receives all messages stored in the mailbox for that account. Note that the account name is the first part of the user's email address. If the address is sam.davis@mydomain.com., then sam.davis is the account name. It is also the mailbox name on the POP server.

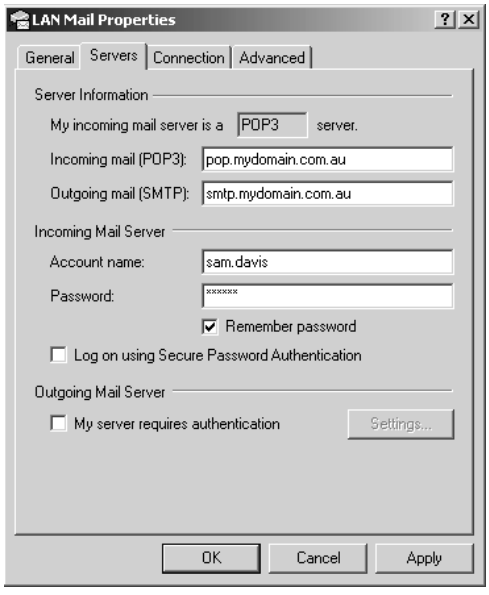


Fig 3.39
Emails are received from a POP server and transmitted to an SMTP server.

So how does email arrive into the mailbox on the POP, or IMAP, server of the recipient? The senders SMTP server establishes an SMTP connection with the recipients SMTP server. To do this it first needs to determine the IP address of the recipients SMTP server. It does this by performing a DNS lookup. DNS stands for domain name server, these are servers that map domain names to IP addresses. For example, the email address fred@nerk.com.au includes the username fred and the domain name nerk.com.au.

A DNS lookup determines the IP address of the email server that stores all mail for the domain nerk.com.au. The email message is sent over the Internet to the machine with this IP address. During this process the sending SMTP server behaves as an SMTP client to the remote receiving SMTP server. Once the message has been sent to the recipient's remote SMTP server it is passed to the corresponding POP, or IMAP server. This server places the message into the mailbox of the recipient ready for collection.

Fig 3.40 shows an email message being sent. The lines commencing with numbers have been received from the remote SMTP server; the sender has entered all other bolded lines. This client-server interaction produces the envelope component used by SMTP to deliver the message. The content component of the message commences after the 'data' command and ends when a full stop (period) is entered on a line by itself. Normally the email SMTP client application automatically generates the commands in *Fig 3.40* based on the header fields within the content of the email message.

```
220 omta03sl.mx.bigpond.com ESMTP server ready Tue, 7 Nov 2006 01:19:08 +0000
ehlo
250-omta03sl.mx.bigpond.com
250-XREMOTEQUEUE
250-ETRN
250-ETRN
250-AUTH LOGIN PLAIN
250-PIPELINING
250-DSN
250-8BITMIME
250 SIZE 15728640
mail from:<sam.davis@pedc.com.au>
250 Ok
rcpt to:<info@pedc.com.au>
250 Ok
rcpt to:<orders@pedc.com.au>
250 Ok
data
354 Enter mail, end with "." on a line by itself
from: sam.davis@pedc.com.au
to: info@pedc.com.au
cc: orders@pedc.com.au
subject: SMTP test message

We'll get this message later with POP.
.
250 Ok
quit
```

Fig 3.40
Sample SMTP client-server session.

GROUP TASK Practical Activity
The SMTP session in *Fig 3.40* was performed on a Windows XP machine using Telnet. To connect to the bigpond SMTP server the command telnet mail.bigpond.com 25 was entered at the run command on the Start menu. Send an email to yourself using Telnet and *Fig 3.37* as a guide.

GROUP TASK Research
SMTP servers accept connections from SMTP clients on TCP/IP port 25, POP servers use port 110. Research what TCP/IP ports are and create a table of commonly used TCP/IP ports together with their purpose.

```
+OK POP3 server ready.
USER sdav8298@bigpond.net.au
+OK please send PASS command
PASS af7rhd3e
+OK sdav8298@bigpond.net.au is welcome here
LIST
+OK 5 messages
1 1912
2 9506
3 25410
4 32896
5 4860
.
RETR 1
+OK 1912 octets
X-McAfeeVS-TimeoutProtection: 0
Return-Path: <sam.davis@pedc.com.au>
Received: from mail62.messagelabs.com ([203.166.119.147])
          by imta05sl.mx.bigpond.com with SMTP
          id
<20061107013310.IAG14880.imta05sl.mx.bigpond.com@mail62.messagelabs.com>
          for <sdav8298@bigpond.net.au>; T
X-VirusChecked: Checked
X-Env-Sender: sam.davis@pedc.com.au
X-Msg-Ref: server-5.tower-62.messagelabs.com!1162863190!6529975!1
X-StarScan-Version: 5.5.10.7; banners=.,-,-
X-Originating-IP: [220.233.16.107]
Received: (qmail 29294 invoked from network); 7 Nov 2006 01:33:10 -0000
Received: from 107.16.233.220.exetel.com.au (HELO envy.hi-speed.com.au)
(220.233.16.107)
  by server-5.tower-62.messagelabs.com with SMTP; 7 Nov 2006 01:33:10 -0000
Received: from pride.hi-speed.com.au (pride.hi-speed.com.au [203.57.144.25
        by envy.hi-speed.com.au (8.11.2/8.11.2) with ESMTP id kA71X3C04121
        for <sdav8298@bigpond.net.au>; Tue, 7 Nov 2006 12:33:03 +1100
Received: from omta02ps.mx.bigpond.com (omta02ps.mx.bigpond.com
[144.140.83.154])
        by pride.hi-speed.com.au (8.9.3/8.9.3) with ESMTP id MAA30755
;
        Tue, 7 Nov 2006 12:32:55 +1100
Received: from [60.229.156.120] by omta02ps.mx.bigpond.com with ESMTP
          id
<20061107013225.PDZP24597.omta02ps.mx.bigpond.com@[60.229.156.120]>;
          Tue, 7 Nov 2006 01:32:25 +0000
from: sam.davis@pedc.com.au
to: info@pedc.com.au
cc: orders@pedc.com.au
subject: Test Message
Message-Id:
<20061107013225.PDZP24597.omta02ps.mx.bigpond.com@[60.229.156.120]>
Date: Tue, 7 Nov 2006 01:32:25 +0000

This message sent using smtp and will be retrieved using pop.


_____________________________________________________________________
This email has been scanned by the MessageLabs Email Security System.
For more information please visit http://www.messagelabs.com/email
_____________________________________________________________________
.
DELE 1
+OK
QUIT
+OK sdav8298@bigpond.net.au POP3 server signing off.
```

Fig 3.41
Sample POP client-server session with client commands in bold.

A sample POP session is reproduced above in *Fig 3.41*. This client-server session was initiated in Windows XP by entering the command telnet mail.bigpond.com 110 in the run dialog on the Start menu. To retrieve messages from a POP server requires the user to verify their identity using their user name and password – that is not my real password in *Fig 3.41*! The username is then used to identify the mailbox. Once this has been done a list of messages including their length can be returned using the LIST command. To retrieve a message the RETR command is used and to delete messages from the POP server the DELE command is used.

Notice the extensive headers added to the message in *Fig 3.41* compared to the original message sent in *Fig 3.40*. Some of these headers have been added by the virus checker, whilst others have been added by each of the SMTP servers. Email to pedc.com.au addresses goes to the pedc.com.au mail server that is hosted by hi-speed.com.au. The hi-speed mail server redirects all pedc.com.au mail to the sdav8298@bigpond.net.au address. This means Parramatta Education Centre needs to POP just one bigpond mailbox to retrieve all its mail.

GROUP TASK Discussion
Identify the path taken by the email sent in *Fig 3.40* and retrieved in *Fig 3.41*. Which server do you think added the virus checking headers?

GROUP TASK Practical Activity
Use Telnet, or some similar program, to POP your own mailbox on your own mail server. Retrieve (RETR) and examine an email encoded using MIME and briefly comment on its MIME header fields.

SMTP, POP, IMAP and DNS are protocols operating at the Application Level. SMTP, POP and IMAP are all part of software applications running on both email clients and email servers. It is possible, and highly likely, that a single machine is an SMTP, POP and IMAP server. In fact many email server applications include all three of these protocols within a single application. DNS servers are usually separate entities to email servers, they provide DNS lookup services to many other Internet applications, not just to email servers.

Consider the following flowchart:

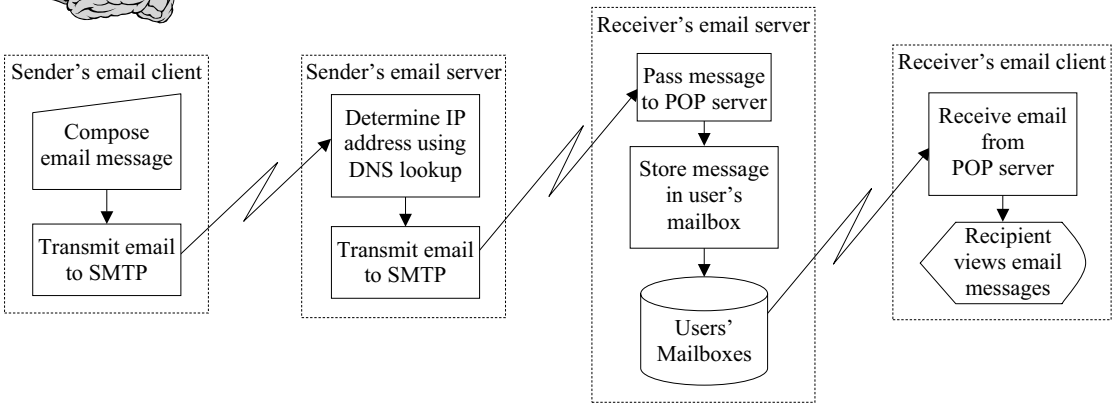


Fig 3.42
Flowchart describing the sending and receiving of email messages.

GROUP TASK Discussion
Suggest modifications to the above flowchart so it more accurately reflects the transmission of the email described in *Fig 3.41*.

SET 3D

1. Most phone lines connecting homes to the local exchange are made of:
   (A) copper.
   (B) aluminium.
   (C) optical fibre.
   (D) steel.

2. The hardware to connect many PSTN telephone lines to a computer is known as a:
   (A) voice modem.
   (B) telephony board.
   (C) ISDN line.
   (D) VoIP broadband modem.

3. Email messages are sent across the Internet using which Application Level protocol?
   (A) SMTP
   (B) POP
   (C) IMAP
   (D) IP

4. Which of the following best describes menus within voicemail systems?
   (A) A linear sequence of OGMs.
   (B) A linear sequence of screens.
   (C) A hierarchical system of screens.
   (D) A hierarchical system of OGMs.

5. The path an email message takes during its journey from sender to receiver can be determined by examining:
   (A) trace fields within the content of the message.
   (B) trace fields within the envelope of the message.
   (C) identification fields within the content of the message.
   (D) identification fields within the envelope of the message.

6. The quickest way to speak to an operator when using an IVR system is to press which key?
   (A) # key
   (B) * key
   (C) 0 key
   (D) 9 key

7. During a telephone call made from a standard PSTN home telephone, which of the following is TRUE?
   (A) Audio is digitised by the home phone.
   (B) Audio is digitised at the exchange.
   (C) The entire connection is digital.
   (D) The entire connection is analog.

8. Why are long distance calls cheaper when using VoIP?
   (A) The PSTN is free.
   (B) The Internet is free.
   (C) Broadband is cheaper than a PSTN line.
   (D) Call quality is poorer using VoIP.

9. An application that allows a computer to be used as a VoIP phone is called a:
   (A) Speech recognition application.
   (B) VoIP gateway
   (C) TTS application
   (D) Soft phone

10. Using MIME base64 encoding, the data 11110000 11110000 would be sent as which series of characters?
    (A) 8PD
    (B) 8PA=
    (C) 4PA=
    (D) 8HA

11. Explain what each of the following acronyms stand for, and describe their purpose.
    (a) OGM (c) VoIP (e) POP
    (b) RTP (d) SMTP (f) IMAP

12. (a) Contrast telephone calls made using a standard PSTN telephone line with calls made using VoIP.
    (b) Prepaid phone cards are used to make cheap VoIP calls from normal phones. Research and explain how Prepaid phone cards work.

13. Compare and contrast storyboards used during the design of software user interfaces with those used during the design of phone information systems.

14. Outline the purpose of each of the following fields within email messages.
    (a) Destination address fields.
    (b) Originator fields.
    (c) Identification fields.
    (d) Informational fields.
    (e) Resent, trace and optional fields.

15. With regard to email, explain each of the following:
    (a) How non-text data and attachments are encoded within messages.
    (b) How email messages are transmitted and received.

ELECTRONIC COMMERCE

Financial transactions that occur over an electronic network are all examples of electronic commerce. We use electronic commerce systems to withdraw cash from ATMs (automatic teller machines), pay for store purchases using EFTPOS (electronic funds transfer at point of sale), buy and sell goods over the Internet and to perform electronic banking transactions over the Internet. The majority of Australians are participants in one or more electronic commerce transactions ever day. Indeed Australia is one country that has enthusiastically embraced all forms of electronic commerce systems. In this section we examine ATMs, EFTPOS, Internet banking and trading over the Internet.

1. AUTOMATIC TELLER MACHINE (ATM)

Today most Australians are familiar with the operation of automatic teller machines (ATMs), at least from the user's perspective. ATMs are present outside banks, within shopping malls, in service stations and numerous other locations. There are a number of different ATM networks in Australia – most are operated by or on behalf of banks. Today all these networks are connected, both within Australia and also to most overseas networks. As a consequence it is possible to make a withdrawal from an Australian bank account from almost any ATM in the world. Similarly tourists, when in Australia can withdraw cash from their home accounts.

Each ATM includes at least two collection (input) devices and at least four display (output) devices (see *Fig 3.43*). Collection devices include a magnetic stripe reader that collects magnetic information from the back of the customer's card. This data is used to identify the customer and their financial institution. A keypad is used to enter the customer's PIN (Personal Identification Number) and to enter other numeric data. Most ATMs include buttons beside the screen that initiate the functions displayed on the screen. Some versions include a touch screen and hence buttons beside the screen are not required.

Display devices include the screen – which is often a CRT although LCD screens are becoming popular. A receipt printer produces a hardcopy record of any transactions performed. A speaker is embedded within the ATM to provide basic audio feedback as keys are pressed. The cash dispenser is a specialised display device that includes many security functions to ensure it delivers the exact amount of cash.

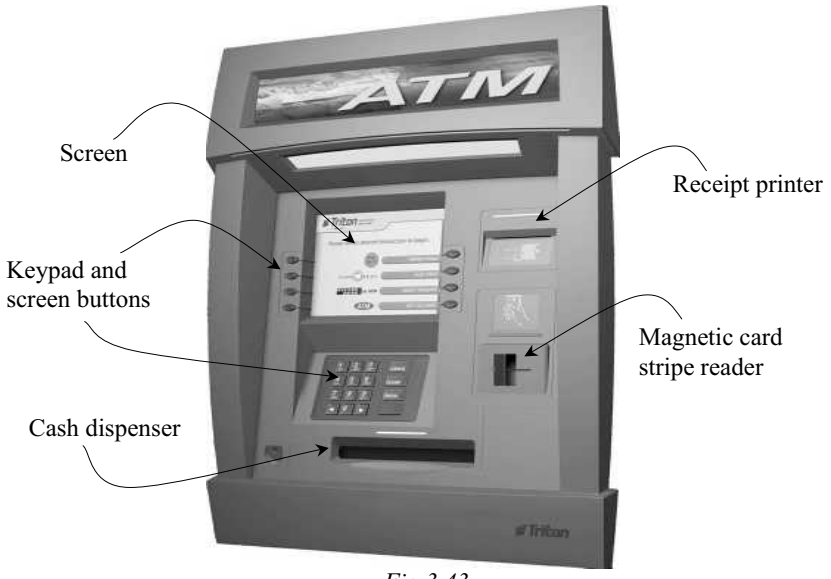


Fig 3.43
Automatic teller machine (ATM) collection and display devices.

Cash dispensers include a safe that contains drawers for each denomination of bank note and another drawer for reject bills. The cash dispenser includes two sensors and various mechanical parts for moving bank notes. One sensor counts the number of bills and the other measures the thickness of each bill. Any bills that do not meet specifications are diverted to the reject drawer at the top of the safe. *Fig 3.44* shows an LG CDM3200 cash dispenser used within many permanent bank ATMs.

Fig 3.44
LG CDM3200 Cash Dispenser

Most modern ATMs are essentially personal computers with specialised peripheral devices housed in secure cabinets. They include a standard PC motherboard and processor running common operating systems such as Windows and Linux.

To approve transactions all ATMs are connected to a network that ultimately must be connected to the customer's bank. ATMs installed outside banks usually include a permanent Ethernet connection to the banks network, those within shopping centres connect using a dedicated phone line, whilst smaller ATMs within service stations include a dial-up modem that only connects when required. The quantity of data transferred during a typical ATM transaction is small.

If the ATM is operated by the customer's bank then the approval process is simplified as the transaction can be completed in real time. For example when an ANZ customer makes a withdrawal from an ANZ ATM the funds are directly debited from the ANZ customer's account without passing through any other accounts. However the process becomes more complex when a customer performs transactions using an ATM operated by some other financial institution. The funds move from the customer's account into the cash account of the financial institution operating the ATM. This transfer must be approved before any cash is dispensed. The process becomes even more complex for privately operated ATMs, such as those found in many service stations and shops. Such transactions are similar to EFTPOS transactions; we shall consider an example during our EFTPOS discussion that follows.

Consider the following:

There have been many successful and unsuccessful attempts to steal money via ATMs. Some examples include:

1. Physically stealing the ATM using "ram raid" style robberies.
2. Observing users entering their PIN and later stealing their card.
3. Installing an additional magnetic stripe reader together with a hidden wireless video camera to record card numbers and PINs.
4. Internal crimes where say a $20 tray is loaded with $50 bills.
5. Intercepting new cards and PINs from customer's mail boxes.

GROUP TASK Research
Research, using the Internet or otherwise, examples of each of the above crimes. Identify and briefly describe security measures in place that attempt to prevent such crimes occurring.

2. ELECTRONIC FUNDS TRANSFER AT POINT OF SALE (EFTPOS)

EFTPOS terminals are now standard equipment at the register of most retail stores. Using the EFTPOS system buyers can pay for goods electronically using either a credit or debit card. In other countries the EFTPOS system is known by various other names. For example in the USA it is known simply as POS, in the UK the term EFTPOS is not used, rather users refer to EFTPOS cards as debit cards. Currently New Zealanders are by far the highest users of EFTPOS. In New Zealand customers are not charged for EFTPOS transactions – as a result EFTPOS is routinely used for purchases of just 10 or 20 cents.

Fig 3.45
OMNI 3200se EFTPOS terminal with built-in thermal printer.

A typical EFTPOS terminal, such as the OMNI 3200se shown in *Fig 3.45*, includes a keypad and magnetic stripe reader for collecting and a monochrome LCD screen and a small thermal printer as display devices. Most EFTPOS terminals transmit and receive transaction data over the PSTN via a built-in dialup modem. Wireless versions that communicate over mobile phone networks and Ethernet versions that communicate over the Internet are also available. In all cases the data is secured during transmission using a public two key encryption system.

In larger department stores it is common for the processes performed by EFTPOS terminals to be integrated with the stores internal register and point of sale systems. Within smaller stores EFTPOS terminals operate independent of the stores register.

GROUP TASK Discussion
Review the operation of public (or two key) encryption systems. Refer to chapter 2 - page 172-173.

GROUP TASK Practical Activity
Observe EFTPOS terminals at various stores and identify their components and in particular the type of cables connecting the terminals to the EFTPOS network and other POS hardware devices in the store.

Consider the following

Consider a typical EFTPOS purchase transaction using an EFTPOS terminal within a store. These processes are similar to making a withdrawal from a privately owned ATM within a store. The store owner is called the merchant hence eventually the funds must move from the customer's account into the merchant's account. If the device is a privately operated ATM then in most cases the merchant is responsible for filling the ATM with cash from their own funds. In Australia it is common for both customers and merchants to be charged for transactions, however merchant charges generally decrease as usage increases. Some private ATM companies will actually pay the merchant a small commission when usage exceeds some agreed limit.

In our example the host server is operated by the private company who supplied the EFTPOS machine to the store. The processes occurring during a typical EFTPOS transaction are described below and are summarised on the DFD in *Fig 3.46*:

- Customer swipes card through magnetic stripe reader and the card number is read.
- Merchant enters sale amount into EFTPOS terminal's keypad.

- Customer selects account and enters their PIN via the keypad.
- EFTPOS terminal dials host server and connects.
- EFTPOS terminal transmits encrypted card number, account type, PIN and sale amount to host server.
- Host server determines the customer's financial institution based on the card number.
- Host server connects to customer's financial institution and transmits encrypted transaction details including card number, account type, PIN and sale amount.
- Financial institution approves the transaction only if it can verify the customer based on their PIN, the customer has sufficient funds in their account and the customer has not used their daily EFTPOS limit.
- If the transaction is approved the financial institution responds to the host by transmitting a unique transaction ID together with an OK. The financial institution reserves the funds to prevent them being used by other transactions.
- The host processor receives the OK from the financial institution and causes the transfer of funds from the customer's account into the host's cash account. This is the electronic funds transfer (EFT) part of the transaction.
- Host verifies the funds have been transferred to its cash account and records all details of the transaction.
- Host sends an OK to the EFTPOS terminal to confirm the transfer is complete and the EFTPOS terminal responds to the host that it has received the message.
- The host receives the OK from the terminal and commits the transaction. If no OK is received then the entire transaction is reversed.
- The EFTPOS terminal prints a receipt for the customer and for the merchant.
- Each evening the host processor calculates the total amount owing to each merchant. These totals are transferred via an automatic clearing house (ACH) from the host's cash account into each merchant's account. Note that this step is not included on the DFD in *Fig 3.46*.

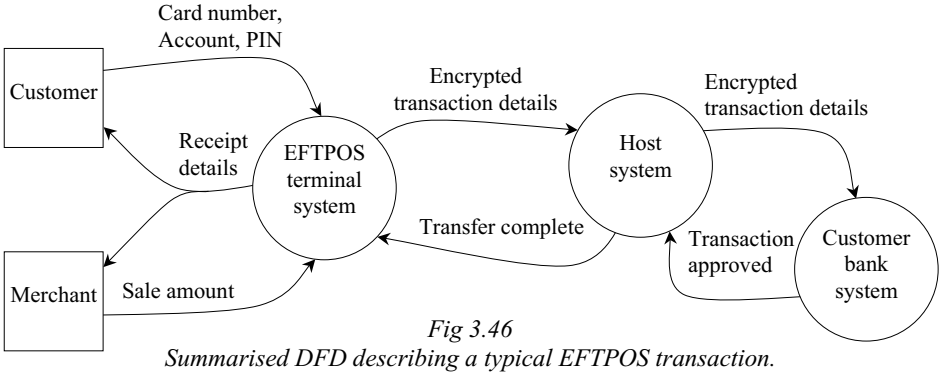


Fig 3.46
Summarised DFD describing a typical EFTPOS transaction.

For ATM transactions a slightly different sequence is involved. In most cases the host system verifies the customer using their PIN prior to the transaction amount and type being entered. This allows ATM customers to complete many transactions without the need to re-enter their PIN. Note that privately operated ATMs do not provide functions for transferring funds between accounts or for performing deposits.

GROUP TASK Activity
Expand the above DFD to include more detail of the processes occurring within the EFTPOS terminal system, host system and customer bank system. Also construct a DFD for the ACH system.

3. INTERNET BANKING

Internet banking allows bank customers to pay bills, transfer money between accounts and perform various other functions from the comfort of their home or office. Most banks and other financial institutions encourage their customers to use Internet banking as it is considerably more cost effective compared to face-to-face or even telephone operator assisted services. Furthermore Internet banking is convenient for customers as they need not travel to a branch and the service is generally available 24 hours a day and 7 days a week.

To access Internet banking the customer must have a computer connected to the Internet, together with a user ID and password from their financial institution. The customer's web browser connects directly to the bank's web server using a URL commencing with https rather than http. The use of https indicates to the web browser that the http protocol is to be used together with SSL (Secure Sockets Layer) or TLS (Transport Layer Security) protocols. SSL and TLS operate within the OSI transport layer just above TCP. Both these Communication Control and Addressing Level protocols use public key encryption to ensure the secure delivery of data in both directions. Most web servers accept https client requests on port 443 rather than the usual port 80 used by http web servers. Once an https session has been secured most web browsers display a small padlock icon in their status bar (see *Fig 3.47*).

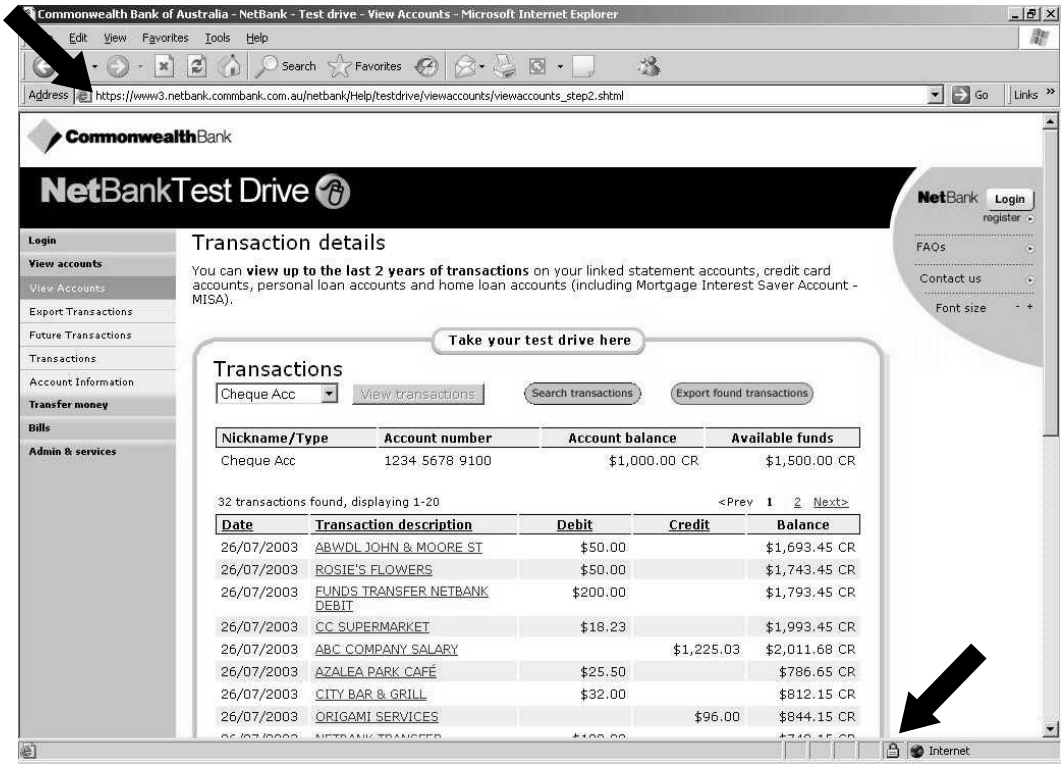


Fig 3.47
Test drive screen of the Commonwealth Bank's Netbank site.

To encourage and train new users most banks include a simulation of their Internet banking functions. *Fig 3.47* is a screen shot from the Commonwealth Bank's Netbank Test Drive. Notice the URL in the address bar commences with https, indicating secure public encryption is being used. Furthermore this URL ends with the file extension .shtml rather than the more usual .htm or .html. The extension .shtml refers to hypertext mark-up language documents with embedded "server-side includes". In this banking example the "server-side includes" cause the bank's web server to add

data specific to the customer prior to transmitting the web page. Clearly this is necessary to customise each page using the customer's account and transaction details. Server-side means that the server executes programming code and the resulting output is sent to the client – in this case the customer's web browser. There are various other server-side systems such as CGI (Common Gateway Interface) and ISAPI (Internet Server Application Programmers Interface). For Internet banking the server-side code causes SQL SELECT statements to execute on the banks database servers. The results returned from the select queries is then combined with the html web page and transmitted securely to the customer's web browser.

GROUP TASK Practical Activity
Work through an Internet banking simulation. Note any security features and identify when the web server is likely to be performing SQL queries prior to transmitting each web page.

Consider the following:

There have been numerous attempts to illegally access Internet banking sites. It is unclear just how many attempts have been successful – banks are reluctant to share such information. Some common examples include:

- Fraudulent emails claiming to be from banks that request user names and passwords. Often such emails are sent randomly to thousands of email addresses in the hope that some unsuspecting users will respond. Such fraud attempts are so common they have been given their own name – "phishing".
- Emails that direct customers to fraudulent web sites that imitate the real site. One such scam opened an SSL page that precisely imitated the real bank's login screen except when the login button was clicked an error message was displayed followed by the real bank's login page. The user name and password were sent to the illegal operators.
- Malicious software that records keystrokes, such as passwords, and sends them to illegal operators. Such software usually installs as part of some other software product and is an example of a Trojan.
- Identity theft where a fraudulent person obtains sufficient information about another so that they can contact the bank, identify themselves as the other person and have the password altered.

GROUP TASK Discussion
Why do you think banks are somewhat reluctant to divulge information in relation to the number of fraudulent Internet banking activities? Discuss.

GROUP TASK Activity
Create a list of recommendations that should be followed by customers to improve the security of Internet banking usernames and passwords.

GROUP TASK Discussion
Many customers unknowingly divulge their passwords. Who is or should be responsible, the customer or the bank? Discuss.

HSC style question:

Read the following article then answer the questions that follow.

Western Australian
16 January 2004

Banks 'killing' the bush

MPs blast branch closures as communities feel the pain

CANBERRA **By Mark Thornton**

BANKS that close branches in rural and remote areas leave "gaping holes" in those communities which lead to their slow deaths, according to a Federal parliamentary report. The report, by the joint committee on corporations and financial services, criticised the closure of rural branches purely on economic grounds, particularly because the action was usually taken without any community consultation. Finance Sector Union figures show banks have closed 2000 branches nationally in the past decade. In WA, the major banks closed 10 branches in 2003.

The committee recommended banks develop comprehensive community consultation procedures before closing any more branches. It said part of the problem had been the Australian Prudential Regulation Authority's inadequate database on the availability of banking services. The committee suggested another government agency take over the work.

"It concerns the committee that there are pockets in the Australian community where competition in the retail banking industry is not strong and where the withdrawal of bank branches has created a void in the provision of banking and financial services," committee chairman Senator Grant Chapman said.

"Time and again we heard that while technology may have ameliorated the difficulties this has created in conducting financial transactions, it has not replaced the gaping hole left in the Community by the departure of the local bank manager, who was not only a trusted financial adviser who knew the local people and local economy, but was also a local community leader."

The Federal parliamentary report included the following recommendations:

- Give a minimum of six months written notice to customers before closing a branch.
- Prepare a community impact statement to help customers understand the reasons behind the closure and help them decide any action.
- Arrange the free transfer of accounts to other institutions of the customers' choice.
- Better education and training programs in the use of new technology so older and indigenous Australians can use Internet and telephone banking services.

The Australian Consumers Association said the recommendations did not go far enough. The Australian Bankers Association said it would consider the report.

(a) Identify and discuss banking services that are difficult to perform, or simply cannot be performed, using Internet and telephone banking.

(b) Closing rural bank branches clearly results in job losses for bank employees. However research shows further job loss occurs within local businesses.

Identify likely reasons for these further job losses.

(c) One of the committee's recommendations was:

Better education and training programs in the use of new technology so older and indigenous Australians can use Internet and telephone banking services.

Identify strategies that could be used to implement this recommendation.

Suggested solution

(a) Impossible to perform cash deposits and withdrawals, also impossible to perform cheque deposits. Any services that cannot easily be described using a rigid procedure are difficult to perform using electronic banking. For example a farmer may default on a loan however they may well be expecting a large cheque at any moment. Such problems are easily explained to a local bank manager who understands the needs and operational realities of small business within his local area. Such understanding is near impossible to replicate electronically.

(b) Likely reasons for further job losses include.

- Local residents now travel to other towns to perform their banking. Therefore fewer customers are in town to spend money within local businesses.
- Banking is performed electronically, hence no need for customers to go to town so local businesses suffer job losses.
- Local people no longer carry cash, so on-the-spot purchases are reduced. This results in lower turnover and consequential job losses.
- A spiralling effect occurs whereby one business closing causes more people to travel to larger centres, which further reduces the clientele for other businesses, and so on.
- Without access to a local bank manager, small business owners are less able to explain their needs in regard to financial problems. As a consequence it is difficult for them to access funds to continue operation.

(c) Possible education and training strategies that could be used include:

- Provision of onsite visits at minimal or no cost when people first apply for Internet or telephone banking services.
- Free classes on the use of the Internet. Perhaps through the local school or TAFE college.
- Creation of a mentoring scheme, whereby current local users are encouraged to provide assistance to elderly or indigenous users.
- Instructional information brochures sent to all elderly or indigenous customers.
- Provide free access to electronic banking through council libraries and community centres. Provide trainers to assist people on a one-to-one basis.
- Free assistance via a 1800 number.

Comments

- Each part of this question would likely be worth 3 marks.
- In part (a) it is necessary to identify banking services that cannot physically be performed over the Internet as well as those that are difficult to perform successfully without face-to-face contact.
- In parts (b) and (c) it is necessary to identify multiple reasons/strategies. It is reasonable to expect that three solid reasons/strategies would need to be identified for full marks.

4. TRADING OVER THE INTERNET

Buying and selling goods over the Internet is booming. Individuals and small business are able to sell to worldwide markets with little initial setup costs. Buyers are able to compare products and prices easily from the comfort of their own home. Online auctions, such as eBay, provide a means for selling and purchasing. Furthermore processing payments for goods is simplified using sites such as PayPal.

Trading over the Internet has resulted in the creation of virtual businesses. These businesses do not require shop fronts and are able to set up operations across the globe without the need to invest in expensive office space. Such businesses are an example of a virtual organisation – other types of virtual organisation exist to complete specific projects, collaborate on new standards or simply to share common interests. For example a database application can be developed using a team of developers who each live in different countries.

Virtual Organisation
An organisation or business whose members are geographically separated. They work together using electronic communication to achieve common goals.

One of the most significant problems facing businesses that sell over the Internet is establishing customer trust and loyalty. Most people feel they are more likely to receive quality service and product support when they purchase from a traditional store. Traditional shopfronts have a permanence about them and furthermore customers are negotiating deals face-to-face. This is not the case when trading over the Internet. In general the only contact is via the website and email messages. Internet only businesses must provide exceptional customer service and support if they are to overcome these issues.

Another significant concern for Internet buyers is security of purchasing transactions. In particular security of account details such as credit card numbers and account numbers. Companies, such as PayPal, resolve this concern by acting as a "middleman" between buyer and seller. The buyer submits their financial details to the middleman who makes the payment to the seller on behalf of the buyer. The seller never receives the customer's credit card or account details. The funds are withdrawn from the buyer's account and deposited into the seller's account by the "middleman".

Consider PayPal:

Currently PayPal is the world's most popular online payment service. PayPal maintains accounts for each of its customers – both buyers and sellers. When making a purchase funds must first be deposited into your PayPal account. These funds are then transferred into the sellers PayPal account. Sellers are then able to transfer the funds from their PayPal account into any bank account throughout the world. All PayPal financial transactions are encrypted using the SSL protocol.

PayPal is currently owned by eBay and hence paying for eBay items using PayPal is the preferred method. PayPal provides their service to all types of online stores and services. Some sellers direct customers to the PayPal site as one payment option whilst others integrate the PayPal system within their site such that all payments are effectively made using PayPal. For sellers the use of PayPal removes the need for them to setup their own secure payment systems and to have them certified according to the legal requirements of their country. Furthermore PayPal can accept payments in almost any currency from people almost anywhere in the world.

Behind the scenes PayPal maintains communication links to banking systems and clearing houses throughout the world. These various systems charge fees to process transactions. PayPal does not charge buyers for a basic account, however they charge sellers a percentage on their sales in much the same way that merchants are charged by banks for credit card sales. PayPal also makes much of their money from interest earned on the money within PayPal accounts.

GROUP TASK Discussion
Identify reasons why buyers and sellers prefer to perform online financial transactions using services such as PayPal rather than more traditional credit card and direct deposit transaction systems.

GROUP TASK Discussion
PayPal is not a bank and therefore the laws and government safeguards with which banks must comply do not apply. Discuss possible implications for PayPal customers.

Consider eBay:

Currently eBay is the most popular online auction and Internet trading system. According to eBay their customers are buying and selling with confidence.

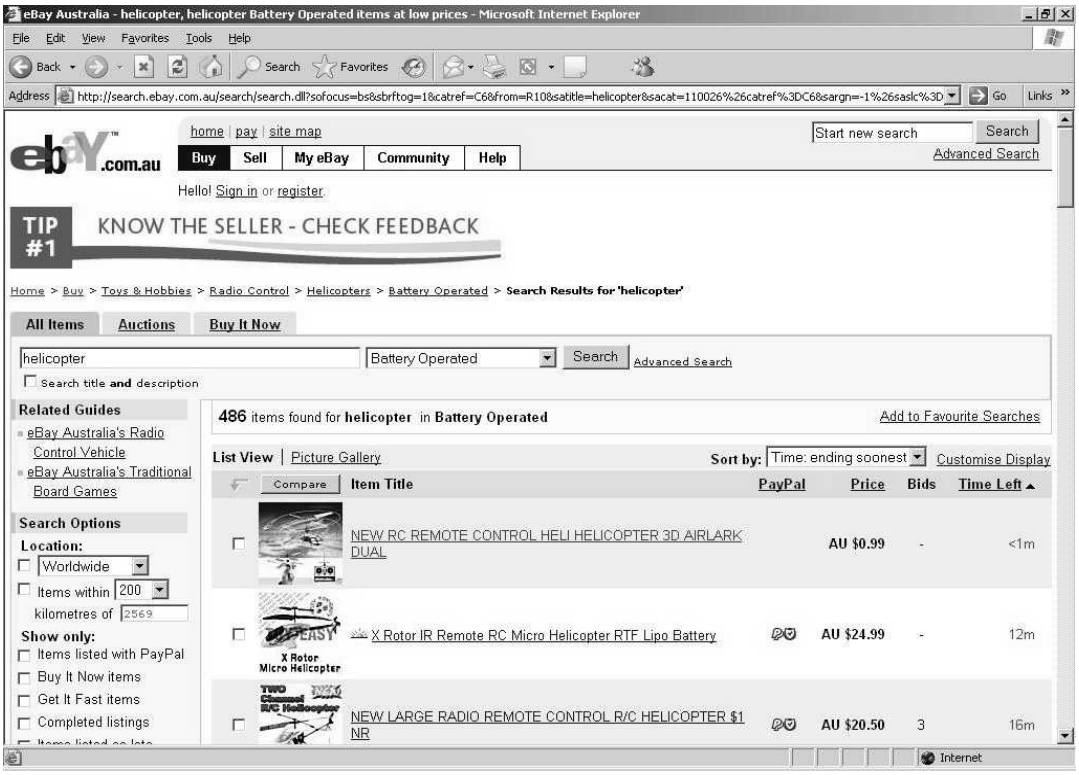


Fig 3.48
eBay's online auction search screen.

GROUP TASK Discussion
Identify and describe features within the eBay system that encourage honest trading between buyers and sellers.

GROUP TASK Discussion
Currently there are millions of people worldwide who earn the majority of their income from eBay sales. Compare and contrast eBay stores with traditional stores.

SET 3E

1. Examples of electronic commerce systems include:
   (A) Fax, telephone, teleconferencing.
   (B) EFTPOS, DBMS, Web servers.
   (C) ATMs, EFTPOS, Internet banking.
   (D) Banks, Building Societies, Credit Unions.

2. Display devices within ATMs include:
   (A) screen, speaker, cash dispenser, receipt printer.
   (B) keypad, touch screen.
   (C) screen, receipt printer, keypad, magnetic stripe reader.
   (D) magnetic stripe reader, barcode scanner, touch screen.

3. Which of the following is TRUE of EFTPOS transactions?
   (A) The customer's PIN is used to identify the customer's account.
   (B) Funds are not immediately credited to the merchant's account.
   (C) Funds are reserved prior to customers entering their PIN.
   (D) Funds leave customer's accounts during the evening following the purchase.

4. The most significant problem for businesses selling over the Internet is:
   (A) establishing customer trust and loyalty.
   (B) verifying customer payments.
   (C) complying with complex taxation laws that apply in different countries.
   (D) maintaining stock in different geographical locations.

5. Examples of "server side" systems include:
   (A) http, https.
   (B) Java and VB applets.
   (C) CGI, ISAPI.
   (D) SSL, TLS.

6. Virtual businesses:
   (A) can trade internationally.
   (B) require shop fronts.
   (C) must rent or buy office space.
   (D) require significant capital to setup.

7. Cash is only dispensed from an ATM after:
   (A) the customer's PIN is verified as correct.
   (B) sufficient funds are available in the customer's account.
   (C) funds are transferred into the financial institution operating the ATM's account.
   (D) All of the above.

8. At the time this text was written the country who used EFTPOS the most was:
   (A) Australia.
   (B) USA
   (C) New Zealand.
   (D) Sweden.

9. Which of the following is TRUE when using SSL or TLS?
   (A) The URL commences with http and public key encryption is used.
   (B) The URL commences with https and public key encryption is used.
   (C) The URL commences with https and private key encryption is used.
   (D) The URL commences with http and private key encryption is used.

10. An organisation where members are geographically separated but work together via electronic communication is known as a(n):
    (A) online business.
    (B) e-commerce site.
    (C) virtual organisation.
    (D) Internet community.

11. Identify and briefly describe the operation of collection and display devices within:
    (a) ATMs
    (b) EFTPOS terminals

12. Explain the processes that occur when making a withdrawal from an ATM.

13. Explain the processes that occur when making an EFTPOS purchase.

14. Research and describe TWO examples where illegal electronic access has been gained to bank accounts.

15. Online auctions sites such as eBay have an enormous following.
    (a) Explain how such sites build trust between buyers and sellers.
    (b) Identify different payment options available on auction sites and assess the security of each option.

NETWORK COMMUNICATION CONCEPTS

In this section we introduce concepts required to understand the design and operation of networks. We shall examine client-server architecture and distinguish between thin and fat clients. We then consider network topologies that describe how network devices are physically and logically connected. Finally we describe different encoding and decoding methods used to represent data as signals suitable for transmission.

CLIENT-SERVER ARCHITECTURE

Client-Server Architecture
Servers provide specific processing services for clients. Clients request a service, and wait for a response while the server processes the request.

As the name client-server suggests, there are two different types of computer present on the network, namely servers and clients. The server provides particular processing resources and services to each client machine. For example, web servers retrieve and transmit web pages, and database servers retrieve and transmit records. The client machines, which are commonly personal computers, also perform their own local processing. For example, web browsers, email clients and database applications. Each server provides processing services to multiple clients.

Client-server processing is a form of distributed processing where different computers are used to perform the specific information processes necessary to achieve the systems purpose. Client-server processing occurs sequentially, this means that for each particular client-server operation just one CPU is ever processing data at a particular time. Many operations may well be occurring simultaneously however each particular operation is processed sequentially.

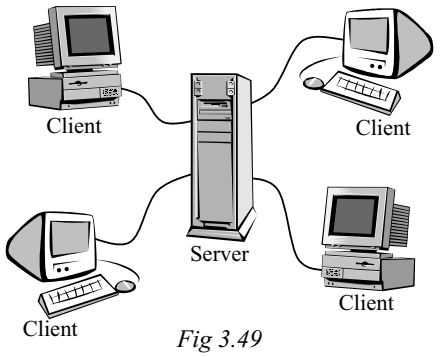


Fig 3.49
Each server provides services to multiple clients.

When a particular operation is being performed either the client is processing or the server is processing, but not both at the same time. Consider *Fig 3.50*, the client machine performs processing and then when it requires the resources of the server it sends a request, the client waits for a response from the server before it continues processing. Between the request being sent and the response being received the server is performing the requested processes.

Notice that the client machines do not need to understand the detail of the server's processes and the server does not need to understand the detail of the processes occurring on the clients. Rather the two machines merely agree on the organisation of requests and responses. Hence a single server can provide processing resources to a variety of different clients running quite different software. For example, a single web server is able to provide resources to client computers of various types running a variety of different web browsers. Similarly a single database server can provide data to a variety of different client applications. As long as the request is legitimate, the server will perform the required processes and generate and transmit a response.

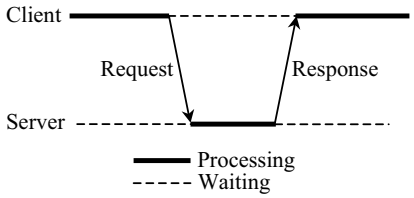


Fig 3.50
Client-Server processing is performed sequentially.

Our discussion so far implies that servers are quite separate computers dedicated solely to server tasks; for large systems with many clients this is often the case,

however it is not a requirement. Consider a small office or even home local area network (LAN). One machine is likely to be connected to the Internet and hence is an Internet server for all other computers on the LAN. Another computer on the LAN is connected to and controls the operation of a shared printer; hence it is a print server. Both these computers are servers, yet they are also clients to each other and even to themselves. In effect a computer can be a server for some tasks and a client for others.

In general client applications provide the user interface, hence they manage all interactions with end-users. This includes collecting and displaying information processes. In many cases the user is unaware of the server's role – indeed many users maybe unaware of the servers very existence. From the user's perspective interactions between client and server are transparent. For example when performing an Internet banking transaction a web browser is the client application that requests data from the banks web server. The banks web server then acts as a client to the banks DBMS server. Users need not be aware of the servers involved and almost certainly are unaware of the specifics of the client-server processes occurring.

Authentication
The process of determining if someone, or something, is who they claim to be.

On larger local area networks (LANs) it is common for all network tasks to be performed by one or more servers using client-server architecture. These servers commonly run a network operating system (NOS) such as versions of Linux, Novell Netware or Window's Server. These network servers control authentication of users to ensure security. Authentication processes aim to determine if users, and other devices, are who they claim to be. Commonly users must log into the network server before they are able to perform any processing. In most cases a logon password is used, however digital certificates and biometric data such as fingerprints are becoming popular methods of authenticating users. NOSs also provide file server, print server and numerous other services to users. We examine NOSs and their capabilities in more detail later in this chapter.

GROUP TASK Discussion
Simple passwords are often compromised. Identify techniques and strategies for maximising the security of passwords.

In our above discussion, the client machine has applications installed that are executed by the CPU within the machine. Such clients are known as "fat clients" or "thick clients". Another strategy that is gaining in popularity is the use of thin clients. A thin client is similar in many ways to the old terminals that once connected to centralised mainframe computers. These terminals only performed basic processing tasks, such as receiving data, displaying it on the screen and also transmitting input back to the mainframe. Thin clients can be implemented in a number of ways. They can be very basic low specification personal computers, often without any secondary storage. These thin clients rely on servers to perform all the real processing. Other thin client implementations are software based. For instance, the RDP (Remote Desktop Protocol) can be used to connect and execute any application running on a remote server. Essentially RDP simply sends the screen display from the remote computer to the thin client. The user at the thin client can therefore log into and operate the remote computer as if they were actually there. This technique is popular with IT staff as it allows them to manage servers from remote locations, such as from home. It is also routinely used to allow employees to access their work network from home or other locations via the Internet. RDP and other thin client protocols also provide a simple technique for making applications available over the Internet.

NETWORK TOPOLOGIES

The topology of a network describes the way in which the devices (nodes) are connected. A node is any device that is connected to the network, including computers, printers, hubs, switches and routers. All nodes must be able to communicate using the suite of protocols defined for the particular network. In general all nodes are able to both receive and transmit using the defined network protocols. Nodes are connected to each other via transmission media, either wired cable or wireless.

Physical Topology
The physical layout of devices on a network and how the cables and wires connect these devices.

The topology of a network describes these connections in terms of their physical layout and also in terms of how data is logically transferred between nodes. The physical connections between devices determine the physical topology. The logical topology describes how nodes communicate with each other rather than how they are physically connected.

Logical Topology
How data is transmitted and received between devices on a network regardless of their physical connections.

There are three basic topologies – bus, star and ring. In addition two other topologies, hybrid and mesh, are common on larger networks. Each of these topologies can describe the physical or the logical topology of a network. Often the logical topology is different to the physical topology. For example a physical star topology has all nodes on the LAN connected by individual cables back to a central node – often a hub or switch. This same network can have a different logical topology, either a logical bus or perhaps a logical ring topology.

Physical Topologies

- **Physical Bus Topology**

All nodes are connected to a single backbone – also known as a trunk or bus. The backbone is a single cable that carries data packets to all nodes. Each node attaches and listens for data present on the backbone via a T-connector or vampire connector. As the two ends of the backbone cable are not joined it is necessary to install terminators at each end. The function of the terminators is to prevent reflection of the data signal back down the cable. On electrical networks, as opposed to fibre optic networks, terminators are resistors that completely stop the flow of electricity by converting it into heat.

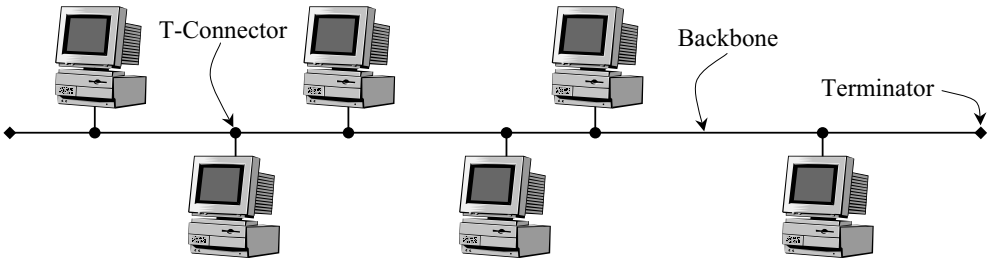


Fig 3.51
Physical bus topologies use a single backbone to which all nodes connect.

In the past physical bus topologies were used for most LANs – in particular Thicknet and Thinnet Ethernet LANs that use coaxial cable as the transmission media. Although these networks require less cable than current star wired topologies they are unable to accommodate the large number of nodes present on many of today's LANs.

Furthermore a single break in the backbone disables the entire network. Today physical bus topologies are used for some high-speed backbones (often using fibre optic cable) and other long distance connections within commercial and government WANs. These high-speed applications have few attached nodes, in many cases just one at each end of the backbone to link two buildings. Where quality of service is critical it is common to install a secondary backbone to provide a redundant connection. If the primary backbone fails for any reasons then the network automatically switches to the secondary backbone.

- **Physical Star Topology**

All nodes connect to a central node via their own dedicated cable. Today the physical star topology is used on almost all LANs, including wireless LANs. In most cases the central node is a switch that includes multiple ports. In the past the central node was likely to have been a hub, multistation access unit (MAU) or even a central computer. We consider the operation of hubs and switches later in this chapter. MAUs are used in token ring networks so that a physical star topology can be used with token ring's logical ring topology. For wireless LANs a WAP (Wireless Access Point) is used as the central node. In terms of physical star topologies the central node is the device that connects all outlying nodes such that they can transmit and receive packets to and from each other node.

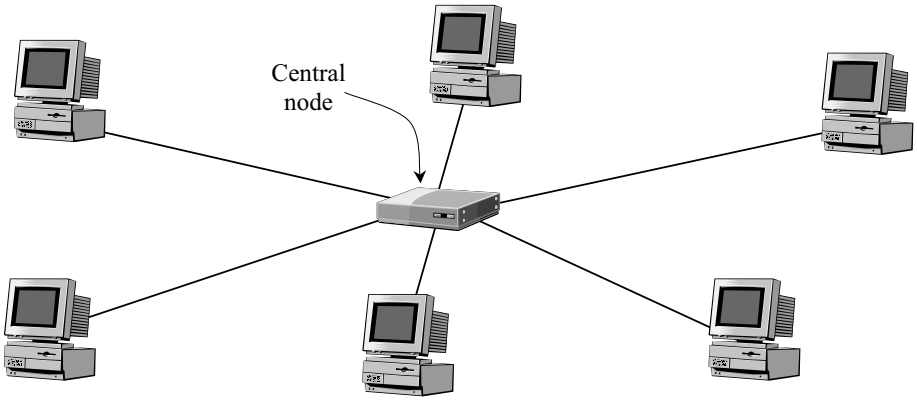


Fig 3.52
In a physical star topology all nodes connect to a central node using their own dedicated cable.

Physical star topologies have a number of advantages over physical bus and ring topologies. This is particularly true for LANs where nodes are physically close – such as within the same room or building. Firstly each node has its own cable and hence can be connected and disconnected without affecting any other nodes. Secondly new nodes can easily be added without first disabling the network. Finally identifying faults is simplified as single nodes can simply be disconnected from the central node in turn until the problem is resolved.

There are however some disadvantages of physical stars. Significantly more cabling is required, however this cable is generally less expensive as it must only support transmission speeds sufficient for a single node. Today UTP (Unprotected Twisted Pair) is the most common transmission media. Also if a fault occurs in the central node then all connected nodes are also disabled.

GROUP TASK Practical Activity
Consider one of your school's computer rooms. Estimate the length of cable required to connect all computers (and other nodes) using a physical bus topology and then using a physical star topology.

- **Physical Ring Topology**

In a physical ring each node connects to exactly two other nodes. As a consequence the cable forms a complete ring. In general data packets circulate the ring in just one direction. This means each node receives data from one node and transmits to the other. If the cable is broken at any point then the entire network is disabled. Therefore removing a node or adding a new node requires the network to be stopped. Furthermore in most implementations each data packet is received and then retransmitted by each node, hence all nodes must be powered at all times if the network is to operate. For these reasons physical ring topologies are seldom used for LANs today.

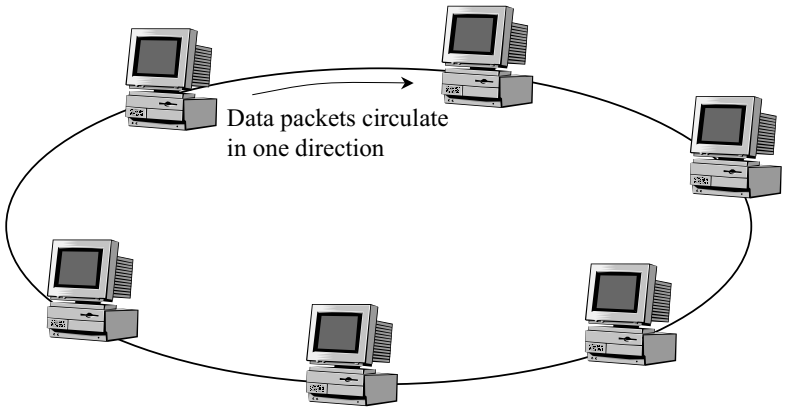


Fig 3.53
In physical ring topologies data packets pass through each node as they circulate the ring.

FDDI (Fibre Distributed Data Interface) and SONET (Synchronous Optical Network) networks are usually configured as physical rings and always operate as logical rings. FDDI can be used for LANs however it is more commonly used for longer distance high-speed connections. As the names suggest FDDI and SONET use optical fibre as the transmission media. FDDI is commonly used to connect an organisation's buildings whilst SONET is used for much greater distances. Both protocols use two physical rings with data circulating in different directions on each ring. Distances between FDDI nodes should not exceed 30km while distances in excess of 100km are common for SONET. For long distance applications the second ring is maintained solely as a backup should a fault occur in the primary ring. In such cases it is preferable to physically route the cabling of each ring separately. The aim being to improve fault tolerance should a cable be broken at any single location. If the cables for both rings are within close proximity (like within the same trench) then chances are that both cables will be broken together. When FDDI is used within a building then both rings can be used for data transmission, which effectively doubles the speed of data transfer.

- **Physical Hybrid Topology**

Hybrid or tree topologies use a combination of connected bus, star and ring topologies. Commonly a physical bus topology forms the backbone, with multiple physical star topologies branching off this backbone (see *Fig 3.54*). The backbone is installed through each building (or room) with a star topology used to branch out to the final workstations – the topology resembles the trunk and branches of a tree.

All hybrid topologies have a single transmission path between any two nodes. This is one reason the name 'tree' is used; consider the leaves on a tree, there is one and only one path from one leaf to another – the same is true for nodes in a physical hybrid or tree network.

Hybrid topologies are the primary topology of most organisations' networks. They allow for expansion – new branches can be added by simply connecting central nodes and branching out to the new workstations. It is common practice to install cabling that supports two or more times the anticipated transmission speed so that future expansion can easily and economically be accomplished. The extra cost of better quality higher-speed cabling being relatively insignificant compared to the installation costs. Consider the tree topology in *Fig 3.54*. It makes sense to install cabling that supports much higher data transfer speeds for the main backbone, whilst the cabling in each of the stars and rings is less critical.

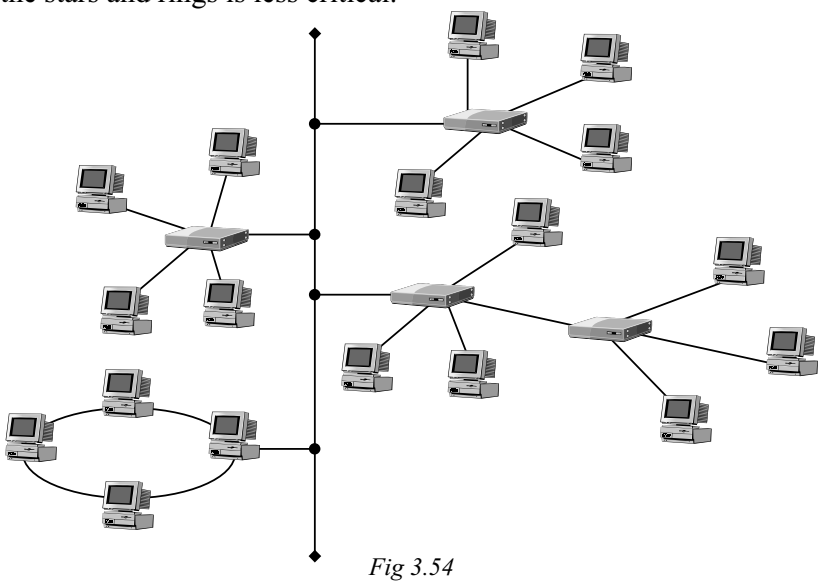


Fig 3.54
Physical tree topologies connect multiple bus, star and/or ring topologies such that a single path exists between each node.

GROUP TASK Practical Activity
Consider your school's physical network. Construct a diagram to describe the physical topology.

GROUP TASK Discussion
Discuss problems that could occur if there is more than one physical path between two nodes on a network.

- **Physical Mesh Topology**

Mesh topologies include more than one physical path between pairs of nodes. This is the primary topology of the Internet, where IP datagrams can travel different paths from the transmitter to the receiver. Mesh topologies require routers to direct each packet over a particular path. Without routers data packets can loop endlessly or they can be reproduced such that two or more copies arrive at the final destination.

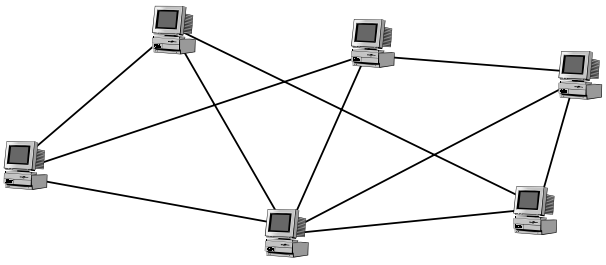


Fig 3.55
Mesh topologies include more than one path between individual nodes.

Commonly the nodes on a mesh network are all routers, and each router connects to further routers or a LAN. Mesh networks provide excellent fault tolerance, as packets are automatically routed around faults. A full mesh topology exists when all nodes are connected to all other nodes. Full mesh topologies are used in high-speed long distance connections where there are relatively few nodes and network performance and quality of service is absolutely critical. When a full mesh is used messages can be rerouted along any other path and hence fault tolerance is maximised.

Logical Topologies

The logical topology of a network describes how data is transmitted and received on a network, regardless of the physical connections. In some references the term "signal topology" is used in preference to the term "logical topology". In many ways this is a more descriptive term as the logical topology describes how signals are transferred between nodes on a network.

It is important to note that both electrical and light signals travel along transmission media at close to the speed of light. This is so fast that when a signal is placed on a wire or fibre it is almost immediately present at all points along the media. The speed of transmission is determined by the rate at which the sender alters the signal – in comparison the time taken for the signal to actually travel down the wire is relatively insignificant.

On an individual LAN the logical topology is in the majority of cases determined at the Transmission Level – the data link layer of the OSI model. The data link layer (layer 2) controls and defines how data is organised and directed across the network. This includes the format and size of frames as well as the speed of transmission. Commonly the unique MAC address of each node is used to direct messages to their destination. In essence the data link layer controls the hardware present at the physical layer (layer 1 of the OSI model).

Multiple LANs are commonly connected to form a WAN at the network layer. In an IP network routers direct messages in the form of IP datagrams to the next hop based on their IP address. Each hop in a datagram's journey may use different data link and physical layer protocols. The logical paths that datagrams follow describe the logical topology of WANs – commonly a logical mesh topology. We restrict our discussion to logical topologies operating within individual LANs.

In this section we discuss bus, ring and star (or switching) logical topologies at the datal link level. For each logical topology we identify common physical topologies upon which the logical signalling operates and we consider the media access controls used to deal with multiple nodes wishing to transmit at the same time.

- **Logical Bus Topology**

A logical bus topology simply means that all transmissions are broadcast simultaneously in all directions to all attached nodes. In effect all nodes share the same transmission media, that is, they are all on the same network segment. All nodes on the same network segment receive all frames – they simply ignore frames whose destination MAC address does not match their own. This presents problems when two or more nodes attempt to send at the same time. When this occurs the frames are said to collide – in effect they are corrupted such that they cannot be received correctly. A method of media access control (MAC) is needed to either prevent collisions or deal with collisions after they occur.

Prior to about 2004 logical bus topologies were by far the most popular – at the time a logical bus was the topology used by all the Ethernet standards. Furthermore switch

technology, which permits more efficient logical star topologies was expensive or simply not available. Currently switches are inexpensive and are required for the current full-duplex Gigabit and faster Ethernet standards.

Ethernet when operating over a logical bus topology uses CSMA/CD as its method of media access control (MAC). CSMA/CD is commonly associated with Ethernet, however in reality it is a MAC technique that is used by a variety of other albeit less popular low-level protocols. CSMA/CD is an acronym for "Carrier Sense Multiple Access with Collision Detection - quite a mouthful, however the general idea is relatively simple to understand.

The "Multiple Access" part of CSMA/CD simply refers to the ability of nodes to transmit at any time on the shared transmission media, as long as they are not currently receiving a frame. Remember that all nodes receive all frames at virtually the same time on a logical bus. If no frame is being received then the transmission media is not being used, therefore nodes are free to send. In *Fig 3.56* the transmission media is free after Node A completes transmission of a frame. This is the "Carrier Sense" part of CSMA/CD – in essence nodes must wait until only the carrier signal is present before sending. Say a node is not receiving and therefore it transmits a frame. Now it is possible that one or more other nodes have also transmitted a frame at the same time – they too were not receiving. If, or when, this occurs a collision takes place on the shared transmission media and all frames are garbled. In *Fig 3.56* a collision occurs when both Nodes B and C transmit at the same time. All nodes are able to detect these collisions and in response a jamming signal is transmitted – this is the "Collision Detection" part of CSMA/CD. In response all sending nodes wait a random amount of time and then retransmit their frames. In *Fig 3.56* Node C waits a shorter time than Node B, hence Node C transmits it's frame prior to Node B.

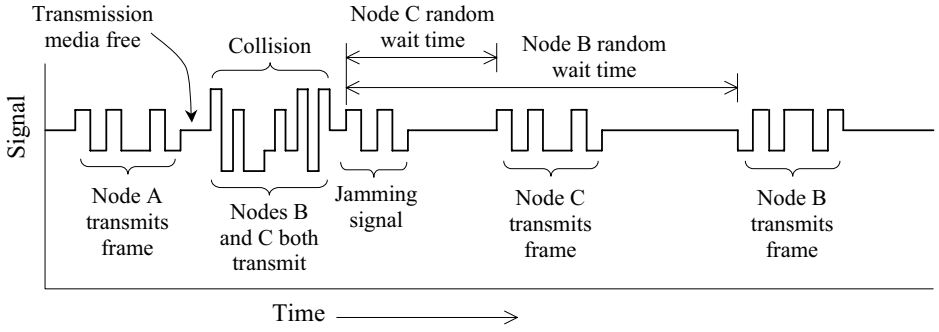


Fig 3.56

CSMA/CD strategy where node B and node C are waiting to transmit after node A has finished.

GROUP TASK Discussion

It is possible for short Ethernet frames to collide after they have been successfully sent. This is more likely where there are large physical distances between the sending nodes. Why is this so? Discuss.

Clearly a physical bus topology supports a logical bus topology. Examples include the earlier Ethernet standards that use coaxial cable, such as 10Base2 (also known as Thinnet) and the earlier 10Base5 standard (also known as Thicknet). There are also Ethernet standards using optical fibre that utilise physical and logical bus topologies. We will examine many of the commonly used Ethernet standards later in this chapter when we consider transmission media and cabling standards in some detail.

Most current Ethernet networks are wired with UTP (Unprotected Twisted Pair) cable into a physical star topology. When connected via a hub a logical bus topology is

being used. Hubs simply repeat all received signals out to all connected nodes; therefore all nodes share a common transmission medium and exist on the same network segment. We examine the operation of hubs in more detail later in this chapter. In terms of logical topologies, conceptually we can think of a hub containing a mini backbone shared by all nodes. 10BaseT and 100BaseT are common Ethernet standards that are wired into a physical star, but use a logical bus topology when the central node is a hub.

Current wireless LANs (WLANs) based on the IEEE 802.11 standard use a logical bus topology. The 802.11 standard specifies two "physical" types of WLAN, those with a central node in the form of a wireless access point (WAP) and "ad-hoc" WLANs where nodes connect directly to each other. Those with a central WAP utilise a physical star topology. Essentially a WAP amplifies and repeats signals much like a wired hub – all nodes hear all messages from the WAP. Ad-hoc WLANs use a physical mesh-like topology that changes dynamically as nodes connect and disconnect.

GROUP TASK Research and Discussion
Why do you think "ah-hoc" wireless LANs have been described as having a physical "mesh-like" topology? Research and discuss.

On all current (2007) 802.11 WLANs all nodes transmit and receive using a single wireless channel – hence a logical bus topology is being used. The characteristics of wireless transmission make CSMA/CD an inappropriate media access control strategy. Wireless nodes are effectively half-duplex as they are unable to reliably listen to a signal whilst they are transmitting – the wireless signal being drowned by their transmission. As a consequence detecting collisions during transmission is difficult. To overcome this issue 802.11 WLANs use CSMA/CA as their media access control strategy rather than CSMA/CD. CSMA/CA is an acronym for "Carrier Sense Multiple Access with Collision Avoidance". As the name implies, CSMA/CA attempts to prevent data collisions occurring rather than dealing with collisions once they have occurred. The CSMA/CA strategy is not new; it was integral to the operation of AppleTalk networks used by early Apple Macintosh computers.

So how does CSMA/CA avoid collisions? Like CSMA/CD each node must first wait for the transmission media to be free. Unlike collision detection nodes must then wait a random amount of time before commencing transmission. In *Fig 3.57* Node C has generated a shorter wait time than Node B so no collision occurs. This simple strategy avoids most of the collisions that occur on CSMA/CD networks. Using CSMA/CD numerous nodes are likely to be waiting for a clear transmission media and as soon as the line is clear they all commence transmission together resulting in collisions such as the one detailed in *Fig 3.56* above. Using CSMA/CA waiting nodes will rarely commence transmitting simultaneously.

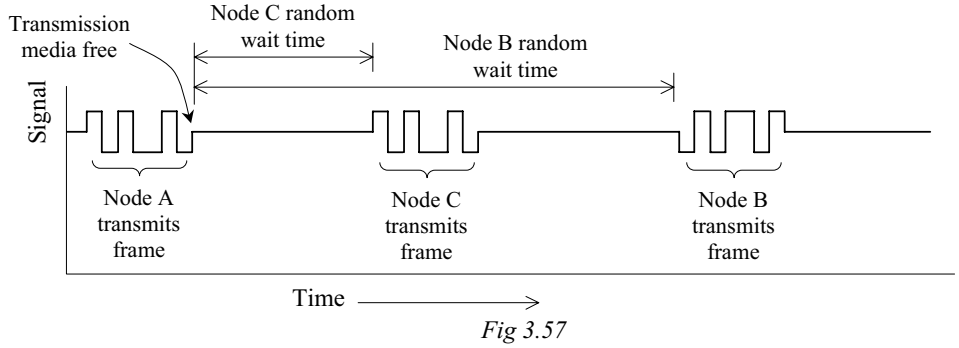


Fig 3.57
CSMA/CA strategy where node B and node C are waiting to transmit after node A has finished.

Further collision avoidance strategies are optionally employed on 802.11 WLANs. One system, known as RTS/CTS, allows nodes to reserve the transmission media in advance. The system can be turned completely off or on, or more commonly the system is used for frames exceeding a preset byte length. Using the RTS/CTS system a node waiting to transmit first sends an RTS (Request To Send) frame. This RTS frame contains a duration ID field that specifies the time the sending node will require the transmission media. In response a CTS (Clear To Send) frame that also contains a duration field is returned. Nodes only send data frames after they have received a CTS frame. Other nodes also receive the CTS frame so that they do not commence sending until sufficient time has elapsed.

GROUP TASK Practical Activity
Examine the configuration screens for a WAP (wireless access points are also included within devices commonly known as "wireless routers"). Identify and describe the purpose of any RTS/CTS settings.

No collision detection or avoidance scheme is 100% perfect – some collisions will not be detected whilst other frames will continue to collide on subsequent transmission attempts. All OSI layer 2 protocols specify some limit to the number of retries that can occur for individual frames. Eventually some frames are simply dropped. Dealing with such failures is left up to the higher OSI layer protocols where definite positive acknowledgement of transmission is required.

There exists media access control (MAC) strategies used over shared transmission media that avoid the possibility of collisions completely. TDMA (time division multiplexing) is used on some fixed and mobile phone networks whilst polling is used for some data networks. The 802.11 WLAN standard includes the option to include polling functionality. Essentially polling gives total control of media access to one node. This node then asks each node in turn if it wishes to transmit.

GROUP TASK Research
Using the Internet, or otherwise, research the essential features and differences between TDMA and polling MAC strategies.

- **Logical Ring Topology**

When a logical ring topology is used each node receives frames from one and only one node and transmits frames to one and only one node. As a consequence all frames circulate a logical ring. Each node receives and transmits each frame so that all frames circulate around the entire ring. The destination or recipient node takes a copy prior to transmitting the frame. Collisions are simply impossible on logical ring topologies.

IBM's original "token ring" protocol was once the most common LAN protocol – Ethernet has largely replaced "token ring". However, the general operation of token ring networks is also implemented within long-distance high-speed networks including FDDI and SONET protocols.

In most logical ring implementations a single frame (known as a token) circulates the ring continuously. When a node wishes to send it must wait for the token. It then attaches its data to the token and sends it on its way. The frame containing the data continues around the ring being received and transmitted in turn by each node until it reaches the recipient. The recipient takes a copy of the data and also sends the frame on to the next node. Eventually the data frame returns to the original sender. The sender then removes the data frame and sends out the token. The token continues to circulate until the next node wishes to send.

Early IBM Token Ring networks were wired into a physical ring topology (see *Fig 3.58*). Later implementations used a physical star topology where the central node was a Multistation Access Unit (MAU) as shown in *Fig 3.59*. Conceptually a MAU can be thought of as containing a miniature ring. MAUs are able to automatically sense when a node is either not attached or is not powered and close the ring accordingly.

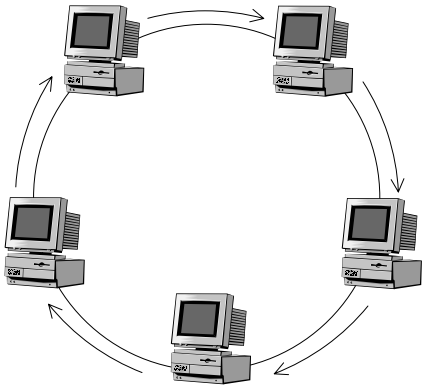


Fig 3.58
IBM Token Ring with physical and logical ring topology.

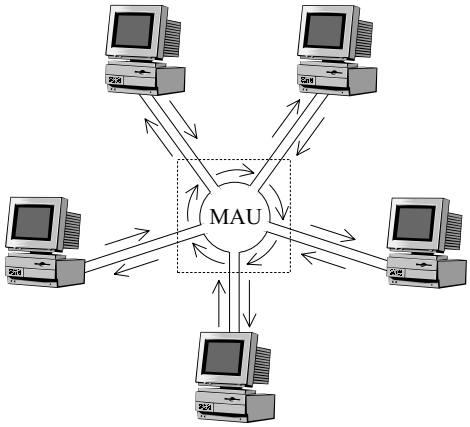


Fig 3.59
IBM Token Ring with physical star topology and logical ring topology.

GROUP TASK Research
Using the Internet, or otherwise, research the data transfer speeds achieved using IBM's Token Ring networks.

FDDI and SONET are both used for long distance communication. In these cases the nodes are routers rather than computers. These routers include connections to other networks not just to adjacent nodes in the ring. In most examples a physical ring topology is used in conjunction with logical ring topologies. Common FDDI and SONET networks are operated by large business, government or telecommunication companies using fibre optic cable. Currently data transfer rates of 40Gbps are achieved using SONET.

GROUP TASK Research
SONET speeds are based on STS levels and Optical Carrier (OC) specifications. Use the Internet to research the speed of SONET based networks based on different STS levels and OC specifications.

SONET rings provide many of the major Internet and PSTN links between major cities. As a consequence such networks must ensure quality of service at all costs. A single physical ring is unsuitable for such networks as a single break in a cable disables the entire network. To solve this problem FDDI and SONET use multiple connected rings. Most FDDI implementation use dual rings – the second existing as a redundant backup should the first fail. Many SONET networks utilise many more than two rings. These multi-ring networks are known as "self-healing rings" and are able to divert data packets around problem areas in a virtual instant. For our discussion we will consider a typical dual-ring FDDI or SONET ring configuration.

When dual rings are used the tokens on each ring rotate in different directions. Say, clockwise for the primary ring and anti-clockwise on the secondary (or standby) ring. Note that under normal conditions the secondary ring is not being used. Imagine a fault occurs in the primary ring – the secondary ring can then become the active ring

whilst the fault is corrected. Now imagine both rings are cut, perhaps by a backhoe physically cutting through the cable. This situation is illustrated in *Fig 3.60* where the cable connecting Node B and Node C has been cut. The new transmission path is shown using dotted arrows. Notice that data still travels in the original direction on both the primary and secondary rings.

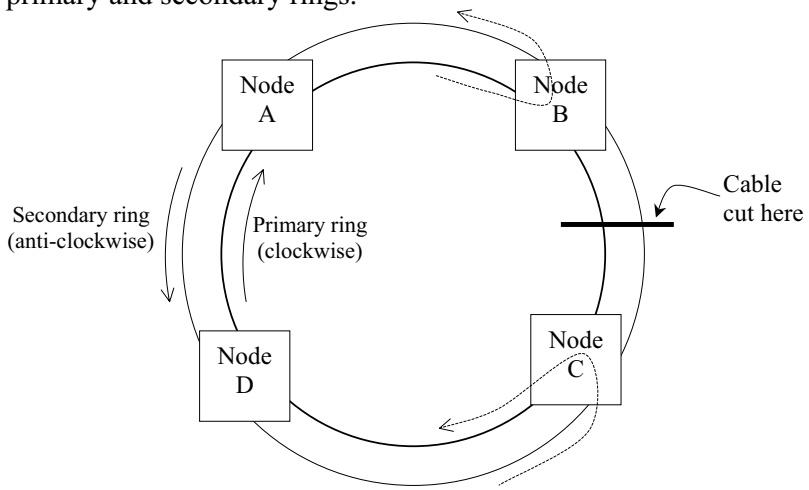


Fig 3.60
Dual ring topology where a cable has been cut causing a new logical ring to be automatically created.

More rings can be added to further improve the fault tolerance or "self-healing" ability of critical ring networks. Note that many complex implementations that more closely resemble a physical mesh topology are used; yet all maintain a logical ring topology.

GROUP TASK Discussion
Identify possible points of failure for each of the physical topologies shown in *Figures 3.58, 3.59* and *3.60*. Suggest how the possibility of such failures could be avoided.

- **Logical Star Topology**

In a logical star topology each node has its own connection to a switch that is the central node. In many references logical star topologies are known as logical switch topologies. Currently all logical star topologies also use a physical star topology.

On a logical star every node exists on its own network segment with the switch. Switches are OSI data link layer 2 devices. In current configurations this connection is full duplex, as it includes two distinct transmission channels – one for sending and one for receiving. Most Ethernet networks use a twisted pair of copper wires (UTP) for each of these channels. Collisions are impossible on logical stars. Frames on each channel always travel in a single direction – either a frame is travelling from node to switch or it is on the other channel travelling from switch to node. Situations where two or more frames exist on a single channel can never occur.

When a node sends a frame the switch detects the destination MAC address and transmits the frame only to the node with that MAC address. Switches are able to process multiple frames simultaneously that are addressed to different nodes. We consider the operation of switches in more detail later in this chapter.

GROUP TASK Discussion
Compare and contrast examples of physical star topologies that use logical bus, logical ring and logical star topologies.

HSC style question:

Luke's Limos is a used car business comprised of three car yards located in adjoining suburbs of Sydney. Currently each car yard has its own Ethernet network that includes a central switch, laser printer and a cable broadband connection to the Internet.

Each of the four salesmen at each car yard has a computer in their office where they record information in regard to their contacts with customers. Currently each salesman is free to record this information in a way they feel best meets their needs.

All computers at each car yard are able to access detailed information in regard to the vehicles for sale at their particular site. This information is stored in a simple flat file database located on the sales manager's computer at each car yard.

All cars currently for sale at all three yards are advertised on a website that is maintained by a web design company. When a car is being prepared for sale an email is sent to the web designer. The email includes the basic details, sale price and an attached photo of the vehicle. When a car is sold the web designer is again emailed so that the vehicle can be removed from the website.

(a) (i) Draw a diagram to represent the physical network topology at one of the car yards.

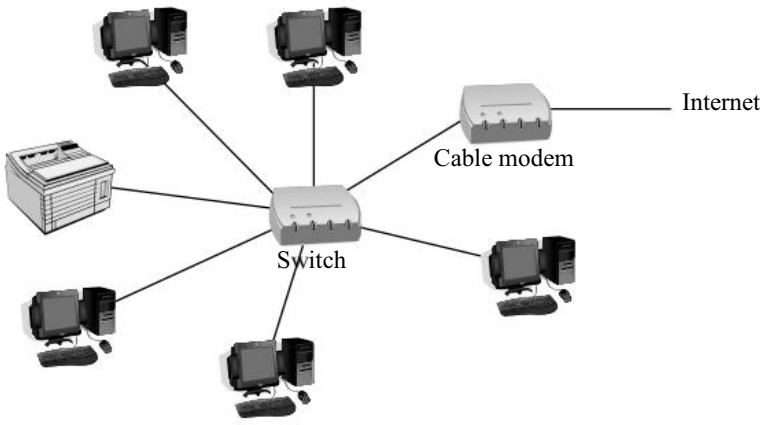
(ii) Explain how data collisions are detected (or avoided) within each car yard's network.

(b) The owner is considering opening a further two car yards within the next year and wishes to explore ways of improving the information flow throughout the business. The owner intends to implement a team approach to selling cars. This requires that all salesmen are able to view the details of all vehicles and all customer contacts within the business.

Discuss suitable modifications and/or additions to the current information system to assist the owner achieve this objective.

Suggested Solution

(a) (i)



(ii) Switches set up a dedicated circuit between sender and receiver. This means it is impossible for collisions to occur. In essence every combination of every pair of nodes is in its own network segment.

(b) Possible modification/additions that could include:

- Use of a single relational database to hold full details of all cars and edited directly by Luke's Limos salesmen.
- The relational database of all car details is held or at least accessed from a web server. Web pages describing each car are produced dynamically using data from the new relational database.
- The web designer sets up and maintains the webpage generation process for the general public. The web pages being generated from the database when requested by a browser.
- Customer contacts could also be included within the database accessed by the web server.
- A replicated database could be used for customer contacts. New contacts being entered at each car yard and then distributed to other yards automatically during the replication process.
- A distributed database could be used for customer contacts where local contacts are physically stored at each individual car yard. However the data from all car yards is still available to all other car yards.
- The customer contact database should include a system where all salesman who have had dealings with a customer are informed of any new contacts with that customer. Perhaps a simple automated email could be sent to all salesman who have had previous contact with a customer whenever a new customer contact occurs.

Comments

- In (a) (i) a star physical topology should be drawn that includes all the devices mentioned in the question.
- In (a) (ii) mention could be made of a star/switched logical topology hence no possibility of collisions occurring.
- Although incorrect, it is likely that some portion of the marks in (a) (ii) would be allocated if CSMA/CD or CSMA/CA was described correctly.
- In part (b) there are many other possible modifications and additions that could be discussed. It is important that each modification/addition is related directly back to the requirements of the new system that are outlined in the question.
- Note that part (b) combines aspects of the database and communication topics.
- Part (b) is an extended response question that would likely be worth 4 to 6 marks in a real examination. Therefore a number of points should be made and explored in some depth. The suggested answer includes many points, however each point could well be explored in greater detail.

SET 3F

1. Which of the following is TRUE of client-server systems?
   (A) Clients must understand the detail of server processes.
   (B) Servers process client requests.
   (C) Clients provide services to servers.
   (D) Servers are always dedicated machines.

2. An employee uses their laptop at home to connect to a server at their work using a thin client RDP Internet connection. Which of the following is TRUE?
   (A) Applications run on the client.
   (B) Applications run on the server.
   (C) The laptop has no hard disk.
   (D) No data is transmitted to the server.

3. The physical topology of a network:
   (A) determines how data is transferred between devices.
   (B) can change when different protocols are installed.
   (C) describes and determines how nodes communicate with each other.
   (D) describes how devices are physically connected to each other.

4. A break in a single cable is more significant when using a:
   (A) physical bus or star topology.
   (B) physical ring or star topology.
   (C) physical ring or bus topology.
   (D) physical mesh topology.

5. Multiple paths between nodes is a feature of:
   (A) physical mesh topologies.
   (B) physical bus topologies.
   (C) physical star topologies.
   (D) physical tree topologies.

6. On an Ethernet LAN each node is connected via UTP to a central hub. Which topology is being used?
   (A) Physical star, logical bus.
   (B) Physical star, logical star.
   (C) Physical bus, logical bus.
   (D) Physical bus, logical star.

7. In regard to topologies and the OSI model, which of the following is generally TRUE?
   (A) Logical topologies for WANs are determined at the data link layer and for LANs at the network layer.
   (B) Logical topologies for LANs are determined at the data link layer and for WANs at the network layer.
   (C) Physical topologies for LANs are determined at the data link layer and for WANs at the network layer.
   (D) Physical topologies for WANs are determined at the data link layer and for LANs at the network layer.

8. All nodes receive all transmissions at virtually the same time when using which logical topology?
   (A) Ring
   (B) Star
   (C) Switched
   (D) Bus

9. What is a data collision?
   (A) Corruption when a nodes starts receiving whilst it is still transmitting.
   (B) A procedure used to ensure transmissions arrive at their destination on logical bus topologies.
   (C) Corruption of messages due to multiple nodes transmitting simultaneously on the same communication channel.
   (D) A fault in the logical topology such that multiple nodes are able to transmit at the same time.

10. Critical ring networks are said to be "self healing", what does this mean?
    (A) Cables are able to repair themselves when broken.
    (B) Each node contains redundant components that take over should the primary component fail.
    (C) Data traffic can be automatically diverted around faults.
    (D) Two or more physical rings are installed.

11. Define each of the following terms and provide an example:
    (a) Client-server architecture (b) Physical topology (c) Logical topology

12. Construct a table of advantages and disadvantages of:
    (a) Physical bus, star and ring topologies. (b) Logical bus, star and ring topologies.

13. Explain how data collisions are prevented, avoided or detected on each of the following networks:
    (a) Ethernet over a logical bus topology.
    (b) IEEE 802.11 wireless LAN.
    (c) IBM Token Ring network.

14. Distinguish between thin clients and fat clients using examples.

15. Maximising fault tolerance of critical networks is a major priority. Describe at least THREE techniques that improve a network's fault tolerance.

ENCODING AND DECODING ANALOG AND DIGITAL SIGNALS

For communication to take place both transmitting and receiving must occur successfully. Transmitting involves the sender encoding the message and transmitting it over the medium. Receiving involves the receiver understanding the organisation of the encoded message – based on the protocols agreed upon during handshaking with the transmitter. The receiver can then decode the message based on the rules of the agreed protocols. In essence both encoding and decoding are organising information processes. Encoding organises the data into a form suitable for transmission along the communication medium. Decoding changes the organisation of the received data into a form suitable for subsequent information processes.

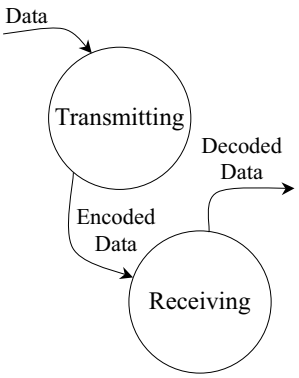


Fig 3.61
Transmitting encodes data and receiving decodes data.

Prior to transmission data is encoded into a signal according to the rules of the transmission protocols being used and suited to the transmission media along which the message will travel. When messages reach their destination the receiver reverses this process by decoding the signal and transforming it back into data.

Data that originates or is stored on a computer is always in binary digital form. Digital data is all data that is represented (or could be represented) using whole distinct numbers – in the case of computers a binary representation is used. Continuous data that usually originates from the real world is analog. Both analog and digital data can be encoded and transmitted on electromagnetic waves. Note that in reality all waves are continuous hence they are analog. For our purpose, it is how we choose to interpret the data carried on these analog waves that we shall use to distinguish between digital signals and analog signals. A digital signal is being used when digital data is encoded onto an analog wave. An analog signal is being used when analog data is encoded onto an analog wave.

To encode analog data into a digital signal requires that the data first be converted into digital using an analog to digital converter (ADC). Similarly to encode digital data into an analog signal the data must be converted to analog data using a digital to analog converter (DAC).

GROUP TASK Discussion
Discuss and develop definitions for the terms "digital data" and "analog data".

Analog Data to Analog Signal

When the data is analog the waveform varies continuously in parallel with the changes in the original analog data. For example microphones collect analog sound waves and encode them as an infinitely variable electromagnetic wave (see *Fig 3.62*). The voltage transmitted from the microphone varies continuously in parallel with the sound waves entering the microphone. An analog signal is produced as the entire analog wave represents the original analog data. All points on the analog wave have significance – this is not true of digital signals.

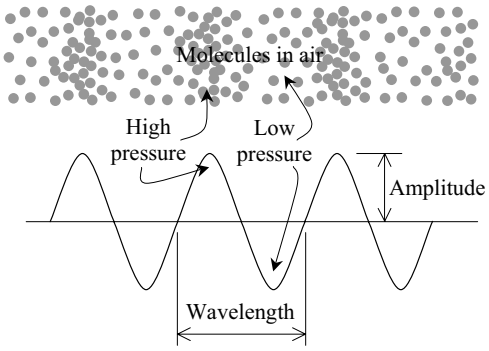


Fig 3.62
Microphones convert analog sound waves into analog signals carried on analog waves.

Analog signals are transmitted along traditional PSTN telephone lines. For voice (audio) microphones are used as the collection device and speakers as the display devices. The microphone encodes the analog data and the speaker performs the decoding process. The electromagnet within the speaker moves in and out in response to the received analog signal. This causes the speaker's diaphragm to move in and out which in turn creates compression waves through the air that we finally hear as sound.

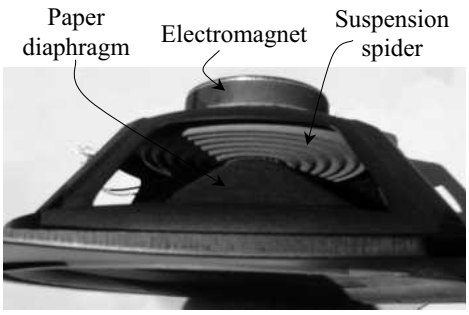


Fig 3.63
Underside of a typical speaker.

Traditional analog radio and analog TV are further examples of analog data transmitted as an analog signal – including broadcasts through air and also analog audio and video cassettes (VHS). In both cases an analog signal is transmitted that varies continuously. This analog signal is decoded and displayed by the receiving radio/stereo or television set.

GROUP TASK Discussion
Brainstorm a list of collection and display devices. Classify the data collected or displayed by the device as either analog or digital.

Digital Data to Digital Signal

Digital signals are produced when digital data is encoded onto analog waves. To decode the wave and retrieve the encoded digital data requires the receiver to read the wave at the same precise time intervals. The receiver determines the characteristics of the wave at each time interval based on the details of the coding scheme. As a consequence each particular waveform can be decoded back into its original bit pattern.

There are two commonly used techniques for encoding digital data. The first alters the voltage present in a circuit to represent different bit patterns. This technique is used over short distances, including communication within a computer and between nodes on a baseband LAN. Note that altering voltage changes the power or amplitude of the wave. The second alters characteristics of a constant frequency electromagnetic wave called a carrier wave. The carrier wave is modified (modulated) to represent different bit patterns by altering a combination of amplitude, phase and/or frequency. The modulation (and subsequent demodulation) process is used for most long distance broadband communication. Both the above encoding techniques create different waveforms (often called symbols) that represent different numbers (bit patterns). The waveforms are changed at regularly spaced time intervals to represent each new pattern of bits.

GROUP TASK Discussion
Encoding schemes that alter voltage between two levels are unsuitable for long distance communication. Why is this? Discuss.

The time between each interval is known as the "bit time". For example on a 100baseT Ethernet network the bit time is 10 nanoseconds. Therefore a transmitting network interface card (NIC) on a 100baseT network ejects one bit every 10 nanoseconds. Similarly all receiving nodes must examine the wave every 10 nanoseconds. On 100baseT protocol networks a single bit is represented after each bit time using Manchester encoding (see *Fig 3.64*) – low to high transitions (waveforms)

represent binary ones and high to low transitions represent binary zeros. The receiver detects the transitions to not only decode the signal but also to remain in synchronisation with the sender.

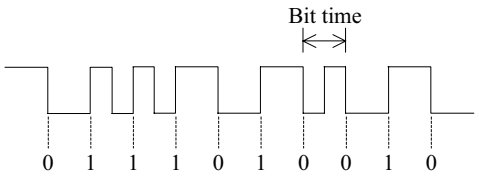


Fig 3.64
Manchester encoding uses the transitions between high and low to represent bits.

Notice that the waveform shown in *Fig 3.64* is essentially square. This is a somewhat simplistic view; in reality the wave is analog and therefore not precisely square. Each transition from high to low or low to high occurs over time. Therefore the actual wave has rounded edges. In *Fig 3.64* we are describing a digital signal where high and low voltages are used to create the transitions that represent ones and zeros – only the detail necessary to detect the encoded digital data in the signal is shown.

GROUP TASK Discussion
Review and discuss how Manchester encoding assists the receiver to remain synchronised with the sender.

Higher speed and/or longer distance protocols represent multiple bits within each distinct waveform. Consider DSL and cable modems which modulate the carrier wave's amplitude and phase within a predetermined range of frequencies. QAM (Quadrature Amplitude Modulation) is currently the dominant protocol. A modem (DSL or cable) that uses 256QAM represents 8 bits after each bit time elapses. As there are 256 different combinations of 8 bits then 256QAM uses 256 different waveforms known as symbols. Each distinct symbol having a unique combination of phase and amplitude. Current cable modems using 256QAM typically transmit (and receive) more than 5Msym/s (5 million symbols per second). As each symbol represents 8 bits then speeds around 40Mbps are achievable. *Fig 3.65* is a conceptual view of 256QAM – notice that each different 8-bit pattern is represented by a different waveform or symbol. In reality each different waveform is repeated continually during each bit time.

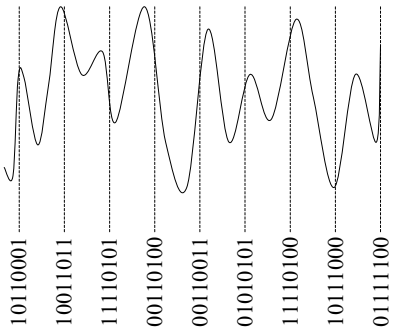


Fig 3.65
Conceptual view of modulation using 256QAM.

Encoding schemes, like QAM, that modulate carrier waves are used within all long distance and/or high-speed low-level protocols (OSI layer 1 and 2). This includes long distance Gigabit and faster Ethernet standards, SONET, FDDI and ATM. These protocols operate on various types of transmission media including wire, fibre optic and wireless mediums.

For digital signals the speed of transmission can be increased in two fundamental ways – by increasing the number of bits represented by each symbol or by decreasing the bit time (equivalent to increasing the symbol rate). The quality of the transmission media and limitations of the transmitting and receiving hardware determine the extent to which distinct symbols can be determined. As the number of symbols increases the difference between each symbol is more difficult to determine. Similarly as bit times decrease the accuracy of synchronisation between sender and receiver must increase.

GROUP TASK Research
Most current modulation schemes use amplitude modulation (AM) and phase modulation (PM) to alter the carrier wave. Using the Internet, or otherwise, research reasons why frequency modulation is seldom used.

Digital Data to Analog Signal

Converting digital data to an analog signal requires the data to first be converted to analog prior to its transmission as an analog signal. A digital to analog converter (DAC) performs this process. Digital to analog conversion is used between video cards and analog monitors and is also used to connect dial-up modems to traditional analog PSTN telephone lines. It is also used when playing audio CDs and DVDs.

As digital data contains distinct rather than continuous data then during digital to analog conversion it is necessary to estimate the intermediate waveforms between each known digital data point. In *Fig 3.66* the dotted lines represent each of the digital samples and the solid line represents the analog signal. Note that the analog signal is a smooth curve produced by estimating the shape of the curve between pairs of digital samples. Audio CDs use PCM (Pulse Code Modulation) to encode the original analog music as a sequence of 16-bit digital sound samples – approximately 44100 per second. When a CD is playing the waveform between each digital sample is estimated based on the values of the adjoining digital samples (refer *Fig 3.66*). For audio CDs the digital samples are so close together that such estimations are imperceptible to listeners.

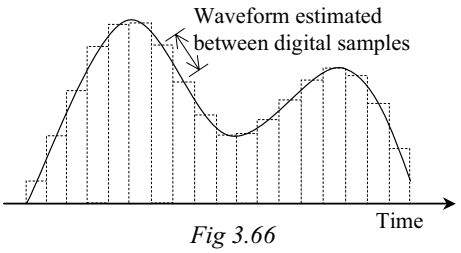


Fig 3.66
Digital to analog conversion estimates waveforms between digital samples.

Today dial-up modems are rarely used to connect to the Internet, however they are routinely used to transmit fax data over traditional PSTN lines. In the past the infrastructure at local telephone exchanges was built to deal exclusively with analog voice signals. A total bandwidth of 3200Hz, ranging from 200Hz to 3400Hz, was used as these frequencies encompass the normal frequencies present in natural speech. Frequencies above 3400Hz were filtered out of the signal completely. As a consequence the signal transmitted (and received) by dial-up modems had to simulate and operate within the same frequency range as analog voice signals. The devices connecting telephone exchanges did not differentiate between voice and other data transmissions. In terms of encoding and decoding processes occurring within the PSTN both voice and data were both transmitted and received identically as analog signals.

Consider the following

Consider the operation of the simple DAC described below in *Fig 3.67*. This DAC makes no formal attempt to smooth its analog output, however some smoothing occurs as the output signal moves from one level to another during switching. In this case each sample contains just 4 bits. Each bit activates a switch that allows current to flow (or not flow) through a resistor. Each resistor allows a different proportion of the voltage through. In the diagram the digital sample 1010 is being processed. If the input voltage is 5 volts then the first 1 in the sample allows five volts through and the next 1 allows just one quarter of 5 volts through – the finally output being 6.25 volts.

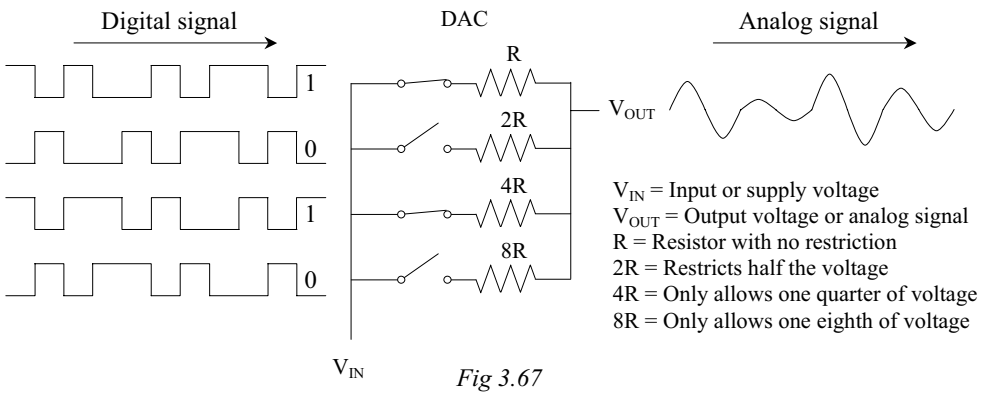


Fig 3.67

A simple binary weighted DAC uses weighted resistors to alter the signal's output voltage or amplitude.

GROUP TASK Discussion

Assume the above DAC has an input voltage of 5. Make up a table listing the voltages output for all possible 4-bit combinations.

Analog Data to Digital Signal

In this case we have continuous analog data that is to be represented digitally during its transmission. Today this routinely occurs when transmitting audio and video analog data within all types of communication networks including the PSTN, VoIP, cable TV network and digital TV network. For analog data to be transmitted digitally first requires the data to be converted to digital using an analog to digital converter (ADC).

Telephone calls from normal home phones are transmitted as analog signals to the local exchange. The analog data is converted to digital data at the exchange where it travels using a digital signal to the receiver's local exchange. At the receiver's local exchange the digital signal is received, the data is converted back to analog and then transmitted as an analog signal to the receiver's residence. Mobile phones convert the analog sound waves to digital within each phone; therefore digital signals are used exclusively to transmit data between mobile phones.

Analog to digital converters (ADCs) repeatedly sample the analog data and convert each sample to a binary number. ADCs are present within many collection devices including sound cards, video capture cards, TV cards, optical mice, scanners and digital still and video cameras. The analog to digital conversion process produces sequences of binary numbers that represent the analog data at particular regular points. For images the sampling points are known as pixels, whilst for audio the sampling points are time based. Video includes both pixel and time based samples.

Consider the following

The components and data connections in a simple ADC within a computer's sound card are shown in *Fig 3.68*; this ADC performs its conversion using the following steps:

- At precise intervals the incoming analog signal is fed into a capacitor; a capacitor is a device that is able to hold a particular electrical current for a set period of time, this allows the ADC to examine the same current repeatedly over time.
- An integrated circuit, called a successive approximation register (SAR), repeatedly produces digital numbers in descending order. For 8-bit samples it would start at 255 (11111111 in binary) and progressively count down to 0.

- The DAC receives the digital numbers from the SAR and repeatedly produces the corresponding analog signal. The analog signals will therefore be produced with decreasing levels of electrical current.
- The electrical current output from the DAC is compared to the electrical current held in the capacitor using a device called a comparator. The comparator signals the SAR as soon as it detects that the current from the DAC is less than the current in the capacitor.
- The SAR responds to the signal from the comparator by storing its current binary number. This number becomes one of the digital samples.
- The SAR resets its counter and the whole process is repeated.

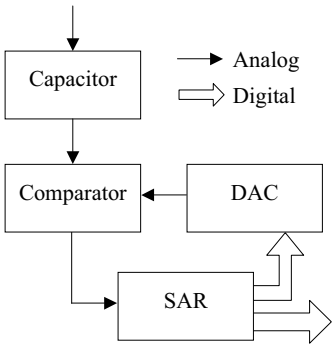


Fig 3.68
Components and data connections for a simple ADC.

GROUP TASK Discussion
Is the ADC described above suitable for use within a mobile phone? What about within a digital still camera? Discuss.

Consider the following

The distinction between digital signals and analog signals is not clear cut. Most would agree that a signal that represents a binary 1 as a high voltage and a binary 0 as no or low voltage is best described as a digital signal. However during transmission this signal is still an analog wave – all waves are continuous by their very nature. Consider a signal that uses hundreds of different symbols to represent different bit patterns. This signal includes a carrier wave encoded with combinations of frequency modulation, amplitude modulation and/or phase modulation to represent digital data. Here we have a finite number of different symbols that are transmitted on a continuous wave.

GROUP TASK Discussion
During our discussion of analog and digital signals we used how these signals are interpreted as the fundamental difference. Do you agree? Discuss and debate the difference between analog signals and digital signals.

NETWORK HARDWARE

In this section we describe:

- transmission media along which signals travel,
- network hardware that connects to the transmission media and
- various types of network servers.

These are the essential hardware components required to connect nodes to form a communication network.

TRANSMISSION MEDIA

Signals are transmitted along a transmission media. The transmission media can either be bounded or wired such as twisted pair, coaxial cable and optical fibre or it can be unbounded such as wireless connections used for satellite links, wireless LANs and mobile phones. The transmission media forms part of the OSI Physical Layer 1.

Wired Transmission Media

Wired or bounded transmission media restricts the signal so that it is contained within a cable and therefore follows the path of the cable. In addition wired media can be shielded to prevent (or at least limit) external electromagnetic forces from affecting the signal. No cable is perfect in this regard which means signals do degrade as distances increase. Different standards are in place, which specify various technical attributes of cables. These attributes determine the maximum recommended distance between nodes.

We restrict our discussion to the three most common types of wired media, namely twisted pair, coaxial cable and optic fibre cable.

- **Twisted Pair**

Twisted pair cable, as the name suggests, is composed of pairs of copper wire twisted together. Each copper wire is contained within plastic insulation and then the twisted pairs of wire are enclosed within an outer sheath. The regular twists in each pair are specifically designed to limit the electromagnetic interference between pairs and also to a lesser extent from outside sources.

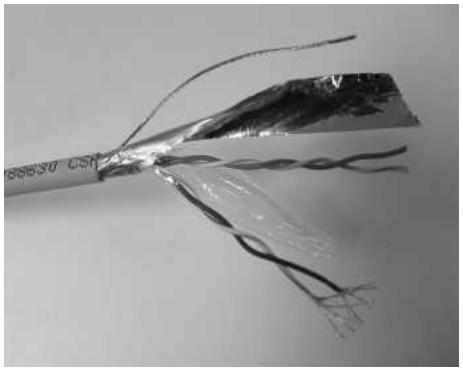


Fig 3.69
Shielded Twisted Pair Cable (STP)

Unshielded Twisted Pair (UTP) is the most common and economical form of copper cable for both LAN and telephone connections. UTP cable does not include any physical shield against outside electromagnetic interference (apart from somewhat limited shielding provided by the twists in each pair). Most UTP cables contain four pairs – a total of eight copper wires. Shielded Twisted Pair (STP) and Screened Twisted Pair (ScTP) includes a metal shield or screen and a drain wire (see *Fig 3.69*). STP and ScTP cable is significantly more expensive therefore its use is limited to applications where a high level of electromagnetic interference is present – primarily industrial applications.

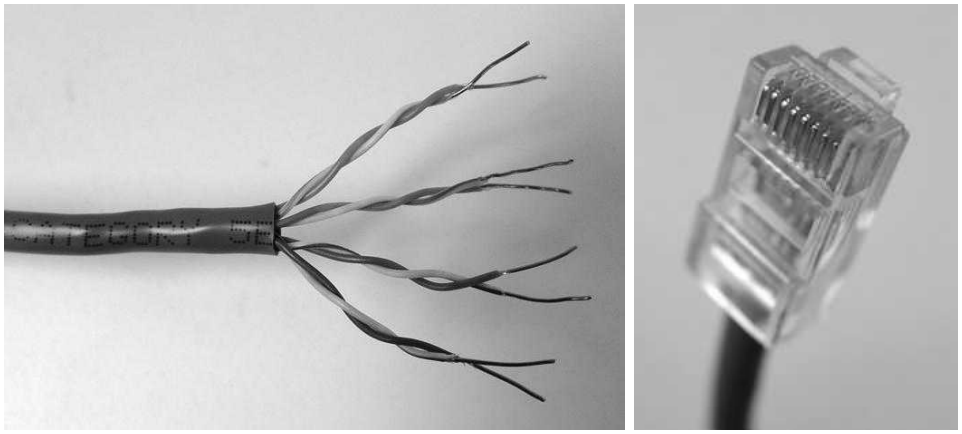


Fig 3.70
Category 5e UTP cable (left) and RJ45 plug (right).

UTP is classified into categories where higher category cable supports higher frequencies and hence high data transfer speeds. Cat 6 cable supports frequencies up to 250MHz whilst the more common Cat-5e cable supports frequencies up to 125MHz. Lower specification Cat 3 cable supports frequencies up to 16MHz and was once popular for 10Mbps networks – today Cat 3 cable is used almost exclusively for telephone lines.

Today (2007) most baseband Ethernet networks use Cat-5 or greater UTP – Cat-5e being the most common although Cat-6 is recommended for new installations. In general individual UTP cable runs should not exceed 100 metres from the central node (usually a switch) to the end node (usually a computer). In permanent installations a maximum run of 90 metres is used so that 10 metres remains to accommodate the patch cables that run from the wall socket to the computer and from the patch panel (see *Fig 3.71*) to the switch. RJ45 female connectors are used on the patch panels, wall sockets and switches. Male RJ45 connectors are used on both ends of the patch cables (see *Fig 3.70* above). Longer UTP cable runs can be accommodated under some circumstances by using higher specification cable.

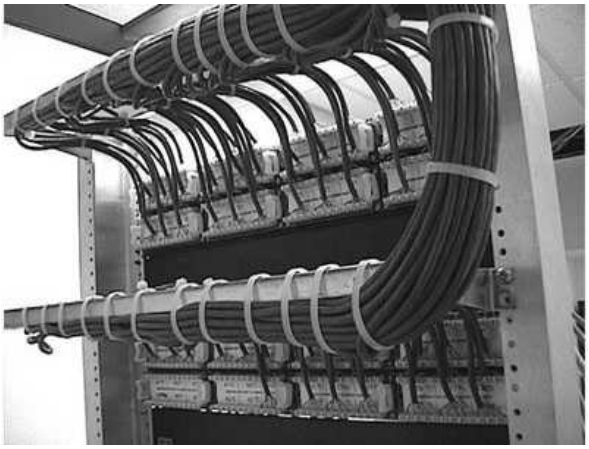


Fig 3.71
Rear view of a typical Cat-5e UTP patch panel.

10baseT Ethernet can operate on Cat-3 or above and 100baseT on Cat-5 and above. Both these standards use just two of the four twisted pairs for data transfer. 1000baseT or Gigabit Ethernet uses all four pairs and operates best on Cat-5e and above cable. Faster Ethernet standards of 10Gbps and above require Cat-6 or Cat-7 cables. The use of higher specification Cat-7 cable allows longer distances between nodes, the specific allowable distances change depending on the speed and configuration of the network.

Cat 3 and even lower specification cable is used to transmit broadband ADSL signals. ADSL splits the total bandwidth into a series of channels. Each channel is assigned a specific range of frequencies – commonly each channel has a bandwidth of 4kHz. Given that Cat-3 supports frequencies up to 16MHz it is more than capable of supporting the hundreds of 4kHz bandwidth channels required by ADSL.

GROUP TASK Research
New network and cabling standards are regularly released. Using the Internet, or otherwise, determine recent Ethernet standards for UTP together with the required cable and recommended distances between nodes.

GROUP TASK Discussion
It is likely that your school is cabled with UTP. Determine the location of any patch panels and also determine the category of cable used.
Discuss likely reasons for the location of the patch panels.

- **Coaxial Cable**

Coaxial cable was originally designed to transmit analog broadcast TV from antennas to television sets. As analog TV stations transmit on frequencies ranging from 30MHz to 3GHz (VHF and UHF bands) the cable also needed to support these high frequencies. Furthermore coaxial cable is relatively immune to outside electromagnetic interference compared to twisted pair.

When computer networks emerged coaxial cable was the natural choice. Early Ethernet standards and also IBM's token ring standards used coaxial cable borrowed from the TV and radio industries. For example 10base5 (Thicknet) and 10base2 (Thinnet) Ethernet both used coaxial cable over a logical bus topology. Compared to UTP, coaxial cable is expensive and furthermore it takes more space and is less flexible. As a consequence coaxial cable is seldom used when cabling new baseband LANs.

Coaxial cable is well suited to broadband applications. Today coaxial cable is used extensively for cable TV where a single cable also carries broadband Internet signals. On cable TV networks each TV station uses a bandwidth of 6MHz. The broadband signal occupies a similar bandwidth and is shared between many users.

The structure of a typical coaxial cable is shown in *Fig 3.72*. Originally all coaxial cables contained a solid copper core, today the core is often steel that is clad with copper. A nylon insulator surrounds the solid core. The insulator is then enclosed within an aluminium foil wrap that is in turn wrapped with braided copper or aluminium. A black plastic sheath covers the entire cable.

Fig 3.72
Coaxial cable.

- **Optic Fibre Cable**

Optic fibre cable is able to support far higher data transfer rates over much greater distances than either twisted pair or coaxial cable. In theory, over 50 billion telephone conversations can be sent down a single hair thin optical fibre! Furthermore optical fibre is completely immune to outside electrical interference. It is therefore not surprising to learn that the majority of major communication links connecting major cities and continents use optical fibre. This includes land based connections and also undersea (submarine) cables connecting continents.

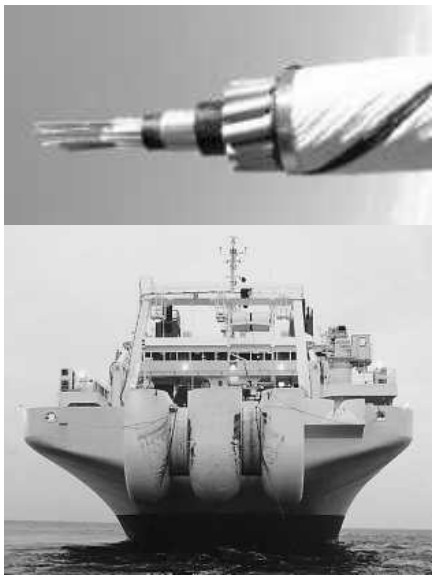


Fig 3.73
Submarine optical fibre cable and purpose-built undersea cabling ship.

Detail of an undersea fibre optic cable together with a purpose built ship are shown in *Fig 3.73*. The cable includes many optical fibres (hundreds in some cables) surrounded by numerous protective coverings including a solid copper sheath, steel cables and many other composite layers. Purpose built ships lay these cables. In shallow water the cable is buried up to 3 metres deep to protect against damage from fishing trawlers, in deeper water the cable is laid directly onto the seabed. Due to impurities in the optical fibres repeaters are installed every 100km or so to amplify the signal.

When making overseas telephone calls or accessing overseas websites the signal is most likely travelling through one of these optical submarine cables. There are numerous optical undersea cables connecting all continents apart from Antarctica. Currently many of Australia's connections originate on the West Coast of the USA and come into Sydney through the Hawaiian Islands, Fiji and New Zealand. Other cables come into Western Australia from Singapore via Jakarta.

Optical fibre is often used for dedicated backbones that connect UTP based networks into a single LAN. Fibre can be utilised as the sole transmission media on LANs, however due to the extra cost involved this is unusual apart from some specialised applications. Industrial applications are one example where complete networks use fibre due to the high levels of electromagnetic interference created by machinery that would cause havoc with UTP or coaxial cables. Most modern aircraft are cabled with optical fibre because of its immunity to interference and also because of its lighter weight. Fibre is used almost exclusively for military networks that carry sensitive information due to the difficulty of tapping optical lines. It is virtually impossible to tap into an optical cable without disrupting the signals.

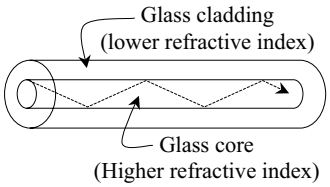


Fig 3.74
Detail of an optical fibre.

A fibre optic cable is composed of one or more optical fibres where each fibre forms a waveguide for containing light waves. The light reflects off the inside of the cladding that surrounds the core (see *Fig 3.74*). Both the core and the cladding are primarily made of pure glass. The cladding has a lower refraction index than the core. As a result light is reflected such that it remains almost totally within the core. The small amount of light that escapes the core is due to impurities in the fibre manufacturing process and is the main reason for current distance limitations. Each fibre's core diameter is usually between 9 and 100 micrometres (millionths of a metre) and the cladding diameter between 125 and 140 micrometres – the diameter of a human hair is around 50 micrometres.

Light waves are really extremely high frequency electromagnetic waves. The light waves used to carry signals within optical fibres reside within the infrared region of the electromagnetic spectrum – just below visible light. Optical fibres are designed to carry specific frequencies or wavelengths of infrared light. Currently fibres designed for wavelengths of 0.85, 1.55 and 1.625 micrometres are common. This equates to frequencies of around 200,000GHz to 350,000GHz. Fibres designed for specific frequencies are known as single-mode fibres. Multi-mode fibre is also available where the refractive index of the cladding varies throughout its diameter to support a range of infrared frequencies. Multi-mode fibre operates reliably over much short distances than single mode fibre.

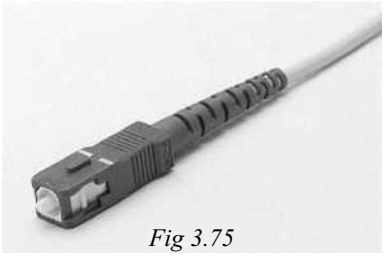


Fig 3.75
SC Connector.

For LAN applications each optical fibre is contained within a protective plastic coating much like that used to protect coaxial cable. This cover is to protect against physical damage and to add strength. The final cable (which may contain a number of optical fibres) is enclosed within a further plastic sheath. It is critical that fibre connections accurately align the optical fibres together. For high-speed links the ends of the fibres are fused together, for LAN applications various types of connectors are used that accurately align the fibres. *Fig 3.75* shows an SC connector commonly used to connect fibre-based Ethernet LANs. The Ethernet 1000baseSX standard specifies multimode fibre over cable runs up to 220m whilst the single mode 1000baseCX standard specifies cable runs up to 2km. In reality much greater distances are possible – up to 30km is not unusual for 1000baseCX connections.

Optical fibre has the potential to support a much larger bandwidth than is possible with copper-based alternatives. When new Ethernet standards are released it is usual for the fibre optic version to be released before the corresponding UTP standard. In terms of data transfer speeds an optical fibre is loafing along at gigabit speeds whilst such speeds are stretching the capabilities of UTP.

GROUP TASK Research
Submarine communication cables have been linking continents since the later 1800s. Research the history and development of submarine cables.

GROUP TASK Discussion
Develop a table of advantages and disadvantages of fibre optic cable compared with UTP and coaxial cables.

Wireless Transmission Media

Wireless or unbounded transmission uses the atmosphere as the medium to carry electromagnetic waves between nodes. Examples of unbounded media include point-to-point terrestrial (ground-based) microwave, satellites, wireless networks such as 802.11, Bluetooth networks, infrared and of course mobile phones. In this section we examine each of these uses of wireless media. Wireless media has distinct advantages over wired media as it can traverse rugged terrain and it allows nodes to move freely about within the coverage area. Unfortunately due to the unbounded nature of wireless media it is particularly susceptible to interference from other sources, which makes it largely unsuitable for critical high-speed connections.

The frequency range used for wireless transmission is from about 10KHz up to 30GHz – just above audio sound and below infrared light in the electromagnetic spectrum. This frequency range is often referred to as RF (Radio Frequency) as currently all wireless signals transmit within the RF range – with the exception of infrared devices, which use frequencies at the lower end of the infrared spectrum. The RF range includes AM and FM radio, analog and digital TV and also each of the unbounded media mentioned above. For example the radio station MIX 106.5 FM transmits its signal by frequency modulating a 106.5MHz carrier signal.

Microwaves occupy frequencies between about 1GHz and 3000GHz. However for most wireless applications frequencies from 1GHz to 30GHz are used, within this bandwidth wavelengths vary in length from 10mm up to 300mm. Due to these relatively short wavelengths microwaves behave somewhat like light. They naturally travel in straight lines and can easily be disturbed by solid objects in their path. However they travel relatively well through the atmosphere. These properties make higher frequency microwaves (those closer to 30GHz) suitable for point-to-point applications including high capacity ground-based microwave and satellite where the waves are aimed precisely from a single transmitter to a single receiver. Microwaves in the middle of the range (closer to 15GHz) are commonly used for satellite to multiple ground applications such as satellite TV. Foxtel currently transmits its digital TV from the Optus C1 satellite at frequencies of around 12.4GHz.

RF waves at lower microwave frequencies are better suited to local broad coverage applications such as mobile phone and WLAN networks. At lower frequencies the waves are better able to penetrate local structures such as buildings. Mobile phones use frequencies of around 1GHz to 2GHz, and Bluetooth and current 802.11g WLANs use frequencies around 2.4GHz. 2.4GHz is within the unlicensed part of the spectrum. For these lower frequency applications the power level of the transmitters can be adjusted to alter the radius of the effective coverage area. To maximise coverage Global Positioning Satellites (GPS) transmit in the range 1GHz to 1.5GHz.

To reduce interference, particular frequency ranges are legally specified for different applications. The Australian Communication Authority (ACA) specifies and enforces how different frequency ranges can be used in Australia. The International Telecommunications Union (ITU) allocates frequencies internationally.

GROUP TASK Discussion
Compare and contrast broadcast communication used for radio and television signals with the communication signals required for telephone and Internet access.

- **Point-to-Point Terrestrial Microwave**

Point-to-point ground based (terrestrial) microwave is used to relay wireless signals across large distances. A direct and uninterrupted line of sight between the transmitter and the receiver is required. Generally sequences of transmitter/receivers, known as transponders, are arranged into a chain. Each transponder receives the signal, amplifies it and transmits it precisely to the next transponder. Distance between transponders varies considerably depending on the terrain, however generally transponders are around 40km apart.

Transponders must be physically located high above the local ground level to avoid trees, buildings and other large obstacles and also to counteract the curvature of the Earth. Microwave transponders are installed on purpose built communication towers. Larger towers can be seen on hilltops (see *Fig 3.76*) and smaller versions on top of large city buildings. Today it is common for these same towers to be shared with mobile phone base station transmitters.

Fig 3.76
Communication tower and microwave transponders.

The use of terrestrial microwave transmission commenced during the 1950s and was commonplace during the 1980s. It was used to relay radio and TV programs between different radio and TV stations and also to relay telephone signals across vast distances. Today optical fibre is replacing many voice and data terrestrial microwave systems with satellite replacing many broadcast radio and TV applications.

- **Satellite**

Satellites use microwaves to carry digital signals from and to both ground based stations and also between satellites. Satellites contain transponders that receive microwaves on one frequency, amplify and then transmit microwaves on a different frequency. A typical communications satellite (see *Fig 3.77*) contains hundreds or even thousands of transponders.

Communication satellites are usually geostationary. This means they remain over the same spot on the Earth at all times. All geostationary satellites are directly above the equator at a height of approximately 35500km. Therefore Earth-based satellite dishes in Australia (southern hemisphere)

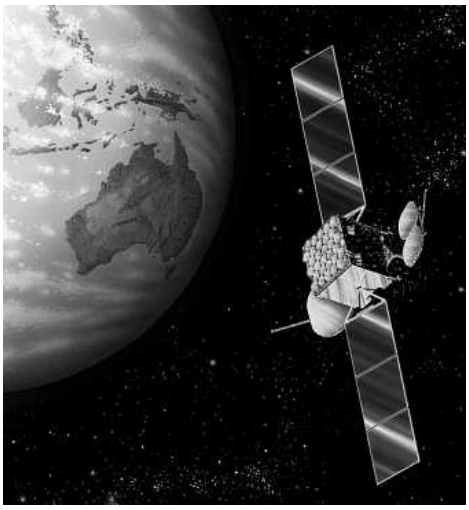


Fig 3.77
Geostationary satellites orbit above the equator at a height of 35,500km.

always face in a northerly direction. In the northern hemisphere such dishes face in a southerly direction. Geostationary satellites are used for satellite TV and also for broadband Internet connections. Satellite is well suited to TV broadcasts however for Internet connections satellite is not the first choice. The time taken for the signal to travel to and from the satellite is in the order of 300 or more milliseconds. For TCP connections this is a significant amount of time and hence satellite Internet is only used in remote locations where land-based ADSL or cable is not available. Cheaper Internet satellite systems use a dial-up link for uploads, as satellite transmitters for two-way satellite systems are expensive. Older style satellite telephones are available that communicate with geostationary satellites. Like satellite Internet, there is a noticeable lag in conversations and hence they are used primarily for emergency land and marine applications.

GROUP TASK Discussion

Even if you live on the equator the round trip to and from a satellite is more than 70,000km. The distance from Sydney to New York is around 16,000km.

Compare satellite and land-based transmission times for an IP datagram travelling between Sydney and New York.

The Global Positioning Satellite (GPS) system currently uses a network of more than 24 satellites that orbit the globe in different directions to form a complete grid (see *Fig 3.78*). Each satellite is continually transmitting a signal from 20,000km above the Earth that includes the satellite's current position and the time the signal was transmitted. Receivers on the ground, such as car and hand-held navigators, receive the signal from multiple GPS satellites within range.

Fig 3.78
The GPS system uses a network of more than 24 satellites.

To pinpoint any position on the globe requires signals from at least 3 satellites, however it is common for up to 8 satellites to be within range at any time. A triangulation system is used to determine the current location of the receiver. The receiving GPS device calculates the time taken for each signal to reach the device. As the signals travel at a constant speed (close to light speed) the distance between each satellite and the receiving device can be calculated. The position of the satellite is known hence a series of spheres can be constructed around each satellite's known position. The point where the spheres intersect on the Earth's surface is the receiver's current position. Most GPS devices are able to plot this position graphically on a map in real time and provide directions both graphically and using synthesised voice (see *Fig 3.79*).

Fig 3.79
TomTom® GPS navigator.

GROUP TASK Research

Research, using the Internet or otherwise, a list of applications where the GPS system is used. Include personal, business, aeronautical, marine and military uses for the GPS system.

Low Earth Orbit Satellites (LEOS) are used for various applications, including mapping and weather forecasting. These satellites travel at high-speed at heights ranging from about 500 to 2000km above the Earth's surface. A typical LEOS orbits the globe about every 1 to 2 hours. Individual satellites are unable to provide uninterrupted coverage at any single position on the globe. Because of the significantly shorter distances from the surface to low Earth satellites they may well have a future in terms of data communication. There are currently (2007) two failed networks of low Earth satellites in operation – Iridium and Globalstar. Both these networks where originally created to provide global mobile phone and data communication services.

GROUP TASK Discussion
Research and discuss reasons for the apparent failure of Iridium and Globalstar. By the time you read this, perhaps these or similar LEOS networks have become economically viable. Research and discuss.

- **Wireless LANs (WLANs)**

We have already discussed much of the detail of current 802.11 series WLANs earlier at part of our discussion of physical star and logical bus topologies. Furthermore we will discuss some of the hardware devices used by WLANs in the next section. In this section we restrict our discussion to frequencies and how they are assigned within 802.11g WLANS – the current 2007 standard.

802.11g WLANs communicate using microwaves with frequencies in the vicinity of 2.4GHz. Currently the range of frequencies around 2.4GHz is unlicensed, which means manufacturers are free to use such frequencies for any purpose they desire. Common applications include cordless phones, Bluetooth devices, remote control toys and even microwave ovens. Such devices can and do influence the performance of 802.11g WLANs. Fortunately the much more powerful waves generated by microwave ovens are largely shielded so in most cases they have little effect. If microwaves were to escape from an oven their high power would effectively drown out any lower powered WLAN signals. Lower powered devices can also cause problems, however such problems usually result in lower data transfer speeds rather than complete loss of WLAN connections.

Each 802.11g WLAN transmits and receives at a maximum speed of 54Mbps on a channel that has a bandwidth of approximately 20MHz. There are 14 possible channels and each channel is assigned a central carrier frequency that is 5MHz from adjoining frequencies. This means that adjoining channel frequencies overlap significantly. It is wise to consider the channels used by adjoining WLANs when interference or poor data transfer speeds are experienced.

Consider the following:

Each of the following industries has strongly embraced WLAN technologies:

- Health care, in particular hospital wards.
- Retail, in particular stock control.
- Education, in particular Universities.

GROUP TASK Discussion
Discuss advantages of WLAN technology for each of the above industries that is likely to have led to its widespread use.

GROUP TASK Discussion
The WiMAX or IEEE 802.16e was released in 2006. Research and describe the essential features and capabilities of WiMAX technology.

- **Bluetooth**

Bluetooth is a communication system for short-range transmission; it was designed to replace the cables that connect portable devices. Bluetooth operates within the unlicensed 2.4GHz part of the spectrum. Many portable and other devices include support for Bluetooth, for example, mobile phones, PDAs (portable Digital Assistants), car and home audio systems, MP3 and MP4 players, laptop computers gaming consoles and numerous other devices. Specialised devices that use Bluetooth are beginning to emerge, for instance the electric motor in *Fig 3.80* is controlled via a Bluetooth connection. Bluetooth devices automatically recognise each other and form an "ad-hoc" network known as a piconet. Up to seven devices can join each piconet, and each device can simultaneously connect to multiple piconets. For instance, a Bluetooth headset can form a piconet with a mobile phone, whilst the mobile phone is transferring data to a laptop over another piconet.

Fig 3.80
Bluetooth electric motor.

All nodes connected to a piconet share a single communication channel. This channel is split into equally spaced time slots. Data packets are placed into one of these slots during transmission. One Bluetooth device is designated as the master and the others are known as slaves – slaves can only communicate directly with the master. The master controls and manages the network. The master alters the frequency used by the channel at regular intervals to avoid interference from other devices and piconets that may be operating close by. The system clock within the master device determines when the frequency is altered and is also used to synchronise the transmission of packets between nodes. Using a single clock for synchronisation is possible because Bluetooth operates over short distances.

The physical distance between Bluetooth devices depends on the power of the transmitter in each device; low power devices must be less than a metre apart whilst around 100 metres is possible with higher powered transmitters. Bluetooth generally supports data transfer speeds of up to 1Mbps, however 3Mbps is possible using Bluetooth's EDR (Enhanced Data Rate) mode.

Bluetooth packets include different error checks depending on the connection being used – some types use a CRC calculated over the entire packet whilst others include error checks over just the packet's header data. The different connection types are designed to efficiently transfer data with different characteristics. For example, some devices, such as remote controls, send very short messages at random times; for these devices an asynchronous connection type is appropriate – in this Bluetooth context asynchronous refers to the random nature of the connection. However, during a phone call the transfer between headset and phone is time sensitive and continuous; hence an isochronous connection is appropriate. The master creates an isochronous connection by reserving a regular number of time slots for the sole use of the headset and phone.

GROUP TASK Activity
Examine the Bluetooth settings present within a device. Explain how each setting affects the operation of the device.

- **Infrared**

Infrared waves occur above microwaves and below visible light. For communication systems, frequencies just above microwaves are used. Infrared waves travel in straight lines hence a direct line of sight is required between source and destination. Currently infrared is only used over short distances. Common applications include remote controls used within many consumer products and for transferring data between a variety of portable devices and computers. The IrDA (Infrared Data Association) maintains a set of IrDA standards. In general, these standards provide a simple and relatively inexpensive means for transferring data between two devices.

GROUP TASK Activity
Create a list of all the devices within your home and school that use infrared communication.

- **Mobile Phones**

In most other countries mobile phones are known as cell phones. This is because mobile phone networks are split into areas known as cells. Each cell contains its own central base station that transmits and receives data to and from individual mobile phones. Each base station is connected to the PSTN (and Internet) using either a cabled link or via a microwave relay link. As users roam from one cell to another the current base station passes the call onto the next base station. Mobile phones automatically adjust the power output by their transmitters based on the signal level received from their current base station – this reduces electromagnetic radiation and also extends battery life.

Fig 3.81
Mobile phone base station.

Both GSM and CDMA digital phone networks are available in Australia. These networks are known as second generation (2G) networks, where first generation refers to the older obsolete analog mobile network. Third generation (3G) networks in Australia are based on UMTS technology. 3G networks combine voice and data at broadband like speeds.

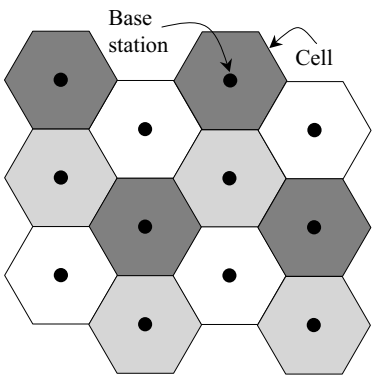


Fig 3.82
Mobile phone networks are composed of cells surrounding each base station.

GSM (Global System for Mobile communication) networks are currently the most popular mobile phone networks in Australia. In GSM networks adjoining cells transmit and receive on different frequencies. At least three different frequency bands are required to avoid overlap between adjoining cells. Each GSM cell supports an equal number of users. In areas of high usage the number of cells is increased and the effective coverage area of each cell is reduced. In large cities and within shopping malls some cells cover areas of just a few hundred metres.

The CDMA (Code Division Multiple Access) network is currently popular in rural areas because of its greater range. CDMA cells all use the same frequencies for all calls and each call is assigned a unique call ID. Calls from many users are multiplexed together. When a user moves from one cell to another it is the call ID that is used as the basis for handing the call to the new base station.

In Australia 3G networks use the Universal Mobile Telecommunication System (UMTS). These 3G networks currently provide wireless connections that transmit and receive voice, video and data at speeds up to 3Mbps. Telstra's 3G network is known as NextG and is used for both mobile phone and mobile Internet connections.

GROUP TASK Discussion
"The distinction between phone and Internet networks is steadily diminishing." Discuss.

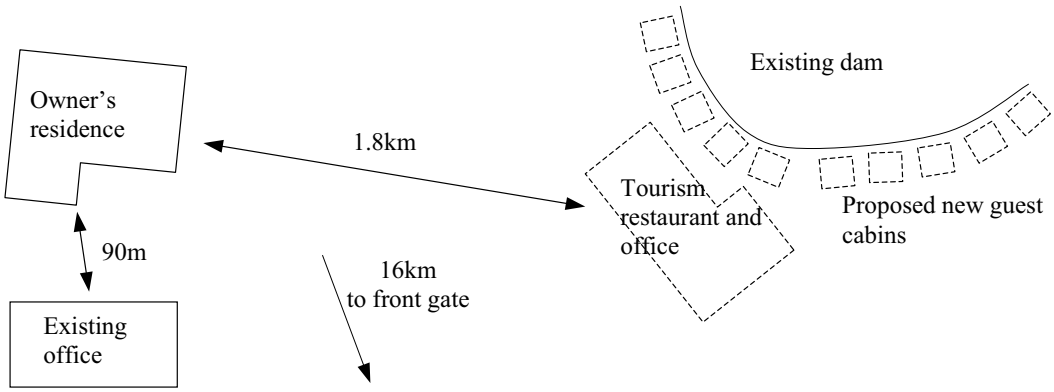
HSC style question:

A large cattle station in a remote area of far north Queensland wishes to update its current information technology to improve both internal and external communication. The cattle station is within a tropical area, hence during the wet season large electrical storms occur almost every day. The cattle station's main income is predictably from cattle sales, however a new tourism venture is growing rapidly.

Currently the cattle station has an office complex where 10 employees share 5 stand-alone computers. The computers are only a few months old and each is connected to its own printer. A computer in the owner's residence has an Internet connection via a standard telephone line. There are three other telephone lines entering the property, currently two are used for voice, and the other for fax.

The owner of the cattle station has created the following 'technology wish list' and sketch of the buildings and distances involved.

- Each office employee is to have their own computer.
- All computers able to share files and access the cattle station's database.
- All computers to have fast Internet access.
- A new website together with an onsite web server.
- Provision for additional Internet connections in each of the 10 new guest cabins.
- A computer in the new tourism restaurant and office that is able to access the main cattle station database.



(a) After researching various high-speed Internet possibilities, it is found that cable, DSL and two-way satellite links will not be available within the foreseeable future. The only available option is to install a one-way satellite link.

Discuss restrictions the use of a one-way satellite link will place on the owner's 'technology wish list'.

(b) Recommend suitable transmission media for each internal network link. Justify each of your recommendations.

Suggested Solution

(a) During electrical storms the satellite link is likely to suffer or not operate at all. Hence the down stream link from the satellite will be lost, in effect all Internet access will be lost. Perhaps one or two dial up downstream links should be maintained so that at least access can continue albeit at slower speeds. Given the number of computers using this link Internet performance would be unacceptably slow.

Although data transfer speeds from satellites are comparable to other broadband connections, the actual time taken to transfer individual IP datagrams is significantly slower. This is due to the distance the data must travel – 35,500km up to the satellite and then 35,500km back down to the cattle station. In this case the extra time is unavoidable as no other suitable option is available, however it does limit the requirement for fast Internet access.

Furthermore as only a fast downstream link is present then having an onsite web server is really out of the question. The upstream link from the web server would be restricted to dial-up modem speeds, which is unsatisfactory. The web server should be attached to fast links both up and down stream, which means it should probably be hosted elsewhere by a suitable ISP.

(b) Fibre optic cable between existing office and tourism facilities. 1.8km is too far for twisted pair (without repeaters) and furthermore the bandwidth required to service 11 computers is more reliably provided using optical fibre. Optical fibre being immune to most forms of interference.

Twisted pair (UTP) within the existing office and to the owner's residence (satellite installed on existing office). Distances between computers within the office are small and the 90m run to the residence is just within the limits of twisted pair. The line to the residence is not critical as it connects to a single node. Twisted pair connected to a switch (or hub) means if a single line is compromised only one node is lost.

Twisted pair running from tourism office to each guest cabin. The distances are small and although the cable would run outside the guest connections are not critical. The node in the tourism office connects to the tourism switch, which in turn is connected to the fibre optic cable, hence loss of connectivity to the tourism office machine is unlikely.

Comments

- Wireless connections using one or more access points could be used to connect the tourism office to the guest cabins. Similarly a wireless link is possible between the existing office and residence.
- UTP would be preferred over wireless for cabling the existing office and tourism office computer. These links being more critical than the guest links and UTP will be less likely to fail during tropical storms.
- Note that guests who are used to broadband speeds are likely to be disappointed with the performance of the one-way satellite link.
- It is likely that part (a) would be worth 3 marks and part (b) would attract 4 to 5 marks in a trial or HSC examination.

SET 3G

1. Most submarine cables used for data are:
   (A) fibre optic cable.
   (B) coaxial cable.
   (C) STP cable.
   (D) UTP cable.
2. Which of the following best describes the difference between analog and digital signals?
   (A) Analog signal – some points on the analog wave are significant. Digital signal – all points on the analog wave are significant.
   (B) Analog signal – all points on the analog wave are significant. Digital signal – some points on the digital wave are significant.
   (C) Analog signal – all points on the analog wave are significant. Digital signal – some points on the analog wave are significant.
   (D) Analog signal – some points on the analog wave are significant. Digital signal – all points on the digital wave are significant.
3. Digital data is encoded as a digital signal using which process?
   (A) modulation or voltage changes.
   (B) demodulation or high/low voltages.
   (C) DAC
   (D) ADC
4. A popular amplitude and phase modulation scheme is:
   (A) SONET
   (B) PSTN
   (C) ADC
   (D) QAM
5. Analog music is encoded on audio CDs using:
   (A) QAM
   (B) DAC
   (C) PCM
   (D) PSTN
6. Analog to digital converters:
   (A) encode the entire wave digitally.
   (B) represent data more accurately because they convert it to digital.
   (C) are used during demodulation of all digital signals.
   (D) sample the wave at regular intervals.
7. When transmitting and receiving, which of the following is TRUE?
   (A) Transmitting decodes, receiving encodes.
   (B) Transmitting encodes, receiving decodes.
   (C) Both transmitting and receiving encode.
   (D) Both transmitting and receiving decode.
8. The twists in UTP cable are designed to:
   (A) prevent all outside electromagnetic interference.
   (B) reduce interference between pairs.
   (C) ensure installers can locate each pair within the cable.
   (D) All of the above.
9. Which best describes the transmission of light through an optical fibre?
   (A) Light reflects off the metallic coating as it moves through the glass fibre.
   (B) The light travels down the centre of the fibre without reflection.
   (C) The light is turned on and off to represent ones and zeros.
   (D) The light reflects off the glass cladding as it moves through the glass core.
10. Which of the following is TRUE of satellites in the GPS system?
    (A) They transmit time and position data.
    (B) They transmit and receive time and position data.
    (C) They receive time and position data.
    (D) They transmit directions to a given location.

11. Define each of the following terms.
    (a) Encoding
    (b) Decoding
    (c) microwave
    (d) infrared
    (e) Analog signal
    (f) Digital signal
12. Describe the nature of the signals used in each of the following.
    (a) A speaker wire
    (b) A 100BaseT Ethernet cable
    (c) The phone cable between a DSL modem and the local telephone exchange.
13. Explain how Bluetooth devices transfer data.
14. Identify strengths and weaknesses and provide examples of where each of the following transmission media is used.
    (a) UTP cable
    (b) Coaxial cable
    (c) Fibre optic cable
15. Explain the operation and uses for each of the following examples of wireless communication.
    (a) Point-to-point terrestrial microwave
    (b) Satellite
    (c) Wireless LANs
    (d) Mobile phone networks

NETWORK CONNECTION DEVICES

In this section we examine devices used to connect nodes to form a LAN and also to transfer data between networks. Each node requires a network interface card that complies with the Transmission Level protocols used by the network. For most LANs a physical star topology is used hence a central node in the form of a hub, switch or wireless access point is required. Gateways connect networks that use different Transmission Level protocols whilst bridges connect networks using the same low-level protocols. Modems allow LANs to communicate with WANs. Routers operate at the Communication Control and Addressing Level to direct data along the most efficient path. For small LANs the functions of many of these devices is combined within a single hardware device generically known as a router.

- **Network Interface Card (NIC)**

Network interface cards convert data between the computer (commonly the PCI bus) into a form suitable for transmission across the network. The conversion uses the rules of the data link and physical link protocols in operation. It is the NIC that negotiates access to the network, including collision detection (or avoidance). Each NIC has its own unique MAC address so that other low-level network devices can uniquely identify the node.

In the past most network interface cards were indeed cards that plugged into the motherboard. Today most computers include the functionality of an Ethernet NIC into the motherboard. An RJ45 port is included for connecting standard UTP patch cables. In addition most laptop computers include built in support for wireless LANs. Wireless NICs that connect via a USB or PCMCIA port are often used when the computer does not have an embedded wireless NIC. NICs for optical fibre networks are usually separate cards that install into a free slot on the PCI bus.

Fig 3.83
Wireless NICs for PCI (top), USB (middle) and PCMCIA (bottom).

- **Repeater**

A repeater is any device that receives a signal, amplifies it and then transmits the amplified signal down another link. Repeaters are used to increase the physical range of the transmission media. Dedicated repeaters are routinely used to extend the reach of fibre optic cable. Most wireless access points can be used as simple repeaters to extend the coverage range of WLANs. Transponders used for ground-based and satellite microwave transmissions are also repeaters.

- **Hub**

When a hub receives a packet of data it simply amplifies and retransmits the packet to all attached nodes. As a consequence hubs are also known as multi-port repeaters. Hubs are dumb devices that operate at the physical layer of the OSI model. They make no attempt to identify the destination node for each message.

Hubs connect nodes together into a single network segment. This means all nodes attached to a central hub are sharing the same transmission channel meaning a logical bus topology is being

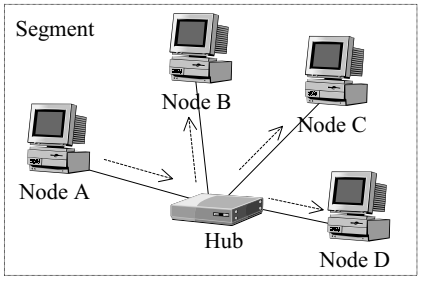


Fig 3.84
Hubs repeat all messages to all nodes on a single LAN segment.

used. Hubs were once the primary devices used to connect UTP Ethernet networks. Today hubs have been largely phased out in favour of more intelligent switches.

- **Bridge**

A bridge separates a network into different segments at the data link layer. Bridges were once used extensively to segment Ethernet logical bus networks – today switches perform this function. Bridges determine the destination MAC address of each frame. If the destination node with that MAC address is on the other side of the bridge then the frame is repeated onto that segment, otherwise the frame is dropped. Essentially a bridge splits a logical bus network into two collision domains.

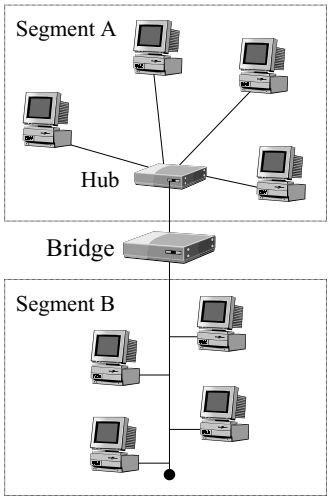


Fig 3.85
Bridges separate networks into separate segments or collision domains..

- **Switch**

A switch can be thought of as an intelligent hub or a multi-port bridge. Switches determine the MAC address of the sender and intended receiver that precedes each message. The receiver's address is used to identify the destination node and forward the message to that node only. In essence, a switch sets up a direct connection between the sender and the receiver; therefore each node exists on its own segment, the switch being the only other device on the segment. As no other nodes exist on each segment each node is free to transmit messages at any time without the need to detect or avoid collisions.

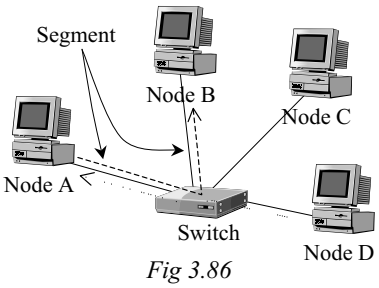


Fig 3.86
Switches forward messages to the destination node only. Each switch – node connection forms a segment.

Switches are able to simultaneously receive and forward messages from and to multiple pairs of nodes. As long as both the sender and the receiver of each message do not conflict with other simultaneous messages then the switch will direct the message correctly. Most switches allow nodes to communicate in full duplex. In *Fig 3.86*, Node A is sending a message to Node B whilst it simultaneously receives a message from Node D, neither message is ever present on Node C's segment. Switches significantly reduce the amount of traffic flowing over each cable resulting in vastly improved data transfer speeds compared to speeds achieved using hubs.

- **Gateway**

A gateway connects two networks together. Gateways can connect networks that use different lower level protocols, however they can also be used to filter traffic movements between two similar networks. Gateways are routinely used to connect a LAN to the Internet, however they can be used to connect any two networks. For example ADSL and cable modems (often called routers) include gateway functionality to convert between the low level Ethernet protocol used by the LAN and the low level protocols used by ADSL and cable connections. Larger LANs often include proxy servers whose task can include gateway functionality as they convert and filter traffic flowing between the LAN and the Internet.

Gateways that connect IP LANs to the Internet have two IP addresses. A local address used for communication within the LAN and an Internet IP address used on the WAN or Internet side of the gateway. The local LAN IP address is used as the default

gateway address for all local nodes wishing to access the Internet. The gateway hides the local IP addresses from the Internet, instead IP datagrams are all sent using the gateway's WAN or Internet IP address. The gateway keeps track of the local IP addresses so that IP traffic from the Internet can be directed to the correct local node.

If a LAN includes a gateway that provides a connection to the Internet then the gateway's LAN IP address must be known to all nodes – in most operating systems this IP address is specified as the default gateway – in *Fig 3.87* 10.0.0.138 is the local IP address of the ADSL router that links to the Internet.

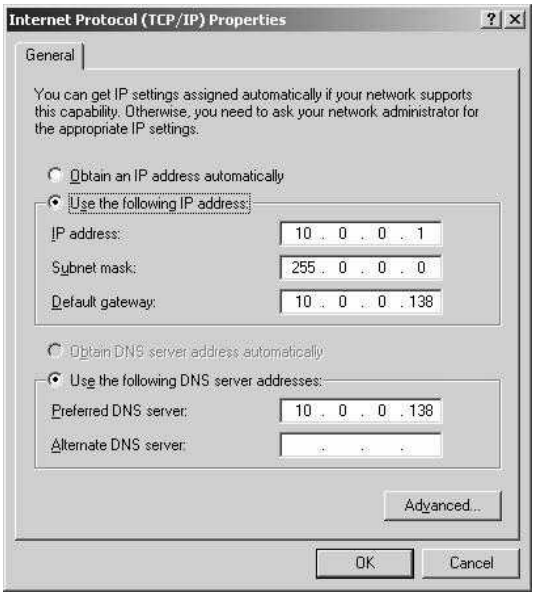


Fig 3.87
The default gateway setting specifies the node acting as the gateway to the Internet.

Like many technology related terms the meaning of the word "gateway" is used differently in different contexts. In general usage the word "gateway" is used to refer to devices that connect a LAN directly to the Internet. However, routers commonly include one or more gateways. As a consequence the general public often use the words router and gateway interchangeably.

- **Wireless Access Point**

Wireless access points (WAPs) or simply access points (APs) are the central nodes on wireless LANs. Access points broadcast to all wireless nodes within the coverage area. On 802.11 WLANs the access point does not direct packets to specific nodes or control the order in which nodes can transmit, rather they simply repeat all packets received. Conceptually an access point performs much like a hub on a wired LAN.

Fig 3.88
Linksys WAP54G wireless access point.

A significant issue with WLANs is security – any user within the coverage range can potentially access the network. To counteract this possibility access points include security in the form of WEP (Wired Equivalent Privacy) and WPA (WiFi Protected Access). WEP uses a single shared key encryption system whilst WPA generates new encryption keys at regular intervals. The WEP system can and has been infiltrated so currently WPA is the recommended system.

No encryption system can work if it is not turned on. This is a major issue for both home and business WLANs. Furthermore the simplicity of creating a WLAN and the ability to access WLANs from outside make security a signifcant issue. Hackers need only to connect a wireless access point to an existing Ethernet connection point and they then have complete access without the need to work around complex firewalls and proxy servers.

- **Modem**

The term modem is a shortened form of the terms modulation and demodulation, these are the primary processes performed by all modems. Today most modems are used to connect a computer to a local Internet Service Provider (ISP); the ISP supplying a high-speed ADSL or cable connection to the Internet. Dial-up modems were once the primary device for connecting users to the Internet. Currently dial-up modems are more often used to send faxes from computers over the PSTN – virtually all dial-up modems are able to both send and receive fax transmissions.

We discussed modulation in some detail earlier in this chapter. Basically modems modulate digital signals by altering the phase, amplitude and/or frequency of electromagnetic waves. That is, modulation is the process of encoding digital data onto an analog waveform. Demodulation is the reverse of the modulation process. Demodulation decodes analog signals back into their original digital form. Clearly both sender and receiver must agree on the method of modulation used if communication is to be successful.

Modulation
The process of encoding digital information onto an analog wave by changing its amplitude, frequency or phase.

Demodulation
The process of decoding a modulated analog wave back into its original digital signal. The opposite of modulation.

Modems are commonly connected to a computer via a USB port or an Ethernet network connection. These interfaces are considered digital links; they do use electromagnetic waves however the data is represented using different voltages. The electronic circuits within the computer can use these voltage changes directly. In contrast modulated analog waves, such as those transmitted down telephone lines or coaxial cables, are not suitable for direct use by the circuits within the computer. Hence the primary role of modems is to provide an interface between the modulated analog waves used for long distance transfer and the digital data suitable for use by computers.

- **ADSL modems**

Asymmetrical digital subscriber lines (ADSL) use existing copper telephone lines to transfer broadband signals. Although these copper wires were originally designed to support voice frequencies from 200 to 3400Hz, they are physically capable of supporting a much wider range of frequencies. It is the various switching and filtering hardware devices within the standard telephone network that prevent the transfer of frequencies above about 3400Hz. To solve this problem requires dedicated hardware to be installed where each copper line enters the local telephone exchange.

ADSL signal strength deteriorates as distances increase, the signal cannot be maintained at all for distances greater than about 5400 metres. Voice lines much greater than 5400 metres are possible using amplifiers. Unfortunately these amplifiers boost only the lower frequencies required for voice, hence ADSL is not currently available in many remote rural areas. Even when distances are short and the copper runs directly into the exchange problems can occur as a consequence of interference. In general phone lines within a building and out to the street are not shielded against interference, this interference is rarely significant enough that a connection cannot be established, however it often reduces the speed of such connections.

So how does ADSL transfer data between an ADSL modem and the local telephone exchange? Using a modulation standard known as Discrete MultiTone (DMT). DMT operates using frequencies from about 8kHz to around 1.5MHz. This bandwidth is split into some 247 individual 4kHz wide channels as shown in *Fig 3.89*. Each channel is modulated using QAM. DMT's task is to specify the channels that are used for actual data transfer. If interference is present on a particular 4kHz channel then DMT will shut down that channel and assign a new channel. This channel switching occurs in real time and is completely transparent to the user. In a sense ADSL is like having 247 dial-up modems all working together, each modem using QAM and DMT ensuring they all work together efficiently. The ADSL modem and the DSL hardware at the telephone exchange communicate to agree on the channels currently being used.

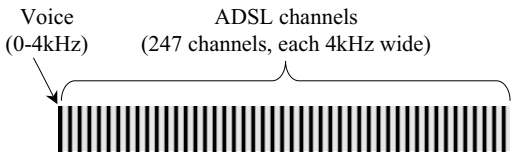


Fig 3.89
ADSL splits higher frequencies into 247 channels, each 4kHz wide.

At the local telephone exchange all the copper wires from the neighbourhood are connected to a splitter (see *Fig 3.90*). This splitter directs the 0-4kHz frequencies to the normal telephone network and the higher ADSL frequencies to a DSL Access Multiplexor (DSLAM). The DSLAM (see *Fig 3.90*) performs all the DMT negotiations with individual ADSL modems and directs data to and from ISPs, where it heads onto the Internet. The term *multiplexor* simply refers to the DSLAM's task of combining multiple signals from customers onto a single line and extracting individual customer signals from this single line.

Fig 3.90
A splitter (left) and DSLAM (right).

In most ADSL systems the lower bandwidth ADSL channels are used for upstream data (from modem to exchange) and higher frequency channels are used for downstream data (exchange to modem). Some channels are able to transfer data in both directions. ADSL is one example of a DSL technology, the A stands for asymmetrical, meaning transmitting and receiving occur at different speeds.

GROUP TASK Research
Research, using the Internet, the upstream and downstream speeds that are achieved using current ADSL connections.

Consider the following

When first installing an ADSL connection it is necessary to install one or more low-pass (LP) filters. Sometimes a single filter is installed where the phone line enters the premises. In this case a qualified technician is required to install a dedicated ADSL line from the LP filter to the location of the ADSL modem. In other cases, the user installs a separate LP filter, like the one shown in *Fig 3.91*, between each telephone and wall socket.

Fig 3.91
Inline LP filter.

GROUP TASK Discussion
What is the function of an LP filter? Describe how the two LP filter installation methods described above achieve the same outcome?

- **Cable modems**

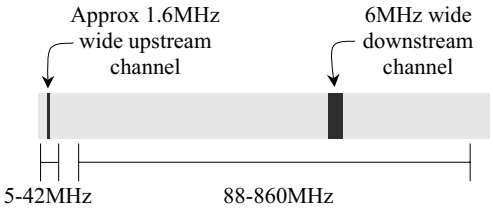


Fig 3.92
Cable modems share a bandwidth of 6MHz downstream and a lower bandwidth upstream.

Cable modems connect to the Internet via coaxial cables; usually the same cable that transmits cable TV stations. *Fig 3.92* describes how the bandwidth within the cable is split into channels. A single 6MHz bandwidth channel is used for downstream data – 6MHz is the width of a single cable TV station. This 6MHz wide channel is assigned within the range 88 to 860 megahertz. A narrower bandwidth channel is used for upstream; commonly 1.6MHz wide however various other bandwidths are supported ranging from 200kHz to 3.2MHz. The upstream channel is assigned within the range 5 to 42 megahertz. The particular frequencies used for both channels are determined by the cable Internet provider and cannot be altered by individual users.

The bandwidth used in a cable system is significantly larger than that used for ADSL. Therefore, one would assume the rate of data transfer would be much larger. In reality cable connections achieve speeds similar to ADSL connections; why is this? Cable connections are shared amongst multiple users. A single 6MHz downstream channel is likely to be shared by hundreds of users. In a sense all the cable modems sharing a particular channel form a local area network. Every cable modem within the network receives all messages; they just ignore messages addressed to other modems. Consequently when only a few users are downloading then higher speeds are possible than when many users are downloading. Clearly the same situation occurs when uploading. This is why cable Internet companies include statements within their conditions stating that speeds quoted are not guaranteed.

Consider the following

Cable modems connect using coaxial cable whilst ADSL systems use standard copper telephone wires. Coaxial cable is shielded to exclude outside interference and also to ensure the integrity of the signal.

GROUP TASK Discussion
ADSL uses DMT and many small bandwidth channels, whilst cable uses QAM and two relatively large bandwidth channels. Discuss reasons for these differences in terms of the transmission media used by each system.

Currently both ADSL and cable Internet providers reduce speeds when an agreed download limit has been exceeded. For cable connections only the upstream speed is reduced whilst both up and downstream speeds are reduced for most ADSL connections.

GROUP TASK Discussion
How can ADSL and cable Internet providers alter speeds? And why don't cable Internet providers reduce downstream speeds? Discuss.

- **Router**

Routers specialise in directing messages over the most efficient path to their destination. Today the large majority of routers operate at the network layer of the OSI model using the IP protocol. Therefore routing decisions are based on each datagram's destination IP address. Routers usually include the functionality of a gateway. They are able to communicate with networks that use different protocols and even completely different methods and media for communication. Many routers also include a variety of different security features. They are able to block messages based on the sender's IP address, block access to specific web sites and even restrict communication to certain high level protocols.

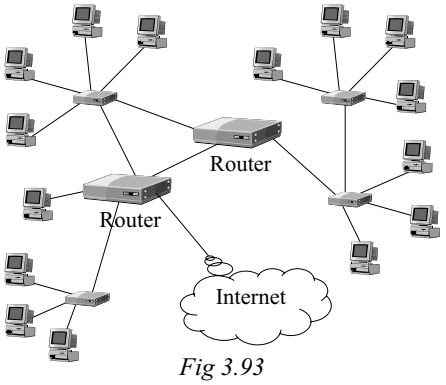


Fig 3.93
Routers forward messages over the most efficient path and can alter this path as needed.

Home or small business routers connect a single LAN to the Internet. For these systems the decision is relatively simple – either the IP datagram is addressed to a local node or it is not. Local datagrams are left alone whilst all others are sent out to the Internet. The routing table maintained by these routers is relatively small and rarely changes. Home and small business routers are commonly integrated devices that commonly include a router, an Ethernet switch and also a wireless access point – these integrated devices are what the general public call routers.

Routers out on the larger Internet connect to many other routers. For these routers deciding on the best path for each IP datagram is considerably more complex. Such routers communicate with other adjoining routers to continually update their internal routing table. The routing table is examined to determine the most efficient route for each IP datagram. However, should any connections within the most efficient path fail then routers automatically direct the message over an alternate path. On larger wide area networks, and in particular the Internet, thousands of routers work together to pass messages to their final destination.

Consider the following

Earlier in this chapter we discussed the operation of the Internet Protocol (IP). During our discussion we learnt that each IP address is composed of a network ID and a host ID. Routers use the network ID as the basis for directing IP datagrams. Network IDs effectively splits the Internet into a hierarchy of sub-networks or subnets. You may have heard the term subnet mask or seen this setting on your own computer. Subnet masks when combined with IP addresses enabled the network ID (and also the host ID) within an IP address to be determined. Routers perform this process on every destination IP address in every datagram to determine the datagrams next hop. The Network IDs and subnet masks are stored in the router's internal routing table.

A routing table is essentially a table that includes records for each Network ID the router knows about. Each record includes a field for the network's IP addresses, the networks subnet mask, the gateway IP address and a metric field. The network IP address and subnet mask are compared with the destination IP address within the current datagram. If the destination IP address is determined to be part of that network then the datagram is sent on the interface with the corresponding gateway IP address.

All routers have multiple IP addresses, one for each gateway. Each gateway provides an interface connecting to another router. The metric field is used to rank records that correspond to the same network ID – higher ranked records being used first.

GROUP TASK Practical Activity
On a Windows machine open a command prompt (type cmd at the run command on the start menu) and type the command ROUTE PRINT. This causes the current routing table to be displayed. Identify each of the fields mentioned above.

GROUP TASK Practical Activity
On a Windows machine open a command prompt (type cmd at the run command on the start menu) and type the command TRACERT followed by a web address, e.g. TRACERT www.microsoft.com. This causes a table showing each hop in a datagrams journey to be displayed. Determine and describe the significance of the fields and records displayed.

SERVERS

Servers provide specific processing services to other nodes (clients). We discussed the general operation of client-server architectures earlier in this chapter. In this section we briefly consider some of the more common services performed by servers. Note that this section is included under the general heading of "Network Hardware"; servers are often distinct computers designed with hardware suited to the services they provide, however what makes them servers is actually the installed software. On large networks dedicated servers are common whilst on smaller networks a server may well perform many tasks including the execution of end-user applications.

Most servers run a network operating system (NOS) to manage user access to the services the server provides. We discuss features of network operating systems in the next section. Most network operating systems include file server and print server functionality as these are the core services that require user authentication and user access rights.

There are numerous different services that servers provide. Examples of servers includes file servers, print servers, database servers, mail servers, web servers and proxy servers. In this section we restrict our discussion to a brief overview of each of these services.

File Servers

A file server manages storage and retrieval of files and also application software in response to client requests. In hardware terms dedicated file servers do not require extremely fast processors, their main requirement being large amounts of fast secondary storage and a sufficiently fast connection to the network.

Fault Tolerance
The ability of a system to continue operating despite the failure of one or more of its components.

Commonly file servers include multiple hard disks connected together into an array – RAID (Redundant Array of Independent Disks). Users are often unaware that multiple disks are being used. RAID uses different combinations of striping and mirroring to both improve data access speeds and also to improve the fault tolerance of the system. Striping stores single files across a number of physical disks and mirroring stores the same data on more than one disk. On larger RAID systems it is possible to replace faulty drives without halting the system – this is

known as hot swapping. To further improve fault tolerance many file servers include various other redundant components including extra power supplies, cooling fans and in some cases the complete server is replicated.

File servers must be able to process multiple file access requests from many users. Consequently the network connection to a file server often operates at a higher speed than for other workstation nodes. For each client request the file server, in combination with the NOS, checks the user's access rights or permissions before retrieving the file. The file server in combination with the NOS ensures the file is retrieved and transmitted according to the user's assigned access rights.

GROUP TASK Discussion
No doubt your school has one or more file servers. Determine the hardware specifications of these machines. Do these machines include any redundant components? Discuss.

Print Servers

A print server controls access to one or more printers for many clients. The print server receives all print requests and places them into an ordered print queue. As the printer completes jobs the next job in the print queue is progressively sent to the printer. Most print servers allow the order or priority of jobs to be changed and they also allow jobs to cancelled. When sharing smaller printers connected directly to a workstation the print server is a software service included within the operating system. In larger networks a dedicated printer server is used.

Dedicated print servers include more advanced functionality. Examples of such functionality includes:

- Ability to prioritise users based on their username. Jobs from higher priority users are placed higher in the print queue.
- Broadcast printing where a single job is printed on many printers.
- Fault tolerance or fall over protection where jobs that fail to print on one printer are automatically directed to some other printer.
- Job balancing where print jobs are spread evenly across many printers.
- Reservation systems where a user can reserve a printer with specific capabilities.
- Ability to reprint documents without the need for the client to resubmit the job. This is particularly useful in commercial environments when a printer jams or has some technical problem.
- Adding banner pages to print jobs. Banners are like cover pages – they commonly include the username, file name and time the job was started. Banners are useful for high volume systems where determining where one job ends and another starts would otherwise be difficult.
- Support for different operating systems and printing protocols. The print server converts client jobs from different operating systems so they will print correctly on a single printer.

GROUP TASK Discussion
No doubt your school has many printers in different locations throughout the school and most users only have access to specific printers. Discuss how printers in your school are shared.

Database Servers

Database servers run database management system (DBMS) software. We discussed the role DBMSs in some detail in chapter 2. Briefly a database server executes SQL statements on behalf of client applications. This can involve retrieving records, performing record updates, deletions and additions. The DBMS provides the connection to the database and ensures the rules defined for the database are maintained. For example ensuring relationships are maintained and performing data validation prior to records being stored.

Mail Servers

We discussed the detailed operation of email earlier in this chapter. Email uses two different application/presentation layer protocols SMTP and either POP or IMAP. These protocols run on SMTP, POP and IMAP servers. It is not unusual for all three protocols to run on a single server machine.

Email client applications, such as Microsoft Outlook, must be able to communicate using these protocols. SMTP (Simple Mail Transfer Protocol) is used to send email messages from an email SMTP client application to an SMTP server. Emails are received by an email client application from a POP (Post Office Protocol) server or IMAP (Internet Message Access Protocol) server.

Web Servers

We discussed the operation of web servers when discussing the HTTP protocol earlier in this chapter. Essentially a web server provides services to web browsers – they retrieve web pages and transmit them back to the requesting client web browser.

Web servers must also include services that allow web pages to be uploaded, edited and deleted. Such services require users to first be authenticated by the web server. Many web servers, particularly those operated by ISPs, host many different web sites. These servers require high speed links to the Internet together with fast access to the files they host.

Proxy Servers

A proxy server sits between clients and real servers. The proxy server tries to perform the request itself without bothering the real server. In essence the proxy server performs requests on behalf of a server. This relieves pressure on the real server and also reduces the amount of data that needs to be transmitted and received. Proxy servers speed up access times when the same request is made by many clients. The proxy server keeps a record of recent requests and responses within its large cache.

Perhaps the most common type of proxy server are those that operate between client browsers and web servers. The proxy server receives all web requests from all clients. If the files are found in the proxy server's cache then there is no need to retrieve it from the original remote web server. Proxy servers that operate between clients and the Internet are also gateways – they provide connectivity between the LAN and the Internet. These proxy servers are also used to censor and filter web content. For example many proxy servers can be set to block access to particular websites or restrict access to particular websites. Most proxy servers can also filter incoming pages to remove pornography and other undesirable content.

GROUP TASK Discussion

It is likely that Internet access at your school is via a proxy server – either within the school or operated by the school system. Determine if this is the case and describe the processes this server performs.

NETWORK SOFTWARE

Network software includes the Network Operating System (NOS) and also network based applications such as those running on the various servers within the network. Most operating systems include network capabilities, however a NOS has many more advanced network management and security features. Network operating systems allow networks to be centrally controlled by network administrators. The ability to centrally control networks improves the security and efficiency of access to the network's various resources. Furthermore it greatly simplifies the tasks performed by network administrators.

In this section we restrict our discussion to an overview of network operating systems and some of the common tasks performed by network administrators.

NETWORK OPERATING SYSTEM (NOS)

Network operating systems operate at the network and above layers of the OSI model. The NOS is installed on one or more servers where it provides various services to secure and support the network's resources and users – one vital NOS service being the authentication of users based on their user names and passwords. Once authenticated the NOS provides the user with access to the network's resources based on their pre-assigned privileges and profiles. Network resources include a variety of hardware and software such as servers, workstations, printers, applications, directories and files. A profile commonly includes details of the desktop configuration, language, colours, fonts, available applications, start menu items and location of user documents. Privileges define the services, directories and files a user (or workstation) can access together with details of how these resources can be used – including file access rights or permissions. Other servers on the network trust the NOS to authenticate users, hence a single login is required.

The NOS allows network administrators to create policies. A policy is used to assign particular resources to groups of users and/or groups of workstations (or clients) with common needs. For example in Windows Server 2003 group policies are created that include profile and privilege details common to groups of users or workstations. Users in a sales department all use similar applications and settings hence the same group policy can be assigned to all users in the sales department. Similarly a group policy can be created for groups of client machines (or workstations), for example workstations in one area may all connect to a particular printer and may connect to the Internet via a particular gateway. Policies greatly simplify the administrative tasks performed by network administrators.

GROUP TASK Research
Using the Internet, or otherwise, find examples of different network operating systems in common use. Research the techniques and tools used to share resources using each of theses NOSs.

NETWORK ADMINISTRATION TASKS

Network administrators are the personnel responsible for the ongoing maintenance of network hardware and software. This includes installation and configuration of switches, routers and other active hardware devices. However on a day-to-day basis network administrators spend much of their time providing support to new and existing users. This includes configuring new workstations (clients) and controlling and monitoring access to network resources as needs change.

Maintaining a LAN is a complex and specialised task performed by professional network administrators. In IPT we can only hope to grasp a general overview of the processes performed by a network administrator. The detail of how each task is accomplished will be different depending on the NOS used. Therefore we restrict our discussion to an overview of some of the more common network administration tasks.

Adding/Removing Users

Each new user has an individual account created that includes their username and password together with details of any assigned policies and privileges. Obviously a user's account is removed or made inactive when a user is removed.

The policies and privileges assigned to a user may be inherited from other existing group policies. Commonly a new user will require access to similar network resources as other groups of existing users, hence the new user is added to one or more existing groups. For example a new salesman requires the same access as existing salespersons. Therefore they are added to the "Sales Group"; as a consequence the new user has access to the same set of network resources as the existing salespersons. When adding a new user they are commonly given a standard password that must be changed when they first log onto the network.

If the network is configured such that users can logon at a number of workstations then their individual profile is configured to be stored on a server. During logon the user is first authenticated and then their individual profile is copied from the server to the local workstation. When they logoff any profile changes, such as desktop settings, are written back to the server.

GROUP TASK Research
Microsoft Windows Server NOSs use domains, domain controllers and active directories. Research and discuss the meaning of these terms and briefly explain the purpose of each.

Assigning printers

Printers can be assigned to specific workstations or to specific users. As printers are physical devices that are installed in specific locations it often makes sense to assign printers to workstations rather than users. This means users will have access to a printer that is physically close to the workstation where they are currently logged on.

Assigning file access rights

File access rights are also known as permissions. On many systems file access rights are a type of privilege. File access rights determine the processes a user can perform on a file or directory at the file level. On most systems the access rights applied to a directory also apply to any files or sub-directories contained within that directory.

Commonly groups of users that perform similar tasks require similar file access rights, which can form part of an assigned group policy. The majority of users will also require full access to a particular directory or folder where their own files and documents are stored.

Typically file access rights are stored by network operating systems within an access control list (ACL). An ACL specifies the user who owns (created) the directory or file, groups who have permissions to access the file and also the access rights assigned to these users. Let us consider typical permissions (access rights) that can be specified for directories (or folders) and also for individual files. The details below relate specifically to systems that use the NT file system (NTFS), which includes all current versions of Microsoft Windows. Other operating systems will have a similar set of permissions.

- **Directory (or folder) Permissions**
  - Full control – Users with full control can change the permissions for the folder, take ownership of the folder and delete any sub-folders and files within the folder. Full control also includes all of the permissions below.
  - Modify – Users can delete the folder and also perform processes permitted by write and read and execute permissions.
  - Read and Execute – Users can navigate through the folder to reach other folders and files. Includes read permission and list folder contents permission.
  - List folder contents – See the names of sub-folders and files within the folder.
  - Read – Users can see the name of sub-folders and files and view who owns the folder. Furthermore users can view all the permissions assigned to the folder but cannot alter these permissions. Users can also view attributes of the folder such as read-only, hidden, archive and system attributes.
  - Write – Users can create new files and sub-folders within the folder. They can change attributes of the folder. Users with write access can view, but not modify folder ownership and permissions.

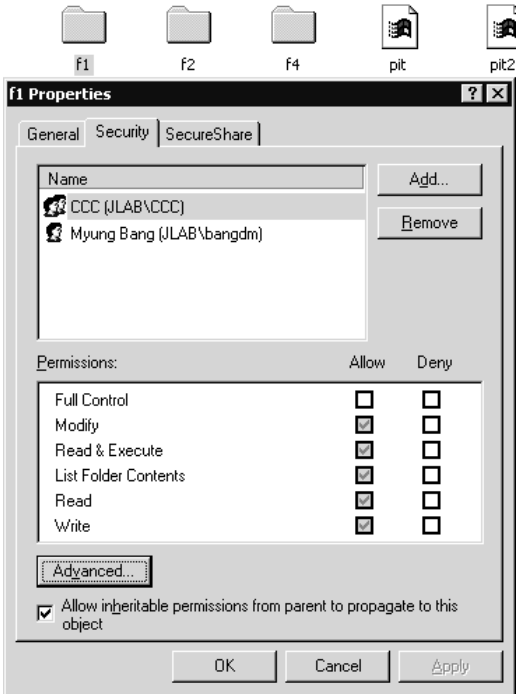


Fig 3.94
Setting NTFS folder permissions.

- **File Permissions**
  - Full control – Users with full control can change the permissions for the file, take ownership of the file. Full control also includes all of the permissions below.
  - Modify – Alter and delete the file and also perform processes permitted by write and read and execute permissions.
  - Read and Execute – Users can run executable files and also read permission processes.
  - Read – Open and display the file. Furthermore users can view all the permissions assigned to the file but cannot alter these permissions. Users can also view attributes of the file.
  - Write – Overwrite the file with a new version. They can change attributes of the file. Users with write access can view, but not modify file ownership and permissions.

GROUP TASK Research
All current operating systems include some form of "file system". Determine the file system used by your school or home computer's operating system. Research available access rights and how they are inherited within this operating system.

Installation of software and sharing with users

Network operating systems are able to automate the installation of software to multiple users. This saves considerable time for network administrators, as they do not need to manually start the installation on numerous client workstations. On large networks where numerous software applications are being used by a wide variety of users in different combinations the automation of software installations is essential.

Software applications can be installed on individual client workstations where they are available for use by any user that logs onto the workstation. In this case the software installs next time the computer starts. This is an appropriate strategy when the software application is widely used – such as a word processor or email client. More specialised applications can be installed for particular users or groups of users. In this case the software installs when the user next logs on.

GROUP TASK Discussion
Think about software applications available for use on your school network. Some are available to all users whilst some are available to just some users. Explain how an upgrade of each of these applications could best be deployed to users.

Client installation and protocol assignment

Every network will have a different specific set of steps for installing new clients. Some require client applications to be installed manually, others automate this process. Some networks require a particular version of the operating system be installed over the network – in these cases it is common for the network settings to also be configured remotely and automatically. Commonly the network administrator or a technician performs these installation steps. Typical steps required to install a new client onto a network include:

1. Ensure the new machine has a compatible NIC (network interface card) installed that supports the data link and physical layer protocols used by the LAN. In most cases new NICs are able to automatically sense the correct speed and protocols being used.
2. Ensure the operating system on the client is compatible with the NOS. Most LANs now use TCP/IP therefore it will be necessary to obtain IP addresses and other parameters needed to configure the connection.
3. Physically connect the NIC to the network using a patch cable. Today this is usually a UTP path cable that connects to an existing network point on the wall. If the point has not been used then the network administrator may need to install a patch cable at the other end to complete the connection from patch panel to switch.
4. The network administrator needs to create the machine within the NOS and assign any profiles – which may include software to be installed. If a new user will use the client then they too will require a user account.
5. After booting the client machine it is necessary to enter a legitimate username and password. A domain or server is also specified. This is used to determine the location of the server used to authenticate the user name and password.

Fig 3.95
Windows Server 2003 logon screen.

HSC style question:

Jack is the network administrator for a company that employs some 50 staff. Each staff member has their own computer connected to the company's LAN. Each staff member has Internet and email access via the company's web and mail servers.

(a) What is a server, and in particular, what are the functions of web and mail servers?

(b) One of Jack's tasks is to assign file access rights to users. What does this task involve? Discuss.

(c) A number of staff are experiencing poor performance when using the LAN. Jack discovers that all these users are directly connected to a single hub and on this hub the data collision light is virtually always on.

Identify the network topology used for this part of the LAN and discuss possible reasons the data collision light is virtually always on.

Suggested Solution

(a) A server is usually a machine on a network that is dedicated to performing a specific task. However what makes these machines servers is the software they execute – hence any machine can be a server. Servers respond to requests from multiple clients. They specialise in performing specific tasks or services.

A web server responds to requests for web pages from clients (usually web browsers). The web server retrieves the requested page and transmits it back to the client (usually over the Internet using HTTP and TCP/IP).

Mail servers store email for each account and are used to set-up these accounts. Mail servers store incoming mail into each user's mail box. The post office protocol (POP) is used by email clients to retrieve mail from mail servers. The Simple Mail Transport Protocol (SMTP) is used to send mail to mail servers and between mail servers. The SMTP mail server checks the email address of all outgoing mail and directs it to the appropriate receiving mail server on the net.

(b) To assign file access rights requires that each user be assigned a user name and password. The user name can be grouped according to access required by different groups of users. Users or groups of users are then given rights to particular directories. These rights could allow them to merely read files or to create, modify and/or delete files within the directories they can access.

(c) As the users are connected to a hub a physical star topology and a logical bus topology is being used. As a consequence all nodes connected to the hub are sharing the same communication channel.

Because collisions are occurring it appears that CSMA/CD is being used. This means that two or more nodes can transmit at the same time resulting in the collisions indicated by the collision light. Reasons for so many collisions include excessive network traffic, which could be caused by a data intensive application, particularly one transferring video, image or audio to many nodes. Perhaps the hub itself is faulty or one node's NIC has a fault such that it is continually trying to send.

SET 3H

1. Which device converts data from a computer into a form suitable for transmission across a LAN?
   (A) NIC
   (B) Repeater
   (C) Switch
   (D) Router
2. Which device extends the range of transmission media?
   (A) Modem
   (B) Repeater
   (C) Bridge
   (D) Gateway
3. Routers direct messages based on which of the following?
   (A) Gateway Addresses
   (B) Collision Domains
   (C) MAC Addresses
   (D) IP Addresses
4. Redundant components in a server:
   (A) cause duplicate data.
   (B) reduce fault tolerance.
   (C) improve fault tolerance.
   (D) increase data access speeds.
5. A central node that repeats messages to all attached nodes is called a:
   (A) repeater.
   (B) switch.
   (C) router.
   (D) hub.
6. Which network device has at least two IP addresses?
   (A) Switch
   (B) NIC
   (C) Router
   (D) WAP
7. A server that operates between clients and real servers is called a:
   (A) mail server.
   (B) proxy server.
   (C) web server.
   (D) file server.
8. A server running SMTP, POP and IMAP is probably a:
   (A) mail server.
   (B) web server.
   (C) file server.
   (D) proxy server.
9. File access rights in many NOSs are known as:
   (A) permissions.
   (B) policies.
   (C) profiles.
   (D) privileges.
10. Policies are used by network administrators:
    (A) to simplify tasks.
    (B) to assign the same rights to many users.
    (C) to assign the same services to many clients.
    (D) All of the above.

11. Outline the processes performed by each of the following devices.
    (a) NIC (d) Bridge (g) WAP
    (b) Repeater (e) Switch (h) Modem
    (c) Hub (f) Gateway (i) Router
12. Outline the services provided by each of the following.
    (a) File server (c) Database server (e) Web server
    (b) Print server (d) Mail server (f) Proxy server
13. A device marketed as an "ADSL Modem" also includes four Ethernet ports and a wireless antenna. Identify and briefly describe the devices integrated within this "ADSL Modem".
14. Outline the steps performed by a network administrator to complete the following tasks.
    (a) Add a new user.
    (b) Install a new client machine.
15. A home network includes three PCs with Ethernet NICS, a laptop with an 802.11 wireless interface, an Ethernet switch, a WAP and an ADSL modem.
    (a) Construct a diagram to explain how these components would best be connected.
    (b) Identify and describe the processes occurring, and the software and hardware used as the laptop browses the web.

ISSUES RELATED TO COMMUNICATION SYSTEMS

Throughout much of this chapter we have concentrated on the technical detail of how data is transferred; in this section we are concerned with the sharing of information and knowledge. After all this is the central purpose of all communication systems.

When communication is face-to-face one's physical appearance, cultural background, gender and physical location are all on display. These factors greatly influence the communication that takes place. When communicating electronically such factors remain largely unknown. In cyberspace relationships can be built on common interests and needs. Information and knowledge is shared between people who may never physically meet. People who would not (or could not) normally communicate face-to-face can freely express and share their ideas and knowledge online. These people are free to converse without prejudice. However all is not perfect, this freedom can easily be abused by the unscrupulous.

Electronic communication systems, and in particular the Internet, allow information to be shared quickly and relatively anonymously. The identity of the author can be hidden or obscured which makes it difficult for readers to verify the source and quality of the information. Unscrupulous persons are able to masquerade as trusted others in order to fraudulently obtain personal information such as credit card or banking details.

Most people presume their email messages to be private; in reality network administrators and others with suitable access rights are able to view and monitor emails. Those in control of networks are able to restrict and monitor the activities of users. Such power relationships are often legitimate, however as is the case with all such relationships power can be abused.

The Internet has removed national and international boundaries. We are free to communicate and trade internationally. Individual governments have little control over international trade and furthermore enforcing international laws is expensive and often ineffective in cyberspace. For example sending spam (mass electronic junk mail) is illegal within Australia, however Australian law has no control over spam sent from off shore locations.

To cover all possible issues arising when using communication systems is clearly not possible. Rather in this section we describe general areas for further discussion and then outline some current and emerging trends in communication.

INTERNET FRAUD

Fraud is a criminal offence in virtually all countries, however Internet fraud when detected rarely results in a conviction. Fraud involves some kind of deception that includes false statements that intentionally aims to cause another person to suffer loss. Unfortunately fraudulent activity using the Internet is the most common form of e-crime. Examples of Internet fraud include:

- Some spam messages try to convince users to purchase goods at discount prices. Users then enter their banking or credit card details, which are later used to make fraudulent withdrawals or purchases. In most cases prices that are "too good to be true" probably are!
- Identity theft is a form of fraud where someone assumes the identity of someone else. Commonly the criminal obtains various personal details about the person so that they can convince organisations that they are that person. This enables the criminal to take out loans, purchase goods and withdraw money from the person's bank accounts. Identity fraud even when discovered can have long term

consequences as the person must restore their reputations with many different organisations.

- Phishing is a form of spam where the email contains a message that purports to be from a trusted source. One common phishing scam uses mass emails purporting to be from a particular organisation and asking recipients to update their details by clicking on a hyperlink. The hyperlink takes them to a site masquerading as the real organisation's login screen. The fraudulent screen collects the user name and password and then forwards the user to the real site. Often users are unaware they are a victim of a scam as the criminals do not use the log in details for some time.

GROUP TASK Research
Using the Internet, or otherwise, research particular examples of Internet fraud. For each example determine if the perpetrators where actually convicted.

GROUP TASK Discussion
Many Internet fraud scams involve banks and other financial institutions. Despite this fact it is rare for such organisations to publicly disclose the extent of such fraudulent activities. Discuss.

POWER AND CONTROL

Those who control access to information are placed in a position of power over the users whose access they control. Not only can access to information be restricted and censored but the activities of users can also be monitored. Often users do not understand the extent to which their online activities can be monitored. Some issues to consider include:

- Parents install Internet filtering software to restrict their children's access to pornography and other inappropriate online information. Essentially parents are acting as censors for their children.
- Employers are able to monitor or even remotely watch and listen into their employee's online sessions and telephone calls. From the employer's perspective they are legitimately monitoring the quality of service provided. Many employees feel such systems imply a lack of trust and infringe upon their right to privacy.
- Email messages, unless securely encrypted, can be freely read by anyone with administrator rights to a mail server through which the messages pass. Many businesses claim they have a right to view messages sent and received on behalf of their company. However there are many cases where this has occurred without the knowledge of the employees.
- Backup copies of messages and web sites can and are stored for extended periods of time. Deleting a message from an email client or a file from a web server is not sufficient. Server archives have been used during investigations and have led to prosecutions.
- Organisations, including most schools, restrict and censor Internet access allowing only "approved" web sites and applications. In theory legitimate reasons exist and in most instances new sites and applications can be added to the approved list upon application. In practice many users find such controls oppressive and react with attempts to circumvent such restrictions.

GROUP TASK Discussion
Consider restrictions placed on Internet access at your school, work or home. Do these restrictions give power to those who administer and control Internet access? Discuss.

REMOVAL OF PHYSICAL BOUNDARIES

In cyberspace one's physical location is of little or no relevance. Individuals and organisations can trade across the globe. This globalisation has many advantages. For instance virtual communities can be created without regard to geographical location. However, there are also legal implications in terms of criminal activity and also in terms of taxation law. Information can be obtained from international sources as easily as from local sources.

- It is difficult to determine the real nature and location of online businesses. A single person can setup a website that appears to represent a large corporation. Such businesses can be setup quickly and they can be dissolved just as quickly. The legal safeguards available in Australia are not present in many other countries. In general Australian law does not apply to international transactions.
- Virtual organisations and communities are created as needs arise. Some are based on common areas of interest, to collaborate on a particular project or to form relationships. Participants in such organisations are largely honest and genuine, however in many cases ethical behaviour cannot practically be enforced.
- Most people speak just one language. As a consequence we seldom communicate with those who speak a different language. This greatly restricts our ability to understand and empathise with other cultures despite the removal of physical boundaries.

GROUP TASK Discussion
Identify particular examples of communication systems you have used that traverse international boundaries. Discuss issues you experienced during such communications.

INTERPERSONAL ISSUES

Electronic communication systems have changed the way many form relationships. Ideas delivered electronically can often appear less forceful and caring when compared to face-to-face communication. During face-to-face communication we continually receive and send non-verbal feedback to confirm understanding and to build relationships. Chat, teleconferencing and other real time communication systems are an attempt to address this issue, however non-verbal clues are not present, which can restrict one's ability to form meaningful personal relationships.

- Online dating sites enable people to present a particular well thought out view of themselves; initial personal contact being made via email. On the surface people feel they have much in common – similar background, culture, job, etc. However when face-to-face meetings subsequently occur people often find there is little or no real attraction.
- Ideas and comments from amateur individuals can appear as legitimate as those from professionals and large trusted organisations. On the Internet uninformed individuals can make their views appear as forceful and influential as experts. This is difficult and rarely occurs with more traditional forms of communication.

- Text based messages delivered via email or chat can easily be misinterpreted. It takes time to receive feedback and even when received it lacks the body language, tone of voice and facial expressions present when communicating in person.
- All are equal when communicating electronically. We need not even be aware that we are communicating with someone with a disability. For example most people have difficulty communicating face-to-face with someone who has a profound hearing disability. On the Internet we may not be aware of such a disability.

GROUP TASK Discussion
Many of us regularly communicate electronically with people we have never met face-to-face. Compare and contrast such relationships with more traditional face-to-face relationships.

WORK AND EMPLOYMENT ISSUES

Electronic communication systems have changed the way many people work and where they complete their work. For many jobs the ability to use electronic communication systems is required. Communication systems have provided the means for many people to work from home or from virtually any other location. They can vary their work hours and they can be contacted anywhere. This is certainly positive for employers and clients, however too often it has led to an expectation that employees are always available.

- Work teams can be setup where team members never or rarely physically meet. Rather they communicate and collaborate electronically using email, forums, teleconferencing and other electronic communication systems.
- Traditional employment is largely based on hours worked. When employees work from home they may well work unusual hours interspersed with other home and personal activities. This presents problems for employers who require reassurance that work is completed. It also presents problems for employees who must balance their intertwined work and personal lives.
- Most research indicates that those who work from home actually work longer hours and are more productive compared to those who travel to a specific work place. Some of the efficiency is due to the travel time saved, however the remainder is largely due to employees having more control and responsibility for the work they do.
- Many employees are provided with mobile phones and laptops that mean they are contactable in various ways 24 hours a day from almost any location. Today many expect to speak directly with people at any time of the day or at least that a response to messages will be made within an hour or so.
- Traditional retail stores are experiencing strong competition form online retailers. Potential customers often view goods in a physical store and then negotiate a better deal with an online retailer. Online retailers have significantly lower operating costs.

GROUP TASK Discussion
Do you know people who work substantially from home? Compare and contrast the nature of work for these people compared to those who travel to a specific workplace.

CURRENT AND EMERGING TRENDS IN COMMUNICATION

Blogs

Blog is short for web log, which is essentially a journal that is made public by placing it on the web. Individuals regularly update their blog to express their personal views and opinions or simply to detail their day-to-day activities. Most blogs are arranged in date order with the most recent entry at the top. It is common for people to include a blog on their personal website – for instance, many people maintain a personal MySpace.com webpage. MySpace.com includes software tools that automate the creation of blogs.

Wikis

A wiki is a website where users are able to freely add new content and edit existing content. Apparently the term "wiki" originated from the Hawaiian phrase "wiki wiki", which means "super fast"; the implication being that the amount of content grows rapidly due to the large number of authors. Probably the most well-known and largest wiki is Wikipedia; an online encyclopaedia created and edited by members of the public. Because the information within a wiki is produced by the general public it should never be accepted on face value; rather alternative sources should be used to verify the accuracy of the information.

GROUP TASK Discussion
Some organisations, including some schools, have blocked access to Wikipedia, whilst others embrace and encourage its use. Discuss and debate both sides of this issue.

RSS Feeds

RSS is an acronym for Really Simple Syndication. Syndication is a process that has been used by journalists and other content creators for many years. When content, such as a news story or TV show, is syndicated it is published in many different places. For instance, a TV show such as Neighbours is produced in Australia but is syndicated and shown in many other countries. RSS feeds implement this syndication process over the Internet. The author offers some content they have created as an RSS feed. Other people can then choose to take up the author's offer of syndication and subscribe to the feed. With RSS feeds the subscription is usually anonymous – the author has no idea of the identity of the people who have subscribed to their RSS feed.

Podcasts are distributed as RSS feeds, however any type of online content can be distributed using this technique, including blogs, wikis, news and even updates to web sites. The feed can contain any combination of audio, video, image and text. In addition, feeds need not contain the complete content; rather a partial feed can be used that includes links to the complete content.

To subscribe to RSS feeds requires newsreader software. The newsreader stores details of each RSS feed you subscribe to. The newsreader then checks each subscribed feed at regular intervals and downloads any updates it detects to your computer. This means the content is sitting on your computer waiting to be read – there is no need to download anything at this time, in fact the computer can be offline.

RSS feeds have become popular largely as a consequence of the excessive quantity of junk mail people receive. Many people are reluctant to enter their email address into web forms out of fear they may receive masses of unwanted email messages. No identifying information, including email addresses, is required to subscribe to an RSS feed.

GROUP TASK Research
RSS feeds are in many ways an extension of newsgroups, which have been around as long as the Internet. Research how newsgroups work.

Podcasts

Podcasting puts users in control of what they listen to, when they listen to it, how they listen and where they listen. Essentially a podcast is an audio RSS feed that is automatically downloaded to your computer and copied to your MP3 player. Aggregator software, such as Apple's iTunes, manages and automates the entire process – from the user's perspective content simply appears on their MP3 player.

The term "podcast" is a play on the words iPod and broadcast, however any MP3 player can be used, not just Apple iPods – a podcast is simply a collection of MP3 files. Podcasters are the people who create the "radio like" audio content, often on a regular basis or as a series of programs. Typically each podcast is a sequence of MP3 files created over time. Commercial media and other organisations are also embracing podcasting as an alternative to more traditional information delivery systems.

GROUP TASK Research
Blogs, wikis and podcasts are often referred to as part of "Web 2.0". Research and discuss the meaning of the term "Web 2.0".

Online Radio, TV and Video on Demand (VOD)

Online radio and TV programs are streamed over the Internet and displayed in real time using a streaming media player. Many traditional radio and TV stations now provide their programs online. Some stations provide a live digital feed, however it is the ability to watch past programs that distinguishes online delivery from traditional broadcasts – users can watch the programs they want, when they want.

Video on demand (VOD) systems are used to distribute video content directly to users over a communication link – much like an online video/DVD store. The aim of all VOD systems is to provide users with high quality video immediately in real time. Unfortunately current (2007) transmissions speeds and compression technologies are insufficient for this aim to be achieved. As a consequence VOD implementations compromise either quality, range of titles or the immediacy of delivery. Streaming systems compromise quality whilst largely achieving the range of titles and real time aims. Cable and satellite pay TV offer a limited range of high quality titles where each title commences at regular intervals – not quite real time. Online VOD stores deliver a large range of high quality movies. However movies must be downloaded prior to viewing – typical downloads take more than an hour.

3G mobile networks

The term 3G refers to third generation mobile communication networks. Essentially 3G networks provide higher data transfer rates than older GSM and CDMA mobile phone networks. As a consequence, access to much richer content is possible. 3G networks support video calls, web browsing and virtually all other Internet applications. Although 3G mobile phones are the primary device used on 3G networks, it is also common to use 3G networks to connect computers to the Internet. Currently high speed 3G coverage is limited to major cities and surrounding areas.

GROUP TASK Research
Research current 3G network speeds, the speed required for high quality VOD and predictions of future mobile network speeds. When will high quality VOD be possible over mobile networks? Discuss.

CHAPTER 3 REVIEW

1. Which list contains ONLY network hardware?
   (A) SMTP server, NOS, DBMS server.
   (B) UTP cables, switch, NIC.
   (C) Router, proxy server, codec
   (D) Ethernet, TCP/IP, HTTP.

2. In regard to error checking, which of the following is TRUE?
   (A) Messages containing errors are discarded.
   (B) Messages without errors are acknowledged.
   (C) Messages with errors are resent.
   (D) All answers – it depends on the protocol.

3. A 16-bit checksum is being used. For an error to NOT be detected what must occur?
   (A) The corruption must be the result of a data collision.
   (B) The sender or receiver has incorrectly calculated the checksum.
   (C) The message is corrupted such that the checksum is still correct.
   (D) The sender and receiver are not synchronised or are using different protocols.

4. The essential difference between the Internet and the PSTN is:
   (A) Internet is packet switched, PSTN is circuit switched.
   (B) Internet is circuit switched, PSTN is packet switched.
   (C) Internet is connection-based, PSTN is connectionless.
   (D) Internet is digital, PSTN is analog.

5. A switch is called a multipoint bridge because:
   (A) it separates a network into different segments.
   (B) it converts between two or more protocols.
   (C) It maintains a send and receive channel for each node.
   (D) it uses a physical and logical star topology.

6. An email includes email addresses within its To and Bcc fields. Which of the following is TRUE?
   (A) The To recipients are unaware of any of the other recipients.
   (B) The Bcc recipients are unaware of any of the other recipients.
   (C) Recipients in the Bcc field will be unaware of the To recipients.
   (D) Recipients in the To field will be unaware of the Bcc recipients.

7. Client-server architecture is best described by which of the following?
   (A) A central server performs all processing on behalf of all clients or workstations.
   (B) A network wired as a physical star where the central node is a server and other nodes are clients.
   (C) Clients request a service, and then the server performs the operation and responds back to the client.
   (D) A system where particular machines known as servers control access to all network resources for client workstations.

8. Networks where all messages are broadcast to all attached nodes utilise which topology?
   (A) Logical bus topology.
   (B) Physical bus topology.
   (C) Logical star topology.
   (D) Physical star topology.

9. A self-clocking code where high to low and low to high transitions represent bits is known as:
   (A) CSMA/CD
   (B) CSMA/CA
   (C) Manchester encoding.
   (D) Ethernet.

10. The ability to stream video of different quality to many participants is commonly implemented over the Internet as:
    (A) multipoint multicast.
    (B) multipoint unicast.
    (C) single-point, unicast.
    (D) single-point, multicast.

11. Compare and contrast:
    (a) MAC addresses with IP addresses.
    (b) ADSL and cable modems.
    (c) Checksums with CRCs.
    (d) Odd parity with even parity.
    (e) Packet switched networks with circuit switched networks.
    (f) Analog data with digital data.
    (g) Wired media with wireless media.
    (h) CSMA/CD with token passing.
    (i) Blogs and wikis.
    (j) Online radio and TV with traditional radio and TV.
12. Outline the operation of:
    (a) Video conferences over the Internet.
    (b) Electronic mail.
    (c) EFTPOS.
    (d) Self-healing dual ring topologies.
    (e) Routers.
    (f) modems
    (g) HTTP.
    (h) RSS feeds
    (i) VOD
13. Explain how messages are transferred over Ethernet networks where a physical star topology is used and the central node is a:
    (a) hub.
    (b) switch
14. Explain how digital data is encoded using:
    (a) Manchester encoding.
    (b) 256 QAM.
15. Outline the processes performed by SSL, HTTP, TCP and IP as a private message passes from source to destination over the Internet.

In this chapter you will learn to:

- recognise and describe a transaction
- identify, describe and use a batch transaction processing system
- distinguish between the storage of collected data and the storage of processed data in a batch system
- identify, describe and use a real time transaction processing system
- compare and contrast batch and real time transaction processing
- analyse an existing transaction processing system to determine its strengths and weaknesses
- design and implement procedures for validating entered data
- assess the work routine of a clerk in a manual transaction system to determine its suitability for automation
- identify participants, data/information and information technology for the given types of transaction processing systems
- describe the relationships between participants, data/information and information technology for the given types of transaction processing systems
- for a scenario diagrammatically represent transaction processing using data flow diagrams
- distinguish between the different types of TPS
- store digital data in databases and other files in such a way that it can be retrieved, modified and further processed
- implement systems to store paper transactions
- select and apply backup and recovery procedures to protect data
- document, including diagrammatical representations, the steps in batch processing
- document, including diagrammatical representations, steps in real time transaction processing
- identify systems for which batch is appropriate and is not appropriate
- distinguish between on-line real time and batch systems
- create and use a transaction processing system
- describe the operation of relevant hardware and how each is used to collect data for transaction processing
- design and justify paper forms to collect data for batch processing
- design user friendly screens for on-line data collection
- identify existing procedures that may provide data for transaction processing
- create user interfaces for on-line real time and batch updating, and distinguish between them
- identify situations where data warehousing and data mining would be an advantage
- assess the impact on participants involved in transaction processing
- identify jobs that have changed and/or jobs that have been created as a result of transaction processing, and report on the implications of these changes for participants in the system
- discuss alternatives for when the transaction processing system is not available and explain why they need to be periodically tested
- identify security, bias and accuracy problems that could arise from the actions of participants
- recognise the significance of data quality

Which will make you more able to:

- apply and explain an understanding of the nature and function of information technologies to a specific practical situation
- explain and justify the way in which information systems relate to information processes in a specific context
- analyse and describe a system in terms of the information processes involved
- develop solutions for an identified need which address all of the information processes
- evaluate and discuss the effect of information systems on the individual, society and the environment
- demonstrate and explain ethical practice in the use of information systems, technologies and processes
- propose and justify ways in which information systems will meet emerging needs
- justify the selection and use of appropriate resources and tools to effectively develop and manage projects
- assess the ethical implications of selecting and using specific resources and tools, recommends and justifies the choices
- analyse situations, identify needs, propose and then develop solutions
- select, justify and apply methodical approaches to planning, designing or implementing solutions
- implement effective management techniques
- use methods to thoroughly document the development of individual or team projects.

In this chapter you will learn about:

Characteristics of transaction processing systems

- a transaction - a series of events important to an organisation that involve a request, an acknowledgement, an action and an outcome
- the components of a transaction processing system, including:
  - purpose
  - data
  - information technology
  - processes
  - participants
- batch transaction processing - the collection and storage of data for processing at a scheduled time or when there is sufficient data
- real time transaction processing - the immediate processing of data
- the significance of data validation in transaction processing
- the historical significance of transaction processing as the first type of information systems

Types of transaction processing systems

- web-based
- non web-based
- on-line real time
- batch
- systems that appear real time, responding as the transactions occur, but where the actual updating is batch processed, such as credit card transactions

Storing and retrieving in transaction processing systems

- storage of digital data in databases and files
- retrieval of stored data to conduct further transaction processing such as printing invoices
- systems to store paper records of transactions
- data backup and recovery, including:
  - grandfather, father, son
  - off-site storage
  - secure on-site storage
  - full and partial backups
  - recovery testing
  - suitable media
  - specialised backup software
  - transaction logs
  - documenting backup and recovery procedures
  - mirroring
  - rollback
- updating in batch systems:
  - historical significance
  - limitations of batch processing
  - technology required
  - steps in a batch update
  - suitable applications
- updating in on-line real time systems:
  - relevance and impact
  - technology required
  - hardware requirements - large secondary storage
  - software requirements – on-line database and user friendly interface
  - steps in on-line real time processing
  - suitable applications

Other information processes in transaction processing systems

- collecting in transaction processing:
  - hardware including
    - Automatic Teller Machines (ATM)
    - barcode readers
    - Radio Frequency Identification (RFID) Tags
  - collection from forms
  - screen design for on-line data collection
  - web forms for transaction processing (real time and batch)
- analysing data, in which output from transaction processing is input to different types of information systems, such as:
  - decision support
  - management information systems
  - data warehousing systems (for data mining)
  - enterprise systems

Issues related to transaction processing systems

- changing nature of work and the effect on participants, including:
  - the automation of jobs once performed by clerks
  - shifting of workload from clerks to members of the public
- the need for alternate procedures to deal with transactions when the TPS is not available
- bias in data collection:
  - when establishing the system and deciding what data to collect
  - when collecting data
- the importance of data in transaction processing, including:
  - data security
  - data integrity
  - data quality
- control in transaction processing and the implications it has for participants in the system
- current and emerging trends in transaction processing
  - data warehousing and data mining
  - Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP)

4

OPTION 1
TRANSACTION PROCESSING SYSTEMS

Transaction processing systems are crucial to the operation of most finance, banking and electronic commerce organisations. Transaction processing is primarily concerned with maintaining data integrity. Such systems can operate at the single database level, but they also operate at higher levels where data in many databases and even many different systems is involved. For example transferring funds from one financial institution to another.

Transaction
A unit of work composed of multiple events that must all succeed or must all fail. Events perform actions that create and/or modify data.

So what is a transaction? A transaction is a series of events that when performed together complete some unit of work that is important to an organisation. Each transaction has two possible outcomes, either it is a complete success or it is a complete failure.

If a transaction is successful then all the events contained within the transaction must have performed their actions successfully. However, if one or more events are unable to complete their actions then the whole transaction must fail, which requires the data to be left in the same state it was in prior to the transaction commencing. This means any events that could successfully perform their actions must be stopped. For example when transferring funds between accounts two events must occur; one account is debited and another credited. If the debit event fails then the credit event must be stopped, similarly if the credit event fails then the debit event must be stopped.

Managing the success or failure of transactions is an essential process performed during transaction processing. Transaction processing systems include mechanisms for ensuring events can be completed successfully, but not yet permanently. Essentially the transaction processing system requests that each event occur and receives a response indicating that the actions performed are guaranteed to succeed or have failed. If a successful response is received for all events then the transaction as a whole can be "committed", meaning each event is requested to store its data changes permanently within the appropriate databases or systems. If one or more events have failed then the transaction is "rolled back", meaning each event is requested to abort all actions. In response each event sends an acknowledgement to confirm they have performed the request.

A transaction can include events that perform actions on a single database, many databases or on a variety of different information systems. These databases and systems can be widely distributed and in some instances they are operated by different organisations. The detail of how such transactions are processed will become clearer throughout the chapter.

GROUP TASK Discussion
Brainstorm a list of typical transactions and their component events and actions. Discuss problems that may cause these transactions to fail.

In this Option we commence by examining characteristics of transaction processing systems. This includes a brief examination of the history of transaction processing, how transaction processing automates manual tasks, the components of transaction processing systems and how such systems maximise the accuracy of data. We then examine real time and batch transaction processing systems. In real time systems each transaction is immediately processed online, whilst batch processing collects input data over time and then at some later time batches of many similar transactions are processed. Backup and recovery strategies and technologies are examined. We then examine collection hardware and forms used for collection. The data within transaction processing systems is used as input to other systems; we briefly consider examples of such systems. Finally we discuss issues related to transaction processing systems.

CHARACTERISTICS OF TRANSACTION PROCESSING SYSTEMS

Transaction processing is one of the earliest commercial uses of computer systems. In this section we examine some early examples of transaction processing to illustrate how such systems automate and improve upon manual transaction processing. We then examine features of manual transaction processing systems that make them well suited to automation. Finally we examine modern transaction systems – their components and how they maintain the accuracy of data.

HISTORICAL SIGNIFICANCE OF TRANSACTION PROCESSING

The operations performed by transaction processing systems were up until the 1950s performed solely by clerks using manual processes. Early computers were originally developed to solve scientific and mathematical problems for government and military. It was during the 1950s that the application of computers to business and financial records emerged.

Prior to the 1980s it was common for complete transaction processing applications to be developed (often in Cobol) for each individual organisation. During the 1980s database management systems emerged to manage and control access to databases. Today most transaction processing systems are based on one or more relational database management systems (RDBMS) with client applications being written to meet an organisation's specific needs. Some of the significant developments that have led to today's transaction processing systems are outlined below.

Fig 4.1
UNIVAC I the first commercially available computer used for transaction processing.

- UNIVAC I (Universal Automatic Computer), released in 1951, was the first commercially produced computer to gain wide acceptance by the public. The UNIVAC I was based on vacuum tubes and was the first computer to be routinely used for batch processing of business transactions. UNIVAC I was designed and built by John Presper Eckert and John William Mauchly. Their company Eckert-Mauchly Computer Corporation was bought by Remington Rand – both Eckert and Mauchly continued to work for Remington Rand after selling the company. The UNIVAC II and UNIVAC III were subsequently released.

- The programming language Cobol (Common Business Oriented Language) was developed in 1959. At the time computers were largely used for scientific and mathematical calculations. Cobol, as its name suggests, was targeted directly at business applications and is still widely used on large mini and mainframe computers. Cobol was the first language for large scale transaction processing.
- In 1964 IBM released its highly successful System/360 range of computers and peripherals (see *Fig 4.2*). These general-purpose systems supported approximately 40 different peripherals and included the ability to include redundant components to improve fault tolerance. Information systems based on System/360 supported real time input and processing from hundreds of attached terminals.
- In 1969 IBM released the first version of CICS (Customer Information Control System). In terms of transaction processing, CICS is a transaction processing monitor or TPM – it manages the processing of transactions from multiple clients to multiple servers. This software product has been continuously upgraded and is widely used today.
- SQL was first developed in the early 1970s by IBM under the name "System R". The design ideas for System R were a direct result of Ted Codd's work – Codd is considered the founder of relational database theory. At the time System R was viewed as a product to allow users to directly interrogate databases. The original designers never intended it to become a language that would be used from within applications.

Fig 4.2
IBM System/360 Model 65 operator consol attached to the CPU.

- In the early 1980s commercial general-purpose relational database management systems (RDBMS) emerged. Oracle just beating IBM's release of DB2. These systems used SQL both to create relational databases using DDL (Data Definition Language) statements and to view and update relational data using DML (Data Manipulation Language) statements.
- SQL first became an ANSI (American National Standards Institute) standard in 1986, however most current database systems, although compliant with most of the standard SQL syntax also include their own non-standard extensions.
- Microsoft entered the RDMS market in the 1990s with its SQL Server product. Microsoft SQL Server evolved from the Sybase DBMS; Microsoft dissolved their partnership with Sybase and renamed their product SQL Server.
- Today Oracle, IBM's DB2 and Microsoft's SQL Server dominate the market, however some open source products such as MySQL have significant market share within small to medium sized organisations.
- Today large enterprises such as banks, large corporations and government departments use transaction processing monitor (TPM) software to manage transactions across a variety of databases and applications of different types, operated by different organisations and in different locations. Common TPMs in use include IBM's CICS and Encina products, BEA System's Tuxedo software and more recently Microsoft Transaction Server (MTS). TPMs are an example of enterprise systems as they manage critical data and processes across an entire organisation.

GROUP TASK Discussion
Initially transaction processing software was written for each specific application. Today it is common to use standard DBMS and TPM software and only the client applications are custom solutions. Propose likely reasons why this has occurred.

GROUP TASK Research
Research and identify examples of enterprise systems. Determine whether these systems process transactions.

AUTOMATION OF MANUAL TRANSACTION PROCESSING

Processing of manual transactions almost always follows a strict sequence of events. Each event must be acknowledged as complete before the next commences and if any event fails then the entire transaction is aborted. In manual systems, events are performed by clerks and other personnel according to strict predefined rules. Indeed in large organisation's it is common for each clerk to repetitively perform just one of the events within each transaction. The transaction is handed onto the next clerk responsible for the next event in the sequence. The strict sequence and rules of such transactions make them particularly well suited to automation using computers.

As a simple example let us consider a manual system used within a small store and then assess the benefits of automating this system. The store is operated by a husband and wife team who have time during the day to perform all sales, purchasing, stocktaking and other transactions manually. The store uses a simple cash register, which is essentially a calculator with an attached cash drawer. The cash register does keep a total of all sales processed during the day. The store has an EFTPOS terminal, which operates as a separate system.

Sales – this transaction occurs to process each customer's purchases.
1. Locate price on item and enter into cash register.
2. Repeat 1 for each product.
3. Calculate total.
4. Receive payment from customer.
5. If EFTPOS payment then wait for approval and hand EFTPOS receipt to customer.
6. Enter payment amount into cash register.
7. If cash payment then calculate and hand change to customer.
8. Hand register receipt to customer.

Stocktake – this transaction collects data to enable the storeowner to calculate the quantity of each product to purchase.
1. Make a photocopy of stocktake sheets. These sheets specify the required number of each product when fully stocked, the product's supplier and also columns for recording current stock in the store.
2. Count and record number of each product on shelves.
3. Count and record number of each product in storeroom.

Purchasing – this transaction produces purchase orders for each supplier.
1. Complete Stocktake.
2. Calculate number of each item to purchase and record on stocktake sheets.
3. Create purchase order for supplier.
4. Work through stocktake sheets recording each product from current supplier.
5. Calculate order total.
6. Repeat steps 3, 4 and 5 for each supplier.
7. Submit all purchase orders to suppliers via fax.

GROUP TASK Discussion
Three transactions are described above, however there are other transactions that need to occur. Propose other likely transactions and outline a set of possible events occurring to perform these transactions.

Notice that much of the data used by all three of the above transactions is the same. It is the information generated by the transaction that is different. Furthermore the output from one transaction is used as data for another transaction. For example each sales transaction reduces the amount of stock, and each stocktake transaction produces the data required for purchasing. Such observations make this system well suited to automation. The flow of data and information entering and leaving each of these transactions is modelled on the data flow diagram in *Fig 4.3*. Note that each of the transactions is represented as a process as they are composed of events that process data in some way. Each of these transactions could be expanded into a lower level DFD or a step-by-step description that details their component events.

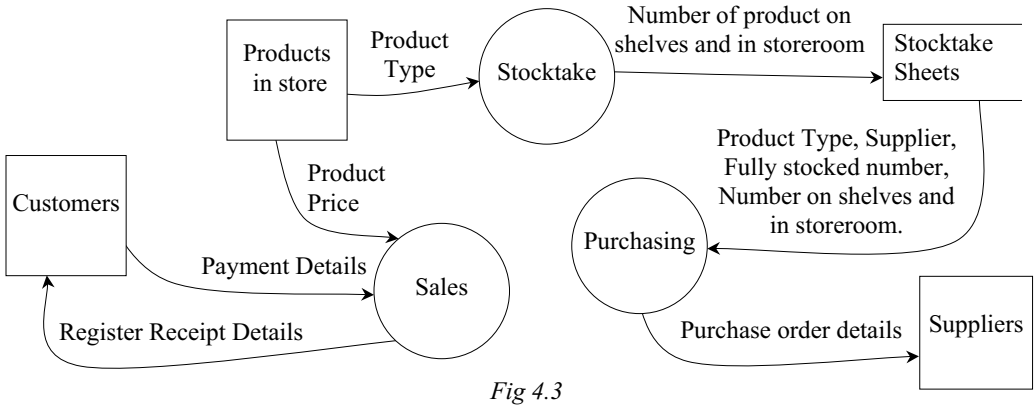


Fig 4.3
Data flow diagram modeling the flow of data between a store's manual transactions.

The stocktake sheets perform many of the tasks performed by a database, hence on the DFD a data store is used. They store all the data required by the purchasing transaction process. In addition the stock take sheets allow processing to halt between the stocktake process and the purchasing process. The Products in Store entity could also have been represented as a data store as each product stores its price in the form of a price tag and its product type. In reality these are the actual products, hence representing them as an external entity makes more sense.

No doubt it is clear that this system could be automated using a relational database to integrate sales, product, supplier, orders and stocktake data. Later in this chapter we shall examine a point of sale (POS) system, which is essentially an automated version of the above manual system. At this stage we are interested in the strengths and weaknesses of manual systems and of automation. Let us consider some general areas relevant to most manual systems together with common strengths and weaknesses of automation. We shall then discuss our local store example in an attempt to assess its suitability for automation.

Manual system strengths:

- Minimal start-up costs – little or no initial capital expenditure.
- Minimal training time and costs.
- Quick response to changing requirements.
- Well suited to small organisations where participants have time and fulfil multiple roles.
- Responds well to human insight and intuition.

Manual system weaknesses:

- Analysis of historical data is difficult and time consuming.
- Transactions take considerably longer to process.
- Difficult to rigidly enforce transaction rules and sequences.
- Redundant or duplicate data is a feature of most manual systems.
- Some human errors are to be expected.
- System becomes more and more difficult to manage as it grows.
- Making backups of data is difficult and is rarely, if ever, performed for all data.

Automated transaction processing strengths:

- Much faster transaction processing.
- Less repetitive work for participants.
- Enforces the sequence and rules for each transaction.
- Calculation errors are virtually eliminated.
- Ability to integrate transaction processing with outside organisations.
- Historical data available for statistical and financial analysis.
- Backups easily made and restored if system fails.
- System easily grows as transaction processing needs grow.

Weaknesses of automated transaction processing:

- Significant start-up costs to purchase information technology.
- Extensive training required to operate the system.
- Changes to requirements often require specialised expertise to implement.
- Rigidly enforces existing transaction rules and sequences for all data.
- Less total work for humans resulting in lower employment.
- Reliance on information technology – failure of one or components can cripple the entire system.

In our local store example the storeowners are a husband and wife team who are currently able to complete the manual transactions. In this case the time saved through automation is unlikely to result in increased profits. Furthermore the cost required to set-up and learn to operate a new automated system is unlikely to be justified. If the storeowners had sufficient expertise to design and develop their own automated system then this would be worthwhile. Without an automated system it is difficult for the owners to accurately monitor sales trends over time. If they were able to perform such historical analysis then perhaps significant savings could be made by maintaining more efficient stock levels of products according to predicted demand at different times of the year. It is likely that this is currently occurring in a somewhat ad-hoc manner for obviously seasonal items – such as ice creams, Christmas decorations and gloves.

Consider the following:

Each of the following businesses currently use a manual system for recording their various transactions.

- A hardware store that stocks thousands of different items and has a staff of 8 employees working at all times.
- A small bookstore that is able to supply any title but maintains minimal stock. The store purchases titles as they are ordered by customers.
- A carpenter who substantially does subcontract work for 3 builders but does do some small jobs for residential customers.
- An eBay store that started out selling approximately 5 items per day, but is now selling 50 items per day.

GROUP TASK Discussion
Assess the suitability of an automated transaction processing system for each of the above businesses. Discuss likely advantage and disadvantages of automation compared to retaining their existing manual systems.

COMPONENTS OF TRANSACTION PROCESSING SYSTEMS

In this section we examine the various components of transaction processing systems. Like all information systems (see *Fig 4.4*), transaction processing systems operate within an environment that includes information processes, participants, data/information and information technology. All these components work together to meet the system's purpose. Each transaction is an information process and is therefore composed of events that are also information processes. For example adding a new customer to a database involves collecting and storing information processes. All information processes are performed using the resources within the system. The system's resources include participants, data and information, and information technology.

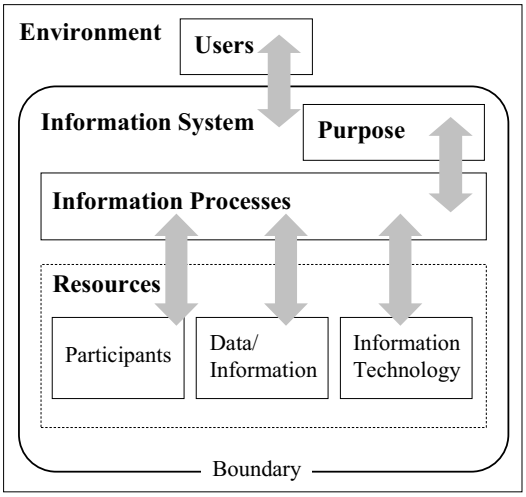


Fig 4.4
Components of an information system.

We have already introduced the general nature of transactions and much of the remainder of this option continues to examine different types of transactions in more detail. Therefore in this section we shall concentrate on participants, data/information and information technology within transaction processing systems. Recall that participants are people who carry out or initiate information processes. Information technology includes the hardware and software that carries out information processes.

Participants

Anybody who interacts directly with a transaction processing system becomes a participant in that system – they are integral to the system's operation. Therefore participants include people who work for the organisation that operates the transaction processing system and also people (often customers of the organisation) who enter data that initiates transactions. For example a bank employee is a direct user (and participant) as they initiate the printing of monthly bank statements. Customers become direct users (and participants) when they use Internet banking to initiate say the transfer of money between accounts. On the other hand indirect users are not participants, they send and/or receive data from the system but do not directly cause its entry or display. For example when a monthly or quarterly bank statement is received by a customer in the mail the customer is not a participant, rather they are an indirect user as they did not initiate the generation of the statement directly.

People in the environment only become participants in online real time transactions. Real time transactions are performed immediately in response to user or participant input. On the other hand for batch processes people from the environment may well provide data, but no transaction processing occurs or is initiated by that person. Rather the transaction is performed along with other similar transactions at some later time or when a sufficient quantity of data is present. Consider the difference between a

transaction performed using an ATM compared with writing a cheque. When using an ATM the user initiates and therefore causes the transaction to be performed immediately – essentially they are performing duties similar to a bank teller. However when writing a cheque the customer has little control over when the transaction is actually processed and furthermore they are not interacting directly with the bank's transaction processing system.

Data/Information

In the majority of transaction processing systems data is stored in databases – usually relational databases. This data is transformed into information by the system's information processes. We studied the organisation and design of relational databases in depth in chapter 2. All the information in regard to tables, records, relationships, referential integrity, data validation, data integrity and data verification applies to transaction processing systems. However in transaction processing systems a further issue exists – how to ensure the integrity (correctness and accuracy) of data during transactions. What if another user or process views or alters data during a transaction? What if the data received from another system has problems? What if the system fails in some way during a transaction? In regard to data and information such issues are resolved by recording the detail of all transactions in a transaction file or log. How these transaction records help to resolve these issues will become clearer in the next section on data validation and data integrity.

Within transaction processing systems additional data is always created to record details of each transaction that occurs. In older systems the actual live data was commonly known as the master file and the details of each transaction was recorded in a transaction file. The application controlled and managed both the transaction file and the master file. All changes being recorded in the transaction file during transaction processing, with changes to the master file only being made when transactions are finally committed. Newer systems still create such transaction data (often called a transaction log), however management of this transaction data is left up to the DBMS and, if used, the transaction processing monitor rather than the application software. Most commercial operating systems also provide transaction capabilities as part of the file system. Such operating systems create transaction records that allow actions on complete files to form part of transactions. These operating system capabilities are also available to other applications, including transaction processing monitors.

To simplify our discussion let us refer to the transaction data or transaction file as a transaction log and the actual data as the master file. Recall that transactions can be committed or rolled back. The transaction log contains the essential data that facilitates this ability. When an event occurs as part of a transaction two possibilities arise:

1. *Fig 4.5* describes the first possibility for an event that modifies a single record. The event occurs however the changed or added records are recorded in the transaction log and no change is

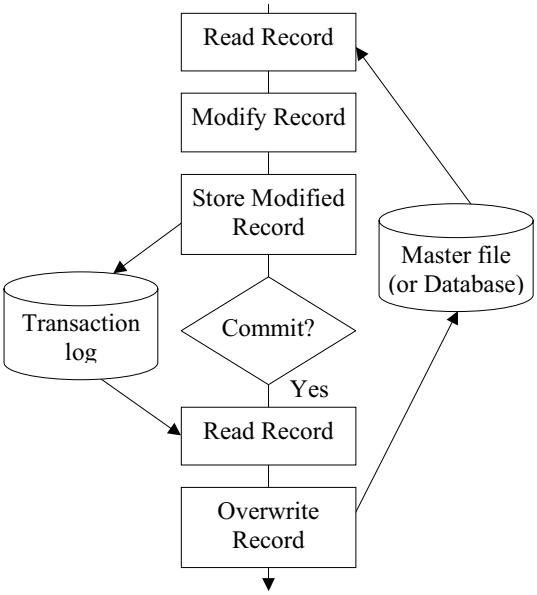


Fig 4.5
Flowchart describing modifying a record as part of a transaction where the master is not altered until the transaction commits.

yet made to the actual data in the master file. If the transaction is committed then the records in the transaction log replace or are added to the master file. If the transaction is rolled back then the records in the transaction log are not written to the master file.

2. Copies of the original unchanged data are recorded in the transaction log and then the changes are made immediately to the actual data within the master file. If the transaction is committed then nothing more needs to occur. If the transaction is rolled back then the record in the transaction file is copied back over the actual data in the master file. When new records are created as part of a transaction the transaction log must contain an entry specifying the record to delete should the transaction be rolled back.

GROUP TASK Activity
Create a flowchart to model the processes occurring to modify an existing record in the master file using the second strategy described above.

In either case the transaction log is used to enable the committing or rolling back (or even rolling forward) of events within transactions. Most current DBMSs actually record both before and after versions of the data within their transaction logs – in essence they allow implementation of both the above possibilities. This means the transaction log is really a log of all the activities performed on the data.

The most compelling reason for maintaining before and after versions of all data changes is to provide a backup of all recent changes since the last backup. The database (or master file) can be restored from the most recent backup and then the transaction file can be used to commit (or roll forward) all transactions performed since the restored backup was made. If at the time of failure some transactions were incomplete then those events that formed part of such transactions can be rolled back. Such restore operations are essentially automated within most modern DBMS and transaction processing monitor software products.

A complete transaction log is also useful during audits as it shows when, what and who performed each transaction. Utilities are available for most DBMS products that allow the transaction log to be analysed in detail. Such utilities also allow transactions in the log to be rolled back and rolled forward individually.

GROUP TASK Research
Transaction log files continually grow in size – sometimes their size can exceed the size of the actual database. Research techniques and strategies for ensuring transaction logs do not grow excessively.

Information Technology

The hardware and software forms the information technology of the system. Transaction processing systems vary enormously in both size and scope. A small database may serve just a few local users, however a similarly small database may serve many more users via the web. Larger critical transaction processing systems perform thousands or even millions of transactions daily. The hardware and software requirements vary enormously; hence in this section we shall introduce some general areas for consideration. Later in this chapter we examine more specific examples where the detail of the hardware and software can be specified more precisely.

- **Hardware**

Possible hardware for transaction processing systems includes:

- Server machines that include redundant components to improve fault tolerance. In medium to large systems multiple servers provide access to the same

database. If one server fails or is taken offline for maintenance or upgrading then the other servers automatically take up the extra load. Larger government and multinational organisations commonly use mainframe computers that are able to support thousands of users and access to enormous databases.

- Storage devices with sufficient capacity and data access speeds to support the size of the database and the number of users. Commonly RAID storage is used. For transaction logs even on relatively small systems a mirrored RAID solution is common to ensure that a single failure of one drive does not result in loss of incomplete and recent transactions.
- Communication devices and transmission media able to support the number of required users and data access speeds to ensure acceptable response times. Many servers include multiple NICs to achieve higher data access speeds.
- Backup devices such as tape drives, tape libraries, CD burners and DVD burners. In some systems complete copies of the data are maintained on mirrored hard disks located in a different location to the operational data.
- Client workstations for running the client applications and interacting with the system's participants and users. These machines may include specialised collection devices such as barcode scanners, RFID readers, touch screens and magnetic ink character recognition (MICR) readers. The client machines may be dedicated devices such as ATMs or EFTPOS terminals or they may be personal computers connected via the Internet or the organisation's network.

- **Software**

Possible software for transaction processing systems includes:

- DBMS software to manage and control the transactions performed on linked databases. We discussed the operation of DBMSs in chapter 2. The DBMS runs on one or more servers and provides services to client applications to enable them to access the databases within the transaction processing system. Each DBMS includes the ability to manage transactions performed on databases under its control and includes a transaction log.
- Client applications that are installed on client workstations and provide the interface for participants to initiate and perform the system's information processes. The client applications make requests, often in the form of SQL statements, to the database servers using a client-server architecture. For larger systems that perform transactions across many servers or many systems the client applications send their requests via the transaction processing monitor.
- Proprietary software applications that are designed and developed to meet the needs of a specific organisation. Such software is common in large transaction processing systems running on mainframe machines. These software applications are written from the ground up – including providing, specifying and controlling access to data directly without the use of a DBMS. The term proprietary means the software is produced for a specific system or organisation and is generally owned by that organisation.
- Transaction processing monitors (TPMs) are software applications that coordinate the transaction processing of large transaction processing systems. *Fig 4.6* describes the general software architecture of transaction processing systems that include transaction processing monitors. These large systems commonly include many database servers that may access the same logical database or may access different databases. In addition transaction processing monitoring applications can connect to systems operated by other organisations.

Each server or system has its own resource manager (refer *Fig 4.6*) that makes available resources to the TPM. A resource manager is essentially a software product that provides an interface between the resource and the transaction processing monitor.

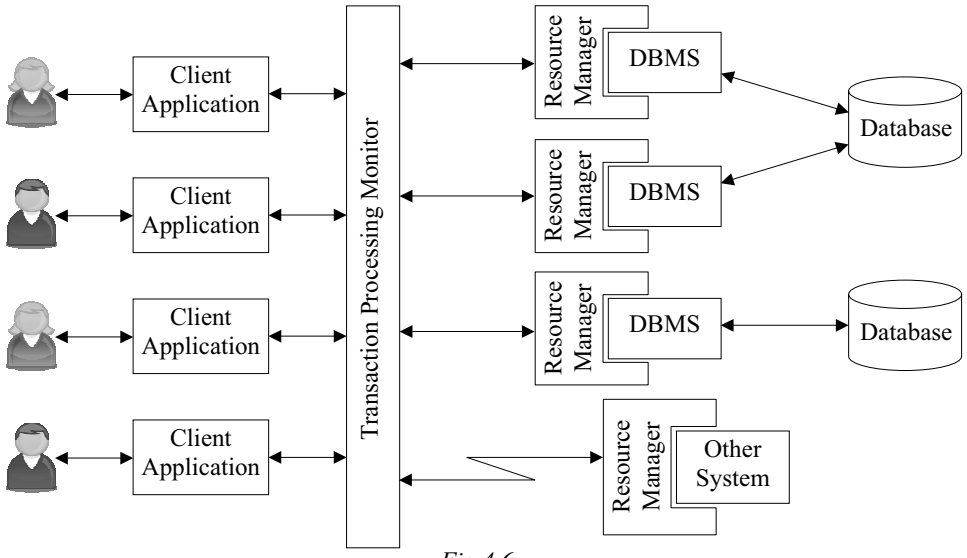


Fig 4.6
General architecture of a system that includes a transaction processing monitor.

The main task of transaction processing monitors is to ensure the integrity of transactions that include events which execute on different servers and/or systems. The TPM controls the commit and rollback of the total transaction in response to requests sent to and acknowledgements received from resource managers. Each server or system performs its own lower level transaction and reports the outcome (success or failure) back to its resource manager, who in turn communicates with the TPM. In addition TPMs are able to balance the load of transactions sent to each server. Transaction processing monitors are also known as transaction managers or transaction processing services. Examples include IBM's CICS and Encina products, BEA System's Tuxedo and Microsoft Transaction Server.

GROUP TASK Discussion
A user initiates the transfer of funds from a Commonwealth Bank account to a Westpac bank account. Discuss the role of transaction processing monitors during this transfer.

DATA INTEGRITY

Data Integrity
A measure of how correct and accurately data reflects its source. The quality of the data.

The integrity of data is critical in all transaction processing systems. Recall from our earlier work on database systems (Chapter 2) that data integrity is a measure of how correct and accurate data is compared to its source. In Chapter 2 we considered three techniques for improving data integrity, namely data validation, data verification and also referential integrity. In this section we briefly discuss examples of each technique within transaction processing systems. We then introduce the ACID properties of transactions and the type of problems they solve.

Data Validation

Data validation checks ensure reasonable data enters the system. In transaction processing systems data that is incorrect at the time of collection is likely to cause a variety of different problems when it is later used as part of transactions. There are two different types of data validation commonly performed. The first ensures the data entered is of the correct data type and format. This is generally performed by the client application. The second is more difficult as it aims to ensure the data entered is correct in terms of the business rules of the enterprise. That is, it determines if the data is correct in terms of its ability to be processed. For example when ordering a book the ISBN is often entered as a unique identifier. Data validation within the client application ensures the correct number of digits are entered. The book store's business rules require that the ISBN must exist within their database. Therefore a query must be executed to validate that this is indeed true.

Data Validation
A check, at the time of data collection, to ensure the data is reasonable and meets certain criteria.

A single data entry error that is undetected can affect numerous transactions across many organisations. For example consider a BPay reference number on a supplier's invoice that is being paid by a customer using Internet banking. Let us assume this invoice must be paid before the goods are shipped. If the BPay reference number is entered incorrectly by the customer then the total transaction will eventually fail. The consequences of this simple data entry error is costly for both the customer and also for the organisations involved in the transaction. The bank must inform the customer of the problem, however the customer is not aware of any potential problem and hence they are unlikely to check their bank messages. The supplier does not receive the funds and therefore will reissue the invoice or simply not supply the goods. The customer is not happy as they are unaware of the error and hence wonder why their goods do not arrive. Resolving the problem involves further time and cost for all parties. These issues could be resolved by validating the BPay reference number prior to the transaction commencing.

Data Verification

Data verification is used to maintain the integrity of data over time. This is a difficult task in most information systems and is rarely 100% successful. For example people and also businesses move location, change their phone numbers, credit card numbers and even change their names. Ensuring that such changes are reflected in the data is the aim of data verification processes.

Data Verification
A check to ensure the data collected and stored matches and continues to match the source of the data.

In large government and commercial transaction processing systems data verification becomes an enormous undertaking. Currently in Australia there is no single unique identifier that can legally be used to identify individuals across all these systems. If such an identifier was available then it would be possible for individuals and organisations to change their details in one place and have these changes replicated to other systems. Privacy concerns prevent such practices. For example in the mid 1980s the federal government attempted to release the "Australia Card", which was to contain a unique number for each Australian citizen and resident. This number was to be used to link records between most government departments and even between commercial organisations. As a result of public outcry over privacy

concerns the legislation was never passed. Currently tax file numbers (TFNs) and Australian Business Numbers (ABNs) are shared between many government agencies albeit with strict controls in regard to how data can be linked and used. In Australia it is illegal for private organisations to use TFNs and ABNs to link data from multiple sources.

GROUP TASK Discussion
Discuss advantages and disadvantages of widespread use of a unique identifier for each Australian citizen and resident.

Referential Integrity

In relational databases referential integrity ensures all foreign keys in linked tables match a primary key in the related table. This means a record in the primary table must exist before records can be added to the table containing the linked data. If referential integrity is not enforced then orphaned records will exist. In general such records cause significant problems when queries are executed on the database.

Within a single database referential integrity is enforceable and hence problems simply cannot occur within the database. When many databases are involved or identifiers are being entered by users then problems are inevitable. Data validation and verification issues can affect referential integrity. For instance, entering an incorrect BPay reference number means that the primary records held in the various organisations' databases cannot be linked to the customer's payment. The Australia Card aimed to provide a primary key for each Australian that could be used as the foreign key in many linked databases. Both systems are attempting to use a unique identifier in an attempt to enforce referential integrity across multiple databases.

GROUP TASK Discussion
Brainstorm real world examples of data validation and data verification that aim to improve the referential integrity (and therefore the data integrity) of databases.

ACID Properties

ACID is an acronym for atomicity, consistency, isolation and durability. The aim is to ensure all transactions comply with these four properties. They ensure that transactions are never incomplete (atomicity), the data is never inconsistent (consistency), transactions do not intrude or affect each other (isolation) and that the results of a completed transaction are permanent (durability). All these properties combine to ensure the integrity of all data is maintained before, during and after each transaction.

To illustrate each of the ACID properties let us use an example transaction – making an airline reservation using a credit card. This transaction includes the following general sequence of events:

1. Reserve a seat on a specific flight.
2. Process and approve credit card payment.
3. Issue and record ticket details.

- **Atomicity**

To be atomic all events within a transaction must complete successfully or none at all. If any single operation fails then the entire transaction is aborted. This involves rolling back all events completely so that the data is returned to its original state. If all events are successful then the transaction is committed, which means the data changes are made permanent or durable.

In our airline transaction imagine what would occur if just one operation failed but the others were committed. If no seat were reserved then the passenger would arrive with a paid ticket but with no available seat. If the payment is not processed and approved then the passenger receives a seat and ticket for free – great for the passenger, but not so good for the airline. If no ticket is issued or recorded then the passenger and airline have no record of the transaction resulting in the passenger being refused a seat.

- **Consistency**

The consistency property ensures transactions take data from one consistent state and then when the transaction completes the data is left in a consistent state. For a single event on a single database this is enforced using referential integrity and validation rules. When the transaction includes many events and spans many databases or systems then consistency must apply across all these databases and systems.

In our airline transaction a business rule is likely to require the total number of reserved seats to be equal to the number of tickets issued. If a seat is reserved but does not result in a ticket being issued then the data is inconsistent in regard to this business rule. Many other rules are also likely, such as, a customer must be assigned to each reserved seat, all tickets must be paid for and each ticket must be assigned to a specific flight and passenger.

- **Isolation**

Transactions must process data without interfering with or being influenced by other transactions that are currently executing. In effect each transaction logically executes in isolation to all other transactions. During the processing of a transaction the data is often placed in an inconsistent state. For example when transferring funds between accounts, money is debited from one account and then credited to another account. After the debit but before the credit the data is in an inconsistent state. This inconsistent state should not be exposed to other transactions. Furthermore the records involved should not be available for other transactions to change until the transaction is completed. If the isolation property is not observed then queries will return inconsistent results and other transactions will process with potentially erroneous data.

In small systems where only one transaction executes at a time the isolation property is simple to achieve as one transaction completes before the next commences. If many transactions can execute at the same time then the solution is more involved. However even the largest transaction processing systems must ensure their method of implementing the isolation property results in the same effect as executing each transaction sequentially.

When multiple transactions can execute concurrently all data involved in a transaction must be locked such that other transaction processes cannot alter it. We discussed record locking strategies used by DBMSs in chapter 2 – these strategies are also used within transactions that span multiple databases and systems. Note that locking does not alter the actual data, rather it prevents other operations from changing the data. As a transaction is committed the actual data is altered. Significantly other processes are aware that a record has been locked by another transaction. Therefore other transactions must wait for the lock to be released before they proceed.

Record locking, transaction logs and the "two-phase commit" nature of transactions all influence each other and combine to implement the isolation property. The term "two-phase commit" refers to events being performed temporarily (phase one) during a transaction and then being committed (phase two) if the transaction completes successfully. The first phase is recorded in the transaction log and also involves the record being locked. The second phase alters the actual data permanently and releases the record lock.

Consider our airline transaction example. Imagine the isolation property is not present and a single seat remains available on a flight. Many passengers are now able to simultaneously reserve this single seat successfully, as long as each transaction commences prior to committing one of the other transactions. Furthermore they will go on to pay and be issued with a ticket. When passengers board the flight the airline will discover there are more passengers than available seats.

- **Durability**

Durability ensures that committed transactions are absolutely permanent. Theoretically this means that even if the whole world crashes the changes made by the transaction will be OK. In real systems durability ensures that during a commit the results are actually written to some physical storage device. Notification of a successful commit can therefore be reasonably relied upon.

At first it may seem that executing an update query when committing will ensure durability of the changes, however in many systems data is held in RAM for a period of time and is only written to secondary storage as required. Such systems improve performance, however if power is lost then the contents of RAM is permanently lost. Therefore durability specifically requires all changes to be written to permanent or secondary storage before the transaction is truly committed.

In our airline example, imagine an example transaction is apparently committed successfully. Now say the durability property is not present within the "issuing and recording ticket" event. Suppose the system fails and this operation is not recorded. When the passenger goes to board the flight their ticket will not exist on the system. However inconsistencies will be present as a reservation will exist for the passenger and a record of payment also exists. Resolving this issue will be costly in terms of time and also in terms of inconvenience for the passenger.

HSC style question:

Define the term *transaction* and explain how data integrity is maintained during processing of transactions.

Suggested Solution

A transaction is a unit of work composed of a sequence of events. All actions performed by all events must succeed for the transaction to be committed permanently. If any single event within a transaction fails then all events within the transaction are aborted or rolled back. Commonly each event within a transaction alters data within a database.

Whenever data is altered the potential exists for inaccuracies to be introduced and the integrity of data to suffer. Transactions avoid such possibilities through their ACID properties. Atomicity ensures a transaction succeeds completely or fails completely. Consistency ensures each transaction takes the data from one consistent or correct state to another consistent or correct state. This means inaccuracies or data integrity issues are only possible during processing of a transaction. This possibility is dealt with by the isolation property. This ensures data changes are not available to other transactions until they have been committed. The durability property ensures all changes made by all events occurring within all committed transactions are permanently written to storage. This increases data integrity as it guarantees the consistency of the data after each transaction completes is maintained permanently.

SET 4A

1. Which of the following best describes a transaction?
   (A) An event that alters or creates a record within a database.
   (B) Multiple events that must all succeed or all must fail.
   (C) A system that controls the execution of many transactions across many databases or systems.
   (D) A process that alters data in different records, databases or systems.
2. Transaction processing using computers first emerged during the:
   (A) 1980s
   (B) 1970s
   (C) 1960s
   (D) 1950s
3. A transaction log contains:
   (A) details of the data added or updated during processing of transactions.
   (B) details of the original data prior to it being updated by transactions.
   (C) sequential copies of the data within the master file.
   (D) Answer A and/or B
4. Manual transactions performed by clerks are often well suited to automation because they:
   (A) are boring and repetitious for participants to perform.
   (B) follow a strict predefined sequence of rules.
   (C) can be performed as batch processes.
   (D) commonly include just one operation that alters data.
5. Bank customers become participants when they:
   (A) write a cheque.
   (B) receive a statement in the mail.
   (C) withdraw cash from an ATM.
   (D) All of the above.
6. Examples of TPMs include:
   (A) SQL Server, Oracle, DB2
   (B) SQL Server, CICS, MTS
   (C) Oracle, Encina, Tuxedo
   (D) CICS, Tuxedo, MTS
7. The data needed during commit and rollback processes is stored within the:
   (A) transaction log ..
   (B) master file.
   (C) operational database.
   (D) data source.
8. Which of the following is the most significant task performed by TPMs?
   (A) Manage access to many remote DBMS servers within an enterprise system.
   (B) Provide an interface between client applications and resource managers.
   (C) Manage and control transactions whose events span multiple databases and/or systems.
   (D) Force all events within a transaction to be permanently committed.
9. Over time existing data becomes less and less accurate. Which of the following is undertaken to improve this situation?
   (A) Data verification.
   (B) Data validation.
   (C) Referential integrity checks.
   (D) Ensure transactions adhere to the ACID properties.
10. Transaction A reads data whilst transaction B is executing. Transaction B is rolled back, however transaction A commits. It is later determined that transaction A has introduced inconsistencies into the data. Which ACID property is NOT present?
    (A) Atomicity
    (B) Consistency
    (C) Isolation
    (D) Durability

11. Define each of the following terms?
    (a) Transaction (c) TPM (e) Data validation
    (b) Data verification (d) Referential integrity (f) Data integrity
12. Recommend suitable data validation techniques when collecting each of the following?
    (a) Exam marks that are out of 100.
    (b) A pair of dates, where the first date must be prior to the current data and the second must be at least 1 week after the first date.
    (c) Adding a product and required quantity to a customer's order.
13. Outline the history of computer based transaction processing systems.
14. Explain how transaction processing systems implement the ability to commit or rollback transactions.
15. (a) Why are most manual transactions well suited to automation? Discuss.
    (b) What data integrity issues are resolved when all ACID properties are enforced? Discuss.

REAL TIME (ON-LINE) TRANSACTION PROCESSING

Real time transaction processing systems complete transactions immediately they have been initiated. In most examples of real time transactions an online user initiates the transaction. Online users include employees of the organisation and also customers entering details via the web or other networks. Note that when a user directly interacts (participates) with the system they become a participant in the system.

Real time transaction processing in many references is known as online transaction processing or OLTP. In real time or online systems, as opposed to batch systems, each transaction must complete within a reasonable amount of time. If an operation takes longer than a second or two then feedback must be provided to assure the user that processing is indeed taking place. Wait times of more than a few seconds are likely to be aborted by users in the belief that an error has occurred – this is particularly true of users who are not members of the organisation. This presents significant problems when transactions are initiated over the web. The organisation has no control over the speed of transmission once packets reach the Internet. However they do have control over the speed of data access and processing performed by their hardware and software. In general, real time transaction processing systems require faster direct (or random) access to secondary storage, faster and more secure communication links and more processing power than batch processing systems.

We mentioned above the need for fast response times when using online data entry forms. We shall discuss the design of such forms in some detail later in this chapter. However it is worth introducing some often used strategies. Commonly online forms collect data that is then validated or used as criteria to search a remote database. If the validation involves simple data type or format issues then it can occur within the data entry form by the client. If validation involves referring back to the remote database then response times become an issue. For example when applying for car insurance a form collects details of the make and model as well as personal and payment details. To maximise response times it is common for data collection to be split into a sequence of forms. The data collected on each form being validated or used for a search prior to display of the next form. In our car insurance example a form is used to collect the make of car, say Holden. The "make" is used as the search criteria so that the next form sent to the user need only contain different models of Holden – Commodore, Berlina, and so on. In addition to reducing the amount of data that needs to be transferred, sequences of forms also mean just a few user inputs need to be validated at a time. This improves user friendliness as problems are identified soon after the user made the input.

Consider the following

Consider a single form that collects say twenty data items over the web. Say two validation problems are found. The user is then presented with these two items along with some messages outlining the nature of the problem. Firstly, it will take longer for the result to be returned to the user and furthermore it may have been many minutes since the user made the problem entries. Consequently the user is forced to readjust their thinking to make the corrections. If just a few items are input then validation messages are returned to the user before their thoughts have moved on.

GROUP TASK Discussion

Are the issues discussed above relevant to online data collected over a LAN from participants who are part of an organisation? Discuss.

In this section we examine three examples of real time (online) transaction processing systems. We examine a reservation system, and two non-web based systems, namely a point of sale system and a library loans system. In each case we will identify the participants, data/information and information technology within the system. We also model some of the information processes performed as part of the system's transactions using data flow diagrams and other system models.

RESERVATION SYSTEMS

Reservation systems are used to collect and process bookings for a variety services. Examples include hotel, motor vehicle rental, airline and concert reservations. These systems, although different in terms of the detail of how they are implemented, process similar transactions. In general a transaction that reserves a service is composed of the following operations (or information processes).

1. Collect and store details of required service.
2. Confirm availability of service and temporarily reserve service.
3. Collect and store customer details.
4. Collect payment details as required.
5. Process and store payment as required.
6. If successful then commit reservation permanently.
7. Create and display confirmation to customer.

Today many systems allow customers to initiate the processing of reservations via the web. Many of these systems still provide phone information services or operator assisted telephone services. The essential processes within each transaction remain similar regardless of the interface used to communicate with the customer. If payment or a deposit is required at the time of reservation then it is common for a separate system operated by a financial institution to be used to process and approve payments. A typical context diagram for a reservation system is reproduced in *Fig 4.7*. On this context diagram participants who work for the organisation are not included as an external entity. This is clearly correct when customers enter data directly via the Internet and it is also correct when data is entered into the computer system by employees of the organisation. In all cases the data originates from the customer whether they enter data directly via the web (as a participant and user) or via an operator (as an indirect user).

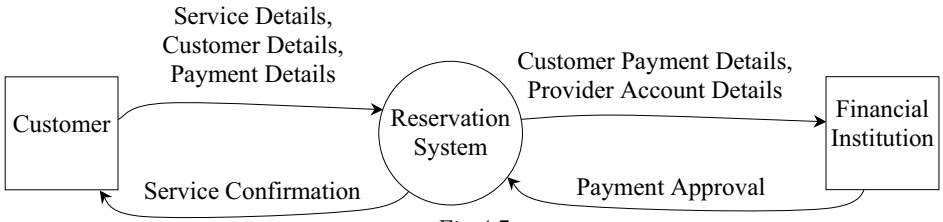


Fig 4.7
Context diagram for a typical reservation system.

There are significant advantages of real time processing of reservations. Customers are provided with a confirmed booking virtually immediately. The service provider is able to allocate their service on a first in first served basis and furthermore they have immediate access to the number of allocated seats, vehicles or rooms on any particular day or for any particular event. This information enables the service provider to tailor their marketing and prices to attract customers to poorly performing days or events.

Let us consider the seven steps outlined above in terms of our understanding of transactions. Do all seven steps need to be successful for the transaction to complete

or be committed? Yes, if any step fails then all processes should be reversed or rolled back. Is the sequence of steps significant? Yes it makes logical sense to ensure availability prior to collecting personal details. It also makes sense to collect personal details prior to processing the customer's payment. Clearly confirmation should not be made until all steps complete correctly.

Consider the following web-based hotel reservation system:

The Hytton is a large city Hotel with a total of 500 rooms. There are four room types, namely, double, queen double, deluxe double and penthouse in ascending size, features and price order. The Hotel has rooms on 13 of its 15 floors. Some rooms have harbour views and generally the view is better on higher floors. Currently the Hotel charges according to room type without regard to view or floor. Rooms with better views are assigned to repeat customers and also based on customer requests.

There are numerous different transactions that occur before, during and after a typical stay at the Hytton. Clearly transactions occur when guests check in, check out, order movies, food and drinks and also as part of routine operations such as cleaning rooms, ordering supplies, payroll, and so on. For our purpose we will restrict our discussion to reserving a room and checking into the hotel upon arrival.

Reservation Transaction (phone-based)

The operations performed during a typical phone reservation transaction are modelled on the systems flowchart in *Fig 4.8*. Notice that the Hytton does not require payment or a deposit at the time a phone reservation is made. The Hotel Database is included twice on the model simply to improve readability.

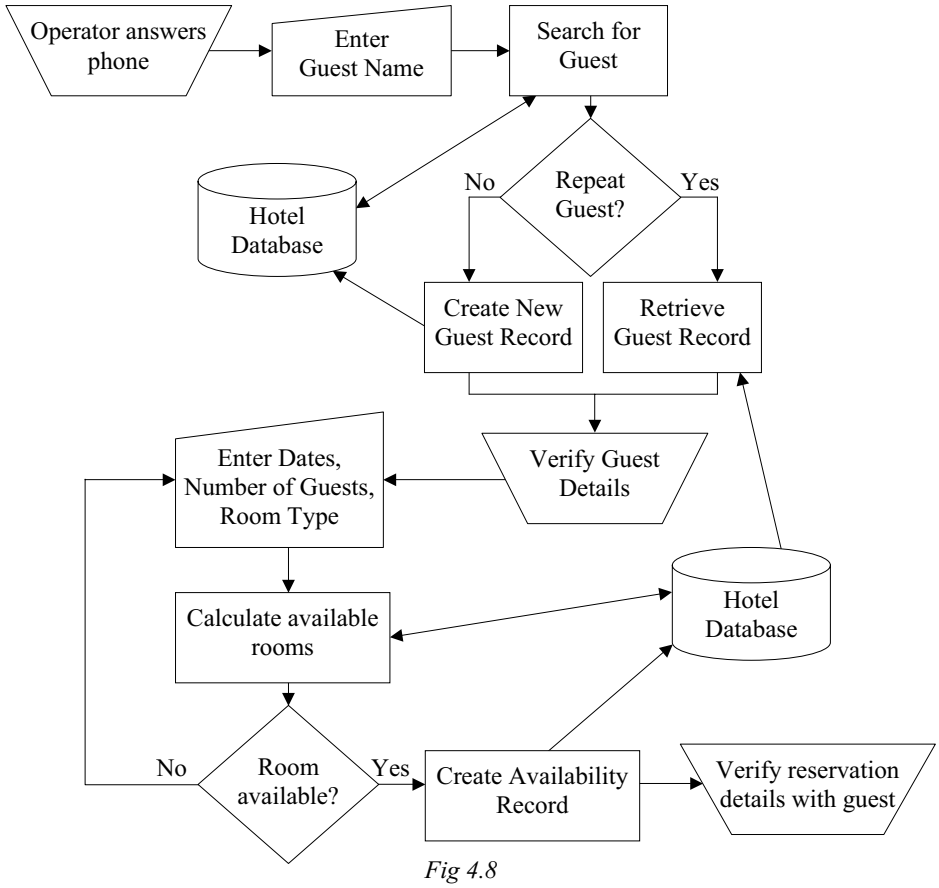


Fig 4.8

The steps performed for a phone reservation transaction at the Hytton Hotel.

GROUP TASK Discussion
Work through the above flowchart (*Fig 4.8*). Is it true that all the processes described must complete successfully for the total transaction to succeed? Discuss.

Reservation Transaction (web-based)

Guests are able to reserve and pay for rooms on-line using the sequence of web forms reproduced in *Fig 4.9*. The final confirmation form is not reproduced; it displays a Reservation Number together with all the confirmed reservation details in a format suitable for printing. A copy of this confirmation form is also emailed to the guest.

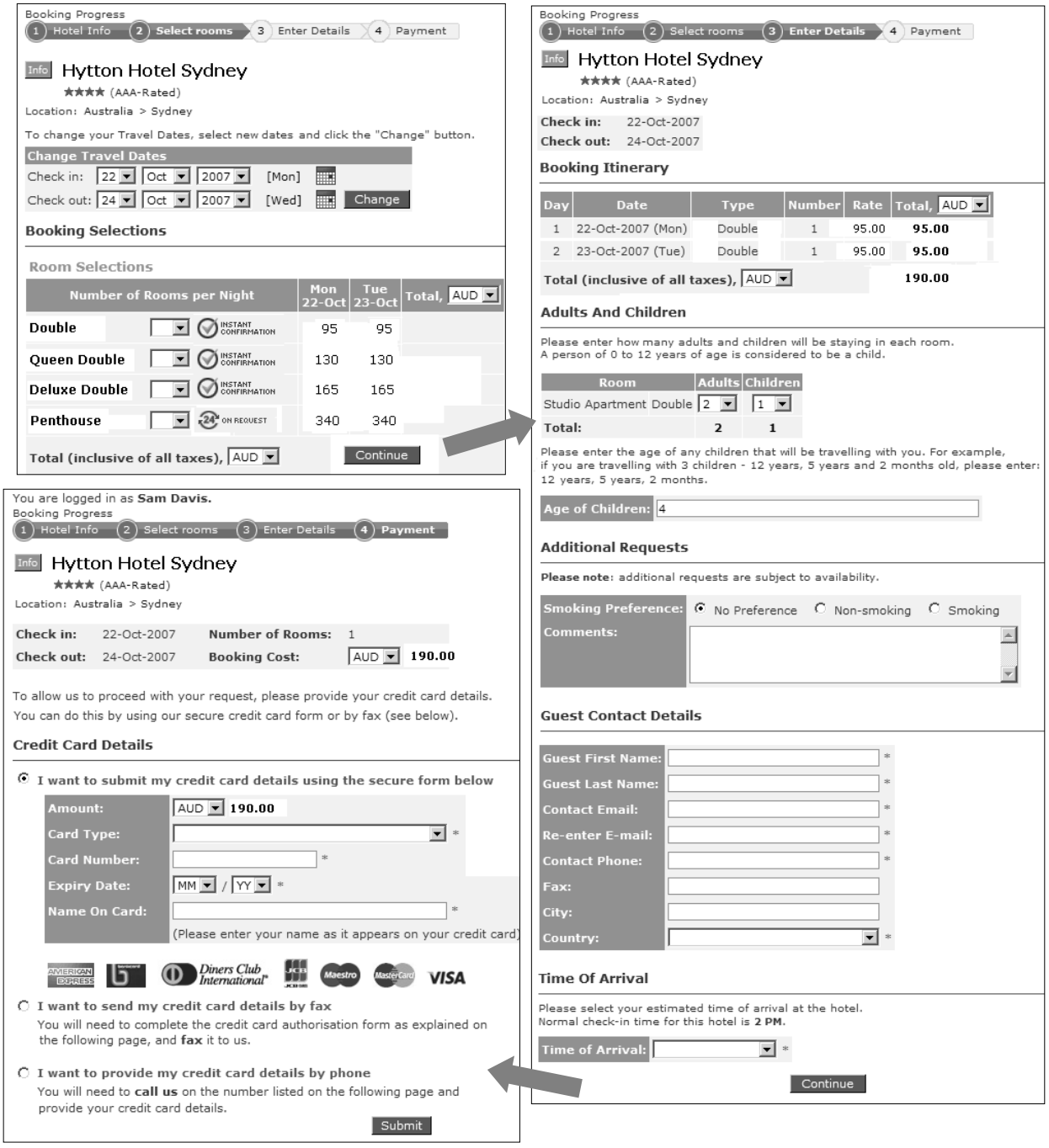


Fig 4.9
Web-based forms for Hytton Hotel reservation transaction.

GROUP TASK Discussion
Analyse the web forms in *Fig 4.9* in terms of their design, the purpose of each of the input fields and the order in which the inputs are collected.

GROUP TASK Discussion
Predict likely processes that are being performed each time the user clicks the continue or submit button on each of the *Fig 4.9* forms. Justify your predictions using evidence on the forms.

GROUP TASK Discussion
Compare and contrast making a phone reservation with making a web based reservation from the perspective of a guest.

Check-in Transaction

Guests check-in at the front desk upon arrival at the Hytton Hotel. The following processes or operations are performed by one of the front desk staff during a typical check-in transaction:

1. Welcome and determine guest's name.
2. Find reservation records – guest record and associated availability record.
3. Complete personal details of guest record – address, phone, other guest names.
3. Determine any specific guest requests in regard to view or floor.
4. Assign specific room to guest, which is stored in availability record. Repeat guests are automatically assigned an available room with the best view.
5. Determine payment method – credit card preferred.
6. If cash or EFTPOS and has not paid then collect and process deposit.
7. If credit card and has not paid in advance then reserve funds for the cost of the room via EFTPOS terminal.
8. Create charge record for deposit, reserved funds or prepayment.
9. Generate electronic swipe card room key.
10. Print check-in details, attach charge receipt and staple to inside of information kit.
11. Hand information kit and swipe card key to guest and verbally verify all details.
12. Arrange porter to deliver luggage to room.

GROUP TASK Discussion
Identify the participants and also the tasks they perform during the phone reservation, web reservation and check-in transactions.

Data/Information

The data for the Hytton Hotel system is stored within a relational database. This database includes the tables and relationships described in the schema shown in *Fig 4.10*. Note there are many more tables that form part of the complete system – only those tables used during the phone reservation and check-in transactions are shown.

GROUP TASK Activity
Refer back to the details of the phone/web reservation and check-in transactions. Determine the records that are accessed, changed and/or created during these transactions.

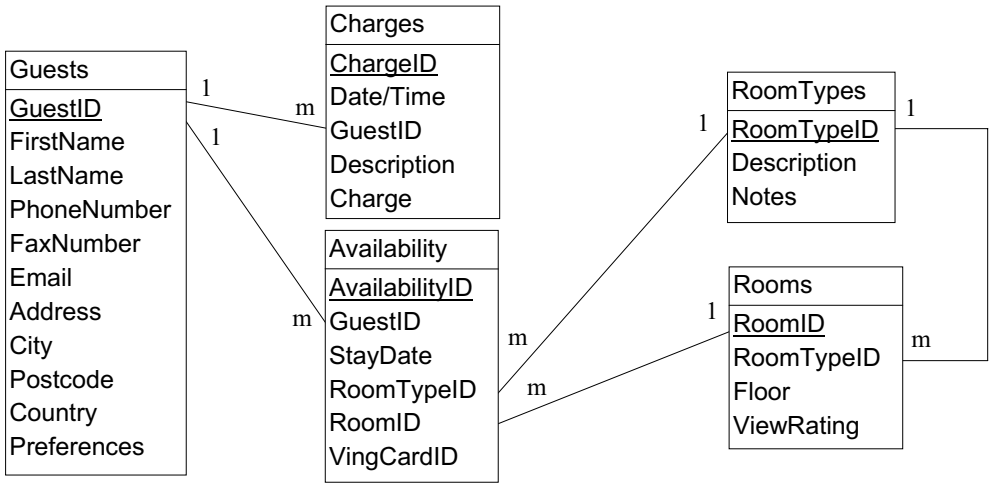


Fig 4.10
Partial schema for the Hytton Hotel database.

Information Technology

The Hytton Hotel's system uses a client-server architecture with the database stored on a RAID storage device attached to the database server. Throughout the Hotel there are a total of 65 workstations with different hardware configurations. Details of the hardware and software include:

- The web and DBMS server software runs on separate Dell PowerEdge 2950 Servers. The servers includes two Intel Dual-Core processors and 32GB of RAM.
- The database is managed by Microsoft's SQL Server DBMS software together with a customised server application.
- Microsoft's Internet Information Services web server software runs on the web server. The web server connects to the Internet via a bank of cable modems. The cable connections also supply pay television and Internet access to guest rooms.
- Although the Hotel's website uses SSL, payments are not processed in house. All online credit card payments are directed to the hotel's bank where they are approved and the funds are deposited directly into the hotel's account.
- The RAID device includes 8 hard disks with a total storage capacity of approximately 5TB. The system uses RAID 5, which includes both striping and mirroring to improve both data access performance and fault tolerance.
- The client application within the Hotel runs on each of the 65 workstations and has been customised to suit the particular needs of the Hotel. The client and server applications are based on a proprietary hospitality application.
- The Hotel uses the VingCard security lock system (see *Fig 4.11*). Each lock has its own unique ID and includes flash memory to store the last 600 entry and exit events. Hotel staff requiring access to rooms are issued with swipe cards, however these do not operate locks on occupied rooms.

Fig 4.11
Generating a guest swipe room card using the VingCard 2800 terminal.

Swipe cards are coded to operate locks in elevators, hotel entrance doors, other hotel facilities such as conference rooms and pools, and of course guest rooms.

- 20 laser printers and 7 small receipt printers are installed throughout the Hotel.
- Each workstation runs Microsoft Windows Vista and includes a 100Mbps Ethernet connection back to the central rack of switches.
- The server connects to a rack of patch panels and Ethernet switches via an Optical Gigabit interface and cable. Connections to all workstations are cabled using Cat 5e UTP.
- Partial backups are performed each night and full backups each week. Backups are written to a small attached tape library capable of auto loading 8 tape cartridges from its built in magazine. Each tape stores 400GB, so total capacity without manual intervention is approximately 3.2TB.

GROUP TASK Discuss
Briefly explain the purpose of each of the hardware and software items listed above in terms of performing reservation and check-in transactions.

POINT OF SALE (POS) SYSTEMS

Point of sale systems process transactions within retail outlets. Retail outlets include small local stores, chains of stores, hotels and clubs, and also large department stores and supermarkets. Although the amount of data increases significantly for larger retailers the general nature of the essential transactions remains similar. Retail stores sell directly to customers and they purchase inventory or stock from suppliers as described on the *Fig 4.12* context diagram. These two processes, purchasing and selling, form the basic transactions performed by all retailers and hence must be present within all POS systems. The retailer sends suppliers a purchase order, the supplier sends the products together with a delivery docket followed shortly by an invoice. Finally the retailer pays the invoice and sends the payment details in the form of a remittance advice. When selling retailers accept payment from customers and provide the customer with a receipt – often in the form of a tax invoice. In addition to these basic transactions most POS systems provide a variety of additional features to enhance security, monitor business performance and assist with marketing.

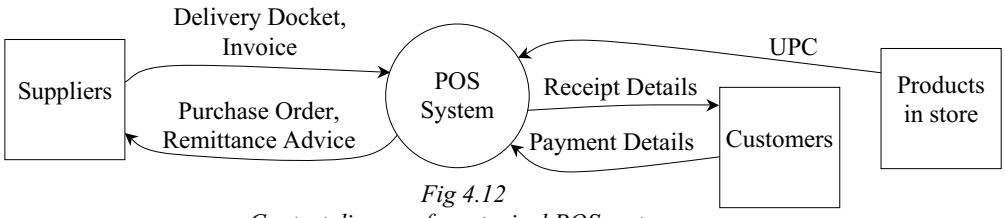


Fig 4.12
Context diagram for a typical POS system.

Particular companies produce and market proprietary POS systems for specific industries. Some companies produce and market complete POS systems for jewellery stores, others specialise in hardware stores, whilst others specialise in fruit and vegetable stores. Commonly these systems include all hardware and software, together with the training required to operate the system.

GROUP TASK Research
Research examples of proprietary POS systems used by various industries. Identify features of the POS systems specific to the industry they serve.

Most packaged products include a printed Universal Product Code (UPC). Each UPC is a 12-digit number usually printed with an equivalent barcode on the product's packaging. The first 6 digits uniquely identify the manufacturer, the next 5 digits uniquely identify each of the manufacturer's products and the final digit is a check digit. For high value items, such as jewellery, a unique identifier and associated barcode is commonly created for individual items by the POS software. For products sold by weight, such as fruit and vegetables, product codes, if used, are added in store once the product has been weighed and packaged.

GROUP TASK Activity
UPC check digits are calculated by summing the 6 digits in odd positions and multiplying by three. This result is added to the sum of the 5 digits in even positions. The check (twelfth) digit is the difference between this total and the next multiple of ten. Examine UPCs on a number of products and confirm the check digit is correct.

Let us expand on our initial POS system context diagram in *Fig 4.12* by considering some typical transactions performed by most POS systems.

Sales –this transaction processes each customer's purchases using the POS terminal.

1. Scan UPC on product packaging.
2. System retrieves product description, price and stock level from database.
3. Stock level reduced by one and stored in database.
4. Repeat 1 to 3 for each product.
5. System calculates total.
6. Process customer payment– EFT, credit card or cash.
7. (i) If EFT then swipe card, have customer enter their PIN and wait for approval.
   (ii) If credit card then wait for approval receipt and collect customer signature. Check signature matches signature on credit card.
   (iii) If cash payment then enter amount tendered and hand change, if any, to customer.
8. Hand receipt (tax invoice) to customer.

Generate Purchase Orders – this transaction creates and submits purchase orders to each supplier electronically.

1. User initiates transaction on a daily basis.
2. System queries database for low stock products. Query returns number of each product to order sorted by supplier.
3. Review each product and confirm order.
4. System generates and submits purchase orders to suppliers via either email or fax.

Receive Delivery – this transaction processes each order when it arrives at the store.

1. Manually check actual products delivered match delivery docket.
2. Enter purchase order number from delivery docket. System retrieves and displays purchase order.
3. If invoice and delivery docket products match then enter date received.
4. System adds number of each item to current stock level of each product.

Enter Invoice – this transaction processes each invoice received from suppliers. Invoices arrive by mail or fax and are often batched entered on a weekly basis.

1. Enter purchase order number from invoice.
2. System retrieves and displays purchase order.
3. If invoice details match purchase order details and products received then enter invoice number and mark for payment.
4. Details, including prices, that do not match require manual override/correction in consultation with supplier.

Pay Suppliers – this transaction produces payments for each supplier based on those invoices that are due for payment. From the user's perspective this is a batch process performed at the start of each month.

1. User initiates transaction at start of each month.
2. System retrieves and displays summary of remittance advice notices for payments due to each supplier. Each remittance advice includes invoice numbers and invoice totals, together with payment total.
3. User confirms each supplier payment.
4. System generates remittance advice notices that include printed cheques.

GROUP TASK Discussion
Identify the participants and the tasks they complete during each of the above transactions.

Based on the above four transaction we create a lower level DFD to describe the flow of data within the above system (refer *Fig 4.13* DFD). On this DFD the store database is included twice – this is simply to improve readability. Clearly other transactions will also occur in most real world POS systems.

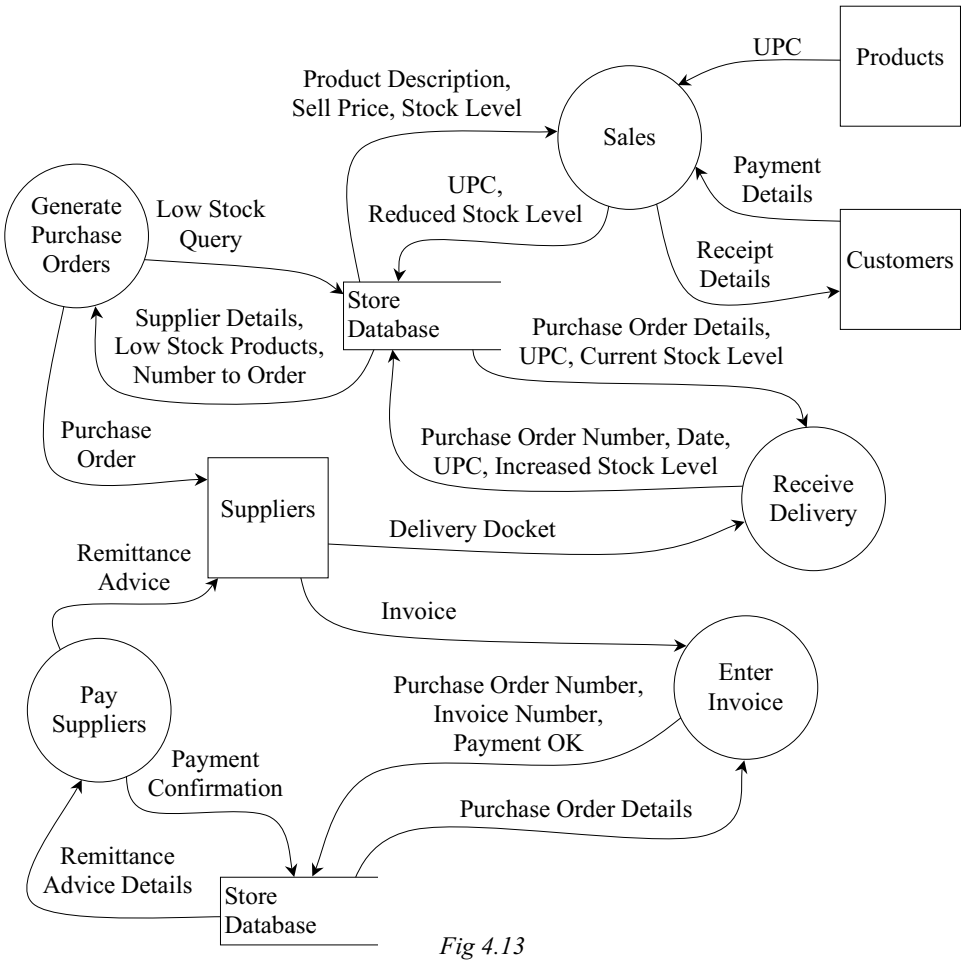


Fig 4.13
DFD for a typical POS system.

GROUP TASK Discussion
Confirm the DFD in *Fig 4.13* correctly reflects the context diagram in *Fig 4.12* and the transaction descriptions on the previous page.

Data/Information

The data entering and used within POS systems is detailed on the context diagram (*Fig 4.12*) and DFD (*Fig 4.13*). These models also show the information output by POS systems – that is, receipt details, purchase orders and remittance advices.

The data within POS systems is almost always stored within a relational database. For the system described above tables for products, suppliers and purchase orders would be required – a possible schema is reproduced below in *Fig 4.14*. In reality the schema would be far more complex to meet additional requirements. For instance, currently no record is maintained of when products were sold and therefore sales trends cannot be analysed. Also each product is assigned a single supplier and cost price. In reality many products are available from multiple suppliers at varying prices. Additions and modifications would also be required if the retailer accepts orders for out of stock products or high value products that are individually coded. Most POS systems also maintain records of each sales assistant and the sales they process.

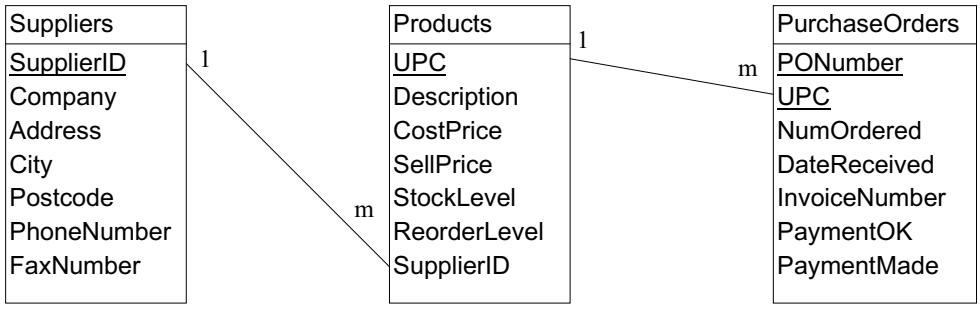


Fig 4.14
Initial schema for a simple POS system.

GROUP TASK Discussion
For each of the four transactions described and with reference to the above schema, identify records that are examined, created and/or modified.

GROUP TASK Discussion
Recommend modifications to the above schema so that records of when each sale was made are maintained.

Information Technology

The essential information technology for POS systems includes a database server that runs DBMS software and includes sufficient storage to secure and maintain the database. For smaller retailers backups are made to CDRs, whilst larger systems include tape drives. One or more POS terminals are installed which run the client application that processes sales transactions. Further personal computers are often present to perform other transactions. Commonly an Ethernet LAN is used to connect to the database server.

Fig 4.15
Touch screen POS terminal.

All hardware, apart from the POS terminals, is common to many other systems. Therefore we restrict our discussion to the detail of POS terminals. Firstly, the use of the word "terminal" is somewhat misleading – in fact most current POS terminals are in fact personal computers that include integrated collection and display devices. In the past POS terminals were indeed terminals where processing was performed centrally. Today POS systems are largely client-server systems and hence much of the processing is performed by the client.

Currently most POS terminals include a standard PC motherboard including Intel processor, RAM and hard disk. Attached or integrated devices include touch screens, magnetic stripe readers, barcode scanners, cash drawers, receipt printers and specialised keyboards.

GROUP TASK Practical Activity

Make a note of POS terminals you observe during the week. Identify the devices present within these POS terminals and comment on the design of each POS terminal.

For most POS systems the size and robustness of POS terminals is at least as important as the technical performance specifications. There is limited space at most checkouts and POS terminals are used continuously for extended periods. POS terminals must be better able to withstand spills and other hazards. The small size of LCD monitors made them popular inclusions in most POS terminals long before their widespread use for other applications.

Fig 4.16
Example restaurant touch screen user interface.

Ergonomic issues for participants using POS terminals are different compared to the issues present for those seated at more traditional computer workstations. POS terminals are commonly used whilst standing for extended periods of time and the collection devices are different. The tasks performed by POS terminal users often include a much broader range of movements as they scan products, use touch screens and interact with customers. Barcode scanners, touch screens and magnetic stripe readers reduce the likelihood of RSI and other health issues associated with keyboard

data entry. The design of user interfaces for touch screen POS applications is quite different to other user interfaces. For example the screen reproduced in *Fig 4.16* includes large coloured buttons and is customised for each restaurant.

GROUP TASK Discussion
Compare and contrast the design of touch screen user interfaces with user interfaces designed for use with a keyboard and mouse.

LIBRARY LOANS SYSTEMS

Library systems perform a variety of processes in addition to transactions required to lend resources. Some of these processes include maintaining the catalogue, searching the catalogue, purchasing, management of library finances, booking of personal computers and small group rooms, managing and charging for printing, sharing of collections with other libraries and integration of digital data within traditional collections. We cannot hope to examine all these areas of library management hence we restrict our discussion to some of the transactions related to the lending (check-out) and return (check-in) of library resources. In library terms these processes are known as "circulation" processes. The desk where resources are checked in and out is thus known as the circulation desk.

Circulation rules specify how books and other resources move between borrowers and the library. Examples include the length of time and number of books that maybe borrowed, the criteria for determining if loans can be extended, which resources can and cannot be borrowed and procedures for reserving resources that are currently loaned. These rules are implemented as transactions where all relevant rules must be observed for a loan to take place.

Consider the following:

The decision table below is used as the basis for approving loans at a particular library. Blanks on the rules grid below indicate either a tick ✓ or cross ✗ is possible.

Conditions	*Rules*						
Borrower is a current library member	✓	✗					
Borrower has overdue fines owing	✗		✓				
Borrower has overdue books	✗			✓			
Borrower has reached their item limit	✗				✓		
Resource is reserved	✗					✓	
Resource can be borrowed	✓						✗
Actions							
Loan approved	✓	✗	✗	✗	✗	✗	✗
Loan rejected	✗	✓	✓	✓	✓	✓	✓

An equivalent decision tree is reproduced below:

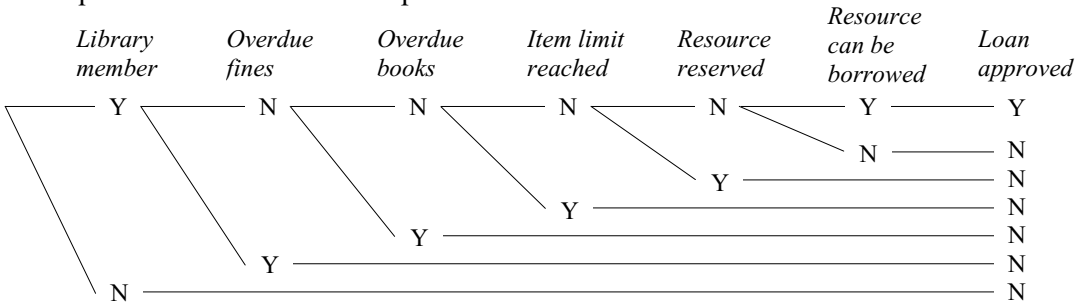


Fig 4.17
Example decision table and tree for approving library loans.

GROUP TASK Discussion
Confirm that the above decision table and tree are indeed equivalent. Is this loan approval process suitable for implementation as a transaction? Discuss with reference to the ACID properties and the ability of the transaction to be rolled back or committed.

Check-out Transaction

The DFD reproduced in *Fig 4.18* models a possible library check-out transaction. This DFD includes processes and data flows to implement the rules described in the decision table (and decision tree) within *Fig 4.16*. Note that the Check book can be borrowed process occurs for each book (or other resource) that a particular member wishes to borrow. The Check member can borrow process occurs once for each check-out transaction, as does the Approve loan process.

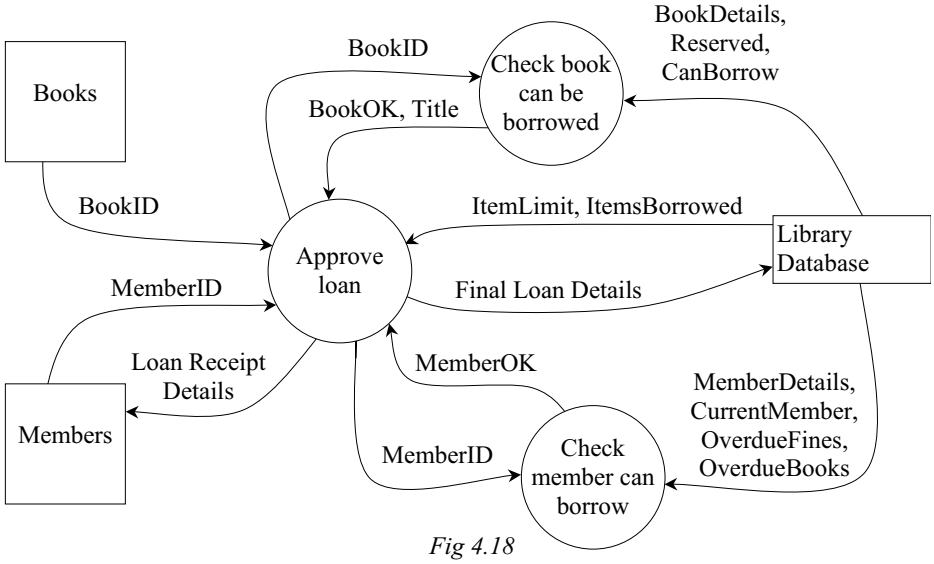


Fig 4.18
Possible DFD for a library check-out transaction.

GROUP TASK Practical Activity
Based on the above DFD, create a step-by-step description of the logical order of processing during a library check-out transaction.

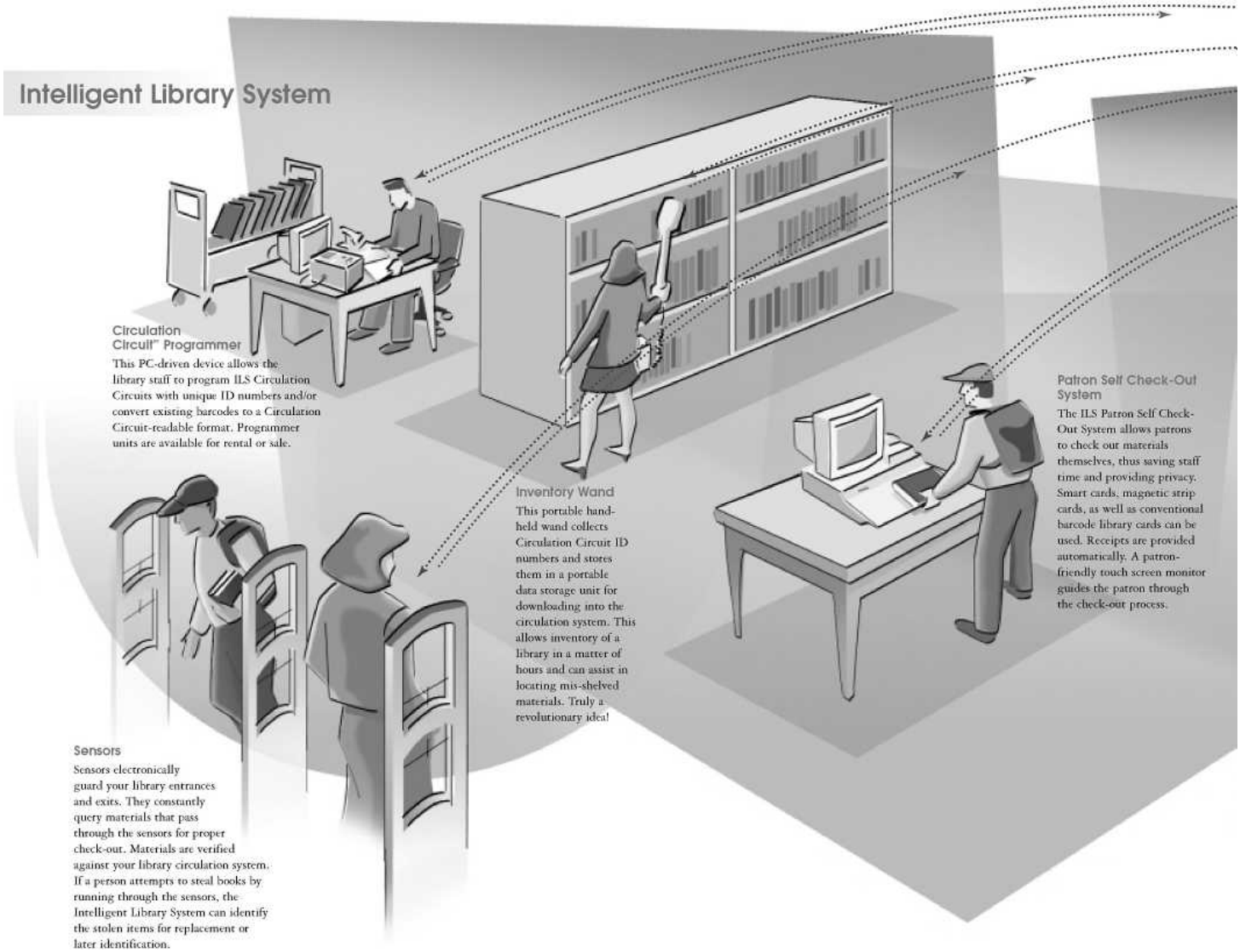
Check-in Transaction

When books are returned or checked into the library the only input data required is the unique identifier for the book. This identifier, say BookID, is sufficient to search the database for the loan record currently associated with that book. The transaction can then update this loan record to record the date the book was returned.

Other processes occur after books have been checked-in. For example the library staff must manually check the condition of returned books as they replace them on the shelves. Any damaged books are repaired and if the damage is excessive then the member may well be expected to pay for repair or replacement. The date returned data is examined when generating overdue notices and overdue fines – these are commonly batch processes performed every few days.

GROUP TASK Discussion
Propose a list of likely information that could be generated from the data created during check-out and check-in transactions. Discuss the purpose of each type of information proposed.

Consider the following:



Checkpoint Meto's Intelligent Library System is based on passive radio frequency identification (RFID). Each book contains a passive RFID tag which permanently stores up to 96 bits of data – the data uniquely identifies each book or other resource. Readers are used to receive the data stored on RFID tags. RFID readers transmit at a frequency specific to the antenna withn the library's RFID tags. The passive tags used within this library system do not contain batteries, rather they are powered by the RF energy from readers within range. The tag responds by transmitting its stored ID back to the reader. Therefore in RFID systems, readers and tags are both transmitters and receivers.

Currently passive RFID tags operate within a range up to approximately 10m depending on the power of the specific reader. Active RFID systems utilise larger battery powered RFID tags and are able to transmit their data over much greater distances – up to 100m is typical. RFID systems are currently replacing many electromagnetic (EM) security systems in both libraries, retail stores and warehouses.

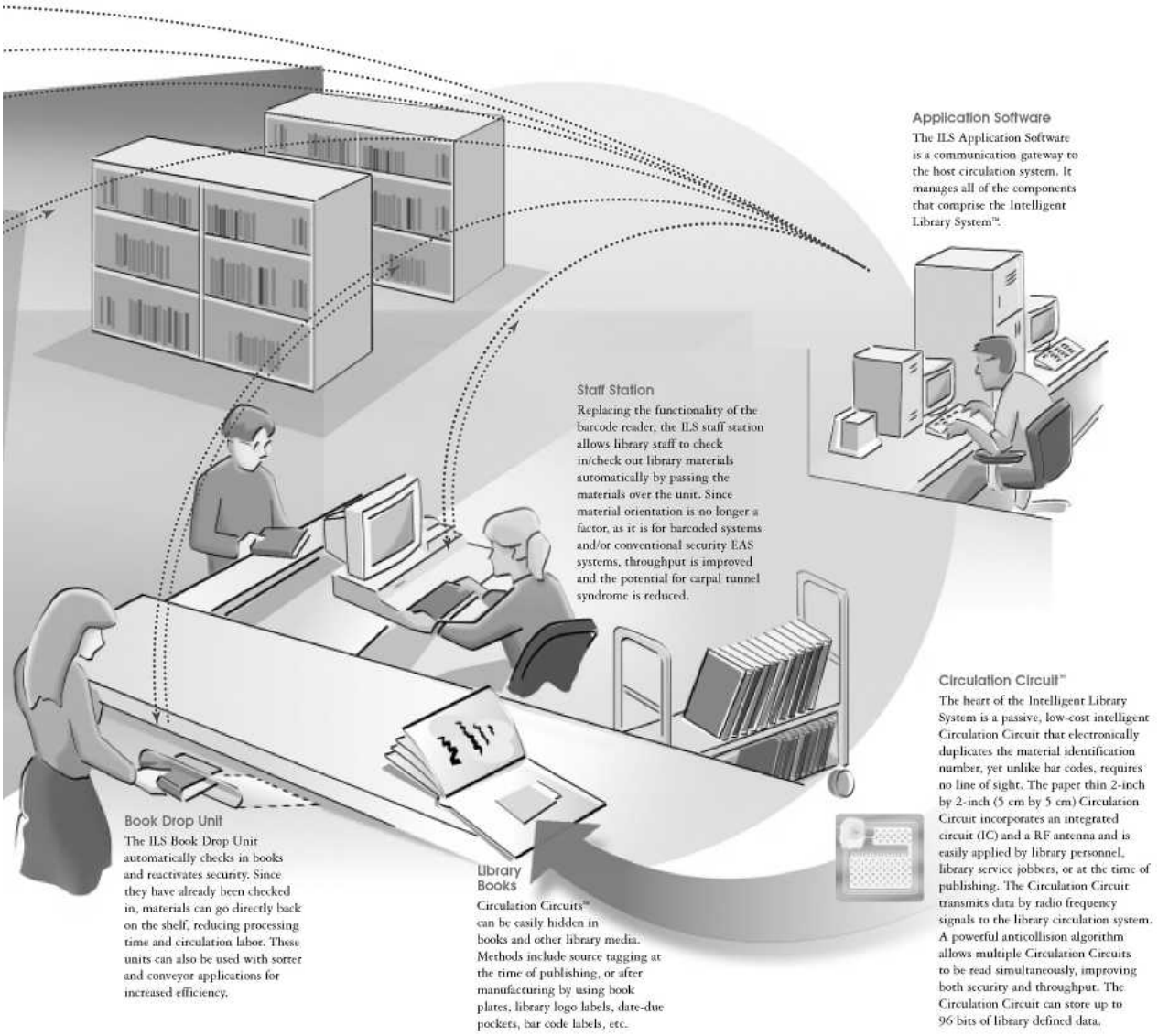


Fig 4.19
Checkpoint Meto's Intelligent Library System
Source: Checkpoint Meto (www.checkpointmeto.com.au)

GROUP TASK Discussion
Identify each of the RFID readers present on the above diagram. Discuss the information processes that use data from each of these RFID readers.

GROUP TASK Research
Research the operation of electromagnetic (EM) security systems. Then compare and contrast EM systems with RFID systems.

GROUP TASK Discussion
Compare and contrast the above RFID collection system with traditional manual and barcode collection systems.

Data/Information

Consider the check-out and check-in transactions described above. The data described in the *Fig 4.20* data dictionary below is either stored within the library database or is generated from data within the library database.

Name	*Data Type*	*Description*
BookID	Integer	PK for each library resource
Book Details	Record	Various attributes including ISBN, Title, Author, Publisher, etc.
Reserved	Boolean	Is the resource currently reserved.
CanBorrow	Boolean	Can this resource be borrowed (True) or is it for in library use only (False).
BookOK	Boolean	True if this resource can be borrowed, otherwise false.
MemberID	Integer	PK and membership number for each library member
MemberDetails	Record	Various attributes including member name, address, phone, membership status and other details.
CurrentMember	Boolean	True if membership exists and is current, otherwise false.
OverdueFines	Currency	Dollar amount of each overdue fine for member.
OverdueBooks	Text	Details of each currently overdue resource.
MemberOK	Boolean	True if the member is able to borrow more resources, otherwise false.
ItemLimit	Integer	Maximum number of resources a member can borrow simultaneously.
ItemsBorrowed	Integer	The number of resources a member currently has on loan.
FinalLoanDetails	Records	Various attributes for a new loan. Includes the MemberID and BookIDs for each resource, together with the date borrowed.
Loan Receipt Details	Various	List of book titles borrowed together with the date borrowed and due date.
ReturnDate	Date	Date each resource is returned or checked-in.

Fig 4.20
Data dictionary for data used by the library check-out and check-in transactions.

Back in chapter 2 we examined the design of a relational database for a library system, refer *Fig 2.17* on page 133. The general nature of this schema meets most of the requirements for our current check-out and check-in transactions.

GROUP TASK Discussion
Identify the data collected, the information produced and the participants involved in the check-out and check-in transactions.

GROUP TASK Discussion
Identify data used by the check-out transaction (*Fig 4.18* DFD) that is NOT stored within the library database.

GROUP TASK Discussion
Modify the schema developed back on page 133 *Fig 2.17* to meet the needs of the check-out and check-in transactions described above.

Information Technology

The majority of library systems store their data within a relational database managed by a database server running DBMS software. This library database is accessed by all users, including library staff, library members and commonly by remote users via the Internet. Clearly the security of this data is critical to the continued operation of all libraries, hence fault tolerant hardware and regular and thorough backup processes are required. The large quantity of data generally requires the use of automated tape backups where tapes are stored securely off site.

A variety of client machines running dedicated applications that perform specific tasks are used. Most libraries have a number of machines dedicated to searching the catalogue (and often the catalogue of other libraries). The library staff perform circulation and also a large variety of other transactions, hence their machines run fully featured versions of library applications.

Staff machines require access to printers, collection devices such as RFID readers and barcode scanners. When new books are purchased it is common for catalogue records to be obtained electronically from outside sources – such as the Australian National Bibliographic Database (ANBD). The ANBD not only provides catalogue data to libraries, it also maintains records of which libraries have copies of particular titles. Today catalogue data is downloaded from the ANBD via the Internet, therefore an Internet connection is required for library staff machines.

GROUP TASK Activity
Analyse the hardware and software present in your school's library system. Identify examples of real time transactions performed within this system.

HSC style question:

SuperBook is an Internet based service that allows customers to make bookings and pay for tickets to major music and sporting events.

When visiting the website customers choose an event, view the currently available seats, choose their desired seats and finally purchase tickets. It is critical that the displayed available seats are wherever possible one hundred percent correct.

In relation to the SuperBook service:

(a) Identify the required information technology.

(b) Analyse the SuperBook service in terms of maximising data integrity.

(c) Construct a data flow diagram for the SuperBook service that describes the data movements between customers, processes and the SuperBook database. Your data flow diagram should include the following processes:
   - Choose Event
   - Display Available Seats
   - Choose Desired Seats
   - Purchase Tickets

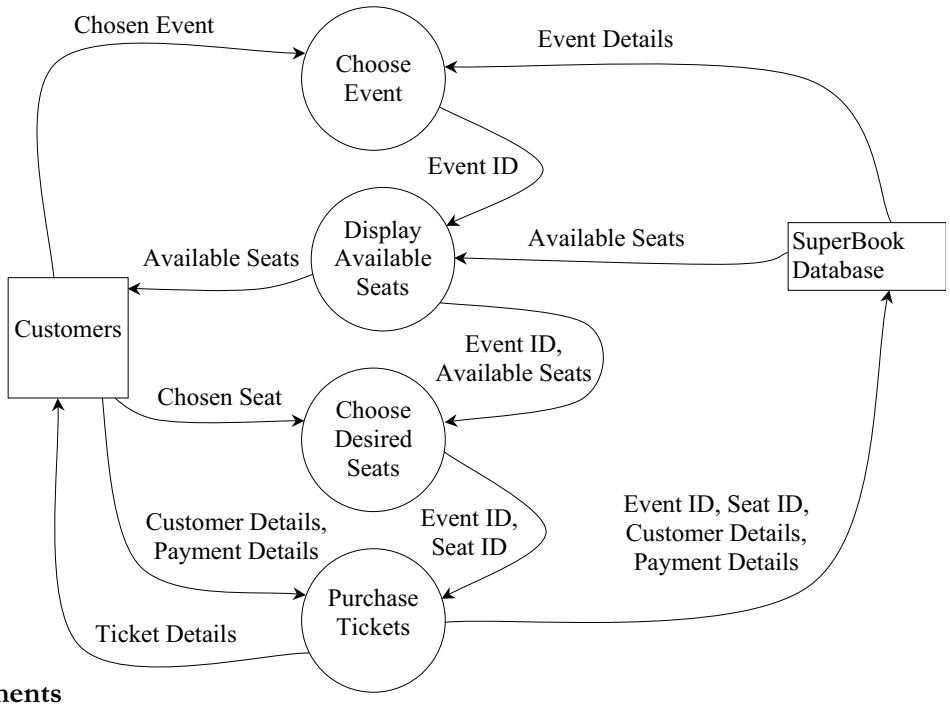
Suggested Solution

(a) Information technology includes:
   - Database server running DBMS server software to access the databases containing the seating and bookings for each event.
   - The server should include redundant mirrored hard disks to ensure fault tolerance.
   - Web server that creates and transmits web pages to each user based on their selections. The data required to create each page is retrieved from the database server.
   - The web server includes encryption software so that payment details are secured during transmission and also once stored.
   - The customer requires a machine with web browser and Internet connection.

(b) To maximise data integrity:

- A transaction should be used that commences when the customer selects a seat or seats and ends when the payment has been confirmed. This ensures the same seat cannot be booked by multiple customers because no other transactions can access the seat unless or until the transaction fails. Furthermore the final booking is guaranteed to be stored permanently due to the durability property of transactions.
- It is impractical to lock all records related to all available seats, hence once a customer selects their seats it is necessary to validate that they indeed remain available as another customer may have selected the same seat after display of available seats. Records for selected seats can then be locked.
- Validation of customer details and payment details is also required to ensure the accuracy or integrity of the data. The payment details are validated during the payment approval operation. The customer details should be displayed to the customer so that they can verify they are correct. Some sites require customers to respond to an automatically generated email – this ensures that at least the email address is correct.

(c)



Comments

- In an actual trial or HSC examination part (a) and (b) would be worth 2 or 3 marks each, part (c) would be allocated 4 or 5 marks. Therefore a total of between 8 and 11 marks would be allocated to this question.
- Answers to parts (b) should specifically address the requirement in the question that available seats are wherever possible one hundred percent correct.
- In part (a) the keyword "identify" requires recognising and naming the information technology likely to be present, however unless absolutely obvious it is worth including a brief justification for inclusion of each item.
- In part (c) no data is written to the database until all processes have completed. This occurs as the transaction is committed. The transaction log and associated data flows could have been included on the DFD, however given the specifics of the question this is unlikely to be required to gain full marks.

SET 4B

1. Generally real time transaction processing requires:
   (A) fast direct access to storage.
   (B) secure communication channels.
   (C) more processing power than batch systems.
   (D) All of the above.
2. On the *Fig 4.7* context diagram (page 382), which of the following is true?
   (A) Each external entity is just a source.
   (B) Each external entity is just a sink.
   (C) Each external entity is both a source and a sink.
   (D) There is one external entity and two processes.
3. Large buttons are preferred on user interfaces for which device?
   (A) Touch screen
   (B) LCD screen
   (C) CRT monitor
   (D) Printers.
4. Consider the schema in *Fig 4.10* on page 386. In addition to the primary key, which attributes in the availability table are populated during the reservation transaction described in the text?
   (A) GuestID, StayDate, RoomID
   (B) GuestID, StayDate, RoomTypeID
   (C) GuestID, RoomID, RoomTypeID
   (D) StayDate, RoomID, RoomTypeID
5. UPCs are often printed as barcodes on the packaging of products. The purpose of UPCs is to:
   (A) identify different products uniquely.
   (B) identify individual items uniquely.
   (C) encode the price of each product.
   (D) improve product security.
6. Most current POS terminals can be best described as:
   (A) dumb terminals that perform only collecting and displaying processes.
   (B) personal computers with specialised collection and display devices.
   (C) a combination of specialised collection and display devices.
   (D) intelligent terminals that perform minimal processing such as data validation.
7. DFDs produced from a context diagram should always include:
   (A) identical data flows entering and leaving the system.
   (B) the same number of processes.
   (C) at least one data store.
   (D) the external entities that are shown on the context diagram.
8. Consider the initial schema for a simple POS system shown in *Fig 4.14* on page 390. When a purchase order is created that includes 4 different products, which of the following is always true?
   (A) 1 record is created in the PurchaseOrders table.
   (B) 4 records are created in the PurchaseOrders table.
   (C) 4 records are created in the PurchaseOrders and Products tables.
   (D) 4 records are created in the PurchaseOrders and Products tables and 1 record in the Suppliers table.
9. On the DFD in *Fig 4.18* on page 393, which of the following best explains why there are two "Check…" processes feeding data to the "Approve Loan" process?
   (A) There are only two decisions to be made prior to approving a loan.
   (B) The DFD would become too complex if all required decisions where detailed on the DFD.
   (C) One process executes for each book and the other executes once for each loan to check the borrower is OK.
   (D) DFDs should not model the intricate details of all processing.
10. Which of the following CANNOT be controlled when collecting data over the web?
    (A) Speed of data access from server secondary storage devices.
    (B) Speed and reliability of Internet connections.
    (C) The number of concurrent users that can be supported.
    (D) The isolation transaction property as many users can read and alter the same data simultaneously.
11. With reference to the DFD in *Fig 4.13* on page 389, construct a lower level DFD to model the Sales process.
12. Outline reasons why many POS terminals use touch screens in preference to keyboards.
13. Visit a local supermarket, hardware or department store. Identify the information technology and participants within the TPS.
14. Analyse your school or local library's check-out transactions. Determine the information technology and describe the operations performed during a typical check-out transaction.
15. Examine a web-based reservation system for an airline or rental car company. Construct a data dictionary detailing the data collected. Include a column explaining the purpose of each data item.

BATCH TRANSACTION PROCESSING SYSTEMS

Batch transaction processing separates data collection from the actual transaction processing operations. This allows transaction processing to be delayed, perhaps to a time when the system is quiet, such as late at night, or perhaps until sufficient data exists, such as generating bills in batches of 100, for example. In all cases the data required for many individual transactions of the same type is collected over time and then at some later time all transactions are processed together without any user interaction. Common examples of batch processing include clearance of cheques, bulk generation of bills and payments, and payroll processing. Batch processing is common when the required data is collected on paper forms.

Commonly the collected data for each individual transaction is added to a transaction file. This can occur over time as data is collected or it can occur just prior to batch processing commencing. When the time comes for batch processing to occur the transaction file is submitted for processing. This file provides all the input data required to process all transactions. During batch processing each transaction executes in much the same way as a single real time transaction. Successful transactions are committed (permanently written) to the master file or database. Transaction problems and errors result in rollback of the transaction. Details of the problems that caused rollback are written to an error file (or log). Writing details to a file allows further transactions within the batch to continue without user interaction. The systems flowchart in *Fig 4.21* describes the typical processes occurring during batch transaction processing.

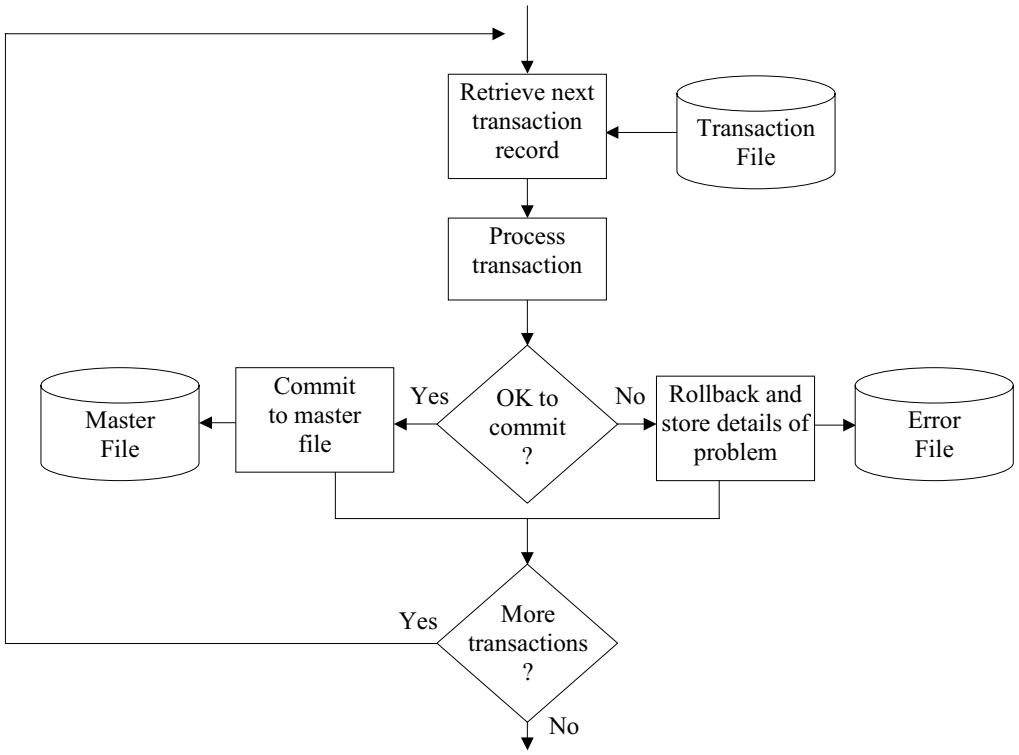


Fig 4.21
Systems flowchart modelling typical batch transaction processes.

GROUP TASK Discussion
Compare and contrast real time and batch processing systems with regard to how transaction problems and errors are resolved.

Historically batch processing was the first type of transaction processing. In the early days of computers all input was via punch cards – this included the actual program code as well as the data to be processed. Each card was manually punched by an operator in preparation for input (see *Fig 4.22*). Completed stacks of punch cards were physically loaded into the computer and processed sequentially. In these early days online real time processing of multiple transactions was simply not possible. The hardware performed a single task at a time and the output was stored sequentially on magnetic tape. As a consequence problems associated with multiple transactions accessing the same data simultaneously did not exist – that is, the isolation ACID transaction property was simply not an issue. Furthermore the processing resources were limited and also costly, therefore batch jobs were scheduled to maximise the use and efficiency of precious processing resources.

Fig 4.22
Operators using key punch machines to create punch cards for batch processing in the 1960s.

Today batch processes are generally performed in parallel with other processes. As a consequence ACID properties must be observed during most batch jobs, including the isolation property. Consider the scenario where a number of different organisations in various locations are processing transactions that access the same data. For example the same credit card number may form part of batch transactions in locations in many different countries. If these transactions happen to overlap then without the ACID isolation property data integrity problems will result.

GROUP TASK Discussion
In old systems each transaction in a batch is performed sequentially. Are any of the ACID properties required within such systems? Discuss.

The processing resources of all computer systems are limited, therefore batch processing even today is scheduled to ensure that each set of batch transactions can complete in a timely fashion. This means many batch processes are scheduled to occur during evenings or weekends when real time processing requirements are lowest. Such scheduling not only ensures CPU processing resources are available, it also reduces the wait time for transactions as it is less likely that other transactions will be simultaneously requiring access to the same data.

Batch transactions that are restricted to a single organisation can be processed offline. This was the normal situation prior to the widespread use of high-speed communication links between organisations. Consider a company's bill generation, all data originates from one single organisation's database. In this case a static snapshot copy of the database can be used to generate bills. Any sales that occur during the bill generation process are not included until the next batch bill generation process occurs.

User interaction with batch processes is restricted to input prior to the commencement of processing and to deal with problems after batch processing completes. Furthermore employees rather than customers commonly initiate batch processing. As a result the design of user interfaces for batch processing is different – they are designed for rapid entry. Often such screens accept numerically coded input via the keyboard's numeric keypad or a barcode scanner. Screen elements designed specifically for mouse input are avoided and keyboard shortcuts are available.

In this section we examine three examples of common batch transaction processing. We examine the processes occurring to clear presented cheques, the generation of bills or invoices and credit card transactions where they appear to be real time but are often processed in batches.

CHEQUE CLEARANCE

Cheques are a paper-based system that prior to the widespread use of computers were the primary method for exchanging funds over long distances. Since the 1970s credit cards and EFTPOS have steadily and significantly cut the total value and number of cheque transactions. In 1994 the daily value of cheque transactions in Australia was approximately $25 billion, in 2004 the daily value had reduced by four fifths to just $5 billion. Despite this reduction, cheques are still expected to form a significant proportion of total financial transactions for the foreseeable future.

A cheque is essentially a promise by a payer (the person or organisation writing the cheque) to pay the payee (the receiver of the cheque) some amount of money. Cheque clearance processes commence when a payee deposits a cheque at their local bank branch. The purpose of cheque clearance processes is to expedite the secure transfer of funds from the payer's account to the payee's account via the banking and clearance network.

However cheques are promises and promises can and are broken. In the case of cheques this regularly occurs when the payer has insufficient funds in their account to cover the value of the cheque. Other problems can also occur, such as forged, unsigned, illegible, altered, lost and stolen cheques. Cheque clearance processes include safeguards to identify and deal with such problems.

Prior to the late 1990s cheque clearance processes did not involve any electronic communication between banks and cheques took 5 or more working days to clear. Today cheque details are exchanged electronically between banks and most cheques clear into payee's accounts in 3 working days. Compared to totally electronic transfers, such as EFTPOS and direct deposit, 3 working days is an eternity. This reality is a consequence of the manual processing inherent in a paper-based system. The value of the cheque must be determined, commonly using scanners and OCR software or in some cases it is manually entered. Signatures and dates require verification and the actual paper cheques are physically exchanged between banks. Until the 1990s all cheques were physically returned to the payer's branch for clearance. Today banks operate their own central facilities that perform cheque clearance processes for many branches.

Financial transaction clearance procedures in Australia are legislated by Government and controlled by the Australian Payments Clearing Association Ltd. (APCA). All major banking and financial institutions, including the Reserve Bank, are APCA members. APCA operates a number of clearance systems, including the Australian Paper Clearance System (APCS) used primarily for the clearance of cheques.

Fig 4.23
MagTek's Mini-MICR cheque reader.

Typical steps for cheque clearance in Australia include:

1. Payee Fred receives a cheque in the mail from payer Freda (who has an account with DEF bank). Fred deposits the cheque into his account at his local branch of ABC bank.
2. In some cases the teller immediately passes the cheque through a MICR (Magnetic Ink Character Recognition) reader to determine the payer's BSB (Bank State Branch) number and account number and the teller

manually enters the value of the cheque. When this occurs the funds are immediately credited to Fred's account as unavailable funds. Usually these funds immediately begin to accumulate interest in Fred's account. More commonly the cheques, together with the deposit slip, are simply filed for later batch processing.

3. During the afternoon all cheques deposited at local branches of ABC bank are physically transported to a central outwards processing facility operated by ABC bank. Some smaller banks share such facilities with larger banks.
4. At ABC bank's outwards processing facility high speed MICR (Magnetic Ink Character Recognition) readers read payer BSB (Bank State Branch) numbers and account details from each cheque. Scanners automatically determine the value of the cheque and the details on deposit slips. Each cheque is encoded with its own unique ID so it can be traced should it later be dishonoured or stopped. Most banks also print the cheque value on the cheque using MICR printers. Payee accounts are credited with funds if this has not already occurred at the branch. Based on the BSB numbers, cheques are automatically sorted into bundles destined for different banks together with the total value of each bundle. *Fig 4.24* shows IBM's 3890 sorter which is able to read MICR characters (*Fig 4.25*) and sort up to 2400 cheques per minute. Note that completion of this batch process provides electronic records of all cheques deposited into payee accounts operated by the bank.

Fig 4.24
IBM 3890 high speed cheque sorter includes MICR reader and optional scanner.

5. Each bundle of cheques is transported to a central check clearing house operated by APCS. Appointed representatives of all banks exchange bundles of cheques. In addition the net difference between exchanged bundles is calculated. For example the representative from ABC bank may hand the DEF bank representative cheques totalling $2.2 million, whilst DEF bank hands ABC bank bundles of cheques totalling $2.5 million. In this case the net difference of $300,000 is transferred from ABC bank to DEF bank. At this stage all cheques are now under the control of the payer's bank. In our example Freda's cheque is now in the hands of her bank – DEF bank.

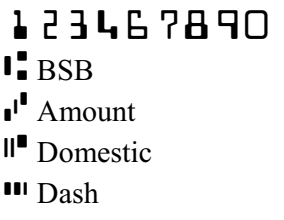


Fig 4.25
Standard MICR characters.

6. Bundles of cheques are now physically transported to the central inwards processing facility of each bank – Freda's cheque goes to DEF bank's inwards facility. Currently most facilities are within major cities such as Sydney and Melbourne. The cheques commence being batch processed. Each cheque again passes through a MICR reader and scanner. The scanner determines the value of each cheque whilst the MICR reader determines the account. For each cheque, the system ensures there are sufficient funds in the payer's account, verifies the authenticity of the cheque and debits the value of each cheque from the payer's account. Problem cheques are diverted for manual examination. Cheques where there are insufficient funds or other problems are dishonoured. The ID encoded by the payee bank is used to determine and inform the payee's bank of such problems.

In the past cheques were sorted into individual branch bundles and physically transported to branches for final batch processing. Today account details and images of account holder signatures are available online, therefore verification can now take place centrally via secure communication links. It is the removal of the need to physically transport cheques back to their branch of origin that has reduced clearance times from 5 days to the current 3 days.

GROUP TASK Discussion
Each cheque passes through two distinct batch processes. Identify and describe the operations performed during these batch transactions.

GROUP TASK Discussion
Identify the information technology, data/information and participants involved in the cheque clearance process.

GROUP TASK Research
Many countries, including Australia, are analysing systems that digitise images of entire cheques – in the banking industry this is known as "cheque truncation". Some countries have already implemented such systems. Research advantages of such cheque truncation systems

BILL GENERATION

In many systems the generation of bills or invoices is well suited to batch processing. When orders for products or records of services provided are already within the system then no extra data collection is required prior to generating invoices. No user interaction is needed and multiple invoices are usually generated at the same time. Often bills are generated during times when the resources of the system are not being used – commonly during the night. Consider telephone, electricity, gas, rates and other regular household bills. The data exists within the organisation's database and therefore batch processing can be used to generate bills.

Even small businesses that process small numbers of orders each day use batch processing. The orders are entered as they are received throughout the day and then in the afternoon all the day's invoices are printed as a batch job. The orders are packed manually using details from the printed invoices. Each order is then dispatched together with the invoice. The invoicing database schema we produced in chapter 2 when describing normalisation is typical of such a system (refer *Fig 2.70* on page 185). This database would be queried to return all invoice details for the current day. This query is then used as the record source for a report that generates and prints the day's invoices.

GROUP TASK Discussion
Analyse the Invoicing database created in chapter 2 to determine the data and processes required to generate invoices for the current day.

Apart from the relatively static product details and prices the data required to generate each invoice is largely independent of the data on all other invoices. This data independence means that invoices can be generated in any desired order and more significantly multiple invoices can be generated simultaneously. This characteristic is particularly significant for large systems that generate many thousands of bills. To generate say monthly telephone bills requires reading each customer's address details and records of all the calls made within the billing period. The batch process does not need to access, update or create data in any other system. Also during processing no

data is updated or created within the telephone company's database that is accessed during the generation of any other customer's phone bill. This processing independence means parallel processing can be used to drastically reduce the total processing time required.

Consider the following:

For large batch systems, where many thousands of bills are generated in a single job, it is common to make a snapshot copy of the live data. This snapshot is an offline copy of the actual data as it was at the end of the billing period – maybe the end of a year, quarter, month, week or even the end of a day. The online version of the database continues to operate without its performance being degraded by the batch processes, and the batch processes are not interrupted by the online processes. Because of the independence of the data and processing the snapshot copy can be split into different parts that are physically stored and batch processed in parallel on different storage devices and using different CPUs (refer *Fig 4.26*). This distributed processing strategy reduces processing time significantly – if a batch job takes two hours to complete on one machine then it will take approximately half this time if two machines are used. Such systems use high-speed digital printers that link with automatic folding and envelope insertion devices.

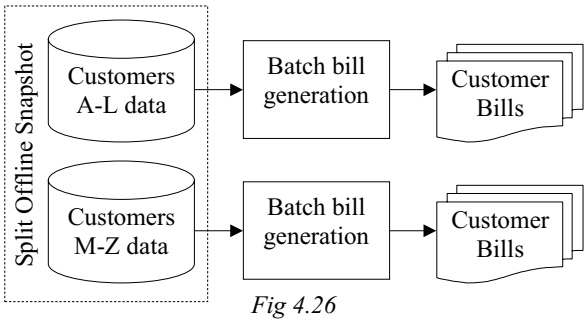


Fig 4.26
Parallel batch processing using a split offline snapshot copy of the online database.

The time taken to retrieve records from secondary storage is a significant limiting factor in terms of improving the performance of large systems. Splitting and physically storing parts of the database on different storage devices is one technique that improves performance. Another significant technique used by large batch systems is sequentially accessing data. This involves accessing data in the order in which it is physically stored on the disk. In the past, data was physically stored sequentially on tape and hence it was either impossible or extremely inefficient to read the data randomly – rewinding and fast-forwarding the tape takes time. Sequential access was a necessity rather than a choice if jobs were to complete in a reasonable amount of time – even for small batch jobs. Today the read/write heads within hard disks are able to quickly jump directly to required records, however this still takes time. Furthermore hard disks read data in complete sectors and commonly multiple adjacent sectors are also read. This data is stored in the drive controller's cache. If random access is used then much of the data that is physically retrieved is not actually processed. If sequential access is used then all data retrieved is subsequently processed and the movement of read/write heads is minimised. Note that significant performance gains are only possible when the transactions processed are independent of each other and the data they access is physically stored sequentially. Transaction processes that use retrieved data as the criteria for searches and that write data require careful analysis and design if the advantages of sequential access are to be maximised. For instance, the order in which processes are performed can be significant or it may be more efficient to remove an operation from a transaction and perform it separately on all the data. For independent processes, such as those required for bill generation, the ACID properties can be relaxed somewhat in order to improve performance.

GROUP TASK Discussion
Many people now receive bills online via the web, in general these bills are still the result of batch processing. Why not generate these bills online and in real time? Discuss.

CREDIT CARD TRANSACTIONS (REAL TIME OR BATCH?)

From the customer's perspective credit card transactions appear to be processed in real time. The customer presents their card to the sales assistant who then swipes the card through an EFTPOS terminal and has the customer sign a receipt. If the signature on the card matches the signature on the receipt then, from the customer's perspective, the transaction is complete. On the web, customers enter their credit card details into a secure web page and then within seconds the transaction is approved – it appears the entire transaction is complete. In reality many credit card transactions are batch processed during the evening following the sale, whilst some are actually completed in real time. Note that a credit card transaction is not complete until the funds have moved into the merchant's account.

All credit card transactions involve at least four significant parties. Customers who are credit card holders, merchants who are generally retailers, card issuers who manage the customer side of credit card transactions and acquirers who manage the merchant side of transactions. Most acquirers and issuers are banks who share the expense of operating the network and technology between issuers and acquirers predominantly via the MasterCard and Visa systems. Let us consider the general sequence of events that occurs to process credit card transactions (refer to *Fig 4.27*):

① Customer gives merchant permission to access credit in their account to pay for goods or services. For "card present" transactions handing over the card and signing verify that permission has been given. For "card not present" transactions, such as telephone and mail order, the verbal or written order and credit card details are sufficient verification of permission.

② Merchant creates and transmits transaction details manually or electronically to their acquirer. This can occur via an EFTPOS terminal, manually by written voucher or over the Internet via a payment gateway. For larger value manual transactions approval (steps ③, ④ and ⑤) are performed over the phone.

③ The acquirer receives the transaction and determines the card issuer. The transaction details are then forwarded electronically to the card issuer.

④ The card issuer checks the customer has sufficient credit remaining to cover the transaction and reserves these funds. An authorisation code is sent back to the acquirer.

⑤ The acquirer receives the authorisation code and electronically forwards it back to the merchant. On EFTPOS terminals the word "APPROVED" is commonly displayed. For manual transactions approval is given over the phone.

⑥ Merchant receives approval message generates a receipt and hands it to customer. If customer is present then they first sign the merchant's copy of the receipt and the merchant verifies the signature against the signature on the card. The receipt includes a

Fig 4.27 Communication during a credit card transaction.

unique number that identifies the transaction within the merchant, acquirer and issuer's systems.

⑦ The card issuer transfers the funds out of their account and forwards the funds to the acquirer. Often many transactions are batch processed together hence a single large transfer takes place together with details of individual transactions.

⑧ The acquirer deposits the value of each of their merchant's transactions into each merchant's account. In most cases this occurs each evening to finalise the days transactions.

Consider the following:

The above sequence of steps occurs for all credit card transactions however there are many different systems that perform these steps at different times and perform some or all of the steps as batch processes. In some cases other organisations are involved that relay data between merchants and acquirers or to perform processing on behalf of merchants.

Let us consider some typical examples and highlight when real time processing and when batch processing is used:

- Retail EFTPOS terminals supplied and connected directly to a particular bank use a combination of real time and batch processing for credit card transactions. When a customer's credit card is swiped the terminal communicates with the bank (acquirer) via a telephone line to authorise the transaction in real time. The bank transmits a reference retrieval number (RRN) back to the EFTPOS terminal and the terminal displays "APPROVED PLEASE SIGN". The customer signs the receipt and the retailer verifies the signatures on the card and receipt match. If the signatures do not match then the transaction is reversed – this reversal is another transaction sent to the acquirer.

  At the close of business each day the EFTPOS terminal "settles" with the acquirer bank. The settlement process transmits details of all transactions to the bank. The bank then batch processes all transactions during the evening resulting in the funds (less any bank charges) being deposited into the retailer's merchant account.

- Most retailers now use EFTPOS terminals for their credit card transactions as described above, however manual systems are still available as a fall back should the EFTPOS terminal or link to the bank fail. Using a manual system the retailer manually takes an impression of the customer's card on a voucher. The voucher is manually completed by the retailer and then signed by the customer Each retailer has a floor limit. If the total value of the transaction is above the floor limit then the retailer telephones the bank for manual authorisation. If authorised the bank reads out an authorisation number, which is manually written on the voucher. Each voucher includes the original, which is later submitted to the bank, a copy for the customer and a copy for the merchant.

  At the close of business the retailer completes a merchant voucher that includes the total number and value of all vouchers. The merchant summary together with the original of all vouchers is then deposited at the retailer's local bank branch (acquirer). The vouchers are batch processed by the bank during the evening.

GROUP TASK Discussion
Why do you think the batch settlement process is performed, why not simply complete each transaction in real time as it occurs? Discuss.

- Some retailers are authorised by their bank to accept mail, phone or fax credit card orders. These are known as MOTO (mail order telephone order) merchant accounts. Banks scrutinise retailers more thoroughly to verify that they are trustworthy and honest before MOTO merchant accounts are approved. Once approved the retailer is able to initiate credit card transactions without the card actually being present – just the credit card number and expiry date is required. The details of the transactions are manually entered into the EFTPOS terminal or can be manually written onto a voucher. As less information is available to verify each transaction the retailer must agree to accept a higher level of risk should transactions be disputed. The transactions are processed similarly to above, however banks often charge higher rates compared to "card present" transactions.
- Internet credit card transactions for large volume businesses are usually processed in real time. Commonly the merchant's website collects details of the purchase, such as products and prices. The website then directs customers to a payment gateway which completes the actual financial transaction such that the funds are moved immediately from the customer's account into the merchant's account. This transfer involves both the authorisation and funds transfer steps occurring simultaneously and immediately.

GROUP TASK Discussion
Banks view Internet credit card transactions as high risk. Propose reasons why this is the case? Does real time processing reduce the risk? Discuss.

- Other Internet credit card transactions, particularly for smaller businesses, are actually processed manually using the retailer's existing EFTPOS terminal and MOTO merchant account. The credit card details and the details of the purchase are transmitted securely to the merchant without any interaction with banks. The merchant then initiates the transaction manually via their EFTPOS terminal. Such transactions are settled, along with any in-store purchases, during the evening using batch processes.
- Businesses that charge customers on a regular basis use batch processing. In this case the business creates a file containing the details of multiple transactions. This file is uploaded to the merchant's acquirer bank where it is batch processed during the evening. The business must hold an authority from each customer to perform each transaction. Such batch systems are used for purchases that require regular payments, for example topping up toll card accounts, making loan repayments and for payment of telephone, electricity, rates and other regular bills.

GROUP TASK Discussion
The above system does not use real time processing at all. The transactions are entirely batch processed. Discuss advantages for the customers, merchants, acquirers and issuers.

- Private companies now provide EFTPOS services and dedicated terminals to retailers. Often these EFTPOS terminals connect via the Internet and are operated by or connected to Internet payment gateways. Transactions performed on such terminals are processed in real time in much the same way as web transactions performed via payment gateways.
- Companies such as PayPal offer credit card processing services that do not require merchants to have their own merchant account with an acquirer. Rather the company uses their own merchant account and acquirer to process transactions on

behalf of other merchants. These systems generally cost more per transaction and hence are used by individuals and businesses that process credit card transactions infrequently.

GROUP TASK Discussion
Analyse each of the above systems and identify where real time processing is being used and where batch processing is being used. Discuss the appropriateness of each type of processing for the given system.

HSC style question:

BigBizzCorp is a medium sized business which uses a traditional batch payroll system to produce weekly payslips for each of its 200 employees who work in one of 10 departments.

Each day when the employees come into work, they 'clock on' by locating their employee time card and punching into a special clocking system, which prints the current time on their time card in today's position. At the end of the day, the employee punches their time card again to allow it to print the time they have just finished for the day.

At the end of each week, the paymaster collects these 200 time cards, and enters the start and end times for each day for each of the employees into the Payroll system.

When the weekly payroll is run, a single payslip is produced for each employee showing their hours worked for this week together with their pay, taxation and superannuation details. An overall summary of the weekly payroll is also produced for use by management in their budgetary processes.

(a) The data entry screen for entering each employee's start and end times into the batch payroll system is reproduced below:

Weekly Payroll for the week ending ../../..

Department: --

Employee number: ------ Employee Name: ----------

	Start Time	End Time	Total Hours
Monday:	--:--	--:--	--:--
Tuesday:	--:--	--:--	--:--
Wednesday:	--:--	--:--	--:--
Thursday:	--:--	--:--	--:--
Friday:	--:--	--:--	--:--

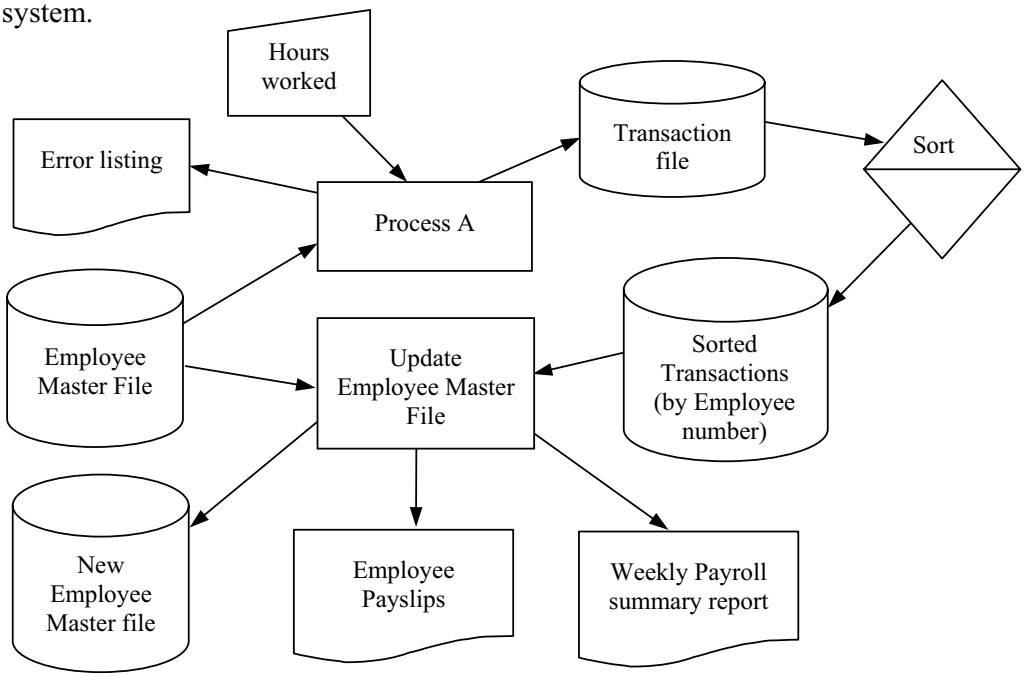
Total Hours for the week: --:--

Done Next

(i) Identify fields on the above screen where data would be entered directly by the paymaster. Explain how the remaining fields would be populated during data entry.

(ii) Propose suitable validation processes that could be performed on the data entered through this screen. Justify your responses.

(b) The systems flowchart originally created during the development of the above batch payroll system is reproduced below. The flowchart diagrammatically represents the steps performed by BigBizzCorp's batch payroll processing system.



(i) Explain the processing likely to be occurring within Process A. Refer to the output produced, including the error listing, as part of your response.

(ii) Describe the method of data access being used each time a file is read from or written to within the above system.

(c) Analyse the strengths and weaknesses of the current batch system and assess the effects of altering this system to a real-time system.

Suggested Solution

(a) (i) The paymaster enters the week ending date and then enters just the start and end time for each day for each employee. The department, employee number and employee name being populated sequentially from the Employee Master File – the data entry process progressing to the next employee each time the Next button is selected.

The Total Hours are generated in real time once each pair of start/end times has been entered. Similarly the Total Hours for the week would be calculated by summing the Total Hours fields – this field being updated as each day's times are entered.

(ii) Fields to be validated include:

- Week Ending – the date must be a valid date, less than today's date. If data is entered incorrectly, the management summary report will have the incorrect date on it.

- Department number – must be a valid department within the company (eg between 1 – 10). If data is entered incorrectly, the management summary report will bill the payroll to an inappropriate or non-existent department.
- Employee number - must be a valid employee number within the company (eg between 1 – 200). If data is entered incorrectly then during the update process, the pay and hours will be accredited to the incorrect employee, or there will be no such employee to match to the transaction, and the update will fail.
- Each day's start and end times – must be a valid time, during valid working hours of business and with end time greater than start time. If data is entered incorrectly, the calculated hours worked on that day will be incorrect and the pay will not be correctly calculated. A common error is to transpose the start and end times, and the validation should check for this.
- Total hours for the week – must calculate to be within a reasonable expected range (say 0 – 60 hours). Users could be prompted to check or re-enter times if the total exceeds the expected maximum.

(b) (i) Process A occurs each time a set of start and end times is entered for each employee record retrieved from the Employee Master File. First validation is performed on each field. If the transaction is correct then it is written to the transaction file – the data would likely include the employee number, the day and a pair of start and end times. If an error is encountered then a line is printed – depending on the operation of the UI, errors could include unknown employee numbers, departments or unreasonable daily total hours. Process A occurs for each transaction, hence all correct transactions are written to the transaction file with an error listing being printed of all problem transactions.

(ii) All data access within the system is sequential. Process A reads each employee record from the Employee Master File one after the other. Process A also writes transactions one after the other so that the transaction file contains a sequence of records where each record is comprised of an Employee Number, day, start time and end time. The sort process arranges these into order based on Employee Number. These sorted records being written sequentially to the Sorted Transactions file.

The Update Employee Master File process would read an employee record (from the Employee Master File) and a sequence of transaction records (from the Sorted Transactions file) that match the employee number. These records are joined, a payslip is printed and a new record for that employee is written to the New Employee Master File. This process repeats sequentially for each employee in the Employee Master File.

(c) Possible strengths include:

- Only a single computer is required.
- Only a single direct user is required.
- Data required only at the end of each week so system is well suited to batch processing.
- Simple system that reflects manual processes.
- No real time queries required which suits batch processing.

Possible weaknesses include:

- Dedicated data entry person required.
- Time delay to process payroll.
- No real time queries are possible.

- Difficult to alter to meet new requirements – i.e. Designed for the specific payroll task.

Effects of altering to a real-time system:

- The Employee master file would need to be rewritten as a direct access file. This would allow queries at any time by management as to who has been absent this week, or how many hours have been worked so far.
- It would be possible to allow employees to enter their own times directly through an interactive system that creates a transaction each time an employee logs on to start or end their day. This eliminates the need for a data entry person with associated costs, possible bias or errors in the data entry.
- Validation can be done instantly at the time of data entry by the employee, without the need for a clerk to look back through the transactions and correct them if they are identified as being in error.

Comments

- In part (a) there are various different ways to interpret the operation of the screen. As this is a batch system, perhaps each employee is displayed one after the other and the user has no control over this order – meaning only start and end times are entered directly as in the above answer. Or possibly the employee number is entered which causes that person's name to be displayed ready to enter their start and end times. Or possibly department could be entered so a sequence of employee's within that department is presented.
- In part (a) (i) it is likely that the total hours data, which is calculated from the start and end times, would not be written to the transaction file. This data is calculated on the screen for use by the data entry user – it performs a data verification role.
- In part (a) (ii) the validation processes described could include checking the employee or department exists within the Employee Master File.
- On the systems flowchart a printed error listing is created during data entry of start and end times. Although this is possible, today it is more likely that during data entry such errors would be displayed on the screen.
- As the master file is updated problems can occur. During batch processing these problems are generally directed to an error log (usually a file). The systems flowchart included in the question does not detail how users are informed of errors that may occur during the Update Employee Master File process.
- In (b) (ii) it is possible to interpret the Process A read from the Employee Master File as random access as the data entry screen can be interpreted to be looking up employees one by one based on the user's employee number inputs. If this is the case then the system is not enforcing the employee order as would occur if access were sequential. Furthermore the transactions are sorted prior to further processing. This implies that Process A does not collect, create and then write the transaction records in the order required by the Update Employee Master File process.
- In part (c) there are many other possible strengths and weaknesses of batch systems, and effects of altering to real time processing that could be discussed. Notice the three parts to the question – strengths of batch processing, weaknesses of batch processing, and effects of altering to real time. In a Trial or HSC examination equal marks would likely be allocated to each of these three parts.
- In a Trial or HSC examination Part (a) (i) and (ii) would likely attract 3 marks each. Parts (b) (i) and (ii) would attract 4 marks each and part (c) would attract a total of 6 marks. Therefore this question would form a complete Transaction Processing Systems questions worth a total of 20 marks.

SET 4C

1. In most batch processing systems the transaction file contains:
   (A) the results or changes made to the master file after transaction processing.
   (B) the data required to process transactions.
   (C) a copy of all data that has been altered or added to the master file.
   (D) details of all transactions that have been successfully committed.
2. User interaction with batch processes includes:
   (A) preparing and/or collecting data prior to batch processing commencing.
   (B) correcting errors after batch processing has completed.
   (C) scheduling when batch jobs should be performed.
   (D) All of the above
3. The isolation ACID property can be relaxed when transactions are:
   (A) processed in parallel.
   (B) processed sequentially.
   (C) performed in real time from multiple sources.
   (D) batch processed.
4. Which of the following is the most significant reason why cheque clearance takes considerably longer than EFTPOS or credit card clearance?
   (A) MICR readers are slow compared to magnetic swipe readers.
   (B) Signatures must be manually verified at the point of sale.
   (C) Ensuring sufficient funds are in the payer's account is performed manually.
   (D) Cheque details are collected from paper documents at different locations.
5. During batch processing, errors detected are commonly written to a file rather than displayed on screen. Which of the following is the best reason why this occurs?
   (A) To permanently record details of all errors encountered.
   (B) It allows batch processes to occur when nobody is present.
   (C) So users are freed to complete real time processes.
   (D) To allow processing to continue without interruption.
6. Recurring household bills are particularly well suited to batch generation because:
   (A) such systems include sequential secondary storage devices.
   (B) the data required to generate the bills already exists within the system.
   (C) large companies have staff dedicated to the bill generation process.
   (D) most households pay such bills using direct deposit or credit cards.
7. Which of the following occurs at check clearance houses operated by the APCS?
   (A) Bundles of cheques are exchanged between banks.
   (B) Cheques are scanned to determine their value.
   (C) The value of each cheque is withdrawn from the payer's account.
   (D) Funds are deposited into each payee account.
8. The four significant parties in all credit card transactions are:
   (A) Customers, retailers, banks and Visa or MasterCard.
   (B) Customers, merchants, clearance houses and banks.
   (C) Customers, merchants, acquirers and issuers.
   (D) Payment gateways, merchants, banks and card companies such as Visa and MasterCard.
9. According to banks, which of the following lists credit card transactions in descending order of risk?
   (A) Internet, MOTO, Card Present
   (B) Card Present, MOTO, Internet
   (C) MOTO, Internet, Card Present
   (D) Card Present, Internet, MOTO
10. Which of the following best describes batch processing?
    (A) Collecting occurs over some time and then many transactions are processed together at a later time.
    (B) Transactions are processed soon after the required data has been collected.
    (C) Many similar transactions are processed in parallel.
    (D) Transactions are added to a queue and are processed in the order in which they were received.
11. Recount the steps that occur once a cheque is deposited until the funds can be withdrawn.
12. Construct a diagram to describe the order of processing occurring to complete a typical "card present" credit card transaction.
13. Sometimes ACID properties can be relaxed during batch processing. Discuss using examples.
14. Compare and contrast the general nature of real time and batch transaction processing.
15. Explain why systems that collect transaction data on paper forms are suited to batch processing.

BACKUP AND RECOVERY

Backup is the process of making a copy of data in case the original is lost or damaged. Recovery is the opposite of the backup process where the backup copy of the data is restored and placed back into the system. We introduced backup and recovery in chapter 2 when discussing techniques for securing data. In this section we shall concentrate on various procedures used to perform backups together with advantages and disadvantages.

Backup
To copy files to a separate secondary storage device as a precaution in case the first device fails or data is lost.

Backups provide a snapshot copy of data at particular points in time. Each backup copy allows the system's data to be recovered back to the state it was in at the precise time the backup copy was made. In the event of total system failure, such as a hard disk crash or a fire that destroys the data completely, it is important to be able to recover to a point as close as possible to the time the failure occurred.

The most common reason for total system failure is hard disk failure – in particular components that move the read/write heads. It is a fact that all hard disks will eventually fail. Research (refer *Fig 4.28*) indicates that rates of failure are high with new hard disks, largely due to manufacturing faults. Failure rates are significantly lower for disks that are approximately 1 to 6 years of age. Failure rates then rise again as components begin to wear out.

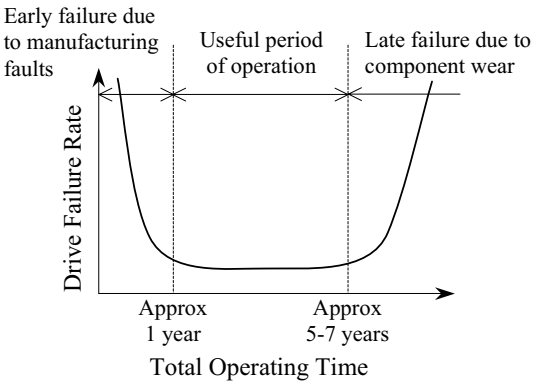


Fig 4.28
Hard disk failure rates over time.

There are many other problems that occur where backup copies made at different times in the past are invaluable. For instance when a user makes changes to a file and later wishes to revert to a previous version. Also viruses are often detected after a period of time has elapsed. In each case having many historical copies of the data allows the system (or a single file) to be restored to a previous state. To recover from the broadest range of possible problems requires backup copies to be made regularly and each backup to be kept for a reasonable amount of time.

GROUP TASK Discussion
Brainstorm lists of possible failures where the most recent backup is most useful and another list where older backups are more useful.

Just how often backups are made and for how long they are kept is dependent on the value and nature of the data. The value of data includes the costs associated with recreating the data, together with the cost of the system being inoperable. Currently recreating 10MB of data is estimated to cost on average about $50,000. Furthermore it is estimated that some 43% of businesses that experience a severe or total loss of data never reopen. Clearly the importance of performing regular backups and ensuring they can be reliably restored is critical.

GROUP TASK Discussion
We have mentioned RAID storage at various times throughout our work. Do mirrored RAID solutions remove the need to make backups? Discuss.

FULL AND PARTIAL BACKUPS

There are three different types of backup that are commonly used within most backup procedures – full backups, incremental backups and differential backups. Both incremental and differential backups involve making partial backups.

Full Backup

As the name suggests, a full backup is a complete copy of all data within the system. This can be a complete image of the entire hard disk(s), including the operating system, program files, configuration settings and of course data. For most transaction processing systems it is the data that is of particular value – the software and configuration settings rarely change and are far easier to restore. Therefore most businesses perform full backups of all their data files on a regular basis.

Full backups are the easiest to restore should failure occur. The full backup is simply copied back into the operational system. Unfortunately copying all files takes a long time and requires large amounts of storage; therefore it is often impractical to perform full backups on a daily basis. Common backup procedures specify that full backups be made on a weekly basis, usually commencing on Friday afternoons and for large systems continuing over the weekend.

Most operating systems store an archive bit along with each file. The archive bit is set to true when changes are made to a file and when a file is first created. When full backups are made all archive bits are set to false indicating that a backup copy of each file has just been made.

Incremental Backup

Incremental backups include making partial backups that copy all files that have changed or been created since the last backup – the last backup may have been a full, or partial backup. An incremental backup therefore includes only those files where the archive bit is true. As a consequence performing an incremental backup is significantly faster and requires significantly less storage compared to a full backup. After each file has been copied its archive bit is set to false. Therefore if incremental backups are performed each afternoon then each incremental backup copies only those files that have been altered or created since the previous afternoon's backup was made.

Before incremental backups can be made it is necessary to first perform a full backup. Commonly faster incremental backups are then made on at least a daily basis. The significant saving in backup time is counteracted by the extra time required to recover the data. During a recover the latest full backup is first restored, then each partial backup is restored in the order in which they were made. Hence files that have changed since the full backup progressively overwrite the older versions as each partial backup is restored.

Differential Backup

A differential backup uses partial backups to makes copies of all files that have been altered or created since the last full backup. If such partial backups are made each day then each will contain copies of all files within all previous partial backups since the last full backup was made. To restore to the most recent backup requires first restoring the full backup and then restoring just the most recent partial backup.

In terms of archive bits, differential backups copy all files where the archive bit is true, however differential backups do not alter any archive bits. Therefore over time one would expect more and more archive bits to be true and hence more and more files are included within subsequent differential backups. The size of the differential backup continues to grow until the next full backup is completed.

TRANSACTION LOGS, MIRRORING AND ROLLBACK

Recall that transaction logs contain historical details of each transaction performed including details of transactions that are currently being processed. These details can be used to restore a transaction processing system back to a consistent state at some precise point in time – completed transactions can be recommitted or rolled back and incomplete transactions can be continued or rolled back. Let us consider some disaster situations where transaction logs, mirroring and rollback assist system recovery.

- Imagine the drive controller on a database server has failed during a busy period – which is when such faults usually occur. At the time of failure numerous transactions were incomplete. To recover from this disaster requires the server to be shut down, a new drive controller installed and the server restarted. The transaction log is then used by the system to automatically rollback all incomplete transactions, which returns the data to its most recent consistent state.

  What about transactions that span multiple servers and systems? In general, most transaction systems automatically abort (rollback) actions that have not been committed after a specified period of time – this deals with most issues. However, most systems include further safeguards to ensure this occurs. If the transaction was initiated by the server that crashed then it sends each system involved in each transaction a message detailing the transaction and the specific actions they should abort. What about actions performed on the crashed server that formed part of a transaction initiated by another system? In this case the server informs the initiating system, who then rolls back the complete transaction.
- A database server is attached to two RAID storage devices. The first RAID device stores the main databases and uses RAID striping to improve data access speeds. The second is a mirrored RAID device and is used to store the transaction logs for all databases. Now suppose a disaster occurs which totally destroys the first RAID device and all but one of the hard disks on the second RAID device. Because the second RAID device was mirrored, the remaining hard drive will contain a complete copy of the current transaction log – I'd copy this to a fresh hard drive!

  In this case recovery first requires installing new RAID devices, installing software and then restoring the data from the most recent set of backup media. We now have consistent data but it is missing all changes made since the last backup. The solution to this problem is to use the transaction log to roll forward and recommit all transactions performed since the last backup.

Consider the following

Backup and recovery protects against each of the following:

- Hardware errors and failure.
- Software errors.
- Physical theft or destruction of hardware.
- Unauthorised or unwanted changes to data, due to viruses or hacking, for example.
- Intentional changes to data that for various reasons need to be reversed.

GROUP TASK Discussion

For each of the above dot points, discuss suitable backup techniques that will protect the data. Consider the use of secure onsite and offsite storage, full and partial backups, mirroring and the use of transaction logs.

BACKUP MEDIA

Magnetic tape remains the dominant media for backing up data on large systems, including most transaction processing systems. Other forms of backup media include hard disks, CDs and DVDs. Compared to magnetic tape, the limited capacity, lower data transfer speed and higher cost of these alternatives makes them unviable alternatives for backup of most large systems. Currently online businesses are emerging where backups can be made over the Internet. Some large organisations maintain their own dedicated high-speed communication links to remote backup sites.

Magnetic Tape

Magnetic tape is a sequential access media contained within cassettes or cartridges and is currently the most convenient and cost effective media for backup of large quantities of data. Magnetic storage, including tape was described in some detail back in chapter 2, therefore we restrict our discussion to their widespread use for backup purposes.

Fig 4.29
Various types of magnetic tape cartridges.

A single inexpensive magnetic tape can store the complete contents of virtually any hard disk; currently magnetic tapes (and tape drives) are available that can store in excess of 500GB of data at just a few cents per gigabyte. Most backup systems compress data prior to it being written to tape, this compression usually doubles the capacity of most tapes – a 500GB tape can actually be used to backup 1TB of system data.

Tape cartridges encase a much larger surface area of storage material than other forms of removable storage. The ability to backup such large amounts of data using just one tape far outweighs the disadvantages of sequential access. In any case both backup and restore procedures are essentially sequential processes. Furthermore tape cartridges are light, portable and do not contain complex electronics. This makes the cartridges suitable for long term and offsite storage.

There are two different technologies currently used to store data on magnetic tape, helical and linear. In the related Preliminary textbook we discussed the detailed operation of helical and linear tape drives.

Tape libraries, such as the one shown in *Fig 4.30*, include multiple tapes and multiple tape drives. A robotic system moves tapes between the storage racks and the tape drives. Such systems allow tapes to be automatically rotated according to the details of the organisation's backup procedures. The tape drives are just normal single drives whose operation has been automated. The use of many standard tape drives improves the fault tolerance of the tape library as complete drives can be replaced without affecting or even halting backup processes.

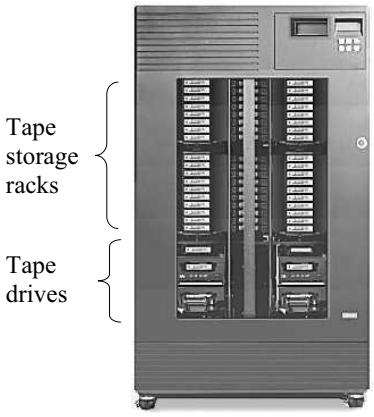


Fig 4.30
Qualstar's TLS-58132 tape library stores up to 340 terabytes of data.

Various different size tape library devices are available to suit the backup demands of different information systems. Small tape libraries are available that hold just four tapes and use a single drive; these devices provide capacities suited to most small businesses. Larger devices hold hundreds or even thousands of tapes and contain many drives. Large government departments and organisations link multiple tape library devices together; such systems hold hundreds of thousands of tapes and many thousands of tape drives.

Hard disks

The use of hard disks for backup has recently become popular for smaller systems. External hard disk devices are available that connect to a computer via high-speed USB or firewire ports, whilst others connect directly to Ethernet networks. In terms of cost these alternatives are still significantly more expensive than tape if the equivalent level of protection is to be achieved – currently tapes costs tens of dollars each whilst similar capacity hard disks cost hundreds of dollars each. For backup processes many hard disks are required. Nevertheless for small business and home backup purposes external hard disks are now a viable alternative. For larger systems the physical size, weight and mechanical complexity of hard disks is significant when the media must be transported to secure offsite storage.

Note that mirrored RAID systems use multiple hard disks to store copies of data. These systems protect data and provide fault tolerance should one of the mirrored drives fail. Such systems do not protect data against total system failure and are of no use when historical data is required to rebuild the system to a prior state. Hard disks used for backup are configured to perform full backups and partial backups such that the system (or individual files) can be restored to previous states.

GROUP TASK Research
Research the current cost of external hard disks with a similar capacity to the hard disks within current personal computers.

Optical Media – CD and DVD

For single machines and small businesses optical media is a popular and low cost backup solution – most computers include CD/DVD read/write drives and rewritable DVD media is relatively inexpensive. A single layer DVD stores 4.7 gigabytes and a double layer DVD approximately 8.5 gigabytes of data, this is sufficient for backing up most hard disks. Even when data compression is used multiple rewritable DVDs will likely be needed for a full backup, however a single DVD is generally sufficient for partial backups.

For backups of larger systems DVD media does not have sufficient capacity to make it a viable alternative to tape. Currently a single magnetic tape cartridge can store more than 500GB of data, this amount of data requires in excess of 50 DVDs. Large government and commercial organisations use tape libraries that backup to hundreds or even thousands of tape cartridges. Clearly using DVDs would require 50 times more DVD disks than tapes. This is unviable in terms of quantity of disks but also in terms of physically moving the disks in and out of drives.

Recordable "write once-read many" (WORM) CDs and in particular DVDs are used for archiving critical data that must be archived in an unalterable form for long periods of time – often exceeding 10 years. In most instances the data must be archived permanently because of governmental or organisational requirements. In most cases this data will rarely be read. Such applications include medical, legal and taxation records that were traditionally stored on paper. For larger systems optical "jukebox" devices are available. These devices include multiple optical drives together with automatic disk changers. The Pioneer DRM-3000 shown in *Fig 4.31* includes two DVD drives and 6 magazines that each holds 50 DVD-R disks.

Fig 4.31
Pioneer DRM-3000 optical disk jukebox.

GROUP TASK Research
Research and then compare and contrast the current capacity and cost of tape cartridges compared to optical disks.

Online Systems

Businesses are beginning to emerge on the Internet that specialise in providing online backup and recovery for individuals and small businesses. These online systems totally automate the backup process for users. All data is transferred via the Internet to a secure remote site. The remote site then manages the secure storage of the data on behalf of the individual or business. Clearly the remote site must use some form of secure and permanent storage. When first using an online backup system a full backup must be made, which is a time consuming process. After the initial backup, incremental backups are made at regular intervals – in some cases every time a file is saved. Such systems enable recovery of different historical versions of individual files as well as recovery of complete systems.

Large organisations that manage large volumes of critical data maintain complete operational copies of their entire system at remote locations. Such copies include the hardware, software, communication lines and data. Data from the original site is continually backed up via online communication lines to the remote site or sites. This is the ultimate in fault tolerance as a complete system failure, such as a fire or terrorist attack, can be recovered instantly by simply activating the backup site.

GROUP TASK Research
Research using the Internet current online backup services. Determine the capabilities and cost of such services.

BACKUP PROCEDURES

The same backup media should not be used continuously to perform backups. Rather multiple sets of backup media should be purchased and used. The aim is to maintain many complete backup copies produced at different times such that the system's data can be recovered back to a variety of different past states. If only a single set of backup media is used then failure of the media can spell disaster. Furthermore many problems, such as viruses, may go undetected for some time. In these cases a backup copy produced prior to the problem occurring is invaluable.

A definite backup procedure is required that is documented and applied consistently. Most backup procedures fail as a result of human error. Therefore it is vital that backup procedures are thoroughly understood and are simple to apply. It is particularly important for the people who perform the backups to be aware of their importance – backups can easily become a chore that are easily overlooked. The procedure should specify which set of media is to be used for each backup and when and where backup copies should be stored offsite.

Backup procedures should also specify how backup copies are to be verified to ensure they will actually work in the event of failure. Commonly the backup software verifies all data on the media as the backup is being made – essentially after writing, the data is read back into RAM and compared to the original. Specialised backup software is available that can be configured to enforce the backup procedure – including verification. However human assistance is still needed to physically change the backup media and to ensure media is stored offsite as required. It is advisable to manually perform a test recovery at regular intervals to ensure recovery operates as expected. Such recovery tests should be performed using a different media drive – it is

possible that tapes or other media will not operate correctly in different drives. All backup copies will be useless if the backup drive itself fails or is destroyed.

GROUP TASK Research
Research backup software included with your operating system and other examples of specialised backup software. Outline the available features.

A simple, albeit costly, backup strategy would be to make a full backup to new media at regular intervals such as every afternoon. Such a system is certainly simple to implement and for some critical or high value data such a strategy may well be an appropriate solution. However for most systems a less costly solution that reuses the backup media is generally preferred. There are three commonly used media rotation schemes; Grandfather, Father Son (GFS), Round Robin and Towers of Hanoi. We shall discuss examples of each of these schemes. To simplify our discussion we assume a single tape is sufficient for completing each backup. In reality each backup may require multiple tapes, DVDs or some other type and quantity of backup media.

Grandfather, Father, Son (GFS)

This is the most commonly used rotation scheme. GFS rotation requires daily or son tapes, weekly or father tapes and monthly or grandfather tapes. Full or partial backups are performed each working day to a son tape, except for the last workday. On the last workday a full backup must be performed to one of the weekly or father tapes. At the end of the fourth week a full backup is made to one of the monthly or grandfather tapes. The set of son tapes is reused each week, the set of father tapes is reused each month and the set of grandfather tapes is reused each year. Usually the monthly or grandfather tapes are stored offsite and the weekly tapes are stored onsite within a safe, however this is varied to suit the needs of the individual organisation.

To implement a GFS rotation within an organisation that operates 5 days per week requires four son tapes, three father tapes and thirteen grandfather tapes. Note there are 13 four-week periods in a year, not 12. The son tapes are labelled Mon, Tues, Wed and Thurs. The father tapes are labelled Week 1, Week 2, Week 3 and the grandfather tapes Month 1, Month 2, … Month 13. After making an initial full backup the schedule in *Fig 4.32* is used to determine the tape media used for each afternoon's backup.

Mon	*Tues*	*Wed*	*Thurs*	*Fri*
Mon	Tues	Wed	Thurs	Week1
Mon	Tues	Wed	Thurs	Week2
Mon	Tues	Wed	Thurs	Week3
Mon	Tues	Wed	Thurs	Month1
Mon	Tues	Wed	Thurs	Week1
Mon	Tues	Wed	Thurs	Week2
Mon	Tues	Wed	Thurs	Week3
Mon	Tues	Wed	Thurs	Month2
Mon	Tues	Wed	Thurs	Week1
Mon	Tues	Wed	Thurs	Week2
Mon	Tues	Wed	Thurs	Week3
Mon	Tues	Wed	Thurs	Month3

Fig 4.32
Grandfather, Father, Son media rotation.

Weekly and monthly backups should always be full backups, however the daily or son backups can be full or partial backups. If a relatively small amount of data is present then full backups can be used throughout. When full backups are used just one tape is required to restore data from the most recent backup, or indeed from any backup. If differential daily backups are made then two tapes are required to restore to the most recent backup – the last weekly full backup is first restored followed by the most recent daily differential backup. If incremental daily backups are used then the most recent weekly full backup is restored followed by restoration of each of the subsequent incremental daily backups. Using full daily backups simplifies the restore process at the expense of longer backups. Using differential daily backups results in slightly more complex restore processes, but reduces the time taken for daily backups significantly – only files changed since the last full backup are copied. Using incremental daily backups

complicates restore processes, but requires less time for each daily backup – only files changed during the day are copied.

GFS rotation means that recovery operations can restore back to any day in the last week, any week within the last month and any month within the last year. Usually the final yearly backup is archived permanently so it is also possible to restore back to the end of a particular year. At first it may seem unlikely that such ancient data would ever need to be restored. This is probably true of the entire data, however it is not unusual for particular files from previous months or years to be required.

Notice that the son tapes are used much more often than the father tapes and the father tapes more often than the grandfather tapes. This means the son tapes will suffer the most wear and grandfather tapes the least wear. Some backup procedures specify that tapes be simply replaced at regular intervals, whilst other GFS procedures promote son tapes to become fathers and fathers to become grandfathers. Such promotion strategies mean new tapes are introduced as sons where they are used actively for a period of time. As they age they are promoted to become fathers where they are less active. Finally the father tapes are promoted to become grandfathers where they go to an offsite retirement home to relax quietly on a shelf!

There are numerous different ways to implement GFS rotation. The number of sons can be increased for organisations that operate 7 days a week or so that backups are performed more than once per day. The frequency of father backups can be increased or decreased, as can grandfather backups. Indeed some complex schemes increase the number of generations to include great or even great great grandfather generations. The detail of the procedure is determined by the needs of the organisation. For some organisations losing even a days work would be catastrophic, whilst for others this is an acceptable risk.

Round robin

A round robin rotation reuses all tapes equally. Each tape is numbered sequentially, say from 1 to 5 or maybe Mon to Fri. Each tape is then used in turn. When all tapes have been used the cycle simply repeats that is, tape 1 is used after tape 5 and the cycle continues. Clearly when just five tapes are used and backups are made daily then it is not possible to restore data to states more than five days old. Clearly each tape added to the cycle extends the ability to restore back a further day.

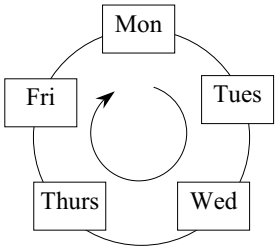


Fig 4.33
Round Robin media rotation.

This simply strategy is only suited to small businesses where restoration of data back to a particular day is a high priority. For instance if 30 tapes are used then it is possible to restore back to any day in the past month. In reality most organisations that use a round robin scheme will (or should) also archive backups permanently at regular intervals.

Towers of Hanoi

This is a somewhat complex method of rotation based on the Towers of Hanoi logical puzzle. In the puzzle a series of disks are stacked in size order on one of three poles as shown in *Fig 4.34*. The aim is to move all disks to the third pole, however you can only move one disk at a time, and larger disks can not be placed on top of smaller disks. In *Fig 4.34* we have six disks labelled A to F in ascending size order, however any number of disks is possible. The solution

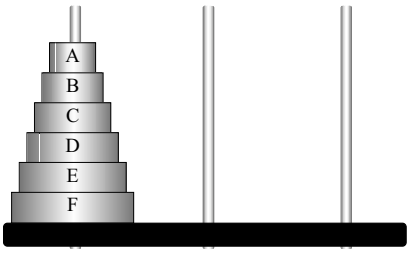


Fig 4.34
The Towers of Hanoi puzzle.

involves moving the smallest disk A every second move, disk B every fourth move, disk C every eighth move and so on. In our example, disk B is first moved on the 2nd move, disk C as the 4th move, and disk F cannot be moved until the 32nd move.

So how does this puzzle relate to backups and tape rotation? Each disk represents a tape and the order in which the disks are moved determines the order in which the tapes are used. Therefore in our example, tapes are used in the order shown in *Fig 4.35* below – this complete sequence repeats continuously every 32 days. Notice that tapes used less often will contain data from the more distant past whilst those used more often contain more recent data.

Day 1 **A**	Day 2 **B**	Day 3 **A**	Day 4 **C**	Day 5 **A**	Day 6 **B**	Day 7 **A**	Day 8 **D**
Day 9 **A**	Day 10 **B**	Day 11 **A**	Day 12 **C**	Day 13 **A**	Day 14 **B**	Day 15 **A**	Day 16 **E**
Day 17 **A**	Day 18 **B**	Day 19 **A**	Day 20 **C**	Day 21 **A**	Day 22 **B**	Day 23 **A**	Day 24 **D**
Day 25 **A**	Day 26 **B**	Day 27 **A**	Day 28 **C**	Day 29 **A**	Day 30 **B**	Day 31 **A**	Day 32 **F**

Fig 4.35
Towers of Hanoi rotation sequence with six tapes.

When performed manually any new tape added to the system becomes the new tape A and all other tapes move up – A becomes B, B becomes C, and so on. Therefore over time each tape will eventually be used the same number of times. Furthermore offsite storage can be specified for particular tapes. For example in our example tape E and F could be stored offsite in the knowledge that they will only be required every 32 days.

GROUP TASK Discussion
Analyse the above Towers of Hanoi rotation sequence to determine the historical data that can be restored at different points within the sequence.

Specialised backup software that operates in conjunction with tape libraries usually supports and recommends the Towers of Hanoi rotation scheme. A magazine is loaded with sufficient tapes for a complete Towers of Hanoi rotation – 6 tapes in our example above. The software performs the daily tape rotation automatically by loading and backing up to the correct tape in the sequence. At the end of the sequence the complete magazine is placed into secure storage and a new (or recycled) magazine is loaded into the library. The backup software is also used during recovery operations. Such backup and restore applications track and are able to display the different versions of data or specific files that are available to be restored. This means it is not necessary for staff to understand the complexity of the system to restore data efficiently. Some advanced systems provide network access to tape libraries for end users. Such systems allow users to restore historical versions of their own files from the tape library as required.

HSC style question:

Big Bad Bikes (BBB) imports bikes from overseas suppliers and sells them to the general public. BBB sells mountain bikes, road bikes, BMX bikes and bicycle clothing. BBB has a transaction processing system (TPS) to process their sales, generate purchase orders, supplier payments and produce stock level reports.

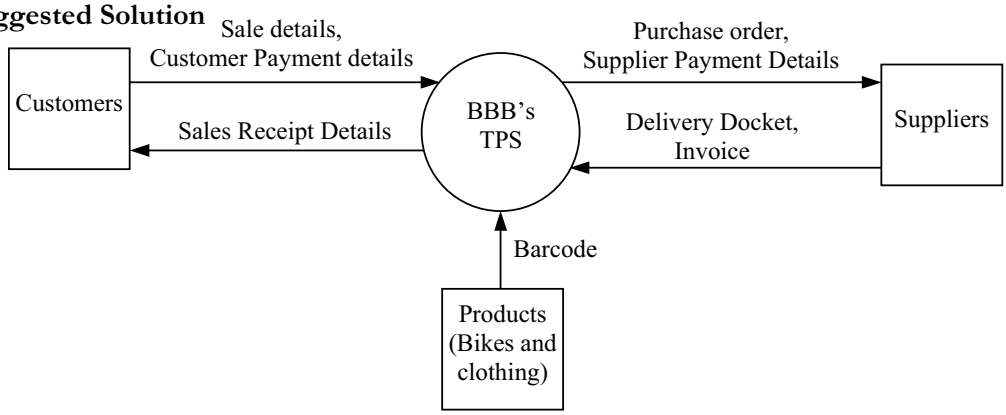
Each stock item in the store has a barcode. When new stock arrives from suppliers the barcode is scanned to update the stock inventory database. A point-of-sale (POS) terminal is used to record all sales and produce customer receipts.

(a) Represent the TPS of Big Bad Bikes using a context diagram.

(b) Propose and describe a suitable backup procedure which may be employed by Big Bad Bikes.

Suggested Solution

(a)



(b) Backup is the process of making a copy of files used by the system in case the original is lost or damaged. There are a number of possible backup procedures that BBB could utilise, however as the data within the system is not enormous a Grandfather, Father, Son tape rotation scheme, full backups and offsite storage is a suitable procedure. To verify backup copies are usable and to simplify recovery in the event of failure it is recommended that two identical tape drives be purchased – one installed on the BBB server and another on the owners home computer.

- As the total data within the TPS is likely to fit on a single tape a total of 20 tapes are required. The first four tapes are the sons and are labelled Mon, Tues, Wed and Thurs. The next 3 tapes are the fathers and are labelled Week 1, Week 2 and Week 3. The remaining 13 tapes are the grandfathers and are labelled Month 1 through to Month 13.
- Each afternoon, except on Friday, a full backup is made of the TPS data to the corresponding son tape. These tapes are stored on a shelf within the office.
- On the first Friday a full backup is made to the Week 1 father tape and then on the next two Fridays backups are made to the Week 2 and then Week 3 father tapes. These weekly father tapes are stored within the safe at the shop.
- Every fourth Friday backups are made to the grandfather month tapes in sequence. That is, the Month 1 tape first, then the Month 2 tape and so on. After each grandfather backup is made the owner of BBB takes the tape home, verifies the tape is readable using their home computer and stores the tape securely until required the next year.
- At the end of each year a further backup is made onto a fresh tape. This backup is placed into permanent storage – perhaps in a safe deposit box at BBB's bank.

Comments

- In a Trial or HSC exam part (a) would likely attract 3 marks and part (b) 4 marks.
- It is not necessary to include employees of BBB on the context diagram as they are part of the system – they perform the system's information processes.
- There are numerous other possible and suitable backup procedures that could have been proposed and discussed.

SET 4D

1. Partial backups are included in:
   (A) Full and incremental backups.
   (B) Full and differential backups.
   (C) Differential and incremental backups.
   (D) Full, differential and incremental backups.
2. Which of the following is always true if 10 media sets are used and backups are made daily?
   (A) Data can only be recovered to a state within the past 10 days.
   (B) Data can only be restored to exactly 10 points in time.
   (C) It is possible to recover to more than 10 points in time.
   (D) More than one set of backup media will be needed to restore data.
3. Which of the following lists is in ascending order by storage capacity?
   (A) CD, DVD, Tape cartridge.
   (B) CD, Tape cartridge, DVD.
   (C) Tape cartridge, DVD, CD.
   (D) DVD, Tape cartridge, CD.
4. When differential backups are made, which of the following occurs?
   (A) All archive bits are set to true.
   (B) All archive bits are set to false.
   (C) Archive bits for changed or new files are set to false.
   (D) No archive bits are altered.
5. Currently the preferred backup media for large systems is:
   (A) Magnetic tape.
   (B) Rewritable optical disks.
   (C) External hard disks.
   (D) Mirrored RAID.
6. It is common practice to purchase two identical tape drives and store one offsite. Why is this?
   (A) If the original is damaged or destroyed along with the data then data can still be recovered.
   (B) So backup copies can be verified on another drive.
   (C) If (or when) the original tape drive fails backups can continue without the need to urgently obtain a new tape drive.
   (D) All of the above.
7. Hard disks are least likely to fail:
   (A) when new.
   (B) during mid-life.
   (C) late in life.
   (D) during the first 5-7 years.
8. A GFS rotation has been used for a full year. This rotation uses 4 daily, 3 weekly and 13 monthly tape sets. Backups are made each afternoon. To what points in time can data be restored?
   (A) Any day in the last month and the end of each month in the last year.
   (B) Any day in the last week, any month in the last year and any day in the previous month.
   (C) Any day in the last week, the end of any week in the last month and the end of any month in the last year.
   (D) Any day in the last year.
9. A round robin rotation is used with 30 sets of backup media. Backups are made each weekday at midday and again at 6pm. Which of the following is true?
   (A) Data can be recovered to any day in the past 30 weekdays.
   (B) Data can be recovered if it was created more than 15 weekdays ago and has not been altered since.
   (C) Data can be recovered if it was deleted from the system more than 15 weekdays ago.
   (D) Data that was altered more than 15 days ago cannot be recovered.
10. The Towers of Hanoi rotation scheme described in *Fig 4.35* has been used for many months. Just prior to the day 25 backup, how old are each of the 6 backups?
    (A) A is 1 day old, B is 2 days old, C is 4 days old, D is 8 days old, E is 16 days old and F is 32 days old.
    (B) A is 1 day old, B is 2 days old, C is 3 days old, D is 4 days old, E is 5 days old and F is 6 days old.
    (C) A is 2 days old, B is 4 days old, C is 8 days old, D is 16 days old, E is 32 days old and F is 64 days old.
    (D) A is 2 days old, B is 3 days old, C is 5 days old, D is 1 day old, E is 9 days old and F is 25 days old.
11. Define the terms “backup” and “recovery”.
12. List and briefly describe a variety of issues and faults that can be resolved when backup copies of data are available.
13. Explain the role of archive bits when performing full, incremental and differential backups.
14. Assess the merits of secure onsite storage and offsite storage of backups.
15. Explain each of the following in relation to backup and recovery.
    (a) Transaction log (b) Mirroring (c) Documenting procedures

COLLECTING IN TRANSACTION PROCESSING SYSTEMS

In this section we consider the information technology used to collect data into transaction processing systems. We examine the operation and use of specialised collection hardware. We then discuss the design of data collection forms including paper, online and web-based forms.

COLLECTION HARDWARE

There are numerous different collection hardware devices used within transaction processing systems. Examples include keyboards, mice, image scanners, touch screens, RFID readers, Automatic Teller Machines and fingerprint scanners. Earlier in this chapter we described RFID readers and tags (refer page 394) within a library's transaction processing system. In chapter 3 we examined ATMs in some detail – refer page 294-295. In this section we restrict our discussion to three specialised collection devices namely, MICR readers, barcode readers and magnetic stripe readers.

GROUP TASK Review
Review the sections on RFID readers (page 394) and also ATMs (page 294-295). Describe their role in transaction processing systems.

MICR for reading cheques

Magnetic ink character recognition (MICR) has been used on cheques since the 1950s and it still remains the primary method for high-speed collection of cheque data. The ink (or toner) used to print MICR characters contains ferromagnetic material – a material that can easily be magnetised. It is the ferromagnetic material within the ink or toner that is used by the readers. This means customers or bank employees can write over the MICR characters without greatly affecting the ability of the MICR line to be successfully read.

There are strict standards that specify the precise design of MICR fonts, paper characteristics and positioning of MICR line data on cheques. An agreed standard is critical as cheques are regularly read by banks and clearance facilities across the world. Furthermore cheques pass through many MICR readers during clearance processes. Current standards require that each cheque can be read at least 30 times. In Australia the Australian Payments Clearance Association (APCA) produces and maintains the MICR specifications for cheques.

Each cheque is encoded with a MICR line positioned across the bottom of the cheque. The MICR line includes fields for BSB, account number and cheque number. The value of the cheque can also be included – for large organisations the cheque value can be printed as the cheque is created, whilst for other cheques the value can be added to the MICR line during outwards processing. In Australia the MICR E-13B font is used, whilst many European countries use the CMC-7 font. The majority of MICR readers include the ability to read MICR lines in either of these fonts. Both these fonts include the 10 digits and 4 additional symbols for BSB, amount, domestic and dash – examples of the E-13B character set was shown back in *Fig 4.25*.

There are two common technologies within MICR readers, waveform and matrix readers. Waveform MICR readers first magnetise the entire MICR line by passing it over a write head. The cheque then moves past a magnetic read head. The read head converts the magnetic data into a very small electrical signal. Each MICR character is designed to be significantly different, therefore the electrical signal for each character has a unique signature waveform. The reader detects each waveform and converts it into the corresponding digital code for that character. Waveform technology is used within smaller low speed MICR readers. Commonly cheques are fed into these readers by hand or via a small hopper.

Matrix reader technology is used within most high-speed MICR readers, including the IBM 3890 and Unisys DP1800 series of reader/sorters. This technology has been in use since the early 1960s and remains the dominant technology used within most bank's readers and sorters. Matrix readers are able to read and sort up to 2400 cheques per minute. Each MICR line is magnetised such that each character is split into many vertical slices. The read head includes 30 mini-heads positioned at right angles to the slices. Therefore each character is split into a matrix of magnetic cells – each slice split into 30 cells. As each cheque's MICR line passes through the reader, each of the 30 mini-heads simply determines whether each cell in each slice is magnetised. The result being a mini bitmap of each character (see *Fig 4.36*). This bitmap is then converted to its corresponding digital code. The data is transmitted back to the system for further processing and storage. The BSB data is commonly used as the basis for sorting cheques in preparation for transport to particular banks or branches.

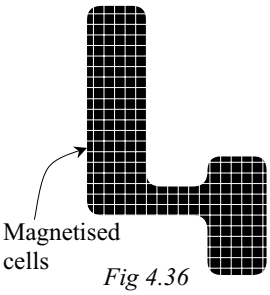


Fig 4.36
Matrix MICR readers read each character as a matrix of magnetised cells.

GROUP TASK Investigation
Examine the MICR line on a cheque. Identify the fields present within the MICR line and the symbols used to separate these fields.

Barcode Readers

Barcode readers or scanners operate by reflecting light off the barcode image; light reflects well off white and not very well off black. This is the basic principle underlying the operation of all types of scanners. A sensor is used to detect the amount of reflected light; so to read a barcode we can either progressively move the light beam from left to right across the barcode or use a strip of light in conjunction with a row of light sensors. Each of these techniques are used for different designs of barcode scanner; those based on LED, laser and CCD technologies dominate the market, *Fig 4.37* shows an example of each. Most barcode scanners incorporate a decoder to organise the data into a character representation that mimics that produced by the keyboard. This means most barcode readers can be installed between the keyboard and the computer without the need for dedicated interface software.

Fig 4.37
Clockwise from top-left: LED wand, multi-directional laser and CCD based barcode scanners.

Barcode wands use a single light emitting diode (LED) to illuminate a small spot on the barcode; the reflected light from the LED is measured using a single photocell. As the wand is steadily moved across the barcode, areas of high and low reflection change the state of the photocell. The photocell absorbs photons (a component of light); as the intensity of photons absorbed increases so too does the current flowing through the photocell; large currents indicating white and smaller currents indicating black. This electrical current is transformed by an analog to digital converter (ADC) to produce a series of digital ones and zeros. The same LED technology is used for slot readers, where the barcode on a card is read by swiping the card through the reader.

Lasers are high intensity beams of light; as such they can be directed very precisely. Laser barcode readers can therefore operate at greater distances from the barcode than other technologies; commonly up to about 30cm away. The reflected light from the laser is detected by the photocell using the same technique as LED scanners. There is no need to manually sweep across the barcode as the laser beam is moved using an electronically controlled mirror. Basic models continually sweep back and forth across a single path, whilst more advanced models perform multiple rotating sweeps that trace out a 'star like' pattern. These advanced models are much more effective as the user need not hold the scanner parallel to the barcode; rather the scanner rotates the scan line until a positive read is collected. Supermarkets often use this type of barcode scanner mounted within the counter top.

Barcode scanners based on charge coupled devices (CCDs) contain a row of photocells built into a single microchip. CCD technology is used within many image collection devices including; CCD barcode scanners, digital still and video cameras, handheld image scanners, and also flatbed scanners. For both barcode and image scanners a single row CCD is used. The light source for these scanners is typically a single row of LEDs; the light being reflected off the image back to a mirror. The mirror reflects the light onto a lens that focuses the image at the CCD. Each photocell in the CCD transforms the light into different levels of electrical current which are converted into bits, using a similar technique to that used in LED and laser barcode scanners. CCDs in image scanners differ slightly; they convert the electrical current from each photocell into a binary number, normally between 0 and 255, using a more complex analog to digital converter (ADC).

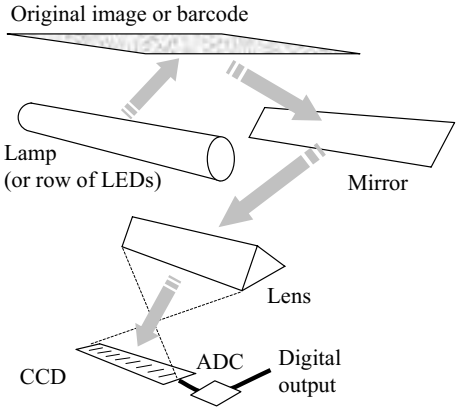


Fig 4.38
The components and light path typical of most CCD scanner designs.

GROUP TASK Research
Research the specifications and cost of different barcode readers. Classify each barcode reader according to its suitability for use within retail transaction processing systems.

Magnetic Stripe Readers

Today almost all plastic cards contain magentic stripes. This includes credit cards, ATM cards, library cards, ID cards, frequent flyer cards, store charge cards and even some temporary and limited use paper cards such as train and bus tickets. Apart from some rare exceptions the magnetic stripes on these cards include three parallel tracks. Each track is approximately 0.11" (2.8mm) wide and 0.025" (0.6mm) apart (refer *Fig 4.39*). Tracks one and three contain data written at a density of 210 bits per inch, whilst track two is written with a density of just 75 bits per inch.

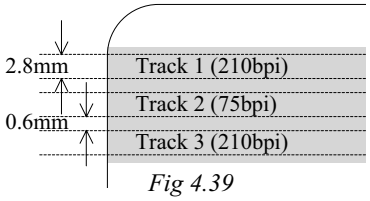


Fig 4.39
Typical layout and density of tracks on magnetic stripes.

Most ATM and credit cards contain data on tracks 1 and 2, however most readers, including those within EFTPOS terminals and ATMs are only capable of reading track 2. Track 2 contains the primary account number and other details unique to the card issuer – many ATM cards contain an encrypted version of the PIN. The PIN cannot be decrypted without a decryption key from the card issuer's bank. Track 1

contains essentially the same data as track 2 plus the card holders name, so if your name ever appears on an offline terminal then the stripe reader must be reading track 1 rather than track 2. Track 3 was originally intended to contain rewritable data such as details of offline transactions. As all EFTPOS and ATMs are online devices track 3 is rarely used.

Track 1 is encoded using a 6-bit subset of ASCII and is able to store 79 alphanumeric characters. Track 2 is encoded using 4-bit BCD (Binary Coded Decimal) and is able to store 40 characters, track 3 also uses BCD encoding and can store up to 107 characters. The 4-bit BCD character set only includes 16 characters - the 10 digits and 6 control characters as shown in *Fig 4.40.* All characters are followed by an odd parity bit, therefore on track 1 a total of 7-bits are used per character, whilst on tracks 2 and 3 just 5 bits are used per character. The data on each track is followed by a longitudinal redundancy check (LRC) character. LRCs calculate an odd parity bit for each corresponding bit (or column) in each character within the data. When a card fails to be read correctly and needs to be swiped again it is generally due to parity check or LRC errors.

Data Bits	Parity Bit	Char-acter	Function
0000	1	0	Data
0001	0	1	Data
0010	0	2	Data
0011	1	3	Data
0100	0	4	Data
0101	1	5	Data
0110	1	6	Data
0111	0	7	Data
1000	0	8	Data
1001	1	9	Data
1010	1	:	Control
1011	0	;	Start Sentinel
1100	1	<	Control
1101	0	=	Field Separator
1110	0	>	Control
1111	1	?	End Sentinel

Fig 4.40
BCD character set used on track 2 and 3 of most magnetic stripes.

GROUP TASK Activity
Using the information above calculate the minimum width of the magnetic stripe so that it is able to accommodate the maximum number of characters on all three tracks. Compare your result with a real card.

GROUP TASK Activity
The data on track two of an ATM card contains a start sentinel followed by the account number 12345678, a field separator, the encrypted PIN 2468 and finally an end sentinel and LRC character calculated over all other characters. Produce the binary string stored on track 2 of this card.

All magnetic stripe readers contain a magnetic read head that operates using the same principles as the read heads on tape drives and within hard disks. Some readers require the user to swipe their card, whilst others require the card to be inserted into a slot. Insertion style machines control the speed at which the magnetic stripe passes the read head and hence tend to produce less errors. Such readers retain the card within the machine until the transaction is completed. In ATMs insertion style readers are used to increase security. For example failure to enter a correct PIN after a set number of attempts or detecting that a card is stolen results in the card being retained within the machine.

GROUP TASK Discussion
Brainstorm applications that use barcode readers and applications that use magnetic stripe readers. Discuss likely reasons why each of these applications uses one type of reader rather than the other.

COLLECTION FROM FORMS

Forms are used to collect data required to process transactions. Forms can be paper based where indirect users manually complete the form and then data is entered into the system and subsequently batch processed at a later time. Common examples of paper forms include Medicare forms, taxation returns, loan applications and enrolment forms. Today many of these organisations also provide alternative web-based data entry screens – this removes the need for participants to manually enter data from paper forms into the system. Screens used for data collection are also forms. These screens can be part of front-end client applications that connect via a local area network to backend DBMS systems or they can be web-based clients where the data travels over the Internet and then via a web server before it arrives for storage within the system's database. In either case the transactions can be processed immediately in real time or they can be stored in a transaction file for later batch processing.

All forms are user interfaces; their purpose is to guide the user through the data collection process such that the data is collected accurately and efficiently. Paper forms are unable to react to user inputs whilst screens are able to provide real time feedback in response to inputs. Data validation is used to improve accuracy. On screens the validation criteria can be enforced. On paper forms validation criteria cannot be enforced rather indicators of the required data are provided, for example instructions, example data and input areas that restrict the length of data.

General form design principles

Some general principles that apply to paper, online and web-based form design include:

- Know who the users are. What are their goals, skills, experience and needs? What motivation has led the user to complete the form? Answers to these questions are critical. The form must be usable given the ability of the users. The form will not be completed if the user has little motivation, therefore the purpose of the forms completion should be clear. Furthermore the purpose should reflect some user goal or need. For example, a web-based form that requests personal details when the user has no idea of how these details will be used or what they will receive in return is unlikely to be completed honestly or at all.
- Identify the precise nature of all data items that will be collected. This includes the data type, length and any other restrictions. Does entry of one data item determine the possible values or alternatives for subsequent data items? Answers to such questions help determine validation rules and the sequence of input fields.
- Consistency with other forms and applications. Capitalise on users past experience and skills by using and arranging form components in familiar ways. For instance on screens radio or option buttons should be round, whilst check boxes should be square. On paper forms, a series of boxes is often used to control the number of characters to be entered and to promote legible handwriting.
- Form components should be readable. Readability is affected by the actual words and fonts used as well as the logical placement and grouping of related fields. Underlined and capitalised text should be avoided; bold text is preferred where extra emphasis is needed. Sans serif fonts are preferred – serif fonts are generally reserved for large blocks of text.
- Forms should include significant areas of white space to visually imply grouping or simply to rest the eyes. Colour and graphics should be used sparingly and only when they improve the readability of the form. In general pastel background colours are preferred with dark text and white input fields. Cluttered forms always appear more complex compared to forms where elements are generously spaced.

Layout of labels and input fields

The layout and alignment of labels and input fields should lead the user through the desired input sequence. In Western countries it is preferred for both labels and input fields to be left justified. It is simpler to scan down a page when there is a hard line down the left hand edge. Therefore all labels and input fields should be left justified.

Some common layouts are shown in *Fig 4.41*. The single column layouts on the left are easier for the eye to scan down and each label is equally close to its corresponding input field, however significantly more vertical space is used. The two column designs require less vertical space, however the differing length of labels causes problems. If the labels are left justified then all smaller length labels are positioned some distance from their corresponding input fields. If the labels are right justified then we have an undesirable jagged left edge. Introducing horizontal lines into the design assists the eye to better link labels with their input fields; however including such lines between all fields reduces the ability to scan downwards. The second single column example groups input fields – groups that make logical sense should be chosen. When designing forms compromises must be made. For large systems various designs should be tested with many users before settling on a final layout.

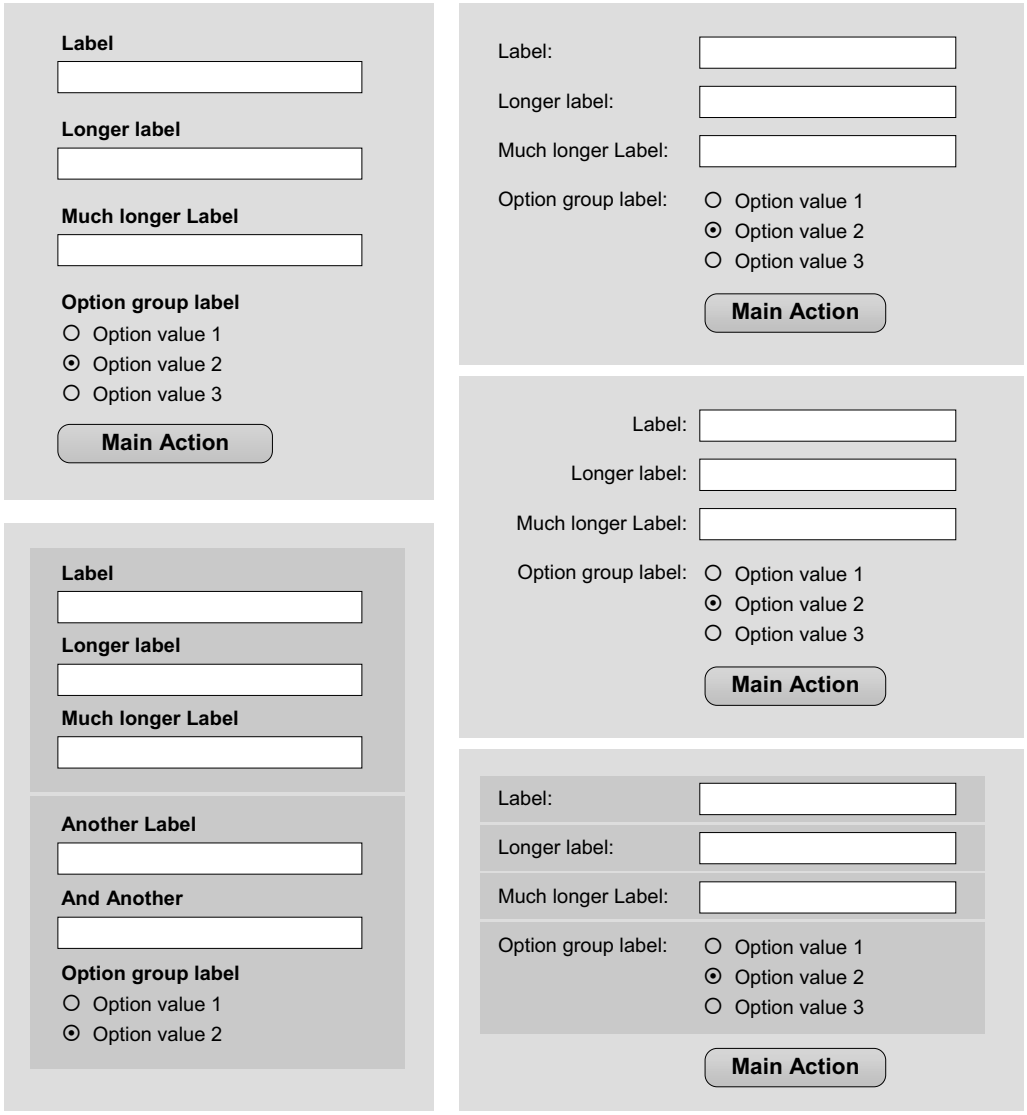


Fig 4.41
Possible label and input field layouts.

GROUP TASK Discussion
Identify and analyse other design features present on the example layouts in *Fig 4.41*.

GROUP TASK Practical Activity
Create a suitable layout for collecting customer names, addresses, phone numbers and email addresses.

Principles particular to the design of paper forms

Additional design considerations specific to paper-based forms include:

- Paper forms are used to collect data that will subsequently be input into a computer system; therefore the paper form and the data entry screen need to be structured to assist the data entry process as well as the manual completion of the paper form. Paper forms should not merely be a printout of the corresponding data entry screen; rather both versions should use the strengths of their respective mediums whilst maintaining consistency in terms of the order of data elements.
- Paper-based forms cannot react to a user's responses; hence instructions must be available and clearly stated. General instructions relevant to the whole form should be placed before the questions commence, whereas instructions for particular items should be present at the point on the form where they are needed. For example if a certain answer means the person must jump to question 9 then this needs to be stated clearly; on a data entry screen the questions that are not needed can be dulled or simply not displayed at all.
- Colour, texture, fonts and the paper itself cannot be altered when using paper forms. Paper forms therefore should be designed so that these elements will work for all, or at least the majority, of users. The paper should be thick enough that type cannot be seen through the page. Consider having large print versions available for sight-impaired users.
- Appropriate space for answers. The space provided for answers on a paper form cannot increase or decrease; most people use the space provided as an indicator of the amount of information they need to supply. On data entry screens it is possible for such space to grow as needed, on paper forms such space needs to be more carefully considered.

GROUP TASK Practical Activity
Obtain a copy of your school's enrolment form. Analyse the design and layout of this form and recommend areas for improvement.

Principles particular to the design of online screens

Additional considerations specific to the design of online screens that form part of software applications include:

- Clearly show what functions are available. Users like to explore the user interface; this is how most people learn new applications, therefore functions should not be hidden too deeply. If a particular function is not relevant then it is better for it to be dulled than for it to be hidden, this allows users to absorb all possibilities. At the same time the user interface should not be overly complex.
- Every action by a user should cause a reaction in the user interface. This is called feedback; without feedback that something is occurring, or has occurred, users will either feel insecure or will reinitiate the task in the belief that nothing has

happened. Feedback can be provided in subtle ways; such as the cursor moving to the next field, a command button depressing or the mouse pointer changing. Tasks that take some time to complete should provide more obvious feedback indicating the likely time for the task to complete.

- User actions that perform potentially dangerous changes should provide a way out. Many software applications include an 'undo' feature, whilst others provide warning messages prior to such dangerous tasks commencing. In either case the user is given a method to reverse their action.
- Operating systems have their own standards for user interface design. These standards should be adhered to wherever possible so that users knowledge and skills can be transferred from other familiar applications.

GROUP TASK Discussion
Forms designed for touch screens are significantly different to those used for keyboard and mouse entry. Identify the essential differences.

Principles particular to web forms

Additional considerations specific to the design of web-based forms include:

- The speed of individual user's Internet connections is unknown. Therefore web-based forms should try to validate data within the downloaded page wherever possible – the aim being to reduce the amount of data transferred. If delays are possible then feedback should be provided or processing delayed for later batch processing.
- It is often possible to design a sequence of forms such that transmission of data required for validation occurs prior to the next form in the sequence being displayed. Commonly web forms validate all input fields together after a submit button is clicked. If validation or other errors occur then users should be informed of what the error is, why it occurred and how it can be rectified. Often the original form is displayed again – if this technique is used then all correctly entered data should be filled in rather then expecting all data to be re-entered. Furthermore the data on one form can be used to determine the available options available on subsequent forms.
- In general the hardware and software used to access web-based forms is largely unknown. As a consequence particular care needs to be taken to ensure the software technologies used will operate correctly on many different combinations of hardware and software. In particular web pages should be tested within all popular web browsers using a variety of different screen resolutions.
- Users are able to set their own preferences within web pages. Labels and input fields can appear differently on different users machines even when they are using identical hardware and software. Therefore web pages need to be designed so that they will automatically format correctly based on the settings within each user's browser.
- Security of personal and other details is critical when using web-based forms. Financial transaction data should always be encrypted during transmission. If users are to feel confident divulging their details then the security measures used during transmission and subsequent storage should be clear. For example, https, which includes the secure sockets layer protocol should be used. Most browsers display a small padlock to indicate that all data transferred will be securely encrypted.

GROUP TASK Practical Activity
Browse the web and examine a variety of different web-based forms. Analyse each form and propose improvements.

Consider the design of the following forms:

Australian Government
Australian Taxation Office

Short tax return for individuals **2006**

1 July 2005 to 30 June 2006

Use the *Short tax return instructions 2006* to fill in this tax return. Please print neatly in BLOCK LETTERS with a black or blue ballpoint pen only. Do not use correction fluid or tape. Print one letter or number in each box. Print X in appropriate boxes.

1 **Are you eligible to use the *Short tax return 2006*?** (pages 1–2 of the loose leaf cover) Yes ☐ No ☐ ► If no, see page 2 of the loose leaf cover for other ways to lodge your tax return.

2 **Your tax file number (TFN)** (see the privacy information on page 52 of the instructions) TF

Sequence number – from your previous notice of assessment (page 2 of the instructions) / (needed if lodging by phone)

3 **Are you an Australian resident?** (page 3 of the instructions) Yes ☐ No ☐ ► If no, you cannot use the short tax return. RE

4 **Your sex** Male ☐ Female ☐ SX

5 **Your full name** Title – for example, Mr, Mrs, Ms, Miss NM

If lodging by phone and any of these details have changed, see page 3 of the instructions.

Surname

Given names

Has any part of your name **changed** since you last notified the Tax Office?

No ☐ Yes ☐ ► Your previous surname

6 **Your postal address**
Print the address where you want your mail sent. PA

Has your address **changed** since you last notified the Tax Office?

No ☐ Yes ☐ ► If lodging by phone and any of these details have changed, see page 3 of the instructions.

Suburb or town

State Postcode PC

Country if not Australia

7 **Is your home address different from your postal address?** HA

No ☐ ► Go to Question 8.

Yes ☐ ► Print your home address.

Suburb or town

State Postcode

Country if not Australia

8 **Your date of birth** DAY MONTH YEAR DB

9 **Your daytime phone number** Area code PN

10 **If you had a spouse at any time during 2005–06, write your spouse's name** (page 3 of the instructions)

Surname SN

Given names

11 **Will you need to lodge an Australian tax return in the future?** Yes ☐ Don't know ☐ No ☐ ► This is my final tax return. FN

12 **Do you want your refund paid directly into your financial institution account?**

No ☐ ► Your refund will be paid by cheque to your postal address.

Yes ☐ ► If you want it paid into the same account as last year, leave the following boxes blank. Otherwise give details of your account.

If lodging by phone and you answer YES, we must have your current bank account details before you can use the phone lodgment service. See page 3 of the instructions.

BSB number Must be six digits Account number EF

Account held in name(s) of

NAT 8416–6.2006 **IN CONFIDENCE** when completed PAGE 1

Fig 4.42
Australian Taxation Office Short Tax Return for Individuals, page 1.

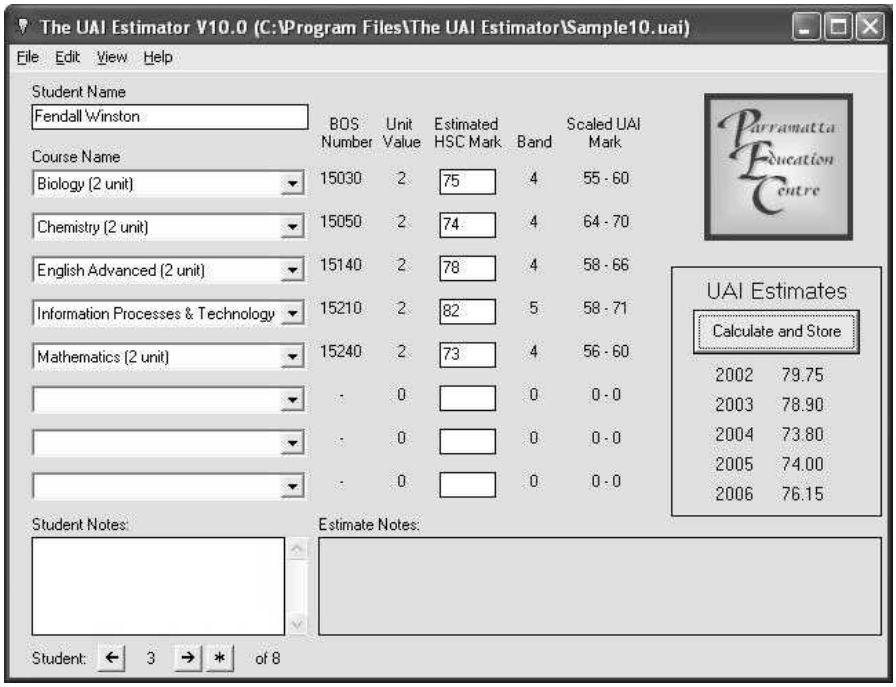


Fig 4.43
Main data entry screen from The UAI Estimator Version 10.0 for Windows.

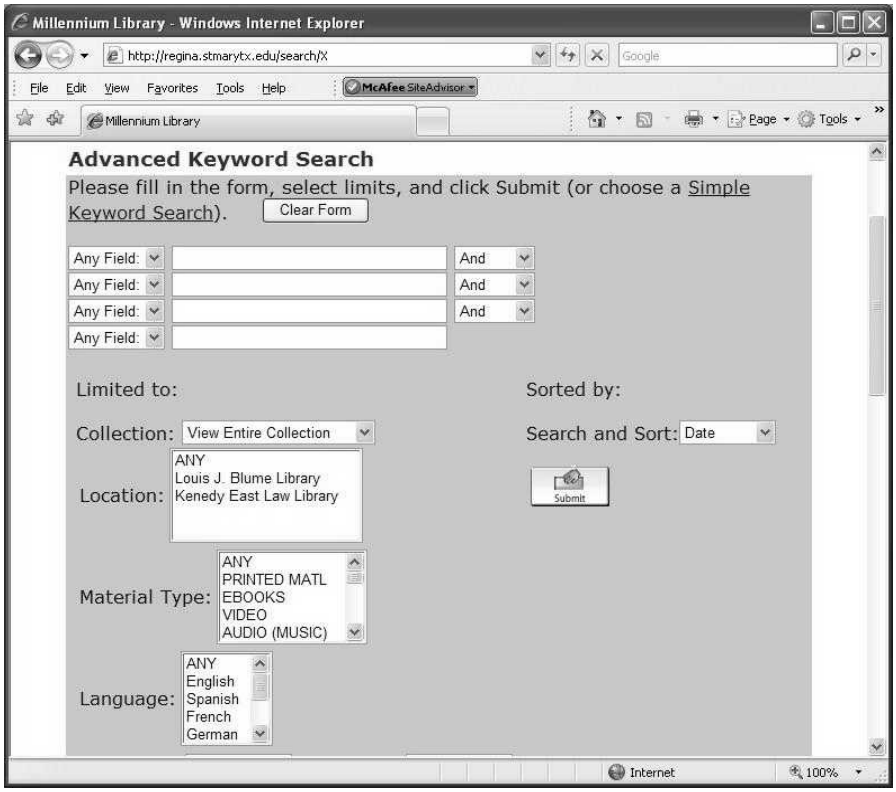


Fig 4.44
Library search web-based form within Microsoft Internet Explorer.

GROUP TASK Discussion
Evaluate the design and layout of each of the above data collection forms. In each case, propose possible improvements.

ANALYSING DATA OUTPUT FROM TRANSACTION PROCESSING SYSTEMS

Transaction processing systems contain large quantities of data that can be analysed to improve the organisation's performance. Past trends can be examined, the current state of the organisation's finances can be analysed and information can be used as evidence to assist decision makers. Such analysis can be performed on the operational data or on a data warehouse. In this section we consider data warehouses, management information systems and decision support systems used to analyse existing transaction data. Finally we consider enterprise systems, which are large systems that perform critical tasks for an organisation.

Data Warehouse

We first considered data warehouses in chapter 2. A data warehouse is a large database that includes historical copies of data from each of an organisation's operational databases. Data warehouses grow as new transaction data is added over time. The data warehouse is not in itself an analysis tool rather it is a data resource that analysis tools access to analyse the historical activities of the organisation.

Data warehouses are large snapshot copies of transaction databases, that is, they are static or read only in nature. This means analysis can take place without concern over simultaneous access or updating of transaction records. Furthermore the data warehouse can act as an archive for the organisation's historical data.

Advantages of data warehouses include:

- Old transaction data can be purged from the operational system and archived within the data warehouse. This improves the performance of the operational system, as less data needs to be examined during transaction processing.
- Analysis processes performed on the data warehouse do not degrade performance of the operational system. Data warehouses are generally maintained on their own hardware and software; hence they have no effect on the performance of the operational systems.
- A data warehouse includes historical transaction data, often over 10 or more years. Systems change completely and are regularly upgraded, however data warehouses are designed such that all data is stored using a similar format. This common format greatly simplifies analysis processes.
- Data warehouses are snapshot copies of the real data. This data does not and should not change. Therefore analysis processes can proceed more efficiently. There is no need to be concerned with record locks, ACID properties and data integrity issues.
- Data warehouses centralise data from within the entire organisation. Commonly this includes customer, sales, employee, payroll, production, marketing and any other data created within an organisation. Having all such data in a central repository means analysis can take place across the entire organisation.
- As a data warehouse is completely separate to the operational data it can be organised differently to the operational data. For instance indexes can be created on particular fields to improve the performance of analysis processes without risk of degrading the operational system's performance.

GROUP TASK Discussion

Often links are provided within the operational system that allow users to view data within the data warehouse. Propose specific scenarios where such access to historical data maybe an advantage.

Management Information Systems

A Management Information System (MIS) transforms data within transaction processing systems into information to assist in the management of business operations. MIS functions include the generation of sales reports, profit and loss statements, graphs of sales trends and a variety of other reports required for the day-to-day operation of organisations. Such reports are essentially summaries or statistical analyses of existing data within the system. These reports are used by managers to plan and direct the operation of the organisation.

In a small business such information is generated directly by the manager, whilst in larger organisations one or multiple departments are dedicated to MIS processes. In small systems the functions of the MIS are often contained within the transaction processing system whilst in larger systems the MIS is a separate system or systems. Large management information systems link to transaction data and perhaps to a data warehouse. For instance reports that compare current productivity with historical productivity require access to current transaction data and also to historical data within the organisation's data warehouse. Within large organisations MISs can include one dedicated to generating information to assist financial management, another to provide information to assist warehouse managers, another to assist production managers and yet another to provide information to assist marketing managers.

The participants who work within management information systems require strong technical computer skills together with a solid grasp of business processes. These personnel must transform the data within the system into information of relevance to decision makers within their organisations. This can only occur when a mix of technical IT skills and business knowledge is present.

Consider the following:

Each of the following is an example of information generated by an MIS. In each case the data source is ultimately transaction data.

- A list of each product a factory produces together with the profit or loss made on each over a 12 month period.
- A table listing each salesperson together with the total monthly value of their sales over the past 12 months.
- The total value of cheques for each bank that pass through a large cheque clearance facility on a particular day.
- A column graph displaying the total number of sick days taken by all employees on each day of the week.
- A line graph for each product showing average total number sold each month over a five year period.

GROUP TASK Discussion
For each of the above examples, identify the transaction data that has been analysed to create the information.

GROUP TASK Discussion
For each of the above examples, discuss how the information could be used by management to assist the operation of the organisation.

Decision Support Systems

A Decision Support System (DSS), like an MIS, provides information to managers to assist the decision making process. However decision support systems do much more than merely summarise current transaction data. The analysis performed by decision support systems presents possible solutions and is able to assess the likely consequences of making particular decisions. For example an MIS creates a graph summarising total sales made by each branch over the last month. A decision support system is used to determine possible reasons why particular branches had higher or lower sales totals. For instance, Online Analytical Processing (OLAP) systems (refer page 224) are a type of online decision support system that allow decision makers to drill down through different levels or dimensions within the data to uncover new relationships and other information. The results can then be used to improve future performance. In essence a decision support system can be thought of as an intelligent kind of MIS. Many decision support systems look to the future, they are able to generate forecasts and predictions based on historical or incomplete data. For example predicting future interest rates or forecasting the weather are problems that do not have a definite single correct solution. Decision support systems analyse the available data to produce or suggest the most likely outcomes.

The second option in this course (Chapter 5) deals exclusively with decision support systems, in this section we are concerned with how decision support systems are used to analyse data generated by transaction processing systems. Be aware that not all decision support systems analyse transaction data – there are various other possible data sources.

Decision support systems that analyse transaction data commonly use a data warehouse as their data source. Clearly the system's hardware and software must be capable of processing enormous amounts of data. Data mining is one decision support technique that examines the raw data in an attempt to discover hidden patterns and relationships. Data mining presents new information that was not originally intended to be present within the data. Creating and querying data marts is another decision support technique – data marts simplify and improve the efficiency of information extraction from large data warehouses. A data mart is essentially a reorganised summary of specific data from the data warehouse and/or transaction database. Each data mart aims to meet specific decision support needs of a particular department. A series of queries are executed either directly by users or via decision support software to retrieve information that assists decision makers.

Consider the following:

To create a data mart select queries are run that create summaries of the data in the transaction database or data warehouse and then the results of the query are used to create a new table within the data mart. For example a query that returns the number of each product sold per day could be used to create a new table. Within large data warehouses that contain many millions or even billions of records the creation of the new table will take some time – perhaps hours or even days. However this new data mart table will be reused and as it contains far less data it can be analysed more rapidly. Unfortunately whenever data is summarised some of the original detail (or granularity) is lost. Therefore such summaries must be chosen carefully so that required detail is retained.

Creating new tables for a data mart requires a corresponding reorganisation of the database schema. This reorganisation aims to optimise the schema for decision

support processing – the original schema was designed to optimise transaction processing. Often a simpler de-normalised schema based around one single large table is preferred for decision support. Many of the reasons for normalising databases are not present in data mart based decision support systems – existing data is never altered and new data is added in bulk. Even when the large table contains a summary of the raw data it can still include many millions of records, therefore the schema should be designed so that querying the large table only occurs when needed. One common strategy is to design the attributes of this large table as a series of foreign keys to smaller tables. For instance a BranchID would be linked to a Branch table that details the location, region and state for each branch. It is the detailed attributes within the smaller tables that are used within queries as the search, sort and grouping criteria. Notice that such a schema forms a star with the large table at the centre linking out to each of the smaller tables (refer *Fig 4.45*). Users are able to efficiently identify criteria for queries by examining the smaller tables. The query is then constructed using these criteria with joins to the larger table added later. Such a simple schema allows users to quickly produce ad-hoc queries without the need to understand the complexities of SQL statements needed to design queries with multiple joins.

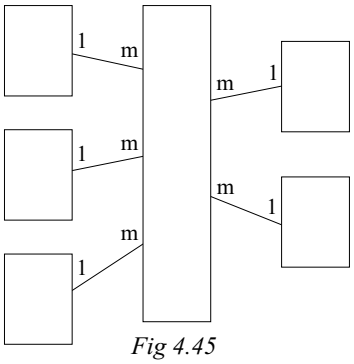


Fig 4.45
Typical star schema used for many data marts.

GROUP TASK Discussion
With regard to decision support systems, identify and list advantages of data marts compared to data warehouses.

GROUP TASK Research
Research, using the Internet or otherwise, specific examples of business decisions that have been made based on the analysis of historical transaction data.

Consider the following:

A supermarket chain has some 200 stores across Australia. Each store's transaction database includes a record for each individual product scanned through a register for each customer purchase. The chain's head office creates a data mart for use by its marketing department. Within this data mart a central table is created that contains a single record for the total number of each product sold each day within each of the 200 stores.

GROUP TASK Discussion
Propose examples of information that can be retrieved from the above data mart.

GROUP TASK Discussion
Identify and describe examples of information that CAN be derived from the transaction database but CANNOT be derived from the data mart.

Enterprise Systems

An enterprise is simply a large organisation, for example government departments, large corporations and universities. An enterprise system is any system that performs processes central to the overall operation of an enterprise. This includes critical hardware, critical software applications and in particular critical data. For instance, a typical university would have a variety of enterprise systems in operation, including a student records system, a finance system, a payroll system, a human resources system and also a content management system. Each of these enterprise systems is central to the running of the university and operates throughout the university.

Consider the following enterprise system case study:

Dimension Data

Customer Size: 8600 employees

Organization Profile

Founded in 1983 and headquartered in Johannesburg, South Africa, Dimension Data is a global IT provider and Microsoft Gold Certified Partner operating in 36 countries across five continents.

Business Situation

Dimension Data needed an enterprise-grade database that supported database mirroring for disaster recovery and database snapshots for reporting.

Solution

Dimension Data upgraded its existing SAP R/3 infrastructure to Microsoft SQL Server 2005 Enterprise Edition running on Microsoft Windows Server® 2003 Enterprise Edition operating system. The company moved to SQL Server 2005 to take advantage of new features and enhanced functions of the database—including the Database Snapshot and Database Mirroring features.

Dimension Data uses SQL Server 2005 Database Mirroring to maintain a continually updated copy of its data on a separate server at each data center. It plans to expand its use of Database Mirroring to include storing a continuously updated database at a geographically separate disaster recovery center.

The Database Snapshot feature of SQL Server 2005 is used for creating copies of the database throughout the day, both for location backup and as a reporting database so that queries can be run without impacting the production database.

A member of the HP Service Provider Program, Data Dimension supports its SAP infrastructure with HP ProLiant servers equipped with Intel Xeon processors. Intel Xeon processors offer an ideal choice for demanding enterprise applications such as SAP.

The SAP deployment architecture, which is identical for Johannesburg and London, includes:

- SAP R/3 data, totaling about 100GB s stored in a data warehouse running on SQL Server 2005 Enterprise Edition.
- Every three hours the Database Snapshot feature of SQL Server 2005 is used to create an updated copy of the SAP database.
- SQL Server 2005 Analysis Services is used to create two multidimensional data cubes, to support faster data access for analytics. The cubes are used by some analysts and other users.
- Dimension Data's worldwide workforce accesses SAP information by logging into a portal supported by Microsoft SharePoint® Portal Server. Microsoft Active Directory® directory service is used to help ensure information is accessible on a role-based basis. SAP data is accessed by about 1,600 users.

Fig 4.46
Modified extract of Dimension Data case study (Source: microsoft.com)

GROUP TASK Discussion
Explain how the mirroring and snapshot features of Dimension Data's new enterprise solution protects their critical data..

GROUP TASK Discussion
SAP, HP and Microsoft are major players in the enterprise system market. Research examples of enterprise systems that use these company's product

SET 4E

1. Ferromagnetic materials used within MICR ink and toner:
   (A) is magnetically charged.
   (B) can be magnetised.
   (C) are encoded with binary data.
   (D) are used during optical scanning.

2. Which of the following is true in regard to the operation of barcode readers?
   (A) Light is reflected off the barcode to one or more sensors.
   (B) Less light is reflected off dark colours.
   (C) The sensor(s) detect the intensity of reflected light.
   (D) All of the above.

3. In regard to the magnetic stripe on most ATM and credit cards, which of the following is true?
   (A) The stripe contains 2 tracks, however for most applications just one track contains data.
   (B) The stripe contains 3 tracks, however for most applications just one track contains data.
   (C) The stripe contains 3 tracks, however for most applications just one track is read.
   (D) The stripe contains 3 tracks, however for most applications two tracks are read.

4 Discovering hidden patterns and relationships within large stores of data is known as:
   (A) data mining.
   (B) data warehousing.
   (C) decision support.
   (D) forecasting.

5. MICR, barcode and magnetic stripe readers use which type of sensors respectively?
   (A) Magnetic, optical, magnetic.
   (B) Optical, optical, magnetic.
   (C) Magnetic, magnetic, magnetic.
   (D) Optical, optical, optical.

6. In general, labels and input fields on forms should be:
   (A) centred.
   (B) right justified.
   (C) left justified.
   (D) fully justified.

7 Check digits and characters encoded on magnetic stripes use:
   (A) odd parity.
   (B) even parity.
   (C) checksums.
   (D) CRCs.

8 In regard to the design of paper forms, which of the following is true?
   (A) The input field order is determined by the corresponding electronic data entry form.
   (B) The form should make extensive use of colour and graphics to motivated users.
   (C) All instructions should be included as a separate document.
   (D) Space for answers provides an indicator of the amount of information required.

9 Designing forms such that they present well in different fonts and screen resolutions is particularly important when designing:
   (A) web forms.
   (B) paper forms.
   (C) online forms.
   (D) forms within software applications.

10. Which of the following reports is most likely to be produced by a DSS rather than a MIS?
   (A) Total sales by branch over the last 6 months.
   (B) Average time to produce each product during the last week.
   (C) Table detailing predicted profits resulting from different upgrade options.
   (D) Line graph displaying the total sales of a product for each month in the previous year.

11. Define the following terms:
   (a) RFID (b) Barcode (c) Magnetic stripe (d) Data warehouse (e) Data mining (f) DSS (g) MIS (h) Enterprise system

12. Describe the operation of each of the following collection devices?
   (a) RFID reader (b) Barcode reader (c) Magnetic stripe reader

13. Contrast the design of paper-forms with the design of online/web forms.

14. A retailer sells personalised T-Shirts over the web. Customers upload their own image files, which are subsequently printed on the T-Shirts. T-Shirts are available in four sizes - S, M, L and XL. Cost is $30 for the first T-Shirt that uses a particular image and $20 for extra T-Shirts using the same image. $15 is charged per order to cover postage and handling.
   (a) Identify the data that needs to be collected to process a sale.
   (b) Design a suitable data entry screen.

15. Distinguish between Management Information Systems and Decision Support Systems. Include examples to illustrate your response.

ISSUES RELATED TO TRANSACTION PROCESSING SYSTEMS

There are numerous significant issues that should be considered when designing and operating transaction processing systems. In this section we restrict our discussion to issues in regard to:

- The changing nature of work.
- The need for alternative non-computer procedures.
- Bias in data collection.
- Data security, data integrity and data quality issues.
- Control and its implications for participants.

THE CHANGING NATURE OF WORK

The nature of work has seen significant change since the 1960s. These changes have been both in terms of the types of jobs available and also in the way work is undertaken. The widespread implementation of computer-based systems, and in particular transaction processing systems, has been the driving force behind most of these changes. In the early 1970s many thought that the consequence of new technologies would be a reduction in the total amount of work needing to be done, this has not occurred, rather new industries and new types of employment have been created. Many people are now working longer hours, in more highly skilled and stressful jobs than ever before.

Industries that once employed significant numbers of clerks have seen the greatest changes. The majority of tasks traditionally performed by clerks are now automated. Consider banks, transaction processing systems have largely replaced the numerous clerks that once worked within each branch. Furthermore the widespread use of ATMs, EFTPOS and credit cards mean customers rarely need to visit the bank. The data entry tasks performed by bank staff are now performed by the customer, in the case of ATMs and by retailers in the case of EFTPOS and credit card transactions. In recent years the Internet has changed how transaction data is collected and processed. It is now common to complete totally automated online purchases. No human employed by the retailer needs to have any direct interaction with customers during the transaction's processing.

THE NEED FOR ALTERNATIVE NON-COMPUTER PROCEDURES

What happens when a transaction processing system fails? Perhaps there is a power failure, lightning strike, fire, theft or communication lines are broken. Maybe the data within the system has been lost or some hardware components are inoperable. Recovery then involves purchasing replacement hardware, rebuilding systems and restoring data. This takes time and during this time an alternative mode of operation is required. For large centralised systems such problems are resolved by maintaining backup power generators and redundant communication lines at complete mirrored sites. For smaller systems, alternative non-computer procedures are needed if the organisation is to continue to operate. Commonly the only alternative is a return to paper based non-computer procedures whilst the problems are corrected.

Alternative non-computer procedures should be trailed and tested at regular intervals to ensure they operate as planned. In particular such tests should ensure all participants understand and are able to correctly implement the procedures. For example when banks supply retailers with EFTPOS terminals they commonly supply stock of manual paper forms. These paper forms allow the retailer to continue trading despite failure of the EFTPOS system. However sales assistants must know how to process sales using these paper forms – this requires training and regular testing.

Consider the following examples of system failure:

- A local post office is broken into and all computers are stolen. Upon phoning Australia post it is determined that it will be one week before replacements arrive.
- A thunderstorm disrupts the communication lines into a large warehouse. The warehouse is informed that the lines are unlikely to be restored for 3 days. The transaction processing systems at the warehouse receives and processes hundreds of orders per day that are subsequently shipped out by a fleet of 20 trucks.
- The ATMs outside a busy bank branch are ram raided and the cash boxes are stolen. It will take at least two weeks for replacement ATMs to be installed.

GROUP TASK Discussion
Propose possible non-computer procedures that could be used to minimise the effects of each of the above system failures.

GROUP TASK Discussion
Explain possible techniques that could be used to train participants and test the procedures proposed above.

BIAS IN DATA COLLECTION

Bias is an inclination or preference that influences most aspects of the collection process; the result of bias during collection is inaccurate data leading to inaccurate outputs from the system. Those involved in collecting data must aim to minimize the amount of bias present.

Bias
An inclination or preference towards an outcome. Bias unfairly influences the outcome.

When deciding on the data to collect bias can be introduced. Often incomplete data is collected with the aim of simplifying the system. For example it is common for loan applications to collect data on a person's income based entirely on their last few tax returns. This data is used to assess each individual's ability to repay the loan; the assumption being that an individual's income is likely to remain relatively constant over time. In fact many people, particularly those who own or operate businesses, are able to adjust their income to suit their expenses. By simply collecting past income data the success of each loan application is biased in favour of salary and wage earners at the expense of business owners.

Locating or identifying a suitable source of data for collection is another potential area where bias can occur. Often efficiency of data collection means that the cheapest or most available source of data is used rather than the best source of data. Consider surveys; the data source for all surveys should aim to be a representative sample of the entire population. However for ease of collection many organisations collect survey data from users over the Internet. Internet users, in most cases, are not a representative sample of the population; in general Internet users are younger, have higher incomes and possess higher technology skills than the general population. Consequently results derived from such surveys will not accurately reflect the entire population.

The collecting process itself should take into account the likely perceptions held by those on whom the data is collected. People answer questions and fill out forms differently based on their perception of how the data will be used. For example a survey conducted by the Australian Taxation Office is likely to yield different results

to a similar survey conducted by the Australian Bureau of statistics. Individuals would likely perceive the tax office as being interested in their individual responses whereas a survey conducted by the Australian Bureau of Statistics is more likely to be viewed as truly anonymous.

DATA SECURITY ISSUES

A summary of strategies we have examined to combat data security issues include:

- Passwords- Passwords are used to confirm that a user is who they say they are. Once verified the user name is then used by the system to assign access rights.
- Backup copies- A copy of files is made on a regular basis.
- Physical barriers- Machines storing important data and information, or performing critical tasks are physically locked away.
- Anti-virus software- All files are scanned to look for possible viruses. The anti-virus software then either removes the virus or quarantines the file.
- Firewalls- A firewall provides protection from outside penetration by hackers. It monitors the transfer of information to and from the network. Most firewalls are used to provide a barrier between a local area network and the Internet.
- Data encryption- Data is encrypted in such a way that it is unreadable by those who do not possess the decryption code.
- Audit trails- The transaction log includes details of who and when transactions were performed. It is possible to work backwards and trace the origin of any problem that may occur.

DATA INTEGRITY ISSUES

A summary of strategies we have examined to maximise data integrity include:

- Data validation- checks, at the time of data collection, to ensure the data is reasonable and meets certain criteria.
- Data verification- regular checks to ensure the data collected and stored matches and continues to match the source of the data.
- Referential integrity- ensuring all foreign keys in linked tables match a primary key in the related table.
- ACID properties- ensuring transactions are never incomplete (atomicity), the data is never inconsistent (consistency), transactions do not affect each other (isolation) and that the results of a completed transaction are permanent (durability).
- Minimising data redundancy- Normalising reduces or eliminates duplicate data within individual relational databases, however when transactions span multiple databases issues will arise. The use of unique identifiers shared between organisations allows individual entities to be accurately identified.

DATA QUALITY ISSUES

Data integrity is about the accuracy of the data – how well it matches and continues to match its source. Data quality takes this one step further, it concerns how reliable and effective the data is to the organisation. For example, responses on survey forms may well be entered accurately into a system, however the quality of the data will be poor if the respondents didn't answer honestly or as intended. The resulting information will be unreliable and ineffective. Other data quality issues occur when combining data from different systems.

Consider creating a data warehouse from many databases. Some records will describe the same entity differently; both may be correct, so which record is best? The organisation of databases is likely to be different; different keys, data types or schemas, for example. The meaning attached to an attribute can change over time; perhaps a client application was modified and now stores different data in some old

field. Combining such data is difficult, unreliable and inefficient. Furthermore the effectiveness and reliability of the information from subsequent data mining and OLAP systems is reduced. Data Quality Assurance (DQA) standardises the definition of data and includes processes that "scrub" or "cleanse" existing data so it meets these data quality standards.

Consider the following data security, integrity and/or quality issues:

- A hacker gains access to an organisation's system. They download customer credit card details and use them to make various purchases over the Internet.
- An RTA employee alters driving test results so that licences are issued to people who failed their driving test.
- A analyst using a data mining application uncovers links between sets of attributes that cannot possibly be true.
- A bank customer determines that a funds transfer has not been completed. The funds have left their account but have not been deposited into the other account.

GROUP TASK Discussion
For each of the above issues, determine the source of the issue and suggest suitable strategies that would assist in preventing the issue re-occurring in the future.

CONTROL AND ITS IMPLICATIONS FOR PARTICIPANTS

Control is the act of influencing or directing activities. In terms of managing the activities of employees some level of control is reasonable. Management assigns tasks and then quite legitimately expects employees to complete these tasks in a timely and accurate fashion. However whenever one has control over another the relationship is open to abuse. Determining precisely when control over participants is excessive is often a grey area. Most would agree it is reasonable for managers to monitor the activities of those they manage, however what level of monitoring is reasonable? Should management control Internet access or be able to read all email messages? Is it reasonable to monitor phone calls or remotely view a users desktop? Audit trails allow management to track which records individuals have accessed; when is such tracking reasonable? Answers to such questions differ considerably according to the management style used and the nature of the tasks participants perform.

Current management theory suggests higher levels of productivity are achieved when participants are motivated. Motivation improves when participants are given responsibility for tasks and how they are completed. Motivated employees are less likely to engage in undesirable activities and are much more likely to focus on work. When employees are assigned boring or repetitive tasks they lose motivation and then quite naturally seek to engage in other non-work related activities. When this occurs management too often imposes authoritative controls such as excessive monitoring in combination with negative consequences in an attempt to enforce control. Such measures further reduce motivation resulting in even stricter controls being imposed – a downward trend emerges. A more sustainable management style encourages trust and motivates employees to take responsibility for work they complete.

GROUP TASK Discussion
As an employee, what level of monitoring by management do you feel comfortable with? Brainstorm scenarios where monitoring and control of participants is necessary (or at least justified)

CHAPTER 4 REVIEW

1. One operation within a transaction fails, what should occur?
   (A) Other operations within the transaction should be committed.
   (B) The system should halt so that the reason for the failure can be corrected.
   (C) All operations within the transaction should be rolled back.
   (D) No further transactions should be performed until the problem is resolved.
2. Participants are those people who:
   (A) are the source of data used by the system.
   (B) receive information output from the system.
   (C) interact directly with the system.
   (D) analyse data within the system.
3. Transaction logs used by most DBMSs include details of records:
   (A) prior to being altered.
   (B) after they have been altered.
   (C) added and deleted.
   (D) All of the above.
4. The file used to store data collected prior to batch processing is commonly called:
   (A) an error file.
   (B) a master file.
   (C) a database.
   (D) a transaction file.
5. Checks to ensure data entered is reasonable are known as:
   (A) data validation checks.
   (B) data verification checks.
   (C) data integrity checks.
   (D) data redundancy checks.
6. Which ACID property ensures either all or no operations within a transaction are committed?
   (A) Atomicity
   (B) Consistency
   (C) Isolation
   (D) Durability
7. Strict sequential processing of transactions ensures which ACID property is observed?
   (A) Atomicity
   (B) Consistency
   (C) Isolation
   (D) Durability
8. What is the main task performed by TPMs?
   (A) Providing an interface between many transaction processing systems.
   (B) Ensuring transactions performed on a database observe the ACID properties.
   (C) Monitoring and ensuring the security of transactions.
   (D) Managing transactions that span multiple database, systems and client applications.
9. At most two sets of backups will be required to completely restore data when which of the following backup types are used?
   (A) Full and incremental.
   (B) Full and differential.
   (C) Incremental and differential.
   (D) Full backups only.
10. High speed MICR readers use which technique to read the MICR line on cheques?
    (A) waveform
    (B) matrix
    (C) CCD
    (D) LED

11. Provide at least TWO examples of systems where each of the following devices is used:
    (a) MICR
    (b) Barcodes
    (c) Magnetic stripes
    (d) RFID readers and tags
    (d) TPMs
    (e) Tape libraries
    (f) Touch screens

12. Compare and contrast each of the following:
    (a) User interfaces for real time processing with user interfaces for batch processing.
    (b) Random (or direct) access with sequential access.
    (c) Data validation with data verification.
    (d) OLAP (Online Analytical Processing) and OLTP (Online Transaction Processing).
    (e) Data warehouses and data marts.
    (f) Data integrity with data quality.

13. (a) Recount the sequence of processes occurring to complete a typical credit card transaction. Assume the transaction is initiated using an EFTPOS terminal supplied by the retailer's bank.
    (b) Describe different uses of transaction logs within transaction processing systems.
    (c) Distinguish between the storage of collected data and the storage of processed data in a batch transaction processing system using an example.

14. A company's mail server records each email sent or received in a separate file. Incremental backups using a round robin rotation occur automatically every hour to an online tape library. All employees have full access to files within the tape library. Full backups are not made, however all archive bits were set to true when the system was first installed. Tapes are changed every year as there is sufficient capacity to store messages for 12 months.
    (a) Critically evaluate the above backup procedure.
    (b) Predict issues that may occur as a consequence of the above backup procedure.
    (c) Propose and justify an improved procedure for backup and recovery.

15. Analyse an online web-based purchasing system of your choice.
    (a) Determine the data items collected,
    (b) Identify the operations performed during a purchasing transaction, and
    (c) Evaluate the design of the data collection web forms.
    (d) Explain how the company could analyse the collected data to identify areas for improving its operations.

In this chapter you will learn to:

- select and recommend situations where decision support systems could be used
- classify situations which are structured, semi-structured or unstructured
- identify participants, data/information and information technology for an example of a decision support system
- describe the relationships between participants, data/information and information technology for an example of a decision support system
- analyse trends and make predictions using an existing spreadsheet model
- extract data, based on known criteria, from an existing database to help make a decision
- recognise appropriate decision support systems for a given a situation
- design spreadsheets by:
  - linking multiple sheets to extract data and create summaries
  - using absolute and relative references in formulae
- implement spreadsheets by:
  - entering data
  - naming ranges
  - creating templates
  - organising data for easy graphing
  - using formulae to link and organise data in cells
- design a set of if-then rules for a particular situation
- diagrammatically represent the if-then rules
- enter rules and facts into an expert system shell and use it to draw conclusions or make a diagnosis
- describe situations better suited to forward chaining and those better suited to backward chaining
- create a simple macro in a spreadsheet
- compare and contrast processing methods used by databases, neural networks and expert systems
- describe the process of data mining to search large databases for hidden patterns and relationships and use these to predict future behaviour
- analyse alternatives using 'what-if' scenarios
- make predictions based on the analysis of spreadsheets
- use a simple neural network to match patterns
- extract information from a database for analysis using a spreadsheet, including charting relevant data
- distinguish between neural networks and expert systems
- describe tools used for analytical processing
- determine the sources of data for a decision support system for a given scenario
- describe the operation of intelligent agents in situations such as search engines for the Internet
- describe the impact on participants in decision support systems when some of their decision-making is automated and recommend measures to reduce negative impacts
- identify situations where user(s) of decision support systems also require knowledge in the area
- determine whether the decisions suggested by intelligent decision support systems are reasonable
- demonstrate responsible use of a decision support system by using its findings for the intended purpose only
- identify situations where decision support systems are of limited value
- recognise the importance of business intelligence based on enterprise systems

Which will make you more able to:

- apply and explain an understanding of the nature and function of information technologies to a specific practical situation
- explain and justify the way in which information systems relate to information processes in a specific context
- analyse and describe a system in terms of the information processes involved
- develop solutions for an identified need which address all of the information processes
- evaluate and discuss the effect of information systems on the individual, society and the environment
- demonstrate and explain ethical practice in the use of information systems, technologies and processes
- propose and justify ways in which information systems will meet emerging needs
- justify the selection and use of appropriate resources and tools to effectively develop and manage projects
- assess the ethical implications of selecting and using specific resources and tools, recommends and justifies the choices
- analyse situations, identify needs, propose and then develop solutions
- select, justify and apply methodical approaches to planning, designing or implementing solutions
- implement effective management techniques
- use methods to thoroughly document the development of individual or team projects.

In this chapter you will learn about:

Characteristics of decision support systems

- decision support systems - those that assist user(s) in making a decision
- the interactive nature of decision support systems
- the nature of decision support systems which model, graph or chart situations to support human decision making

Categories of decision making

- structured:
  - decisions are automated
  - decision support systems are not required
- semistructured:
  - there is a method to follow
  - requirements are clear cut
- unstructured:
  - there is no method to reach the decision
  - judgements are required
  - requires insights into the problem

Examples of decision support

- semistructured situations, such as:
  - a bank officer deciding how much to lend to a customer
  - fingerprint matching
- unstructured situations, such as:
  - predicting stock prices
  - disaster relief management
- the use of systems to support decision making, including:
  - spreadsheets
  - databases
  - expert systems
  - neural networks
  - data warehouses
  - group decision support systems
  - Geographic Information System (GIS)
  - Management Information Systems (MIS)

Organising and decision support

- designing spreadsheets:
  - creating a pen and paper model
  - identifying data sources
  - planning the user interface
  - developing formulas to be used
- the knowledge base of if-then rules in an expert system

Processing and decision support

- structure of expert systems
  - knowledge base
  - database of facts
  - inference engine
  - explanation mechanism
  - user interface
- types of inference engines, including:
  - forward chaining
  - backward chaining
- certainty factors as a means of dealing with unclear situations
- pattern matching in neural networks
- the use of macros to automate spreadsheet processing

Analysing and decision support

- data mining
- extracting summary data from a spreadsheet
- comparing sequences of data for similarities and differences
- spreadsheet analysis, including:
  - what-if models
  - statistical analysis
  - charts
- On-line Analytical Processing (OLAP)
  - data visualisation
  - drill downs

Other information processes

- collecting
  - identification of data for decision support systems
  - the role of the expert in the creation of expert systems
  - the role of the knowledge engineer in the creation of expert systems
- storing and retrieving using intelligent agents to search data

Issues related to decision support

- the reasons for decision support systems, including:
  - preserving an expert's knowledge
  - improving performance and consistency in decision-making
  - rapid decisions
  - ability to analyse unstructured situations
- responsibilities of those performing data mining, including:
  - erroneous inferences
  - privacy
- responsibility for decisions made using decision support systems
- current and emerging trends of decision support systems including
  - data warehousing and data-mining
  - Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP)
  - group decision support systems and the communication it facilitates

5

OPTION 2
DECISION SUPPORT SYSTEMS

Decision Support Systems assist people in making decisions. A decision occurs when a choice is made between two or more alternatives. The alternatives aim to meet some objective or goal – presumably some alternatives will prove to be better than others. Decision Support Systems can assist in generating possible alternatives, however more importantly they provide mechanisms for assessing and predicting how successfully each alternative is likely to meet the problem's objective or goal. Decision Support Systems supply evidence to assist decision makers determine alternatives and then prioritise one alternative over other possible alternatives.

Decision
A choice between two or more alternatives. Committing to one alternative over other alternatives.

A decision occurs when a decision maker commits to one alternative. The decision results in resources being allocated and some activity occurring to implement the chosen alternative. Once a decision is implemented then uncertainties come into play. Uncertainties are the uncontrollable elements that affect the ultimate achievement of the objective or goal. The selected alternative together with any uncertainties combines to produce the final outcome. The final outcome may totally achieve, partially achieve or it may totally fail to achieve the goal. Decision Support Systems attempt to predict uncertainties using various techniques such as "rules of thumb", certainty factors, the experience of experts and statistical analysis of historical data. These techniques do not alter the uncertainty; rather they attempt to predict the uncertainty by reporting the range of likely outcomes or the probability of each occurring.

GROUP TASK Discussion
What is one and one? Possible alternatives include 1, 2, 10 and 11. Explain how each of these alternatives is possible. Prioritise the alternatives in order from most to least likely. Decide on one alternative.

GROUP TASK Discussion
Identify and describe the uncertainty that makes each of the above answers possible. If your decision is later found to be incorrect, does this mean your decision was wrong? Discuss.

Decision-making is critical when solving all types of problems, however for many problems decision-making is a difficult and imprecise task. Decision support systems aim to simplify the decision-making process by automating the assessment of different alternatives or conclusions. The solution to some problems can be clearly and definitely determined, which implies all variables are clearly and thoroughly understood. Such structured situations do not require decision support systems as the best alternative can be objectively determined. Indeed these structured decisions can be totally automated. Many other decisions are less precise. The variables are unknown or it is not possible to be certain about their value or influence. Decision

support systems are most useful in semi-structured situations where some mix of certainty and uncertainty is present. Unstructured situations are those characterised by significant or even complete uncertainty, therefore determining, recommending and prioritising alternatives is particularly difficult. In these situations there is no structured method for reaching a decision, there are too many variables, many are unknown and their interactions are highly complex and poorly understood. For these rather unstructured situations decision support systems are often designed to simulate the human brain. The aim being to assess the situation using insight, intuition and judgements much like a human thinker.

We can think of structured, semi-structured and unstructured situations as lying on a continuum (refer *Fig 5.1*). More structured decisions can be made reliably using machines, whilst at the other end of the continuum are totally unstructured decisions that require human intuition, feelings, emotions and insight. For example finding the average of a set of numbers is highly structured whilst deciding on the merits of a piece of art is highly unstructured. Decision support systems are most useful when the decision lies somewhere between these two extremes.

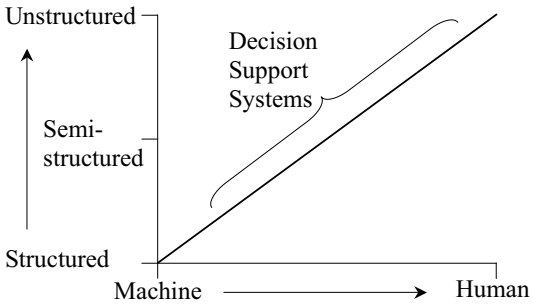


Fig 5.1
Decisions lie on a continuum. Decision support systems are most useful when the decision lies between extremes.

Consider the following:

- A business owner is trying to decide which of two products they should produce and market. Both products require an initial investment of $100,000 and there are insufficient funds to produce both products. It is determined that the chance of product A failing is virtually zero, however it is also unlikely that it will make a substantial profit. Most likely product A will make a comfortable profit. On the other hand product B is a far riskier alternative. It has a significant chance of total failure, however it is also fairly likely that it will produce significantly larger profits than product A.
- Doctors perform tests and examinations and they ask patients questions. They do this in an attempt to diagnose (or decide on) the nature of the illness. Once the most likely illness is determined the doctor decides on the most suitable treatment. They may prescribe medication and suggest diet or lifestyle changes in an attempt to cure the diagnosed illness.

GROUP TASK Discussion
Identify the alternatives present in each of the above decisions.
Discuss the different data and information that is likely used to formulate these alternatives.

GROUP TASK Discussion
In terms of the structured/unstructured continuum, think about how the alternatives are or can be prioritised. Does the level of uncertainty influence the decision? Discuss.

CHARACTERISTICS OF DECISION SUPPORT SYSTEMS

Decision support systems (DSS) are a form of information system that assist users make decisions. The user is involved in the decision making process, they answer questions posed by the system and they control the operations performed. Most decision support systems are interactive – they require input and direction from users.

The following seven points describe typical characteristics of most decision support systems. This list is based on *Power, D., What are the characteristics of a Decision Support System? DSS News, Vol. 4, No. 7, March 30, 2003*. Daniel Power operates the DSSResources.com website – a valuable resource for anybody with an interest in decision support systems.

1. Facilitation. DSS facilitate and support specific decision-making activities and/or decision processes. That is, DSS simplify and make easier the process of decision making.
2. Interaction. DSS are computer-based systems designed for interactive use by decision makers or staff users who control the sequence of interaction and the operations performed. The participants use the DSS interactively – the DSS seeks data and requires guidance during its execution.
3. Ancillary. DSS are not intended to replace decision makers rather they are tools that assist decision makers. The information output from a DSS is used as evidence to help and direct decision makers rather than being an absolute decision making solution.
4. Repeated Use. DSS are intended for repeated use. A specific DSS may be used routinely or used as needed for ad hoc decision support tasks. The effort and costs associated with the design and development of a DSS is substantial. Such costs are justifiable when the DSS can be reused to assist in the support of similar decisions.
5. Task-oriented. DSS provide specific capabilities that support one or more tasks related to decision-making. These tasks may include intelligence and data analysis, identification and design of alternatives, choice among alternatives and decision implementation.
6. Identifiable. DSS are information systems in their own right, they have a distinct and clear purpose. DSS may be independent systems that collect or replicate data from other information systems or they can be subsystems within a larger, more integrated information system.
7. Decision Impact. DSS positively contribute and affect the decision making process. DSS are intended to improve the accuracy, timeliness, quality and overall effectiveness of a specific decision or a set of related decisions.

Decision support systems use a combination of models, analytical tools, databases and automated processes to assist decision-making. Computer models are a simulation of a real system. For example weather forecasters build complex computer models that attempt to simulate and predict the behaviour of weather. Models use various analytical tools to process data. For example a spreadsheet includes many statistical functions that can be applied to historical data in an attempt to predict future behaviour. The analytical tools operate on data, often from a database, but other data sources such as documents or rules can be used. Within many DSS automated processes are used. Automated processes within DSS commonly attempt to simulate human intelligence – in particular human decision making processes. For example expert systems model the reasoning of a human expert and neural networks are able to learn and make decisions by detecting patterns within data.

EXAMPLES OF DECISION SUPPORT SYSTEMS

In this section we consider examples of semi-structured and unstructured situations where decision support systems are routinely used. We restrict our discussion to a general overview of each DSS example rather than a detailed discussion of the information processes and information technology used. Our aim is to introduce situations where DSS are used and in each case identify the participants, data/information and possible information technology together with the relationships between these system resources.

SEMI-STRUCTURED SITUATIONS

Semi-structured situations are those where the requirements that must be met to make a decision are clearly understood and well defined. Furthermore there is a recognised method or sequence of steps that can be followed to determine if the requirements for the decision have been met.

Approving Bank Loans

When a bank is deciding how much money to lend to a customer they are really making a prediction in regard to how confident they are that the customer will be able to make repayments. Furthermore they are predicting the likely consequences for the bank should the customer fail to meet their repayments.

There are three basic requirements used by most banks when assessing loans:

1. The customer's income is sufficient to meet the regular loan repayments.
2. The customer's income will continue at current levels for the term of the loan.
3. The bank will be able to recover their funds if the customer is unable to meet their repayment obligations.

The bank must be sufficiently satisfied that all three criteria are met before they will approve a loan. For example a customer that has just started work in a higher paid job may now have an income able to satisfy the first criteria and they may only be asking to borrow 50% of the purchase price of a home. However as the customer has no history of earnings at this higher level they may fail to satisfy the second criteria and therefore the loan is refused.

GROUP TASK Practical Activity
Represent the above three criteria for assessing a loan using a decision table. There are two possible actions either the loan is approved or the loan is refused.

Consider how a bank officer can assess the validity of each of the three criteria. For each criteria a series of rules are developed where each rule is evaluated using data specific to the individual loan and customer. Let us assume the loan is for a home where the customer will live, although similar rules could be established for other purchases, such as cars, holidays or investment loans.

1. The customer's income is sufficient to meet the regular repayments.

A possible (and common) rule of thumb used by many banks when assessing home loans requires that the regular payment amount is less than or equal to 35% of the customer's gross income. Such a simple rule does not account for existing loans, bills and other regular expenses the customer may have. Furthermore customers must have sufficient funds remaining from their income to pay for incidental weekly expenses such as groceries, petrol, clothes and so no. For our purpose let us simplify our system by adding an additional rule. After subtracting tax, the loan repayment and other

regular expenses from the customer's weekly gross income, at least $100 plus an extra $50 for each dependant must remain to cover incidental weekly expenses. Our rules are more logically stated within the decision table in *Fig 5.2* below.

Conditions	**Rules**			
Weekly Loan Repayment <= 35% of Weekly Gross Income	✗	✓	✗	✓
Weekly Gross Income – Weekly Tax – Weekly Loan Repayment – Other Regular Weekly Expenses >= $100 + ($50 × Number of Dependants)	✗	✗	✓	✓
Actions				
The customer's income is sufficient to meet the regular loan repayments.	✗	✗	✗	✓

Fig 5.2
Decision table showing rules for assessing criteria 1 when approving a home loan.

Analysing the above decision table we find a total of five data items are required to assess the first of our three criteria. Let us consider the source of each of these data items.

- Weekly Loan Repayment – calculated using the loan amount and term of the loan requested by the customer together with the current interest rate charged by the bank. If the customer fails to meet the criteria the loan amount can be lowered in an attempt to approve the loan.
- Weekly Gross Income – collected directly from the customer and requires supporting documents to verify correctness.
- Weekly Tax – can be calculated based on income or collected directly from customer pay slips or tax office documents.
- Other Regular Weekly Expenses – collected directly from the customer and requires supporting documents to verify correctness.
- Number of Dependants – collected directly from the customer and requires supporting documents to verify correctness.

GROUP TASK Discussion
Propose techniques and documents suitable for verifying the data supplied by customers on loan applications.

2. The customer's income will continue at current levels for the term of the loan.

There is no way of knowing what a customer's future income will be, hence most banks use customer's employment history as an indicator of likely future income. If a customer has worked for the same employer for the past 20 years then they are more likely to continue to be employed in this job for the foreseeable future. On the other hand a customer who has recently (and regularly) changed jobs is a riskier proposition, particularly if their income has varied considerably.

Commonly banks require a customer's last two tax returns. The bank averages the income declared on these tax returns and compares the result to the customer's current income. The aim is to determine how secure the customer's income has been in the past. The assumption being that past income security is a strong indicator of future income security.

Customers who own and operate businesses or have various other sources of income are often able to adjust their personal income to meet their expenses. For such individuals personal tax returns can be misleading indicators of likely future income. In these cases banks require business tax returns and other financial documents as evidence to predict likely or possible future income.

GROUP TASK Discussion
Suggest possible data and rules that could be used to predict whether a customer's income will continue at current levels for the term of the loan. Consider other possible data and rules in addition to past income.

3. The bank will be able to recover their funds if the customer is unable to meet their repayment obligations.

If the customer is unable to make repayments then the bank needs to be confident that they can recover their funds. Possible reasons for customers defaulting on repayments include unemployment, death or disablement, rises in interest rates and a variety of other financial difficulties. Ultimately banks are businesses that aim to make profits for their shareholders, they are obliged to ensure that funds they lend can be recovered in the unfortunate event that the customer is unable to make their repayments.

The primary technique for ensuring the banks funds are recoverable is to take out a mortgage over the property – virtually all home loans require a mortgage. A mortgage is a legal pledge that essentially means the customer offers the property as security should they default on their loan obligations. In effect a mortgage means the bank can sell the property should the customer fail to make their loan repayments.

A mortgage does not protect the bank's funds if property prices fall. To account for this possibility most banks calculate a loan to value ratio (LVR) to assess their ability to recover funds. The LVR is the percentage of the value of the property that has been loaned. For example if a property is valued at $300,000 and the customer wishes to borrow $240,000 then $240,000 divided by $300,000 produces an LVR of 80%. Commonly banks are happy to fund loans where the LVR is less than or equal to 80%. When the LVR exceeds 80% most banks require the customer to pay for lenders' mortgage insurance. Lenders' mortgage insurance (LMI) covers the bank for any short fall between the sale price of the property and the balance of the loan account. Currently LMI costs between 1% and 3% of the purchase price of the property – the amount increases as the LVR increases. In general most banks do not approve loans where the LVR exceeds 95%. A decision tree based on the above LVR and LMI discussion is reproduced in *Fig 5.3*.

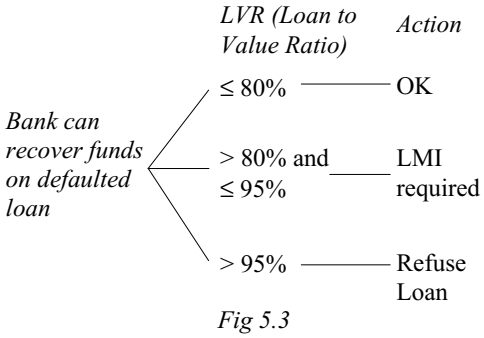


Fig 5.3
Decision tree showing rules for assessing criteria 3 when approving a home loan.

Notice that all the criteria and rules we have discussed have been determined precisely. These rules combine to describe a method for solving the problem and hence making a decision whether to approve the loan. Furthermore the data used to assess each criteria is well understood and defined. Such characteristics are typical of all semi-structured situations.

GROUP TASK Discussion
Propose suitable software that could be used to implement the above decision support system.

GROUP TASK Discussion
Identify the people involved during the operation of the above loan approval system. Who are the system's participants? Discuss.

Fingerprint Matching

There are numerous types of biometrics used to identify individuals including fingerprints, DNA, face, ear, retina, iris, hand veins, voice patterns and also signatures. Signatures are used extensively however they are relatively easier to forge compared with other biometrics. Many biometrics are difficult to collect and complex to analyse, such as DNA. Fingerprints have been used to identify individuals since the late 1600s and more recently have become a common biometric used to authenticate computer users.

It is theoretically possible for two individuals to have the same fingerprint, however the probability of this occurring is so small that it is reasonable to assume that all fingerprints are unique identifiers. Fingerprints form prior to birth and develop using a combination of genetic and environmental factors within the womb – even identical twins have different fingerprints.

There is a significant difference between authenticating (verifying) that a person is who they claim to be and attempting to identify an individual by comparing their fingerprint to a large database of fingerprints. When using a fingerprint for authenticating a user the fingerprint replaces a traditional password as described on the left in *Fig 5.4*. The user enters their username and their fingerprint is scanned. A single comparison is made between the scanned fingerprint and the existing fingerprint stored alongside the username. A single decision is required, either the fingerprints are sufficiently similar or they are not. For criminal investigations and other identification systems a single fingerprint is compared to a database of fingerprints (flowchart on the right in *Fig 5.4*). In this case many thousands of comparisons maybe required in an attempt to identify an individual – the FBI maintains fingerprint records for more than 200 million individuals.

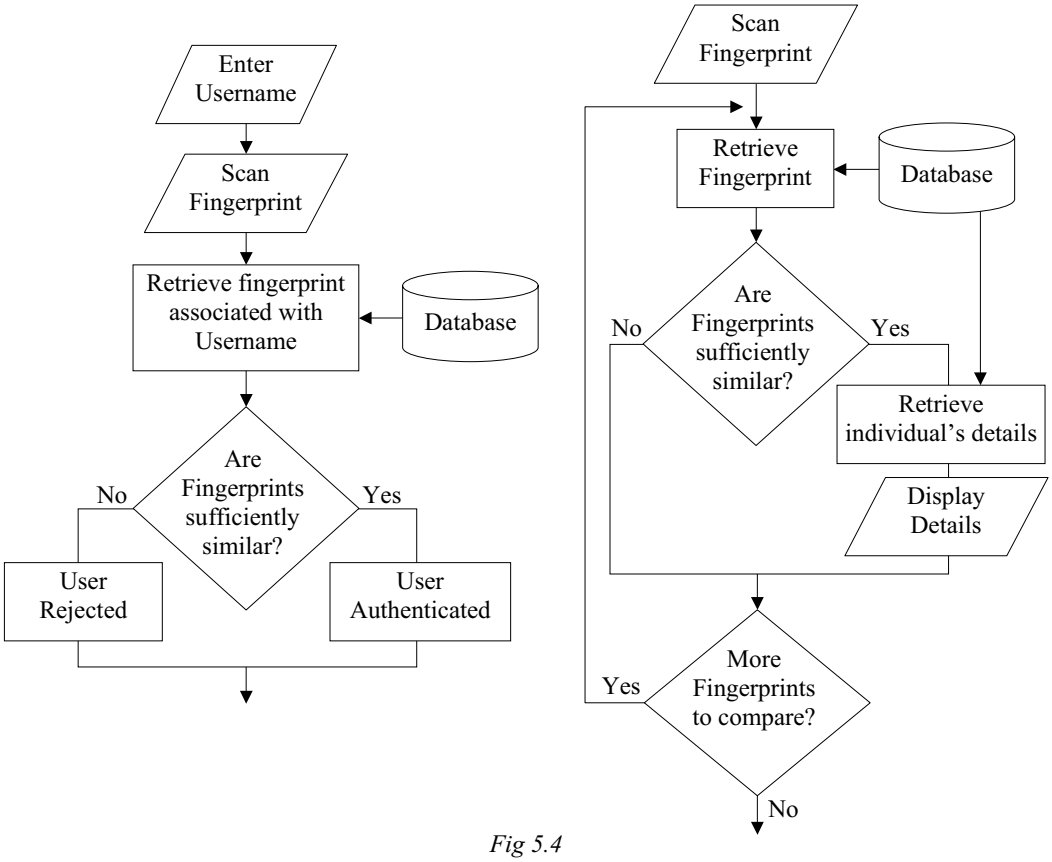


Fig 5.4
Flowcharts modelling authenticating (left) and identifying (right) using fingerprints.

GROUP TASK Discussion
An ATM card is something you have, a PIN number or password is something you know, whilst a fingerprint is something you are. You can change things you have or know but not something you are.
Discuss in relation to security and also privacy.

In terms of fingerprint matching decision support systems the significant decision is deciding whether two fingerprints are from the same finger on the same individual. In *Fig 5.4* we expressed this decision using the question "Are fingerprints sufficiently similar?" The meaning of "sufficiently similar" varies according to the ultimate purpose of the system. In criminal trials many more similarities between the two fingerprints must be present compared to a system that authenticates users within say a library loans system.

Investigators preparing evidence for criminal trials use a wide range of techniques and strategies for comparing fingerprints, we restrict our discussion to techniques used by computer fingerprint matching systems. Computer matching techniques are largely based on one of three techniques.

1. Identifying minutiae and comparing their relative positions. Minutiae are local occurrences of specific features. This is the most common technique used by the large majority of fingerprint matching systems. In most systems the minutiae identified are restricted to ridge endings and ridge bifurcations as shown in *Fig 5.5*. The location and direction of each minutia is recorded together with their position relative to each other or to some obvious feature. In *Fig 5.6* each minutiae is indicated by a circle and a small line indicating its direction. The two squares indicate a significant feature – the details of each minutiae are stored relative to this feature. When a new user is enrolled into the system the details of the minutiae are determined and stored as a template – the scanned fingerprint image is no longer required. When a user is being authenticated minutiae in the newly scanned image are compared to the user's existing fingerprint template.
2. Ridge feature matching is used for systems where the resolution of images is insufficient to accurately determine minutiae. Such systems record more general features such as ridge shape, number of ridges and orientation of ridges. Today such techniques are rarely used for computer applications however they remain an important technique for criminal investigations where poor quality fingerprints are lifted from articles at a crime scene.

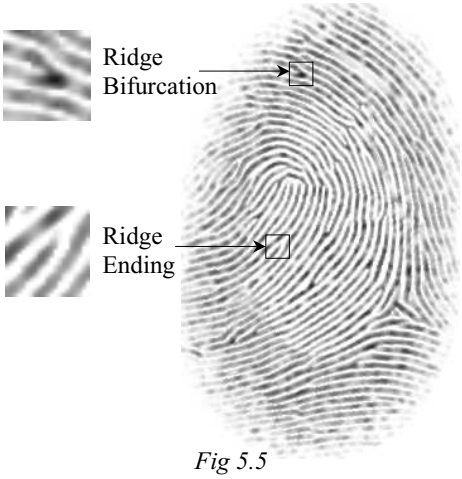


Fig 5.5
Examples of minutiae determined by many fingerprint matching systems.

Fig 5.6
Minutiae identified within a typical fingerprint image.

3. Comparing the images or bitmaps of the fingerprints directly. This involves translating and rotating the images and then superimposing the images over each other. The intensity of pixels at corresponding positions within each image is compared. Such systems require that images of fingerprints are stored within the system and hence significant storage is required. Furthermore such systems are less accurate as they are more susceptible to poor image quality, sweat, finger pressure, background lighting and other factors that affect image quality.

Fingerprint templates based on identifying and comparing the relative position of minutiae (the first technique described above) commonly includes details of some 200 minutiae per fingerprint. However for a successful match to take place far fewer minutiae need to be matched – typically just 10 to 20 matches are required to positively authenticate the user. Some smartcards contain templates of the owner's fingerprints. In this case storage is limited and the templates often contain details of just 40 of the most significant minutiae. *Fig 5.7* shows an integrated Smartcard reader and fingerprint scanner unit used to activate a door lock. This system does not require the device to be connected to a host computer. The template stored on the Smartcard is compared to the scanned fingerprint.

Fig 5.7
Smartcard reader and fingerprint scanner used to operate a door lock.

GROUP TASK Research
Research a number of fingerprint authentication systems. Determine if minutiae are used and also determine the amount of storage required for each fingerprint template.

GROUP TASK Discussion
Identify and briefly describe the data and information technology required to implement fingerprint matching as the technique for authenticating users of a single computer.

UNSTRUCTURED SITUATIONS

Unstructured situations are those where requirements upon which the decision is based are less clear and there is no definitive method for reaching a decision. Such decisions require human qualities such as insight and judgements to be made. Often the resulting decision is made based on available evidence, experience and understanding.

Predicting Stock (Share) Prices

Shares are initially issued by public companies to raise funds to finance their business operations – this is known as a float or initial share offering. Existing shares in companies are traded between the current owner (seller) and buyers. Individuals (or companies) purchase shares in a company with the expectation they will later be able to sell them to some other individual (or company) at a higher price. Essentially the seller and buyer agree on a price and the shares are sold (traded) for the agreed sum of money.

In Australia shares in all public companies are traded at the Australian Stock Exchange (ASX) – other countries have their own stock exchanges. Individuals (and companies) buy and sell shares in public companies via stockbrokers. Stockbrokers process the trade of shares on behalf of buyers and sellers. For instance Fred may wish to sell 1000 shares in ABC Ltd. at a price of $7.00 per share. Fred's stockbroker enters details of his requested sell order into the ASX system. Jack on the other hand wishes to purchases 1000 shares in ABC Ltd. and is willing to pay up to $7.10 per share. Jack contacts his stockbroker who enters Jack's buy order into the ASX system. The ASX system matches sell and buy orders on a first in first served basis. In our example Fred and Jack's orders are linked and the sale is processed at a price of $7.00 per share – Jack pays Fred $7000 and ownership of the shares is transferred from Fred to Jack.

As with any purchase the buyer wishes to buy at the lowest price and the seller wishes to sell at the highest possible price. The stock market, like most markets, operates on the principle of supply and demand. If there is strong demand for a company's shares and few existing shareholders wish to sell then sellers can raise their selling price. Conversely if few people wish to purchase then sellers will have to lower their price if they are to complete a sale.

Deciding on which company's shares to buy and sell and precisely when to buy and sell them is critical. This is not a simple decision – it involves predicting the future. To make matters even more difficult it involves predicting the future better than others who are trading. Trading on the stock market is often referred to as a game, where the aim is to outsmart the opposition. Buyers are willing to pay a higher share price because they predict future price rises. At the same time sellers are only willing to sell when they predict that the share price has reached a peak and is likely to fall. Various different decision support systems are used by traders in an attempt to predict market rises and falls better than other players in the market.

Some of the data inputs to stock market prediction decision support systems include:

- Past sale prices and quantity of shares traded for each public company's shares. The monthly, weekly, daily and even hourly highest and lowest sale prices are freely available in daily newspapers and online from the ASX.
- Various data specific to individual companies. The aim is to predict whether a company is likely to increase or decrease its profits. Perhaps they have just acquired new assets or they have a new board of directors. Some analysts consider and track the past performance of chief executive officers (CEOs) and other high level management.
- Industry specific data. For example changes in the reserve bank's interest rates cause a corresponding change in mortgage rates. When mortgage rates rise people have less money to spend on retail goods, resulting in lower retail company share prices. Share prices for companies who import or export goods are more likely to be influenced by changes in currency exchange rates than companies who trade solely within Australia.
- Overall historical measures of stock market performance. In Australia the All Ordinaries (All Ords) is a measure of the performance of a sample of major companies listed on the ASX. Other stock markets throughout the world generate similar measures, such as the Dow Jones for the New York stock exchange, the FTSE 100 for the London stock exchange and the Nikkei Dow for the Tokyo stock exchange. The Australian stock market is affected by changes in global markets hence it is reasonable to consider the performance of other markets when attempting to predict the Australian market.

- Advice and predictions from politicians and stock market experts. It is likely that other traders will be influenced by comments made by such people and will then trade accordingly. Often predictions made by significant persons can become self-fulfilling prophecies. For example if an expert publicly predicts that a stock's price will double then many people will scramble to purchase these shares. As a consequence of the mad scramble the share price indeed doubles. Considering such advice and predictions allows your own predictions to better account for the possible actions of competing traders.

The above list is by no means complete, however it does illustrate the unstructured nature of stock markets. Let us consider the desired output from such a decision support system. Essentially the aim is to predict future movements in a company's share price. *Fig 5.8* shows a typical graph of a company's share price fluctuations over time. A typical DSS uses historical known share prices as part of the data input to make predictions about the future fluctuation of the share price. As it is impossible to generate such predictions with absolute certainty the output generally recommends

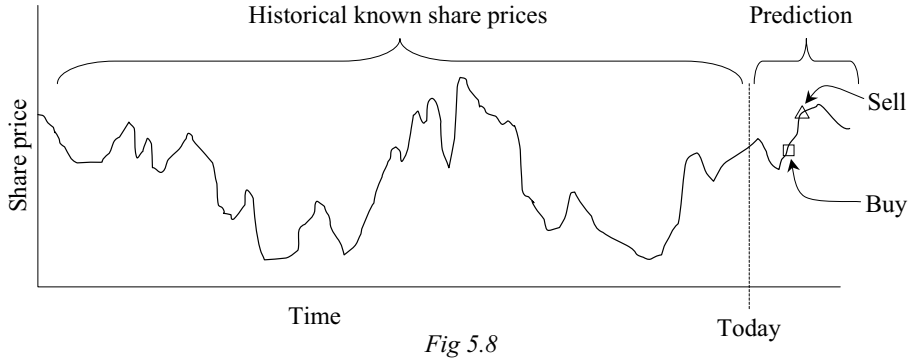


Fig 5.8
Typical graph of a company's share price fluctuations over time.

possible actions with different degrees of certainty. In *Fig 5.8* the system may be 60% certain that buying when the price reaches the level indicated by the small square (say $4.10) and then selling when the price increases to that indicated by the triangle (say $4.50) is the best strategy. However such predictions are usually accompanied by further instructions. In our example the DSS may recommend that if the price falls, rather than rising as predicted, then the shares should be sold immediately the price reaches $4.00 to minimise the loss. There is one certainty with which most stock market experts agree; playing the stock market game over the short term is certainly a risky business!

GROUP TASK Research
There are numerous software applications and online sites that claim to be able to predict share prices. Research some of these systems and their claims. Comment on the nature of the DSS used and the likelihood of such systems being able to accurately predict share prices.

Disaster Relief Management

When disasters occur the overriding aim is to provide assistance as soon as possible. Unfortunately when war or natural disasters strike it is often difficult to immediately determine the precise effects or extent of the disaster. The first response aims to minimise the loss of life, however this requires at least some understanding of the severity of the disaster and its impact on those involved. Those managing disaster relief operations must balance the need to act promptly against the need to determine what assistance is required.

To further complicate matters assistance from the international community is often delayed – governments are often reluctant to admit their need for international assistance. Once international assistance is requested further issues emerge such as, who controls the operation, delays due to customs restrictions and certifying medical staff to operate in foreign countries. Also it is not uncommon for inappropriate aid deliveries to cause bottlenecks at major airports.

Many disasters do not strike suddenly, rather there is often significant prior information indicating the impending build up, for example HIV-AIDs, tsunamis and locust plagues. In each case early warning systems that are able to inform the affected population and also the wider world are critical. Often information is able to save more lives than physical resources applied after the event.

Every disaster is different and therefore requires a unique response. A list of issues confronting those managing disaster relief operations include:

- Determining the extent and nature of the disaster.
- Cooperation between aid organisations so relief resources are used effectively and are not duplicated unnecessarily.
- Determining and then managing the timing and delivery of relief supplies.
- Early warning systems combined with education, particularly for those in remote areas and in poorer countries.
- Obtaining approval to enter a foreign country to provide relief.
- Relaxing import laws to allow speedy entry of relief supplies.
- Speedy approval of urgent medication for use by medical aid staff.
- Certifying medical and other relief staff to work in foreign countries.
- Approval to allow military relief staff and appliances to enter foreign soil.
- Understanding and respecting foreign laws and customs.

Consider the following extract:

World Disasters Report 2005 - Introduction (Partial Extract)

Information: a life-saving resource

Good information is equally vital to ensure disaster relief is appropriate and well targeted. After the tsunami (December 2004), women's specific needs were often overlooked. Large quantities of inappropriate, used clothing clogged up warehouses and roadsides across South Asia. Assessing and communicating what is not needed can prove as vital as finding out what is needed – saving precious time, money and resources.

First, aid organizations must recognize that accurate, timely information is a form of disaster response in its own right. It may also be the only form of disaster preparedness that the most vulnerable can afford.

Markku Niskala
Secretary General
International Federation of Red Cross and Red Crescent Societies

GROUP TASK Discussion

Discuss how the unstructured and complex nature of disaster relief management makes "saving precious time, money and resources" difficult to achieve. Explain how "timely information" can assist.

So how can decision support systems assist disaster relief management efforts? Many general decision support systems are used to efficiently allocate personnel and other resources to particular disaster relief tasks. These systems store data describing the details of the disaster, the actions required to relieve the situation and the resources available to perform these actions. For example imagine contaminated drinking water is found at a location. Actions require temporary water to be urgently brought to the site to ensure the health of the local people. The system resources required to implement this action include water trucks and drivers, pumps, containers to distribute the water to individuals and a clean water source to refill the trucks. The DSS aims to efficiently assign particular resources to particular actions. Assigning resources to actions is a task common to most disaster relief efforts. Other decision support systems perform more specialised tasks such as determining efficient and safe search and rescue patterns or predicting the effect of particular actions.

Note that not all DSS used for diaster relief are totally unstructured. Examples of specific decision support systems used for disaster relief management include:

- SiroFire is a DSS developed by the CSIRO that simulates the spread of bushfires. The user can enter details of fire breaks and other fire controls and then simulate the resulting effect on the fire. *Fig 5.9* shows a SiroFire simulation where a fire commenced at a single point and has been burning for nearly three hours. The system uses data describing the terrain, fuel type and current weather conditions.

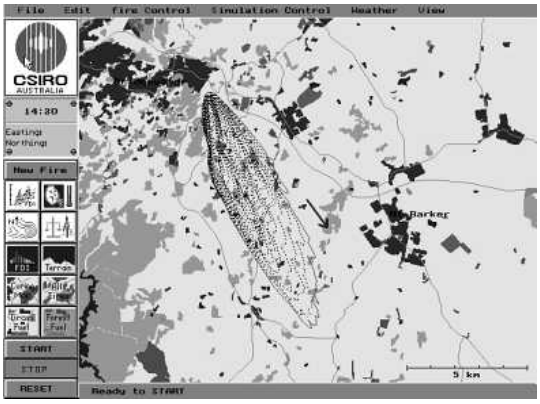


Fig 5.9
SiroFire software developed by the CSIRO for predicting the growth of bushfires.

- Co-OPR (Collaborative Operations for Personnel Recovery) is a Group Decision Support System that allows multiple personnel to collaborate and contribute to decision making processes. *Fig 5.10* shows the central command for Co-OPR. The system assists decision-making processes during the recovery of injured personnel from remote locations. Co-OPR includes teleconferencing together with instant messaging capabilities. This DSS assigns tasks to personnel in the field in real time.

Fig 5.10
Co-OPR is an example of a group decision support system for recovering personnel.

- Cúram software produces a product known as SEM (Social Enterprise Management). SEM integrates the provision of services from many different aid organisations, such as those providing health, social security, housing and security, via a single collection point. This means those in need can be assessed for a variety of different benefits based on the data collected during a single interview. Cúram's Intelligent Evidence Gathering™ interface collects data using an intelligent question, response system. If eligibility for a particular service is detected then the system intelligently asks relevant questions.

GROUP TASK Discussion
Consider each of the above examples of decision support systems in terms of the semi-structured to unstructured continuum. Discuss whether each is best described as a semi-structured or unstructured situation.

GROUP TASK Research
Research, using the Internet or otherwise, examples of decision support systems used to assist in the management of disaster relief. Briefly describe the nature of the decision making assistance each system provides.

HSC style question:

A sales analysis package is under development for use within the hotel industry. This package uses historical data including details of each past guest stay in the hotel. External data particular to each hotel's location is also imported or entered into the database. For example major sporting and entertainment events, weather forecasts and school holidays.

The package is to be used by the management of the hotel to allow them to better predict the number of guests likely to use the hotel on a week-to-week basis. Management can then adjust staffing levels more efficiently. The sales team will use the product to predict times of low occupancy. Advertising and other marketing strategies can then target these times.

(a) Identify the data used by this decision support system.

(b) Identify participants in this decision support system and for each provide an example of a decision where the system would be of assistance.

(c) Is this hotel decision support system best described as a structured, semi-structured or unstructured situation? Justify your answer.

(d) The results obtained from this system should, in theory, improve the profitability of hotels. However it is possible that results could be erroneous. Discuss the effects of negative results and who would be responsible for these negative results.

Suggested Solution

(a) Data collected from external sources includes details of major sporting and entertainment events within a reasonable distance of the hotel, weather forecasts for the area and school holiday periods. Data obtained from the hotel's existing information system includes various historical data with regard to past guest stays. This would likely include the dates of each stay, the number of guests per stay, their total spend and whether they are a repeat guest. It is likely that historical data with regard to staffing levels and details of past local events and past weather conditions would also be used.

(b) Participants would include management of the hotel and the sales team. Management use the system to predict guest numbers in order to make better informed decisions about required future staffing levels. The sales team uses the system to predict times of low occupancy. This helps the sales team to decide

which periods of time should be specifically targeted as part of their marketing and advertising strategies.

(c) This hotel DSS is best described as a semi-structured situation. The data used by the DSS is clearly defined. The system uses historical data to formulate the most likely future guest numbers. Presumably tests, using the historical data, have been undertaken to confirm the ability of the system to predict guest numbers. This means the input data is sufficient to determine reasonable predictions of future guest numbers. Therefore the system includes a clear method for transforming these inputs into reasonable guest number predictions.

(d) Negative results will lead to lower occupancy rates within the hotel which in turn results in reduced profits. The system is only as good as the data and rules it contains to transform the data into future guest number predictions.

The outputs are predictions rather than definite statements of fact. Hotels need to be aware that not all predictions will come true – this is not the fault of the system rather it is due to uncertainties that the system is unable to account for. For example the weather forecast may predict fine weather on the day of a large outdoor event. In reality it may rain on this day and the concert could be cancelled. It is not really anybodies fault for such incorrect predictions – assuming predictions and decisions are based on sound and tested data and rules.

Comments

- In an HSC or Trial examination this question would likely be worth a total of eleven marks – two marks for (a) and three marks each for parts (b), (c) and (d).
- In part (a) it is not necessary to identify external and internal data, however this is reasonable additional detail that is included in the scenario.
- In part (a) the suggested solution elaborates on the likely details of the historical data used by the system. It is reasonable to assume these details given the context of the scenario and the predictions it makes.
- In part (b) the participants are included within the scenario, hence simply naming management and the sales team would attract minimal marks. Most marks would be awarded for correctly describing examples of decisions made by each group that is assisted through the use of the DSS.
- In part (c) it would be possible to argue that the DSS is unstructured and obtain most of the marks. For instance one could argue that many additional variables that are largely unknown or that cannot be reliably determined affect the certainty of the predictions. Such variables could include local and global economic conditions, competitors and their marketing efforts and other variables that are simply unknown.
- In part (c) the system is certainly not structured as the output is a prediction and also the weather forecast details input are also predictions. There is no single definite correct answer.
- With regard to part (d) it is true that no decision support system will produce perfect results. If a definitely correct output was possible then a decision support system would not be required.

SET 5A

1. Decision support systems are used when:
   (A) the method of solution is clear.
   (B) conclusions are reached with complete certainty.
   (C) the decision includes uncertainty.
   (D) all variables affecting the decision are known.

2. Which of the following is the most structured situation?
   (A) Finding the range of a set of marks.
   (B) Deciding on a DVD player to purchase.
   (C) Forecasting the weather.
   (D) Selecting your favourite song.

3. Which of the following is the most unstructured situation?
   (A) Finding the range of a set of marks.
   (B) Deciding on a DVD player to purchase.
   (C) Forecasting the weather.
   (D) Selecting your favourite song.

4. When a bank approves a loan, which of the following is TRUE?
   (A) The bank knows the customer will meet their repayment obligations.
   (B) The bank is confident the customer will be able to meet the repayments.
   (C) The bank is unsure of the customer's ability to repay the loan.
   (D) The customer has agreed to the terms of the loan.

5. The goal of stock market prediction decision support systems is to:
   (A) accurately predict what and when to buy and sell shares.
   (B) submit sell orders and buy orders to stockbrokers.
   (C) trade shares from the current owner to buyers.
   (D) analyse market trends and chart historical fluctuations in share prices.

6. When assessing housing loans, what is a LVR used for?
   (A) To determine if the customer's income is sufficient to meet the repayments.
   (B) To predict if the customer's income will continue at current levels.
   (C) To assess the ability of the bank to recover funds if the customer fails to meet their repayment obligations.
   (D) To ensure the bank can recover all its funds if the customer fails to meet their repayment obligations.

7. Authenticating users based on their fingerprints commonly uses which of the following techniques?
   (A) Comparing minutiae.
   (B) Ridge feature matching.
   (C) Comparing bitmaps directly.
   (D) A combination of all of the above.

8. Predicting share prices is best described as a:
   (A) structured decision situation
   (B) semi-structured decision situation.
   (C) unstructured decision situation.
   (D) game of chance.

9. The minutiae commonly used by fingerprint matching systems are:
   (A) ridge shape and orientation.
   (B) ridge endings and bifurcations.
   (C) number of ridges and location.
   (D) All of the above.

10. Inputs into disaster relief decision support systems include:
    (A) delivering relief supplies and determining the extent of the disaster.
    (B) Relaxing import laws and identifying relief personnel.
    (C) Cooperation between relief agencies and certifying medical staff.
    (D) Determining the extent of the disaster and identifying available resources.

11. Define each of the following terms with regard to decision support systems:
    (a) Decision (b) Alternatives (c) Uncertainty

12. Outline the significant features of structured, semi-structured and unstructured situations.

13. Explain reasons for each of the following using examples from the text:
    (a) Why is approving a bank loan considered to be a semi-structured situation?
    (b) Why is predicting stock prices considered to be an unstructured situation?

14. Explain how fingerprints are collected and then processed to authenticate users.

15. Research a specific and significant disaster. List at least 3 decisions that needed to be made as part of the disaster relief management effort. Describe possible or actual decision support tools that could or were used to assist making each of the decisions in your list.

TOOLS THAT SUPPORT DECISION MAKING

In this section we describe a variety of different tools that assist or support decision making, namely:

- Spreadsheets,
- Expert systems,
- Artificial Neural Networks (ANN),
- Databases (including DBMSs and operational databases),
- Data warehouses and data marts,
- Data mining,
- Online Analytical Processing (OLAP),
- Online Transaction Processing (OLTP) systems,
- Group Decision Support Systems (GDSS),
- Intelligent agents,
- Geographic Information Systems (GIS) and
- Management information systems (MIS).

Note that not all these tools are specific Decision Support System (DSS) tools – refer *Fig 5.11*. Some are data sources for DSSs; such as operational databases, data warehouses and data marts – OLTP creates the operational databases that in turn are the data source used to create data warehouses and data marts. Others are tools that assist when making structured decisions; in particular MISs. Spreadsheets are also commonly used in structured decision situations, calculating averages and other statistics, for example. OLTP systems generate reports from operational data for use by structured decision makers. For instance, weekly product sales summaries are used when deciding how much stock should be ordered from suppliers.

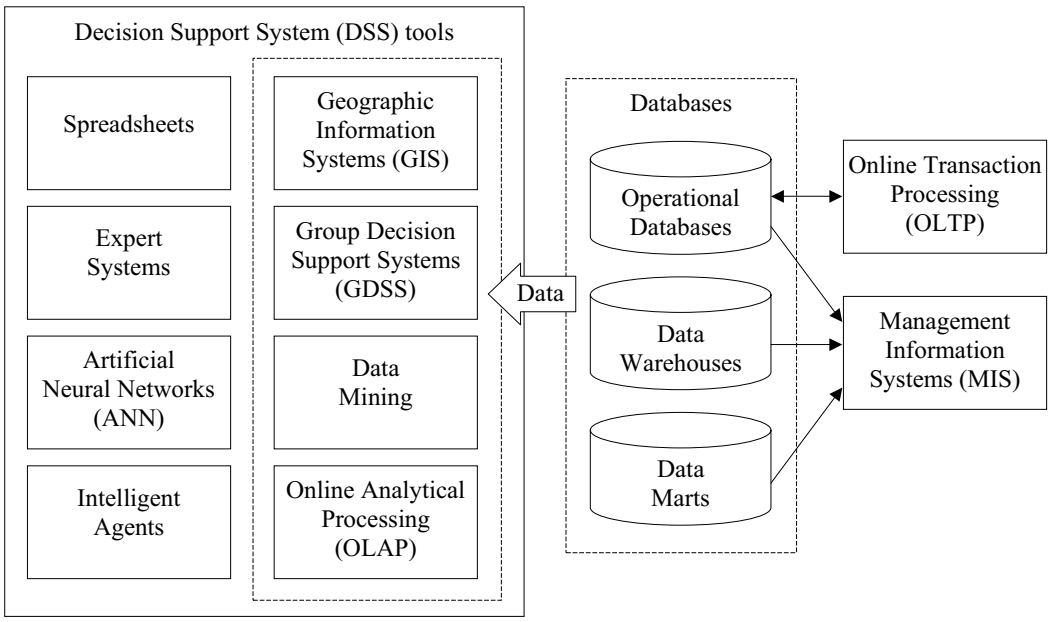


Fig 5.11
Tools that support decision making.

Recall that Decision Support Systems are required when the decision situation is semi-structured to unstructured. In these situations the variables and their influence on the decision are unclear or there is no clear method of solution. *Fig 5.11* classifies tools for these decision situations as Decision Support System tools. It is these Decision Support System tools (refer *Fig 5.11*) that are the major focus of this option topic – in particular spreadsheets, expert systems and artificial neural networks. Often a combination of DSS tools is used within a single DSS. For instance, data mining can

use artificial neural networks and intelligent agents often operate in the background when performing OLAP.

In later sections we explain the detail of spreadsheets, expert systems and artificial neural networks. Hence in this section we restrict our discussion to a brief outline of their general characteristics.

SPREADSHEETS

Spreadsheet applications organise data into one or more worksheets. Each worksheet is a 2-dimensional arrangement of columns and rows. The intersection of a column and row is called a cell. Each cell holds text, numeric or formula data independent of other cells. Formulas refer to other cells using their cell address.

Presumably you have already covered the Information Systems and Databases core topic, so your understanding of databases should be clear, however it is worth briefly considering the essential difference between spreadsheets and databases. Unlike rows within a spreadsheet the records within a database table are all composed of the same fields. All records in a table contain the same set of fields and each field has a single data type. Databases process records as complete units whilst spreadsheets process cells as complete units. In a database records have no predetermined order whilst in a spreadsheet each cell has a specific location and order in relation to other cells – cell A1 is always above cell A2 and cell B2 is always to the right of cell A2.

In terms of decision support systems, spreadsheets are particularly valuable tools for performing "what-if" analysis – altering inputs and viewing the effect on the outputs. The opposite process, known as "goal seeking" allows a desired output (the goal) to be entered, the spreadsheet then calculates the inputs required to achieve this output. Most spreadsheets include an extensive set of statistical functions that allow complex statistical analysis of data. Modern spreadsheet applications include powerful charting features for displaying results in a more human friendly form. In addition processes within current spreadsheets can be automated using macros. A macro is essentially a symbol or shortcut that causes a sequence of processes or a program code routine to execute.

GROUP TASK Discussion

Presumably you have used spreadsheets previously in a variety of different situations. Consider these past situations. Was a decision involved? If so, what role did the spreadsheet play in the decision making process?

EXPERT SYSTEMS

An expert system is a software application that simulates the knowledge and experience of a human expert. The knowledge of the expert is coded by a knowledge engineer into a series of rules that are stored within a knowledge base. The expert describes how he or she would act or respond to different conditions and the knowledge engineer translates these responses into rules. When the completed expert system is executed it asks questions in a logical order much like a human expert. Deciding on the order and questions to ask is based on user responses and is determined by the inference engine. Questions and answers continue until the expert system determines one or more conclusions or is unable to reach a conclusion.

Commonly expert systems are used when the knowledge of a human expert needs to be reproduced for many users. For example troubleshooting computer hardware problems, diagnosing medical conditions or even playing chess. In general an expert system is a suitable choice when a human expert can solve the problem or make the decision during a consultation over the telephone.

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANN) are an attempt to simulate the complex structure and processes performed by the human brain. The human brain is a neural network composed of billions of neurons connected to each other in complex ways via synapses. As we learn the synapses grow or contract to alter the electrical signal passing between neurons. Each neuron receives inputs from other neurons, if it likes what is hears then it fires its output on to other neurons.

Artificial neural networks use far fewer neurons than the human brain – commonly fewer than a hundred. Like the human brain artificial neural networks are able to learn. Furthermore they do not need to be instructed on how to precisely solve problems. Rather artificial neural networks are trained using sets of sample inputs together with known outputs. Once trained the network is able to determine the most likely outputs based on new unseen input data.

Artificial neural networks are well suited to unstructured situations. They are particularly useful when the relative importance or certainty of the inputs is unknown or there is no prescribed method for solving a problem.

GROUP TASK Discussion
Distinguish between expert systems and artificial neural networks based on the above brief outlines.

DATABASES

Database Management Systems (DBMSs) include the ability to extract and analyse data within databases using SQL statements. Many decision support tools and systems use the services of DBMSs to obtain data for further analysis. Some import data from operational databases, whilst others link to databases directly via the DBMS. For example spreadsheet based DSSs often query databases and then import the results for further analysis. During analysis the spreadsheet summarises the imported data; perhaps creating charts to analyse business trends, for example. When developing neural networks training and testing data is often sourced from databases. Some expert systems connect to databases that act as an extension of the system's database of facts. For instance, an expert system designed to recommend products will likely attach to a database containing details (facts) about each available product. Data from operational database systems, such as online transaction processing (OLTP) systems, is extracted to create data warehouses and data marts.

Consider the use of databases in the following situations:

Approving Bank Loans

Earlier we identified three basic requirements used by banks when assessing loans:

1. The customer's income is sufficient to meet the regular loan repayments.
2. The customer's income will continue at current levels for the term of the loan.
3. The bank will be able to recover their funds if the customer is unable to meet their repayment obligations.

GROUP TASK Discussion
Discuss how databases could be used to confirm that each of the above requirements is a reasonable indicator of a customer's ability to repay a loan.

Fingerprint Matching

Earlier we identified three techniques used for matching fingerprints, namely:

1. Identifying minutiae and comparing their relative positions.
2. Ridge feature matching.
3. Comparing the images of the fingerprints directly.

GROUP TASK Discussion
Initially all fingerprints are scanned as images. If minutiae or ridge feature matching is used should these features be determined in advance or determined during fingerprint matching? Discuss in terms of databases.

DATA WAREHOUSES AND DATA MARTS

Data warehouses are large separate databases that include data imported from operational databases across an enterprise. These large databases commonly contain many years of historical data. Many DSS tools, such as data mining tools, analyse these vast data warehouses repeatedly. We discussed data warehouses, including how they are created back in chapter 2 (page 222).

Data Warehouse
A large separate combined copy of different databases used by an organisation. It includes historical data, which is used to analyse the activities of the organisation.

Some DSSs are developed using evidence from large data warehouses or data marts. Once the DSS is completed it connects to the operational databases. For example a neural network for assessing customer needs is trained and tested using data extracted from the organisation's data warehouse. Once implemented the neural network assesses customer needs as they are entered into the organisation's online transaction processing (OLTP) system.

To improve the performance of data mining and OLAP (Online Analytical Processing) systems relevant data is often extracted into a data mart; either from the enterprise's data warehouse or directly from their operational databases. Preparing data for data mining and OLAP changes its organisation and perhaps even some of its content. We don't want to change the original data source so creating a data mart is often the preferred solution – the general nature and organisation of data marts is described in chapter 4 (page 438) as part of the transaction processing systems option. For data mining and OLAP applications creating a dedicated data mart, running on its own database server is generally a worthwhile investment. Two common strategies for creating data marts in preparation for OLAP or data mining are shown in *Fig 5.12*, the first extracts data from an existing data warehouse and the second extracts data directly from the organisation's operational databases.

Data Mart
Reorganised summary of specific data extracted from a larger database. Data marts are designed to meet the needs of an individual system or department in an organisation.

Data mining involves the creation and testing of numerous models. Such model creation and testing requires fast data access and fast processors – many data mining tools are designed to take advantage of parallel processing. The ability to develop better models is greatly enhanced if each model can be created and tested in minutes rather than hours or days. Furthermore using a separate database system, such as a data mart provides, means the intensive processing performed by data mining will not affect the efficiency of the organisation's other information systems.

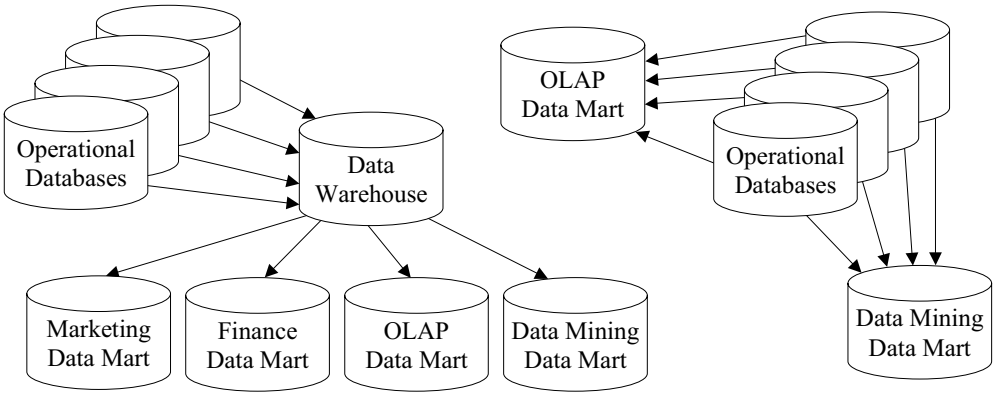


Fig 5.12
Two common strategies for creating a data mart.

OLAP systems allow users to analyse large amounts of data quickly and online. Creating a dedicated data mart means that no other systems are sharing data access and furthermore the organisation of the data can be altered to suit the particular analysis processes supported by the OLAP system.

GROUP TASK Research
Data warehouses and data marts are not just used by data mining and OLAP systems. Research other systems that use these large data stores.

DATA MINING

Data Mining
The process of discovering non-obvious patterns within large collections of data.

Data mining aims to discover new knowledge through the exploration of large collections of data – data mining is also known as knowledge discovery. It is a process that uses a variety of data analysis tools to discover non-obvious patterns and relationships that may prove useful when making predictions. These patterns and relationships are models that describe characteristics or trends within the data.

Different data mining tools create different types of models and will likely discover different patterns and relationships. Some common tools include artificial neural networks, decision trees, rule induction, linear and non-linear regression, genetic algorithms and K-nearest neighbour reasoning. There are many others and most commercially available data mining systems include a variety of different tools.

Data mining is not an automatic process that trolls through data warehouses (or data marts) and miraculously makes predictions and recommendations. Rather data mining requires guidance and a thorough understanding of the data. This is by far the most time consuming task, often consuming around 90% of the total data mining costs and time. The data to be mined will first need to be reorganised, cleansed and summarised to suit the particular data mining tools being used. Cleansing removes redundant data and also corrects other data integrity and data quality issues such as missing or incorrect data items. Unusual atypical data items, known as outliers, should be analysed – perhaps they are incorrect or maybe they represent some one off occurrence. Maybe they should be edited or even removed. When using some data mining tools outliers can have an unwarranted influence on the results.

Let us consider a sample of data mining tools from the wide range of data mining tools available. We will briefly describe decision tree algorithms, rule induction, linear and non-linear regression and K nearest neighbour tools. The detailed operation

of neural networks including genetic algorithms is discussed later in the chapter – both these tools are also used for data mining.

Decision Tree Algorithms

The main goal of decision tree algorithms is to find conditions that clearly distinguish between groups of data that all possess similar attributes. The best conditions are those that maximise the differences between the data in each group. Decision tree algorithms look for common characteristics upon which the data can be split as they work to determine the best conditions. Once a best condition is found the data is split into distinct groups. Each of these groups is then examined to determine further splits and hence create sub-groups. The process continues categorising the data into smaller and smaller groups. The result is essentially a decision tree; a model that categorises the data into progressively smaller groups where each group possesses particular characteristics in common.

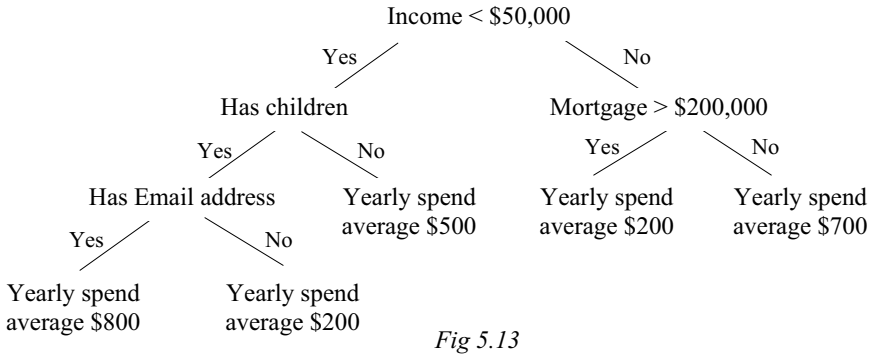


Fig 5.13
Sample decision tree model resulting from data mining.

Consider the sample decision tree in *Fig 5.13*. In this example the database being mined includes details of all the organisation's past and current customers – including some personal details and details of their past purchases. The design of the tree is the result of data mining – the decision tree algorithm determined each of the conditions. During data mining the algorithm first determined that the best way to split the data was based on incomes above and below $50,000. It determined this by analysing all attributes of each customer. In a real world situation there could be millions of records (one for each customer in this example) and each record may contain hundreds of attributes. Eventually after detailed analysis the decision tree algorithm concluded that Income < $50,000 was the best condition to split the customers into different groups. The split was made and then the process was repeated with each group to generate further conditions. Notice that final tree does not recommend a particular action, rather it simply splits the data into groups. Management of the organisation could use this knowledge in various ways. Perhaps marketing efforts could target new customers who have an email address, have children and have incomes below $50,000. Perhaps they could devise more effective strategies to encourage customers with high incomes and high mortgages to increase their spend. Or perhaps the knowledge can be used as part of further data mining processes.

Rule Induction

Rule induction determines sets of rules that do not form a single decision tree. Think of a rule as an IF THEN selection. These rules are the results of rule induction and they do not necessarily split all the data into distinct groups. For instance the rule "If Customers purchase a hammer then they are likely to also purchase nails" says nothing about the group of people who do not purchase hammers, perhaps some of them are also likely to purchase nails. The resulting model categorises data into groups, however each group will likely intersect with other groups (see *Fig 5.14*).

Unlike decision trees, rule induction determines each rule independently of other rules. Often rule induction will produce rules that would not be found using a decision tree algorithm. Decision trees are forced to only consider data in each sub-group – rule induction is free to consider as much data as is needed to induce and verify a rule. Using rule induction the entire data set or any subset of the entire set can be used to determine many different rules. Sometimes two conflicting rules will be produced, for instance "Women under 30 prefer sports cars" conflicts with "People over 25 prefer luxury cars." When this occurs then further analysis and investigation is needed to verify the validity of the rule.

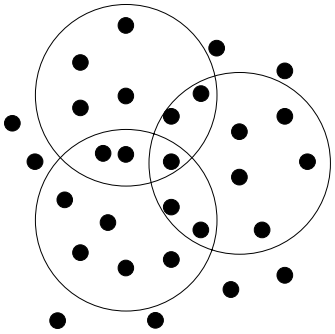


Fig 5.14
Rule induction groups data into independent groups.

Non-Linear Regression

Regression is the process of fitting a model to data. No doubt in science you've drawn a "line of best fit" or "trendline" through a graph of sample points, this is an example of regression analysis. The line is a simple model that allows other values to be predicted. If the line is a straight line then linear regression is being used. In terms of data mining few data sets can be accurately modelled using straight lines hence various non-linear regression techniques have been developed. For example, most spreadsheets include the facility to automatically fit various standard families of curves to data – log, exponential and polynomials are common examples. Non-linear regression tools used for data mining perform similar, albeit considerably more complex processes. *Fig 5.15* is a sample regression curve in just two dimensions, regression tools can generate models over three, four or many more dimensions – we just can't draw the model on paper as a simple curve. Regression tools are often used to model changes that take place over time. Data mining produces the model, which can then be used to predict future values.

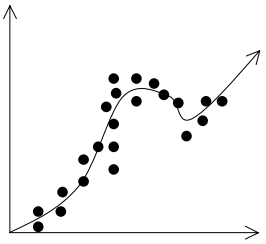


Fig 5.15
Sample two-dimensional non-linear regression curve.

In reality artificial neural networks (ANNs) are a complex form of non-linear regression. They create a model based on sample data that can then be used to predict outputs for unknown inputs. However the models produced by ANNs are difficult to interpret – explaining why a particular ANN works is extremely difficult. Other non-linear regression tools are able to supply some reasoning for why they work.

K-Nearest Neighbour (K-NN)

K-nearest neighbour is a classification technique that compares each data item to previously classified items. It searches for K existing items that are most similar to the current item. In other words the algorithm identifies K items that are the nearest neighbours to the new item. In *Fig 5.16* the circle encloses the 10 nearest neighbours to the new data item N. It then determines how these K items have been classified and counts the number of items in each class. The new item is placed in the class with the highest count. In *Fig 5.16* the new data item N is classified as belonging to class A as there are more data items within the circle in class A than in class B or class C. This is extremely simple example, in reality most K-NN systems consider the distance of each existing data item from the new data item – those closer have more influence on how the new item is classified.

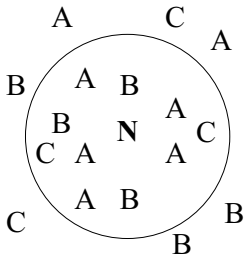


Fig 5.16
K-Nearest Neighbour where N is placed in class A.

The significant difficulty with K-NN systems is determining how the closeness or distance between data items can be sensibly determined. Each attribute needs to be considered. Determining distances between numeric values is simple but how do you determine the distance between text attributes? For example what is the distance between pets? How do you measure the distance between a dog and a cat or between a cat and a parrot? A consistent scheme needs to be devised that will result in meaningful distance measures for the particular situation. Perhaps the expected life span could be used or the average yearly food cost. When data mining a veterinary suppliers database possibly the average yearly vet bill could be used to determine such distances.

GROUP TASK Discussion
Many data mining tools classify data into new non-obvious groups that all possess similar characteristics. How can this new classification lead to new knowledge about the data? Discuss.

GROUP TASK Discussion
The models produced by data mining do not always hold true in the real world. Discuss how the validity of new rules and classifications can be tested in the real world.

ONLINE ANALYTICAL PROCESSING (OLAP)

Online Analytical Processing (OLAP) systems allow decision makers to analyse large data stores visually, online, as needed and as quickly as possible. For this to occur requires fast processors and fast data access and response times. To meet these requirements large enterprise OLAP systems include their own dedicated OLAP servers linked to databases that are organised specifically to optimise the efficiency of the system's analysis processes. Users interact with the system using OLAP client applications installed on their personal computers. Small commercial OLAP software is also available, these applications analyse much smaller quantities of data for small and medium sized organisations – often using a standard desktop computer to analyse data in the organisation's operational database.

In chapter 2 (page 224) we described the organisation of OLAP data (OLAP cubes) and the general nature of functions performed by OLAP systems. In this section we focus on two essential features of OLAP systems, namely data visualisation and drill downs.

GROUP TASK Review and Discuss
Reread the description of OLAP on page 224 of chapter 2. Explain how OLAP data is organised and describe the aim of OLAP systems.

Data Visualisation

Data Visualisation
Displaying data, summary information and relationships graphically, including charts, graphs, animation and 3D displays.

Displaying information in a visual and interactive format is a feature of OLAP systems. Most OLAP systems are able to interactively generate a variety of graphs and charts in real time based on user input – often in the form of simple mouse clicks. Some systems are able to generate animations and three dimensional graphics. Far more information can be represented within a graphical display than is possible within tables and text. Furthermore, relationships between data and other significant information is much easier for people

to grasp when presented graphically. Examine the highly complex Sales Dashboard in *Fig 5.17* – this screen was the winner of DM Review's 2005 data visualisation contest. The screen contains an enormous amount of information, however even a brief glance uncovers numerous relationships and trends – revenue and profit are rising, whilst market share declines and order size slowly increases, for example. Now imagine even attempting to uncover such relationships and trends if all this data and all these statistics were presented as a series of tables – definitely a very difficult, laborious and inefficient task. Data visualisation is what makes OLAP intuitive and usable for decision makers. They can concentrate on the information they need to make informed decisions, rather than being swamped by masses of data and statistics.

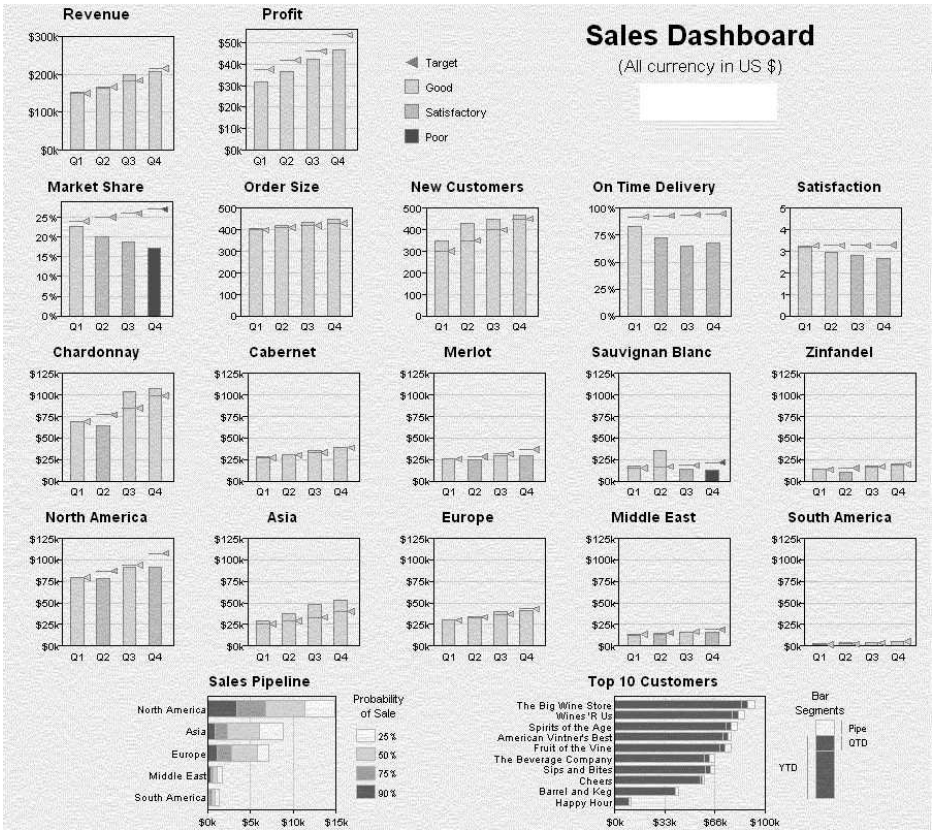


Fig 5.17
Sales Dashboard developed by Robert Allison of SAS Institute.
(Winner of DM Review's 2004 Data Visualisation contest).

OLAP automates data visualisation. Other systems require statistics to be calculated and graphs created individually. Using OLAP, these largely analytical processes are automated. The user selects the data or characteristics of the data that they wish to explore and OLAP takes over to perform the hard work of calculating the statistics and generating the graphical models.

Drill Downs

Drill down refers to the ability to progressively focus in on more and more detailed information. This is much like exploring the files on a hard disk; you start at the root directory, open a sub-directory, then open a further directory with this sub-directory, this process continues until you locate the required file. In OLAP systems

Drill Downs
Progressively moving from summary information to more detailed information. Each move focuses and expands particular information.

drill downs are performed on data and characteristics of data. For example, an enterprise may have operations in say Australia, New Zealand and China. Say the first graph displays profit for each of these countries. Drilling down on Australia causes a graph of profits for each Australian branch to be displayed. If the user then drills down on Sydney they uncover the profits made by each department within the Sydney branch.

OLAP takes drill down one step further – at any stage the displayed data can be changed. For instance, instead of profit for individual Sydney departments the user might explore Sydney's payroll costs, and then the number of Sydney employees whose salaries are above $100,000. They then examine salesmen within this category and drill down to uncover an individual's monthly sales figures. They can then compare these monthly figures to salesmen throughout the entire organisation, and then filter the results to include only salesmen on similar incomes. This free form exploration of information is known as "slicing and dicing" – in terms of OLAP cubes each slice or dice conceptually splits the cube along one or more dimensions.

Consider the following sequence of OLAP drill down screens:

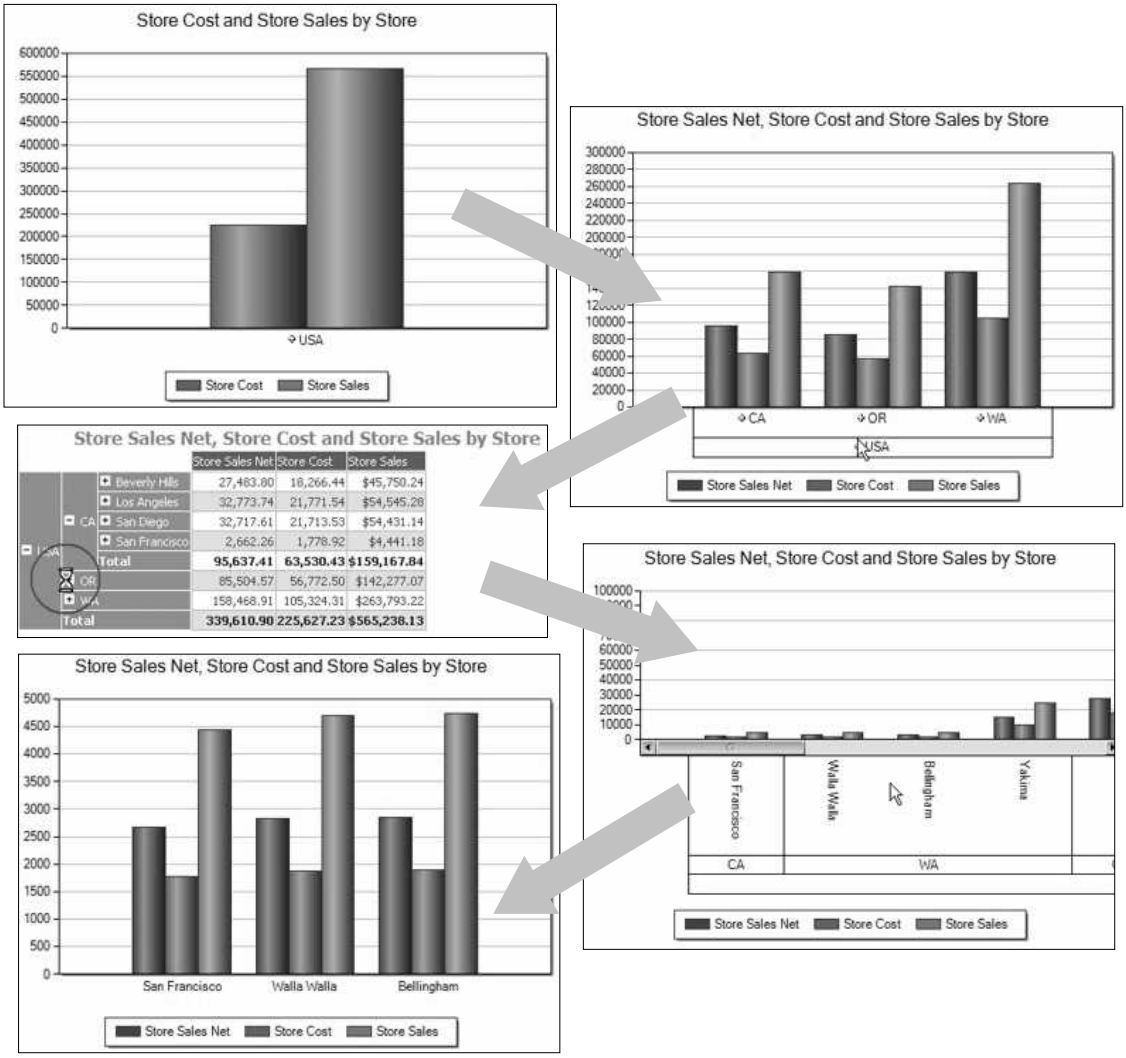


Fig 5.18
Data visualisation and drill down example using Dundas OLAP Services for .NET.

GROUP TASK Activity
Analyse the sequence of screens in *Fig 5.18* to determine the information uncovered at each step in the drill down

GROUP TASK Practical Activity
Using the Internet, download or view an online demonstration of an OLAP application in operation.

ONLINE TRANSACTION PROCESSING (OLTP)

Much of the transaction processing systems option topic (chapter 4) is about OLTP systems – in particular the work on real time and online transaction processing. In general, OLTP systems create and manage the operational databases present in most large organisations. For instance, you are interacting with an OLTP system when you make an online purchase, transfer funds between bank accounts, withdraw money at an ATM or make a purchase using EFTPOS. In fact, if the same bank account is used then the same large OLTP system is involved in all these transactions. OLTP systems also interact with each other so that transactions that span many systems are completed in their entirety or not at all. For example, when making an online purchase your funds must leave your account at your bank and then be deposited into the sellers account held at their bank. The OLTP systems of both banks must complete their actions for the total transaction to be a success.

So how do OLTP systems relate to Decision Support Systems? Their main role is to provide the operational data that is then analysed by DSSs. However, most OLTP systems also perform rudimentary data analysis tasks that assist decision makers – but they are not decision support systems. For instance Internet auction sites, such as eBay, use OLTP systems to process bids and other types of transactions. These sites calculate and display various statistics and graphs that help users assess the reliability and honesty of other buyers and sellers.

GROUP TASK Reading
Read the first page of the Transaction Processing Systems option topic (page 365). Define the term "transaction".

GROUP DECISION SUPPORT SYSTEMS (GDSS)

Group decision support systems (GDSS) are information systems that facilitate decision-making activities between multiple participants. They provide computer-based tools to assist participants to contribute to the decision making process. Commonly GDSS is used during business meetings to improve the ability of the meeting participants to reach consensus and make informed decisions. The GDSS can operate over the Internet, a LAN and/or within a meeting room. A GDSS uses many of the tools present in most teleconferences or video conferences together with tools specifically designed to assist the decision making process – teleconferencing and video conferencing was discussed in chapter 3. GDSSs can be used in small meetings with just a few participants, however they are particularly useful for large meetings with tens or even hundreds of participants. The technology aims to allow everyone to contribute whilst maintaining a meeting structure that promotes efficient decision-making.

Typical GDSS features that specifically assist decision-making include:

- Voting and ranking systems that automatically collect and tally votes from all participants.

- Ability for all participants to contribute - often anonymous contribution is possible.
- Comments shared with all other participants, commonly using an electronic whiteboard feature.
- Flexibility to incorporate external information as required.

GROUP TASK Research
Group Decision Support Systems are a relatively new technology. Research examples of GDSS and outline the features each example includes.

INTELLIGENT AGENTS

Intelligent agents operate in the background to complete tasks that assist people. They act intelligently and on behalf of the person, for example a travel agent does all the legwork needed to assist people plan and book vacations. The travel agent makes intelligent decisions to best meet your needs. For instance, they may know you have young toddlers so they will tend to suggest hotels that cater to young families.

In terms of information systems, there are many different types of software agents but not all are intelligent agents. The defining feature of all software agents is their ability to act without human intervention. That is, they begin processing data based on changes they perceive or recognise. There are numerous examples of such software agents, for example email clients are usually set to POP a user's email account at regular intervals, say every five minutes and the spell checker in a word processor automatically underlines misspelt words. Both these agents are operating on their own, however they are not displaying human-like intelligence. The email client agent simply recognises that five minutes has passed and then blindly performs a predefined action. Each time a word is entered the spell checker checks its dictionary. Intelligent agents are also known as daemons or bots. Daemon was originally a UNIX term referring to processes that run unattended in the background. Intelligent agents are a particular type of agent (or daemon) that responds in an intelligent and human-like manner.

In general, intelligent agents possess the following characteristics:

- Autonomous – Intelligent agents operate independently without constant guidance from users. They make decisions to determine how to solve problems and solve them on their own.
- Proactive – Intelligent agents do not wait to be told, rather they act and often make suggestions to the user.
- Responsive – Intelligent agents recognise changes in their environment that indicate changes in user needs and they alter their behaviour accordingly.
- Adaptive – Intelligent agents can change their behaviour or learn new behaviour over time to account for changing user preferences.

Often many intelligent agents communicate with each other to make decisions and solve problems.

Consider the following examples of intelligent agents:

Some areas where intelligent agents have been used to filter Internet content include:

- Intelligently monitoring website changes and reporting back to users when relevant changes occur.
- Enhancing the results returned by search engines based on user preferences and past behaviour.

- Compiling a personalised daily newspaper from multiple sources based on the user's preferences and past interests.
- Filtering incoming email messages and detecting and informing the user based on experience when critical messages are received.
- Finding the best prices and products from online auctions and retailers.
- Filtering web content to remove adult material, popup adds and other unwanted material.

Some other examples where intelligent agents are used include:

- Checking documents for plagiarism.
- Air traffic control systems to monitor individual aircraft and detect those off course or in danger of collision.
- Medical monitoring systems used in intensive care wards. Used to monitor patient vital signs and respond accordingly.
- Monitoring activity on network servers.
- Human like simulators that present information using animated characters.
- Personal assistants that telephone or email details of appointments and other information they find or determine.

GROUP TASK Research
Research, using the Internet or otherwise, specific examples of intelligent agents. Briefly describe features that indicate the agent is able to act intelligently.

GEOGRAPHIC INFORMATION SYSTEMS (GIS)

Geographic Information Systems represent data using maps. Probably the most well known example is Google Earth, which includes satellite photographs of the entire planet. GISs plot features, landmarks and other information on top of maps in the form of layers, for example, one layer may use different colours to indicate population density, whilst another layer shows the major communication links, and another overlays the location of a company's customers. Most GISs include zoom and pan features that allow users to focus in on particular areas of interest. For instance, if a business sells 3G phones then they can zoom in on areas with 3G coverage and then examine areas where they have few customers but there is a high population density. Commonly textual tags are displayed as the mouse hovers over a particular feature. The tag may display the underlying data or perhaps statistics or even a graph related to the current map location.

Many GISs can also operate in conjunction with GPS receivers so real time location data can be displayed within the GIS. For instance, a courier company can track the location of their drivers. They use this information to more efficiently allocate jobs such that travel times and distances are minimised.

GROUP TASK Activity
Read the article on the use of a GIS during the Sydney 2000 Olympics. Whilst reading, note any inputs to the GIS, functions performed by the GIS and information output from the GIS.

GROUP TASK Research
GISs are used within many industries, including transport, environmental services, wildlife monitoring, real estate, surveying, mining, etc. Research and outline at least three examples of such GIS systems.

GIS Strikes Gold at Summer Olympics

Hosting the XXVII Olympiad was no small task, even for Australia, a country where everything seems to be just a little larger than the norm. For instance, the entire sailing event for Olympics 2000 was staged for the first time in a harbour, allowing spectators to view races that have, up until this point, always been staged in the open ocean.

Sydney Harbour officials took it all in stride. The harbour is big, and a sailing event, even if it is the largest in the world, would not dramatically affect its busy shipping lanes and cross-harbour traffic.

However, to allow itself this measure of complacency, the Sydney Organizing Committee for the Olympic Games (SOCOG) joined forces with the government and private enterprise immediately after being informed of its successful bid to stage this year's Olympics. Together, they mapped out a comprehensive strategy to stage a truly millennium-scale sporting event.

"Our responsibility for the Olympics was to provide SOCOG with uninterrupted areas to run their sailing events," comments Rob Colless, graphical systems manager at Waterways. "We had a major role in helping to plan for this part of the Olympics, and that's where GIS technology comes in."

To accomplish this required integrating a variety of sources of information, for which Waterways used MapObjects, ESRI's developers' software that includes embeddable mapping and GIS components, allowing the creation of dynamic live maps with GIS capabilities. Maps were viewed with ArcView GIS, which also hot linked related photos, videos, and text to the map display.

Waterways made extensive use of its Intranet, which was powered by ArcView Internet Map Server and used to automatically distribute vital information concerning harbour activities to those monitoring the various Olympic events. This allowed them instant access to ongoing races so that, in the event of any disruptions to a race, an immediate response could be mounted.

"We built a three-dimensional model of the whole of Sydney Harbour," Colless continues, "which is based on hydrographic soundings from more than 100 years worth of soundings records."

Because the model includes position and depth information, Waterways was able to use it to set up a series of exclusion zones around the harbour. The coordinates were then supplied to the crews laying buoys around the exclusion zones, who used GPS to locate each position. Because depth information was included, they could easily cut the right length of rope, attach it to the buoy, and drop it into position.

"The model really saved us a lot of time," Colless continues. "Previously you would have to go out to the location, check the depth, cut the rope, and then lay each buoy in position. Also, if any of the buoys got pulled out of position, the model allowed us to easily get them back into position because we had captured the coordinates."

Because Sydney Harbour continued its commercial shipping and other operations during those periods in which Olympic events were not scheduled, the Sydney Harbour Operations Center (SHOC) was set up. Waterways Authority, water police, and the National Parks and Wildlife Service, as well as representatives from other harbour-affiliated organizations, staffed SHOC to manage all activity in the harbour.

"We have radio communication and GPS tracking devices on most of our major vessels, and information was relayed back to SHOC headquarters via a mobile telephone network," explains Colless. "That information was read directly into the GIS and our Incident Management software so that we could see instantly where our vessels were and where they should have been to properly monitor, manage, and respond, if necessary. We recorded anything that could possibly have an effect on the racing events such as a whale entering the harbour, a capsized boat, or spectators breaching the exclusion zones. This is where the GIS mapping was very important, because with real-time GPS we could pinpoint where an incident occurred and then instantly create a map of it to assist in taking remedial actions, as well as include the map in our incident record."

Fig 5.19
Modified extract of an article in ArcNews on ESRI.com.
(ESRI produce and market ArdGIS and related GIS software).

MANAGEMENT INFORMATION SYSTEMS

The general nature of Management Information Systems is described on page 436 of the transaction processing systems option. Management Information Systems summarise data within an organisation's systems into information to assist in the management of the organisation's day-to-day operations. MIS functions are "programmed" into the MIS in advance, they meet predetermined requirements and follow structured processes to solve well understood structured problems. For example, an MIS function used by the sales department would likely produce monthly reports of total sales across different regions. A standard report is used, which extracts data using an SQL select statement. These processes are repeated each time the sales report is generated; only the data changes. MISs solve structured problems, they provide decision makers with information but they are not classified as Decision Support Systems.

Consider the following:

Each of the following is an example of information generated by an MIS. In each case the data source is transaction data.

- A list of each product a factory produces together with the profit or loss made on each over a 12 month period.
- A table listing each salesperson together with the total monthly value of their sales over the past 12 months.
- The total value of cheques for each bank that pass through a large cheque clearance facility on a particular day.
- A line graph for each product showing average total number sold each month over a five year period.

GROUP TASK Discussion

For each of the above examples, identify the transaction data that has been analysed and discuss how the information could be used by management to assist the day-to-day operations of the organisation.

SPREADSHEETS

In this section we design a spreadsheet-based decision support system for the scenario outlined below. Throughout the design process we will introduce specific spreadsheet concepts of relevance when developing all types of spreadsheet-based information systems and others of particular relevance to decision support systems.

Consider the following scenario:

Management of ABC Corporation wishes to generate forecasts of their corporation's performance over the next five years. The system should meet the following requirements.

The decision support system shall:

- Generate accurate predictions of after tax profit for each of the next five years.
- Alter the predictions appropriately for different forecast inputs for total sales increases or decreases.
- Alter the predictions as the user increases or decreases the cost of goods, administration and marketing relative to total sales made.
- Use and display real data from at least the previous year to verify that inputs and predictions are realistic.

- Detail costs associated with producing goods, administration of the business and marketing for each prediction.
- Express each of the above costs relative to total sales.
- Detail actual sales totals required to meet predicted profits.
- Forecasts will take account of two external variables, inflation and taxation rates.

Identifying inputs and data sources

The data sources determine the accuracy of the inputs into the decision support system. These inputs are processed by the spreadsheet application using formulas to produce the outputs. Data sources for each of the inputs should be chosen carefully to ensure they are accurate.

Typically the outputs of a decision support system are displayed directly to the user of the system. In our example scenario all the outputs will be displayed in a format suitable for use by ABC Corporation's management.

The inputs and their associated data sources for our example are:

- Past year sales records from the company's sales database.
- Past year cost records from the company's accounts databases.
- Current and future predicted inflation rates sourced from the Reserve Bank.
- Company tax rates from the Australian Taxation Office (ATO).
- Percentage increase or decrease in sales from user.
- Percentage of total sales for each cost category, namely goods, administration and marketing from user.

The outputs to the user will include:

- Predicted after tax profit for the next five years adjusted for inflation.
- Total sales, Goods costs, administration costs and marketing costs required to achieve each profit prediction.

The above inputs, outputs and their associated data sources and sinks are detailed on the context diagram in *Fig 5.20*. The user will be able to interactively alter their inputs and immediately view the changes reflected in the outputs.

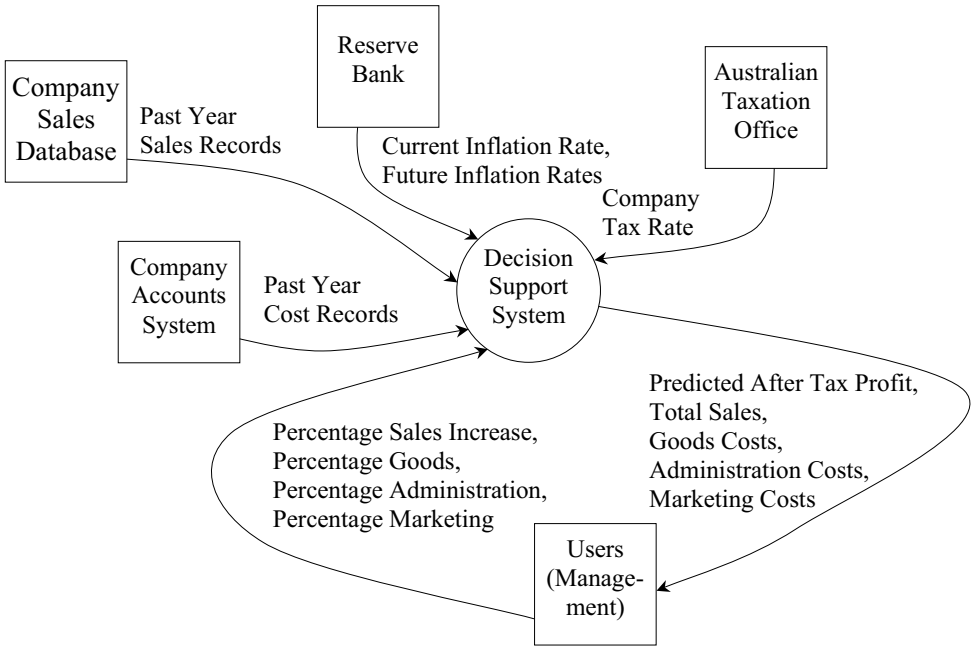


Fig 5.20
Context diagram for ABC Corporation's decision support system.

GROUP TASK Discussion
Discuss how the quality of the data from each of the data sources in *Fig 5.11* could be verified.

Developing formulas to be used

The inputs into spreadsheets are transformed into outputs using formulas. At this stage let us consider the basic formulas that transform the inputs in our ABC Corporation example into the required outputs.

First the Past Year Sales Records will need to be summed to obtain the total sales for the previously completed year. This could be done by executing an SQL query directly with the Company Sales Database or the data for each sale could be imported into the spreadsheet and the sum calculated within the spreadsheet. For our example we will import this data into a separate worksheet.

The Total Sales for the current year is calculated by increasing the Previous Year Total Sales by the Percentage Sales Increase entered by the user. This occurs five times – once for each prediction. Note that we will require a column in our spreadsheet for each of our five prediction years.

Total Sales = (1 + Percentage Sales Increase) * Previous Year Total Sales ..... (1)

This somewhat simplistic calculation does not take account of the effects due to inflation. For example if the total value of sales increase by 4% per year but inflation is running at 6% per year then in real terms total sales will actually have decreased. We have the predicted inflation rate so we can adjust the predicted Total Sales down accordingly to determine the equivalent total sales value in today's money.

Inflation Adjusted Total Sales = Total Sales/(1 + Inflation Rate)$^{\text{Prediction Year}}$..... (2)

The Prediction Year value is the number of years into the future that the prediction is being made. For example for the fifth year prediction Prediction Year = 5 therefore the formula (2) reduces the Total Sales by the Inflation Rate five times.

The Goods Costs is determined by multiplying the Total Sales by the Percentage Goods entered by the user. Similar calculations are made to calculate the Administration Costs and Marketing Costs. We choose not to calculate these values using the Inflation Adjusted Total Sales values so that in the future these figures can be compared with the actual figures.

Goods Costs = (1 + Percentage Goods) * Total Sales ..... (3)
Administration Costs = (1 + Percentage Administration) * Total Sales ..... (4)
Marketing Costs = (1 + Percentage Marketing) * Total Sales ..... (5)

Total Costs is calculated by adding the three values calculated in (3), (4) and (5). Profit for each year is calculated by subtracting Total Costs from Total Sales.

Total Costs = Goods Costs + Administration Costs + Marketing Costs ..... (6)
Profit = Total Sales – Total Costs ..... (7)

We now need to calculate and subtract tax from the company's profit to determine the net profit. Finally we adjust the net profit for inflation to provide more realistic and comparable values. This inflation adjustment formula is similar to that in (2) above.

Tax = Profit * Company Tax Rate ..... (8)
Net Profit = Profit – Tax ..... (9)
Inflation Adjusted Net Profit = Net Profit/(1 + Inflation Rate)$^{\text{Prediction Year}}$ ..... (10)

The significant predictions are the Inflation Adjusted Net Profit values for each of the five years, however our spreadsheet will also display each of the values calculated by all of the above formulas.

Planning the user interface

Each cell within a spreadsheet contains labels, formulas or values. Values are the data that provides input to the formulas. Note that in some instances text can also be processed by formulas. Labels contain text that describes the data and results, together with some instruction on how to use the spreadsheet. Formulas are calculations performed on the values. It is often wise to design the user interface so that there are distinct instruction, input, calculation and output areas. However cells containing formulas also display the output, hence it is common for these areas to overlap significantly or even completely.

For our ABC Corporation example we initially create a pen and paper model (see *Fig 5.21).* Our model essentially splits the spreadsheet into three zones – a calculation and output area, an input area and an output area that contains a chart. Instructional labels are included as titles to describe the nature of the adjoining inputs or outputs.

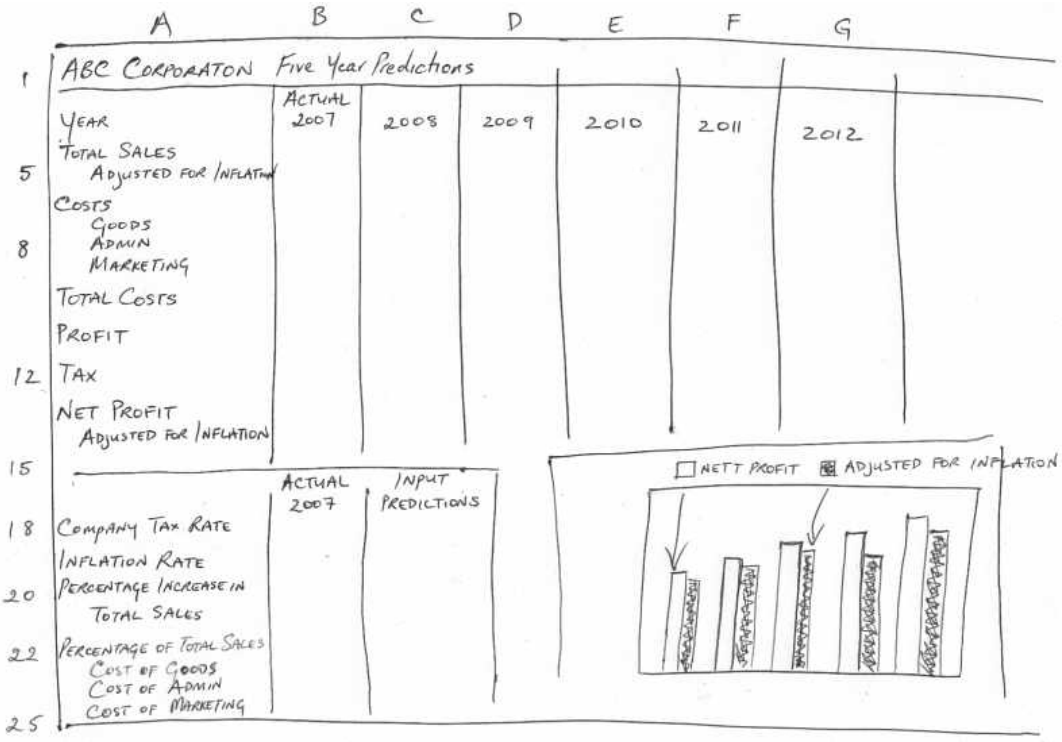


Fig 5.21
Pen and paper design for ABC Corporation's decision support system.

GROUP TASK Discussion
Identify cells on the pen and paper model in *Fig 5.21* that contain labels, cells that will contain formulas and cells that will contain values. Analyse the design of this user interface.

GROUP TASK Practical Activity
Open a new spreadsheet and enter the labels shown in *Fig 5.21*. Save your work, as we will use this spreadsheet throughout this section.

GROUP TASK Discussion
On the previous page we developed a total of 10 formulas. Explain where each of these formulas would be entered within the spreadsheet. Identify any other cells that will require formulas that we have not yet developed.

Extracting information from a database for analysis using a spreadsheet

There are various techniques for extracting data and including it within a spreadsheet. Some possible techniques include:

- Use the DBMS to create a select query, then copy and paste the results into the spreadsheet via the clipboard. This technique is only suitable when importing relatively small amounts of data. Furthermore the user must have the software and sufficient permissions to be able to create and execute queries. This technique can be somewhat inefficient if data is to be extracted on a regular basis.
- Use a front-end application to export to a text file containing the required data. This file is then imported into the spreadsheet. This is a common technique when the application that accesses the database is a commercial product, the user does not have direct access to the database or the user does not have the skills to create their own SQL queries.
- Connect to the database from the spreadsheet application and import the data directly into the spreadsheet. Most spreadsheet applications include such facilities for many common DBMS systems. Essentially an SQL query is written within the spreadsheet – most spreadsheets include a wizard to guide and simplify the query creation and import process. Once the initial connection and query have been created the data can be simply refreshed. During a refresh the connection is made, the query is run and the spreadsheet data is automatically updated to reflect the current data. Furthermore ODBC (Open Database Connectivity) drivers are available for most DBMS systems – ODBC provides a common interface so that various applications can communicate with databases created by specific DBMSs.

GROUP TASK Discussion

There are other techniques for extracting data from databases for analysis using a spreadsheet. Discuss other possible techniques.

For our ABC Corporation example we need to extract the past year's sales records from the company sales database and the past year's cost records from the company accounts database (refer to the context diagram in *Fig 5.20*). We shall connect to these data sources using ODBC connections. In this instance the databases are maintained using Microsoft's SQL Server DBMS. We will use Microsoft Excel as the spreadsheet application. By default Microsoft's Windows operating system includes a suitable SQL Server ODBC driver.

The connection to the Sales and Accounts databases can be created within Excel or they can be created using the ODBC Data Source Administrator included with Windows – in Windows XP open control panel then select administrative tools and open data sources. In either case a DSN (Data Source Name) is created that can be reused to connect to the databases by other applications. *Fig 5.22* shows our two DSNs in the ODBC Data Source Administrator after they have been created. The process and inputs required to create a DSN differ depending on the DBMS and ODBC driver being used.

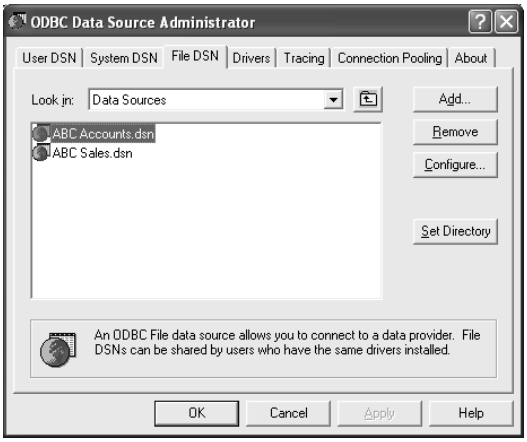


Fig 5.22
Windows's ODBC Data Source Administrator

Within spreadsheet applications it is possible to have more than one worksheet within a single spreadsheet file. When importing large amounts of data into spreadsheets it generally makes sense to import into a new worksheet. For our ABC Corporation example we require two extra worksheets – one for the past year sales data and another for the last year costs data. In Excel choose worksheet from the insert menu – rename each worksheet to reflect its contents (refer *Fig 5.23*).

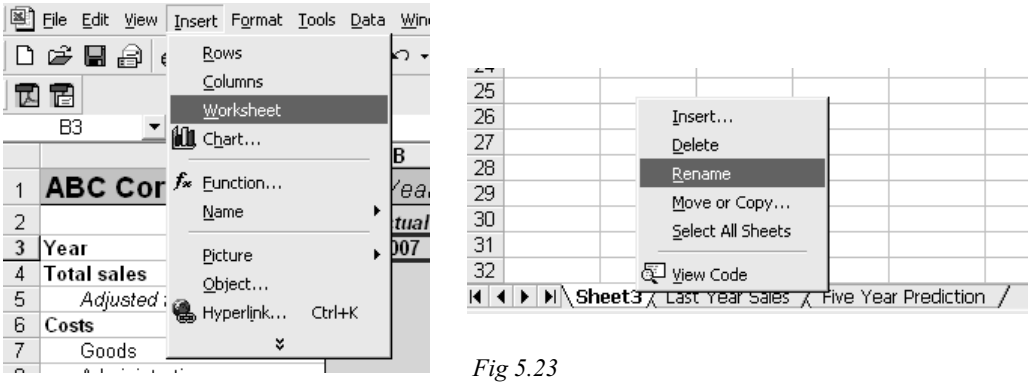


Fig 5.23
Inserting and renaming worksheets in Microsoft Excel.

We can now create the query required to extract the required data using our previously created ODBC data sources. In Excel select New Database Query from the Data then Get External Data menus (*Fig 5.24*). After creating the query the data is imported into the specified worksheet. *Fig 5.25* shows some of the data imported into the two worksheets named Last Year Sales and Last Year Costs. The query is saved within the spreadsheet file. To refresh the data each time the spreadsheet is used simply requires the Refresh Data command to be selected from the Data menu – this command can be seen in *Fig 5.24*. When spreadsheets will be reused (as is usually the case with DSS) the ODBC/query method of extracting data is far more efficient than the user manually performing the import each time the spreadsheet is used.

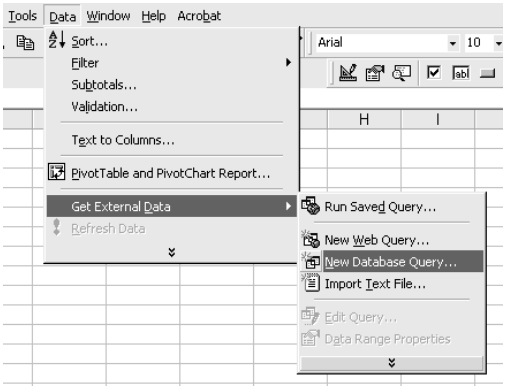


Fig 5.24
Creating a new database query in MS-Excel.

	A	B	C
1	InvoiceNumber	InvDate	InvoiceTotal
2	395	2/07/2006 0:00	3222
3	396	3/07/2006 0:00	2134
4	397	4/07/2006 0:00	1343
5	398	5/07/2006 0:00	678
6	401	8/07/2006 0:00	3200
7	402	9/07/2006 0:00	2560
8	403	10/07/2006 0:00	3323
9	404	11/07/2006 0:00	124
10	405	12/07/2006 0:00	567
11	406	13/07/2006 0:00	890
12	407	14/07/2006 0:00	4100
13	408	15/07/2006 0:00	345
14	409	16/07/2006 0:00	890
15	410	17/07/2006 0:00	678

	A	B	C
1	CostDate	Category	Cost
2	2/07/2006 0:00	Administration	5460
3	5/07/2006 0:00	Marketing	5675
4	8/07/2006 0:00	Goods	4450
5	11/07/2006 0:00	Marketing	3200
6	14/07/2006 0:00	Goods	2560
7	17/07/2006 0:00	Goods	3323
8	20/07/2006 0:00	Goods	5065
9	23/07/2006 0:00	Administration	7850
10	26/07/2006 0:00	Marketing	6450
11	29/07/2006 0:00	Goods	4100
12	1/08/2006 0:00	Goods	2450
13	4/08/2006 0:00	Goods	1890
14	7/08/2006 0:00	Administration	4798
15	10/08/2006 0:00	Marketing	1700

Fig 5.25
Last Year Sales worksheet (left) and Last Year Cost worksheet (right) with sample imported data.

GROUP TASK Discussion
Analyse the sales and costs data imported into the two worksheets in *Fig 5.25*. Discuss how this data can be summarised to calculate the total sales and each of the three costs for the current year.

Spreadsheet Formulas

Formulas within spreadsheets are built using a combination of operators, functions, values and/or cell references. A selection of common operators and functions, together with simple examples are reproduced in *Fig 5.26*, most spreadsheets include a vast list of built-in functions and also have the ability for users to create their own functions. When entering formulas into cells an equals "=" sign is used to indicate a formula, rather than a label or value. The cell references are the addresses of one or more cells; these references provide the links to the data processed by the operators and functions.

	A	B	C
1	**Name**	**Year**	**Score**
2	Peter	11	66
3	Marlene	11	81
4	John	12	70
5	Melissa	12	65
6	Kim	12	78
7	Janine	11	90
8	Aileen	11	81
9	Louise	12	64
10	Ted	11	50.8
11	Margaret	12	63
12	Sam	12	83
13	Luke	11	54

Operator	Description	Example Formula	Result
Arithmetic			
+	Addition	=C1+C2	147
-	Subtraction	=C3-C1	4
*	Multiplication	=B3*B4	132
/	Division	=C2/B2	6
^	Exponentiation	=C2^2	121
Relational			
=	Equals	=B2=B3	TRUE
<>	Does not equal	=B2<>B3	FALSE
>	Greater than	=C2>C3	FALSE
<	Less than	=C2<C3	TRUE
>=	Greater than or equal to	=B2>=B3	TRUE
<=	Less than or equal to	=B3<=B4	TRUE

Function	Description	Example Formula	Result
SUM	Adds values in range	=SUM(C2:C13)	845.8
MAX	Maximum value in range	=MAX(C2:C13)	90
MIN	Minimum value in range	=MIN(C2:C13)	50.8
COUNT	Number of values in range	=COUNT(C2:C13)	12
ABS	Absolute value	=ABS(-60)	60
SQRT	Positive square root	=SQRT(B2)	3.3166…
INT	Integer part	=INT(C10)	50
AVERAGE	Mean value of a range	=AVERAGE(C2:C13)	70.483…
STDEV	Standard deviation of a range	=STDEV(C2:C13)	12.146…
RANK	Position of a value within a range	=RANK(C5,C2:C13)	8
IF	Decision based on a logical test	=IF(C2>C3,A2,A3)	Marlene
SUMIF	Adds values that meet a given criteria	=SUMIF(B2:B13,11,C2:C13)	422.8
HLOOKUP	Horizontal lookup searches for a value in the top row of a range and returns a value in that column but in a different specified row. The top row must be sorted.		
VLOOKUP	Vertical lookup searches for a value in a column and returns a value in that row but in a different specified column. The first column must be sorted.		

Fig 5.26
Selection of common spreadsheet operators and functions.

GROUP TASK Practical Activity
Enter the sample data in *Fig 5.26* into a spreadsheet. Now enter each of the example formulas and check the displayed result matches the result in *Fig 5.26*. Sort the data on column A and create a VLOOKUP formula.

Consider the components of the formula =IF(A1+A2>10,B1,B2). This formula contains four cell references – A1, A2, B1 and B2. It also contains the addition "+" arithmetic operator, the greater than ">" relational operator and the built-in logical IF function. This formula, in English says, "If the sum of the contents of cells A1 and A2 is greater than 10 then return the contents of cell B1 else return the contents of cell B2." The IF function has three parameters, the first is a logical test (or condition) that evaluates to either true or false. The second parameter specifies the calculation to perform if the condition evaluates to true and the last parameter specifies the calculation to perform when the condition if false. In our =IF(A1+A2>10,B1,B2) example both the second and third parameters simply return the contents of a cell, however these parameters can themselves be complex formulas that include other operators, functions and cell references.

The parameters used by many functions refer to a range of cells, for example the formula =SUM(A1:A500) adds up and returns all the values found in cells A1, A2, A3,… A500. A range of cells that forms a rectangle or block on a worksheet is specified using the cell address of the upper left hand corner, followed by a colon ":" and then followed by the cell address of the bottom right hand corner. In most built-in functions a range can include multiple blocks or cells separated by commas. For example the formula =SUM(A1:B4,D5:E7) will add the values in a total of 14 cells.

All spreadsheet formulas are functions – they accept one or more parameters but they always return exactly one value. The above =IF(A1+A2>10,B1,B2) formula has three parameters whilst the =SUM(A1:A500) has a single parameter composed of 500 inputs, however in both cases a single value is returned.

Linking multiple worksheets

Many spreadsheets are composed of multiple worksheets; commonly one sheet contains the formulas and the output whilst other sheets contain data, much like our ABC Corporation example. When a formula includes a reference to cells that are on another worksheet the cell reference must include the name of the worksheet in addition to the address of the cells.

In our ABC Corporation example we need to calculate the total sales for the previous year. The required data is contained in column C on the worksheet we named Last Year Sales (refer *Fig 5.25*). We could construct the following formula within cell B4 of our main worksheet.

=SUM('Last Year Sales'!C1:C10000)

In this formula we have used the range C1:C10000, we use C10000 simply because we anticipate never having more than 10000 rows in our data source. The SUM function ignores cells that do not hold a value so our range can include the heading in cell C1 without affecting the result. Notice that single quotes surround the name of the worksheet, these quotes are only required because the name of our worksheet contains spaces.

It is also possible to construct references that extract data from other spreadsheet files (workbooks). If the Last Year Sales worksheet were stored in a separate file with the path C:\ABCCorp\Sales.xls then the required formula would be:

=SUM('C:\ABCCorp\[Sales.xls]Last Year Sales'!C1:C10000)

Naming ranges

When a range of cells will be used in many formulas it is convenient to give the range a more meaningful name. This is particularly so when the range refers to cells in another worksheet or workbook. In Excel a range is named using the Name command on the Insert menu.

In our ABC Corporation DSS example we require formulas in cells B7, B8 and B9 to determine the total cost of goods, administration and marketing for the previous year (refer to our pen and paper model in *Fig 5.21*). The input data for these formulas is in the Last Year Costs worksheet in columns B and C. Each formula will use the SUMIF function. SUMIF has three parameters – the range of cells to search, the search criteria and the range of cells to sum. The first and third parameters are ranges that are common to all three formulas. We create two named ranges called CostCategories and LastYearCosts that refer to ranges B2:B1000 and C2:C1000 respectively within the Last Year Costs worksheet. The completed formulas together with others that also use named ranges are reproduced in *Fig 5.27*.

	A	B
1	**ABC Corporation**	*Five Year Predictions*
2		***Actual***
3	**Year**	**=YEAR(MAX(LastYearInvoiceDates))**
4	**Total sales**	=SUM(LastYearInvoiceTotals)
5	*Adjusted for inflation*	
6	**Costs**	
7	Goods	=SUMIF(CostCategories,A7,LastYearCosts)
8	Administration	=SUMIF(CostCategories,A8,LastYearCosts)
9	Marketing	=SUMIF(CostCategories,A9,LastYearCosts)
10	**Total Costs**	
11	**Profit**	
12	**Tax**	
13	**Net profit**	
14	*Adjusted for inflation*	

Fig 5.27
Formulas using named ranges.

GROUP TASK Practical Activity
Create the two worksheets Last Year Sales and Last Year Costs and enter (or import) some sample data similar to that shown in *Fig 5.25*. All dates should be from the same financial year, that is, from the beginning of July to the end of June the next year.
Create the named ranges and then the formulas shown in *Fig 5.27*.

GROUP TASK Discussion
Explain how the formula in cell B3 operates to return the year that is being used to generate the actual sales and costs totals.

Absolute and relative references

The ability to copy formulas and have their cell references change automatically to reflect the new location is a powerful feature of spreadsheets. A single formula can be written and then filled down or across to occupy tens, hundred or even thousands of cells. Cell references that change when copied are called relative references; those that do not change when copied are called absolute references.

Relative Reference
A cell reference that refers to a cell in relation to the current location. The cell pointed to changes when the reference is copied to a new location.

Absolute Reference
A cell reference that points to a specific cell. It does not change when copied to a new cell.

Absolute cell references are specified by including a dollar sign "$" before the column and/or row reference. For example the cell reference $A$1 does not change

when copied to a new location whilst the relative reference A1 changes when copied to reflect the new location. A single cell reference can include a relative column reference and an absolute row reference, for example A$1 – in this case the column reference changes relative to the new location but the row reference always point to row 1. Similarly the cell reference $A1 when copied always points to column A, however the row changes to reflect the new location.

	A	B	C	D
1	10	2		
2	20	4	=$A$1+$A1+A$1+A1	=$A$1+$A1+B$1+B1
3			=$A$1+$A2+A$1+A2	=$A$1+$A2+B$1+B2

Fig 5.28
Absolute and relative reference example.

Consider the sample spreadsheet reproduced in *Fig 5.28*. The original formula was entered into cell C2 as =$A$1+$A1+A$1+A1, this formula has then been copied and pasted into cells C3, D2 and D3. In the original C2 formula all references point to cell A1, which is located one row up and two columns to the left of cell C2. When copied all relative row references point to rows one up from the cell containing the formula. Similarly all relative column references point to the cell two columns to the left of the cell containing the formula. Clearly absolute references do not change when copied. For instance in cell D3 we have the formula =$A$1+$A2+B$1+B2, all references preceded by a dollar sign have not changed. All relative row references have changed to point to row 2, as row 2 is one row above the formula's current location in row 3. All relative column references have changed to point to column B, as column B is two columns to the left of the formula's current location in column D.

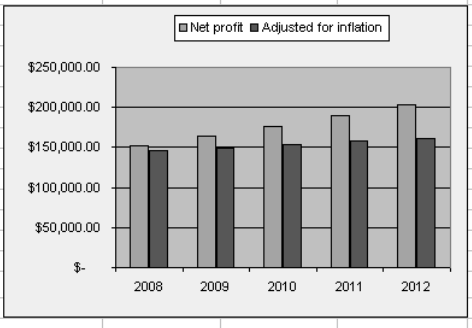
The completed ABC Corporation decision support system spreadsheet is reproduced in *Fig 5.29* and the formulas are shown in *Fig 5.30*. Let us consider how absolute and relative referencing assists when entering these formulas.

Microsoft Excel - ABCCorporationFinal.xls

File Edit View Insert Format Tools Data Window Help Acrobat

C21 = 7.5%

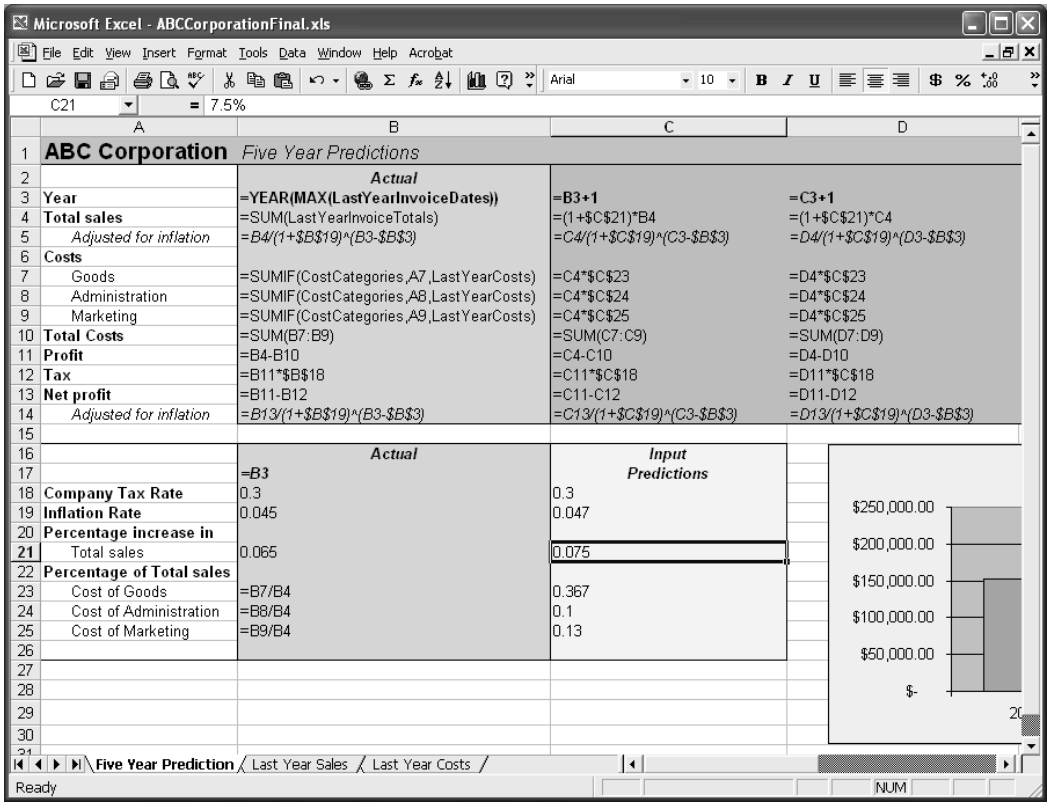
	A	B	C	D	E	F	G
1	**ABC Corporation**	*Five Year Predictions*					
2		*Actual*			*Predictions*		
3	**Year**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
4	**Total sales**	$ 502,207.60	$ 539,873.17	$ 580,363.66	$ 623,890.93	$ 670,682.75	$ 720,983.96
5	*Adjusted for inflation*	*$ 502,207.60*	*$ 515,638.18*	*$ 529,427.93*	*$ 543,586.46*	*$ 558,123.63*	*$ 573,049.57*
6	**Costs**						
7	Goods	$ 186,832.00	$ 198,133.45	$ 212,993.46	$ 228,967.97	$ 246,140.57	$ 264,601.11
8	Administration	$ 51,828.00	$ 53,987.32	$ 58,036.37	$ 62,389.09	$ 67,068.28	$ 72,098.40
9	Marketing	$ 67,299.00	$ 70,183.51	$ 75,447.28	$ 81,105.82	$ 87,188.76	$ 93,727.91
10	**Total Costs**	$ 305,959.00	$ 322,304.28	$ 346,477.10	$ 372,462.89	$ 400,397.60	$ 430,427.42
11	**Profit**	$ 196,248.60	$ 217,568.89	$ 233,886.55	$ 251,428.05	$ 270,285.15	$ 290,556.54
12	**Tax**	$ 58,874.58	$ 65,270.67	$ 70,165.97	$ 75,428.41	$ 81,085.54	$ 87,166.96
13	**Net profit**	$ 137,374.02	$ 152,298.22	$ 163,720.59	$ 175,999.63	$ 189,199.60	$ 203,389.57
14	*Adjusted for inflation*	*$ 137,374.02*	*$ 145,461.53*	*$ 149,351.62*	*$ 153,345.74*	*$ 157,446.68*	*$ 161,657.28*
15							
16		*Actual*	*Input*				
17		**2007**	*Predictions*				
18	**Company Tax Rate**	30.0%	30.0%				
19	**Inflation Rate**	4.5%	4.7%				
20	**Percentage increase in**						
21	Total sales	6.5%	7.5%				
22	**Percentage of Total sales**						
23	Cost of Goods	37.2%	36.7%				
24	Cost of Administration	10.3%	10.0%				
25	Cost of Marketing	13.4%	13.0%				
26							



Five Year Prediction | Last Year Sales | Last Year Costs

Ready NUM

Fig 5.29
Completed ABC Corporation decision support system spreadsheet.



Microsoft Excel - ABCCorporationFinal.xls

File Edit View Insert Format Tools Data Window Help Acrobat

C21 = 7.5%

	A	B	C	D
1	**ABC Corporation** *Five Year Predictions*			
2		***Actual***		
3	**Year**	**=YEAR(MAX(LastYearInvoiceDates))**	**=B3+1**	**=C3+1**
4	**Total sales**	=SUM(LastYearInvoiceTotals)	=(1+$C$21)*B4	=(1+$C$21)*C4
5	*Adjusted for inflation*	*=B4/(1+$B$19)^(B3-$B$3)*	*=C4/(1+$C$19)^(C3-$B$3)*	*=D4/(1+$C$19)^(D3-$B$3)*
6	**Costs**			
7	Goods	=SUMIF(CostCategories,A7,LastYearCosts)	=C4*$C$23	=D4*$C$23
8	Administration	=SUMIF(CostCategories,A8,LastYearCosts)	=C4*$C$24	=D4*$C$24
9	Marketing	=SUMIF(CostCategories,A9,LastYearCosts)	=C4*$C$25	=D4*$C$25
10	**Total Costs**	=SUM(B7:B9)	=SUM(C7:C9)	=SUM(D7:D9)
11	**Profit**	=B4-B10	=C4-C10	=D4-D10
12	**Tax**	=B11*$B$18	=C11*$C$18	=D11*$C$18
13	**Net profit**	=B11-B12	=C11-C12	=D11-D12
14	*Adjusted for inflation*	*=B13/(1+$B$19)^(B3-$B$3)*	*=C13/(1+$C$19)^(C3-$B$3)*	*=D13/(1+$C$19)^(D3-$B$3)*
15				
16		***Actual***	***Input***	
17		***=B3***	***Predictions***	
18	**Company Tax Rate**	0.3	0.3	
19	**Inflation Rate**	0.045	0.047	
20	**Percentage increase in**			
21	Total sales	0.065	0.075	
22	**Percentage of Total sales**			
23	Cost of Goods	=B7/B4	0.367	
24	Cost of Administration	=B8/B4	0.1	
25	Cost of Marketing	=B9/B4	0.13	
26				
27				
28				
29				
30				

Five Year Prediction / Last Year Sales / Last Year Costs

Ready NUM

Fig 5.30
Completed ABC Corporation decision support system spreadsheet showing formulas

In Excel the keyboard shortcut Ctrl+~ toggles viewing formulas and viewing results. Notice in *Fig 5.30* that all the formulas in column C are, in a relative sense, the same as the formulas contained in column D (and also columns E, F and G). Therefore it is only necessary to construct the formulas once, in column C. These formulas can then be filled to the right into columns D to G using Excel's Edit-Fill-Right command.

Consider the formulas developed earlier:

Total Sales = (1 + Percentage Sales Increase) * Previous Year Total Sales ..... (1)
Inflation Adjusted Total Sales = Total Sales/(1 + Inflation Rate)$^{\text{Prediction Year}}$ ..... (2)
Goods Costs = (1 + Percentage Goods) * Total Sales ..... (3)
Administration Costs = (1 + Percentage Administration) * Total Sales ..... (4)
Marketing Costs = (1 + Percentage Marketing) * Total Sales ..... (5)
Total Costs = Goods Costs + Administration Costs + Marketing Costs ..... (6)
Profit = Total Sales – Total Costs ..... (7)
Tax = Profit * Company Tax Rate ..... (8)
Net Profit = Profit – Tax ..... (9)
Inflation Adjusted Net Profit = Net Profit/(1 + Inflation Rate)$^{\text{Prediction Year}}$ ..... (10)

GROUP TASK Discussion
Compare each of the above formulas with the corresponding spreadsheet formulas on the screen in *Fig 5.30*.

GROUP TASK Discussion
Identify and discuss the source and purpose of the data in each of the cells within the range B18:C25.

SET 5B

1. Which type of decision support system simulates the structure of the human brain?
   (A) Spreadsheets
   (B) Expert Systems
   (C) Artificial Neural Networks
   (D) Databases

2. Which tool specialises in reproducing a person's specialised expertise in a particular knowledge area?
   (A) Spreadsheets
   (B) Expert Systems
   (C) Artificial Neural Networks
   (D) Databases

3. Software operates in the background to automatically delete spam based on a list of email addresses entered by the user. This is an example of an:
   (A) intelligent agent.
   (B) agent but not an intelligent agent.
   (C) email client application.
   (D) POP client application.

4. Which of the following is true of all spreadsheet formulas?
   (A) A single output is produced from one or more inputs.
   (B) One or more outputs are produced from one or more inputs.
   (C) A single output is produced from a single input.
   (D) One or more outputs are produced from a single input.

5. During data mining records are classified into groups with similar characteristics. Some records are classified into more than one group. Which data mining tool is possibly being used?
   (A) Decision tree algorithm
   (B) Rule induction
   (C) Non-linear regression
   (D) K-nearest neighbour

6. Which of the following lists an arithmetic operator first, a logical operator next and finally a function name?
   (A) <, /, COUNT
   (B) SUM, =, +
   (C) *, >=, IF
   (D) =, MIN, ^

7. All cells in the range A1:B3 contain the value 5, all cells in the range D2:G4 contain the value 3 and all other cells in the range A1:G4 are empty. What value would be displayed in cell A5 if it contains the formula =COUNT(A1:G4)?
   (A) 18
   (B) 28
   (C) 66
   (D) 12

8. Cell D1 contains =$A2-B$5. When copied into cell F6 it will appear as:
   (A) =$A2-B$5
   (B) =$A3-B$5
   (C) =$A7-D$5
   (D) =$C2-B$10

9. Naming a range of cells is recommended under which of the following circumstances?
   (A) The cells are on a different worksheet.
   (B) The named range will be used in many formulas.
   (C) To improve the readability of formulas that reference the named range.
   (D) All of the above.

10. When designing the user interface of spreadsheets it is common practice to:
    (A) combine input and output areas.
    (B) separate input and outputs areas.
    (C) combine instruction and calculation areas
    (D) separate calculation and output areas.

11. Define each of the following spreadsheet terms:
    (a) cell
    (b) worksheet
    (c) value
    (d) label
    (e) formula
    (f) range
    (g) named range
    (h) cell reference

12. Distinguish between each of the following:
    (a) absolute references and relative references.
    (b) expert systems and neural networks.
    (c) data warehouses and data marts.
    (d) DSSs and MISs,
    (e) OLAP and charts created using spreadsheets
13. Outline the essential characteristics of:
    (a) Data warehouses
    (b) OLAP
    (c) MIS
    (d) Data marts
    (e) GDSS
    (f) Intelligent agents
    (g) OLTP
    (h) GIS
    (i) Data mining
14. Construct a spreadsheet formula to calculate and return each of the following:
    (a) The range of values within the range of cells A1 to A500.
    (b) If the average of the values in cells B3 to B10 is less than 50 return the word "Fail", otherwise return the word "Pass".
    (c) Return True if the value in cell A1 is a whole number and False if it is not.
15. Construct a spreadsheet to record class marks for an assessment task. The spreadsheet is to calculate each student's position in class and the class mean, mode and median.

Charts and graphs

Charts and graphs are used to visually illustrate the relationships between two or more sets of data. For example the rainfall each month for a particular town contains two sets of data, the months and the rainfall figures. Consider the example table and column graph in *Fig 5.31*; within the table the precise value of each data item can be seen, however the graph more effectively shows the distribution of rainfall throughout the entire year. Different information is highlighted on the graph compared to the table.

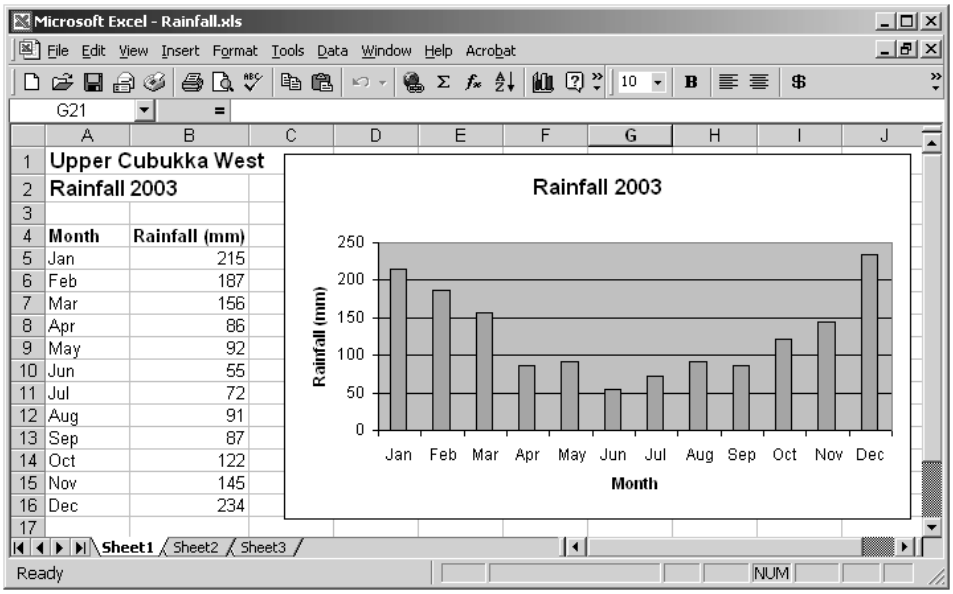


Fig 5.31
Rainfall data displayed in a table and as a column graph using Microsoft Excel.

Different types of graph or chart emphasise different types of information; let us consider examples of the more common graph types together with their major purpose in terms of communicating information.

- **Column and bar graphs**

Column graphs display data values vertically whereas bar graphs display data values horizontally. Both column and bar graphs are well suited to sets of data where the categories or entities are not numeric or have no inherent order; in this context the set of numeric values measure the same thing for various different entities. For example in *Fig 5.32* each state is a different entity; the order in which these entities appear is not important, whereas each numeric value is a measurement of the same quantity. A line graph would be inappropriate for graphing this data, as points on the lines between different states have no meaning.

The graphs in *Fig 5.32* are based on a single data series. Column and bar graphs can be created to graph multiple data series for each entity. Each data series can be shown as a separate column or bar, or they may be stacked together to show the total for each entity.

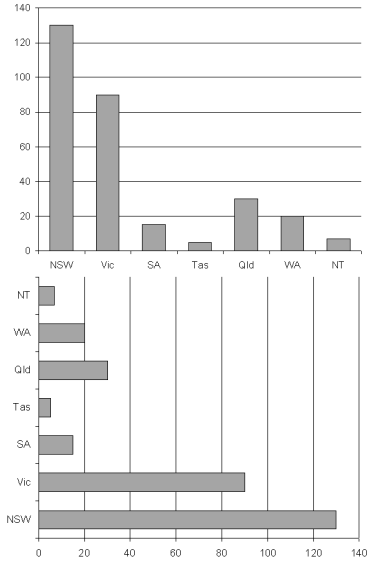


Fig 5.32
Column graphs and bar graphs display the relative differences between data values.

- **Line graphs**

Line graphs are commonly used to display a series of numeric data items that change over time. They are used to communicate trends apparent in the data. Lines connecting consecutive data points highlight the changes occurring; when all such lines are plotted overall trends emerge.

When using line graphs the source data must be sorted by the data to be graphed along the horizontal or x-axis. For example in *Fig 5.33* the horizontal axis contains the months of the year, if this data were not sorted correctly then the trends communicated by the lines connecting each data value would be incorrect.

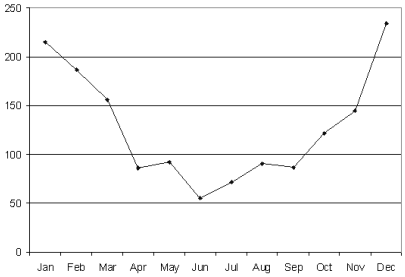


Fig 5.33
Line graphs highlight trends in a data series. Both axes should contain ordered data.

- **Pie charts**

Pie charts show the contribution or percentage that each data item makes to the total of all the data items. For example *Fig 5.34* clearly communicates that NSW contributes far more to the total than any of the other states and that Tas. and NT contribute the least.

The nature of pie charts means they are only able to plot a single data series. Pie charts do not provide information on the precise value of each data item rather they communicate the relative differences between each discrete category on the graph.

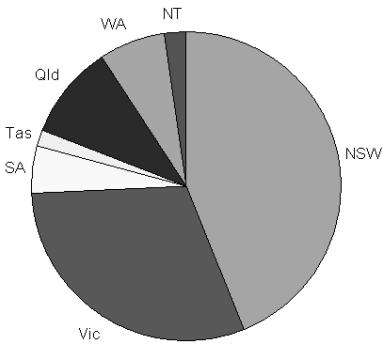


Fig 5.34
Pie charts highlight the contribution each data item makes to the total.

- **XY graphs**

XY graphs are used to plot pairs of points. The source data being composed of a series of ordered pairs. Each ordered pair is composed of an X coordinate and a Y coordinate used to determine the position of a single point on the graph. When these points are connected using a series of smooth curves a continuous representation of the relationship between the X and Y coordinates is produced.

In contrast to line graphs, it is not necessary for the X coordinates to be evenly spaced. It is quite common to obtain samples at random times which can then be connected to form a continuous curve. Furthermore the curve can be extrapolated in an attempt to describe trends outside the range of the sample data.

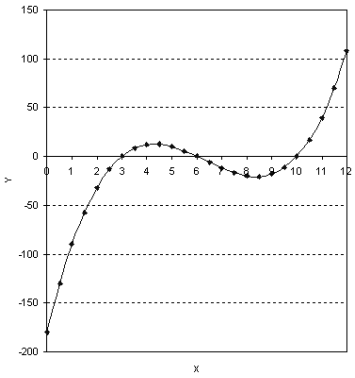


Fig 5.35
XY graphs are used to plot a series of ordered pairs.

GROUP TASK Discussion
Assess the suitability of the column graph used on the completed ABC Corporation DSS spreadsheet in *Fig 5.29*.

GROUP TASK Practical Activity
The range and scale of the axes on a graph can skew the graph and introduce bias. Explore using a chart within a spreadsheet.

Spreadsheet macros

Macro
A short user defined command that executes a series of predefined commands.

Macros are used to automate processing in all types of applications including spreadsheets. A macro is a single command or keyboard shortcut that causes a set of predefined commands to execute. The set of commands can be created by recording a sequence of user keyboard and mouse actions or the commands can be entered directly as programming code. Applications that allow keyboard and mouse actions to be recorded actually convert these actions into equivalent lines of programming code. When the macro command (or its assigned shortcut key combination) is initiated the lines of program code are executed.

The use of macros allows common sequences of commands to be stored and then reused many times. Let us consider two macros for our ABC Corporation DSS Excel spreadsheet. The first ResetInputs macro will reset all the Prediction Inputs (C18:C25) to the same values as the actual values from the previous year (B18:B25). The second Zoom macro will change the scale on the y-axis of the chart to more obviously show the profit differences between each prediction year. We shall assign each macro to a command button on the spreadsheet.

In Excel we can create the first ResetInputs macro by recording keystrokes. Essentially we copy and paste the values from B18:B25 to C18:C25 (refer *Fig 5.29*). The following steps are performed in Microsoft Excel:

1. On the Tools menu select Macro then Record New Macro...
2. In the Record Macro dialogue name the macro ResetInputs and assign the shortcut key combination Ctrl+r (see *Fig 5.36*).
3. Select the range of cells B18:B25 and then type Ctrl+C to copy these cells.
4. Select cell C18 and choose Paste Special from the Edit menu. The dialogue in *Fig 5.37* is displayed. Select the option in the dialogue so that just values rather than the formulas are pasted.
5. Hit the Escape key to remove the selection around cells B18:B25.
6. Use the mouse to select cell C21 as this is the primary input cell. We wish to have this cell selected after the macro executes.
7. Finally stop recording using the on screen stop button or via the Stop command on the Tools- Macro menu.

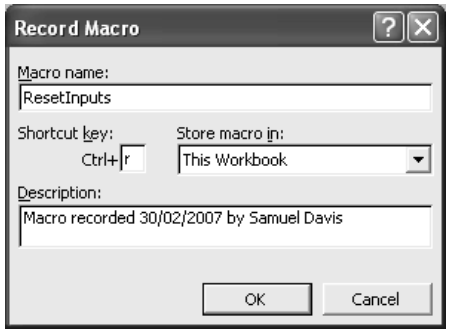


Fig 5.36
Microsoft Excel Record Macro dialogue.

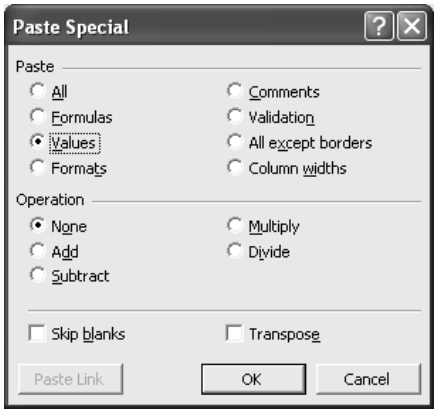


Fig 5.37
Microsoft Excel Past Special dialogue.

GROUP TASK Practical Activity
Create the above macro in Excel. Test the macro operates as expected using the Ctrl+r shortcut. Suggest reasons why this macro would be useful for users of the ABC Corporation DSS spreadsheet.

We now add a command button to the spreadsheet that will also activate our ResetInputs macro. In Excel first display the Microsoft Office Control Toolbox toolbar (see *Fig 5.38*) – choose Toolbars on the View menu. Select the Command Button icon and draw a command button under the prediction inputs. Clicking view code (with the new command button selected) opens the Visual Basic editor and creates a sub program that will execute when the command button is clicked by a user. We simply enter the command ResetInputs – the name of our previously recorded macro. Now close the Visual Basic editor and click the design mode icon to exit design mode. The command button when clicked now runs the ResetInputs macro. Each control has a variety of properties that can be altered – select the command button and click the control properties icon on the Control Toolbox toolbar – change the caption property to "Reset Inputs".

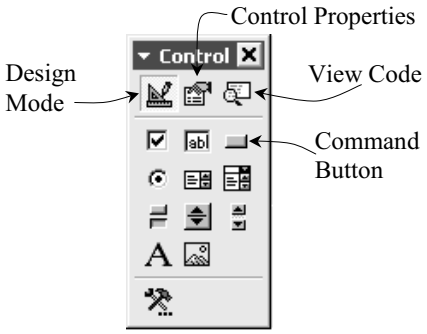


Fig 5.38
Microsoft Office Control Toolbox toolbar.

Creating the second Zoom macro is beyond the requirements of the IPT course, therefore we will simply describe the general operation of the completed macro. By default column graphs created in Excel have a y-axis that commences at zero and automatically adjusts to suit the largest value to be graphed (refer to the graph in the *Fig 5.20* screenshot). The zoom macro assigns a new minimum value to the y-axis of the graph using a command button. The new minimum value is calculated in cell C29 on the spreadsheet using the formula =ROUNDDOWN(MIN(C14:G14),-4)-10000. This formula finds the smallest prediction, rounds it down to the nearest 10000 and then subtracts 10000. Clicking the command button toggles the minimum value between the calculated minimum in cell C29 and zero. *Fig 5.39* shows an example graph where the minimum y-axis value has been set to $130,000.

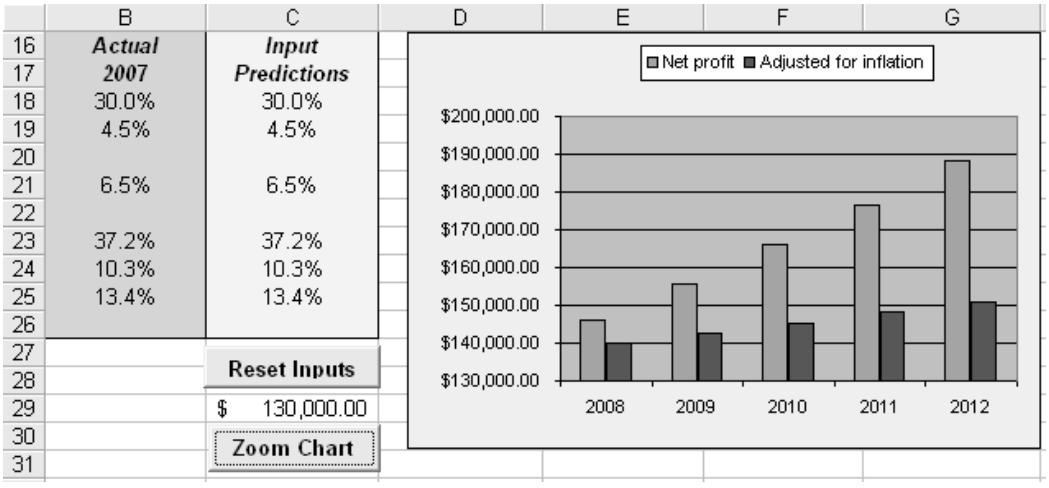


Fig 5.39
Extract of ABC Corporation spreadsheet showing zoomed chart and macro command buttons.

The Visual Basic code to adjust the minimum y-axis value on the chart is reproduced in *Fig 5.40*. When the existing MinimumScale value for the y-axis of the chart is zero the Zoom procedure sets the MinimumScale value to the value in cell C29. If the MinimumScale value is not zero then it is set to zero. The screenshot in *Fig 5.40* also includes the code created when the ResetInputs macro was recorded.

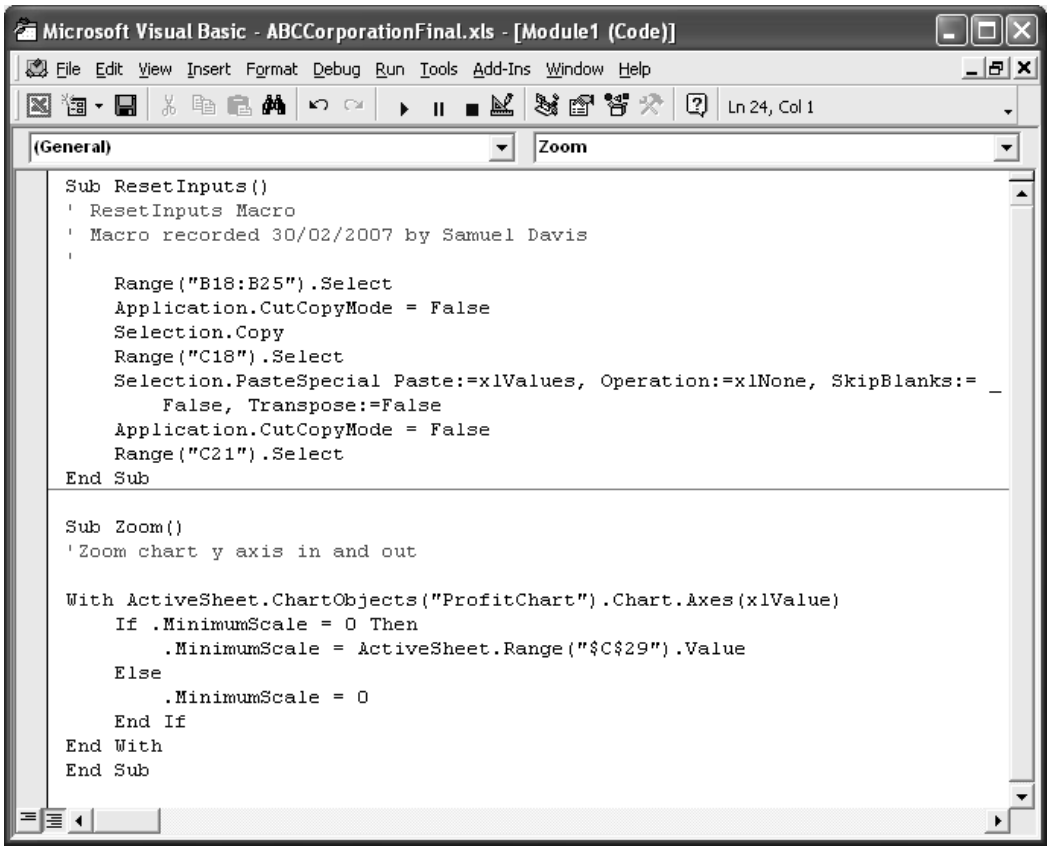


Fig 5.40
Visual Basic code for the ResetInputs and Zoom macros.

GROUP TASK Discussion
Examine the code in the ResetInputs procedure in *Fig 5.40*. Determine the recorded keyboard and mouse actions that correspond to each of the lines of code.

GROUP TASK Research
Using the Internet, or otherwise, research a variety of other examples of macros used within spreadsheets.

Spreadsheet templates

A spreadsheet template is simply a reusable spreadsheet that includes all the required headings, titles, formulas, formatting, charts, external links, macros and other components needed to solve a particular problem. The user opens the template and enters their own data, the spreadsheet then performs its processing based on these new inputs. Professional templates are available that make extensive use of custom formatting and macros. It is often more cost effective to purchase a professionally designed template rather than reinvent the wheel by creating the spreadsheet from scratch.

Many users simply open an existing version of the spreadsheet, change the data, make other changes and save the result using a different name. Using this technique it is possible that the user will inadvertently overwrite their original file. To overcome this problem it is possible to save the original version specifically as a template file. New spreadsheets can then be created based on this template. The original template is not

altered rather its content is copied into the new spreadsheet. In Excel the available templates are displayed when a new spreadsheet is created using the new command on the file menu. A range of professional templates is available commercially to accomplish common tasks and many businesses create their own templates for use by their employees. Such professionally developed templates often include custom toolbars, menus and other advanced functionality that is difficult and time consuming for casual spreadsheet users to develop.

GROUP TASK Research
Research and briefly describe the functionality of some different spreadsheet templates that perform decision support tasks.

GROUP TASK Practical Activity
Save a completed spreadsheet as a template file. Create and save a number of worksheets based on this template.

ANALYSING USING SPREADSHEETS

What-if analysis and scenarios

What-if scenarios allow you to consider the effect of different inputs. Different sets of inputs are processed and analysed to determine a corresponding set of resulting outputs. The "What-if" analysis process aims to produce the most likely outputs for each set of inputs. The aim being to predict the likely, or at least possible, consequences for each particular set of decision inputs; these predictions can then be used to make more informed decisions.

When performing 'What-if' analysis it is the inputs or data that is changed; the processing that transforms these inputs into predictions remains the same. Therefore when designing a what-if scenario it is vital to understand the detailed nature of the analysis processes for all possible sets of inputs. In most cases these processes operate on numeric data using various mathematical and statistical calculations, for this reason spreadsheets are particularly suitable software tools for what-if analysis. Spreadsheets automatically recalculate each formula immediately after any input data is altered, therefore the information displayed is updated to reflect the current data.

In most spreadsheet applications sets of inputs can be saved as a scenario. Each scenario can be retrieved for further analysis and the primary outputs for all scenarios can be generated – usually on a new worksheet.

Consider ABC Corp. spreadsheet:

In Excel different scenarios can be created. *Fig 5.41* shows three scenarios within the Scenario Manager dialogue. Each of these scenarios has a different set of inputs for cells C21, C23, C24 and C25. The show command button causes the scenario to be displayed within the spreadsheet. Edit allows the inputs and their values to be altered. Summary is used to execute all scenarios and produce a table of their inputs together with their critical outputs (*Fig 5.42*). In our example the outputs have been specified as the Net Profits Adjusted for Inflation – cells B14:G14.

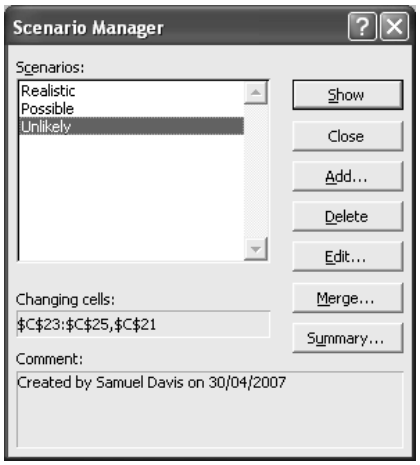


Fig 5.41
Scenario Manager in MS-Excel.

Scenario Summary				
	Current Values:	Realistic	Possible	Unlikely
Changing Cells:				
$C$23	37.2%	36.0%	34.0%	30.0%
$C$24	10.3%	9.0%	10.0%	8.0%
$C$25	13.4%	12.0%	10.0%	10.0%
$C$21	6.5%	7.0%	10.0%	12.0%
Result Cells:				
$B$14	$ 137,374.02	$ 137,374.02	$ 137,374.02	$ 137,374.02
$C$14	$ 140,003.19	$ 154,780.86	$ 170,221.94	$ 195,923.44
$D$14	$ 142,682.67	$ 158,483.75	$ 179,180.99	$ 209,984.93
$E$14	$ 145,413.44	$ 162,275.23	$ 188,611.57	$ 225,055.62
$F$14	$ 148,196.48	$ 166,157.41	$ 198,538.50	$ 241,207.94
$G$14	$ 151,032.77	$ 170,132.47	$ 208,987.89	$ 258,519.51

Notes: Current Values column represents values of changing cells at time Scenario Summary Report was created. Changing cells for each scenario are highlighted in gray.

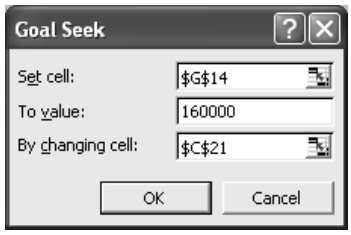
Fig 5.42
Scenario summary for the ABC Corporation DSS Spreadsheet.

GROUP TASK Discussion
Brainstorm possible situations where scenarios and scenario summaries would be of use during decision-making processes.

GROUP TASK Practical Activity
Using the ABC Corporation spreadsheet or some other spreadsheet, create various "what if" scenarios. Store these scenarios and create a scenario summary similar to the one reproduced in *Fig 5.42*.

Goal seeking

Goal seeking starts with a desired output and then determines the required inputs. It is essentially the opposite of performing "What if" analysis. Within spreadsheets a desired value is specified for a cell that calculates an output. The spreadsheet application then determines the input required to calculate the desired value.



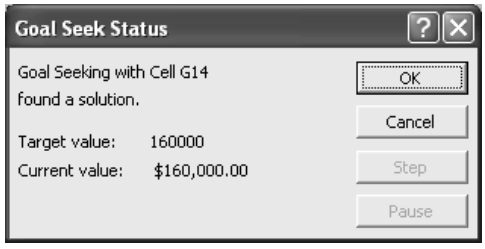


Fig 5.43
Excel's Goal Seek input and result dialogues.

In Excel a Goal Seek function is available. We can use this function to perform goal seeking in our ABC Corporation spreadsheet. Say, the goal is to achieve an inflation adjusted profit of $160,000 in the fifth prediction year. Cell G14 contains the fifth year inflation adjusted profit. We wish to achieve this goal by altering the percentage increase in total sales within cell C21. Refer *Fig 5.43*, clicking the OK button causes the goal seeking function to execute. In this case a solution is found and cell C21 is set to the required input value – 7.7% for the current data as shown in *Fig 5.44* on the next page.

GROUP TASK Practical Activity
Using Excel's built-in goal seeking function only one input is altered. Experiment using the ABC Corp. spreadsheet to determine different sets of inputs that achieve the same goal.

	A	B	C	D	E	F	G
13	**Net profit**	$ 137,374.02	$ 148,000.77	$ 159,449.56	$ 171,784.00	$ 185,072.58	$ 199,389.11
14	*Adjusted for inflation*	$ 137,374.02	$ 141,627.53	$ 146,012.74	$ 150,533.73	$ 155,194.71	$ 160,000.00
15							
16		***Actual 2007***	***Input Predictions***				
17							
18	**Company Tax Rate**	30.0%	30.0%				
19	**Inflation Rate**	4.5%	4.5%				
20	**Percentage increase in**						
21	Total sales	6.5%	7.7%				
22	**Percentage of Total sales**						
23	Cost of Goods	37.2%	37.2%				
24	Cost of Administration	10.3%	10.3%				
25	Cost of Marketing	13.4%	13.4%				
26							
27							
28			Reset Inputs				
29			$ 130,000.00				
30			Zoom Chart				
31							

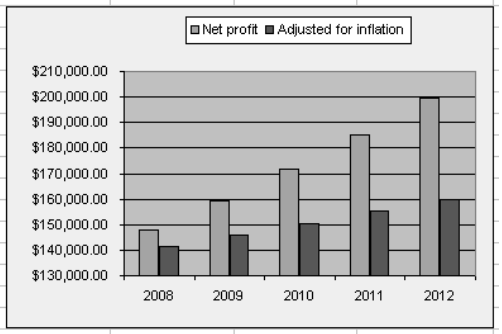


Fig 5.44
ABC Corporation DSS example after goal seeking.

Consider the following:

The UAI Estimator is a program developed by Parramatta Education Centre for estimating UAIs based on historical data from the Board of Studies and the Technical Committee on the Scaling of the HSC. The estimates reflect what UAI a student would have achieved had the entered set of results been achieved in each of the five HSC prediction years. For example in *Fig 5.45* the marks entered for Christopher Eclectus would have achieved approximate UAIs of 77.10 in the 2002 HSC, 76.75 in the 2003 HSC, 73.20 in the 2004 HSC, 72.35 in the 2005 HSC and 74.15 in the 2006 HSC.

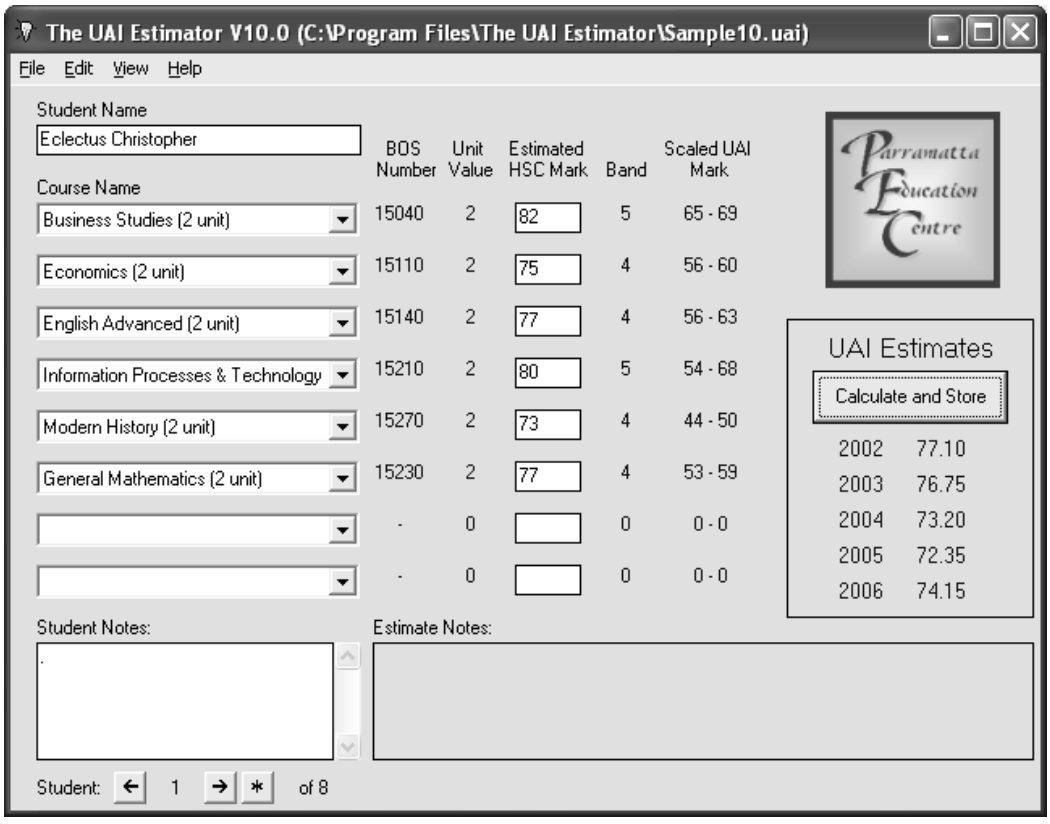


Fig 5.45
Sample UAI Estimator Version 10.0 screen.

Christopher wishes to achieve a UAI of 80, hence he would like some indication of the HSC marks required to achieve this result. The UAI Estimator application includes a Reverse function to assist Christopher – the Reverse dialogue is reproduced in *Fig 5.46*. Christopher feels he cannot improve his Economics and his General Mathematics results; hence these marks are not ticked on the dialogue and will not be altered by the reverse function.

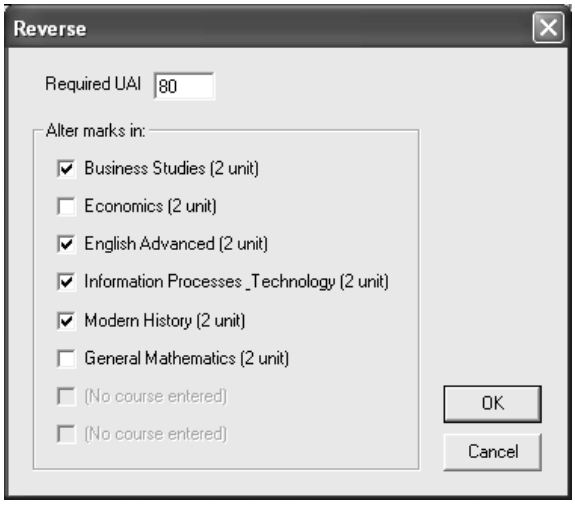


Fig 5.46
The UAI Estimator Reverse dialogue.

The results after the Reverse function has executed are displayed in *Fig 5.47*. The reverse function has altered Christopher's HSC mark estimates for Business Studies, English, IPT and Modern History proportionally to achieve his goal of an 80 UAI. The Reverse function seeks a 2006 UAI estimate of 80 without altering the HSC marks entered for either Economics or General Mathematics.

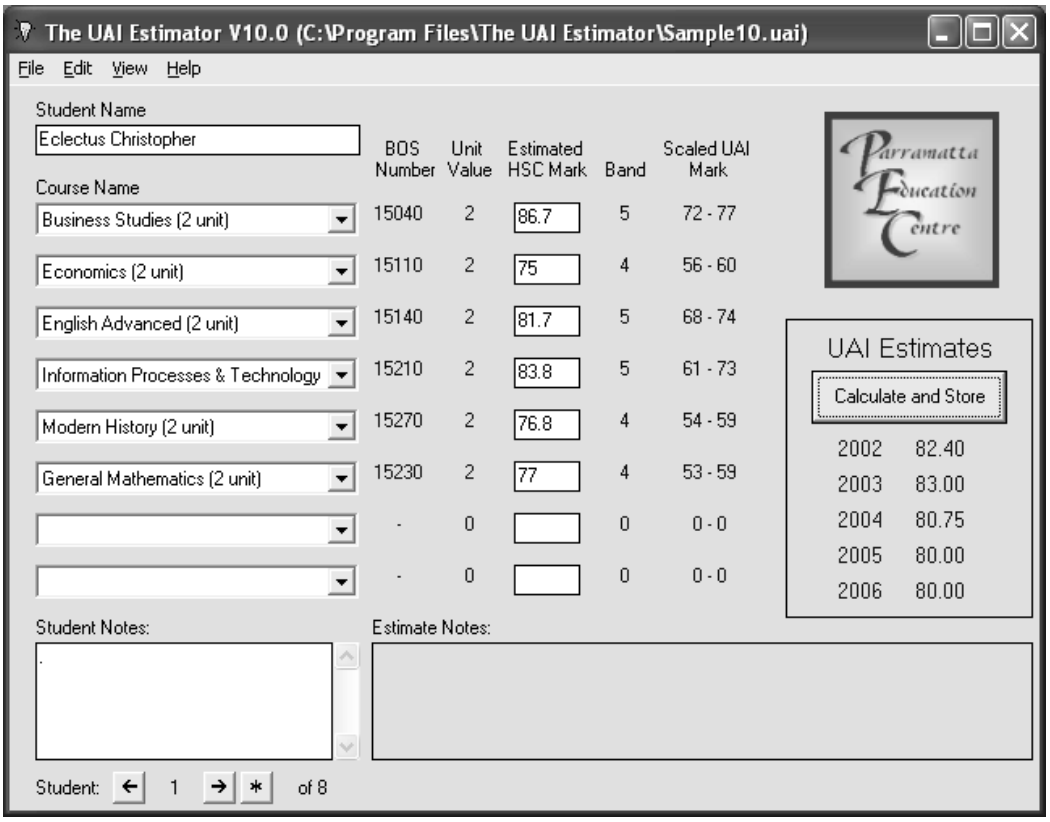


Fig 5.47
UAI Estimator screen after the Reverse function has run.

GROUP TASK Discussion
Compare and contrast the goal seek function in Excel with the Reverse function in the UAI estimator.

Statistical analysis

Statistical analysis is a broad field that aims to summarise and make generalisations about data. Statistical analysis is a branch of applied mathematics used by experts in almost all fields of endeavour. In this section we can only hope to briefly describe some of the simpler statistical analysis techniques. In general statistical analysis is performed over one or more sets of real world data to produce statistical measures that help describe the data as a whole. These statistical measures can then be used to comment on characteristics of the data, make comparisons with other data sets or make predictions.

Some commonly used statistical techniques and measures include:

- Charting or graphing data series. Often sample data is collected that describes a small proportion of the total population; in these cases frequency distributions are often generated and then charted as frequency or cumulative frequency histograms. Such charts show the general shape of the underlying data and are useful to visually identify relationships and general trends within data.
- Charted sample data can be used to generate trend lines that can then be used to determine the most likely values for unknown data inputs. Trendlines can be extrapolated forwards and backward to allow predictions to be made that are outside the range of the known data values. Trendlines can also be used to estimate the value of outputs between known data items – this process is known as interpolation. Most spreadsheets are able to automatically generate trendlines either directly on charts or using various statistical formulas. Before creating a trendline the general shape of the distribution should be determined – Excel is able to generate linear (straight lines), logarithmic, exponential and polynomial trendlines.
- Measures of central tendency such as average (mean), mode and median. The mean is the sum of the data items divided by the number of data items. The mode is the most commonly occurring data item. The median is the middle data item when all data items are sorted.
- Measures of spread such as range, variance and standard deviation. The range is the difference between the highest and lowest data items. Variance and standard deviation are measures used to describe the average amount by which each score differs from the mean.
- Comparisons between two or more data sets by comparing measures of central tendency and spread or using measures such as correlation. The range of possible correlations is from –1 to 1. A correlation of 1 means the data sets increase or decrease together perfectly. Negative correlations mean that as one data set increases the other decreases (or vice versa). As the correlation gets closer to 1 (or –1) the relationship between the data sets becomes stronger. Conversely as correlations approach zero the relationship between the data sets becomes weaker. A correlation of 0 means there is no relationship between the data sets.
- Probability measures such as confidence coefficients and confidence intervals for predictions. For example a prediction may be made with a confidence coefficient of 90% which essentially means the probability of the prediction coming true is 90%. Confidence intervals are also often quoted, for example "I am 90% sure that profit will be within the interval $150,000 to $160,000." Confidence intervals are often quoted with 90%, 95% or 99% confidence coefficients. In general confidence intervals are smaller for larger data sets and are larger for smaller data sets. Similarly, data sets with smaller standard deviations have smaller confidence intervals, whilst data sets with larger standard deviations result in larger confidence intervals.

Consider the following:

Fred is an IPT student who has a theory that there is a close relationship between a student's HSC IPT result and their results in English and Maths. He has collected marks from each of his IPT classmates. Fred's spreadsheet is reproduced in *Fig 5.48*. Fred intends to predict other student's English and Maths results based entirely on their IPT result.

	A	B	C	D
1	*Freds HSC IPT Predictor of HSC English and Maths Results*			
2				
3	**SAMPLE DATA**			
4	**Name**	**IPT**	**English**	**Maths**
5	Fred	70	72	69
6	Mary	65	65	61
7	John	76	77	72
8	Sue	58	68	59
9	Ali	81	76	76
10	Mark	66	70	64
11	Sharon	71	78	73
12	Mike	72	73	78
13	Jane	86	81	74
14	Michelle	92	93	83
15	Perry	69	75	77
16	Melissa	82	85	75
17	Kim	78	73	68
18	Luke	66	62	60
19	Louise	77	74	71
20				
21	**SAMPLE DATA STATISTICS**			
22	**Average**	73.9	74.8	70.7
23	**Mode**	66	73	#N/A
24	**Median**	72	74	72
25	**Range**	34	31	24
26	**Std Deviation**	9.0	7.7	7.1
27	**Correlation**		0.85	0.77
28				
29	**SAMPLE PREDICTIONS**			
30		50	57.2	56.0
31		60	64.6	62.1
32		70	71.9	68.3
33		80	79.3	74.4
34		90	86.6	80.5
35		100	93.9	86.7

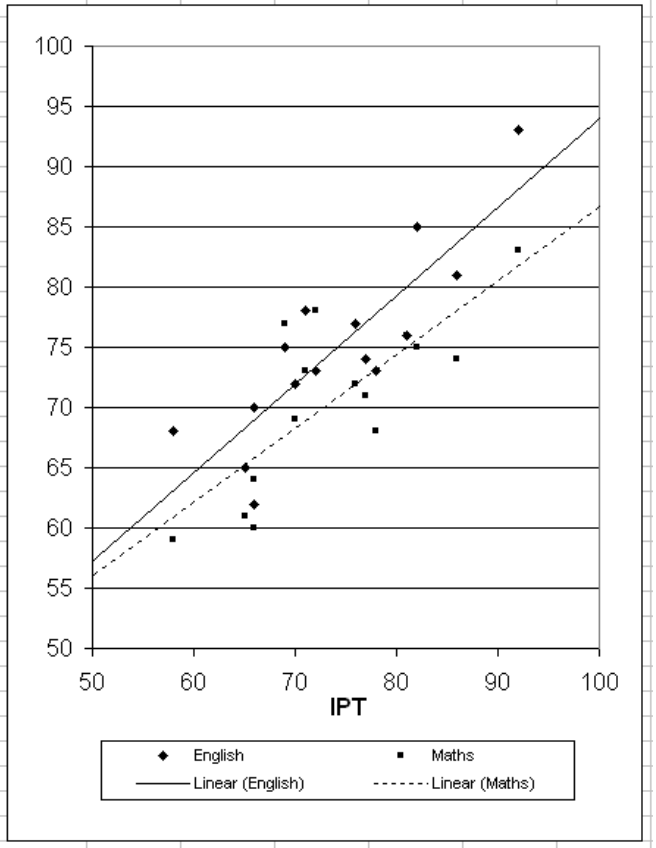


Fig 5.48
Fred's HSC IPT Predictor of HSC English and Maths Results.

GROUP TASK Discussion
Do you think Fred's theory is reasonable for both English and Maths? Discuss using evidence from his spreadsheet in *Fig 5.48*.

GROUP TASK Practical Activity
Reproduce Fred's spreadsheet. Note that the sample predictions are calculated using the TREND function in Excel. For example cell C30 contains the formula =TREND(C$5:C$19,$B$5:$B$19,$B30)

GROUP TASK Practical Activity
Collect HSC mark estimates for IPT, English and Maths from your classmates. Assess the validity of Fred's theory using this data.

HSC style question:

A school textbook supplier specialises in supplying complete sets of textbooks for each student in the schools they service. As part of this service they purchase all second-hand books from students and distribute them to the next intake of students. The second-hand books are collected at the end of each year from each student – at this time students are paid 40% of the current retail price.

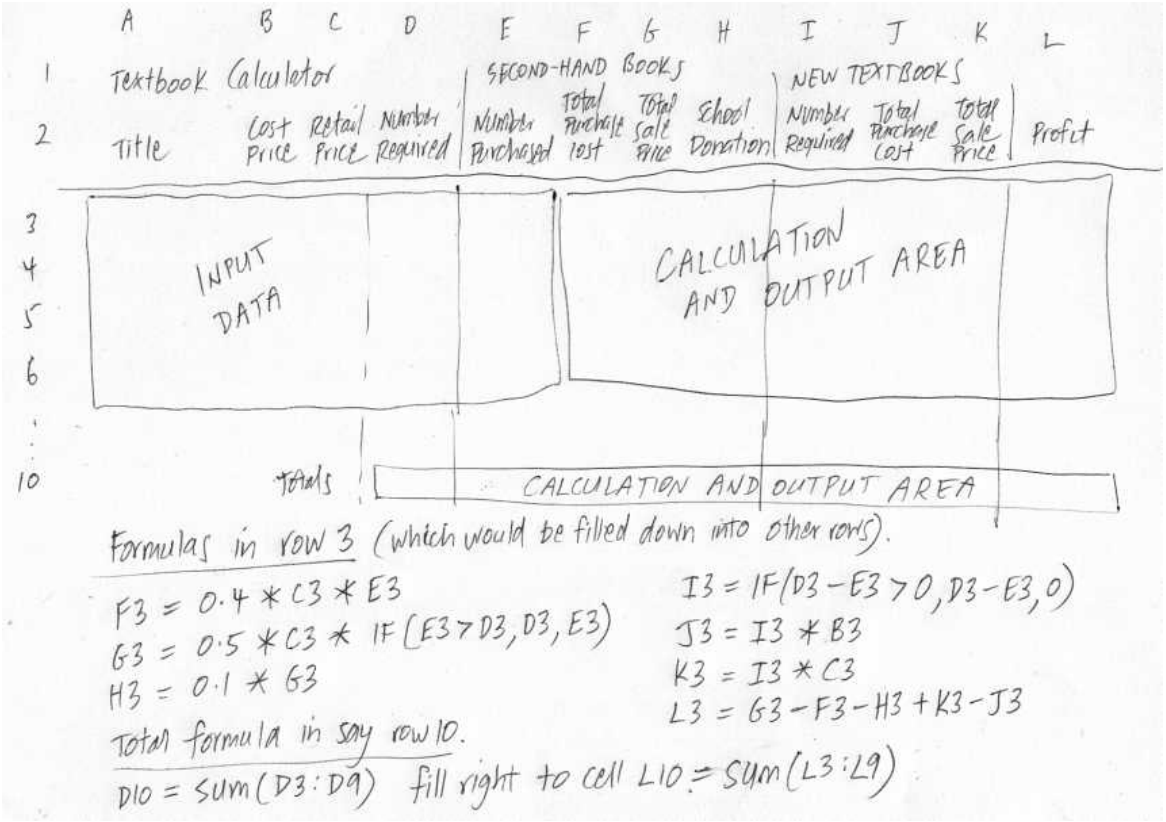
Prior to school commencing each year schools provide the textbook supplier with estimates of the total number of students completing each course. The textbook supplier needs to purchase sufficient new books to make up the shortfall between the second hand books and the estimated number of books required.

During the first week of the school year the book supplier attends the school to distribute the texts and collect payments. Students manually pick their required textbooks and pay the textbook supplier directly.

Students are charged 50% of the retail price for second hand books and full retail price for new textbooks. 10% of the total proceeds from second-hand book sales are donated to the school.

Your task is to develop a pen and paper spreadsheet model to assist the textbook supplier determine approximately how many new copies of each text they need to order and to estimate their likely profit. Your spreadsheet should be designed as a template that performs these processes for a single school. Include examples of each required formula on your spreadsheet model.

Suggested Solution



	A	B	C	D	E	F	G	H	I	J	K	L
1	**Textbook Calculator**				*Second Hand TextBooks*				*New Textbooks*			
2	**Title**	**Cost Price**	**Retail Price**	**Number Required**	**Number Purchased**	**Total Purchase Cost**	**Total Sale Price**	**School Donation**	**Number Required**	**Total Purchase Cost**	**Total Sale Price**	**Profit**
3	IPT The Prelim Course	$ 39.00	$ 48.00	24	9	$ 172.80	$ 216.00	$ 21.60	15	$ 585.00	$ 720.00	$ 156.60
4	IPT The HSC Course	$ 40.00	$ 50.00	18	14	$ 280.00	$ 350.00	$ 35.00	4	$ 160.00	$ 200.00	$ 75.00
5	SDD The Prelim Course	$ 39.00	$ 48.00	9	12	$ 230.40	$ 216.00	$ 21.60	0	$ -	$ -	-$ 36.00
6	SDD The HSC Course	$ 40.00	$ 50.00	10	7	$ 140.00	$ 175.00	$ 17.50	3	$ 120.00	$ 150.00	$ 47.50
7	Mathematics Prelim	$ 39.50	$ 45.00	76	33	$ 594.00	$ 742.50	$ 74.25	43	$ 1,698.50	$ 1,935.00	$ 310.75
8	Mathematics HSC	$ 41.00	$ 50.90	64	29	$ 590.44	$ 738.05	$ 73.81	35	$ 1,435.00	$ 1,781.50	$ 420.31
9	The Club	$ 17.00	$ 28.00	95	52	$ 582.40	$ 728.00	$ 72.80	43	$ 731.00	$ 1,204.00	$ 545.80
10	Macbeth	$ 18.00	$ 26.50	95	45	$ 477.00	$ 596.25	$ 59.63	50	$ 900.00	$ 1,325.00	$ 484.63
11												
12		**Totals**		**391**	**201**	**$ 3,067.04**	**$ 3,761.80**	**$ 376.18**	**193**	**$ 5,629.50**	**$ 7,315.50**	**$ 2,004.58**

Fig 5.49
Suggested solution to Textbook Supplier question implemented in Excel.

Comments

- On a trial or HSC examination this question would likely be worth approximately 8 to 10 marks.
- The suggested solution identifies a reasonable good set of inputs needed to perform the task. It is likely that further inputs would also be included such as ISBN, the year level, the course name and perhaps the author and the book's publisher.
- The input areas have been kept together within the suggested solution.
- Columns dealing with second books are grouped together, as are columns dealing with new textbooks. Other grouping schemes could also have been used such as grouping purchase columns together and grouping sales columns together.
- The IF formula in cell G3 of the suggested solution is needed to account for the possibility that more second hand textbooks are purchased than are actually required.
- The IF formula in cell I3 ensures that negative numbers of books will not be generated in column I. This could potentially occur when more second hand books have been purchased than are required for the next year's students.
- Correctly identifying the need for IF formulas and then implementing them correctly would likely be used by markers to distinguish between very good and excellent answers.
- Column I contains the number of new books the supplier needs to order. Column L together with its total contains the total profit figures. This is the only information required by the question. The remaining calculation columns are really intermediate calculations. It is possible to develop more complex formulas that do not require so many columns of intermediate formulas.

GROUP TASK Practical Activity
Implement the suggested solution to the above HSC style question within a spreadsheet. An example of a complete implementation is reproduced in *Fig 5.49* as an indication of typical data inputs.

SET 5C

1. Which type of chart is most appropriate for graphing yesterday's maximum sell prices for 10 different companies shares?
   (A) Pie chart
   (B) Line graph
   (C) XY graph
   (D) Column graph
2. The relative differences between quantities is most clearly highlighted on which type of graph?
   (A) Pie chart
   (B) Line graph
   (C) XY graph
   (D) Column graph
3. Investigating the relationship between a company's daily sales and daily costs would be best represented using which type of chart?
   (A) Pie chart
   (B) Line graph
   (C) XY graph
   (D) Column graph
4. Which of the following best describes the spreadsheet term "macro"?
   (A) A shortcut that executes a series of predefined commands.
   (B) A recorded sequence of keystrokes and mouse actions that can be replayed.
   (C) Visual Basic code that executes when a command button is clicked.
   (D) A formula whose results are determined only when the user presses the corresponding key combination.
5. A reusable spreadsheet that includes headings, titles, formatting, charts, formulas, macros, etc but no actual data is known as a:
   (A) worksheet
   (B) template
   (C) model
   (D) original file
6. Altering inputs and observing the effect on outputs is known as:
   (A) scenario management
   (B) goal seeking
   (C) what-if analysis
   (D) trendline analysis
7. The built-in goal seek function in a spreadsheet is only able to alter a single input to achieve its goal. Why is this?
   (A) Each goal (output) is determined by one and only one input.
   (B) There are potentially many different combinations of inputs that achieve the same goal.
   (C) There is insufficient demand from users for a more comprehensive goal seek function.
   (D) Generating values for many inputs is beyond the capabilities of current hardware and software.
8. The correlation between a set of predictions and their actual values is found to be 0.97. Which of the following is True?
   (A) The predictions are totally inaccurate.
   (B) The predictions are rather inaccurate.
   (C) The predictions are very accurate.
   (D) The predictions are totally accurate.
9. Measures of central tendency include:
   (A) mean, mode, median.
   (B) range, variance, standard deviation.
   (C) correlation, probability, confidence intervals
   (D) average, maximum, minimum.
10. The UAI Estimator's Reverse function described in the text is an example of:
    (A) spreadsheet analysis
    (B) what if analysis
    (C) statistical analysis
    (D) goal seeking
11. Compare and contrast each of the following:
    (a) Line graphs with XY graphs
    (b) Column graphs with line graphs
    (c) What-if analysis with goal seeking
12. Define each of the following terms and provide an example.
    (a) Spreadsheet macro (b) Spreadsheet template
13. Outline common statistical measures and explain how these measures can be used to make predictions based on historical data.
14. I have a theory that the success of the sport's team someone supports is an indicator of that person's ability to predict tomorrow's temperature.
    (a) Recommend suitable data, data sources and collection techniques for gathering data to test my theory.
    (b) Construct a pen and paper model of a spreadsheet suitable for analysing the test data in an attempt to confirm (or I suspect refute) my theory
15. Construct a spreadsheet that will graph functions of the form $y = Ax^3 + Bx^2 + Cx + D$.

EXPERT SYSTEMS

Expert systems are intelligent software applications that simulate the behaviour of human experts as they diagnose and solve problems. Expert systems are often described as being "goal oriented" – they operate best when they have one or more definite goals to pursue. The expert system can then formulate a logical sequence of questioning that most efficiently pursues these goals. Conclusions are made when a goal is achieved. For example the initial goal for doctors is to diagnose illness, they ask questions and perform tests in a logical fashion to achieve this goal. Achieving the goal results in a conclusion – a particular illness is diagnosed.

Conclusions that can be made by a human expert asking a logical sequence of questions over the telephone are well suited to expert systems. Human experts possess extensive knowledge and experience in a particular area. For example a motor mechanic who has been working in the field for many years is able to systematically and also intuitively diagnose problems with motor vehicles. Although formal training is often the basis of an expert's knowledge they also develop certain intuitive heuristics that they apply. In many instances the expert may not be able to explain precisely why they choose to explore a particular possibility – they just know with some degree of certainty that the chosen path of enquiry generally leads to a correct diagnosis or solution. For example an experienced doctor may know that it is more likely for infants presenting with a runny nose to then succumb to an ear infection. As a consequence the doctor more closely examines infant ear canals and is more likely to prescribe antibiotics to treat potential ear infections in infants. Expert systems allow the knowledge of human experts to be used repeatedly without the need or expense of the human expert being present.

HUMAN EXPERTS AND EXPERT SYSTEMS COMPARED

Let us consider the processes occurring as a human expert makes decisions and compare these processes to that used by a computerised expert system. The expert asks questions and the responses are used by the expert to determine the next question asked. Each response provides the expert with another fact they can use to assist their decision-making. In an expert system these facts are stored in a database of facts. The expert analyses the facts and determines the next question to ask. In expert systems the reasoning used to determine the next question is performed by the inference engine as it examines coded rules within a knowledge base.

Often a line of questioning will lead to a dead end. In this case the expert backtracks and commences another line of questioning. The next line of questioning can still use the known facts determined from previous responses. In an expert system the inference engine simulates the brain of the human expert – it decides on the most logical line of questioning to pursue, including backtracking and using existing facts.

Eventually the human expert reaches a conclusion – a decision is made or recommended and the goal is achieved. In some cases the conclusion is definite, but in many cases the conclusion is expressed as one or more likely possibilities. Each possibility is expressed with a certain level of confidence. For example a human expert may conclude, "I'm fairly certain that the problem is in the Widget module, however it could be an issue with the timing of the Woggle." The human expert determines their level of confidence in each conclusion during the question and response exchange. Conclusions emerge throughout the exchange with varying degrees of certainty. Those with low levels of certainty are ruled out completely, whilst those with high levels of certainty become recommendations or conclusions. Expert systems perform similar processes by assigning certainty or confidence values to possible conclusions. Each response causes one or more rules to be evaluated. Each

rule alters the confidence or certainty factor for one or more of the possible conclusions. The final conclusions are presented based on the final confidence or certainty values.

At the end of the question/answer exchange human experts are able to explain how they reached their conclusions by repeating the logic upon which each conclusion was based. Expert systems are also able to provide such explanations. This facility is known as the explanation mechanism. This mechanism essentially displays the facts compiled during the question/answer session together with the rules that where used as a consequence of each fact.

Consider the following "Extra Clothes" scenario:

The following scenario will be used throughout our discussion of expert systems:

We decide on what extra clothes to take with us each day based on what we perceive the most accurate weather forecast to be. We may consider professional forecasts, base our forecast on recent weather or we may simply look out the window. Probably a combination of these strategies is used. Based on our predicted forecast we decide to pack extra warm clothes and/or rain protection.

GROUP TASK Discussion

Identify the general goals or conclusions from the above decision support scenario. Now identify the inputs upon which these conclusions are based. Discuss possible rules that could process the inputs into conclusions.

STRUCTURE OF EXPERT SYSTEMS

There are five components of expert systems, the knowledge base, database of facts, inference engine, explanation mechanism and user interface. Expert system shells contain all five of these components. Particular expert systems are created by adding rules to the knowledge base and perhaps adding some facts to the database of facts. The expert system shell's inference engine uses this data as it executes. The context diagram in *Fig 5.50* models the flow of data between the user and these four components.

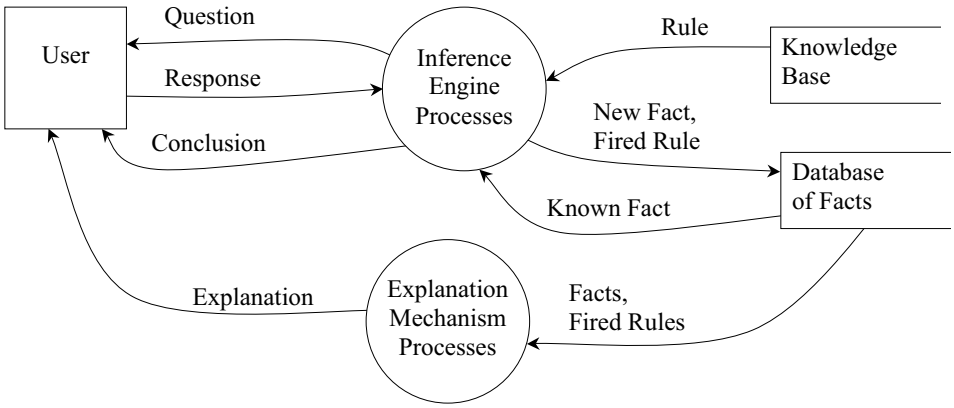


Fig 5.50
General context diagram for an expert system.

In this section we describe the first four of these components using examples from the Extra Clothes scenario. The user interface is included as needed during our discussion. We complete this section on Expert Systems with various points to consider when developing expert systems.

Knowledge Base

Rule (Expert System)
A single IF…THEN decision within an expert system's knowledge base.

The knowledge base is a data store that contains all the rules used by the inference engine to draw conclusions. Each rule is a simply an IF…THEN... statement. A condition that evaluates to be either true or false follows the IF. In expert systems this condition is known as a "premise". If the premise is found to be true then the statement (or statements) following the THEN are executed. Each statement following the THEN is known as a consequent. When the premise is found to be true the rule fires and all consequents in the rule are executed. In the Extra Clothes system an example rule could be "IF Rain is expected THEN Take an umbrella". The premise is "Rain is expected" and the rule has a single consequent "Take an umbrella". In its current form this rule cannot be directly entered into the knowledge base – it must be modified by the knowledge engineer to suit the required syntax that is understood by the inference engine.

Knowledge Engineer
A person who translates the knowledge of an expert into rules within a knowledge base.

When we develop our Extra Clothes expert system we will act as both the human expert and also the knowledge engineer. When developing real expert systems these people are different. The human expert explains their reasoning to the knowledge engineer. The knowledge engineer first translates the expert's reasoning into a series of "English like" IF…THEN… rules. There could well be hundreds or even thousands of such rules. The knowledge engineer then codes these into the syntax understood by the expert system shell. Different expert system shells use a different syntax and include different techniques for dealing with uncertainty.

- **Rules, attributes and facts**

In the Extra Clothes expert system the English like rule "IF Rain is expected THEN Take an umbrella" could be coded in the knowledge base as:

IF [ChanceOfRain] = "Expected"
THEN [RainGear] = "Umbrella"

This rule, together with details of a prompt (question) specification and goal are shown in *Fig 5.51* – this knowledge base operates in conjunction with expertise2go's e2gLite expert system shell. Two variables, known as attributes have been used – ChanceOfRain within the premise and RainGear within the consequent. In many expert systems attribute names are enclosed within square brackets. If the attribute ChanceOfRain holds the value "Expected" then the premise is true and the rule fires causing the attribute RainGear to be set to the value "Umbrella". All consequents set the value of an attribute. Assigning a value to an attribute establishes a fact – facts are stored in the database of facts. If the rule in our example has fired then the two facts [ChanceOfRain]="Expected" and [RainGear]="Umbrella" will be present within the database of facts.

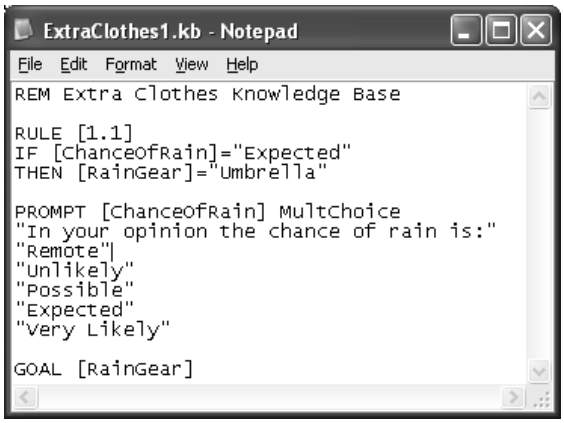


Fig 5.51
Initial Extra Clothes knowledge base for expertise2go's e2gLite expert system shell.

If a premise contains an attribute whose value is not yet known (no fact in regard to the attribute yet exists) then the inference engine can examine other rules whose consequent establishes a relevant factor it can ask the user for the value. Therefore both rules and questions establish facts. Once a fact exists for an attribute any future premise that includes that attribute can be automatically evaluated.

In the simple knowledge base in *Fig 5.51* the attribute ChanceOfRain can take a single value from the set of possible values "Remote", "Unlikely", "Possible", "Expected" and "Very Likely". In addition to rules, the knowledge base contains specifications of acceptable values for each attribute. *Fig 5.51* shows how such values are specified in knowledge bases for the e2gLite expert system shell.

If no fact already exists to determine the validity of the premise [ChanceOfRain] = "Expected" the inference engine may ask the user a question to determine a value for ChanceOfRain. In this case a multiple choice question would be asked – commonly radio buttons are used as shown in the expertise2go example in *Fig 5.52*. If the user selects "Expected" as their answer then the rule fires. Even if they choose one of the other options (except I don't know…) a fact in regard to ChanceOfRain is still established and stored in the database of facts.

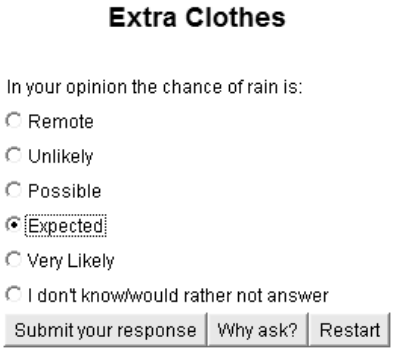


Fig 5.52
Multiple choice question displayed within expertise2go.

There are many other ways for the knowledge engineer to code each rule. We could have coded our example rule as:

IF [RainExpected] = TRUE THEN [TakeUmbrella] = TRUE, or as

IF [ForecastRainExpectation]>50% THEN [UmbrellaConfidence] = 40

In the first version two Boolean attributes, RainExpected and TakeUmbrella, are used. These attributes can hold values of either TRUE or FALSE. In the second version numeric attributes have been used. The attribute ForecastRainExpectation could store the probability of rain obtained from a professional weather forecast – perhaps via an online connection. Numeric attributes are used for continuous quantities such as temperature or length, and also for integral quantities such as the number of items, or age in years. In the second rule above, the attribute UmbrellaConfidence is a confidence variable used by the system to determine the degree of confidence that an umbrella should be taken.

GROUP TASK Discussion
Consider some possible rules for the Extra Clothes expert system. Identify the premise, consequent and also the attributes for each of these rules.

- **Dealing with uncertainty**

Confidence Variable
An attribute whose value is determined mathematically by combining its assigned values.

The use of confidence variables is one technique for dealing with uncertainty within expert systems; another common technique uses certainty factors. Let us consider each of these techniques.

Confidence variables operate differently to other attribute types, each time a value is assigned to a confidence variable it is mathematically combined with the existing value of the variable – commonly by simply adding the new value to the existing value. For instance say the following rules are in the knowledge base:

IF [ForecastRainExpectation]>50% THEN [UmbrellaConfidence] = 40

IF [OutsideView]="SomeClouds" THEN [UmbrellaConfidence]= 30

If the premise in both these rules has been evaluated as true, then the confidence variable UmbrellaConfidence would hold the value 70. *Fig 5.53* shows these rules within the logic block screen from ExSys© CORVID™ expert system shell. This user interface is used to enter rules into CORVID knowledge bases. We can have other rules that if true will reduce the value of this confidence variable – for instance if it hasn't rained for months we may wish to rule out taking an umbrella altogether. In this case our rule could include the consequent UmbrellaConfidence= –100 to effectively remove the possibility of recommending an umbrella.

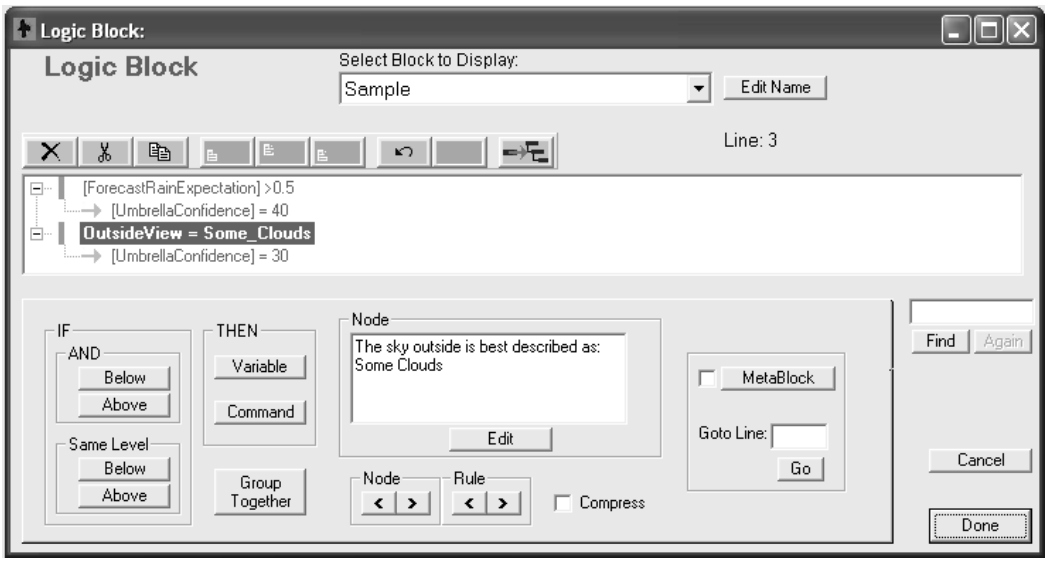


Fig 5.53
ExSys© CORVID™ expert system shell logic block user interface for entering rules.

Each confidence variable typically represents one of the possible conclusions the expert system will select from. Therefore all values assigned to all confidence variables should be scaled similarly so that comparisons of their final values are legitimate – an important consideration when developing rules that use confidence variables. Commonly confidence variables are assigned values such that higher final values correspond to higher levels of certainty in that conclusion. Unlike other variable types, confidence variables are rarely used within the premise of a rule. This is because their value is not set permanently and hence does not establish a definite fact, rather the value changes as new rules fire.

Certainty Factor
A value, usually in the range 0 to 1, which describes the level of certainty in a fact or conclusion.

Certainty factors describe the perceived probability or more accurately the level of certainty that a fact or a consequent is correct. Certainty factors are specified directly as part of each consequent and they can also be entered by the user as they answer questions. When users enter a value for a certainty factor they are indicating their level of certainty that their response is correct. The knowledge base includes a threshold value used to determine the level of certainty required for rules to fire. For example say a user answers a question and indicates they are 70% certain their answer is correct then the associated rule will only fire if the premise is true and the threshold value is less than 70%. Even when the rule does not fire the user's answer together with the certainty factor entered is stored as a fact. Like probabilities,

certainty factors are generally expressed on a scale from 0 to 1 or as percentages from 0% to 100%. 0% implies complete uncertainty meaning it is considered to be definitely false, 100% means it is considered completely true and values between 0 and 100% indicate varying degrees of certainty in the correctness of the fact or conclusion.

Heuristic
A rule of thumb considered true, usually with an attached probability or level of certainty.

Confidence variables and certainty factors allow decisions to be made that are not definitely true or definitely false. Certainty factors are assigned values based on the expert's experience of what usually occurs or the user's confidence in their answer. Such rules are known as "rules of thumb" or heuristics. Heuristics allow expert systems to reach likely conclusions rather than definite conclusions. This is an example of "fuzzy logic" where results are not simply correct or incorrect, rather one result can be a bit correct, another maybe kind of correct and others can lie anywhere on the continuum between true and false. To implement fuzzy logic expert systems commonly allow a single attribute to take multiple values at different levels of certainty. For example, the single attribute ChanceOfRain may hold the value "Expected" with 70% confidence and also hold the value "Possible" with 80% confidence. Each of these facts may cause different rules to fire that in turn cause the system to reach different conclusions with different levels of confidence. When there are many attributes that each hold many values at many different levels of confidence the inference engine processing becomes complex as each combination of possible values and confidence levels is used in an attempt to fire rules.

Consider the following:

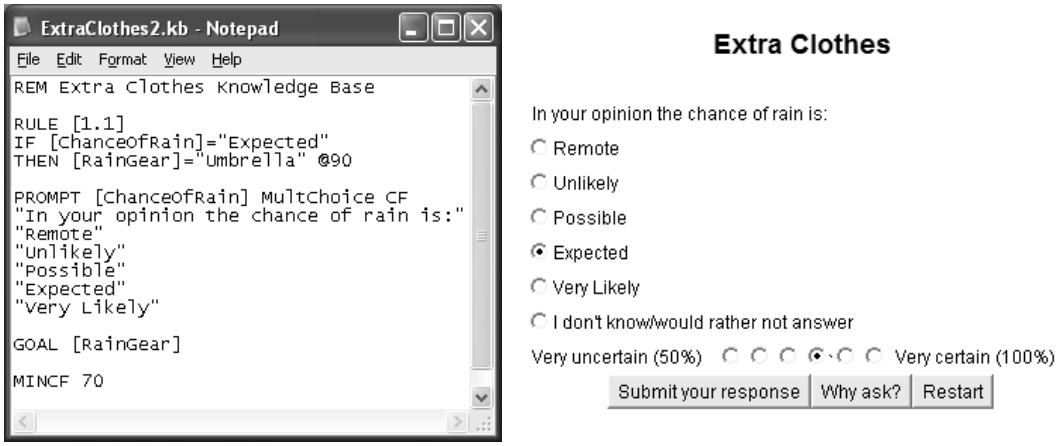


Fig 5.54
Initial Extra Clothes knowledge base and question with certainty factors added.

In *Fig 5.54* three additions have been made to the initial knowledge base from *Fig 5.51* to implement certainty factors. In the knowledge base in *Fig 5.54* above a certainty factor for the consequent of 90% has been added, CF has been added to the PROMPT statement and a minimum CF threshold of 70% has been specified. CF is a common abbreviation used in many expert systems to specify confidence factors and in this knowledge base MINCF specifies the minimum confidence factor value required for rules to fire. When the expert system is executed the question shown at right in *Fig 5.54* is displayed. If the user answers the question as indicated the system concludes that RainGear should be an Umbrella with 72% confidence.

GROUP TASK Discussion
Explain how the system has calculated the result with 72% confidence. Discuss what would occur if the user answered at each of the other levels of confidence.

GROUP TASK Discussion
Options for users to specify values for confidence factors range from 50% up to 100%. Why don't the options range from 0% to 100%? Discuss.

ExtraClothes3.kb - Notepad

File Edit Format View Help

```
REM Extra Clothes Knowledge Base

RULE [1.1]
IF [ChanceOfRain]="Expected" AND
   [DaysSinceLastRain]<25
THEN [RainGear]="Umbrella" @90

PROMPT [ChanceOfRain] MultChoice CF
"In your opinion the chance of rain is:"
"Remote"
"Unlikely"
"Possible"
"Expected"
"Very Likely"

PROMPT [DaysSinceLastRain] Numeric CF
"How many days since rain last fell:"
"0"
"1000"

GOAL [RainGear]

MINCF 60
```

ExtraClothes4.kb - Notepad

File Edit Format View Help

```
REM Extra Clothes Knowledge Base

RULE [1.1]
IF [ChanceOfRain]="Expected" OR
   [DaysSinceLastRain]<25
THEN [RainGear]="Umbrella" @90

PROMPT [ChanceOfRain] MultChoice CF
"In your opinion the chance of rain is:"
"Remote"
"Unlikely"
"Possible"
"Expected"
"Very Likely"

PROMPT [DaysSinceLastRain] Numeric CF
"How many days since rain last fell:"
"0"
"1000"

GOAL [RainGear]

MINCF 60
```

Fig 5.55
Edited versions of initial Extra Clothes knowledge base

The initial rule has been edited to include the numeric attribute DaysSinceLastRain. In version 3 (left in *Fig 5.55*) the premise contains the logical AND operator and in version 4 (at right in *Fig 5.55*) the logical OR operator is used.

When the expert system is executed the following observations are made:

- In version 3 if "Expected" is entered for ChanceOfRain with 90% confidence and 25 is entered for DaysSinceLastRain with 80% confidence the conclusion recommends an Umbrella with 64.8% confidence.
- In version 3 if "Expected" is entered for ChanceOfRain with 80% confidence and 25 is entered for DaysSinceLastRain with 80% confidence then no conclusion is possible.
- In version 4 if "Expected" is entered for ChanceOfRain with 50% confidence and 25 is entered for DaysSinceLastRain with 70% confidence the conclusion recommends an Umbrella with 63% confidence.
- In version 3 both questions are always asked, whilst in version 4 often just one question is asked.

GROUP TASK Discussion
Explain why each of the above observations occurs. Describe example inputs that the system will be unable to process into conclusions.

GROUP TASK Practical Activity
Create a each of the four versions of the knowledge base from *Fig 5.51*, *5.54* and *5.55*. Do you get the same results as those described in the text?

Database of Facts

As the name implies, the database of facts contains all the known facts accumulated during the current session. However it also includes any facts known prior to execution. In many expert systems a series of previously known facts is added or imported into the database of facts prior to the inference engine commencing its work. These facts could be from a linked database, spreadsheet or some other data source. Clearly this means the user will not need to answer questions about attributes for which such facts already exist. For example expert systems that recommend products often import facts that apply to each product. In our example Extra Clothes system an online connection to the weather bureau could be used to determine initial facts in regard to professional forecast attributes.

The database of facts also stores a detailed history of which rules have fired and in which order they fired. This information together with the facts is used by the explanation mechanism to justify conclusions the system makes. Furthermore the ability to view the specific sequence of rules that fired is of great assistance when knowledge engineers are debugging the knowledge base.

In reality, for all but the largest systems the database of facts is maintained within RAM during processing. If the user wishes to halt execution then the database of facts must be saved so the session can be continued at a later time. Some web based systems store the database of facts as a "cookie" on the users machine. In large systems the database may well be an actual database stored on secondary storage.

GROUP TASK Discussion
Describe the essential differences between the knowledge base and the database of facts. Why not simply store facts within the knowledge base?

Inference Engine

The inference engine is the brain of the expert system; its processes simulate the reasoning of a human expert. The aim of the inference engine is to reach conclusions that satisfy the goal or goals of the expert system. It logically applies the rules and facts to efficiently reach conclusions that meet these goals.

There are two fundamental strategies used by inference engines – backward chaining and forward chaining. These strategies determine the order in which rules are tested. We shall describe examples of both these strategies using the following version of our Extra Clothes knowledge base.

Consider the following decision tree:

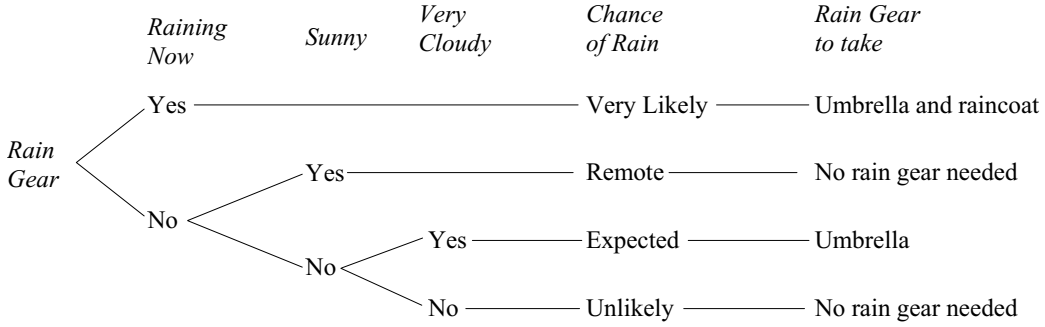


Fig 5.56
Decision tree for sample Extra Clothes expert system.

The decision tree in *Fig 5.56* has been implemented as a knowledge base (see *Fig 5.57*) for use with the e2gLite expert system shell. The premise for rule 1.1 means the rule will fire if the variable ChanceOfRain has either of the values "Remote" or "Unlikely".

This version of the knowledge base does not ask the user for the ChanceOfRain, as this is a somewhat subjective question. Instead the user is asked more objective questions, namely "Is it raining outside now?", "Is it sunny outside now?" and "Is it very cloudy outside now?" – answers to one or more of these questions is used to determine a value for ChanceOfRain.

GROUP TASK
Discussion

Compare the decision tree in *Fig 5.56* with the knowledge base in *Fig 5.57*.

There are three PROMPTs in the knowledge base. Is there a logical order in which these questions should be asked? Discuss.

It is preferable to ask users objective questions. Why is this? Discuss.

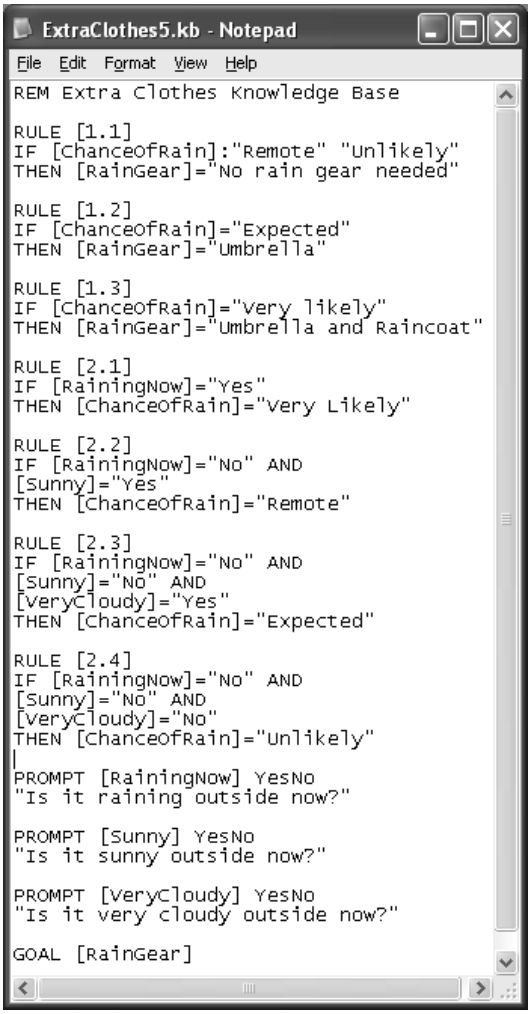


Fig 5.57
Sample Extra Clothes knowledge base.

- **Backward Chaining**

Backward chaining is what causes expert systems to ask questions in an order that gathers more and more detailed information to achieve goals. This behaviour closely reflects the questioning performed by human experts – they pursue a line of questioning that is focused on a particular goal. Questions that are irrelevant to the current goal are not asked and questions of relevance to the current goal are asked in a logical order. Backward chaining is known as a goal driven strategy, essentially the inference engine only considers rules whose consequent will set a value for the current goal attribute.

During backward chaining the inference engine maintains a goal list (also known as a goal stack). The lowest goal in the list is the overall goal of the system – in the knowledge base in *Fig 5.57* determining a value for RainGear attribute is the overall goal. As backward chaining progresses sub-goals are added to and removed from the top of the goal list. The inference engine is always trying to determine a value for the goal attribute at the top of the goal list. If a fact is determined (or already exists) that achieves the top goal then that goal is removed from the goal list and the next goal in the list becomes the new aim of the inference engine. Goals are also removed from the goal list if the inference engine cannot determine a value for the goal attribute.

To achieve the top goal in the goal list the inference engine first looks in the database of facts to see if a value for the goal attribute is already known – if a fact already exists then the goal is achieved and is removed from the top of the goal list. If no such fact exists it then looks for rules that set a value to this goal variable within their consequent. If all such rules fail to set a value for the goal variable (establish a fact) the inference engine will then ask the user. If the user is unable to answer (or asking the user is not an option) then the goal cannot be achieved and is removed from the goal list. If one of the relevant rules fires or the user answers then the goal is achieved, a fact is added to the database of facts and the goal is removed from the top of the goal list. Note that this strategy means the user will never need to answer the same question twice.

In our *Fig 5.57* knowledge base the overall goal is to determine a value for RainGear. Let us work through an example session from the point of view of the inference engine – *Fig 5.58* describes the changing state of the goal list. Initially the goal list contains just the overall goal to determine a value for RainGear (Goal list 1 in *Fig 5.58*) and initially the database of facts is empty. We examine the rules and find the consequent of Rule 1.1 assigns a value to our overall goal RainGear. For this rule to be evaluated (and hopefully fire) requires a value for ChanceOfRain, hence ChanceOfRain is added to the top of the goal list(Goal list 2 in *Fig 5.58*). The new goal of the inference engine is to determine a value for ChanceOfRain. The inference engine first looks in the database of facts to see if it already has a value for ChanceOfRain – currently no such fact is present. It now looks for rules that include ChanceOfRain in their consequent. Rule 2.1 is one such rule, however for this rule to fire we need a value for RainingNow. Therefore RainingNow is added to the top of the goal list (Goal list 3 in *Fig 5.58*) and becomes the new goal of the inference engine. Significantly the inference engine remembers where it was up to when attempting to achieve the goal ChanceOfRain – when ChanceOfRain later becomes the current goal once more processing will proceed from this point.

Goal List 1 Determine value for:
RainGear

Goal List 2 Determine value for:
ChanceOfRain RainGear

Goal List 3 Determine value for:
RainingNow ChanceOfRain RainGear

Goal List 4 Determine value for:
ChanceOfRain RainGear

Goal List 5 Determine value for:
Sunny ChanceOfRain RainGear

Goal List 6 Determine value for:
ChanceOfRain RainGear

Goal List 7 Determine value for:
RainGear

Fig 5.58
Goal lists for example Extra Clothes backward chaining example.

Our current top goal in Goal List 3 of *Fig 5.58* is to determine a value for the attribute RainingNow. There are no facts and no rules that can be used, therefore the inference engine asks the user. Let's assume the user answers "No" to the question "Is it raining outside now?" This answer establishes the fact RainingNow="No", which is stored in the database of facts. Our goal to determine a value for RainingNow is achieved, so this goal is removed from the top of the goal list.

We are back to determining a value for ChanceOfRain as our goal (Goal list 4 in *Fig 5.58*). Previously, processing of this goal was considering Rule 2.1, however this rule fails to fire as the premise [RainingNow]="Yes" is found to be false. We now consider Rule 2.2 – the consequent of this rule also sets a value for ChanceOfRain. To evaluate the premise of Rule 2.2 requires values for RainingNow and for Sunny.

We have a fact that states RainingNow="No" so that part of the premise is true. Determining a value for Sunny is added to the top of the goal list (Goal list 5 in *Fig 5.58*) and becomes the current goal. No facts or rules exist to achieve this goal so the user is asked "Is it sunny outside?" – we'll assume the user answers "Yes" to this question. The fact Sunny="Yes" is stored in the database of facts, hence the Sunny goal is achieved and is removed from the goal list.

We return once more to the ChanceOfRain goal (Goal list 6 in *Fig 5.58*) where we last left it evaluating the second part of the premise of Rule 2.2. As Sunny="Yes" is now a known fact we find the whole premise of Rule 2.2 is true, hence the rule fires causing the consequent to be executed. This establishes and stores the fact ChanceOfRain="Remote". The ChanceOfRain goal is achieved and subsequently removed from the goal list.

Our goal list now contains just our overall goal to determine a value for RainGear (Goal list 7 in *Fig 5.58*). Recall that we left this goal at the point where it was processing Rule 1.1. We now have the fact that ChanceOfRain="Remote" so the premise of Rule 1.1 is true. The rule fires causing RainGear to be set to "No rain gear needed". This fact finally achieves our overall goal and is displayed to the user.

Notice that there was no need to ever determine a value for the attribute VeryCloudy during our sample session. This demonstrates a significant characteristic of backward chaining compared to forward chaining – only those questions directly required to reach a conclusion that achieves the goal are asked.

GROUP TASK Discussion

Consider the Extra Clothes knowledge base in *Fig 5.57*. Using a backward chaining strategy, describe the inference engine processing occurring using different user inputs to those described in the above discussion.

- **Forward Chaining**

Forward chaining starts with facts (what is known) and uses this data to reach conclusions. Forward chaining is often referred to as a data driven strategy – data is supplied in the form of facts without any specific goal being specified. The inference engine attempts to fire each rule in turn using the known facts. Each rule that fires creates new facts and these facts are then available when evaluating subsequent rules. Although goals are achieved using forward chaining this is not the inference engine's focus like it is when backward chaining.

Many expert systems, when forward chaining, work sequentially through all the rules repeatedly so that new facts determined by later rules can be used to evaluate earlier rules on future passes through the knowledge base. Other expert systems are set so they will stop and ask the user for values each time a rule's premise cannot be evaluated using the available facts. This can result in questions being asked that could have been inferred by later rules within the knowledge base – the order in which rules appear in the knowledge base becomes significant. In general backward chaining is used for interactive sessions whilst forward chaining is used when facts are known in advance. Forward chaining is recommended for expert system's that import data into their database of facts prior to the inference engine commencing.

In reality a combination of backward and forward chaining is often used. Existing known facts are forward chained to infer new facts, whilst backward chaining is used to interactively infer facts in conjunction with user inputs. Forward chaining existing facts first often minimises the number of questions users need to answer. Backward chaining uses facts determined by forward chaining and vice versa. For example expert systems are used to suggest products based on customer's requirements. The

data that describes each individual product is stored in an attached database – the data in this database can be thought of as an extension of the database of facts. Backward chaining determines the customer's requirements whilst forward chaining is used to suggest products. Such systems can forward then backward chain or vice versa.

Forward chaining is a far simpler strategy to understand compared to backward chaining. The rules within the knowledge base are simply tested in the order in which they occur within the knowledge base. If a rule doesn't fire it is discarded and the inference engine simply moves onto the next rule. If a rule does fire then the consequents are executed and the resulting facts are stored in the database of facts.

Consider the processing performed using a forward chaining strategy with the knowledge base in *Fig 5.57* above. We will assume the inference engine first asks each question specified by a PROMPT statement and then forward chains to reach a conclusion. Say the user answers the questions as indicated in *Fig 5.59*. The database of facts now contains RainingNow="No", Sunny="No" and VeryCloudy="Yes".

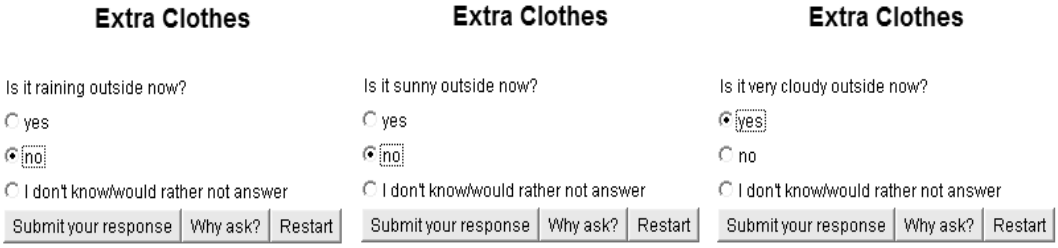


Fig 5.59
Sample user interface and responses prior to forward chaining commencing.

Forward chaining now commences by examining each rule in the *Fig 5.57* knowledge base in turn. Rules 1.1, 1.2 and 1.3 cannot be evaluated and so they discarded. The premise for Rule 2.1 is false and so to is the premise for Rule 2.2 – neither rule fires. The premise of Rule 2.3 is true so the rule fires and ChanceOfRain="Expected" is added to the database of facts. Rule 2.4 does not fire. We have now reached the end of the rules, we need to repeat if we are to use our new inferred fact to determine a value for RainGear. Commencing at Rule 1.1 again we work through all the rules in sequence. Rule 1.2 fires causing RainGear="Umbrella" to be stored in the database of facts. Rule 2.3 will also fire which provides no new information and simply reasserts the existing fact ChanceOfRain="Expected". We reached the conclusion, namely that we should take an umbrella, but the inference engine does not stop once this goal is achieved, rather it continues until it is unable to generate any new facts.

In our rather simple Extra Clothes example we had just one goal, in many systems there are many varied goals. Forward chaining continues attempting to fire rules and produce new facts until it finds no more new facts. The inference engine does not search out particular goals; rather forward chaining produces facts that the user interprets as the conclusions that achieve goals.

GROUP TASK Discussion
Using a forward chaining strategy and the knowledge base in *Fig 5.57*, describe the inference engine processing occurring using different user inputs to those described in the above discussion.

GROUP TASK Discussion
In the above discussion it was necessary to work through the rules twice. Why was this necessary? Suggest changes to the knowledge base in *Fig 5.57* such that only a single examination of the rules would be needed.

Explanation Mechanism

Expert systems are able to explain how they reached conclusions. Essentially the explanation is a replay of the inferences made by the inference engine. Inferences occur every time a rule fires and new facts are established. This information is contained within the database of facts, so the input to the explanation mechanism is simply the contents of the database of facts – refer to the context diagram in *Fig 5.50*.

Simply displaying each rule that fired to assert each fact is not very user friendly. An example of a standard explanation provided by e2gLite is reproduced in *Fig 5.60*. This is a rather technical explanation of the operations performed by the inference engine and is really unsuitable for display to users. In real expert systems text is included within the knowledge base to explain the purpose of each rule and consequent. The explanation mechanism is therefore able to generate explanations in plain English, such as the example Camcorder recommendation in *Fig 5.60*.

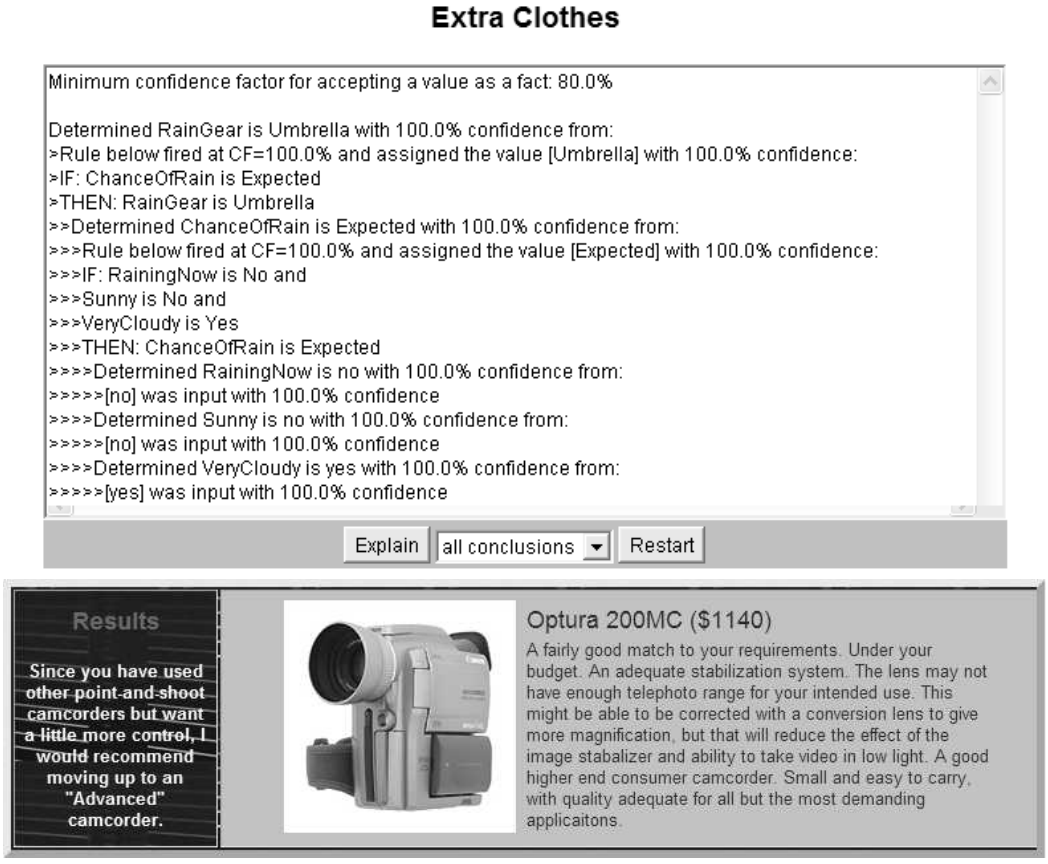


Fig 5.60
Examples of a technical explanation (top) generated by e2gLite and a user friendly CamCorder explanation (bottom) generated by CORVID™.

GROUP TASK Discussion
Consider the Extra Clothes explanation in *Fig 5.60*. Discuss more appropriate wording that could be used for the explanation.

GROUP TASK Discussion
Analyse the Camcorder explanation in *Fig 5.60*. Discuss likely rules and facts that would have contributed to the generation of this explanation.

DEVELOPING EXPERT SYSTEMS (KNOWLEDGE ENGINEERING)

In this section we detail common tasks performed during the development of an expert system. These tasks are in addition to the general project management and development tasks undertaken when constructing information systems of all types.

Understanding the Problem and Planning

The aim is to establish the goals of the proposed expert system and decide whether it is possible and worthwhile developing the system. The following points are of particular significance when developing expert systems.

- **Identify the precise goals of the system.**

What is the problem that needs to be solved and when the problem has been solved what conclusions are made? Answers to such questions are needed to establish the goals of the expert system. As expert systems are goal oriented this is a vital step. If it is difficult or not possible to precisely define the goals of the system then probably an expert system is the wrong type of decision support system to use for this problem. There should be a finite set of possible conclusions or recommendations for each goal. In practice each goal attribute should have a predefined number of possible values – each value corresponds to a particular conclusion or recommendation.

In our Extra Clothes expert system we identify two goals – determining the rain gear to take and determining which warm clothes to take. Possible conclusions for our rain gear goal are to take an umbrella, take both an umbrella and a raincoat or take no rain gear. For our warm clothes goal the conclusions could be to take a jumper, take a jacket, take both jumper and jacket or take no extra warm clothes.

- **Ensure human experts can solve the problem and are available.**

Expert systems are not able to solve problems that human experts cannot solve. If no human experts are able to solve the problem then an expert system cannot be created. The aim is to create a system that reaches the same conclusions as a suitably qualified expert, therefore the human expert or experts must be available as the rules are developed.

In our Extra Clothes example we are acting as the human expert as well as the knowledge engineer. Most people are able to make reasonable decisions about suitable warm cloths and rain gear to take each day. Therefore we, as the human expert, can solve the problem and furthermore we are available! For real world problems it is often clear that the human expert can solve the problem, as this is often a substantial part of their job. However such experts are often busy people and making themselves available to the knowledge engineer is often difficult. To further complicate matters the best human experts are usually the busiest. If experts are not busy then it is worth considering whether there will be a market for the expert system.

- **Observe and analyse examples of human expert consultations**

Firstly, access to consultations must be possible – this is often difficult when a consultation involves disclosure of sensitive private information such as during medical diagnosis sessions. Observing human expert consultations confirms that human experts can actually solve the problem. Many such consultations should be observed so that the goals of the expert system are confirmed and the ability of these goals to be achieved can be reliably assessed. Is it worthwhile developing an expert system that will only reach a conclusion in a small number of cases?

Expert systems are best suited to problems that human experts solve using essentially verbal information. If human experts make extensive use of non-verbal cues, such as eye contact, touch, voice inflexion and posture, then this will be difficult or impossible to simulate using an expert system. Expert systems are simpler to develop

when questions and responses can easily be translated into text. Image and video data is possible, however its use requires complex technical analysis techniques that can add substantial time and cost to the system's development and ongoing maintenance.

Observing many human expert consultations helps establish common heuristics used to solve the problem. It is often useful to record audio or video footage of consultations in addition to observing live consultations. Furthermore taped consultations allow the knowledge engineer to analyse the interactions more closely as they design rules.

Designing Solutions

With regard to expert systems, designing the solution is primarily about creating the knowledge base of rules. This is the essential task performed by the knowledge engineer. In general, the best approach is to start with the overall goals and work to progressively add more detailed rules. Eventually the detailed rules will include attributes whose values can be established by asking the user questions. This design technique reflects the backward chaining strategy used when the system is executed. We focus on the top-level goals, develop more detail in the form of rules that achieve these goals, we then focus on the sub-goals of our new rules to design more detailed rules. This process continues until we reach a point where the users are able to objectively provide responses. This process is commonly known as top-down design.

A results and explanation display that simply shows the facts and rules is useful for testing the system during the design of the knowledge base. However, once the knowledge base is completed the format of the results and explanation displays can be specified so the display is more user friendly for the system's users.

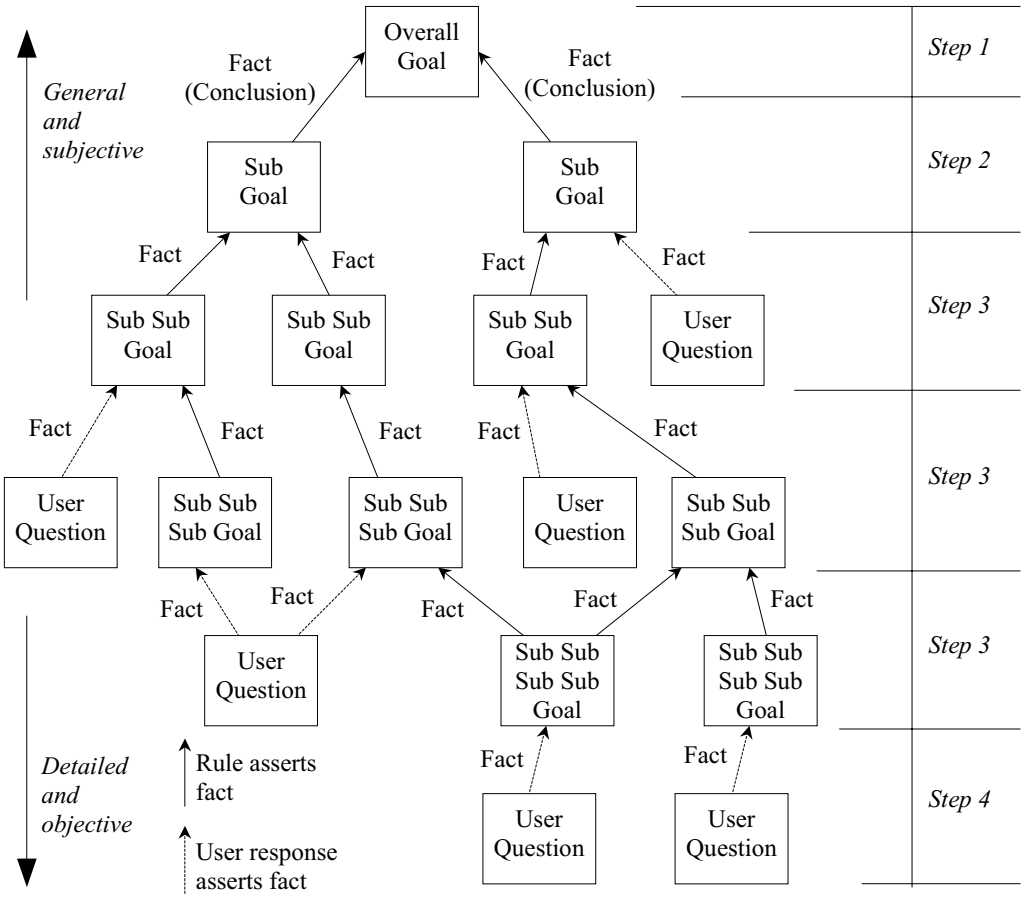


Fig 5.61
Overview of a recommended strategy for designing a knowledge base.

A general overview of a recommended design process for developing a knowledge base is modelled above in *Fig 5.61*. This model should be read in conjunction with the sequence of recommended steps and tasks that follows:

1. Assign attribute names to each goal and values to represent each conclusion

Each goal is represented by an attribute and the goal is achieved when the attribute has been assigned a value. Assigning a value to an attribute establishes a fact. For the overall goals each fact is really a conclusion that forms part of the displayed results.

In our Extra Clothes example we use the attribute names RainGear and WarmClothes to represent our two goals of determining the rain gear to take and determining which warm clothes to take. There are three possible conclusions determined for our RainGear goal, either take an Umbrella, take both an umbrella and a raincoat or take no rain gear. We represent each conclusion by specifying possible values for the attribute RainGear, namely "Umbrella", "Umbrella and Raincoat" and "No rain gear needed". Similarly we specify possible values for our WarmClothes attribute of "Jumper", "Jacket", "Jumper and jacket" and "No warm clothes needed".

In our example Extra Clothes system both our goal attributes have a text data type that is restricted to a list of particular values. It is possible for goal attributes to be numeric types, Boolean types or even confidence variables. Recommending the number of items to purchase in an estimating expert system would require a numeric attribute, whilst recommending whether a purchase should be made could be represented as a Boolean attribute.

Confidence variables do not require lists of values rather the confidence variable itself represents the level of confidence in a single conclusion. In our Extra Clothes example we could use three confidence variables to represent each of our rain gear conclusions – say with attribute names UmbrellaConf, BothRainGearConf and NoRainGearConf. The confidence variable with the highest final value is recommended above those with lower final values.

2. Design rules with consequents that assign values to goal attributes

Based on observation and consultation with the human expert, the knowledge engineer produces a series of high level rules that each result in one or more of the conclusions. This means each consequent will assign a value to one of the goal attributes. If the human expert is not 100% confident about a consequent then certainty factors should be included as part of the consequent.

These high level rules are based on rules of thumb used by the human expert as they make decisions. For example, in our Extra Clothes example we, as the expert, think it is best to take both an umbrella and a raincoat if we feel rain is very likely. This rule of thumb is coded in the knowledge base as the rule:

IF [ChanceOfRain]="Very likely" THEN [RainGear]="Umbrella and Raincoat"

Notice that the consequent is one of our conclusions. Also note that the premise includes new attributes that must be defined – including their possible values. These new attributes become sub-goals during backward chaining.

In our example asking the user what they think is the chance of rain is quite a subjective question – different users will no doubt supply different answers based on their own experience and how they value the available evidence. In the final expert system we aim to only ask users objective questions, questions where the majority of people given the same evidence will provide the same response. Subjective issues should be dealt with using further rules that determine values based on the knowledge of the human expert.

3. Design further rules with consequents that assign values to sub-goal attributes

Attributes within the premise of each rule developed in the previous step become our new goals. We then develop further rules whose consequents assign values to these attributes to achieve each of our new goals. Again if the expert is not 100% certain then certainty factors should be included. In our Extra Clothes example we, as the expert, decide that rain is very likely if it is currently raining. This rule of thumb is added to the knowledge base as:

IF [RainingNow]="Yes" THEN [ChanceOfRain]="Very likely"

Now consider whether it is appropriate to ask the user a question to determine a value for each new attribute. In the above example rule asking the user "Is it raining outside now?" is an objective question – presumably all users will answer the same way given the same evidence. Answers to objective questions are not affected by the users personal emotions or bias, rather the answers are based on something concrete, known or observable. Once such objectivity is achieved we can create a question for the attribute and there is no need to develop further rules to achieve that sub-goal.

4. Repeat step 3 for all attributes where objective questions cannot be asked

If there are attributes where objective questions cannot be asked then step 3 needs to be repeated – perhaps numerous times. Further rules are developed until objective questions can be asked. Note that the number of rules added will likely increase each time step 3 is completed until objective questions begin to emerge.

In some cases the nature of the problem requires that some subjective questions are appropriate or even necessary. Or it maybe that the level of detail required to achieve such objectivity is unwarranted or it is not possible to totally remove all subjectivity from questions. Attributes with these characteristics should be assigned certainty factors so that the user can indicate their level of confidence in their responses.

The knowledge base is complete once facts required to fire all rules can be determined either using questions or as a result of another rule firing. This does not necessarily mean that all sets of user responses will result in a conclusion, it is often appropriate for some combinations of answers to fail to reach a conclusion – as occurs during consultations with real human experts.

Consider the following general types of Expert Systems

- Software and hardware trouble-shooters and wizards.
- Product recommendation systems on retailer's web sites and on information kiosks in retail stores.
- Travel planners that suggest times and routes or recommend combinations of travel destinations, flights and accommodation.
- Medical expert systems to assist doctors diagnose disease and/or prescribe suitable medication.
- Interactive voice response (IVR) expert systems that collect answers over the telephone and make automated recommendations for products, troubleshooting and other types of information.

GROUP TASK Research

Research and identify at least one specific example of each of the above types of expert system. For each specific example determine the goals of the system and list an example of a conclusion that achieves each goal.

HSC style question:

Consider the following partially developed knowledge base included in an expert system used to determine whether a security license should be issued to an applicant. The expert system is used during interviews between a clerk and an applicant.

1	Security License issued	IF	Correct age AND Conditions met
2	Conditions met	IF	Good references AND Australian resident
3	Australian resident	IF	Valid document sighted
4	Valid document sighted	IF	Australian birth certificate OR Australian citizenship OR Evidence of Resident Status
5	Good references	IF	2 Authorised referees AND < 12 months old
6	Authorised referees	IF	Supplied by doctor, teacher , JP or religious leader
7	Correct age	IF	>= 18

(a) Construct an equivalent decision tree for the logic in this system.

(b) We now wish to modify the above rules to include information relating to the completion of required training.

- Training must be offered by an organisation which is a Registered Training Organisation (RTO) approved by the Commissioner of Police. Such organisations have a current RTO Master License number.
- Once an applicant has attended a suitable training course, they must sit a test. Their result must be 100% in the section on Relevant Law, and greater than 50% overall in the remaining sections.

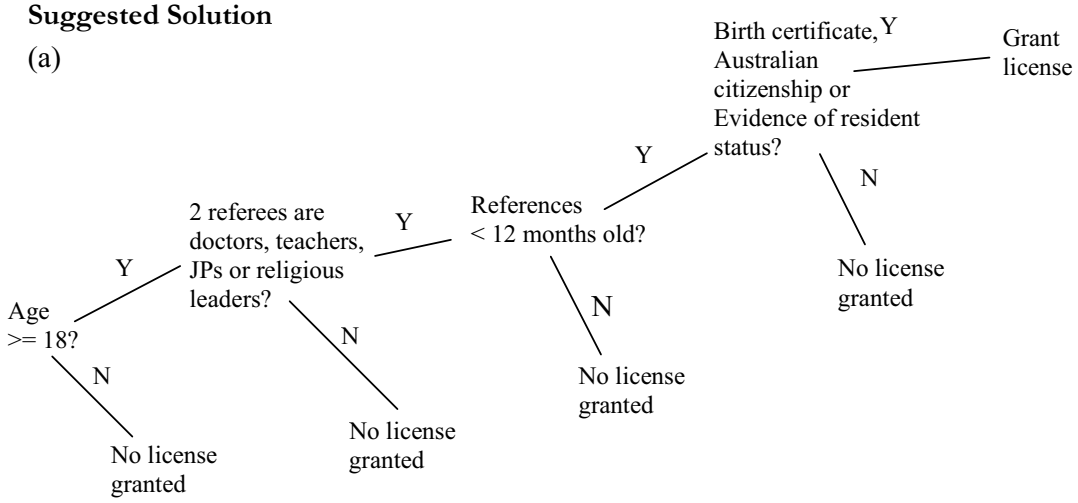
Modify any relevant rules in the knowledge base and include any new required rules to incorporate this extra information.

(c) Distinguish between the role of the human expert and the knowledge engineer in the development of this expert system.

(d) Describe the Processing information processes occurring during a consultation as the inference engine backward chains and eventually concludes that a security licence should be issued. Refer to the original knowledge base during your discussion.

Suggested Solution

(a)



(b) Modify Rule 2 as follows:

2	Conditions met	IF	Authorised training AND Good references AND Australian resident

Add rules 8 and 9 as follows:

8	Authorised training	IF	Course offered by RTO AND 100% in Law section AND >50% in other sections
9	RTO	IF	On list approved by Commissioner of Police AND current RTO Master License number

(c) The human expert is a person who is recognised as being knowledgeable in the area of this expert system. In this particular system, they would likely be a senior manager in the relevant Police department who is highly knowledgeable in the rules relating to the requirements for granting of a security license.

The knowledge engineer would talk to this human expert to elicit the required facts and rules necessary to issue security licences. They would need to identify any inconsistencies or gaps in the information received and resolve these in conjunction with the human expert. The knowledge engineer encodes the information into rules using the required syntax of the expert system shell being used. Finally the knowledge engineer specifies the text of questions and the format of the user interface and results.

(d) The overall goal is to determine if a security licence can be issued.

- This goal means Rule 1 will be considered first, to fire this rule requires the applicant to be the correct age.
- Rule 7 is now examined and the age of the applicant is asked – an age >= 18 is entered. The inference engine then returns to its initial goal and again examines rule 1. It now has a fact that age is >=18 so it must determine a value for conditions met.
- Rule 2 is examined next as it determines if conditions have been met. The premise requires a value for Good references so Rule 5 is examined.
- Rule 5 requires a value for Authorised referees so Rule 6 is now examined.
- Rule 6 requires the clerk to confirm that the referees are doctors, teachers, JPs or religious leaders. This is true so the fact Authorised referees=True is created and the inference engine returns to Rule 5.
- The first condition in Rule 5's premise is true so the second condition is tested. This requires the clerk to confirm that the references are less than 12 months old. This occurs and therefore Rule 5 fires which creates the fact Good references=True.
- The inference engine returns to Rule 2, Good references is true so it now considers Australian resident. Therefore Rule 3 is considered.
- Rule 3 fires when Valid document sighted is True, to determine this Rule 4 is examined.
- Rule 4 requires at least one of the three options in the premise to be true. No further rules exist so the clerk is asked and the premise is found to be true. Valid document=True is created and we return to Rule 3.
- Rule 3 fires so the fact Australian resident=True is created.
- The inference engine returns to Rule 2. Both conditions are now true so the rule fires causing the fact Conditions met=True to be asserted.
- Finally the inference engine returns to Rule 1. Both conditions are now True so a security licence can be issued.

Comments

- In a trial or HSC examinations part (a), (b) and (c) would likely be awarded 3 marks and part (d) would be awarded 5 or 6 marks.
- In part (a) the suggested solution correctly describes the logic of the knowledge base, however it does not need to include every attribute that would be created within a coded knowledge base. The decision tree does not need to detail intermediate attributes whose values are inferred from facts collected directly from the user. Although the logic in the decision tree within the suggested solution is correct, it is not formatted according to the method described in chapter 1.

  In general, the logic of any knowledge base can be described using only those attributes that are collected by questioning the user or that are part of the initial facts. Values for all other attributes are ultimately derived from facts in regard to these attributes. In this knowledge base there are four questions that the user may have to answer, hence the decisions based on the answers to these four questions will form the basis of the system's logic. Other equally correct answers could be constructed that do include other intermediate decisions, however such detail would not be needed to gain full marks.
- In part (b) there are many ways to correctly modify the knowledge base using a variety of extra rules. It makes logical sense to include an extra condition within Rule 2 so that the new rules are linked to the existing rules and hence to the overall goal.
- The suggested solution in part (b) does not specifically test that the applicant has attended a suitable training course. This is a reasonable assumption given that the new Rule 8 tests the applicant achieved the required results in the test – presumably attending the course is required to sit the test.
- In part (d) the question states that the system concludes that a security licence should be issued. This means we can assume the clerk enters answers that lead to this conclusion. Without this information it would be difficult to describe the precise processes performed by the inference engine.
- In part (d) the suggested solution uses the terms attribute, fact, premise and consequent. The use of these terms is not required for full marks, however it is far easier to describe this complex processing when these terms are used.
- The suggested part (d) solution does not indicate that when processing returns to a rule it commences from the point it was previously at. A minor criticism that would be unlikely to result in a lost mark.

GROUP TASK Activity
Reformat the decision tree in the part (a) suggested solution using the method described on page 71, then construct a similar decision table.

GROUP TASK Discussion
"Forward chaining would be a reasonable inference strategy for this situation." Do you agree? Discuss using examples from the decision tree in the suggested solution.

GROUP TASK Discussion
Say, licence applicants complete an online form that causes their answers to be stored in a database - determining whether to grant licences occurring at a later time. Assess the suitability of forward chaining compared to backwards chaining in this situation.

SET 5D

1. In an expert system rules are stored within the:
   (A) knowledge base
   (B) database of facts
   (C) inference engine
   (D) explanation mechanism
2. Which of the following is TRUE for the rule "If streetlights are on then it is probably night"?
   (A) "streetlights are on" is the consequent and "probably night" is the premise.
   (B) "streetlights are on" is the premise and "probably night" is the consequent.
   (C) Both "streetlights are on" and "probably night" are premises.
   (D) Both "streetlights are on" and "probably night" are consequents.
3. Tasks performed by knowledge engineers include:
   (A) consulting with human experts.
   (B) designing rules.
   (C) coding rules using the syntax required by the expert system shell.
   (D) All of the above.
4. Facts can be established by:
   (A) Asking the user questions.
   (B) Firing rules.
   (C) Entering them into the initial system.
   (D) All of the above.
5. In an expert system the order in which rules are examined is determined by the:
   (A) knowledge base
   (B) database of facts
   (C) inference engine
   (D) explanation mechanism
6. Backward chaining results in which of the following?
   (A) The ability of the system to explain its conclusions.
   (B) Reasoning that closely reflects that used by a human expert.
   (C) Each rule being tested in the order it appears in the knowledge base.
   (D) A complete knowledge base describing the rules that control the system's logic.
7. What occurs each time a rule fires?:
   (A) One or more rules are added to the database of facts.
   (B) One or more facts are added to the database of facts.
   (C) One or more rules are added to the knowledge base.
   (D) One or more facts are added to the knowledge base.
8. When designing rules for a knowledge base, which of the following strategies is generally used?
   (A) Commence with the overall goals and progressively add more detailed rules. Include questions only when they can be answered objectively.
   (B) Produce rules as required and finally edit their consequents to achieve the goals. Questions can be asked for any unknown attributes.
   (C) Identify the overall goals and user questions and then develop rules that link the goals with the questions.
   (D) Commence with questions, develop the rules that fire in response to these rules, continue developing rules until finally the goal or goals are achieved.
9. During backwards chaining which of the following does NOT occur?
   (A) Facts are established when rules fire.
   (B) If no fact about an attribute within a premise is known the inference engine first looks for rules with the attribute in their consequent.
   (C) During inference processing the overall goal is always at the top of the goal list.
   (D) The user is asked questions only when no fact in regard to the attribute can be established using rules.
10. Which of the following is true in regard to confidence factors?
    (A) They are added together during inference engine processing.
    (B) Their value is attached to attributes.
    (C) Their value is attached to facts.
    (D) Their value cannot be altered by users.
11. Explain the purpose of each of the following components of expert systems.
    (a) Knowledge base
    (b) Database of facts
    (c) Inference engine
    (d) Explanation mechanism
12. Define each of the following expert system terms.
    (a) Rule
    (b) Premise
    (c) Consequent
    (d) Fact
    (e) Attribute
    (f) Certainty factor
13. Distinguish between backward chaining with forward chaining. Provide an appropriate example where each inference strategy would be used.
14. Outline the tasks performed by a knowledge engineer as they develop an expert system.
15. Recount the backward chaining inference processes occurring to achieve the RainGear goal using the knowledge base in *Fig 5.57*. Assume it is not raining, it is very cloudy and it is sunny outside.

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANNs) simulate the organisation, analysis and processing information processes performed by the human brain. Like the human brain, ANNs are able to learn by experience and then apply their new knowledge to new unseen problems. These characteristics make artificial neural networks particularly well suited to complex unstructured decision situations where the method of solution is poorly understood. Unlike the human brain, artificial neural networks are designed to solve specific types of problems and are not easily able to transfer their knowledge to the solution of largely unrelated problems. In general ANNs are trained using sample data that includes the desired results. It is only when training is completed that the ANN is able to solve similar unseen problems. For instance during training for an OCR (Optical Character Recognition) application the artificial neural network is provided with numerous bitmaps of words written using different handwriting and fonts together with the actual words in each bitmap. Once trained the ANN can recognise the words within bitmaps even when the handwriting and fonts are different. OCR is largely a pattern matching exercise – problems that involve such pattern matching decisions are well suited to solution using ANNs.

BIOLOGICAL NEURONS AND ARTIFICIAL NEURONS

The human brain is a highly complex biological neural network composed of some 10^{11} (1,000,000,000,000) individual neurons – even an ant's brain has more than 20,000 neurons. Neurons are the main information processing cells within the brain. In the human brain each neuron can connect to around 100,000 other neurons. Furthermore these connections are created, deleted and altered as we learn. In simple terms each neuron receives electrical inputs from other neurons, decides whether to fire and if it does fire then an electrical signal is output to adjoining neurons. Artificial neurons (also known as processing elements or PEs) perform similarly – all the inputs to the neuron are mathematically combined and if the result exceeds some threshold the neuron fires causing an output to adjoining neurons.

Consider the biological neuron in *Fig 5.62*. The soma is the processing centre of the neuron; it contains the cell's nucleus. The dendrites are the inputs from adjoining neurons and the axon transmits the output. There is a single axon that transmits the single output to many axon terminals. The axon terminals are in close proximity to the dendrites of adjoining neurons. The space between each axon terminal and adjoining dendrite is called a synapse. This space determines how much of the signal from one neuron is received along the dendrite and into an adjoining neuron – the smaller the synaptic space the larger the electrical signal and vice versa.

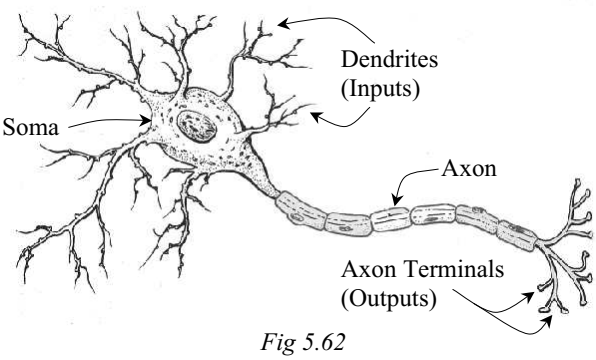


Fig 5.62
Biological neuron.

Let us now consider the relatively simple model of an artificial neuron shown in *Fig 5.63* – in reality the mathematics is significantly more complex. For our discussion the model in *Fig 5.63* illustrates the basic principles in sufficient detail. Like the biological neuron there are many inputs labelled I_1, I_2, I_3,…I_n, and a single output labelled *S*. Each *I* input comes from a single prior neuron's *S* output. Similarly the *S* value output from one neuron connects directly to an *I* entering one or more subsequent neurons.

The function of the biological neuron's synaptic space is performed by weighting each input using the W_1, W_2, W_3,… W_n values in *Fig 5.63*. During training of the network it is the value of these W weights that changes. In most artificial (and also biological) neurons the outputs are continuous values, however in our simplified neuron discussion let us assume the outputs S and therefore also the inputs I are either 0 or 1. This means either a neuron has fired (1) or has not fired (0). It also simplifies deciding whether the neuron should fire as we simply need to determine if the calculated value x is greater than some threshold value T – creating a step function as shown in *Fig 5.64*. In reality T is a mathematical function that smooths the step function so that neurons fire with different levels of intensity – often S-shaped sigmoid functions similar to *Fig 5.65* are used.

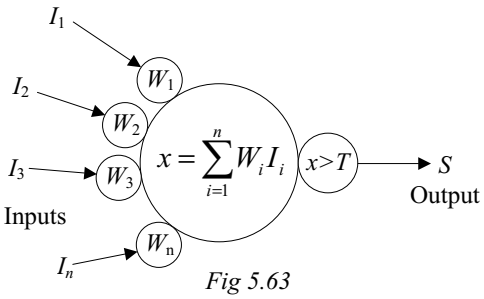


Fig 5.63
Simplified Artificial Neuron.

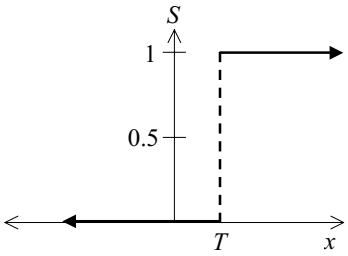


Fig 5.64
Binary step function.

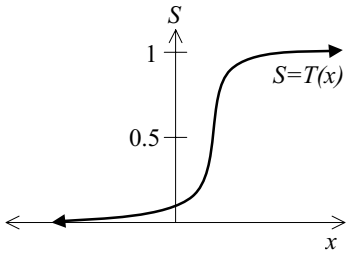


Fig 5.65
S-shaped sigmoid function.

During training various W values between –1 and 1 are allocated to each input. For example, say a neuron with three inputs I_1, I_2, I_3 was allocated weight values during training of W_1 =0.9, W_2 = –0.3 and W_3 =0.7 respectively. The neuron's activation value is calculated – the x value in *Fig 5.63*. In general this x value is the sum of the products of each input/weight pair. In our example, say the first and second inputs I_1, I_2 come from neurons that fired and the third input I_3 is from a neuron that did not fire. In this case x is calculated to be 0.6 as shown in *Fig 5.66*. If this x value is greater than the neuron's threshold value T then the neuron fires and the output S is set to 1. In our example if the neuron's threshold (T value) is 0.5 then the neuron will fire as x is indeed greater than T ($0.6 > 0.5$). As a result a binary 1 is output as the neuron's S value. This S value is transmitted and subsequently received as an input (I value) to further neurons.

$$
\begin{aligned}
x &= \sum_{i=1}^{3} W_i I_i \\
&= W_1 I_1 + W_2 I_2 + W_3 I_3 \\
&= (0.9 \times 1) + (-0.3 \times 1) + (0.7 \times 0) \\
&= 0.6
\end{aligned}
$$

Fig 5.66
Activation value calculation.

Compared to artificial neurons each biological neuron within the brain takes a significant amount of time to fire – approximately 0.01 seconds for a human brain neuron and some 0.000000001 seconds for artificial neurons. However within the human brain thousands or even hundreds of thousands of neurons fire in parallel – the human brain is really a massively parallel processor. Most artificial neural networks contain less than a hundred neurons, many far fewer, and even on the most advanced CPUs only a few neurons can fire at the same time.

GROUP TASK Discussion
Compare and contrast each component of the biological neuron in *Fig 5.62* with its corresponding component within an artificial neuron.

STRUCTURE OF ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANNs) contain multiple neurons arranged into layers. *Fig 5.67* shows the organisation of a typical feedforward ANN, this is the most common type of ANN, however be aware that other types do exist. Feedforward ANNs contain an input layer and an output layer and most implementations contain just one or two hidden (or middle) layers. The output from each neuron connects to each of the neurons in the subsequent layer – some arrows are missing in *Fig 5.67*. In most ANNs

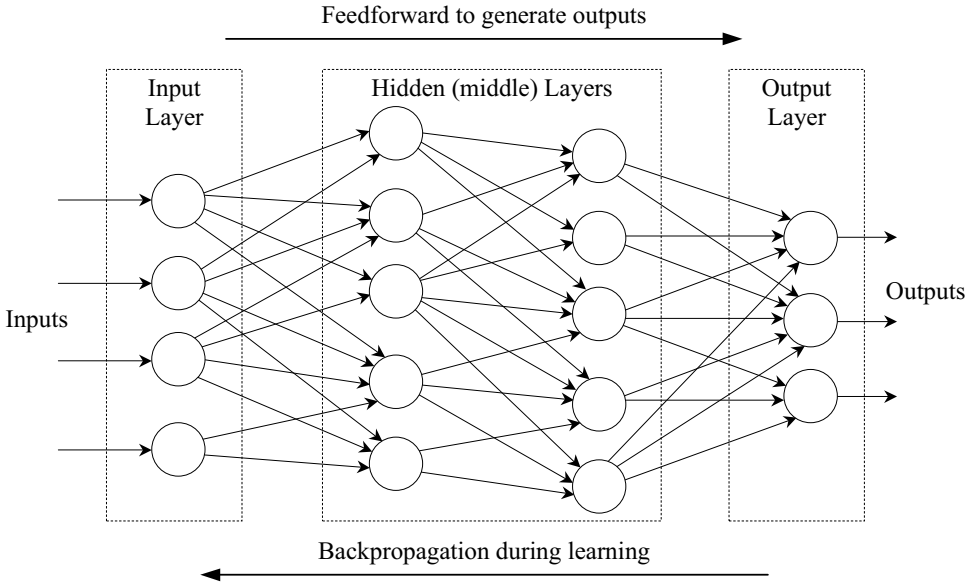


Fig 5.67
Typical artificial neural network with two hidden layers.

the input layer simply passes its inputs onto each neuron in the first hidden layer. The hidden and output layers are where the real processing occurs. The hidden layers are composed of neurons that each produces their own distinct output that feeds into each neuron in the next layer. The final hidden layer's outputs feed into the final output layer, which also contains neurons. Outputs from the output layer are the final results of the ANN. This is known as feedforward processing and hence the design is known as a feedforward ANN.

The outputs from an ANN are really predictions based on the neural network's past experiences. The past experience is learnt during training and stored within each neuron as its individual weights and threshold details. The combination of many neurons allows the ANN to make generalisations such that it can generate accurate predictions for new sets of data inputs.

Enough theory, let us now consider possible structures for two example ANNs – a simple OCR (Optical Character Recognition) neural network and a basic market price prediction neural network.

Simple OCR Neural Network

This neural network aims to detect the digits 0 to 9 within monochrome bitmaps with a resolution of 8 pixels by 8 pixels – a pattern matching exercise. In *Fig 5.68* the bitmap clearly represents the digit 3, however there are numerous other ways to create a bitmap that we would consider to represent a 3. The network will be trained with many such examples of each of the digits 0 to 9 together with the expected outputs.

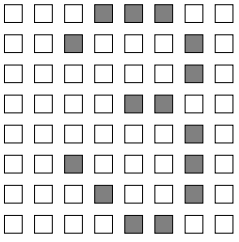


Fig 5.68
Monochrome bitmap

Clearly the input to this network is an unseen bitmap and the final output will be the digit the network "thinks" it recognises within the bitmap. Consider the input bitmap in *Fig 5.68*, it contains a total of 64 pixels and each pixel is either white (0) or black (1). The network could be designed to include 64 input neurons – one for each pixel. This would work well with the simplified neuron design we described above. However, we could also consider each row (or column) of pixels as a single input. In this case the input layer would contain 8 input neurons each receiving an integer from 0 to 255. Say, our network encodes each row such that each column is represented by a power of two. *Fig 5.69* shows how the example bitmap from *Fig 5.68* would be encoded using this system.

	128	*64*	*32*	*16*	*8*	*4*	*2*	*1*	*Neuron input values*
Neuron 1	0	0	0	1	1	1	0	0	16 + 8 + 4 = 28
Neuron 2	0	0	1	0	0	0	1	0	32 + 2 = 34
Neuron 3	0	0	0	0	0	0	1	0	2
Neuron 4	0	0	0	0	1	1	0	0	8 + 4 = 12
Neuron 5	0	0	0	0	0	0	1	0	2
Neuron 6	0	0	1	0	0	0	1	0	32 + 2 = 34
Neuron 7	0	0	0	1	0	0	1	0	16 + 2 = 18
Neuron 8	0	0	0	0	1	1	0	0	8 + 4 = 12

Fig 5.69
Example encoding using 8 input neurons.

Now consider the output layer. It could contain a single neuron that outputs values from 0 to 9 directly. This is possible, however such a design would only provide the most likely digit recognised. A more useful design would use 10 output neurons, one for each digit. Each output generates a number representing the likelihood or probability that each digit has been recognised. For our example bitmap one would expect the 4th output neuron, representing the probability of the digit 3, to output the highest value, whilst the 6th and 9th neurons representing the digits 5 and 8 would likely output a significant but lower probability.

Deciding upon the structure of the hidden or middle layers is more difficult. Even in real world systems the number of layers and the number of neurons within each hidden layer is largely a trial and error exercise. If there are too few neurons the network will not be able to detect sufficient detail to generalise. However if too many neurons are used then the network becomes too sensitive to minute insignificant details within the training data. In both cases the results will be poor. A common strategy is to progressively add more neurons, retrain the network and then use unseen test data to determine the accuracy of the results. Eventually a point is reached where adding more neurons decreases the accuracy of the results, in theory the previous version should be close to the optimal network. Often minor tweaking will further improve the results. Once the hidden layer (or layers) and training are complete the neural network is ready to predict digits present in unseen bitmaps.

GROUP TASK Activity
Construct a diagram similar to *Fig 5.67* for the Simple OCR Neural Network described above. Assume a single hidden layer containing four neurons is sufficient for an acceptable level of accuracy.

Market Price Prediction Neural Network

Earlier in this chapter we discussed the nature of predicting stock prices. We described this situation as unstructured and identified some of the possible data inputs that could be used to predict future prices. The following extract is reproduced from California Scientifics' website, it describes an ANN produced with their BrainMaker software that can apparently predict mutual fund prices with some accuracy.

Mutual Fund Prediction

Dr. Judith Lipmanson of CHI Associates in Bethesda, Maryland, publishes technical business documents and newsletters for in-house use at technical and advisory firms. She also is a technical analyst who uses a neural network to predict next week's price of 10 selected mutual funds for personal use.

For the past several months, she has been using a BrainMaker neural network. The network gets updated with new data every week, and takes only minutes to retrain from scratch on a PC.

Results have been good. Currently, the network is producing outputs which are about 70% accurate. Although the network is not perfectly accurate in its predictions, she has found that the neural network make predictions which are useful.

Dr. Lipmanson's network relies on historically-available numerical data of the kind typically found in back-issues of the Wall Street Journal. These indicators include such factors as the DOW Industrial, DOW Utilities, DOW Transportation and Standard & Poor's 500 weekly averages. Several years worth of data was gathered for the four initial conditions (the inputs) and the ten results (the outputs). The results were shifted by a period of one week and the information was used to train the network.

The network looks something like this:

Inputs:

The DOW Industrial
The Dow Utility
Dow Transportation
S & P 500

Outputs:

Fund # 1 next week	Fund # 6 next week
Fund # 2 next week	Fund # 7 next week
Fund # 3 next week	Fund # 8 next week
Fund # 4 next week	Fund # 9 next week
Fund # 5 next week	Fund # 10 next week

She collects the closing weekly averages on Friday and uses the new data to predict prices of the 10 mutual funds for the next week. Making forecasts with a trained network requires only a few seconds, and the network can be readily updated with new information as it arises.

(Source: extract of an article on California Scientifics' website)

GROUP TASK Discussion

Brainstorm other possible inputs that may improve the performance of this network. Describe the nature of the training data and testing data for this neural network.

GROUP TASK Research and Discussion

Research examples of stock price prediction neural network software. Do you think these software applications are really able to predict stock market prices with a high degree of certainty? Discuss.

HOW BIOLOGICAL AND ARTIFICIAL NEURAL NETWORKS LEARN

Within the brain learning occurs as the size of the synaptic spaces change and as a consequence the significance or weight allocated to each input changes. In general, connections where the neuron fires more often tend to strengthen, which means the synaptic space closes and a stronger signal passes to the receiving neuron. As the neuron is now receiving different input levels it fires under a different set of input conditions. This is occurring in numerous synapses connecting numerous neurons and ultimately results in changes to how we react to new inputs. The outputs have changed and therefore learning has occurred and is subsequently applied to new problems. When teaching a small child to recognise letters of the alphabet we present the child with training data in the form of an ABC book. If the child correctly recognises a letter we praise their efforts. If they are incorrect, we provide them with the correct letter. Learning occurs in either case. Over time and based on many different training sets the child learns to recognise each letter – even when different fonts or handwriting styles are used.

Within artificial neural networks similar training processes occur. The network is presented with inputs, which the network processes through its neurons into outputs. Initially most outputs will be incorrect, however the network, much like the child, learns by its mistakes. When training an ANN each generated output is compared to the known or desired outputs. If they don't match then weights in some neurons are adjusted such that the results begin to approach the desired results. If the output closely matches the desired result then these weights are given higher significance. During training we expect the network to progressively recognise patterns in the input data with increasing accuracy. However this does not always occur. Commonly various parameters such as the number of hidden layers, number of neurons in each layer and even the addition of new inputs will be tried. The new network must then be retrained from scratch. The aim being to predict the outputs with sufficient levels of accuracy that the ANN can be relied upon to make predictions based on new unseen inputs.

Not all unstructured decisions can be reliably solved by ANNs. For an ANN to learn how to process the inputs into outputs there needs to be some relationship between the inputs and the outputs. ANNs are particularly useful tools when the designer suspects such a relationship exists but determining the mathematics of the relationship is not possible or is extremely complex to determine using more traditional techniques. If this relationship (method of solution) can be determined then there is no need to use an ANN. When designing an ANN all inputs that one suspects will have an influence on the outputs should be included – during training the ANN will eventually learn to ignore any inputs that are irrelevant.

GROUP TASK Discussion
Identify and briefly describe features of decisions that make them possible candidates for solution using artificial neural networks.

Consider the following (Extension).

How are the weights and threshold parameters within each artificial neuron altered during Training? There are many standard techniques that are available and often a variety of techniques are tried. Back propagation and genetic algorithms are two common training techniques.

Back Propagation

Back propagation works backwards from neurons in the output layer, through the neurons in the hidden layers and finally to the input neurons. It first compares the current output from each output layer neuron with the desired output from the training data. Initially there will most likely be a significant difference. The back propagation algorithm then considers the inputs received from hidden layer neurons to each output layer neuron – stronger inputs are assumed to have higher significance. The weights are then adjusted temporarily so that the output neurons produce results closer to the desired outputs.

The above process is repeated on the hidden layer neurons such that they now have new temporary weights. These new hidden layer weights will also affect the output layer, so the process must be repeated for the output layer. If the results are closer to the desired results then the algorithm works backwards again until it eventually reaches the input layer. If the training inputs are similar then all weight changes are made permanent. The entire process is repeated hundreds or even thousands of times using the entire set of training data. Over many such repetitions (known as epochs) better solutions begin to emerge. In general accuracy continues to improve for a while and then begins to decrease. Obviously the system retains the best solution.

Genetic Algorithms

Genetic algorithms use techniques based on the changes that take place as plants and animals evolve. There are two significant techniques that occur, the first simulates sexual reproduction and the second simulates mutations. Different genetic algorithms use sexual reproduction and mutations in different sequences. The following discussion describes one possibility, however the detail of sexual reproduction and mutation is similar in all implementations.

For sexual reproduction, genetic algorithms determine two possible solutions (complete sets of neuron weights and other parameters) that both have merit in terms of achieving the desired results. These solutions are known as chromosomes, reflecting their purpose during biological sexual reproduction. Each weight is like a gene within a real chromosome. Initial chromosomes are produced either randomly, using back propagation or some other technique. To implement sexual reproduction the genetic algorithm takes a random set of genes (weights) from one chromosome and overwrites these genes within a copy of the other chromosome. This produces a new chromosome that possesses characteristics of both parent chromosomes. This possible solution is tested using the training data. If it produces better results than its parents then it is retained as a new parent for subsequent "breeding". But what if breeding has been attempted many times but no better solution emerges? In this case the system will try mutating chromosomes. This simply means some of the genes are randomly changed in the hope that a better solution will emerge. Mutations that do not produce better solutions are discarded – just like in nature. The entire process repeats until a sufficiently accurate solution (chromosome) has evolved.

GROUP TASK Research
Research examples of ANN software applications. Determine whether these applications use back propagation and/or genetic algorithms during training of the network.

GROUP TASK Research
Genetic algorithms are not just used to train ANNs. Research and briefly outline other applications where genetic algorithms are used.

HSC style question:

Women are advised to have a Pap smear done each year, to detect cells that might develop into cancer of the cervix. A sample is taken of cells from the surface of the cervix and this sample is placed onto a slide, spayed with a fixing chemical, and sent to a laboratory for examination. Detected early, cervical cancer has an almost 100% chance of cure.

Papnet is the name of a neural network system designed to assist in the process of analysing these slides to detect abnormal cells.

> Since a patient with a serious abnormality can have fewer than a dozen abnormal cells among the 30,000 - 50,000 normal cells on her Pap smear, it is very difficult to detect all cases of early cancer by this "needle-in-a-haystack" search. Imagine proof-reading 80 books a day, each containing over 300,000 words, to look for a few books each with a dozen spelling errors! Relying on manual inspection alone makes it inevitable that some abnormal Pap smears will be missed, no matter how careful the laboratory is. In fact, even the best laboratories can miss from 10% - 30% abnormal cases

Source: http://ww.openclinical.org/neuralnetworksinhealthcare.html

> "The traditional Pap smear has contributed to the dramatic decrease in deaths from cervical cancer. Nevertheless, its accuracy has been limited to-date because of its dependence on the microscope and human manual screening," said Dr. Gary Goldberg, professor and director of gynecology, associate director of gynecologic oncology, Albert Einstein College of Medicine and Montefiore Medical Center.
>
> "This study demonstrated that this applied computer technology **(Papnet)** can assist the cytologist/cytopathologist in detecting a greater number of abnormal smears and, therefore, help the clinician prevent invasive cervical cancer."

Source: http://www.pslgroup.com/dg/3ecb2.htm

(a) Describe the sources of data for the Papnet neural network system, and the role of training in the effectiveness of this system.

(b) Critically analyse the responsibility of those involved in decision making using this Papnet neural network system.

(c) Justify why an expert system was determined to be an unsuitable or inappropriate solution to this problem.

(d) Critically evaluate the use of this neural network system compared to the previous manual system where trained pathologists would manually check each slide.

Suggested Solution

(a) The digital image of each slide needs to be made available to the Papnet system, presumably by either being scanned into the system, or digitally imported from the machine that reads the slides. In addition, during the training cycle a trained person also needs to enter for each selected slide the result of whether it is pre-cancerous or not.

The effectiveness of the system depends very strongly on how many slides are submitted during the training cycle, how different they are, and how accurate the results for each slide are as they are entered into the system. The Papnet system uses this information to set the weightings and threshold values for each of its neurons within its neural network so that all of the slides that have been input produce a correct output result.

The hope is that when the system is operational with previously unseen slides, it will continue to use these set weightings and threshold values to produce an output for each slide that is equally correct.

(b) The laboratory staff or cytologists must ensure that the results produced by the system are reasonable, and that they do not make erroneous inferences from the output of this system – in the early stages of the use of the system and at regular intervals thereafter, they should manually check the results of specific slides to check that the results are consistent with a manual check of the slide. The images of positively identified slides should be reviewed by a cytotechnologist.

They should not use the Papnet system for any purpose other than that which it was designed for, particularly as it has been trained solely for the purpose of detecting the existence of pre-cancerous cells. They should not use the system to predict any other relationship or factor.

They should be very aware of privacy issues and not allow any personal or identifying data to be stored with the digital slide images and the result, in case the data is subsequently accessed or made available to others for a purpose other than that of simply identifying pre cancerous cells.

(c) An expert system requires the definition of facts and rules to be developed by a knowledge engineer in consultation with an expert in the field. These facts and rules are entered into the expert system software using the required syntax. In a case such as the Papnet system, it is probably very difficult to formulate a consistent, reliable, comprehensive set of rules that will accurately predict the state of the cells. It is much more likely that the trained laboratory technicians use their experience and intuition to identify positive slides, without being able to verbalise the 'rules' they apply to arrive at their decision. They intuitively use pattern matching to identify relevant slides, which is exactly what a Neural Network does best.

(d) The previous manual system has some real deficiencies.

- There is a large error rate with the manual system, approx 10% – 30%.
- There is the need to train laboratory staff, whose job then entails looking at thousands of slides a day which must become tedious and ergonomically stressful.

On the other hand, the Papnet system has some very real advantages

- It has greatly improved consistency – it does not get tired or bored and is not impacted by personal stress or emotion like human laboratory staff.
- It has greatly improved performance – it is able to produce results much faster, and therefore processes many more slides per day
- It has the ability to respond accurately to previously unseen samples – no preconceptions as to what might constitute a positive sample, it merely performs a pattern matching test based on its previous training.
- It preserves the expert knowledge that was used at the time of training. If that person (or team of people) is no longer available, patients and doctors can still be assured that they are utilising the benefit of their expert experience in subsequent diagnosis.

The disadvantages of the Papnet system are

- No possibility of an explanation for its decisions (unlike a human expert) – once a result for a slide is output, it must just be accepted without any supporting explanation.
- The output will likely be less accurate if slides that are very different from those used in the training cycle are assessed. In this case, it may be necessary to retrain the network using a range of samples including such 'unusual' slides in the new training cycle.
- A trained cytologist should always check the results produced by the Papnet system to ensure that it continues to produce consistently accurate results.
- If the technology fails or malfunctions, it will be difficult to fall back to a manual system as the human laboratory workers may become deskilled.

The cost of such an automated system is not insignificant, but of course it must be compared with the alternative salaries and incidental expenses required to train and employ human cytologists to do the same job, and the accuracy levels associated with each approach. The articles indicates there is strong support for the new neural network system, and that it appears to offer significant advantages over the current manual system.

Comments

- Part (b) would likely attract a total of 12 to 16 marks with each sub-part attracting 3 or 4 marks each.

SET 5E

1. Which of the following is TRUE of the inputs and outputs into and out of neurons?
   (A) Many inputs and many outputs.
   (B) One input and many outputs.
   (C) Many inputs and one output.
   (D) One input and one output.

2. The real processing within a feedforward ANN occurs in which layers?
   (A) Input and output layers.
   (B) Middle and output layers.
   (C) Input and middle layers.
   (D) Middle or hidden layers.

3. An artificial neuron has a negative weight for one of its inputs. Which of the following best describes the effect on the neuron's output?
   (A) The neuron will always fire with less intensity than would have occurred if the weight was positive.
   (B) The neuron will always fire with greater intensity than would have occurred if the weight was positive.
   (C) The neuron will fire with less intensity for greater input values and with greater intensity for lower input values.
   (D) The neuron will fire with less intensity for larger input values and there will be less effect on the output for lower input values.

4. Most feedfoward ANNs contain how many layers of neurons>
   (A) 1 or 2
   (B) 3 or 4
   (C) 5 or 6
   (D) More than 6.

.5. Two sets of neuron weights and threshold values are randomly combined. What is most likely occurring in this ANN?
   (A) Learning using a genetic algorithm and sexual reproduction.
   (B) Learning using a genetic algorithm and mutations.
   (C) Learning using back propagation.
   (D) Learning using backward and forward chaining.

6. The weights attached to an input into a biological neuron is determined by:
   (A) the input value.
   (B) the synaptic space.
   (C) a combination of the input value and the synaptic space.
   (D) the soma.

7. Each weight in an artificial neuron corresponds to which part of a biological neuron?
   (A) Soma
   (B) Synaptic space
   (C) Axon
   (D) Dendrite

8. Common training strategies for ANNs include:
   (A) back propagation and genetic algorithms
   (B) rule induction and regression techniques.
   (C) decision tree algorithms and K-nearest neighbour.
   (D) genetic algorithms and data mining.

9. Why is a biological neural network able to process data faster than an artificial neural network?
   (A) Biological neurons take less time to fire compared to artificial neurons.
   (B) Artificial neurons take less time to fire compared to biological neurons.
   (C) Biological NNs use parallel processing and artificial NNs do not.
   (D) Artificial NNs use parallel processing and biological NNs do not

10. When selecting the inputs for an ANN, which of the following is the most appropriate advice?
   (A) Only include inputs that can clearly be processed into the desired outputs.
   (B) Include as many inputs as possible.
   (C) There is no point including inputs that have an obvious effect on the outputs.
   (D) Include all inputs that may possibly have an effect on the outputs.

11. Describe each of the following.
   (a) Artificial neuron
   (b) Feedforward processing
   (c) Backpropagation
   (d) Genetic algorithms

12. Compare and contrast biological neurons with artificial neurons

13. Explain how an artificial neuron processes its inputs into an output.

14. Describe the structure of a typical artificial neural network.

15. Compare and contrast human learning with ANN learning

ISSUES RELATED TO DECISION SUPPORT SYSTEMS

Decision support systems are intelligent systems that improve the decision-making ability of individuals and groups. Although there are many advantages of using decision support systems, they should be used with caution. The decisions made by such systems, particularly when the consequences are of a critical nature, need to be rigorously tested rather than being blindly accepted as fact. Most decision support systems deal with uncertainty; hence the results should be interpreted within the level of uncertainty of the system.

REASONS FOR INTELLIGENT DECISION SUPPORT SYSTEMS

Throughout this chapter we have examined many different types of decision support systems that all aim to make intelligent decisions. These systems largely simulate the decisions made by humans, in particular expert systems. In this section we consider some of the reasons for developing intelligent systems.

Preserving an expert's knowledge

Many intelligent systems model the knowledge of human experts. This is particularly true of expert systems, where the rules and facts are developed by the knowledge engineer to specifically create a system that will make decisions based on the same heuristics used by the human expert. The expert system can then be distributed and used by many users without requiring an expert to be in attendance. For example expert systems for diagnosing problems with vehicles, aircraft and other complex machinery can be developed by the manufacturer and distributed to service centres throughout the world.

An expert's knowledge is also modelled in other types of decision support systems. For example an experienced statistician can create a spreadsheet, which can then be used to assist others to make decisions. In this case the expert's knowledge is encoded into formulas, macros and other features within the spreadsheet model. The model can then be used by less experienced users to perform similar statistical analysis processes without the statistician being present. Neural networks can also be designed that preserve expert knowledge when the expert is unable to explain or rationalise their reasoning. This is particularly apparent for unstructured situations where the expert is known to be able to solve the problem due largely to their extensive experience, however the complexity of their reasoning makes it difficult to establish rules upon which they base their decision.

GROUP TASK Activity

Flick back through this chapter and identify examples of decision support systems that preserve an expert's knowledge. In each case explain advantages of preserving the expert's knowledge.

Improving accuracy and consistency in decision-making

Decision support systems reach exactly the same conclusion each time they are presented with the same set of inputs. This is not always the case when humans make decisions. Humans are affected by emotions, stress, tiredness and many other factors. These factors affect their ability to objectively make decisions. The problem is further exacerbated when many different people must reach conclusions. Decision support systems will consistently reach the same conclusion given the same inputs.

Assuming the decision support system is able to produce accurate results it will continue to do so indefinitely. On the other hand human decision-making ability changes over time – perhaps becoming more accurate or perhaps becoming less accurate. Some decision situations change over time – relevant inputs change and

even the rules used to decide can change. It is important that the accuracy of all decision making, both automated and human, is regularly validated to ensure accuracy and consistency of results.

GROUP TASK Discussion
"Decision support systems are only as good as their developers".
Do you agree? Discuss.

Rapid decisions

Decision support systems can, in most situations reach conclusions many times faster than humans. In some cases, such as data mining applications, the amount of data that needs to be analysed is enormous, this means manual analysis by humans is simply not viable. The speed with which computers can analyse vast quantities of data means that many more possible conclusions can be investigated much more thoroughly than would be practical using manual techniques. A decision support system can keep track of many hundreds of different attributes and their relationships to each other, whilst even the most capable expert will manage to understand and process just a few relationships. For example we humans can fairly accurately determine trends between two variables when presented with a 2 dimensional graph, however we have difficulty determining such trends in 3 dimensions, let alone 4, 5 or 100 or more dimensions. Computers can easily analyse and determine such trends.

GROUP TASK Discussion
"Computer based decision support may be faster at suggesting possible solutions, but ultimately the human brain is the real decision maker"
Discuss in terms of the development and also the use of DSSs.

Ability to analyse unstructured situations

Unstructured situations are characterised by significant uncertainty. In general conclusions are difficult to justify objectively as no clear method of solution is known. This can be due to the complexity of the problem or a lack of understanding of the variables influencing or affecting the outcome. Therefore, such conclusions are expressed with different levels of confidence or probability. As humans we too express conclusions with varying levels of confidence, indeed often the levels of confidence produced by decision support systems are calculated based on a human's perception of how reasonable a conclusion appears to be.

The ability for decision support systems to deal with unstructured situations is largely about their ability to find patterns in data and develop generalisations that model these patterns. When the DSS is presented with further data it attempts to match or categorise the data according to its existing known patterns. In some cases the known patterns continually change whilst in others training and implementation are quite separate processes. The human brain develops complex links and associations between all the knowledge it stores. New knowledge is continually being added and existing knowledge modified and updated. All our combined knowledge is available each time we make a decision. In contrast a DSS is limited in its ability to make inferences, it is restricted to knowledge (perhaps vast quantities of knowledge) in one particular area.

GROUP TASK Discussion
In unstructured situations the conclusions made by DSSs can prove to be unworkable or inaccurate in the real world. Why is this? Discuss.

PARTIPANTS IN DECISION SUPPORT SYSTEMS

Decision support systems are designed, developed and operated by people. They are designed to assist decision-making by suggesting solutions rather than providing definitive answers. It is the responsibility of the system's participants to use the conclusions from such systems for the purpose for which they were intended.

Some issues that can arise in regard to DSS participants include:

Erroneous inferences

Decision support systems are dealing with semi-structured to unstructured situations. As a consequence it is inevitable that some inferences made will prove to be incorrect. It is critical that participants understand that decision support systems are there to support decision-making; they add evidence to assist people to make more informed decisions. People should not blindly accept the recommendations made by a DSS.

Data mining is particularly likely to infer relationships that are either incorrect or are of little or no use in the real world. Those performing data mining must have a clear understanding of the data and also of the business whose data is being mined. In most cases it is necessary to run numerous data mining experiments using a variety of different data mining tools. Each tool may suggest different conclusions - some conclusions will be useful whilst others will be useless. Those conclusions or inferences that appear useful should be tested in the real world. For example, a group of 100,000 customers likely to be interested in a range of new products may be determined as a consequence of data mining. Rather than send out a catalogue to all 100,000 customers it would be prudent to first validate the recommendation by sending out say 1,000 catalogues and monitoring the response.

GROUP TASK Discussion

Why is data mining particularly likely to produce incorrect inferences or inferences that are of little use in the real world? Discuss.

Privacy concerns

The ability of information systems to store and process large quantities of data about individuals for a variety of different purposes raises privacy concerns. Furthermore data is traded between organisations much like any other product. In regard to decision support systems data mining often raises more significant privacy issues than other types of DSS. Data mining makes and/or requires connections between records from different sources. For example details collected when a customer signs up for a store loyalty card can be linked to details of each of their future purchases. The store may purchase further data from other stores and organisations in an attempt to link customer records and build more complete profiles of their customers. The attributes used to link an individual's records often contain private information such as names, addresses, phone numbers and so on. This information may not be of relevance in terms of the conclusions and inferences made, however it is required if the data is to be successfully linked.

In general, organisations have legitimate reasons for maintaining records of their customer's details and interactions with the organisation. However privacy laws require that customers be informed of the purpose of collecting any private data, including whether the data will be sold or otherwise provided to other organisations. This information often forms part of an agreement entered into between the customer and the organisation when the customer first opens an account.

Decision support systems that collect and process sensitive information, such as medical records, racial or ethnic background or criminal records, have much more

stringent privacy requirements. Individuals must explicitly give their consent before sensitive data is collected and the organisation must explain precisely how the data will be used. Apart from specifically approved research activities, sensitive data cannot be used as part of data mining activities. Even when consent has been given the organisation must implement extra security measures to ensure others cannot access the data. Such security measures include restricting access by outside organisations and individuals and also restricting access within the organisation. Internally organisations, such as the police and health department, create audit trails that record the user and time each record is accessed. This ensures employees only access sensitive information when required to complete their duties.

The Internet has created further difficulties as privacy laws in different countries vary considerably and enforcing such laws is difficult. Many countries have entered into reciprocal agreements where each agrees to uphold the privacy laws of the other. In general the responsibility for maintaining privacy of individual's personal details is largely the responsibility of individual organisations. People are becoming more aware of such concerns and are reluctant to divulge their personal details unless they feel confident the organisation is trustworthy. It is often in the interests of commercial organisations to make explicit statements that guarantee personal information will not be shared or divulged.

GROUP TASK Discussion
Brainstorm organisations that hold your private details, including sensitive information. Discuss how all this data could potentially be linked.

Responsibility for decisions

Decision support systems produce recommendations and evidence to assist decision makers. The responsibility for the decision lies with the person who makes the final decision. For instance a medical expert system may recommend a particular medication be prescribed. The doctor will then use their experience to confirm the recommendation and make the final decision. Decision support systems are dealing with situations that include uncertainty; hence the conclusions are not definite statements of fact. All decision support systems are influenced by the experiences of the people who created the system. Different or new situations and different users will have different experiences upon which they base their final decision – the recommendations from a DSS may well influence this decision but the final responsibility rests with the decision maker.

Consider the following decision situations:

Assume a decision support system has been produced to perform each of the following decision tasks:

- Deciding on which model washing machine to purchase.
- Determining which of three quotations to accept for a house renovation.
- Diagnosing a medical condition.
- Deciding on which company's shares to purchase.
- Deciding who to vote for in a federal election.

GROUP TASK Discussion
For each of the above situations, would you trust a DSS and simply implement its recommendations? Discuss.

Impact of decision support systems on participants

Decision support systems aim to improve the decisions participants make. They provide automated tools for gathering evidence, understanding relationships and making inferences. In most situations more accurate decisions result that take account of circumstances that would not otherwise have been practical to consider. The way decision makers make decisions changes; they no longer need to rely solely on their own experience, rather they can more easily obtain evidence from experts or based on patterns within historical data. The DSS provides further evidence to support and assist the decision maker, which allows them to make more confident, appropriate and justifiable decisions.

When new decision support systems are implemented they change the nature of work for existing participants and in rare cases they can also reduce the total number of participants required. For example a software company may implement an expert system to provide support and troubleshooting advice to users. As a consequence less support personnel are required and those that remain will need more advanced skills – presumably they will only be contacted to support issues not covered by the expert system. Overall the number of participants with expertise in the area is reduced.

GROUP TASK Discussion

A bank provides each loan officer with a new automated DSS. Each loan officer first uses the DSS in an attempt to approve loans. If unsuccessful the existing guidelines used previously can be considered in unusual situations. Discuss changes in the nature of the loan officer's work as a result of the introduction of this DSS.

CHAPTER 5 REVIEW

1. Which type of decision support system is able to learn?
   (A) Spreadsheets
   (B) Expert systems
   (C) Neural networks
   (D) Databases

2. Decision support systems are used when the decision situation is:
   (A) structured or semi-structured.
   (B) semi-structured or unstructured.
   (C) structured or unstructured.
   (D) structured, semi-structured or unstructured.

3. When no method for reaching a decision is known the situation is described as:
   (A) unstructured.
   (B) semi-structured.
   (C) unstructured.
   (D) unbounded.

4. Within fingerprints a ridge bifurcation occurs where:
   (A) a ridge ends.
   (B) significant features are apparent.
   (C) ridges are close together.
   (D) a single ridge splits into two ridges.

5. The reasoning used during consultations is best simulated using:
   (A) Spreadsheets
   (B) Expert systems
   (C) Neural networks
   (D) Databases

6. Information processes in an expert system are performed by the:
   (A) inference engine.
   (B) knowledge base
   (C) explanation mechanism.
   (D) Both A and C.

7. A formula in cell A2 contains a single cell reference and is copied into cell B5. In cell B5 the row in the reference has changed but the column remains the same. Which of the following is true of the formula's cell reference?
   (A) It is a relative reference.
   (B) The row reference is relative and the column reference is absolute.
   (C) The column reference is relative and the row reference is absolute.
   (D) It is an absolute reference.

8. In an expert system the premise of a rule has just been found to be true, what happens next?
   (A) The inference engine evaluates the next rule.
   (B) Rules that include this premise in their consequent will be evaluated.
   (C) Facts will be established based on the rule's consequent.
   (D) The goal has been achieved so the results will be displayed.

9. A genetic algorithm randomly alters some values within a possible solution. Which term best describes this process?
   (A) Decision making
   (B) Sexual reproduction
   (C) Learning
   (D) Mutation

10. Data from many operational databases is imported into a large database on a regular basis. This has been occurring for a number of years. The large database is called a:
    (A) Data mart
    (B) Data mine
    (C) OLTP
    (D) Data warehouse

11. Define the following terms.
    (a) Cell
    (b) Worksheet
    (c) Formula
    (d) Goal seeking
    (e) Backward chaining
    (f) Forward chaining
    (g) Rule (expert system)
    (h) Hidden layer
    (i) GDSS
    (j) GIS
    (k) MIS
    (l) OLAP
    (m) Artificial neuron
    (n) Data warehouse
    (o) Data mart
    (p) Data mining
    (q) Regression
    (r) Intelligent agent
    (s) Back propagation
    (t) Genetic algorithm

12. Describe the organisation of each of the following.
    (a) Spreadsheets
    (b) Expert systems
    (c) Neural networks
    (d) Data marts

13. Explain each of the following spreadsheet analysis techniques.
    (a) What-if analysis
    (b) Goal seeking
    (c) Statistical analysis
    (d) Charts

14. For each of the following types of DSS, identify a specific example of a decision situation where data is extracted from a database.
    (a) Spreadsheets
    (b) Expert systems
    (c) Neural networks
    (d) Data mining
    (e) GIS
    (f) GDSS

15. Critically evaluate the suitability of each of the following DSS types for predicting stock market prices.
    (a) Spreadsheets
    (b) Expert systems
    (c) Neural networks
    (d) OLAP system
    (e) Data mining

In this chapter you will learn to:

- use multimedia systems in an interactive way and to identify how they control the presentation of information
- identify multimedia software appropriate to manipulating particular types of data
- compare and contrast printed and multimedia versions with similar content
- summarise current information technology requirements for multimedia systems
- distinguish between different approaches to animation including path-based and cell-based through practical investigations
- describe the roles and skills of the people who design multimedia systems
- identify participants, data/information and information technology for one example of a multimedia system from each of the major areas
- describe the relationships between participants, data/information and information technology for one example of a multimedia system from each of the major areas
- discuss environmental factors that will influence the design of a multimedia system for a given context, and recommend ways of addressing them
- critically evaluate the effectiveness of a multimedia package within the context for which it has been designed
- interpret developments that have led to multimedia on the World Wide Web
- discuss multimedia systems that address new technological developments
- compare and contrast multimedia presentations
- describe how relevant hardware devices display multimedia and use a variety of devices
- implement features in software that support the displaying of multimedia and explain their use
- use available hardware and software to display multimedia and interact with it
- summarise the techniques for collecting, storing and displaying different forms of media and implement these in practical work
- create samples of the different media types suitable for use in a multimedia display
- describe the process of analog to digital conversion
- plan a multimedia presentation using a storyboard
- diagrammatically represent an existing multimedia presentation with a storyboard
- design and create a multimedia presentation
- combine different media types in authoring software
- design and create a multimedia World Wide Web site that includes text and numbers, hypertext, images, audio and video
- identify standard file formats for various data types
- recommend an appropriate file type for a specific purpose
- describe the compression of audio, image and video data and information
- decide when data compression is required and choose an appropriate technique to compress data and later retrieve it
- capture and digitise analog data such as audio or video
- evaluate and acknowledge all source material in practical work
- use Internet based multimedia presentations in a responsible way
- predict and debate new technological developments based on advancements in multimedia systems
- cross-reference material supplied in multimedia presentations to support its integrity

Which will make you more able to:

- apply and explain an understanding of the nature and function of information technologies to a specific practical situation
- explain and justify the way in which information systems relate to information processes in a specific context
- analyse and describe a system in terms of the information processes involved
- develop solutions for an identified need which address all of the information processes
- evaluate and discuss the effect of information systems on the individual, society and the environment
- demonstrate and explain ethical practice in the use of information systems, technologies and processes
- propose and justify ways in which information systems will meet emerging needs
- justify the selection and use of appropriate resources and tools to effectively develop and manage projects
- assess the ethical implications of selecting and using specific resources and tools, recommends and justifies the choices
- analyse situations, identify needs, propose and then develop solutions
- select, justify and apply methodical approaches to planning, designing or implementing solutions
- implement effective management techniques
- use methods to thoroughly document the development of individual or team projects.

In this chapter you will learn about:

Characteristics of multimedia systems

- multimedia systems - information systems that include combinations of the following media, including:
  - text and numbers
  - audio
  - images and/or animations
  - video
  - hyperlinks
- the differences between print and multimedia, including:
  - different modes of display
  - interactivity and involvement of participants in multimedia systems
  - ease of distribution
  - authority of document
- the demands placed on hardware by multimedia systems, including:
  - primary and secondary storage requirements as a result of:
    - bit depth and the representation of colour data
    - sampling rates for audio data
  - processing as a result of:
    - video data and frame rates
    - image processing, including morphing and distorting
    - animation processing, including tweening
  - display devices as a result of:
    - pixels and resolution
- the variety of fields of expertise required in the development of multimedia applications, including:
  - content providers
  - system designers and project managers
  - those skilled in the collection and editing of each of the media types
  - those skilled in design and layout
  - those with technical skills to support the use of the information technology being used

Examples of multimedia systems

- the major areas of multimedia use, including:
  - education and training
  - leisure and entertainment
  - information provision, such as: information kiosk
  - virtual reality and simulations such as flight simulator
  - combined areas such as educational games
- advances in technology which are influencing multimedia development, such as:
  - increased storage capacity allowing multimedia products to be stored at high resolutions
  - improved bandwith allowing transmission of higher quality multimedia
  - improved resolution of capturing devices
  - increases in processing power of CPUs
  - improved resolution of displays
  - new codecs for handling compression of media while improving quality

Displaying in multimedia systems

- hardware for creating and displaying multimedia, including:
  - screens including CRT displays, LCD displays, Plasma displays and touch screens
  - digital projection devices
  - speakers, sound systems
  - CD, DVD and Video Tape players
  - head-up displays and head-sets
- software for creating and displaying multimedia, including:
  - presentation software
  - software for video processing
  - authoring software
  - animation software
  - web browsers and HTML editors

Other information processes in multimedia systems

- processing:
  - the integration of text and/or number, audio, image and/or video
  - compression and decompression of audio, video and images
  - hypermedia as the linking of different media to another
- organising presentations using different storyboard layouts, including:
  - linear
  - hierarchical
  - non-linear
  - a combination of these
- storing and retrieving:
  - the different file formats used to store different types of data, including:
    - JPEG, GIF, PNG and BMP for images
    - MPG, Quicktime, AVI and WMV for video and animations
    - MP3, Wav, WMA and MID for audio
    - SWF for animations
  - compression and decompression
- collecting:
  - text and numbers in digital format
  - audio, video and images in analog format
  - methods for digitising analog data

Issues related to multimedia systems

- copyright: the acknowledgment of source data and the ease with which digital data can be modified
- appropriate use of the Internet and the widespread application of new developments
- the merging of radio, television, communications and the Internet with the increase and improvements in digitisation
- the integrity of the original source data in educational and other multimedia systems
- current and emerging trends in multimedia systems, such as
  - virtual worlds

6

OPTION 4
MULTIMEDIA SYSTEMS

Multimedia systems combine different types of media into interactive information systems. Due to the significant quantities of data required to deliver image, audio and in particular video efficiently most multimedia systems were distributed on CD-ROM and then DVD, however relatively recent increases in Internet communication speeds and capacities have allowed multimedia presentations to be routinely distributed and viewed over the Internet within web browsers. Today most websites include a combination of text, images and animation and many also include audio and video. The integration of various media into a single presentation is a defining feature of multimedia. Information is more effectively conveyed when different media are combined than is possible using each media type in isolation. Furthermore the interactive nature of most multimedia presentations allows users to explore the content in any order and at their own pace.

Professionally developed multimedia systems require a broad range of expertise. This includes personnel skilled in artistic design, those with expertise to collect each media type to those with technical skills to compress and combine the content into an effective integrated presentation. Project managers supervise the scheduling and allocation of funds to ensure the system is delivered on time and within budget.

Multimedia systems are used to educate, train, entertain or simply to enhance the provision of information. Flight simulators are used to train pilots and computer games are a popular form of escape for many. Schools and universities use a variety of multimedia systems to enhance the learning experiences of students. Information kiosks are dedicated hardware and software systems that provide interactive yet specific information about particular services. There are numerous other examples of multimedia systems and new applications are continually emerging.

The widespread use of multimedia systems is largely a consequence of the ever increasing speed of processing and communication technologies, together with advances in compression and decompression techniques. The result being the ability to deliver higher quality content using smaller file sizes over faster communication links.

We structure our study of multimedia systems under the following broad areas:

- Characteristics of each of the media types
- Hardware for displaying multimedia
- Software for creating and displaying multimedia
- Examples of multimedia systems
- Expertise required during the development of multimedia systems
- Other information processes when designing multimedia systems
- Issues related to multimedia systems

GROUP TASK Investigation

Examine various multimedia systems. Note the different media types included. Can you identify the file formats used to represent each of the media types or is a single format used to integrate the various media types?

CHARACTERISTICS OF EACH OF THE MEDIA TYPES

In this section we consider characteristics of each of the following media types:

- Text and numbers
- Hyperlinks
- Audio
- Images
- Animation
- Video

We review the general nature of each of these media types and describe examples of how each is represented in binary. Due to the large size of raw audio, image, animation and video data the files used to store and distribute these media types are often compressed to reduce file size and transmission times. We describe common compression/decompression techniques for examples of each of these media types.

TEXT AND NUMBERS

Most multimedia systems include significant amounts of text. In many systems most of the information is presented as text and the images, sound, video and other media are used to reinforce the textual information. Other multimedia systems such as games limit the use of text to user instructions. Numbers are less commonly used except as part of the underlying code that controls the presentation. For example when a user selects an option by clicking a command button, choosing a radio button or ticking option boxes the underlying code records the input as a number – often as an integer or Boolean (True/False) value. When numbers are actually displayed on the screen they are often represented as text rather than as numeric values.

The two most commonly used methods for digitally representing text are systems based on ASCII (American Standard Code for Information Interchange) such as Unicode systems, and EBCDIC (Extended Binary Coded Decimal Interchange Code). IBM mainframe and mid-range computers, together with devices that communicate with these machines use EBCDIC. The Unicode system of coding text is used more widely and has become the standard for representing text digitally. Standard ASCII represents the English language characters using decimal numbers in the range 0 to 127 – requiring seven bits per character. For example the decimal number 65 represents ‘A’, which is equivalent to the seven bit binary number 1000001. Unicode systems extend the ASCII character set to include characters from other languages as well as various other special characters.

The number media type is used to represent integers (whole numbers), real numbers (decimals), currency, Boolean (True/False) and also dates and times. Boolean values are represented using a single bit where 1 normally represents True and 0 represents False. Quantities that can be expressed on a numerical scale are represented using numbers. Numbers have magnitude, that is, the concept of size is built into all numbers, for example, ‘15 is bigger than 10 but smaller than 20’. The digits that make up numbers have a place value based on their position within the number, for example the 2 in 123 has a different meaning to the 2 within 2345. These attributes are not present in other types of media. For example, images do not have magnitude and nor does text, to say that a photograph of a bird is greater than one of a building or to say this sentence is greater than the last is meaningless.

In multimedia systems both text and numbers are displayed as images using fonts. Each font describes how each character will be rendered when displayed. There are two broad types of font – outline fonts and raster fonts. More common outline fonts, such as TrueType, describe characters using mathematical descriptions of the lines

and curves within each character. Raster fonts simply store a bitmap of each character. As a consequence outline fonts can be scaled to any size without loss of quality whilst raster fonts become jagged (pixelated) when enlarged. *Fig 6.1* shows large versions of the Times New Roman TrueType outline font together with the raster Courier font. Outline fonts should be used wherever possible, particularly when the display will be printed. Furthermore users with sight impairment often use screen magnifiers that operate best with outline fonts.

Fig 6.1
Outline and raster font example.

It is critical to ensure that fonts used within a multimedia presentation will be available on end-user's machines – in general fonts are installed within the operating system. If the specified font is not available on a user's machine then a different font will be substituted with unpredictable effects on the readability of the display. Some presentation and multimedia authoring software packages include the ability to embed font definitions within the presentation. If this functionality is not available then font selection should be restricted to those included within the target systems.

GROUP TASK Activity
Examine the installed fonts on your computer. Identify examples of outline fonts and examples of raster fonts.

Consider lossless RLE and Huffman compression:

Many compression techniques include Run Length Encoding (RLE) and/or Huffman compression. Both these techniques are examples of lossless compression, meaning no data is lost during compression and subsequent decompression. For text and numbers it is critical that all the original data is retained, whilst for audio, images and video some loss of detail is acceptable in the interests of significantly reducing file sizes and transmission times. Common audio, image and video compression techniques use a combination of lossy and lossless compression techniques whilst text and numbers are compressed using just lossless techniques.

Run Length Encoding (RLE) looks for repeating patterns within the binary data. Rather than including the same bit patterns multiple times the pattern is included just once together with the number of times it occurs. RLE is a simple system used within many compression systems. Let us consider a simple example using the string of text "AAAABBBBBBBBBBCDDDDDDDDD". This string contains a total of 24 characters and would typically be represented using 24 bytes of data – 1 byte per character. Using RLE this string could be encoded as "4A10BC9D" – a total of just 8 characters requiring 8 bytes of storage. In this example the data has been compressed by a factor of 3.

Huffman compression looks for the most commonly occurring bit patterns within the data and replaces these with shorter symbols. For example the text string "ABACBAAB" contains 8 characters often represented using 8-bytes or a total of 64-bits of binary data. In 8-bit ASCII A is 65 or 01000001 in binary, B is 66 or 01000010 and C is 67 or 01000011 in binary. In our example we notice that "A" appears 4 times, "B" appears 3 times and C just once. Using Huffman compression we choose short symbols to represent more common bit patterns. So in our example we could construct a symbol table to represent A as say 0, B as say 01 and leave C as it was. Our 64-bits can therefore be represented using just 4 bits for the As, 6 bits for the Bs

and 8 bits for the C. The data has been compressed from 64 bits down to just 18 bits. Clearly there is some overhead required to store our symbol table, however in real examples this overhead is minor compared to the savings. Huffman compression is used when compressing into ZIP, JPEG and MPEG files.

GROUP TASK Research
Research and identify examples of compressed file formats that use RLE and/or Huffman compression techniques.

HYPERLINKS

The organisation of hypertext and hypermedia is based on hyperlinks. Hypertext is a term used to describe bodies of text that are linked in a non-sequential manner. The related term, hypermedia, is an extension of hypertext to include links to a variety of different media types including image, sound, and video. In everyday usage, particularly in regard to the World Wide Web, the word hypertext has taken on the same meaning as hypermedia.

The user clicks on a hyperlink and is taken to some related content; this new content may also contain hyperlinks to further content. Within multimedia systems hyperlinks are routinely constructed to transfer the user to other parts of the presentation. For example an image of a map of Australia may contain hyperlinks that when clicked take the user to further information about the selected area. Hyperlinks connect related information in complex and often unstructured ways. This organisation allows users to freely explore areas of interest with ease. It closely reflects the operation of the human mind as we discover and explore new associations and detail. Our thoughts move from one association to another; hyperlinks reflect this behaviour.

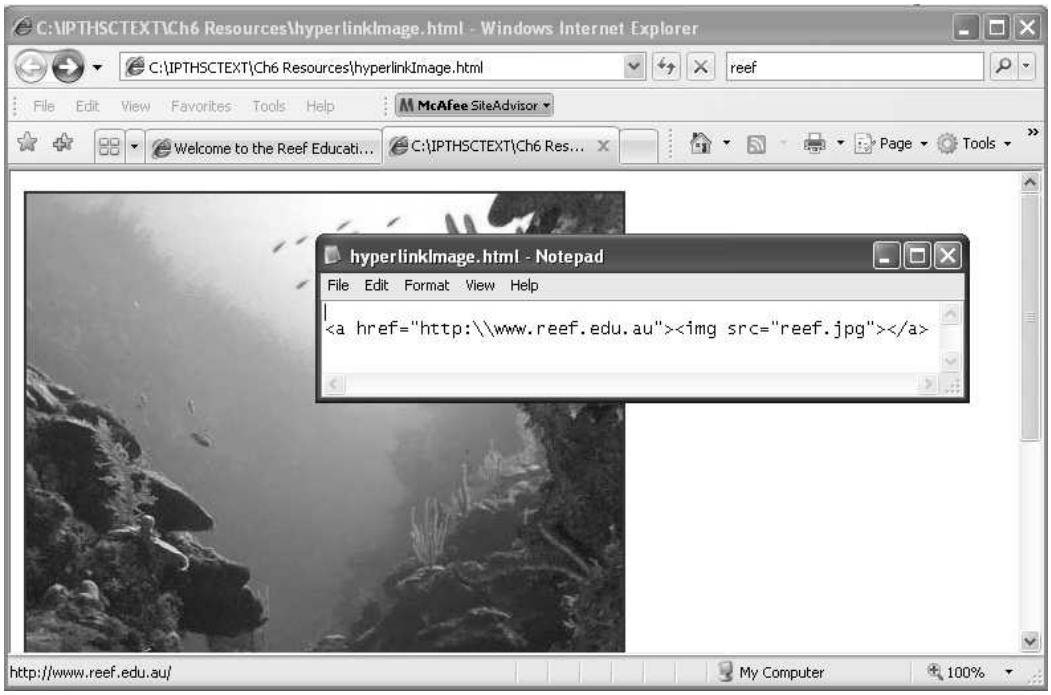


Fig 6.2
Simple HTML image hyperlink example.

Documents accessed via the World Wide Web make extensive use of hyperlinks; these documents are primarily based on HTML. Clicking on a link within an HTML document can take you to a document stored on your local hard drive or to a document stored on virtually any computer throughout the world. From the user's

point of view, the document is just retrieved and displayed in their web browser; the physical location of the source document is irrelevant.

Let us briefly consider HTML tags used to create hyperlinks. For example a hyperlink to the Parramatta Education Centre web site is specified in HTML using:

```
<a href="http://www.pedc.com.au/">Parramatta Education Centre</a>
```

The start tag for a hyperlink commences with `<a` , followed by `href=`, then the URL to the required content within double quotes, and finally the end bracket `>`. Actually more than just the URL can be specified; you can specify a particular HTML document, a particular position within an HTML document or even some other file type such as an image, audio or video file. Following the end of the start tag is the text or image to which the hyperlink is applied; in the above example the text is `Parramatta Education Centre`. The end tag `</a>` finalises the hyperlink. When viewed in a web browser, all text, and any other elements, contained between the start and end tags become the clickable hyperlink. *Fig 6.2* above contains the simple HTML image hyperlink:

```
<a href="http://www.reef.edu.au"><img src="reef.jpg"></a>
```

The HTML file `hyperlinkImage.html` as well as the image `reef.jpg` is stored in a folder on the local hard disk. When this HTML file is opened in a browser the image `reef.jpg` is displayed as a hyperlink. When the image is clicked the browser retrieves and displays the website `www.reef.edu.au`.

In general, HTML documents and also many other documents that contain hypertext are organised as follows:

- All HTML documents are stored as text files.
- Pairs of tags are used to specify hyperlinks and other instructions. Pairs of tags can be nested inside each other.
- Tags are themselves strings of text, they have no meaning until they are analysed and acted upon by software such as web browsers.
- In HTML, tags are specified using angled brackets < >. Text contained within a pair of angled brackets is understood by web-enabled applications to be an instruction; all other text is displayed.
- Web browsers, and other web enabled software applications, understand the meaning of each HTML tag.

GROUP TASK Practical Activity
Create various HTML files containing hyperlinks using both text and images. Alter the hyperlink to point to local and remote files. Experiment by linking to different types of files, such as images, audio and video files. Comment on any problems you encounter.

AUDIO

The audio media type is used to represent sounds; this includes music, speech, sound effects or even a simple 'beep'. All sounds are transmitted through the air as compression waves, vibrations cause the molecules in the air to compress and then decompress, this compression is passed onto further molecules and so the wave travels through the air. Our ear is able to detect these waves and our brain transforms them into what we recognise as sound. The sound waves are the data and what we recognise as sound is the information. File formats for storing audio include MP3, WAV and WMA for sampled sounds and MID which represents individual notes much like a music score.

All waves have two essential components, frequency and amplitude. Frequency is measured in hertz (Hz) and is the number of times per second that a complete wavelength occurs. Sound waves are made up of sine waves where a wavelength is the length of a single complete waveform, that is, a half cycle of high pressure followed by a half cycle of low pressure. In terms of sound, frequency is what determines the pitch that we hear, higher frequencies result in higher pitched sounds and conversely lower frequencies result in lower pitched sounds. The human ear is able to discern frequencies in the range 20 to 20000Hz, for example, middle C has a frequency of around 270Hz.

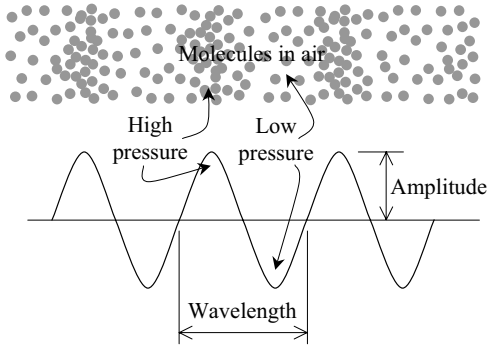


Fig 6.3
Sound is transmitted by compression and decompression of molecules.

Amplitude determines the volume or level of the sound, very low amplitude waves cannot be heard whereas very high amplitude waves can damage hearing. Amplitude is commonly measured in decibels (db). Decibels have no absolute value; rather they must be referenced to some starting point. For example, when used to express the pressure levels of sound waves on the human ear, 0 decibels is usually defined to be the threshold of hearing, that is, only sounds above 0 decibels can be heard, sounds above 120 decibels are likely to cause pain.

Let us now consider how audio or sound data can be represented in binary. There are two methods commonly used, the first is by sampling the actual sound at precise intervals of time and the second is to describe the sound in terms of the properties of each individual note. Sampling is used when a real sound wave is converted into digital, where as descriptions of individual notes is generally used for computer generated sound, particularly musical compositions.

Sampled Audio

The level, or instantaneous amplitude, of the signal is recorded at precise time intervals. This results in a large number of points that can be joined to approximate the shape of the original sound wave. There are two parameters that affect the accuracy and quality of audio samples; the number of samples per second and the number of bits used to represent each of these samples. For example, stereo music stored on compact disks contains 44100 samples for each second of audio for both left and right channels and each of these samples is 16 bits long. This means that an audio track that is 5 minutes long requires storage of 44100 samples × 300 secs × 16 bits per sample × 2 channels, this equates to approximately 50MB of storage. A normal audio CD can hold about 650MB of data, therefore it is possible to store up to around 65 minutes of music on an individual CD. 44100 samples are taken each second because this ensures at least two samples for each wave within the limits of human hearing; remember humans can hear sounds up to frequencies of about 20000Hz, so 40000 samples would ensure at least two samples for all sound waves less than this frequency. Note that the sample rate can also be expressed in hertz, for example 44100 samples per second is equivalent to 44100Hz or 44.1kHz.

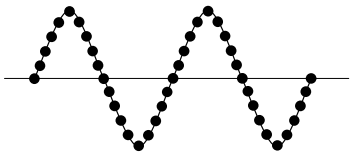


Fig 6.4
Samples are joined to approximate the original sound wave.

It is now common for music and other sound data to be recorded using 6 channels (surround sound), without compression these recordings require three times the storage of a similar stereo recording. Consequently various compression techniques

have been devised to reduce the size of sampled sound data; however greater processing power is required to decompress the sound prior to playback.

Consider MP3 audio files

The Moving Picture Expert Group (MPEG) sets standards for compression of both video and audio. Currently the most popular audio compression format is MP3 – short for MPEG audio layer 3. MP3 files contain compressed sampled audio such that file sizes are reduced by a factor of between 10 and 14, therefore a 50MB file from a CD will compress to an MP3 file of less than 5MB. MP3 is a lossy compression technique meaning that some detail of the original sound is lost during compression. MP3 is designed to remove parts of the sound that will not be noticed by most listeners, hence MP3 files sound very much like the original CD quality sampled sound. Essentially frequencies outside the range of normal human hearing are removed and quiet background sounds imperceptible to most humans are removed. MP3 compression uses complex techniques based on the perceived sound heard by the human ear. Those parts of the music or sound that would not be perceived by the average human ear are removed. The resulting file is then compressed further using lossless compression techniques.

There are many different MP3 compressors for different types of music and sounds. It is the compression process that largely determines the quality of the final MP3. All these compressors produce standard MP3 files that can be decompressed and played on almost any device capable of playing MP3 files.

GROUP TASK Discussion
MP3 files are often ripped from existing audio CDs. Research and discuss the legalities of copying and distributing MP3 files.

Individual Notes

This type of music representation is similar to a traditional music score (see *Fig 6.5*). The vertical position of each note on a music score determines its pitch and the symbol used determines its duration. Different parts of the score are written on their own staff (set of five horizontal lines). Notes vertically above and below each other are played together. Time is indicated horizontally from left to right.

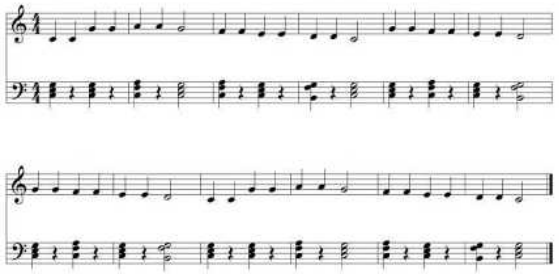



Fig 6.5
Traditional music scores are represented digitally as a series of individual notes.

In binary each note or tone in the music is represented in terms of its pitch (frequency) and its duration (time). Further information for each note can also be specified such as details in regard to how the note starts and ends, and the force with which the note is played. These extra details are used to add expression to each note. Particular instruments can be specified to play each series of notes. The most common storage format for such files is the MIDI (Musical Instruments Digital Interface) format; most digital instruments, including computers, understand this format. Extra files are available that either specify the distinct tonal qualities of a particular instrument or that contain real recordings (digital sound samples) of the instrument playing each note. These files are used in conjunction with the notes to electronically reproduce the

music. Dedicated digital instruments and specialised music software includes actual real recordings whilst most computers simply use generic sounding digital sounds.

GROUP TASK Research
MIDI files can be created using instruments or entirely using software. Research and identify examples of instruments that can collect MIDI data.

IMAGES

The image media type is used to represent data that will be displayed as visual information. Using this definition all information displayed on monitors and printed as hardcopy is ultimately represented as images. All screens and printers are used to display image media, however text and numbers are organised into image data only in preparation for display. Photographs and other types of graphical data are designed specifically for display; this is their main purpose. In these cases the method of representing the image is chosen to best suit the types of processing required. For example, the representation used when editing a photograph to be included in a commercial publication is different to that used when drawing a border around some text in a word processor. There are essentially two different techniques for representing images; bitmap or vector. File formats for storing bitmap images include JPEG, GIF, PNG and BMP. For vector images file formats include SVG, WMF and EMF.

Bitmap

Bitmap images represent each element or dot in the picture separately. These dots or pixels (short for picture element) can each be a different colour and each colour is represented as a binary number. The total number of colours present in an image has a large impact on the overall size of the binary representation. For examples, a black and white image requires only a single bit for each pixel, 1 meaning black and 0 meaning white. For 256 colours, 8 bits are required for each pixel so the image would require 8 times the storage of a similarly sized black and white bitmap image. Most colour images can have up to 16 million different colours, where each pixel is represented using 24 bits. The number of bits per pixel is often referred to as the image's colour or bit depth; the higher the bit depth, the more colours it includes and the larger the storage requirements for the image will be.

Fig 6.6
The resolution of bitmap images should be appropriate to the display device.

The other important parameter in regard to bitmap images is resolution. Resolution determines how clear or detailed the image appears. Resolution is usually expressed in terms of pixel width by pixel height. The image of the Alfa Romeo in *Fig 6.6* has a resolution of 505 pixels by 391 pixels, when the image is enlarged each pixel is merely made larger, for example the jaggy looking grille inset at the top right of the *Fig 6.6* photo. Higher resolution images include more pixels resulting in larger file sizes.

To calculate the uncompressed storage requirements for a bitmap calculate the total number of pixels and then multiply by the colour or bit depth. For example if an image has a resolution of 800 by 600 pixels then the total number of pixels is 480,000. If the bit depth is 24-bits then each pixel requires 3-bytes of storage therefore the total file size in bytes will be 480,000 times 3-bytes per pixel – a total of 1,440,000 bytes.

To convert this figure to kilobytes divide by 1024, so 1,440,000 bytes equals approximately 1406kB. Divide by 1024 again to convert to megabytes, in our example the image requires approximately 1.37MB of storage.

When using bitmap images within multimedia projects it is vital to consider the likely resolution of the end users' display device to determine the most suitable resolution for the bitmap. Typically screens have resolutions ranging from 800 by 600 pixels up to larger widescreen monitors with resolutions of 1920 by 1200 pixels or greater. For screen display there is little point including images with resolutions larger than these sizes. Conversely it is important that images are of sufficient resolution so that will display with sufficient quality.

Bitmap images are often compressed to reduce their size prior to storage or transmission. Many different bitmap image file formats are available; some reduce the size of the image file without altering the image (lossless compression) whilst others alter the image data as part of the compression process (lossy compression).

Consider JPEG image compression

The Joint Photographic Experts Group was the name of the original committee who developed the JPEG specification. JPEG is designed for the compression of realistic natural photographic type images rather than images produced artificially. The JPEG compression technique tends to blur hard edges within artwork, for example the edge of lettering – such sudden colour changes rarely occur in photographs.

JPEG compression aims to reduce file sizes with minimal loss of perceived image quality. To do this requires a basic understanding of how the human eye perceives changes within images. In general changes in brightness or intensity are more noticeable to the human eye than changes in colour, therefore brightness levels should be maintained whilst colour inaccuracies will have less effect on image quality. This is particularly true for blues and to a lesser extent reds. The human eye perceives different greens more accurately than other colours. Therefore degrees of blue and red colour information can be removed during compression with less effect on image quality than brightness and green colour information.

Most raw full colour images are collected by hardware in 24-bit RGB form where each pixel is composed of an 8-bit red component, 8-bit green component and an 8-bit blue component. Most JPEG compression systems first convert the RGB colour representation into a YCbCr representation – Y is the brightness component, Cb stands for chrominance blue and Cr for chrominance red. Each pixel is converted using the following formulas:

$Y = 0.299R + 0.5876G + 0.114B$
$Cb = -0.1687R – 0.3313G + 0.5B + 128$
$Cr = 0.5R – 0.4187G – 0.0813B + 128$

Notice that in the above Y formula the green component has significantly more effect on brightness than the red or blue components. We don't want to lose information from this Y channel. The value of the blue and red components is now largely maintained within the Cb and Cr channels. It is these Cb and Cr channels where we can afford to lose information during JPEG compression.

Once converted to the YCbCr colour system the image is split into a grid of 8 by 8 pixel blocks. Each block is then passed through a complex mathematical process known as Discrete Cosine Transformation (DCT). In simple terms DCT results in a waveform representing the changes in Cb and Cr values. Analysing this wave results in the two chrominance channels Cb and Cr of each pixel within each block being

altered to approximate the values of adjacent pixels. The result being many pixels which have the same or similar Cb and Cr values. These new values can be significantly compressed using standard lossless compression techniques. Note that the Y or brightness level of pixels within the image can also be compressed using lossless compression, however all data in the Y channel remains.

Different levels of compression can be specified within most applications. The application achieves these different levels by altering the range of new Cb and Cr values that can be produced. In most applications JPEG compression is entered as a percentage, for example specifying 90% results in a high quality but large file size whilst 10% creates a small file but a poor quality image. There is no single standard for these percentages – each photo editing application uses its own system.

GROUP TASK Investigation
Load an uncompressed photograph into a photo editor. Save the photo as a JPEG using different levels of compression. Construct a table comparing the level of compression, file size and also your perception of the quality of the image on a scale from poor to excellent.

Vector

Vector images represent each portion of the image mathematically, much like outline fonts. The stored data used to generate the image is a mathematical description of each shape that makes up the final image. Each shape within a vector image is a separate object that can be altered without affecting other objects. For example, a single line within a vector image can be selected and its size, colour, position or any other property altered independent of the rest of the image. For example, the body of the cat in *Fig 6.7* has been drawn using a single filled line whose attributes can be altered independently from the rest of the image.

Fig 6.7
Vector images are represented as separate editable shapes.

The total size of the data required to represent a vector image is, in most cases, less than that for an equivalent bitmap image however the processing needed to transform this data into a visual image is far greater. Vector images can be resized to any required resolution without loss of clarity and without increasing the size of the data used to represent the image. Vector graphics are generally unsuitable for representing photographic images – the detail required is difficult and inefficient to reproduce mathematically.

Consider SVG and WMF/EMF file formats:

Microsoft's window's metafile (WMF) and enhanced metafile (EMF) formats are commonly used vector graphics file formats used within Window's applications. The relatively new scalable vector graphics (SVG) format has been widely accepted as the standard for representing vector graphics on the web. SVG files are text files that include an XML (Extensible Markup Language) description of each of the shapes that form the image. It is likely that all browsers will soon be able to recognise and display SVG images – currently plug-ins are needed to view SVG images in many browsers.

GROUP TASK Activity
Download a simple SVG file and open it within a text editor. Try to make sense of some of the XML code contained within the file.

Consider image distortion and warping:

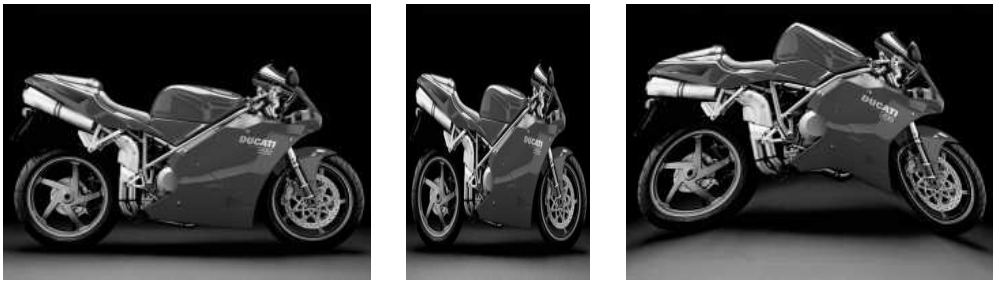


Fig 6.8
Original image (left), distorted version (centre) and warped distorted version (right).

Distorting an image changes the image from its natural shape. This includes bending, twisting, stretching or otherwise altering the proportions of all or part of the image. The term warping is commonly used when the distortion alters parts of an image rather than the entire image. The centre image in *Fig 6.8* has been distorted by altering the proportions of the entire image so that the aspect ratio is changed. The image at right in *Fig 6.8* is best described as a warp as the distortion has been applied to specific parts of the image. Many software applications that produce warps can also produce animations that show the transformation of the original image into the warped version.

GROUP TASK Practical Activity
Use a photo editing or dedicated warping program to distort an image in various ways. If possible produce an animation from the original image to the distorted version.

ANIMATION

Animation is achieved by displaying a sequence of images, known as cels or frames, one after the other. The content of each image is changed slightly from one image to the next. If the images are displayed at a sufficient speed then the human brain merges the images together in such a way that we perceive continuous movement. Commercial feature films display 24 fps (frames per second), however speeds of 12 to 15 fps provide reasonably fluid movement for most simple animations displayed on computer screens. Clearly higher speeds require many more frames and greater storage space and faster transmission speeds.

Prior to computer animation each image was drawn on a sheet of clear celluloid material – the term "cel" (or sometimes cell) is short for celluloid. The clear celluloid allowed a single background image to be reused by overlaying each cell in turn. Furthermore previous cels could be seen through the celluloid as a guide when drawing subsequent cels. The process of placing a series of cels on top of each other is known as "onion skinning" – many animation software applications include an onion skin function that performs the same function electronically.

Traditionally each cel was photographed in turn on film to form an individual frame within the animation. Significant cels, known as key frames, were drawn by the main animator and in between cels were drawn by less experienced animators – this process

is known as "tweening". Automatic tweening is now a function present within most animation software. Key frames are drawn using typical image tools and then the tweening function produces a sequence of intermediate cells that progressively alter the first key frame into the second key frame.

Fig 6.9
Cel-based animation.

Animations are often produced using a combination of cel-based and path-based approaches. Cel-based animation involves creating a sequence of individual cels where each cel is slightly different to the previous cel. For example in *Fig 6.9* walking involves altering the position of the feet, hands and body such that when played the character appears to walk. Cel-based techniques can be used to create the entire animation as a sequence of complete images or it can be used to create small animations of individual characters. For example cel-based techniques can be used to create a library of small animations for each character, say a person walking, sitting down, turning around and so on. These small cel-based animation sequences can then be reused within different parts of the final animation.

Cel-based Animation
A sequence of cels (images) with small changes between each cel. When played the illusion of movement is created.

Path-based Animation
A line (path) is drawn for each character to follow. When played each character moves along their line in front of the background.

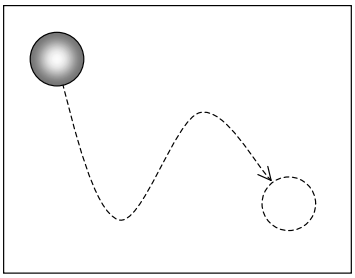


Fig 6.10
Path-based animation.

Path-based animation is used to cause a character to follow a path or line across the background. In most software applications the path the character follows is first drawn as a line (see *Fig 6.10*), the software then creates the animation by causing the character to follow this path across the screen. Characters animated using path-based techniques can themselves be small cel-based animations, such as a character walking, or they can be static images. Most applications allow characters to be rotated, flipped or transformed in various other ways as they follow the path. Professional animation software includes the facility for characters to follow paths in 3 dimensions.

Let us briefly consider two example technologies commonly used to create animations; animated GIF and flash. Animated GIFs are essentially organised as a series of cel-based bitmaps whereas flash organises animation as vectors and can include both cel-based and path-based animation techniques.

Animated GIF

GIF is an acronym for 'Graphics Interchange Format', GIF is a protocol owned and maintained by CompuServe Incorporated. The GIF protocol can be used freely as long as CompuServe is acknowledged as the copyright owner. As a consequence of CompuServe making its specifications freely available GIF files are one of the most commonly used graphic formats on the web. The GIF specification includes the

ability to store multiple bitmap images within a single file, however sound cannot be included and the number of different colours within an individual image is limited to 256. When an animated GIF file is decoded the images are displayed in sequence to create the animation. The GIF specification includes simple compression that is also described within the protocol. The ability to decode all types of GIF files is built into many common software applications, including most web browsers. Many other animation formats and compression methods require their own dedicated software when decoding and decompressing files in preparation for display.

Animation software that produces animated GIF files organise data as a sequence of bitmap images, together with colour palette, timing and various other settings. There are numerous software applications dedicated to the production of animated GIFs; *Fig 6.11* shows the main screen from one such application called 'Easy GIF Animator'. Notice that each cel, or frame, in the animation is shown as a filmstrip down the left hand side of the screen.

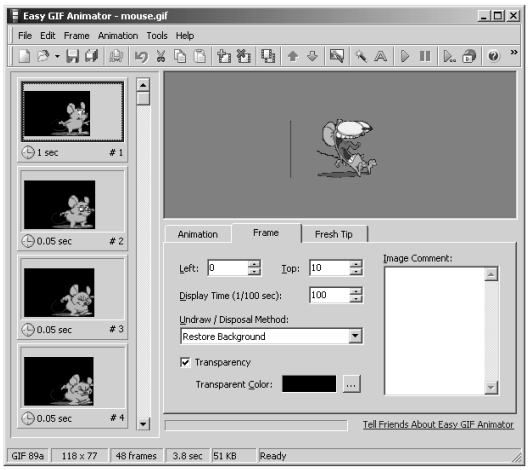


Fig 6.11
Main screen from 'Easy GIF Animator' by Bluementals Software, a Latvian company.

In 'Easy GIF Animator' when a particular frame is selected various properties in regard to the animation can be altered via the frame tab, for example the display time for the frame and a possible transparent colour. The display time is specified in one hundredths of a second and is the time that elapses after a frame has been displayed and prior to the next frame being displayed. Setting a colour as transparent means that when the frame is displayed the background will not be replaced for all pixels of that colour. Each of these properties relates directly to settings specified in the GIF protocol. The transparency check box seen in *Fig 6.11* sets the transparency flag and the colour selected as 'Transparent Colour" sets the transparency index. The transparency index specifies the index of a colour within the colour table. The GIF protocol specifies a colour table as simply a list of RGB colour values; the first set of RGB values being colour 0, the next colour 1, and so on up to the number of colours specified.

GROUP TASK Investigation
Examine the properties of various animated GIF files. Determine the resolution and number of frames used within each of these files.

Flash (SWF file format)

Flash is a standard developed by Macromedia, which is currently owned by Adobe. In early 2000 Macromedia released details of the flash file format (SWF files) to the public, together with details required to play these files. Flash is now an open standard, as a consequence other software development companies are now free to produce applications that create SWF files. For example SWiSH is one such application developed by Sydney software company SwiSHzone.com Pty. Ltd.

All files created with applications based on the flash specifications must be able to play without error in Adobe's Flash player. Studies have shown that more than 96% of Internet users have this player installed; in fact it comes packaged with most operating systems and web browsers. With such a large audience Flash has become

the de facto standard for delivering rich interactive multimedia content that includes animation and sound on the web.

Many websites include Flash animations incorporated within web pages. The web page shown in *Fig 6.12* is composed of a Flash file on the left, which includes animation, together with HTML code for the text down the right hand side. In some cases complete websites are built using flash – particularly those that make extensive use of complex animation.

Fig 6.12
Web page incorporating a flash animation.

Let us consider the organisation of flash data within SWF files, within Macromedia's flash player and finally for display. Flash or SWF files organise data by arranging it into definition tags, control tags and actions; an SWF file is a sequence of such tags and actions. Definition tags are commands to the flash player to create and modify characters; a character is like an actor, prop or even the sound track in a movie, they are elements within the animation that will be displayed. Control tags are used to place instances of these characters on a display list held in memory. The order in which characters reside on the display list determines their order when placed on a frame. For example if a display list has a circle, then a square and then a line added to it in that order, then the circle will be drawn first, followed by the square on top and then the line on top of the square. Portions of the circle covered by the square will not be seen and similarly portions of both the circle and the square covered by the line will not be seen. A special control tag called ShowFrame is used to instruct the flash player to actually create a bitmap of the frame based on the display list; finally the frame is displayed. Creating interactive flash animations involves responding to user input; in flash this is implemented using events and actions; actions occur in response to events such as clicking the mouse. For example an action to restart the animation may occur in response to clicking a button.

In summary, SWF files are organised as a sequence of definition tags, control tags and actions. The flash player reorganises the data into a dictionary based on the definition tags and a display list based on the control tags. Each ShowFrame command encountered causes the current contents of the display list to be reorganised into a bitmap image and then displayed. This method of organisation reduces the size of flash files considerably compared to other animation formats. In most other formats each frame is stored individually within the file rather than being created on the fly as the animation is played.

GROUP TASK Discussion
Compare the organisation of animated GIFs and Flash files with bitmap images and vector images respectively.

Consider morphing:

A morph progressively and smoothly transforms one image into another different image. Flash as well as many other animation software applications are able to produce simple morphs, however more detailed morphs that transform from one photographic image to another require specialised morphing applications. A simple morph may transform a circle into a square, whilst a more complex morph may transform a child's face into the face of their parent or George Bush into Tony Blair as shown in *Fig 6.13*.

Fig 6.13
Morph of George Bush into Tony Blair.

GROUP TASK Practical Activity
Free and shareware versions of morphing software are available. Create an animated morph using one of these software applications.

VIDEO

The video media type combines image and sound data together to create information for humans in the form of movies or animation. Like animation the illusion of movement is created by displaying images or frames one after the other in sequence. Images entering the human eye persist for approximately one twentieth of a second, therefore for humans to perceive smooth movement requires displaying at around 20 images per second – most movies are recorded at 24 frames per second. Video data is composed of multiple images together with an optional sound track. The images and sound must be synchronised for the overall effect to work convincingly.

Motion pictures, as viewed in most cinemas, still use 35mm photographic film to represent the images. Each image or frame measures approximately 35mm wide by 19mm high, hence each second of the movie requires a piece of film 24 × 19mm = 456mm long. Consider the length of film required for a two hour movie; there are 2 × 60min × 60sec = 7200sec in two hours and each second requires 0.456m of film, so the total length for the film is 0.456 × 7200 = 3283.2m or approximately 3.2832km of film.

Let us now consider techniques used to represent video in binary. Like film binary video data is also a sequence of multiple images combined with a sound track. The images, in their raw form, are represented as bitmaps; this results in enormous amounts of data. Consider 1 minute of raw video; if there are 24 frames per second then 1440 frames (24 frames/sec × 60 sec) or bitmaps are needed. If each bitmap has a resolution of 640 by 480 pixels and each pixel is represented using 3 bytes (24 bits) then a single minute of video requires a staggering 1,327,104,000 bytes, or more than 1.2GB of storage (see *Fig 6.14*). Plus we have neglected to include the sound track; the sound track uses sound samples, so if the sound track were recorded at CD quality we'd need to add a further 5MB or so; our total becomes approximately 1.7GB. A two-hour movie, even at this rather meagre resolution, would therefore require some 200 gigabytes of storage. Clearly this data, particularly the images, must be represented more efficiently.

Total Frames	= 24 frames/sec × 60 sec
	= 1440 frames
Data/frame	= 640 × 480 pixels × 3 bytes/pixel
	= 921600 bytes
Total storage	= 1440 frames × 921600 bytes
	= 1327104000 bytes
	= 1327104000 ÷ 1024 kilobytes
	= 1296000 ÷ 1024 megabytes
	= 1265.625 ÷ 1024 gigabytes
	≈ 1.2 gigabytes

Fig 6.14
Calculating the total storage for one minute of raw video image data.

We require an efficient method of compressing and more importantly decompressing the data. Various standards exist for carrying out this process, perhaps the most common being the set of compression standards developed by the Moving Picture Experts Group (MPEG). For example Apple's current Quicktime format (MOV) uses the MPEG 4 standard known as the H.264 codec – codec is short for compression and decompression. Most of the commonly used video formats utilise MPEG standards for compressing and decompressing video, it is the detail of how these techniques are implemented that is different. Video file formats include MPG, MOV, AVI and WMV.

Compressing video involves removing repetitive data and also removing data from parts of images that the human eye does not perceive. Some of these codecs compress data at a ratio of 5 to 1 whilst others can compress by as much as 100 to 1. Compression is somewhat of a balancing act; too much compression and the quality of the video deteriorates, not enough and the size of the file will be too large.

GROUP TASK Activity
A video file with a resolution of 640 by 480 pixels and a bit depth of 24 bits contains 30 seconds of video that plays at a speed of 20 frames per second.. Calculate the approximate size of the file if the codec used has compressed the raw video at a ratio of 25:1.

GROUP TASK Research
The H.264 video codec is used for high definition TV, QuickTime and also for delivering video to 3G devices such as mobile phones. Research and identify reasons why the H.264 codec has such wide acceptance.

Consider block based video compression:

The most common technique used to compress video data is known as 'block based coding'; this technique relies on the fact that most consecutive frames in a sequence of video will be similar in most ways. For example, a sequence of frames where a dog runs across in front of the camera will have a relatively stationary background, that is, the data representing the portions of the background not obscured by the dog is virtually the same for all frames, so why store this data multiple times? Block based coding is the process that implements this idea.

Let us consider a simple block based coding process:

- The current frame is split up into a series of blocks; each block contains a set number of pixels, commonly 16 pixels by 16 pixels.
- The content of each block is then compared with the same block in a past frame.
- If the block in the past frame is determined to be a close match then presumably no motion has taken place in that area of the frame, and a zero vector is stored as an indicator. Vectors indicate direction as well as size of movement, so a zero vector indicates no motion at all.
- Should the blocks not match then other like sized blocks, in the past frame, within the general vicinity of the original block are examined for possible matches (refer *Fig 6.15*). If a match is found then a vector is stored indicating the change in position of the block.
- If no match is found within the search area then the block in the current frame must be stored as a bitmap.

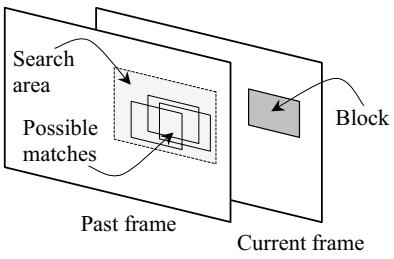


Fig 6.15
Block based coding compares blocks in each frame with those in a similar position on past frames.

Once a complete frame has been coded it is further compressed using various compression techniques commonly used for any binary data. Each frame of data is therefore represented separately but requires that past frames be known before the frame can be reconstructed and displayed. Notice that each frame is still a separate entity including its compression; this means each frame can be decompressed in turn at display time. There is no need to decompress the entire video or to have received the entire file prior to playback commencing.

The first frame, and also other frames at regular intervals, must be stored in their entirety. These are known as key frames. When a user jumps forward or backward within a video the video player must locate a key frame before it can create future frames. For many videos it is unlikely users will perform such actions on a regular basis, however if this is likely to occur then extra key frames should be included.

When video is streamed over the Internet the limiting factor is the speed of the network link. As a consequence many video editing applications allow the user to specify the desired bit rate prior to compressing the video. The application then determines the amount of compression and even the resolution to use during creation of the final movie file.

GROUP TASK Discussion
Most video players first download a reasonable amount of data before playing commences. This is known as buffering. What do you think is the purpose of buffering? Discuss.

SET 6A

1. Higher quality and smaller file sizes for multimedia are largely a result of:
   (A) compression techniques.
   (B) faster processors.
   (C) faster communication links.
   (D) larger storage capacities.

2. The character "D" is represented in ASCII as:
   (A) 1000001
   (B) 1000011
   (C) 1000100
   (D) 1000101

3. An uncompressed bitmap image measures 1000 by 1000 pixels and each pixel can be one of 256 possible colours. What is the approximate storage size of this image file?
   (A) 256kB
   (B) 256kb
   (C) 1MB
   (D) 1Mb

4. Why does JPEG compression represent colour using the YCbCr system rather than RGB?
   (A) Less bits are required per pixel using YCbCr compared to RGB.
   (B) YCbCr has a smaller total palette which in itself reduces the file size.
   (C) Cb and Cr components are less noticeable to the human eye, hence they can be compressed more heavily.
   (D) The Y components is less noticeable to the human eye, hence it can be compressed more heavily.

5. Significant factors that affect the storage size of video files include all of the following EXCEPT:
   (A) resolution.
   (B) fps.
   (C) colour depth.
   (D) bit rate.

6. Which of the following is TRUE with regard to image resolution?
   (A) Images displayed at higher resolution require larger file sizes than when they are displayed at low resolution.
   (B) High resolution bitmaps require larger amounts of storage compared to low resolution bitmaps.
   (C) A low resolution bitmap includes fewer colours than a higher resolution bitmap.
   (D) The resolution of the image determines the size of the displayed image.

7. Which of the following are lossless compression techniques?
   (A) JPEG and MPEG compression.
   (B) RLE and Huffman compression.
   (C) SVG and GIF image compression.
   (D) Sampled audio and scanned images.

8. When animating, what is the process that creates frames between key frames?
   (A) Cel-based animation
   (B) Path-based animation
   (C) Tweening
   (D) Characterisation

9. Which term best describes an animation that transforms one image into another?
   (A) Warp
   (B) Morph
   (C) Distortion
   (D) Transformation

10. A video file contains 10 seconds of footage when played at 12 frames per second. Each frame has a resolution of 320 by 240 pixels and a colour depth of 24-bits. The video file occupies approximately 1.8MB. What is the approximate compression ratio?
    (A) 5:1
    (B) 10:1
    (C) 15:1
    (D) 25:1

11. Briefly explain how computers represent each of the following:
    (a) Text
    (b) Numbers
    (c) HTML hyperlinks
    (d) Sampled audio
    (e) Bitmap images
    (f) Vector images
    (g) animated GIF
    (h) video

12. Compare and contrast each of the following:
    (a) Raster fonts with outline fonts
    (b) Lossless compression with lossy compression.
    (c) MP3 files with MIDI files.
    (d) Cel-based animation with path-based animation.

13. (a) Explain how JPEG images are compressed.
    (b) Explain the process of "block based" video compression.

14. Outline relevant considerations when preparing images for inclusion within multimedia presentations.

15. Lossy compression is often used for image, audio and video but is seldom, if ever, used for other media types. Why is this? Discuss.

HARDWARE FOR CREATING AND DISPLAYING MULTIMEDIA

In this section we examine the operation of common hardware used to display or support the display of multimedia. We consider CRT, LCD, Plasma and touch screens, projectors, and also sound cards and speakers. During display multimedia titles are often retrieved from optical storage, hence we describe the operation of CD-ROM and DVD drives. We also discuss specialised devices, in particular head-up displays and headsets.

SCREENS (OR DISPLAYS)

Information destined for the screen is received by the video system via the system bus. In most applications the video system retrieves this data directly from main memory without direct processing by the CPU. The video system is primarily composed of a video card (or display adapter) and the screen itself. The video card translates the data into a form that can be understood and displayed on the screen.

Video cards (display adapters)

A typical video card contains a processor chip, random access memory chips (often called Video RAM or VRAM) together with a digital to analog converter (DAC). The card shown in *Fig 6.16* receives data via an Advanced Graphics Port (AGP) on the motherboard and transmits video data in either digital or analog form. AGP is a bus standard originally developed by Intel, it allows video cards to directly access main memory independent of the CPU. An AGP bus operates similarly to a PCI bus, but is dedicated to the transmission of video data.

Fig 6.16
Video card with DVI (left) and VGA (right) connectors.

Digital screens are now popular that use the digital visual interface (DVI) standard. DVI video cards are designed to operate with digital screens, using totally digital signals. The DVI standard includes the provision to send both digital and analog signals out of a single DVI connector. Such connectors require a VGA adaptor if the analog outputs on the DVI connector are to be used to connect an analog screen. Many video cards, such as the one in *Fig 6.16*, include a separate VGA connector so that older analog screens can be connected directly to the video card. The DVI standard is also used to connect some widescreen and high definition televisions to set top boxes and DVD players however HDMI (High Definition Multimedia Interface) connectors tend to be used on most non-computer devices.

CRT (cathode ray tube) based monitors

Let us consider the components and operation of a typical cathode ray tube based monitor. The cathode is a device within the CRT that emits rays of electrons. Cathode is really just another name for a negative terminal. The cathode in a CRT is a heated filament that is similar to the filament in a light globe. The anode is a positive terminal; as a result

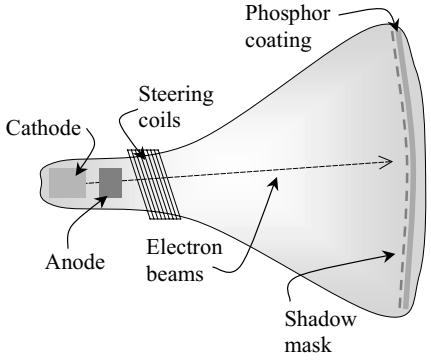


Fig 6.17
Detail of a Cathode Ray Tube (CRT).

electrons rush from the negative cathode to the positive anode. In reality, a series of anodes are used to focus the electron beam accurately and to accelerate the beam towards the screen at the opposite end of the glass vacuum tube. The flat screen at the end of the tube is coated with phosphor. When electrons hit the phosphors they glow for a small amount of time. The glowing phosphors are what we see as the screen image.

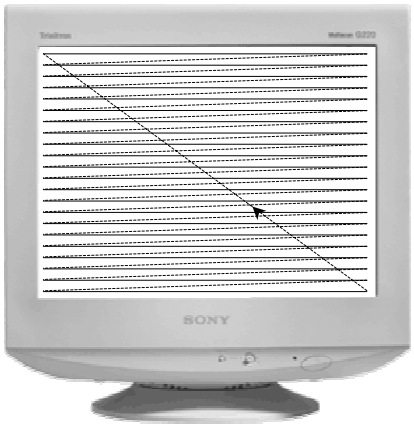


Fig 6.18
The screen is refreshed at least 60 times per second using a raster scan.

To accurately draw an image on the screen requires very precise control of the electron beams. Most CRTs use magnetic steering coils wrapped around the outside of the vacuum tube. By varying the current to these coils the electron beams can be accurately aimed at specific phosphors on the screen. To further increase accuracy a shadow mask is used. This mask has a series of holes through which the electron beam penetrates and strikes the phosphors. There are various types of phosphors that give off different coloured light for different durations. In colour monitors there are groups of phosphors. Each group contains red, green and blue phosphors. When a red dot is required on the screen the red electron gun fires electrons at the red phosphors. To create a white dot all three guns fire. Firing electrons at different intensities allows most monitors to display some 16.8 million different colours.

Colour Depth (Bits per pixel)	Number of colours
1	2 (monochrome)
2	4 (CGA)
4	16 (EGA)
8	256 (VGA)
16	65,536 (High colour)
24	16,777,216 (True colour)

Fig 6.19
Colour depth table showing number of bits required per pixel.

The entire screen is drawn at least 60 times each second; this is known as the refresh rate or frequency and is expressed in Hertz. Each refresh of the screen involves firing the red, green and blue electron beams at each picture element (pixel) on the screen. A screen with a resolution of 1280 by 1024 has approximately 1.3 million pixels to redraw 60 or more times every second. The electron guns fire in a raster pattern commencing with the top row of pixels and moving down one row at a time.

Most CRT monitors are *multisync*, meaning that they can automatically detect and respond to signals with various refresh, resolution and colour-depth settings. The software driver for the video card allows changes to be made to the refresh rate, resolution and colour-depth. Faster refresh rates, increases in resolution or increases in colour-depth require more memory and processing power. Often compromises need to be made between refresh rate, resolution and colour depth to maintain performance at a satisfactory level.

LCD (liquid crystal display) based monitors

Flat panel displays, such as LCD and plasma, have largely replaced CRT based monitors. This is occurring for both computer and television monitors. Currently the most common flat panel technology for computers and also smaller television applications is based on liquid crystals.

GROUP TASK Discussion
LCD monitors have largely replaced CRT monitors. Why do you think this has occurred? Discuss.

Liquid crystals have been used within display devices since the early 1970s. We see them used within digital watches, microwave ovens, telephones, printers, CD players and many other devices. Clearly the technology used to create the LCD panels within these devices is relatively simple compared to that contained within a full colour LCD monitor, however the basic principles are the same. Hence we first consider the operation of a simple single colour LCD panel and then extrapolate these principles to a full colour computer monitor.

So what are liquid crystals? They are substances in a state between liquid and solid, as a consequence they possess some of the properties of a liquid and some of the properties of a solid (or crystal). Each molecule within a liquid crystal is free to move like a liquid, however they remain in alignment to one another just like a solid (see *Fig 6.20*). In fact the liquid crystals used within liquid crystal displays (LCDs) arrange themselves in a regular and predictable manner in response to electrical currents.

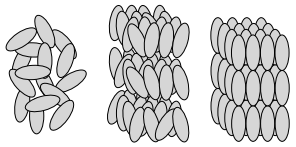


Fig 6.20
The molecules within liquid crystals are in a state between liquids and solids.

LCD based panels and monitors make use of the properties of liquid crystals to alter the polarity of light as it passes through the molecules. The liquid crystal substance is sandwiched between two polarizing panels. A polarizing panel only allows light to enter at a particular angle (or polarity). The two polarizing panels are positioned so their polarities are at right angles to each other. For light to pass through the entire sandwich requires the liquid crystals to alter the polarity of the light 90 degrees so it matches the polarity of the second polarizing panel. Each layer of liquid crystal molecules alters the polarizing angle slightly and uniformly, hence if the correct number of liquid crystal molecule layers are present then the light will pass through unheeded. This is the resting state of LCDs.

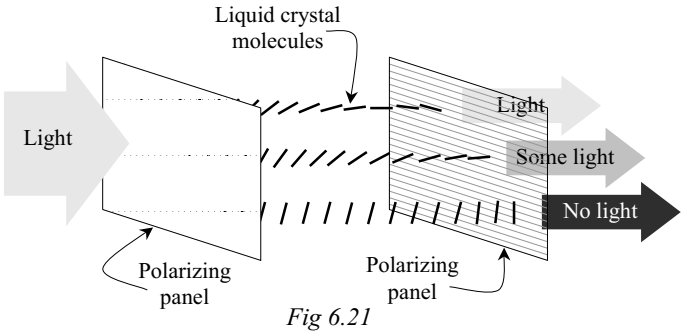


Fig 6.21
The primary components within a LCD.

To display an image requires that light be blocked at certain points. This is achieved by applying an electrical current that causes the liquid crystal molecules to adjust the polarity of the light so it does not match that of the second polarizing panel. Furthermore different electrical currents result in different alignments of the molecules and hence varying intensities of light pass through. In *Fig 6.21* above the first sequence of molecules has no electrical current applied and therefore most of the light passes through. A medium electrical current has been applied to the second sequence of molecules therefore some light passes through. A larger current has been applied to the third molecule sequence, so virtually no light passes through to the final display.

In a CRT monitor light is produced by glowing phosphors, hence no separate light source is required. Within an LCD no light is produced, hence LCD based panels and monitors require a separate light source. For small LCD panels, such as those within microwave ovens and watches, the light within the environment is used. A mirror is installed behind the second polarizing panel, this mirror reflects light from the room back through the panel to your eye. LCD based monitors include small fluorescent lights mounted behind the LCD, the light passes through the LCD to your eye. Such monitors are often called ‘backlit LCDs’.

So how are liquid crystals used to create full colour monitors? Each pixel is composed of a red, green and blue part. A filter containing columns of red, green and blue is contained between the polarizing panels (see *Fig 6.22*). A separate transistor controls the light allowed to pass through each of the three component colours in every pixel.

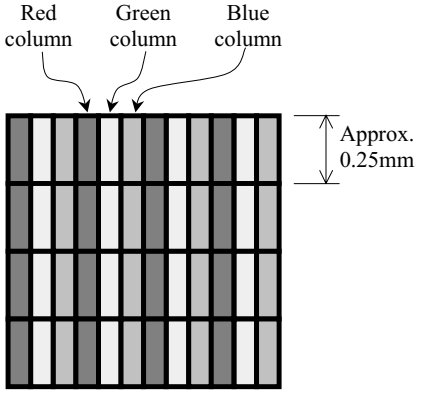


Fig 6.22
Section of the filter within a colour LCD based monitor.

In current LCD screens transistors known as 'Thin Film Transistors' or TFTs are used. A two dimensional grid of connections supplies electrical current to the transistor located at the intersection of a particular column and row. The transistor activates a transparent electrode, which in turn causes electrical current to pass through the liquid crystals (see *Fig 6.23*). However, as each transistor is sent electrical current in turn, usually rows then columns, there is a delay between each transistor receiving current. To counteract this delay storage capacitors are used, each capacitor ensures the electrical current to its transparent electrode is maintained between each pixel refresh.

Consider an LCD monitor that contains 1600 by 1200 pixels – a total of nearly 2 million pixels. Three transistors control each pixel so there is a total of approximately 6 million transistors within this screen. Each of these transistors is refreshed approximately 70 times per second, this means $6 \text{ million} \times 70$ or approximately 420 million transistors are being refreshed each and every second!

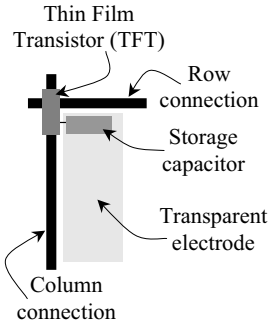


Fig 6.23
Components within each colour of each pixel in a TFT display.

The actual size of each pixel depends on the physical resolution and also the physical size of the screen's viewing area. Screen sizes are traditionally quoted as the diagonal distance across the screen in inches. For example a 17 inch screen at the normal 4:3 aspect ratio actually measures 13.6 inches wide by 10.2 inches high. If this screen contains 1600 by 1200 pixels then there are approximately $1600 \div 13.6 \approx 118$ dpi (dots per inch or pixels per inch). Widescreen monitors use a ratio of 16:9 so a 17 inch widescreen monitor measures 14.8 inches wide by just 8.3 inches high. Due to the different aspect ratio widescreen monitors have there own standard resolutions – 1600 by 900 and 1920 by 1200 being typical examples. A 17 inch widescreen monitor with a physical resolution of 1600 by 900 pixels has approximately $1600 \div 14.8 \approx 108$ dpi.

Because LCD screens contain a precise number of pixels they look best when the resolution of the signal sent from the computer exactly matches the physical resolution of the LCD screen. When lower resolutions are set within the computer the screen must artificially create values for the extra pixels or not use the entire screen. Conversely if a higher resolution signal is sent to a monitor then some detail is lost during display.

GROUP TASK Research
Research the physical resolution of various LCD screens of various sizes. Determine or calculate the number of dots per inch (pixels per inch) that these monitors are able to display.

Plasma Screens

Plasma screens are common within large televisions – currently competing with large LCD screens. Plasma screens, like LCD screens can also be used as computer monitors and also for large advertising displays. In general, LCD screens dominate the computer monitor market, whilst LCD and plasma screens compete in the large wide-screen television market.

A plasma is a state of matter known as an ionised gas. It possesses many of the characteristics of a gas, however technically plasma is a separate state of matter. When a solid is heated sufficiently it turns to a liquid, similarly liquids when heated turn into a gas. Now, when gases are heated sufficiently they form plasma; a fourth state of matter. Plasma is formed as atoms within the gas become excited by the extra heat energy and start to lose electrons. In gases, liquids and solids each atom has a neutral charge, but in a plasma some atoms have lost negatively charged electrons, hence these atoms are positively charged. Therefore plasma contains free-floating electrons, positively charged atoms (ions) and also neutral atoms that haven't lost any electrons. The sun is essentially an enormous ball of plasma and lightning is an enormous electrical discharge that creates a jagged line of plasma – in both cases light (photons) is released. Photons are released as all the negative electrons and positive ions charge around bumping into the neural atoms – each collision causes a photon to be released. In summary, when an electrical charge is applied to a plasma substance it gives off light. Within a plasma screen the gas is a mix of neon and xenon. When an electrical charge is applied this gas forms plasma that gives off ultraviolet (UV) light. We can't see ultraviolet light, however phosphors (like the ones in CRT screens) glow when excited by UV light. This is the underlying science, but how is this science implemented within plasma screens?

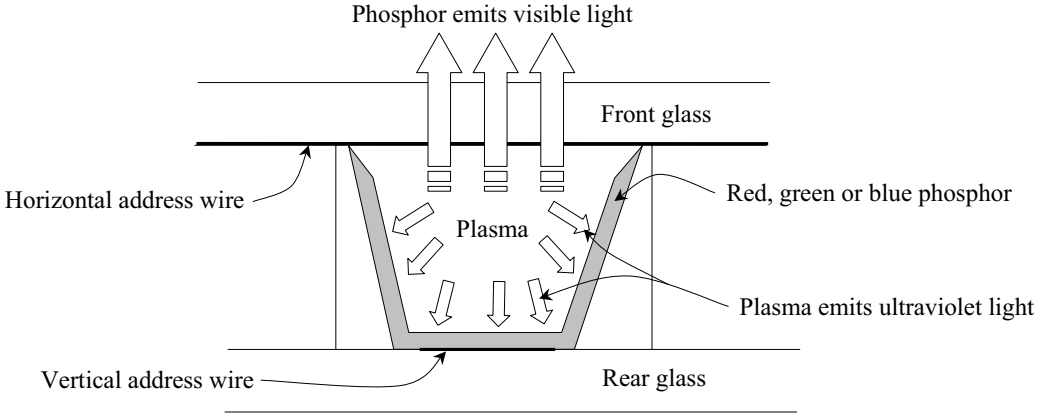


Fig 6.24
Detail of a cell within a plasma screen.

A plasma screen is composed of a two dimensional grid of cells sandwiched between sheets of glass. The grid includes alternating rows of red, green and blue cells – much like a colour LCD screen (refer *Fig 6.22*). Each set of red, green and blue cells forms a pixel. Each cell contains a small amount of neon/xenon gas and is coated in red, green or blue phosphors (refer *Fig 6.24*). Fine address wires run horizontally across the front of the grid of cells and vertically behind the grid. When a circuit is created between a cell's horizontal and vertical address wires electricity flows through the neon/xenon gas and plasma forms within the cell. The plasma emits ultraviolet light, which in turn causes the phosphors to glow and emit visible light. By altering the current passing through the cell the amount of visible light emitted can be altered to create different intensities of light. As with other technologies, the different intensities of red, green and blue light are merged by the human eye to create different colours.

GROUP TASK Discussion
LCD screens require a separate light source, whilst CRT and plasma screens do not. Why is this? Discuss.

Consider the relationship between image and physical screen resolution:

An image is stored with a resolution of 640 by 480 pixels.

- The image appears larger on a small screen that has a small physical resolution than on a larger screen that has a much larger resolution.
- The image appears larger on a large screen with a low physical resolution than on a small screen with a higher physical resolution.
- The image appears larger on one 17 inch screen than it does when viewed on another 17 inch screen.
- Two screens are both set to display 1600 by 900 pixels, however the image is a different size on the two monitors.

GROUP TASK Discussion
How can each of the above dot (pixel) points be explained? Discuss.

Touch Screens

Touch screens are routinely used within ATMs, point of sale terminals, game consoles and also information kiosks. They are also used within tablet computers, PDAs, mobile phones and many other portable devices. A touch screen is both a collection and a display device. Typically touch screens emulate the behaviour of a mouse. Moving your finger across the screen changes the location of the mouse pointer and tapping on the screen corresponds to clicking the mouse button.

Fig 6.25
Information kiosk with integrated touch screen.

The use of touch screens negates the need for a separate keyboard and mouse. This makes them particularly useful devices for installation in public areas where other types of collection device are easily damaged. Furthermore there are no moving parts and the user interface is simpler to use for those who are not familiar with traditional keyboard and mouse input devices. In general touch screen user interfaces should include oversized buttons with space between each button.

There are three major components of all touch screens, the touch sensor panel that overlays the actual screen, a controller that converts signals from the sensor panel into a form suitable for collection usually via a serial or USB port and a software driver so the computer can communicate with the touch panel. There are various different technologies used within touch sensor panels, however in general the sensor panel has an electrical current flowing through it and when the panel is touched this current is interrupted or altered. This change is detected and subsequently used by the controller to determine the location where the touch occurred. In addition most panels are also able to detect pressure. Most touch screens detect just one touch at a time, however multi-touch panels are

available that are capable of detecting the location of simultaneous touch inputs. Touch screens are available as complete units and kits are also available to convert standard CRT and LCD screens into touch screens.

Currently there are three primary technologies used to create touch screen sensors, namely resistive, capacitive and surface acoustic waves (SAW). All three of these technologies are used to determine the coordinates where the touch occurred and also the pressure applied during each touch.

- Resistive sensor panels contain two electrically conductive layers separated by a small gap. One layer contains conductors running vertically and the other has conductors aligned horizontally. When pressure is applied to a point on the screen the outer layer flexes slightly so the gap physically closes between the two layers. This decreases the resistance between the layers and hence an increased electrical current flows at that point.
- Capacitive sensor panels use a single electrically charged panel, usually made of glass with a fine conductive coating. There are sensors located in each corner of the screen that continually and accurately detect the charge present. When a finger touches the screen it absorbs some of the charge. Therefore the charge detected by each of the corner sensors changes slightly. More significant changes occur within sensors closer to the point of contact. As a result the controller can determine the position on the screen where the touch occurred.
- Surface acoustic wave (SAW) touch sensors generate ultrasonic waves that travel from transducers via reflectors and into receivers on the other side. The waves are reflected such that they cover the entire surface of the screen. Generally one pair of transducers and receivers operates horizontally and another pair operates vertically. When the screen is touched the wave is interrupted at that point causing a corresponding change in the received wave pattern.

GROUP TASK Research

Using the Internet or otherwise, research various different systems that include touch screens. In each case explain why a touch screen has been used and identify the technology used within the screen's sensor panel.

DIGITAL PROJECTORS

Digital projectors use a strong light source, usually a high power halogen globe, to project images onto a screen. In this section we consider the operation and technology used within such projectors. There are two basic projection systems; those that use transmissive projection and those that use reflective transmission. Transmissive projectors direct light through a smaller transparent image, whereas reflective projectors reflect light off a smaller image (see *Fig 6.26*). In both cases the final light is then directed through a focusing lens and then onto a large screen.

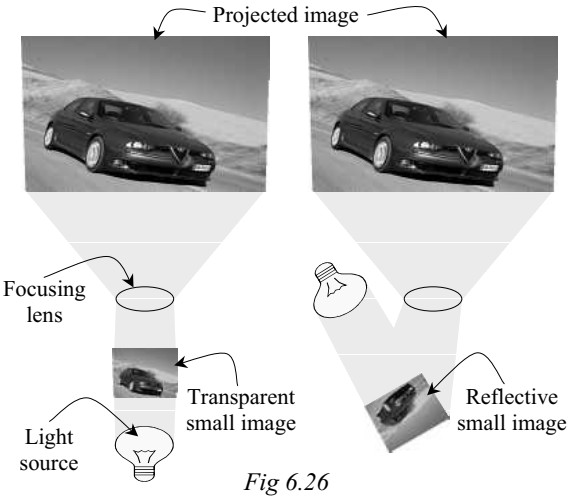


Fig 6.26
Transmissive (left) and reflective (right) projector systems.

Older projector designs are primarily transmissive, the oldest operate similarly to CRTs. Currently CRT based projectors are being phased out, and transmissive LCD projectors are marketed to low-end applications such as home theatre and other personal use systems. For high-end applications, such as conference rooms, board rooms and even cinemas, reflective technologies are common. Let us briefly consider three technologies used to generate the small reflective images within reflective projectors, namely liquid crystal on silicon (LCOS), digital micromirror devices (DMDs) and grating light valves (GLVs).

Fig 6.27
LCOS chip suitable for use in a mobile phone or PDA.

- **LCOS (Liquid Crystal on Silicon)**

Liquid crystal on silicon is essentially a traditional LCD where the transistors controlling each pixel are embedded within a silicon chip underneath the LCD. A mirror is included between the silicon chip and the LCD, hence light travels through the LCD and is reflected off the mirror and back through the LCD to the focusing lens. LCOS chips, such as the one shown in *Fig 6.27*, are also used in devices such as mobile phones and other devices where a small screen is required. For these applications the two polarizing panels are included as an integrated part of the LCOS chip. When used within projectors the polarizing panels are usually independent of the LCOS chips (see *Fig 6.28*). This means the light must only pass through each polarizing panel once on its journey to the screen. LCOS is a relatively new technology that appears to be gaining a larger part of the projector market. Projectors for high quality digital cinema applications use a separate LCOS chip to generate each of the component colours.

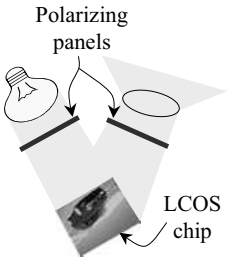


Fig 6.28
Most LCOS based projectors use two independent polarizing panels.

- **DMD (Digital Micromirror Device)**

DMDs are examples of micro-electromechanical (MEM) devices. As the name suggests, DMDs are composed of minute mirrors where each mirror measures just 4 micrometres by 4 micrometres and are spaced approximately 1micrometre apart. Each mirror physically tilts to either reflect light towards the focusing lens or away from the focusing lens. *Fig 6.29* shows just 16 mirrors of a DMD, in reality millions of individual mirrors are present on a single DMD chip (one mirror for each pixel). Each mirror is mounted on its own hinge and is controlled by its own pair of electrodes.

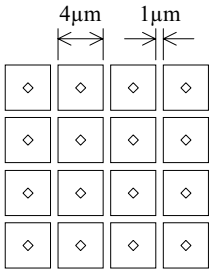


Fig 6.29
DMDs are composed of tilting mirrors.

Dr. Larry Hornbeck at Texas Instruments developed DMD chips and they are produced and marketed by Texas Instrument's DLP™ Products Division. DLP is an abbreviation for "digital light processing", hence DMD based projectors are often known as DLP projectors. Currently DLP projectors are the most popular and widely used of all the projector technologies. To produce a full colour image most DMD projectors include a colour filter wheel between the light source and the DMD as shown in *Fig 6.30*. This wheel

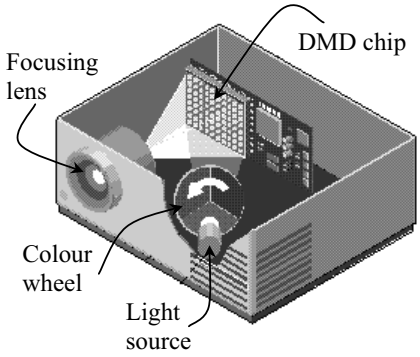


Fig 6.30
Components within a typical DLP projector.

alternates between red, green and blue filters in time with the titling of the mirrors. To produce different intensities of light each mirror is held in its "on" position for varying amounts of time. The human eye is unable to detect such fast changes and hence a consistent image is seen. DMD based projectors currently produce better quality images from lower powered light sources due to their much larger percentage of reflective surface area compared to competing technologies. For example DMD manufacturers currently claim the reflective surface is approximately 89% of the chips surface area compared to LCD devices where the figure is less than 50% of the total surface area. Currently DLP based projectors are available to suit home cinema, classroom, auditorium and even large screen movie theatre applications. The NEC NC2500S in *Fig 6.31* is an example of a DLP based projector designed for use in movie theatres.

Fig 6.31
NEC's NC2500S cinema DLP based projector.

- **GLV (Grating Light Valve)**

Grating light valves were first developed at Stanford University and are currently produced by Silicon Light Machines, a company founded specifically to produce GLV technologies. At the time of writing Sony was developing a high quality GLV based projector for use in cinemas. It is likely this promising technology will also be used within consumer level projectors.

GLVs are another example of a MEMs device. A single GLV element consists of six parallel ribbons coated with a reflective top layer (see *Fig 6.32*). Every second ribbon is an electrical conductor and the surface below the ribbon acts as the common electrode. Applying varying electrical voltages to a ribbon causes the ribbon to deflect towards the common electrode. Hence the light is altered such that it corresponds to the level of voltage applied.

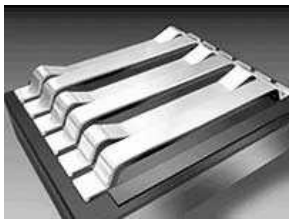


Fig 6.32
A single GLV element.

The major advantage of GLVs is their superior response speed compared to other current technologies. Some GLV chips apparently have response times 1 million times faster than LCDs. This superior response speed allows GLV based projectors to use a single linear array or row of GLVs rather than a 2-dimensional array. For example high definition TV has a resolution of 1920 × 1088 pixels, this resolution can be achieved using a single linear array of 1088 GLV elements, compare this to other technologies that require in excess of 2 million pixel elements. In reality current GLV projectors utilise a separate linear array of GLVs for the red, green and blue components of the image (see *Fig 6.33*). The light source for each GLV linear array being a similar linear array of lasers generating red, green and blue light respectively. The red, green and blue strips of light are combined using a light multiplexer. Finally a rotating mirror directs each strip of light to its precise location on the screen.

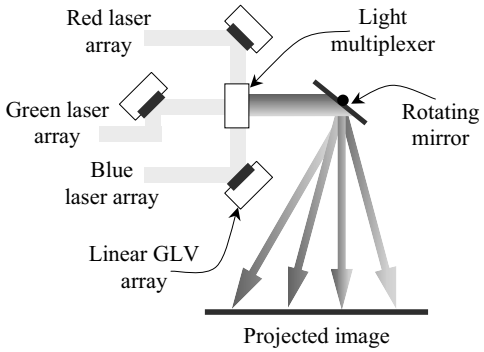


Fig 6.33
Major components of a GLV projector.

GROUP TASK Research
Resolution, brightness, weight, the underlying technology and of course cost are important criteria to consider when purchasing a digital projector. Research and compare examples of currently available digital projectors based on these criteria.

HEAD-UP DISPLAY

Head-up displays, as the name implies, allows the user to keep their head up and looking forward. The display is superimposed on a transparent screen such that the user can view critical information without the need to look down at gauges. This allows the user to concentrate on the real view of the world and at the same time monitor other functions. Without a head-up display the user must look down to read gauges, which involves focusing the eyes on the relatively close gauges and then refocusing again as the look up again. The image projected on head-up displays is designed so the display can be read without the need to refocus.

Fig 6.34
Head-up display within an FA18 Hornet.

Head-up displays have been used within military aircraft and various other military vehicles for many years. In military applications targeting systems utilise head-up displays that superimpose the target area over the actual view. In addition information describing the operation and position of the vehicle can also be displayed. For example in *Fig 6.34* the head-up display within an FA18 fighter jet displays airspeed, altitude and also details of the aircraft's attitude relative to the horizon. The pilot is able to select the specific information displayed to suit their needs at any time.

Fig 6.35
Anaesthesiologist using a head-up display to monitor patient vital signs during surgery.

Head-up displays are available for other applications, such as for motorcycles, racing cars, commercial aircraft, production cars and also for some medical applications. *Fig 6.35* shows an experimental head-up system being used by an anaesthesiologist during surgery and *Fig 6.36* is a head-up display available as an option in current BMW 5 series sedans.

Fig 6.36
Head-up display within a BMW 5 series sedan.

GROUP TASK Research
Head-up displays aim to produce an image that does not interfere with the user's normal view whilst being clearly readable – technically this is very difficult to accomplish. Research current applications of head-up displays to establish how well they achieve this aim.

AUDIO DISPLAY

Digital audio files are first converted to analog signals before being output to speakers. Most computers include a sound card, which is able to perform digital to analog conversion during display processes and also analog to digital conversion during collection of audio data. The processes occurring to display audio are essentially the reverse of the processes occurring during audio collection, therefore many of the components present on sound cards are used during both audio collection and display.

Sound card

Most computers today include the functionality of a sound card embedded on the motherboard, however it is common to add more powerful capabilities through the addition of a separate sound card that attaches to the PCI bus via a PCI expansion slot. In either case similar components are used to perform the actual processing.

In regard to displaying the purpose of a sound card is to convert binary digital audio samples from the CPU into signals suitable for use by speakers and various other audio devices. Most current audio devices, including speakers, require an analog signal, hence we restrict our discussion to the generation of analog audio signals. Analog audio signals are electromagnetic waves composed of alternating electrical currents of varying frequency and amplitude. The frequency determines the pitch and the amplitude determines the volume (we discussed this representation early in this chapter). An alternating current is needed to drive the speakers, as we shall see later.

The sound card receives binary digital audio samples from the CPU via the PCI bus and transforms them into an analog audio signal suitable for driving a speaker. The context diagram in *Fig 6.37* models this process. On the surface it would seem a simple digital to analog converter (DAC) could perform this conversion. In reality audio data is time sensitive, meaning it must be displayed in real time. To achieve real time display sound cards contain their own RAM which is essentially a buffer between the received data and the card's digital signal processor (DSP). The DSP performs a variety of tasks including decompressing and smoothing the sound samples. The DSP then feeds the final individual samples in real time to a DAC. The DAC performs the final conversion of each sample into a continuous analog signal.

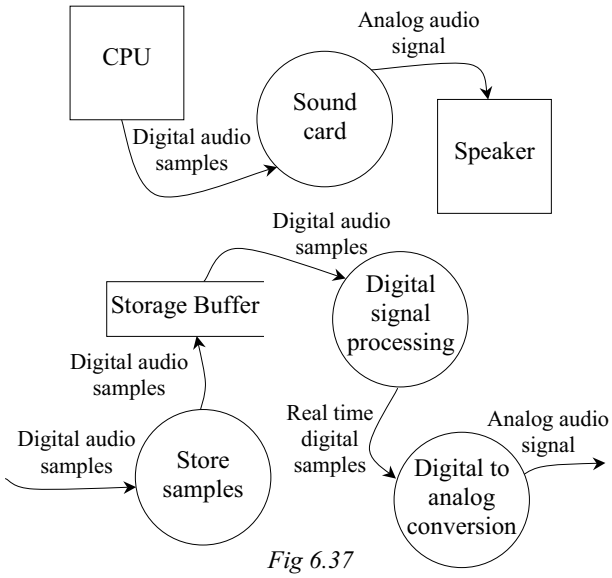


Fig 6.37
A sound card's display processes modelled using a context and dataflow diagram.

The analog signal produced by the sound card's DAC has insufficient power (both voltage and current) to drive speakers directly. This low power signal is usually output directly through a line out connector and a higher-powered or amplified signal is output via a speaker connector. Obviously the line out connector is used to connect display devices that include their own amplifiers, such as stereo and surround sound systems.

Speakers

Speakers are analog devices that convert an alternating current into sound waves. Sound waves are compression waves that travel through the air. An electromagnet is the essential component that performs the conversion into sound waves. Essentially an electromagnet is a coil of wire surrounded by a magnet. As current is applied to the coil it moves in and out in response to the changing magnetic fields. As an alternating current is used to drive the speaker the coil vibrates in time with the fluctuations present within the alternating current. The coil is attached to a paper diaphragm, it is the diaphragm that compresses and decompresses the air forming the final sound waves. The coil and diaphragm are held in the correct position within the magnet using a paper support known as a 'suspension spider'.

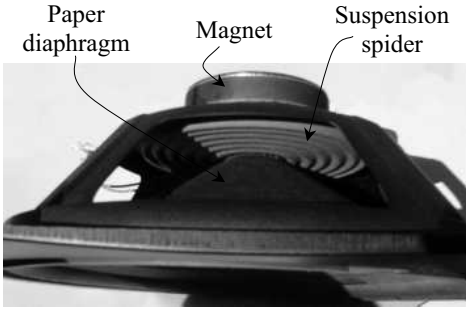


Fig 6.38
Underside of a typical speaker.

The size of the diaphragm in combination with the coils range of movement determines the accuracy with which different frequencies can be reproduced. Large diameter diaphragms coupled with coils that are able to move in and out over a larger range are suited to low frequencies (0Hz to about 500Hz). Such speakers are commonly used within woofers. Smaller diameter diaphragms are tighter and hence respond more accurately to higher frequencies. Speakers with very small diameter diaphragms respond to just the higher frequencies and are known as tweeters. Commonly speaker systems include a separate low frequency woofer or sub-woofer, combined with a number of speakers capable of producing all but the lowest frequencies. Just a single large woofer is sufficient as low frequency sound waves are omnidirectional, that is they can be heard in all directions. Conversely high frequency sounds from say 6000Hz up to 20000Hz are very directional, hence tweeters need to be arranged to produce sound in the direction of the listener.

GROUP TASK Research
Most sound cards include a variety of different input and output ports – some digital and some analog. Examine the audio ports on your school or home computer and determine the nature of the data input or output from each of these ports.

Head-sets

Head-sets integrate a microphone and speakers into a single device worn on the head. Analog head-sets such as the one in *Fig 6.39* connect to analog inputs and outputs. If the headset is connected to a computer then the plugs connect to analog ports on the sound card. Digital versions are now available that connect to USB ports or operate wirelessly, such as the Bluetooth version in *Fig 6.40*.

Head-sets are routinely used in conjunction with telephone systems, particularly for users who spend extended periods of time on the phone. Because the microphone is

close to the user's mouth the amount of external noise introduced into the microphone is greatly reduced. This means lower quality microphones can be used more effectively. In addition the use of head phones means that feedback from the speakers back into the microphone is virtually eliminated. Furthermore users can be in close proximity to each other without the sound interrupting adjoining users. Many multimedia systems, in particular games, make extensive use of music and sound effects that can be distracting for people close by. The use of a head-set allows the user to immerse themselves in sound without interrupting others.

Fig 6.39
Analog head-set including stereo speakers

Audio visual (AV) head sets are available that add video and images to sound. Some contain a single screen viewed by just one eye, whilst others include two screens. Three-dimensional images are possible when the software sends different images to the left and right hand screens. Commonly these AV headsets are used to view traditional movies and music videos rather than interactive content.

Fig 6.40
Bluetooth head-set designed for use with a mobile phone.

Consider Virtual Reality:

Virtual reality (VR) head-sets add sensors to monitor the position of the users head. This allows the displayed images to move fluidly as the user looks around. The user feels totally immersed in the action as they explore virtual worlds. Older virtual reality headsets where large and heavy, more recent models are much smaller and lighter yet their screens are much higher resolution. Icuiti's VR920 shown in *Fig 6.41* includes two 640 by 480 pixel LCD screens, together with stereo speakers, microphone and head tracking sensors.

Fig 6.41
Icuiti's VR920 virtual reality headset.

Although most VR head-sets are designed for gaming they are also finding applications in other areas, such as design, scientific and medical fields. For example patients when undergoing lengthy or painful procedures have been found to experience less discomfort when immersed in a virtual world.

Additional devices attached to virtual reality systems include gloves and even complete body suits. Such devices not only collect the users movements but some also include pressure devices so that virtual objects can be manipulated and felt. Currently devices that provide touch feedback are largely experimental. Simple examples in widespread use include vibrating batteries, force feedback game controllers and steering wheels.

GROUP TASK Research
Using the Internet or otherwise research current examples of virtual reality systems and their applications.

OPTICAL STORAGE

CD-ROM and DVDs store digital data as a spiral track composed of pits and lands. We discussed the nature of the pits and lands back in chapter 2. The single track on a CD-ROM is able to store up to 680 megabytes of data. DVDs contain similar but much more densely packed tracks so each track can store up to 4.7 gigabytes of data. DVDs can be double sided and they can also be dual layered. Therefore a double sided, dual layer DVD would contain a total of four spiral tracks; in total up to 17 gigabytes of data can be stored. Such large amounts of storage make optical disks well suited to the storage and distribution of multimedia software and data.

Retrieving data from an optical disk can be split into two processes; spinning the disk as the read head assembly is moved in or out to the required data and actually reading the reflected light and translating it into an electrical signal representing the original sequence of bits. To structure our discussion we consider each of these processes separately, although in reality both occur at the same time.

Spinning the disk and moving the read head assembly

To read data off an optical disk requires two motors, a spindle motor to spin the disk and another to move the laser in or out so that the required data passes above the laser. The spindle assembly contains the spindle motor together with a clamping system that ensures the disk rotates with minimal wobble. The read head assembly is mounted on a carriage, which moves in and out on a pair of rails. In modern optical drives the motor that moves the carriage responds to tracking information returned by the read head. This feedback allows the carriage to move relative to the actual location of the data track.

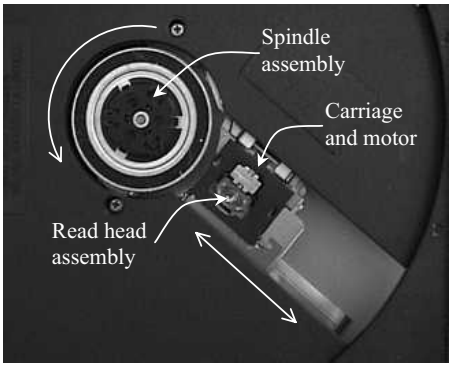


Fig 6.42
Detail of a CD/DVD drive from a laptop computer.

At a constant number of revolutions per minute (rpm) the outside of a disk rotates much faster than the inside. Older CD drives, and in particular audio CD drives, reduce the speed of the spindle motor as the read head moves outwards and increase speed as the read head moves inwards. For example a quad speed drive spins at 2120 rpm when reading the inner part of the track and at only 800 rpm when reading the outer part. The aim being to ensure approximately the same amount of data passes under the read head every second; drives based on this technology are known as CLV (constant linear velocity) drives.

Most CD and DVD drives manufactured since 1998 use a constant angular velocity (CAV) system, which simply means the spindle motor rotates at a steady speed. CLV technology is still used within most audio drives, which makes sense, as there really is no point retrieving such data at faster speeds. However for computer applications, such as installing software applications or viewing video faster retrieval is definitely an advantage. As a consequence of CAV such drives have variable rates of data transfer, for example a 24-speed CAV CD drive can retrieve some 1.8 megabytes per second at the centre and 3.6 megabytes per second at the outside. Quoted retrieval

speeds for CAV drives are often misleading; for example a CAV drive designated as 48-speed can only retrieve data from the outside of a disk at 48 times that required for normal CD audio. These maximum speeds are rarely achieved as very few CDs have data stored on their outer edges.

Current CAV drives have spindle speeds in excess of 12000 rpm; faster than most hard disk drives. Such high speeds produce air turbulence resulting in vibration. When most drives are operating the noise produced by this turbulence can be clearly heard. Furthermore the vibration is worst at the outside of the disk, just where the data passes under the read head at the fastest speed, hence read errors do occur.

Reading and translating reflected light into electrical signals

There are various different techniques used to create, focus and then collect and convert the reflected light into electrical signals. Our discussion concentrates on the most commonly used techniques.

Let us follow the path taken by the light as it leaves the laser, reflects off the pits and lands, and finally arrives at the opto-electrical cell (refer to *Fig 6.43*). Firstly, lasers generate a single parallel beam. This beam passes through a diffraction grating whose purpose is to create two extra side beams; these side or tracking beams are used to ensure the main beam tracks accurately over the pits and lands. Unfortunately the diffraction grating causes dispersion of the beams. To correct this dispersion the three beams pass through a collimator lens; whose job is to make the beams parallel to each other. A final lens is used to precisely focus the beams on the reflective surface of the disk.

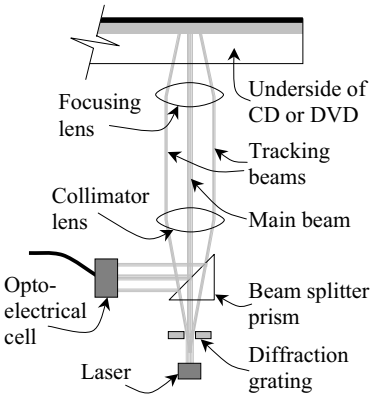


Fig 6.43
Detail of a typical optical storage read head.

As the disk spins both tracking beams should return a constant amount of light, as they are reflecting off the smooth surface between tracks (see *Fig 6.44*). If this is not the case then the carriage containing the read assembly is moved ever so slightly until constant reflection is achieved. In essence the tracking beams are used to generate the feedback controlling the operation of the motor that moves the read head in and out.

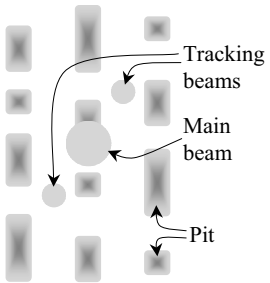


Fig 6.44
Magnified view of main and tracking laser beams.

The reflected light returns back through the focussing and collimator lenses and then is reflected by a prism onto an opto-electrical cell. The prism is able to split the light beam based on its direction; light from the laser passes through, whereas light returning from the disk is reflected. The term 'Opto-electrical' describes the function of the cell; it converts optical data into electrical signals. Changes in the level of light hitting the cell cause a corresponding increase in the output current. Constant light causes a constant current. Hence the fluctuations in the electrical signal correspond to the stored sequence of bits. No change in light entering the cell indicates a zero whilst a change in reflected light indicates a one as a transition from pit to land or land to pit occurs.

The stored binary data on both CDs and DVDs is encoded so that long sequences of either ones or zeros cannot occur. Tracking problems would result when the pits or lands are too long, as would occur when a large number of zeros are in sequence. The

distance between pits and lands would be too small to be reliably read when many ones appear in sequence. The solution is to avoid such bit patterns occurring in the first place. The eight to fourteen modulation (EFM) coding system is used; EFM converts each eight-bit byte into fourteen bits such that all the bit patterns include at least two but less than ten consecutive zeros. This avoids such problems occurring within a byte of data, but what about between bytes? For example the two bytes 10001010 and 11011000 convert using the EFM coding system to 1001001000001 and 01001000010001. When placed together the transition between the two coded bytes is …0101…; our rule of having at least two zeros is broken. To correct this problem two merge bits are placed between each coded byte; the value of these merge bits is chosen to maintain our at least two zeros but less than ten rule.

The electrical signal from the opto-electrical cell is then passed through a digital signal processor (DSP). The DSP removes the merge bits, converts the EFM codes back into their original bytes and checks the data for errors. Finally the data is placed into the drive's buffer where it is retrieved via an interface to the computer's RAM.

GROUP TASK Discussion
When viewing a video file a user notices that the drive light flashes indicating the drive is stopping and starting, yet the video plays smoothly. How can this be explained? Discuss.

GROUP TASK Discussion
Until recently it was common for the CD or DVD to be within the optical drive during execution of multimedia titles. Although this still occurs, it is becoming less common. Indeed many multimedia titles are now accessed directly via the Internet. Discuss reasons to explain these changes.

HSC style question:

(a) Identify the hardware and the processing occurring as a video is retrieved from CD-ROM and displayed on an LCD screen.

(b) Define the term resolution and describe its effect on the storage and display of images.

Suggested Solution

(a) On CD-ROM the video is stored as a sequence of pits and lands on the spiral track of the CD. During retrieval bits are read at regular time intervals. Transitions from pit to land are read as binary ones whilst no transition is read as a binary zero. This data is decoded from its EFM representation within the CD drive and is sent on to RAM.

From RAM the video is decompressed by the CPU into individual frames – most video files are decompressed using a block-based codec such as MPEG. Each frame is then sent to the video system where the video card renders the frame into a bitmap suitable for display.

Each rendered frame is sent from the video card to the screen as a sequence of individual pixel data composed of a red, green and blue value. LCD screens contain three thin film transistors (TFTs) for each pixel corresponding to red, green and blue. The current received by each TFT changes the polarity of the LCD crystals, which in turn causes varying amounts of light to pass through the

screen at that point. As each new frame in the video is displayed in sequence the illusion of movement is created.

(b) Resolution is a measure of the quality of a displayed image. It describes the width and height of the image or screen in pixels. Higher resolution images contain more pixels than lower resolution images. The higher the number of pixels in the image the better the quality of the image and the less pixelated it appears. This affects the storage as higher resolutions require greater file sizes to store the extra pixel data. Images that are intended for screen display require far lower resolution than those destined for printing – the number of physical dots on a screen is significantly less than is produced by printers.

Comments

- In an HSC or Trial HSC examination part (a) would likely attract 4 marks and part (b) would attract 3 marks.
- In part (a) the solution could have described the decompression process in more detail. For example a brief explanation of block-based decoding, such as "The video data includes key frames that are complete bitmaps of the image to be displayed. Subsequent frame data describes just the changes that have occurred from the previous key frame rather than detailing all pixels."
- In part (a) the interface between the video card and the LCD screen could be digital (DVI) or it could be analog (VGA). If analog then the signal is converted to analog by the video card and then converted back to digital by the LCD screen.
- In part (a) mention of the red, green, blue filter covering the TFTs would enhance the solution. In addition the TFTs receive a binary value ranging from 0 to 255 – the above solution implies an analog varying current signal is received.

Consider Microsoft Surface:

During 2007 Microsoft released Microsoft Surface™. The following information is reproduced from Microsoft's 2007 press release:

Microsoft Surface

Product Overview:

Surface is the first commercially available surface computer from Microsoft Corp. It turns an ordinary tabletop into a vibrant, interactive surface. The product provides effortless interaction with digital content through natural gestures, touch and physical objects. In essence, it's a surface that comes to life for exploring, learning, sharing, creating, buying and much more. Soon to be available in restaurants, hotels, retail establishments and public entertainment venues, this experience will transform the way people shop, dine, entertain and live.

Description:

Surface is a 30-inch display in a table-like form factor that's easy for individuals or small groups to interact with in a way that feels familiar, just like in the real world. Surface can simultaneously recognize dozens and dozens of movements such as touch, gestures and actual unique objects that have identification tags similar to bar codes.

GROUP TASK Research

Research Surface computing. Identify the hardware and software used and briefly describe applications where Surface computing is used.

SET 6B

1. Screens that receive analog signals commonly connect to which of the following?
   (A) VGA connector.
   (B) DVI connector.
   (C) HDMI connector.
   (D) USB connector.

2. A device that projects a transparent image over the real view of the world so the user need not change their focus is known as a:
   (A) head set.
   (B) virtual reality system.
   (C) head-up display.
   (D) visual display unit.

3. How is the image on an LCD screen maintained between screen refreshes?
   (A) The phosphors glow for a period of time sufficient to maintain the image between refreshes.
   (B) The liquid crystals remain in alignment between screen refreshes.
   (C) A filter ensures the image remains stable whilst the screen is refreshed.
   (D) Each pixel has its own capacitor that holds the electrical current between screen refreshes.

4. A touch panel flexes slightly as it is touched. The underlying technology is most likely which of the following?
   (A) Resistive
   (B) Capacitive
   (C) SAW
   (D) Transitive

5. DLP projectors form images using which of the following?
   (A) Small LCD screens.
   (B) Miniature tilting mirrors.
   (C) Tiny reflective ribbons.
   (D) Transmissive CRT technology.

6. The volume of sound waves is determined by their:
   (A) frequency.
   (B) wavelength.
   (C) bit depth.
   (D) amplitude.

7. Which of the following produces the light illuminating an LCD screen?
   (A) Liquid crystals.
   (B) Polarising panels.
   (C) Phosphor coating.
   (D) Fluorescent tube.

8. Speakers perform which of the following conversions?
   (A) Digital signal to analog sound wave.
   (B) Analog signal to digital sound wave.
   (C) Analog signal to analog sound wave.
   (D) Digital signal to digital sound wave.

9. Most current optical drives use a CAV system. A consequence of this technology is:
   (A) data located further from the centre of the disk is read more rapidly.
   (B) data is read at a constant rate regardless of its location on the disk.
   (C) more data can be stored on disks with the same size and density.
   (D) the spindle motor must alter its speed depending on the current position of the read head.

10. What is purpose of EFM encoding of data on optical disks?
    (A) To correct read errors efficiently as the data is being read.
    (B) So that transitions between pits and lands can be used to represent binary digits.
    (C) To avoid long pits and lands which are difficult to read reliably.
    (D) To convert each byte of data into fourteen bits.

11. Define each of the following terms.
    (a) Video card
    (b) Liquid crystal
    (c) Polarising panel
    (d) TFT
    (e) Touch screen
    (f) plasma
    (g) DMD
    (h) Head-up display
    (i) Head-set

12. Explain how each of the following devices displays images:
    (a) CRT monitor
    (b) LCD monitor
    (c) DLP projector
    (d) GLV based projector
    (e) Plasma display

13. Identify the components and describe the processes occurring as sampled audio is played through a computer's speakers.

14. Explain the processes' occurring as data is read from an optical disk.

15. Distinguish between virtual reality head-sets and other types of head-sets. Include examples to illustrate the differences.

SOFTWARE FOR CREATING AND DISPLAYING MULTIMEDIA

In this section we examine software applications used to create and then display multimedia. These software applications are able to combine different media types into a single multimedia presentation. There is a broad and diverse range of such software applications available so in this section we can only hope to outline the functionality within some common examples.

We shall examine examples of each of the following categories:

- Presentation software
- Applications such as word processors with sound and video
- Authoring software.
- Animation software.
- Web browsers and HTML editors.

PRESENTATION SOFTWARE

Presentation software is used to produce high quality multimedia presentations designed for display to groups of participants. Commonly such presentations are in the form of a slide show where each slide supports a talk given by a presenter. Presentations can also be printed, uploaded to a website or stored on CD or DVD for display at other times.

Most presentation applications use templates or themes that specify the format and overall design of the slides. Media of all types can be entered or imported into individual slides. Animation can be created to improve the presentation. For example text can float in from the side and different transitions can be used to animate the change from one slide to the next.

Consider the following examples of presentation software:

Apple's iWork Keynote

Keynote includes an extensive collection of 3 dimensional transitions and effects. It also includes "spreadsheet like" tables that can be used as the source for producing charts and graphs. Keynote is able to produce high-resolution output suitable for display on large high definition projectors and screens.

Fig 6.45
Screenshot from Apple's iWork Keynote presentation software for Mac computers.

Microsoft's PowerPoint

PowerPoint is the presentation software included within the Microsoft Office suite of integrated applications. It is currently the most widely used presentation software application. A master slide is used to specify general formatting and design for each slide in the presentation. Like other presentation software, PowerPoint is able to import a wide variety of media types in a wide range of formats. Versions of PowerPoint are available for both Windows and Mac operating systems.

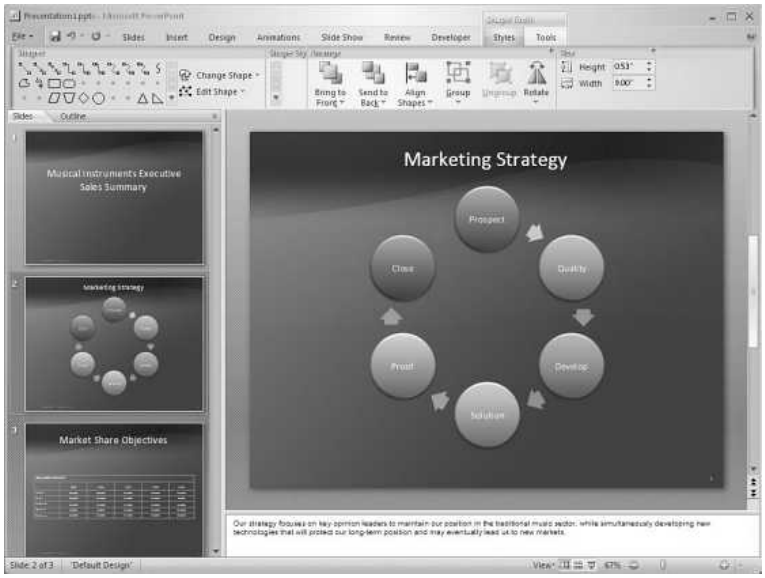


Fig 6.46
Microsoft's PowerPoint presentation software.

OpenOffice.org's Impress

OpenOffice is a suite of integrated software applications – Impress is the presentation software application. OpenOffice is an open source product and can be downloaded and used free of charge. Impress operates similarly to other presentation software and is able to save and open PowerPoint files. In addition, Impress is able to create Flash files (SWF) of presentations that can be distributed via the web for viewing in Adobe Macromedia's popular flash player.

Fig 6.47
OpenOffice.org's Impress presentation software.

GROUP TASK Research
Using the Internet find and read reviews of current presentation software applications. Identify features that differentiate between each product reviewed.

GROUP TASK Practical Activity
Use a presentation software application to produce a simple slide show describing your findings from the above research tasks. Include text, images and at least one video within the presentation.

APPLICATIONS SUCH AS WORD PROCESSORS WITH SOUND AND VIDEO

Many software applications are able to combine media from a variety of sources. For example an image can be included within a text document, a chart created in a spreadsheet can be included in a word processor document, even sound and video files can be linked or embedded within files produced by a variety of applications.

In this section we consider embedding and linking. These are the two commonly used techniques for combining information of different types and from different sources.

Embedding

In many applications it is possible to import files created within a variety of other applications into an existing file. The existing file is known as the "destination file" and the file being imported is known as the "source file". This process is known as embedding as information within the source file becomes part of the destination file. In essence a copy of the source file is inserted into the destination file. For example the paste command within most applications embeds a copy of the information currently on the clipboard into the current document. The current document is the destination file and the content of the clipboard is the source.

Once the embedding process is complete there is no connection maintained between the original source file and the destination file. The effect being that any future changes made to the original source file will not be reflected within the destination file. The embedded data can be edited from within the destination file using either the same software application or a similar software application to that used to create the original source file. The size of the destination file increases to reflect the additional storage required to store the embedded content.

Linking

Linking does not make a copy of the source file, rather it establishes a connection within the destination file to the source file. Therefore any alterations that are made to the original source file will automatically be reflected within the destination file. For example a linked spreadsheet within a word processor file will automatically be updated to reflect any alterations made within the source spreadsheet. Linking is used when the most current version of the source data needs to be displayed. Furthermore when linking it is possible for many destination files to link to the same source file. This is common practice when many users require access to current data within the same file and also within most websites where the same source image is routinely used on multiple web pages.

HTML hyperlinks are an example of linking. The HTML document is the destination document that contains tags that specify the location and name of the linked source files. In addition to web pages most word processors and many other applications are also able to include such hyperlinks. When a user clicks on a hyperlink the application responds by retrieving and displaying the linked source file.

Consider the following:

A word processor file called WP.doc contains an imported bitmap image. The source image file is called Image.jpg. Consider whether the image has been embedded or linked for each of the following:

- Image.jpg is edited but it does not change within WP.doc.
- Image.jpg is edited and the changes are seen within WP.doc.
- The image is opened and edited within WP.doc. Later Image,jpg is found to have also changed.
- The image is opened and edited within WP.doc, however when Image,jpg is opened it has not changed.

GROUP TASK Discussion
Propose scenarios where linking would be appropriate and situations where embedding would be more appropriate.

GROUP TASK Practical Activity
Create two copies of a simple file using a word processor. Using one copy insert links to a sound and/or video file. In the second copy embed the sound and/or video file. Compare the size of the two word processor files.

Consider differences between print and multimedia:

In general most application software is designed specifically to produce output to printers or specifically to produce multimedia output for screen display. However specialist print applications, such as word processors and desktop publishing applications, often include the ability to produce multimedia output and conversely multimedia applications are able to produce printed output.

Some of the essential differences between print and multimedia display include:

- Higher resolution of print compared to screen displays.
- Interactive nature of multimedia compared to static nature of printed output.
- Ability to use hyperlinks, sound, animation and video in multimedia systems.
- Printed output cannot be altered and is relatively slow to distribute compared to online multimedia that once changed is available immediately.
- To read printed output does not require any information technology, whilst multimedia requires access to information technology and skills to use it.
- Professionally published books are more readily trusted compared to multimedia.

GROUP TASK Discussion
Contents pages and indexes are a form of navigation aid present in many printed publications. Discuss similar navigation aids present within many multimedia publications.

GROUP TASK Discussion
Newspapers are now available as online multimedia publications, however many people still prefer to purchase printed newspapers. Discuss.

AUTHORING SOFTWARE

Multimedia authoring software packages are used to design and create multimedia systems. They import and combine different media types into a single interactive system. There is an enormous range of authoring software packages available. Many specialise in the production of specific types of multimedia systems, whilst other more complex packages can produce a broader range of multimedia systems. Commonly specialised applications include templates, are simpler to learn and contain limited functionality compared to more general and complex packages. For example a specialised authoring package for creating quizzes may contain 10 different question types – multiple choice, fill-in the blank, etc. The user is limited to these question types, however the software is simple to use. In contrast a more general authoring package requires advanced programming skills to create a similar quiz, however the developer has more control over the design and behaviour of the final system.

We cannot hope to examine all the possible multimedia authoring packages available, therefore we restrict our discussion to three common examples:

Articulate's Quizmaker

Quizmaker efficiently creates graded and survey type quizzes as flash files. The current version includes 11 different types of graded questions and 10 different types of survey questions. *Fig 6.48* shows the multiple choice data entry screen. Graded tests can provide instant feedback to users or feedback can be provided at the conclusion of the test. The final quiz can be uploaded to a learning management system (LMS), where student tests are automatically delivered and results recorded.

Fig 6.48
Question entry screen from Articulate's Quizmaker authoring package.

The package includes standard colour schemes that can be customised. Many different effects are included to animate and add sound to the transitions between questions. The fonts, colours, images used for buttons and other active user interface elements can be easily customised. However the layout of each question type is fixed and larger images must be zoomed – images cannot be resized for individual questions.

GROUP TASK Practical Activity
There are many different quiz creation authoring packages available. Create a simple quiz using a trial version of one of these packages.

NeoSoft's NeoBook

NeoBook creates fully compiled and self-contained Windows applications as either executable EXE files, screensavers or as browser plug-ins. Interactive multimedia programs such as electronic books, brochures, training, games, CD interfaces and many other applications can be developed without learning and writing any programming code. A master page is used to specify components common to the whole applcation. On the screenshot in *Fig 6.49* the previous page and next page buttons will appear on all pages as they were added to the master page.

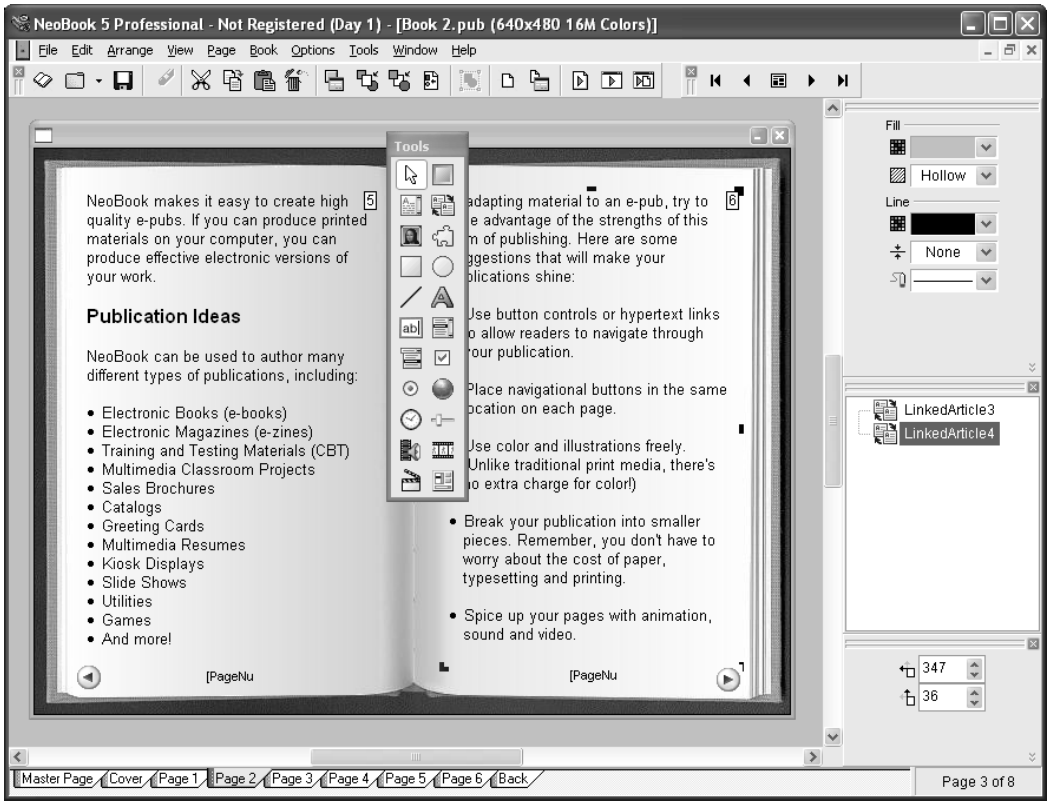


Fig 6.49
Creating an electronic book using NeoSoft's NeoBook multimedia authoring software.

NeoBook includes a tool palette of commonly used objects including text fields, check boxes, lists, image boxes, drop down menus and also a media player object. Each of these objects includes events, such as clicking the mouse, which activates actions. For example when a user clicks an image it could cause a video to play. Unlike many other authoring packages these events and actions can be specified without the need to understand or enter complex programming code. Experienced users are also catered for as they can enter or edit programming code to implement more advanced functionality.

Adobe Flash CS3 Professional

Adobe's Flash CS3 Professional forms part of Adobe's Creative Suite 3 (CS3) and is currently the leading authoring software package for creating rich interactive Flash files for the web. Flash files require the user to have installed the free Flash player from Adobe. Adobe claims more than 96% of browsers already have their player installed and furthermore many mobile devices now include the ability to display flash video content.

We introduced Flash earlier in this chapter when discussing animation. Although Flash is an excellent format for animation it can also integrate each of the other media

types. Indeed many online video repositories, including the popular YouTube.com site, use Flash to deliver streamed video over the Internet. Such Flash video files are not usually produced using Adobe's Flash authoring software, rather proprietary software converts the uploaded videos into the Flash file format.

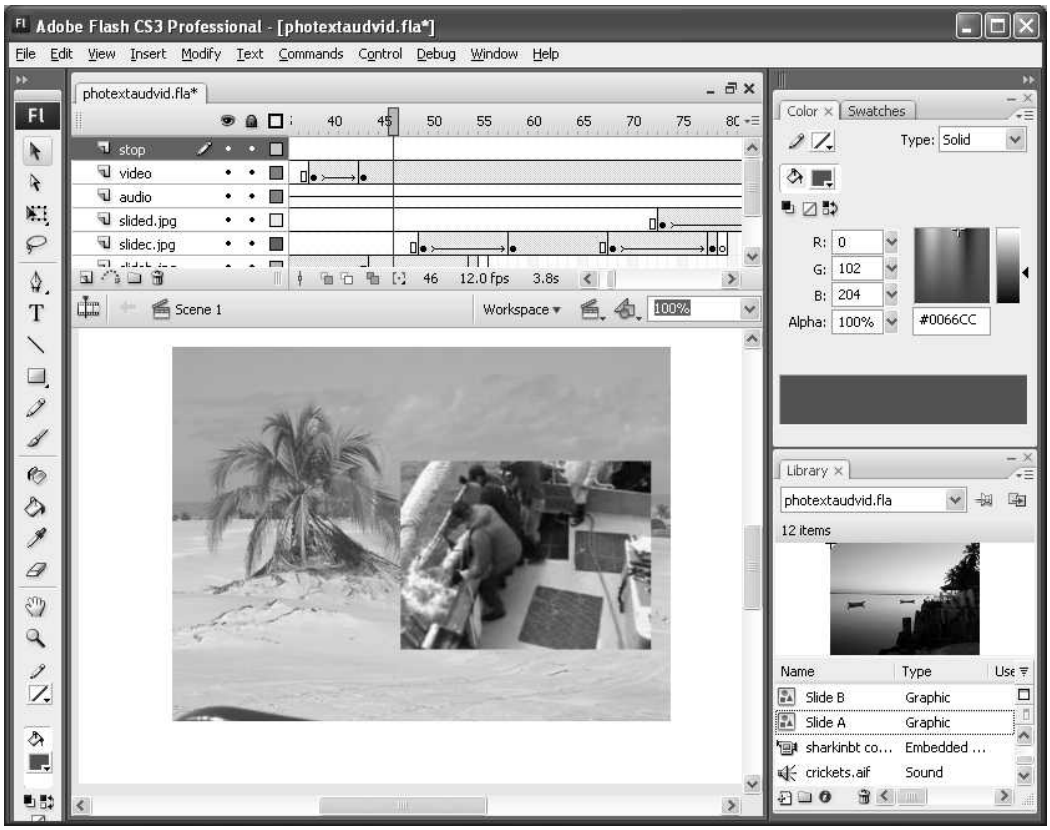


Fig 6.50
Creating an interactive movie using Adobe's Flash CS3 Professional.

Flash projects created with Flash CS3 Professional are known as movies even if they do not contain video. The screen in *Fig 6.50* shows the work area. There are four main areas of the work area known as the stage, timeline, tools and library. The stage is where the media is combined and can be previewed.

In *Fig 6.50* the stage contains an image of a desert island together with an overlayed video. The timeline is divided into frames and includes a play head so you can navigate through the frames within the project – in *Fig 6.50* frame 46 is currently being displayed and the movie is set to play at a speed of 12fps (frames per second). Each of the black dots on the timeline indicate a key frame and the horizontal arrows indicate a tween from one key frame to the next. Each row within the timeline represents a layer within the movie. In general each layer contains a single media item, which frames it is displayed in and also any animation or other effects applied to the layer. Layers higher on the timeline are displayed on top of lower layers. When the final movie is created Flash Professional combines all the layers into a single movie.

The toolbar can be seen down the left hand side of the *Fig 6.50* screenshot. The toolbar contains typical selection, text and drawing tools. All external media must first be imported into the library before it can be used within a movie. In *Fig 6.50* there are 12 items in the library and a graphic is currently selected. Once an item is within the library it can be used multiple times within the Flash project.

GROUP TASK Research
New multimedia authoring packages are constantly being produced. Research and briefly describe three examples of such packages.

ANIMATION SOFTWARE

Earlier in this chapter we considered animated GIF files and also Flash files, both these file types are used to store animation. During our discussion we mentioned Easy GIF Animator, a simple GIF animation software product. Above we considered Adobe's Flash CS3 Professional software, which is also be used to produce animations. Indeed most presentation and authoring software packages include the ability to animate transitions, buttons, menus and a variety of other objects. There are also numerous other applications that specialise in the creation of animation. In this section we restrict our discussion to two examples, Xara3D a text animation tool and Toon Boom Studio used to produce traditional cartoons and other forms of animation.

Xara3D

Xara3D is used to create 3 dimensional animations using a combination of text and/or vector images. The software is simple to use and is aimed at users who wish to create quality animations without the need to learn a complex animation software product. Once created animations can be saved as animated GIFs, Flash files, AVI video files or even as Window's screensaver files.

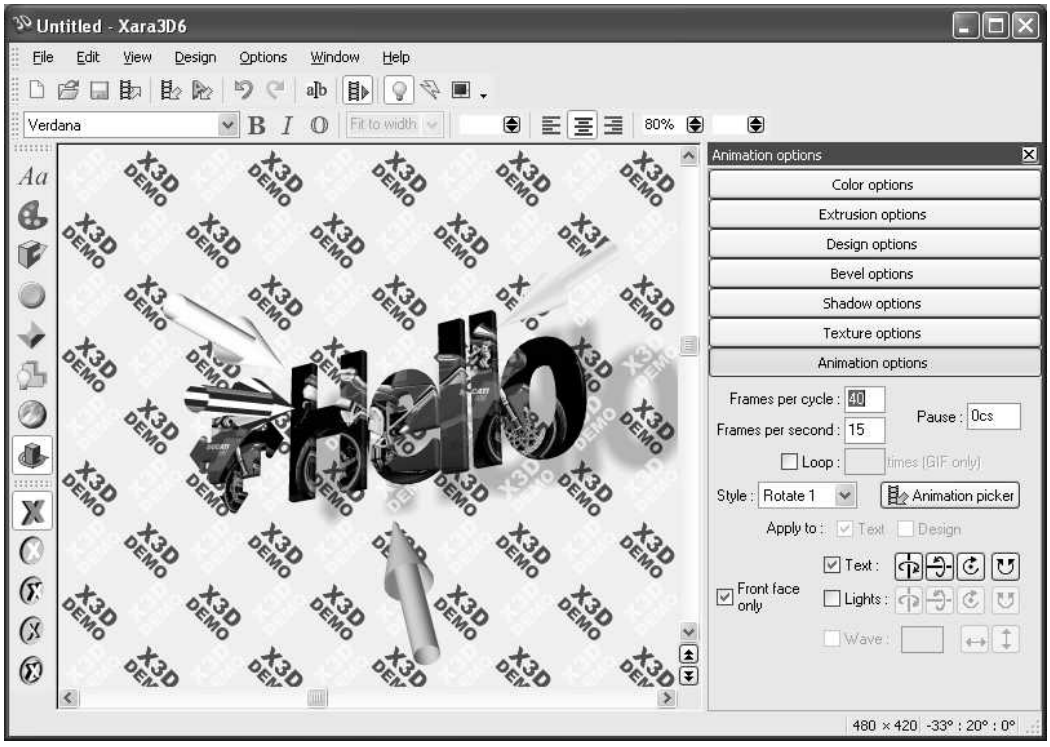


Fig 6.51
Xara3D simplifies the creation of 3 dimensional text animations.

The screenshot in *Fig 6.51* shows the main work area and option bar – the toolbar down the left hand side essentially duplicates this option bar. Each of the three large arrows are light sources that can have their colour and various other attributes altered. The animation option is open in the *Fig 6.51* screen showing attributes of the selected animation style. Currently the Rotate 1 style is selected hence the "Hello" text on the screen rotates in 3 dimensions whilst light reflects off the surface of the text and arrow

image. Each individual character or vector image can have a different animation style, for example one letter could swing left to right whilst another rotates. There are many other attributes that can be changed including altering the depth or extrusion of the text or image and various textures can be applied using JPG images. In *Fig 6.51* a bitmap image of a motorcycle has been used as the texture.

GROUP TASK Discussion
Animation software, such as Xara3D, simplifies the creation of animations but includes limited functionality compared to more complex applications. Is this an acceptable compromise? Discuss.

Toon Boom Studio

Toon Boom Studio is an animation package for producing quality cartoon style animation. Toon Boom's professional animation software products are used by many leading animation studios to create high quality animation for film, television, games, web sites and many other applications. Toon Boom Studio includes cel-based animation functions for producing vector graphic based characters; these characters are then combined into a cartoon using path-based techniques. Different cameras can be added that are able to zoom in and out or pan left/right and up/down. The final animation can be output in high resolution and can be compressed using various common video formats and codecs.

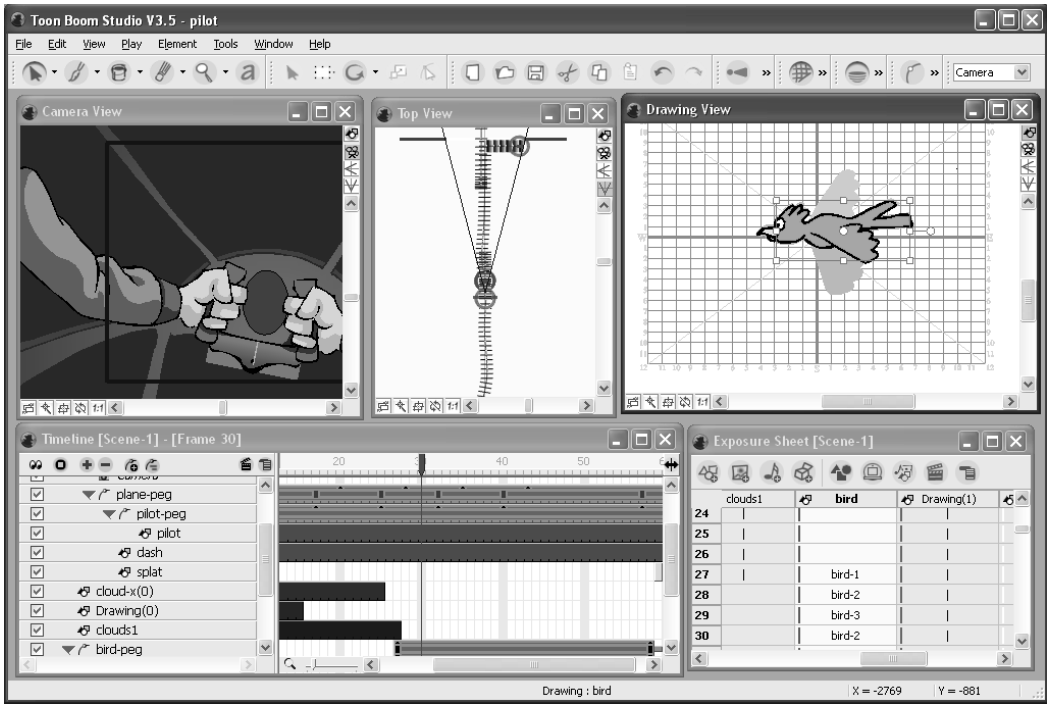


Fig 6.52
Creating an animation within Toon Boom Studio.

The bird in the drawing view window of *Fig 6.52* is a vector based cel animation. The lighter shaded images show the position of the bird in the previous and next image. This process is known as onion skinning and is a traditional technique used to ensure correct positioning of each cel within the sequence. The top view window shows the camera and also the paths each character follows within the animation. The horizontal line with dashes at the top of this window specifies the bird's path. Each dash corresponds to a single frame – in *Fig 6.52* the bird flies from right to left in front of the plane's windscreen. Examining the exposure sheet and timeline windows we see

that the bird first enters the animation at frame 27. When the dashes on a path are close together the character moves more slowly, conversely when the dashes are further apart the character moves through the scene more rapidly.

The vertical path in the top view window specifies the path the plane flies through the scene. The V shaped line shows the field of view of the camera. In the *Fig 6.52* screenshot we are at frame 30 on the timeline, hence both the camera view and top view windows are displaying details corresponding to this frame.

Toon Boom is a complex software package that aims to automate many traditional manual animation techniques. For example when a character is speaking its mouth must move in correspondence with the spoken words in the sound track. Toon Boom includes a lip-sync function that automatically analyses the spoken sound track and accurately suggests suitable mouth shapes that correspond with the sound track. The animator then draws each mouth shape and the software automatically synchronises these shapes with the sound track for each frame in the animation. In *Fig 6.53* the software produced the mouth shapes in the debSpeaks column and has assigned the character mouth shapes within the mouth column. Commonly a total of just nine unique mouth shapes are used to produce the illusion of convincing speech – commonly these shapes are labelled A to H and X is used for a closed mouth. As mouth shape A corresponds to an almost closed mouth the example in *Fig 6.53* uses the same mouth-a image for both X and A mouth shapes.

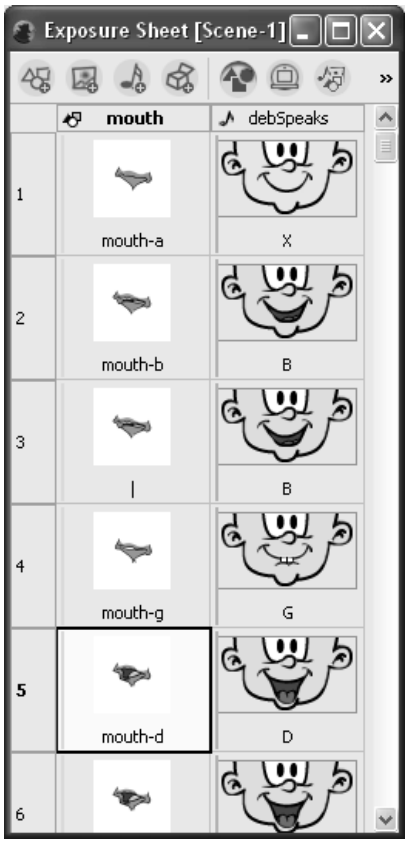


Fig 6.53
Lip Sync function in Toon Boom.

GROUP TASK Practical Activity
View a cartoon style animation that includes speech on your computer. Analyse the scene to identify individual characters and the paths that each follows through the scene. View a speech sequence frame by frame to identify the different mouth shapes used.

WEB BROWSERS AND HTML EDITORS

In chapter 2 we introduced HTML, including various examples of HTML tags. In this section we consider web browsers used to view HTML documents and examples of HTML or web editors used to create HTML files and various other file types commonly used within web sites.

Web browsers or simply browsers are such common software applications that virtually every computer with an Internet connection has a browser installed. Browsers provide the human interface between users and the vast store of information out there in cyberspace. Browsers allow users to navigate and explore the web with virtually complete ignorance in regard to the underlying processes occurring. From the user's perspective browsers provide access to a vast store of information, furthermore they assist users to locate specific information via search engines. In terms of the design of online multimedia systems designers must ensure their web

pages and content will display correctly in a wide variety of web browsers running on a wide variety of hardware and software combinations. For example screen resolutions and the speed of Internet connections will vary considerably. Web sites should be designed so they will display correctly and promptly for the broadest possible range of hardware, software and settings. For multimedia systems such issues are particularly critical as image, sound and video media are often large. A balance between storage size and quality is often needed – users will often browse to a competitor's site if they are made to wait more than a few seconds.

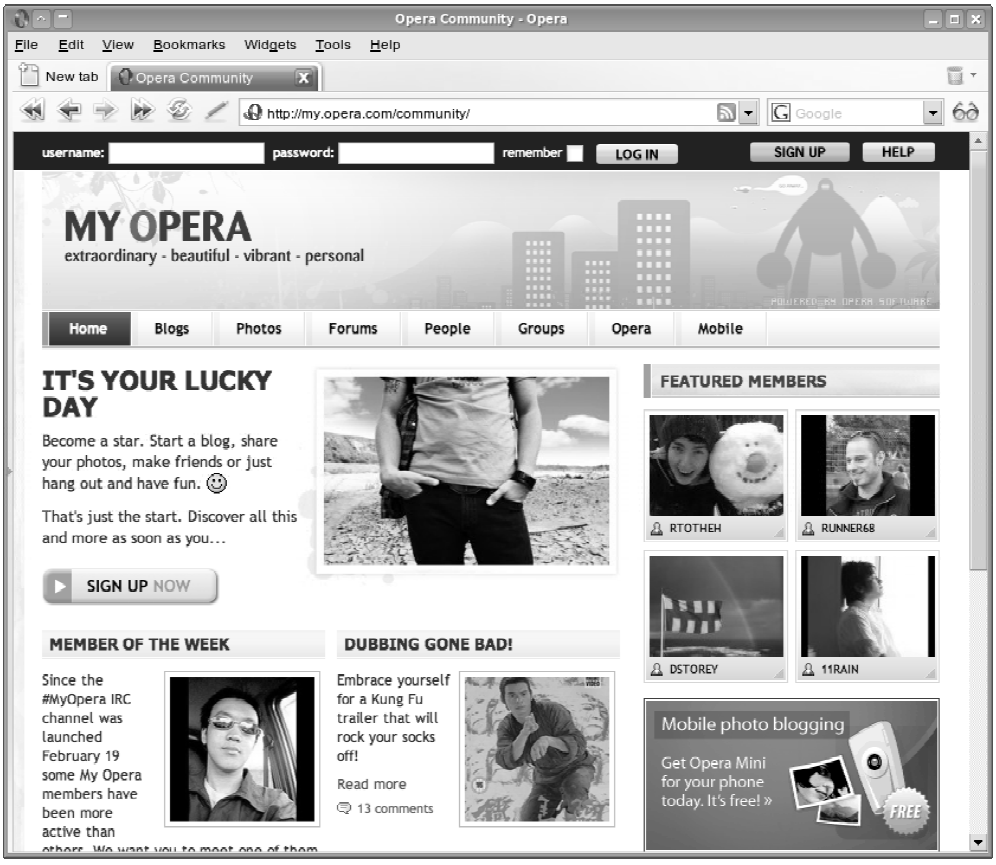


Fig 6.54
Screenshot of the Opera browser on a machine running the Linux operating system.

Microsoft's Internet Explorer is currently the most popular browser – it is included with all current versions of Microsoft's Windows operating system. Apple Macintosh computers come preinstalled with Apple's Safari browser. There are many other browsers including Mozilla's Firefox and also Opera. Versions of both Firefox and Opera are available for a wide variety of operating systems – versions of Opera are also produced for many mobile devices and some game consoles.

GROUP TASK Research
Using the Internet, or otherwise, identify different browsers available for your operating system. Comment on any significant differences between the browsers you find.

GROUP TASK Discussion
Perform a simple survey to determine the browsers used by members of your class. Determine reasons why particular browsers are used.

There is an enormous range of applications for creating HTML and web pages, from simple text editors, such as notepad to professional web development packages for developing complex Internet applications. Clearly we cannot examine all such applications hence we restrict our discussion to a brief overview of three examples, namely Window's Notepad, Coffee Cup HTML Editor and Adobe DreamWeaver CS3.

Window's Notepad

Notepad is a simple text editor included with all versions Microsoft's Windows operating system. As web pages, including HTML tags, are ultimately stored as text it is possible to view and edit the underlying source document using Notepad. In *Fig 6.55* the home page for Sydney University's Vet Science faculty is displayed within Internet Explorer and the source code is displayed within Notepad.

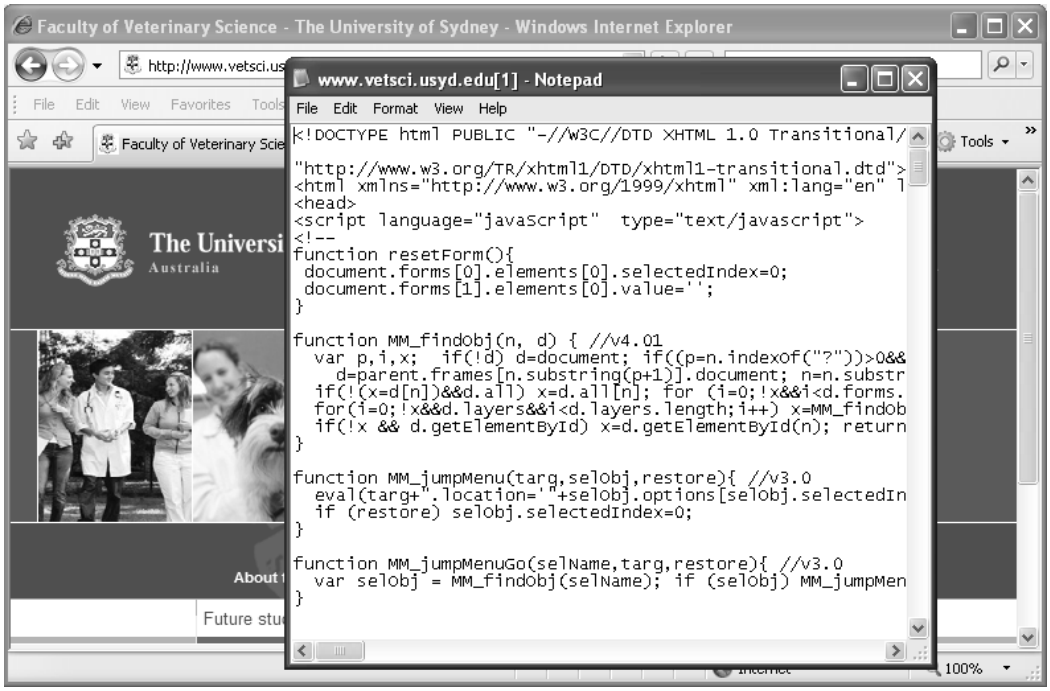


Fig 6.55
Web page and source code for Sydney University's Faculty of Veterinary Science.

Text editors, such as Notepad, are suitable for making minor edits to web pages, however they are unable to check the syntax of HTML tags and other code – all such checks must be performed manually. In *Fig 6.55* we can see that the code apparently conforms to the W3C's XHTML 1.0 standard, also note that the displayed source code includes embedded JavaScript programming code. Creating such code within a text editor would be a difficult task.

GROUP TASK Practical Activity
Browse to various web pages on the Internet and view the underlying source code within a text editor such as notepad. Identify examples of hyperlinks, images and videos within the source code.

GROUP TASK Research
W3C develops and approves various web standards including HTML and XHTML. Research who the W3C is and identify differences between HTML and XHTML.

Coffee Cup HTML Editor

Coffee Cup develops and distributes a number of inexpensive web design software applications including Flash Firestarter for creating simple Flash animations, Web Video Player for converting most video formats into Flash files for use on the web and an FTP client that allows direct editing of files on web servers. Coffee Cup's HTML Editor was their first application and has been upgraded regularly since 1996.

Coffee Cup HTML Editor includes two editing views, namely Code Editor and Visual Editor. The Preview view displays the current page as it should appear within a browser. The Tools menu includes functions that display the current web page in any browser installed on the system. This allows the page to be validated in Internet Explorer, Opera, Firefox or any other installed browser.

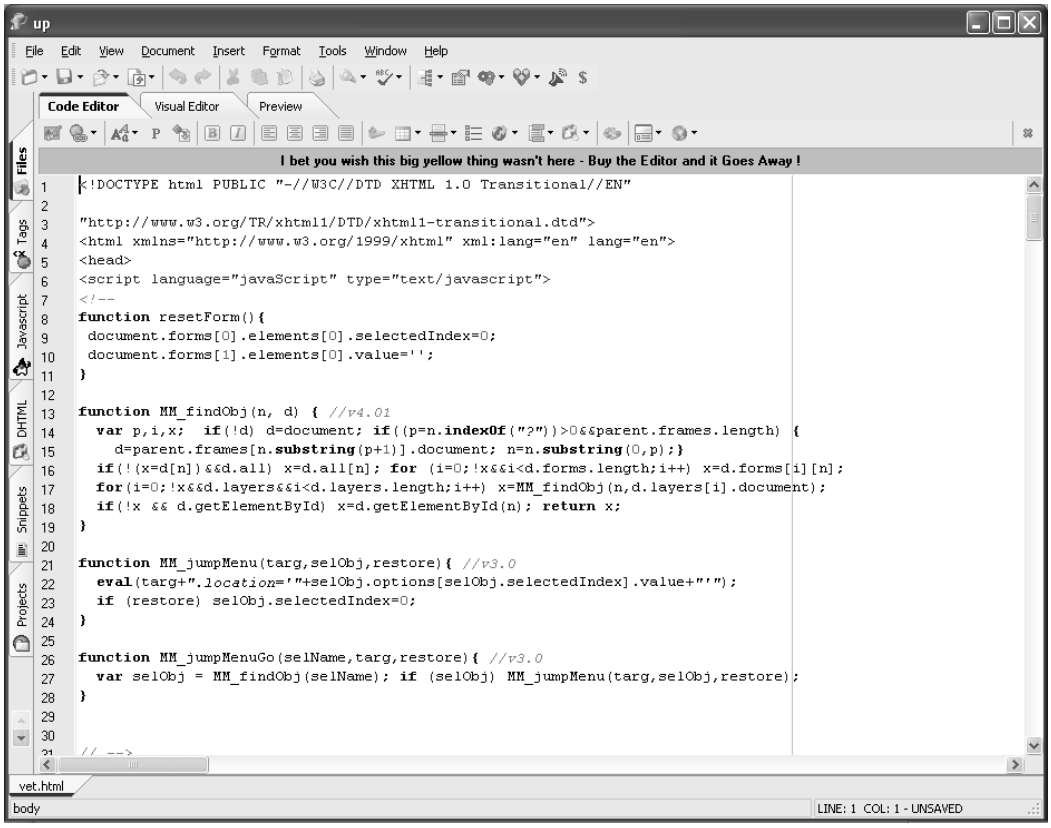


Fig 6.56
Code Editor within Coffee Cup HTML Editor.

The code editor view in *Fig 6.56* shows the actual source code and includes syntax checks that ensure HTML, JavaScript and other code within a web page is correct. In *Fig 6.56* keywords within the source code are highlighted and different colours are used to visually highlight different elements. The tabs down the left hand edge of the screen when opened show lists of elements that can be used within the code. These code elements and snippets can be dragged into the appropriate place in the code to add new statements. For instance the "Tags" tab contains a listing of all the available HTML tags. In general the Code Editor is used by experienced designers to create unusual code that cannot easily be developed using the Visual Editor view.

GROUP TASK Discussion
Compare and contrast the Notepad screenshot in *Fig 6.55* with the Coffee Cup HTML Code editor shown in *Fig 6.56*.

The Visual Editor view shown in *Fig 6.57* is a WYSIWYG (What You See Is What You Get) style editor. It allows the designer to edit web pages graphically without the need to understand the detail of the underlying code. Note that both *Fig 6.56* and *Fig 6.57* show the same web page – Sydney University's Vet Science home page. Clearly editing within the Visual Editor view is a much more user-friendly experience.

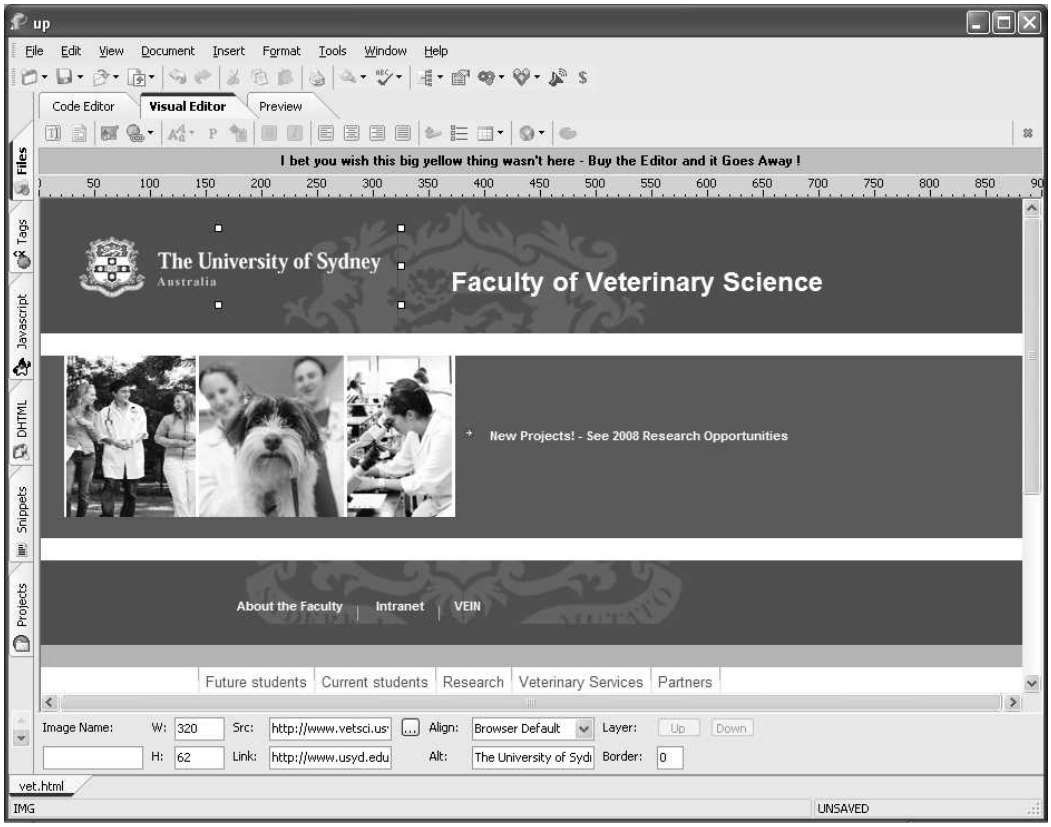


Fig 6.57
Visual Editor within Coffee Cup HTML Editor.

In *Fig 6.57* the University of Sydney logo is selected, therefore the attributes of this image file are displayed across the bottom of the screen. These details correspond to the following HTML code:

```
<a href="http://www.usyd.edu.au/"><img height="62" alt="The University of Sydney"
src="images/frontpage/usyd_logo.gif" width="320" border="0"></a>
```

When inserting an image within the Visual Editor a dialogue is presented where each of the attributes required to create the HTML code is specified. The software then automatically creates the HTML code.

Adobe DreamWeaver CS3

DreamWeaver is a professional web design and development application. It includes support for most current web technologies and is able to integrate media created in a wide range of other applications. Complete web sites can be designed and developed using the WYSIWYG design interface, however code can also be entered and edited directly via the code window. Often developers work with both the design and code windows open using a split view – the cursors in both windows are synchronised so that say clicking on an image causes the corresponding HTML code to be selected in the code window. DreamWeaver includes extensive support for cascading style sheets (CSS) which allows the design and the content to be separated. Furthermore CSS files can be reused so many pages can share the same design.

We cannot hope to even scratch the surface of the functionality included within DreamWeaver, hence we restrict our discussion to a brief overview of the user interface as we introduce the concept of cascading style sheets (CSS).

In the screenshot in *Fig 6.58* both the code window and the design window are open. The heading "Sample Entertainment" was highlighted using the mouse within the design window – notice the corresponding HTML code in the code window has been automatically selected. At the bottom of the screen the properties window shows that the selected heading is formatted as heading 1 (or h1). The other settings in the property window have been determined from within the associated mm_entertaiment.css file – this file is also open and can be viewed by clicking on its tab towards the top of the screen.

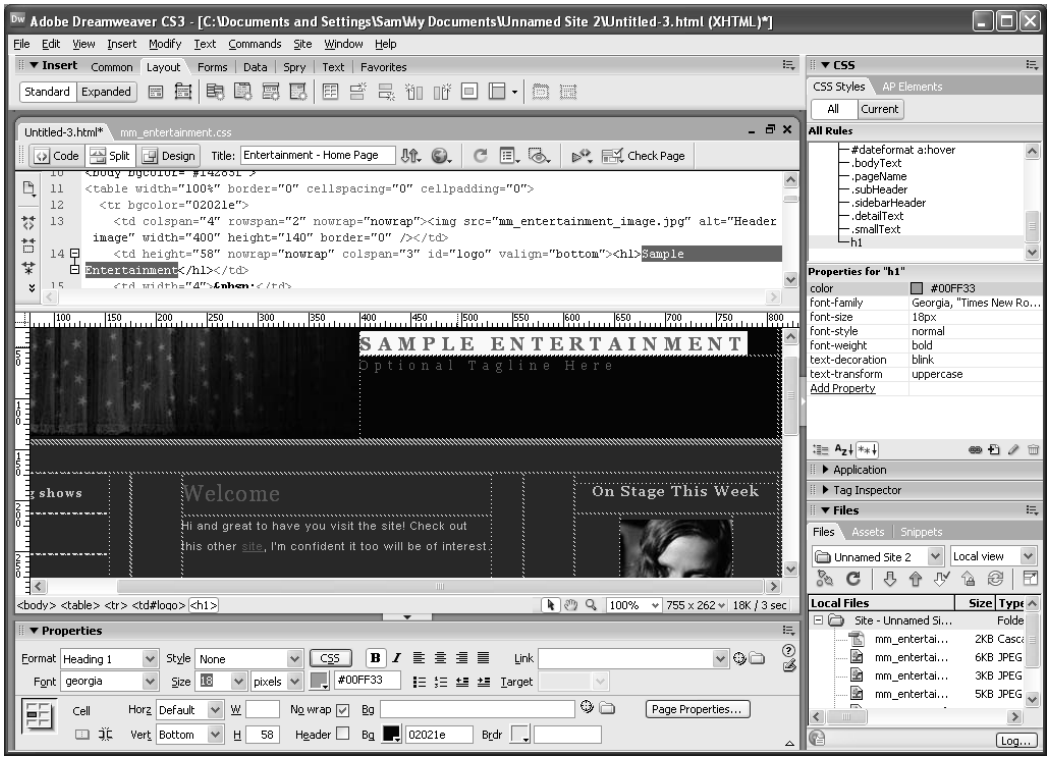


Fig 6.58
DreamWeaver includes extensive support for cascading style sheets.

Consider the CSS styles panel at top right of the *Fig 6.58* screenshot. This panel shows all the currently defined styles used on the page. In the screenshot the "h1" style has been highlighted, therefore the panel below displays properties for the "h1" style. These are the properties contained within the associated CSS file. Because the "Sample Entertainment" text is formatted as "h1" in the HTML code it inherits all the CSS settings of this style – the most obvious being the "uppercase" setting. Any changes made to the CSS file are immediately reflected in the design window.

Versions of DreamWeaver CS3 are available for Microsoft's Windows and Apple's Mac OS X operating systems. Earlier versions of DreamWeaver were produced by Macromedia, which was purchased by Adobe in 2005. DreamWeaver version 1.0 was first released in 1997 and DreamWeaver CS3 was released in 2007.

GROUP TASK Research
In 2007 DreamWeaver was considered to be the leading professional web design and development application. Research if this is still the case.

HSC style question:

Refer to the following image of the home page of Orange County Choppers website when answering parts (a) and (b).

(a) Identify FOUR examples of different multimedia elements, or links to them, on this website.

(b) Propose suitable types of software that you would use to design and create a website such as Orange County Choppers website. Justify your selection of each type of software.

Suggested Solution

(a) Multimedia elements include:

- Text – The events section is primarily composed of text where each event includes a bold heading, the location of the event and the event date.
- Image – Including the background behind the web page as well as a banner image at the top of the screen, thumbnails for each news item and thumbnails of different choppers produced by OCC.
- Hyperlinks – There are numerous hyperlinks that link to other pages on the OCC web site. For example the horizontal menu items below the top banner, and links to enlarge images and news items.
- Video – A video is embedded on the page (lower left) together with pause/play button and a button for adjusting or muting the audio.

(b) Possible software to design and create such a website include:

- Video editing software – to edit the video segments and also to compress them to a lower quality and frame rate suitable for display over the Internet.
- Photo or bitmap editing software – to touch up photographs of choppers used on the site and to combine images such as the tee-shirt and cap advertisement.
- Vector image editing software – Used to create and edit the OCC logo on the right hand side of the banner. This logo would be produced entirely within the vector software by drawing, manipulating and filling mathematically described vector objects.
- HTML editor or professional web design application – to combine all the media types into a single coherent design. The layout of each of the design elements could be created using cascading style sheet technology. This allows the content to be altered on a regular basis without the need to laboriously apply the layout and formatting to each edited element.

Comments

- In an HSC or Trial HSC examination parts (a) and (b) would likely attract 3 or 4 marks each.
- In part (a) there are many other examples that could have been described. The elements chosen should all be different media types to attract full marks.
- In part (b) many different software applications could have been described. Some other possible examples include FTP software for uploading the completed web site to the web server. Content management software (CMS) would likely be used where the content is stored and edited from within a database.
- In reality this site makes extensive use of Flash movies – the banner at the top, the advertisement containing the tee-shirt and cap, and also the video at bottom left are actually Flash files. This is not obvious on the static image of the web page, however the use of a Flash file creation tool could be discussed in part (b).

SET 6C

1. Multimedia slide shows are generally produced using:
   (A) presentation software.
   (B) word processors.
   (C) authoring software.
   (D) HTML editors.

2. A word processor document includes a video. When the document is emailed the recipient is unable to view the video. Which of the following has likely occurred?
   (A) The video was removed by the recipient's ant-virus software.
   (B) The recipient does not have a video player installed on their computer.
   (C) The video was embedded within the word processor document.
   (D) The video was linked within the word processor document.

3. Which of the following are properties of print media that distinguishes it from multimedia?
   (A) High resolution and not interactive.
   (B) High resolution and interactive.
   (C) Low resolution and interactive.
   (D) Low resolution and not interactive.

4. Which of the following best describes authoring software?
   (A) Create various different media types and compress them in preparation for distribution and display.
   (B) Create websites that incorporate a variety of different media types.
   (C) Import and combine different media types into a single interactive system.
   (D) Develop systems that collect data from users and process this data into information that is displayed.

5. What is the name of the process animators use to view the current cel over lighter versions of previous and future cels?
   (A) tweening
   (B) onion skinning
   (C) warping
   (D) morphing

6. The most basic form of HTML editor would most accurately be classified as a:
   (A) Text editor.
   (B) Word processor.
   (C) Web browser.
   (D) Code editor.

7. Which of the following best describes the purpose of cascading style sheets?
   (A) To integrate the formatting, layout and content within a single document.
   (B) To retrieve content and display it using predefined styles.
   (C) To link content from a variety of data sources for display on a single screen.
   (D) To define formatting, layout and styles separately to the actual content.

8. Which of the following lists contains only examples of web browsers?
   (A) Internet Explorer, Safari, Opera.
   (B) Notepad, Coffee Cup HTML Editor, Dreamweaver.
   (C) Toon Boom Studio, Xara3D, Adobe Flash Professional.
   (D) Articulate Quizmaker, Neobook, PowerPoint

9. Software that manages the automatic delivery of educational material to students, including records of activities completed and results of quizzes is known as which of the following?
   (A) Content management system.
   (B) Multimedia system.
   (C) Database management system.
   (D) Learning management system.

10. When creating an animation what is the purpose of a "timeline"?
    (A) To specify the length of individual clips within the animation.
    (B) To assist whilst animating each character.
    (C) To specify when characters enter and leave the animation.
    (D) To define camera angles used for the final animation sequence.

11. Distinguish between each of the following:
    (a) Presentation software and authoring software.
    (b) Web browsers and HTML editors.
    (c) Cel-based and path-based animation.
    (d) Embedding and linking.

12. Open a new document within a word processor. Identify at least FIVE specific examples of simple animations used on the word processor's user interface.

13. Most HTML editors include a WYSIWYG view and a code view. Identify specific editing tasks best accomplished using each of these views.

14. Compare and contrast a printed copy of today's newspaper with the equivalent online version of the same newspaper.

15. List factors that affect how web pages are displayed on individual computers.

EXAMPLES OF MULTIMEDIA SYSTEMS

In this section we examine examples of multimedia systems within the following major areas:

- Education and training
- Leisure and entertainment.
- Provision of information.
- Virtual reality and simulation.

Many multimedia systems cannot be categorised within just one of these areas, rather their content combines multiple areas. For example an educational game or an informational presentation for a new building that includes a simulated walk through the building. The speed of Internet connections now means that many multimedia systems are delivered over the World Wide Web – we consider these and other technological advancements as we consider each of the above areas.

EDUCATION AND TRAINING

Multimedia systems are routinely used to enhance education and training within the home, schools, universities and businesses. Some general examples include:

- Preschool and infants school age interactive educational games include large buttons, bright colours and often a game style format. Some are distributed on CD-ROM or DVD and install as applications whilst others are distributed online and viewed within a web browser. In general input is collected using the mouse. These multimedia systems often introduce reading and number skills using highly interactive content. Animated characters together with audio are often used to lead the child through the presentation. Many titles include a variety of activities that can each be completed in a short period of time. Commonly the difficulty of the activities increases automatically as the child completes sections correctly.

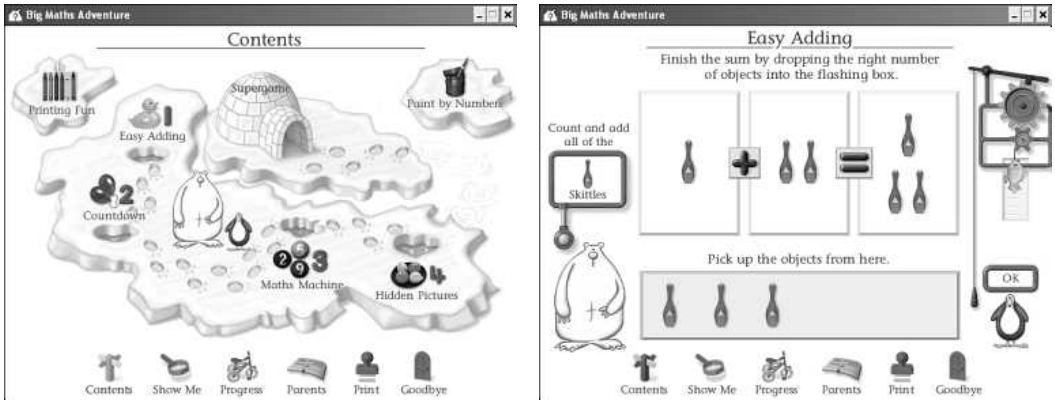


Fig 6.59
Screenshots form Bear and Penguin's Big Maths Adventure published by Dorling Kindersley.

Bear and Penguin's Big Maths Adventure (see *Fig 6.59*) includes an animated Bear and Penguin that lead the child through each of the activities on the graphical menu. Each game style activity introduces basic number skills where the level of difficulty changes as the child progresses. The animated characters help the child with spoken hints if they are unable to answer correctly. Unlimited attempts to answer are allowed and all feedback is positive. This title is designed for five to seven year olds, however there are many other related titles in the Bear and Penguin series. These titles are distributed on CD-ROM and install and execute as applications on 486 or better CPU computers running Microsoft's Windows that include a sound card and speakers.

- Learning management systems (LMS) are used by many schools, Universities and commercial training organisations to manage the distribution of multimedia and other learning resources to their students. An LMS allows different multimedia titles and quizzes to be assigned to particular classes. The student logs into the LMS where they are presented with the activities they need to complete. Commonly learning activities are viewed within a browser over the Internet or intranet. Once an activity is completed the results are communicated back to a database managed by the central LMS. The results could simply be that the student has completed the activity or they could be detailed test scores from an online test. Examples of currently popular learning management systems includes both open source products such as Moodle and commercial products such as Angel's Learning Management Suite. In addition to online multimedia many LMSs also include support for email, blogs, wikis, podcasts and various other technologies.

GROUP TASK Research
Using the Internet, investigate and briefly describe features of an open source and a commercial LMS. Determine the minimum information technology requirements to run each LMS.

GROUP TASK Research
Most LMSs are SCORM compliant. What is SCORM and what are the advantages of using a SCORM compliant LMS?

- Businesses now commonly use multimedia systems to train their staff. General training courses include occupational health and safety, customer support, communication, sales skills and computer skills. Larger corporations develop their

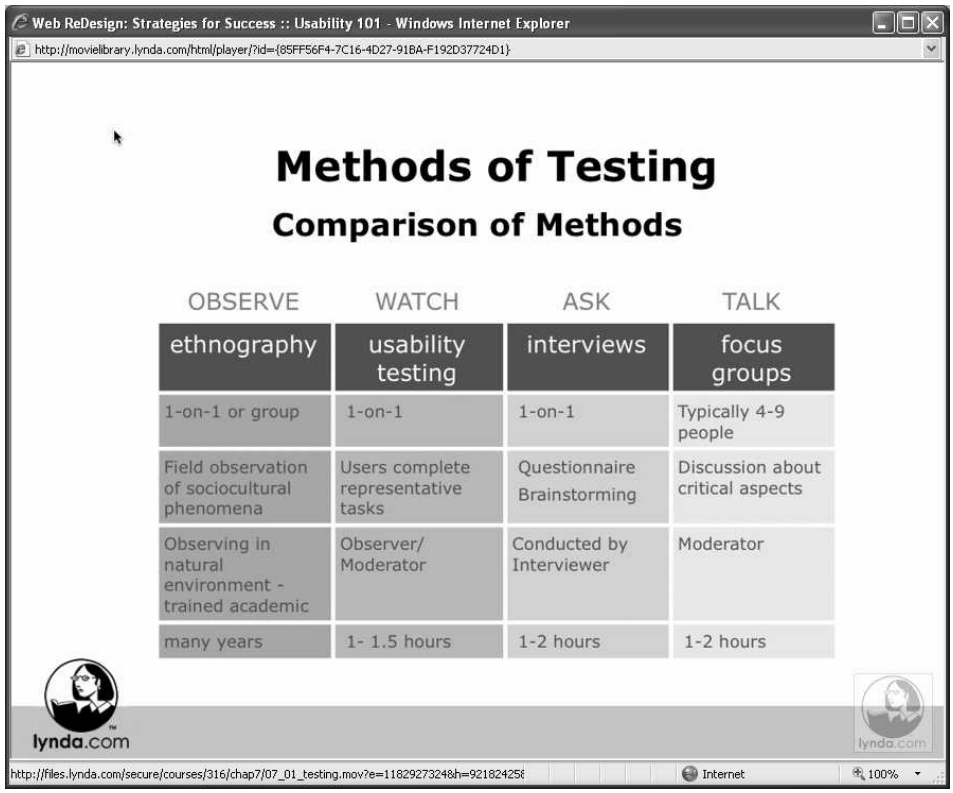


Fig 6.60
Sample narrated training material distributed by Lynda.com.

own training material, however many courses can be purchased on CD-ROM or for delivery over the Internet. *Fig 6.60* is a screen shot from a narrated training course describing usability testing. This particular course runs within Apple's Qucktime player and is distributed by Lynda.com either on CD-ROM or online.

- Software training is one of the most common forms of online multimedia training. Most large software companies produce multimedia tutorials and tours to assist users develop skills to use their products effectively. These tours and tutorials can be installed along with the software application or provided online. The screenshot in *Fig 6.61* is from a multimedia tour of Internet Explorer 7. Software companies provide such tours and tutorials not only to train users but also as marketing tools to increase sales. More detailed training for particular software products is also produced and sold by commercial training businesses.

Fig 6.61
Screenshot from a multimedia tour of Microsoft's Internet Explorer 7 browser.

GROUP TASK Practical Activity
Complete a tutorial within a software application that is used to create multimedia. Comment on the media types used and the ease of use of the tutorial.

LEISURE AND ENTERTAINMENT

Multimedia systems for leisure and entertainment in the form of games are now implemented on a variety of hardware devices including personal computers, dedicated arcade machines and game consoles, hand held consoles, PDAs and mobile phones. Many movies on DVD also include interactive multimedia content including menus and special features.

Let us consider some of the general types of games available rather than examining particular titles. There are an endless variety of different types of games, and many fit into more than one of the following categories. Nevertheless the following categories or genres provide an introduction to the range and diversity of available titles.

Action Games

In these games the player uses their reflexes to control the action in real time. Often the game involves fighting or shooting where the player controls the actions of an individual character or machine. Often such games include high levels of violence in graphic three-dimensional detail. Action adventure games extend this genre to include exploration and discovery as the player gathers equipment and materials as they fight and move to solve puzzles and navigate through mazes.

Role Playing Games

Often role playing games are set within a science fiction or fantasy setting. Each player controls one or more characters which each possess different characteristics. For instance one character may specialise in logic skills, another in magic and another in one to one combat. Characters can be computer controlled, however often a human player controls each character. Such games often run for an extended period of time with characters developing skills and specialisations as the storyline progresses. In many role playing games players take turns and have time to consider strategies and tactics before acting. Other role playing games operate in real time and rely on quick decisive actions.

Massively Multiplayer Online (MMO) Games

As the name suggests these games can include potentially thousands of players interacting over the Internet. Most examples operate within an ongoing virtual world hosted on a dedicated and powerful server. The virtual world continues to exist as players log in and out of the game. Players from across the world can combine their resources to combat opponents or achieve other game objectives.

Fig 6.62
Screenshots from World of Warcraft.

Consider the following:

Many popular MMO games are also role playing games. For example *Fig 6.62* shows screen shots from Blizzard Entertainment's World of Warcraft, a currently popular MMO role playing game. The game is available in versions for Windows and Macintosh computers. World of Warcraft players pay a monthly subscription fee.

GROUP TASK Discussion

It is likely that some members of your class play MMO and/or role playing games. Identify the games played and determine the typical number of participants and also the hardware required to run each game.

Platform Games

Within platform games each player causes a character to jump, bounce, swing, climb or otherwise travel between onscreen platforms. Platform games are one of the earliest forms of console game. Perhaps the most popular early example being Donkey Kong, which introduced Nintendo's Mario character who remains the companies mascot today. Versions of Donkey Kong were copied and implemented on many platforms including Nintendo's popular "Game and Watch" series produced during the 1980s (see *Fig 6.63*).

Fig 6.63
Nintendo's Game and Watch featuring Mario in the game Donkey Kong.

Traditionally the animation used in platform games was two-dimensional, however recently three-dimensional platform games have emerged. Platform games once dominated the commercial market. They now occupy a small part of commercially produced games but remain popular as freeware and shareware titles implemented as flash files for display within web browsers.

Simulation Games

Simulation games mimic a real world situation. The most popular examples include flight simulators, driving simulators and life simulators such as the popular "The Sims" series. Other examples involve economics where players create and manage simulated businesses or run their own country, including planning cities and collecting taxes. Computer simulations of traditional card and board games as well as many sport simulations, such as golf and football, are also popular.

Fig 6.64
Screenshots from Tiger Woods PGA Tour 07 (top) and Railroad Tycoon 2 (bottom).

Two screenshots are shown in *Fig 6.64*. The top screen is from the Xbox version of Tiger Woods PGA Tour 07 by EA Sports. Versions are produced for all major game consoles and also for Windows computers. The bottom screenshot is from Railroad Tycoon 2 developed by PopTop software. This is an economic simulation where the objective is to build and successfully manage the operation of a railroad network. Versions are available for Windows, Macintosh and Playstation. The animation in both these games is almost photo quality, however on personal computer versions the actual quality of the display is heavily influenced by the speed of the CPU, amount of RAM and more significantly the specifications of the video hardware.

GROUP TASK Research
Using the Internet or otherwise determine the specifications of currently popular game consoles. Identify details such as the CPU, RAM video RAM and secondary storage used.

Consider the following:

Computer games are just one use of computers for leisure. Other examples include researching hobbies such as family history, sport statistics, photography, bush walking, music or model railroading. Many of us now use computers as a primary medium for communicating with family and friends, for example instant messaging, blogs, forums, email or web cameras.

GROUP TASK Discussion
Survey your class to determine how each member uses computer technologies for leisure. Identify leisure activities that include multimedia.

PROVISION OF INFORMATION

The integration of a variety of different media types makes multimedia systems well suited to the delivery of information. Users can make selections to filter and search the content for specific information. The general aim of most websites is to provide information to users. The information may be provided to advertise products, promote services or simply to inform users.

Examples of multimedia specifically designed to provide information include:

- Information kiosks are dedicated multimedia systems that usually include a touch screen together with a secured personal computer (see *Fig 6.65*). Some contain magnetic swipe card readers, printers and Internet connections. They are used in foyers of larger commercial buildings to provide basic introductory information about the organisation, within shopping malls in the form of a directory providing information about each store and its location within the mall. Many clubs include information kiosks that incorporate a loyalty system where the club member swipes their card to obtain loyalty points and discounts.
- Multimedia brochures, reports, presentations and business cards for business are created and distributed on CD-ROM. Small diameter, business card size and irregular shaped CDs are possible (see *Fig 6.66*). As CDs contain a single spiral track it is the smallest dimension of the CD that determines the maximum storage capacity.

Fig 6.65
Information kiosk examples.

Fig 6.66
CD-ROMs can be produced in a wide range of sizes and shapes.

- Multimedia encyclopaedias make extensive use of hyperlinks and different media types. Electronic encyclopaedias were first distributed on CD-ROM and DVD, however many are now delivered over the Internet. The introduction of multimedia encyclopaedias replaced printed encyclopaedias virtually overnight.

  Users should carefully research the source of information within online encyclopaedias. For example the popular Wikipedia although a valuable resource is a collaborative collection of articles contributed by individuals. Much of Wikipedia is indeed fact however inaccuracies do exist, as individuals are free to edit articles as they please. There can be significant delays before such inaccuracies are detected and corrected. Content within professionally compiled encyclopaedias is rigorously verified prior to being published.

GROUP TASK Discussion
Brainstorm further examples of multimedia systems whose primary purpose is the delivery of information.

GROUP TASK Research
Information kiosks are often installed in public areas where they are subjected to extreme environmental conditions, such as a wide range of temperature, rain and vandalism. Research features of information kiosks that allow them to operate under such adverse conditions.

VIRTUAL REALITY AND SIMULATION

Virtual reality systems that simulate the real world are finding applications in a wide range of industries. These multimedia systems allow participants to experience an environment as close as possible to the real world environment being simulated. Many of these simulators are used for training personnel where it is impractical in terms of safety and/or cost to perform training in the real world, for example aircraft simulators and medical training simulators. Virtual reality systems are also used to present new architectural designs to clients.

Some example applications of virtual reality include:

- Aircraft flight simulators allow pilots to experience and deal with aircraft failure and other possible disasters. Use of a simulator is much more cost effective (and obviously far safer) than using a real aircraft. *Fig 6.67* shows an exterior and

Fig 6.67
Exterior (left) and interior (right) view of one of CAE SimuFlite's aircraft simulators.

interior view of an aircraft simulator. The entire simulator sits upon hydraulic struts that move in three dimensions to accurately simulate the current attitude of the simulated aircraft. The cockpit faithfully reproduces the layout of the real aircraft and includes multiple screens behind the entire windshield.

GROUP TASK Research
Virtual reality systems are used extensively within the military for both training and also during operations. Research examples of such systems.

- Medical schools have traditionally used textbook images and cadavers to train students. Currently virtual reality simulators are becoming the training method of choice. Such simulators allow students to explore the human body in detail including stripping away layers to examine tissue and organs both externally and internally. Dextroscope is one such VR system (see *Fig 6.68*); the user wears stereoscopic glasses and is able to manipulate three-dimensional images under the transparent screen using intuitive hand and finger movements. Surgeons are able to practice surgical techniques prior to performing the actual procedure on patients.

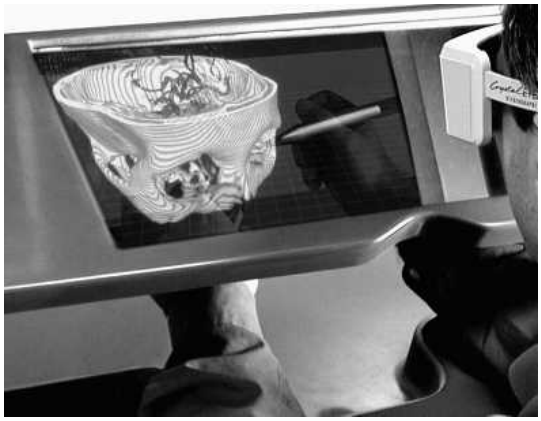


Fig 6.68
Dextroscope is a virtual reality system used to train surgeons and other medical students.

- Surgeons use virtual reality systems to assist during many surgical procedures. Transparent screens are used within the VR headset so the surgeon sees both the real view of the patient overlayed with the virtual view. Accurate sensors are used to ensure the real and virtual views remain accurately aligned as the surgeon moves.
- Experimental virtual reality systems are being used to treat various phobias and to alleviate pain. For instance a patient with a fear of heights can be exposed to a virtual cliff or someone with an extreme fear of spiders can be exposed to a virtual spider. Research into pain relief indicates that immersing patients in a relaxing but engaging virtual environment greatly reduces the amount of pain they experience during medical procedures.

GROUP TASK Research
Research and briefly describe further examples of virtual reality systems used to assist medical practitioners.

- Virtual walkthroughs of new architectural designs can be created and analysed prior to construction commencing. Many CAD (Computer Aided Design) software applications are able to create simple virtual reality displays directly from the CAD drawings. This enables both designers and clients to better visualise the completed building. Some systems are viewed on standard computer screens whilst more advanced systems utilise virtual reality headsets to produce a more realistic three-

Fig 6.69
Screen from Vision House's VR Kichen.

dimensional walkthrough. Software applications are available for home use, home owners can design kitchens (see *Fig 6.69*), bathrooms or even complete homes then view and move through their designs in three dimensions.

- Virtual tours of houses, buildings and other landmarks are routinely used by real estate agents and also as informational guides. Tours, such as the Bavarian Church tour in *Fig 6.70*, are created by collecting a number of 360-degree photographic sequences. Each sequence of images is stitched together electronically into a continuous view. Hotspots are added so the user can move from one 360-degree view to another adjoining view. The user is able to rotate and zoom in and out within each view.

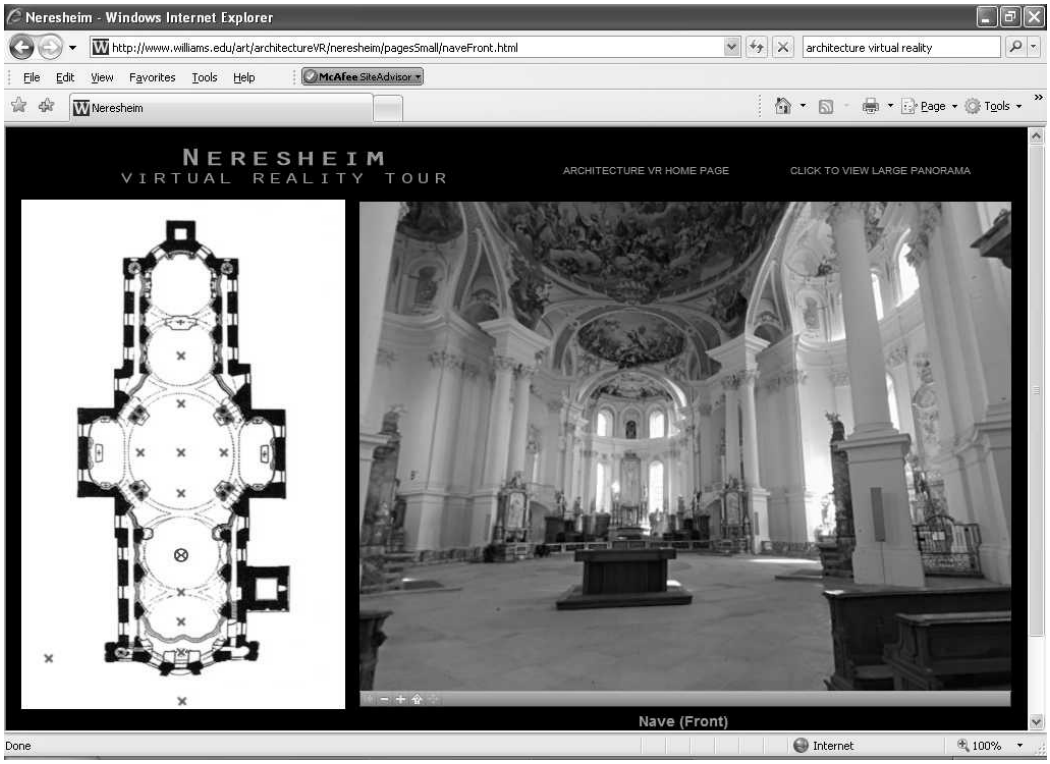


Fig 6.70
Online virtual tour of a Bavarian church produced by the Art History Department of Williams College in Williamstown Massachusetts USA.

GROUP TASK Research
Research further examples of virtual reality systems used to simulate new and existing designs.

- The military makes extensive use of virtual reality systems for training, planning and during actual operations. Complete training exercises can be completed within networked simulators. The soldiers sit in realistic vehicles such as tanks and armoured vehicles. When planning a real operation virtual reality can be used to visually and intuitively describe each detail of the mission. During missions virtual reality systems allow soldiers to better visualise their environment and the positions of their comrades and enemies.

GROUP TASK Research
Research and briefly describe specific examples of military applications of virtual reality.

EXPERTISE REQUIRED DURING THE DEVELOPMENT OF MULTIMEDIA SYSTEMS

A large variety of specialised skills and expertise are required during the design and development of multimedia systems. For small systems a single person may take on a variety of different roles, however for larger commercial systems many different specialists are used. Often existing content, such as stock photographs, clipart, video footage and sounds are used. If such content is covered by copyright then a licence needs to be acquired from the content provider. Creating new media from scratch requires skills relevant to the particular media type. For example professional video production requires a vast array of experts including directors, cameramen, sound engineers, editors and producers – consider the credits for a typical TV show. Once the various media content has been acquired or created they need to be combined to form the multimedia system. Those skilled in layout and design, together with personnel possessing suitable technical expertise with the information technology are required. Project managers coordinate and manage the activities of all these personnel and systems designers create the overall design of the multimedia system and oversee the entire operation to ensure the development remains true to their design in terms of both content and use of technology.

Some general areas of expertise, together with specific examples, include:

Content Providers

Content providers, as the name suggests, provide ready to use content. There are organisations that specialise in the provision of stock photographs, animations, video and also text articles. These content providers act on behalf of the copyright holders and negotiate fees so that the content can be legally used. The content provider retains a portion of the fee and the remainder is forwarded to the copyright owner. FotoSearch is one such content provider (see *Fig 6.71*), they currently manage the licensing of over 2 million photographs, videos and audio clips.

Fig 6.71
FotoSearch is a searchable content provider of image, video and audio.

The license fees charged and the method for calculating such fees varies widely. Some content providers charge a flat fee that allows unlimited use of any of their media. Others negotiate fees based on the number of copies that will be made, the length of time the content may be used or on the significance of the content within the context of the entire presentation. In some cases royalties are paid to the copyright holder over time based on actual sales of the multimedia product.

Some individual photographers, writers, graphic artists, etc negotiate licence fees on their own behalf – often these people will also create original content to meet a specific need. For example a writer may contract to provide a series of articles on a particular topic, in most cases the writer retains all copyrights so they are free to licence the work to others. It is generally far less expensive to negotiate a licence to use existing content than it is to create the content from scratch.

GROUP TASK Discussion

Ensuring copyrights are respected is difficult when content is distributed in digital form over the web, however the web is an excellent medium for marketing such content. Discuss advantages and disadvantages of distributing content over the Internet.

System Designers

There may be a single system designer on smaller multimedia projects and a team on larger projects. System designers are the personnel who work through the stages of the system development cycle. They identify the purpose of the system, make decisions on the most suitable and feasible solution and design the overall solution. This includes determining the hardware and software that will be used and also preparing specifications that detail the information processes that will form part of the solution.

Project Managers

Project managers develop the project plan and ensure it is followed during development. Often adjustments to the plan will need to be made as some sub-tasks run over time or over budget. It is the responsibility of the project manager to schedule and also monitor each of the other development personnel and the tasks they complete. Project managers must be able to communicate and negotiate with other members of the team.

Writers

Writers produce the textual content within multimedia systems and they also create storylines upon which videos, animations and other aspects of the presentation will be based. Writers are selected based on both their writing ability and also on their knowledge of the subject matter. For example writing multiple-choice questions for a medical training quiz requires quite different skills and knowledge compared to writing the storyline for a new adventure game.

Video production personnel

Video can be produced using a simple digital video camera or it can involve a large crew of specialists. For most commercial multimedia systems a crew comprised of at least a director, camera operator, sound engineer and perhaps actors and editors is required. The director visualises the script and then directs the other personnel so that their vision of the final production is realised. Directors are responsible for all artistic aspects of the production. Prior to filming a scene the director approves set designs and costumes and coaches the actors. During filming the director decides on camera angles, lighting and how actors should deliver their lines. After filming they oversee the final editing of the production.

Audio production personnel

Sound engineers specialise in the recording and editing of audio. This includes music, voice and special effects. Much of a sound engineers job is highly technical as they adjust levels and mix different digital audio clips together. The aim of live recording is to reproduce the original live sound as accurately as is possible. Digital audio recordings greatly assist in this regard as unlike analog recordings, digital sound files do not lose quality as they are copied and manipulated.

In many multimedia presentations the audio elements are created or significantly altered using the computer. For instance in many multimedia presentations small audio files are used to add sound effects that provide feedback when the user clicks interactive elements such as buttons and menus. Many computer games make extensive use of sounds that are entirely computer generated or are radically altered. As a consequence sound engineers working on multimedia titles require creative and artistic skills in addition to their technical skills.

Illustrators and Animators

Both illustrators and animators are artists who draw figures and scenes. Illustrators often produce original drawings to supplement the accompanying text. Today most illustrations and animations are created using computer software; therefore illustrators and animators must be proficient computer users. Most illustrators and animators work using vector graphic software applications rather than bitmap software applications.

Graphic designers

Graphic designers improve the readability of multimedia by enhancing the visual appeal of the presentation. They organise the layout of screens, adjust colour, typography and size, and they also develop a consistent look and feel for the presentation. Traditionally graphic designers were employed to layout print media, including magazines, newspapers, advertising and packaging. Today many graphic designers work on the design of websites and multimedia. Graphic designers must have an eye for colour and balance, and they must be able to target their designs to particular audiences. Almost all graphic designers create using computers, hence skills in the use of popular multimedia software applications is required.

Technical personnel

Technical personnel working on multimedia systems ensure the final system will operate correctly on user's machines. They need to consider the hardware configuration and communication speed of typical user's machines. Some considerations and tasks performed by technical personnel include:

- Multimedia delivered over the Internet is reliant on the speed of the user's Internet connection. Different levels of compression, lower resolutions and streaming can be used to ensure the presentation can be delivered in a timely fashion over slower Internet connections.
- When distribution is on CD-ROM then there is physical limit to the total size of the presentation. Images, audio and video can be compressed to reduce their size. However the technical personnel must ensure the required codecs are present on the end-users computers.
- For many multimedia systems, such as most games, software developers and programmers are employed to code the interactive elements of the presentation.
- Copy protection and product registration techniques are often added to commercial multimedia to reduce the likelihood of illegal copies being made.

HSC style question:

A shopping mall is investigating the possibility of installing a series of information kiosks. Each kiosk will include an interactive map of the shopping mall together with a list of stores grouped into categories. Touching on a store either on the map or within the list brings up detailed information about the store.

The kiosks will also include a buyer loyalty awards system. Receipts from all stores will include a barcode. Buyers who join the loyalty awards system are given a card that includes a unique barcode. To accumulate points the buyer first scans their card and then scans their store receipts. Awards points can be redeemed for vouchers that provide discounts on goods within the store.

(a) Identify the information technology and data/information required to implement the above information kiosks.

(b) Describe the roles and skills of TWO different people involved in the development of the information kiosks.

Suggested Solution

(a) Hardware would include touch screens, barcode readers, personal computers and secure enclosures for each kiosk. A network connection to link each kiosk back to a server that hosts a database would be needed together with a switch to connect all nodes.

Software would include the operating system and network software together with the multimedia application that includes the image map of the shopping mall. The application accesses detailed images and textual information about each store from the central database server. The database would also store information about buyers who have joined the loyalty program. Such data would include their name and contact details, the barcode number from their loyalty card and details of points awarded and redeemed.

(b) A graphic designer would create the layout of the screens. Their role is to create a consistent design that is readable, visually appealing and can be used intuitively. As a touch screen is used the interface should use large hotspots and buttons. The graphic designer will need to use their design skills as well as their computer skills to achieve usable and attractive screens.

The database will need to be created by a person with technical skills in regard to the creation and use of databases. This person's role includes creating the schema for both the store details database and the awards system database. They will also be involved in writing queries to retrieve the categorised list of stores, details of individual stores and details of buyer's award points. Furthermore developing strategies and systems for regular backup and techniques for securing the database would form part of their role.

Comments

- In an HSC or Trial HSC examination each part would likely attract 3 marks.
- A database that is networked to each kiosk is needed for the awards system and also simplifies updating of store details.
- In part (b) there are numerous different development personnel that could have been described.

SET 6D

1. Multimedia systems designed for preschool children should:
   (A) include large colourful buttons.
   (B) present information as text.
   (C) include game style activities that take time to master.
   (D) use the keyboard in preference to the mouse.
2. Most computer games use which of the following media types?
   (A) Video and hypertext.
   (B) Animation and audio.
   (C) Text and video.
   (D) Audio and images.
3. Which of the following best describes how royalties are paid to copyright owners?
   (A) A license fee negotiated and paid prior to use of the content.
   (B) A percentage of the total revenue from actual sales as they occur over time.
   (C) A flat fee paid to the copyright owner.
   (D) The fee charged to create original content for a particular project.
4. A dedicated touch screen console within a shopping mall that includes a categorised and searchable list of the stores within the centre would be best described as:
   (A) an information kiosk.
   (B) a simulation.
   (C) a training system.
   (D) a multimedia brochure.
5. Which of the following terms describes an organisation that manages the use of original media on behalf of copyright owners?
   (A) Content provider.
   (B) Legal firm.
   (C) Graphic designer.
   (D) Project manager.
6. Who is responsible for all artistic aspects during commercial video production?
   (A) The producer.
   (B) The actors.
   (C) The director.
   (D) The video editors.
7. Which of the following are the primary participants of all multimedia systems?
   (A) Content providers.
   (B) System designers.
   (C) Technical personnel.
   (D) End users.
8. Which of the following best describes a platform game?
   (A) A game that executes on a dedicated game console such as Xbox or Playstation.
   (B) A game where characters jump, swing or are otherwise moved from one onscreen platform to another.
   (C) A game that will only execute on a specific hardware and software platform.
   (D) A game where the user progresses through a increasingly more difficult sequence of levels.
9. Accurately reproducing a real world environment is the ultimate aim of:
   (A) simulations.
   (B) virtual reality.
   (C) artificial intelligence.
   (D) computer games.
10. Tasks performed by graphic designers include all of the following EXCEPT.
    (A) Designing screen layouts.
    (B) Choosing colour schemes.
    (C) Specifying information technology.
    (D) Developing a consistent look and feel.

11. List THREE specific examples of multimedia systems you have seen within each of the following major areas.
    (a) Education and training
    (b) Leisure and entertainment
    (c) Provision of information
    (d) Virtual reality and simulation

12. Describe the roles and skills of each of the following people during the development of multimedia systems.
    (a) Content providers.
    (b) System designers.
    (c) Project managers.
    (d) Technical personnel.

13. Identify personnel skilled in the collection, creation and/or editing of each of the following media types.
    (a) Text
    (b) Image
    (c) Audio
    (d) Video

14. Identify and describe advances in technology that have enabled multimedia to be routinely distributed over the World Wide Web.

15. Research an example of a virtual reality system. Identify the participants, data/information and information technology for this system.

OTHER INFORMATION PROCESSES WHEN DESIGNING MULTIMEDIA SYSTEMS

In this section we focus on the information processes occurring during the Designing stage of the system development lifecycle. This is when the multimedia system is actually built – in simple terms the media is collected and combined to form the final multimedia presentation such that it operates effectively to display the presentation using the system's information technology.

All of the information processes are used during the design of new multimedia systems. The overall organisation of the presentation is designed using storyboards – the storyboards specify the screen designs, links and individual media elements – text, images, audio and video. The media must be collected which often involves analog to digital conversion. Once the data has been collected it is then processed and reorganised into a form suitable for storage and retrieval. The collected digital files will need to be compressed and reorganised into a format of suitable quality and size. Suitable authoring software can now be used to combine and process the media, create hyperlinks and format the presentation into a form suitable for distribution and final display.

Throughout the discussion that follows we will refer to the design of an example multimedia presentation based on the 1960's Thunderbirds television series. An introduction to this proposed multimedia presentation follows.

Consider the Thunderbirds multimedia presentation:

The "Thunderbirds" is a 1960s television series portraying the activities of the fictitious organisation known as "International Rescue". International Rescue is located on Tracy Island somewhere in the Pacific Ocean. The organisation is headed by Jeff Tracey and includes his five sons Scott, Virgil, Alan, Gordon and John Tracey. Each of the Tracy boys pilots one of the five Thunderbird vehicles. The Tracey's live on Tracey Island along with Brains who designed the Thunderbird vehicles as well as numerous other unusual rescue vehicles and other contraptions. Lady Penelope, with assistance from her butler Parker, is International Rescue's London agent.

The multimedia presentation will use images, sound and video of a toy Tracey Island produced by Soundtech (see *Fig 6.72*). Toy models of each of the Thunderbird vehicles and also of each of the five Tracey boys are included with the island. The island toy includes interactive features such as buttons which play audio of each member of International Rescue, audio of each vehicle and also buttons that launch Thunderbird's one, two and three from the toy island.

Fig 6.72
Soundtech's toy Tracey Island.

The presentation will be designed for use by young children and hence will make extensive use of images, video and sound in preference to text. The image of Tracey Island in *Fig 6.72* will be used as the main menu where the child clicks on different areas to access further media. A separate page of the presentation will be dedicated to each significant member of International Rescue.

ORGANISING PRESENTATIONS USING STORYBOARDS

We considered storyboarding earlier within chapter 2. Storyboards describe the layout of each individual screen together with any navigational links between screens. Often storyboards are hand drawn sketches used to plan the overall design of a multimedia presentation. For smaller presentations links between screens can be indicated directly on the individual screen designs. For presentations with many screens a separate navigation map can be sketched. Such navigation maps show each screen as a simple rectangle with hyperlinks to other screens shown using arrows.

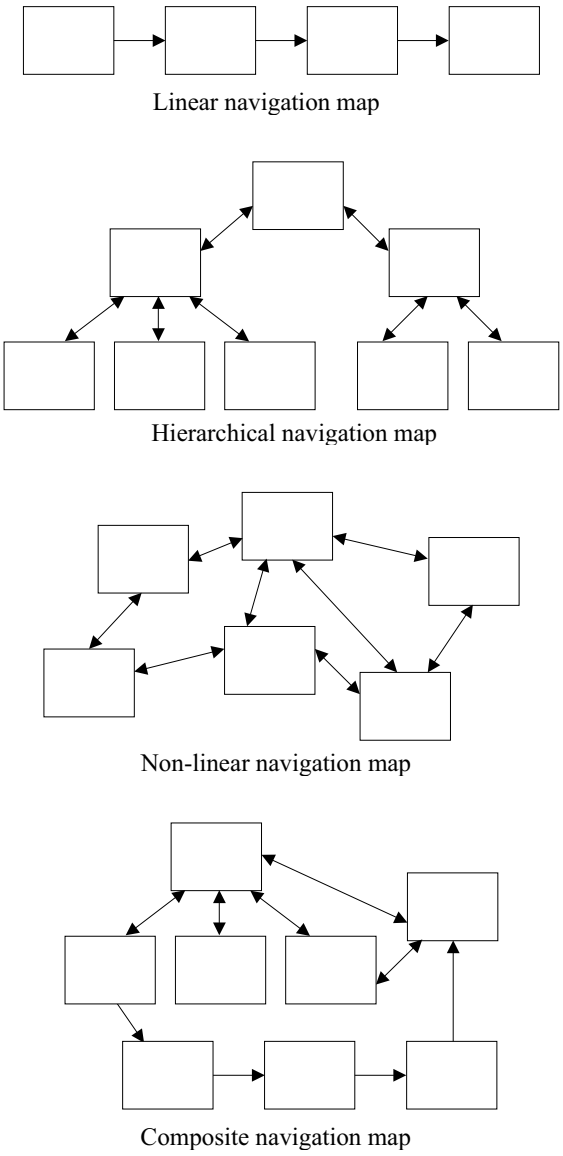


Fig 6.73
Common navigation structures used on storyboards.

There are four commonly used navigation structures namely linear, hierarchical, non-linear and composite (see *Fig 6.73*). The nature of the information largely determines the selection of a particular structure. For example a research project has a very different natural structure compared to an online supermarket. There are two somewhat conflicting aims when designing a navigation structure. Firstly the structure must convey the information to users in the manner intended by the author and secondly the users should be able to locate information without being forced to manually search through irrelevant information. Designers of multimedia must balance the achievement of these two aims as they choose the most effective navigation structure.

Linear storyboards are used when there is a strict logical sequence or order to the presentation. This is common for slide show presentations that tell a story or progressively introduce information to users. For example PowerPoint presentations used during lecture style presentations are almost always linear. Slides are presented in sequence to reinforce the information presented by the speaker.

One of the most important differences between printed and multimedia presentations is the ability for users to easily navigate in a variety of different ways. Within printed books footnotes and indexes allow users to locate related content. In multimedia presentations hyperlinks automate these connections. However users also need to understand the overall navigation structure so they can return to information or explore in a logical sequence. In general it is advisable to base most multimedia presentations on a well understood navigation structure. Commonly a hierarchical

navigation map is used together with a displayed menu that so the user can easily create a mental picture of their current position within the overall presentation. Hierarchical navigation maps categorise the content into progressively more detail. Other links can still be added within the hierarchical structure to allow unstructured browsing. The menu system describes the categories within the hierarchical tree. Some presentations use a separate clickable navigation pane whilst others simply display a list of the higher level screens above the current screen.

GROUP TASK Practical Activity
Examine a number of different multimedia products and websites.
Determine the structure used for navigation and the menus used.
Comment on the ease of navigating through each system.

The individual screen layouts should clearly show the placement of navigational items, titles, headings and content. It is useful to indicate which items exist on multiple pages – such as contact details and menus. Notes that describe elements or actions that are not obvious should be made. Each layout should not just include the functional elements; it should also adequately show the look and feel of the page. Commonly a theme for the overall design is used – this can be detailed separately to each of the individual page designs.

Consider the Thunderbirds multimedia storyboard:

This system is designed for use by young children hence all screens will be composed of images where different images and regions within images link to further screens. It is envisaged that HTML image maps will be used so that different parts of an image can link to different screens or other media files such as audio and video clips.

The main menu will be constructed using a single image of Tracey Island with hyperlinks from the launch areas for the first three Thunderbirds and a further link to a control room screen that includes the Tracey boys and also Lady Penelope. Each of the control room images links to an individual screen for each Thunderbird. The individual Thunderbird screens will have an image of the Thunderbird vehicle and its pilot together with a small image of Tracey Island. The vehicle will link to a video of the Thunderbird launching, clicking on the pilot will play a random audio clip of the pilot speaking and the small island image will link back to the main island screen.

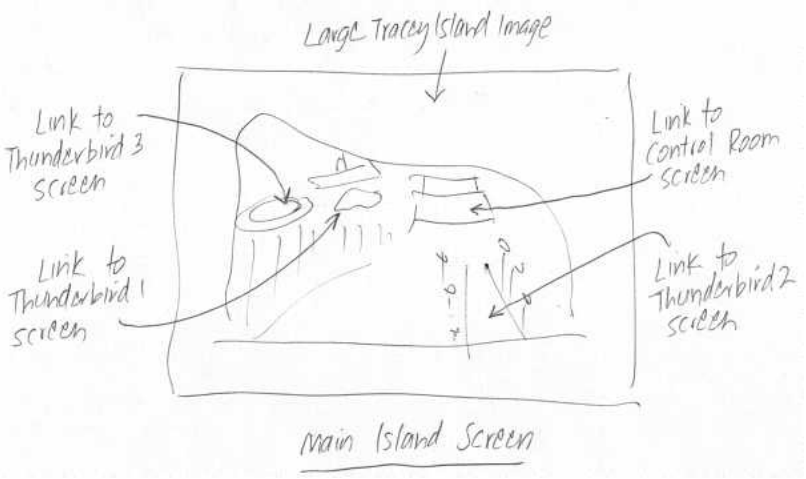


Fig 6.74
Main Island screen design for Thunderbirds multimedia presentation.

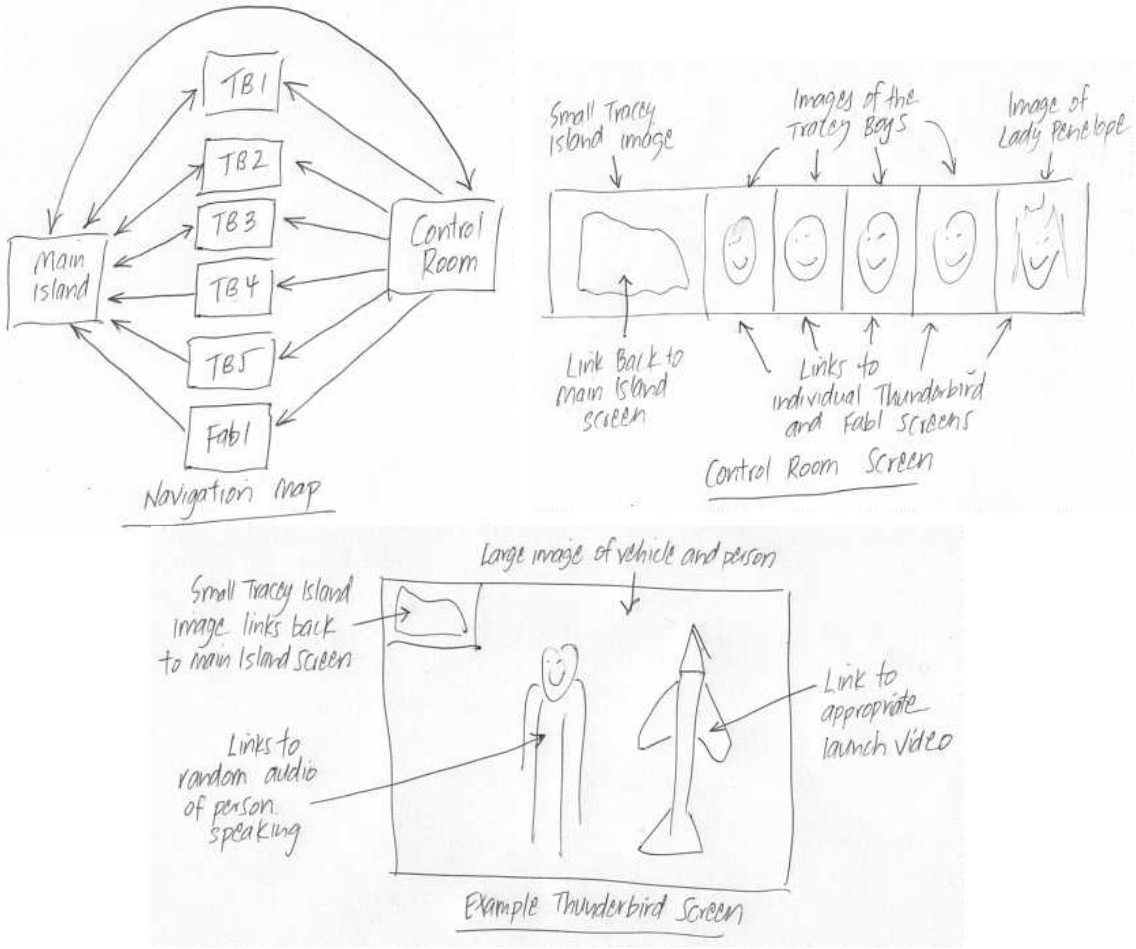


Fig 6.75
Thunderbirds storyboard including screen designs and navigation map.

GROUP TASK Discussion
Critically analyse the above storyboard for the Thunderbirds example. Consider issues such as screen resolution, suitable authoring software and possible file formats for each media element.

COLLECTING MULTIMEDIA CONTENT

Text and numbers are usually input directly using the keyboard. Collecting audio, video and images requires the raw analog data to be converted to digital. In this section we discuss flatbed scanners, digital cameras, microphones and sound cards and video cameras. We conclude by describing typical analog to digital converters (ADCs) within these devices.

Flatbed Scanner

There are various different types of image scanner; all collect light as their raw analog data and transform it into binary digital data. This digital data may then be analysed, organised and processed using optical character recognition software into numbers or text. In relation to multimedia flatbed scanners are more often used to collect image data in the form of bitmaps. Note that barcode scanners use similar technology to flatbed scanners. Essentially light is reflected off the image and one or more sensors, known as photocells, are used to detect the intensity of the reflected light. Each photocell outputs a varying current in response to the amount of light it detects. This

current is converted to a binary number as it passes through an analog to digital converter (ADC) – commonly the output is an 8-bit number from 0 through to 255.

The most commonly used photocells are known as charge coupled devices (CCDs). CCDs contain one or more rows of photocells built into a single microchip. CCD technology is used by many image collection devices including; CCD barcode scanners, digital still and video cameras, handheld image scanners, and also flatbed scanners. For both barcode and image scanners a single row CCD is used.

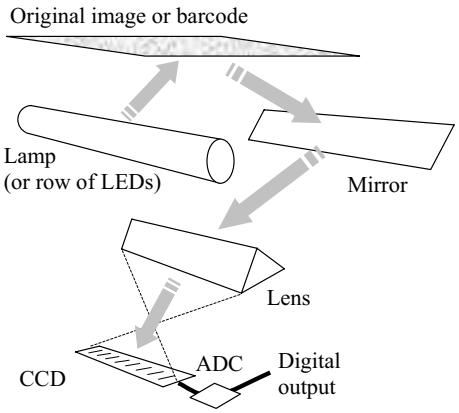


Fig 6.76
The components and light path typical of most CCD scanner designs.

The light source for flatbed and many other scanners is typically a single row of LEDs; the light being reflected off the image back to a mirror as shown in *Fig 6.76*. The mirror reflects the light onto a lens that focuses the image at the CCD. Each photocell in the CCD transforms the light into different levels of electrical current that are fed into an ADC.

Flatbed scanners based on CCDs are by far the most common; scanners based on other technologies are available, but currently they fall into the higher quality and price ranges. We mentioned above that the 8-bit binary numbers returned from a flatbed scanner's ADC range from 0 to 255. If white light is used then these numbers will represent shades of grey, ranging from black (0) to white (255). So how do flatbed scanners collect colour images? They reflect red light off the original image to collect the red component, green to collect the green component and blue for the blue component. Therefore three 8-bit numbers representing the intensity of red, green and blue respectively are used to represent each pixel. Some early scanners performed this action by doing three passes over the entire image using a different coloured filter for each pass; this technique is seldom used today. Today most scanners use an LED light source that cycles through each of the colours red, green, blue; hence only a single pass is needed.

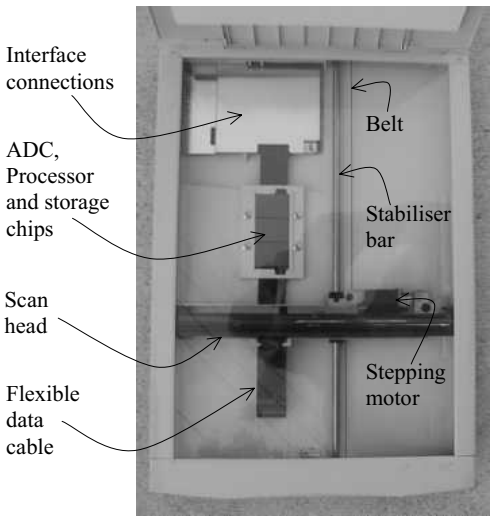


Fig 6.77
Components of a flatbed scanner.

The LED lamp, mirror, lens and CCD are all mounted on a single carriage; these components are collectively known as the scan head (refer *Fig 6.77*). All the components on the scan head are the same width as the glass window onto which the original image is placed; this means a complete row of the image is scanned all at once. The number of pixels in each row of the final image is determined by the number of photosensors contained within the CCD; typical CCDs contain some 600 sensors per inch, predictably this results in images with horizontal resolutions of up to 600 dpi (dots per inch). After each row has been scanned the scan head is precisely moved to the next row. Due to the rapid speed of modern flatbed scanners it is usually difficult to detect this stop start movement.

The following operations occur as a colour image is scanned using a flatbed scanner:

- The current row of the image is scanned by flashing red, then green, then blue light at the image. If you open the lid of a scanner you'll predominantly see white light, this is due to the colours alternating so rapidly that your eye merges the three colours into white. After each coloured flash the contents of the CCD is passed to the ADC and onto the scanner's main processor and storage chips.
- The scan head is attached to a stabilising bar, and is moved using a stepping motor attached to a belt and pulley system. The stepping motor rotates a precise amount each time power is applied; consequently the scan head moves step by step over the image; pausing after each step to scan a fresh row of the image. The number of times the stepping motor moves determines the vertical resolution of the final image.
- As scanning progresses the image is sent to the computer via an interface cable. The large volume of image data means faster interfaces are preferred; commonly SCSI, USB or even firewire interfaces are used to connect scanners. Once the scan is complete the scan head returns back to its starting position in preparation for the next scan.

GROUP TASK Research
Advertising for flatbed scanners often claim to output higher resolutions than should be possible based on the number of physical photosites on their CCDs. Research how manufacturers justify such claims.

Digital Camera

Digital cameras have completely transformed the photographic process. Traditional mechanical and chemical processes using film have been in use since the 1830s; they have now been largely replaced by electronic and digital processes.

Virtually all digital cameras are currently based on either charge coupled devices (CCD) or complementary metal oxide semiconductors (CMOS). These technologies are at the heart of digital camera design; both are image sensing technologies, that is, they detect light and transform it into electrical currents. Currently CCDs provide better image quality, however they cost more to produce and require significantly more power to operate. CMOSs use similar production methods to other types of microchips, hence they are inexpensive to produce and have far lower power requirements. Unfortunately the quality of images produced with CMOS based cameras is currently inferior to CCD produced images. CCD technology is used in almost all dedicated digital cameras where the need for high quality output more than justifies the extra cost and power requirements. CMOS technology is currently used for applications such as security cameras and mobile phone cameras; image quality being sacrificed to minimise critical cost and power requirements.

We discussed CCD technology previously in relation to flatbed scanners; the CCDs used in digital cameras operate in precisely the same manner, they convert light into a varying electrical charge. At our level of discussion this is also the primary function of CMOS chips, the only significant difference being that CMOS chips combine the image sensing and ADC functions into a single integrated chip. Our remaining discussion will focus on CCD based cameras, however much of the discussion is equally true of CMOS based cameras.

Unlike scanners, who generate their own constant light source, cameras must control the amount of light used to generate the image. In a traditional film camera this is accomplished using a shutter. The shutter alters the size of the hole or aperture

through which the light passes and also alters the time the aperture is open (shutter speed). Digital cameras use the same principles; many models do have mechanical shutters whilst others do away with mechanical shutters altogether. Adjusting the time taken between the CCD being reset and the data being collected is used to produce the equivalent process in a digital camera.

Digital cameras must be able to collect an entire image in a virtual instant. This means a two dimensional grid of photosensors is needed; the CCD shown in *Fig 6.78* contains some 2 million photosensors, or photosites, resulting in images with resolutions up to 1600 by 1200 pixels. Digital cameras are often classified according to the number of photosites on their CCDs, cameras based on the CCD in *Fig 6.78* would be classified as 2 megapixel cameras; some CCDs contain 20 million or more photosites.

Fig 6.78
A CCD from a digital camera.

Recall our flatbed scanner, it collected colour using red, green and blue light; this same principle is used by digital cameras. There are various ways of implementing this principle:

- Take the picture three times in quick succession, first with a red filter then a green and finally a blue filter. The three images can then be combined to produce the final full colour image. This approach is seldom used as even slight movement leads to blurred images.
- Use three CCDs where each is covered by a different coloured filter. A prism is used to reflect the light entering the camera and direct it to all three CCDs. This approach is obviously more expensive as three CCDs and various other extra components are needed, however the resulting images are of excellent quality. This technique is generally restricted to high quality professional cameras.
- By far the most common approach is to cover each photosite with a permanently coloured filter. The most common filter is called a bayer filter; this pattern alternates a row of red and green filters with a row of blue and green filters.

The Bayer filter is the most common approach (see *Fig 6.79*), let us continue our discussion based on this technique. A Bayer filter has two green photosites for each red and each blue photosite. The human eye is far more sensitive to green light, hence using extra green sensors results in more true to life images. So the raw analog data from the CCD represents the intensity of either red, green or blue light in each of its photosites. This analog data is then digitised using an analog to digital converter (ADC). Earlier we discussed how 2 megapixel cameras produce final images with resolutions containing approximately the same number of full colour pixels (1600 × 1200 = 1,920,000 ≈ 2 million pixels); how is this possible when the initial digital data from the ADC contains information representing the intensity of one single colour per pixel? A process known as demosaicing is used to produce the final colour values for each pixel. Examining the Bayer filter in *Fig 6.79*, we see that each red photosite is surrounded by four green and four blue photosites, averaging the four green values gives us a very accurate approximation of the likely actual green

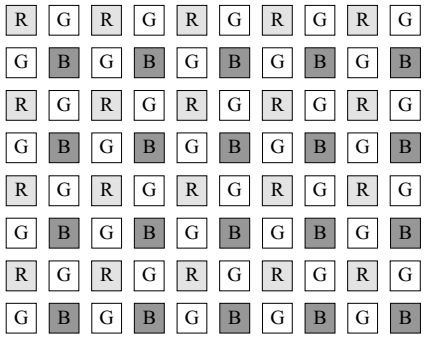


Fig 6.79
Bayer filters alternate red and green rows with blue and green rows.

value, similarly averaging the blue values gives us the most likely blue value. Combining the original 8 bit red value with the calculated 8 bit green and blue values give us the final 24-bit colour value for the pixel. This processing occurs for every pixel, resulting in the output of a complete 24 bits per pixel image with a resolution similar to the number of photosites on the CCD.

The resulting image is usually compressed, to reduce its size prior to storage; commonly a lossy technique, such as JPEG, is used. The file is then stored on a removable storage device, most cameras use removable flash memory cards. A computer later reads these cards, either directly or via an interface cable, which stores the images on the computer's hard disk.

GROUP TASK Discussion
"A camera with say 6 million photosites is not really a 6 megapixel camera." Discuss the validity of this statement.

Microphone and Sound Card

Microphones are, predictably, used to collect data in the form of sound waves, they convert these compression waves into electrical energy. In digital systems, this varying analog electrical energy is converted, using an analog to digital converter (ADC), into a series of digital sound samples. In this section we examine the operation of microphones and consider the operations performed by a typical sound card to process the resulting analog electrical energy into a sequence of digital sound samples.

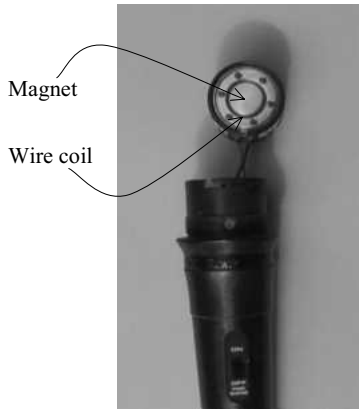


Fig 6.80
A dynamic microphone element. This one has the magnet mounted within the wire coil.

There are a variety of different microphone designs, the most popular being dynamic microphones and condenser microphones. All these designs contain a diaphragm which vibrates in response to incoming soundwaves. If you hold your hand close to your mouth whilst talking you can feel the effect of the sound waves; the skin on your hand vibrates in response to the sound waves in exactly the same way as the diaphragm in a microphone vibrates.

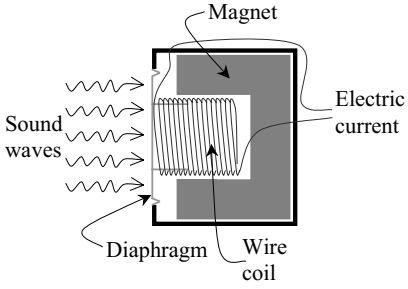


Fig 6.81
Detail of a dynamic microphone.

A dynamic microphone has its diaphragm attached to a coil of wire; as the diaphragm vibrates so too does the coil of wire (see *Fig 6.81*). The coil of wire surrounds, or is surrounded by, a stationary magnet; as the coil moves in and out the interaction of the coil with the magnetic field causes current to flow through the coil of wire. This electrical current varies according to the movement of the wire coil, hence it represents the changes in the original sound wave.

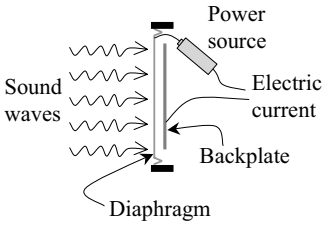


Fig 6.82
Detail of a condenser microphone.

Condenser microphones alter the distance between two plates (see *Fig 6.82*). The diaphragm is the front plate; it vibrates in response to the incoming

soundwaves, whereas the backplate remains stationary. Therefore the distance between the diaphragm and the stationary backplate varies; when the two plates are close together electrical current flows more freely and as they move further apart the current decreases, hence the level of current flowing represents the changes in the original sound waves. Condenser microphones require a source of power to operate; this can be provided from an external source via the microphone's lead or by using a permanent magnetically charged diaphragm. In either case the signal leaving the microphone is an analog signal, this signal must be converted to digital before it can be stored as a sequence of digital sound samples.

Let us now consider the processes taking place once the analog signal from the microphone reaches the computer's sound card. The analog signal is fed via an input port into an analog to digital converter (ADC), which predictably converts the signal to sequences of binary ones and zeros. The output from the ADC is then fed into the digital signal processor (DSP). We shall consider the operation of ADCs later in this section, at this stage we consider what happens to the raw digital sound samples once they reach the DSP. The DSP's task, in regard to collected audio data, is to filter and compress the sound samples in an attempt to better represent the original sound waves in a more efficient form. The DSP is itself a powerful processing chip; most have numerous settings that can be altered using software. Most DSPs perform wave shaping, a process that smooths the transitions between sound samples. Music has different characteristics to speech, so the DSP is able to filter music samples to improve the musical qualities of the recording whilst removing noise. The DSP uses the sound samples surrounding a particular sample to estimate its likely value, if these estimates do not agree then the sample can be adjusted accordingly. Once the sound samples have been filtered the DSP compresses the samples to reduce their size. Some less expensive sound cards do not contain a dedicated DSP, these cards use the computer's CPU to perform the functions of the DSP. The final sound samples are then placed on the computer's data bus. The data bus feeds the samples to the main CPU, where they are generally sent to a storage device.

GROUP TASK Research
Stereo sound contains two distinct channels and many movie sound tracks contain five, six or more audio channels. How are such sound tracks collected and created?

Video Camera

Most video cameras combine image collection with audio collection; the result being a sequence of images that includes a sound track. The term 'video camera' is commonly used to describe devices that combine a video camera and microphone for collecting, with a video/audio recorder/player for storage and retrieval; perhaps the alternate 'camcorder' term better describes such devices. Analog video cameras, or camcorders, have been available for more than twenty years, however digital versions now dominate the market. There are also PC cameras or web cameras that really are just cameras, their sole task being to collect image data and send it to the computer via an interface port.

Both analog and digital camcorders use CCDs to capture light and microphones to capture sound. CCDs and microphones both collect analog data; they convert light and sound waves into electrical current. Digital video cameras convert these electrical signals into digital within the camera, whereas the output from an analog video camera must be converted to digital before a computer can process it.

PC or web cameras, in most cases, use inexpensive complementary metal oxide semiconductor (CMOS) chips. The single CMOS chip in a web camera contains photosensors, an ADC, and all the circuitry necessary to communicate and transmit digital image data to the computer's port. As these cameras are designed to collect images and video for display over the Internet, the poor image quality derived from CMOS photosensors is less significant.

Let us consider the operation of a typical camcorder in more detail. To collect video effectively it is crucial to control the changing nature of the light entering the lens. As the camera and/or subject moves the camcorder needs to respond by altering the amount of light entering the lens and also by refocussing this light onto the CCD. The CCD provides a perfect indicator of the amount of light entering the lens; if most of the photosites on the CCD record strong light intensities then too much light is entering the lens, so the diameter of the aperture is reduced; conversely if the light intensities are weak then the aperture is opened. Focusing is not so simple, the camcorder needs to know the distance to the subject of the current frame. Some camcorders bounce an infrared beam off an object in the centre of the frame; the time taken for this beam to reflect back to the camera is used to calculate the distance to the object. The camcorder uses a small motor to move the lens in or out to focus the image onto the CCD based on the calculated distance. Other camcorders compare the intensity of light detected at adjacent photosites within a rectangle of pixels in the middle of the frame; gradual changes indicating blurred images and larger differences indicating the image is focused. The lens is then moved slightly in or out and the intensities are again compared; the process repeats until the maximum difference in intensities is achieved.

Fig 6.83
The Hitachi DZ-MV100 camcorder stores video on recordable DVDs.

Each photosite in a camcorder CDD and in a digital still camera CCD collect light in precisely the same way; however a video camcorder must be able to collect some 25 to 30 images or frames every second. To accomplish this task the CCD in a camcorder has two layers of sensors, one behind the other; the front layer collects the light and then transfers the electrical current to the lower layer. Whilst the lower layer is being read, the upper layer is collecting the next image.

In all analog camcorders, and in many older digital camcorders, the lower layer of the CCD is split into two distinct fields; the first field being the odd numbered rows and the second being the even numbered rows. The data from one of these fields is read for each frame, the fields being alternated for each successive read; in effect only half the total image is retained. The images are collected in this way to reduce the amount of data and also to mirror the operation of most analog televisions. Televisions display video by alternately painting the odd rows and then the even rows; this process is known as interlacing. Most digital camcorders now use 'progressive scan CCDs', this somewhat obscure term means the contents of the entire CCD is read as a single complete image. Camcorders using progressive scan CCDs require faster processors to manipulate the extra data, however they do produce higher quality video; a further positive is their ability to collect high quality still images.

Within digital camcorders the data passes through an ADC; the resulting digital data is then compressed into a format suitable for storage – usually MPEG. Currently most digital camcorders use magnetic tape, recordable DVD or hard disks for storage.

Models using tape or hard disks require connection to the computer via an interface cable; most connect using either USB or firewire ports. Models using DVD storage also include ports to connect to computers, however DVDs are often more convenient as their contents can be played directly using set-top DVD players or the data can be accessed via the DVD drive on a computer. Most digital camcorders also include analog outputs and inputs allowing transfer of video data to and from analog sources.

GROUP TASK Discussion
In general, digital video cameras capture image data at much lower resolution than digital still cameras. Furthermore most digital still cameras can also capture video albeit at much lower resolution than the images they collect. Discuss reasons for these differences. Research why people purchase video cameras when digital still cameras can capture video.

Analog to Digital Conversion

Analog to digital converters (ADCs) repeatedly sample the magnitude of the incoming electrical current and convert these samples to binary digital numbers; for audio data the size of the incoming current directly mirrors the shape of the original sound wave, hence the digital samples also represent the original wave. The ADCs used in scanners and digital cameras perform essentially the same function as those found on sound cards. The CCDs in image and video collection devices produce varying levels of electrical current that represent the intensity of light detected at each photosite. The ADC converts these varying analog signals into binary digital data.

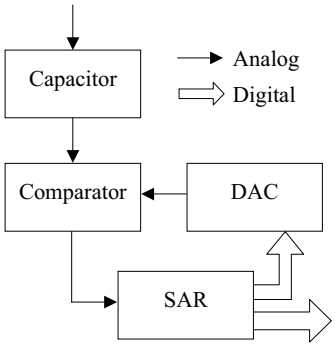


Fig 6.84
Components and data connections for a typical ADC.

Most analog to digital converters contain a digital to analog converter (DAC); on the surface this seems somewhat strange, however the digital to analog conversion process is significantly simpler than the corresponding analog to digital conversion process. The components and data connections within a typical ADC are shown in *Fig 6.84*; this ADC has been reproduced from chapter 3 (page 325) where its operation was described in some detail.

GROUP TASK Discussion
Review both the ADC and DAC processing described in chapter 3 (page 325-325). Are these ADC and DACs suitable for use within sound cards, digital still and video cameras and flatbed scanners? Discuss.

Collecting audio, image and video for the Thunderbirds system:

The images, audio and video for the Thunderbirds system will all be collected using a single digital camera. The 4 megapixel Sony camera used collects images at a resolution of 2304 by 1728 pixels and video at a resolution of 640 by 480 pixels. Each JPEG image requires approximately 1.7MB of storage on the removable Memory Stick. The MPEG video is captured at 25 frames per second and includes a single audio track recorded using 16-bit samples at a frequency of 32kHz.

The Sony camera is unable to capture just audio, therefore video will be captured and the audio track will be extracted at a later stage. Image and video files to be collected include the main island image, images of each of the Tracey boys with their vehicles

and video of each vehicle launching. In addition audio will be extracted from video of each of the various sounds made by the vehicles and also from the control panel.

GROUP TASK Discussion
The Thunderbirds presentation will be uploaded to a web server for use over the WWW. Discuss likely issues if the collected files are used without further processing.

GROUP TASK Practical Activity
Using the Internet, or otherwise, identify suitable software applications that are able to extract the sound track from MPEG video files. Use one of the applications to extract audio from a video file.

STORING AND RETRIEVING MULTIMEDIA CONTENT

When designing multimedia presentations choosing suitable file formats and applying suitable compression is a significant consideration. Firstly the end user's information technology must be able to decompress and display the selected formats and secondly the display must occur in a timely fashion. This is particularly critical when the presentation will be distributed over the World Wide Web where communication speeds vary considerably and the display hardware is largely unknown. Response times that exceed more than a few seconds should be avoided. When this is not possible then feedback in the form of progress bars should be considered – progress bars can take many forms, however their main purpose is to indicate the total wait time remaining.

During our earlier discussion on each of the media types we introduced many of the more common file formats and compression techniques. The following tables summarise these and other file formats according to media type and compression. Recall that lossy compression techniques remove some of the original data whilst lossless compression retains all of the original data.

Bitmap image file formats

File format	Compression	Comments
Windows Bitmap (BMP)	Lossless	Microsoft Windows default bitmap format. Usually files are not compressed however run length encoding (RLE) is supported. Bit depths of 1, 4, 8 and 24 bits/pixel are supported.
Joint Photographics Expert Group (JPG or JPEG)	Lossy	Popular compressed format for photographic images on the web and elsewhere. All web browsers are able to display JPEG images. 8-bit greyscale and 24-bit true colour bit depths are supported at various levels of compression.
Graphics Interchange Format (GIF)	Lossless	Common format for banners and logos on the web. Supports only 8 bits/pixel in either greyscale or colour. All GIF files are compressed using LZW lossless compression – a system similar to RLE named after its developers Lempel, Ziv and Welch.
Portable Network Graphics (PNG)	Lossless	Originally designed to replace GIF on the web. Supports up to 48-bit true colour and 16-bit greyscale. Includes a variable transparency alpha channel so graphics can have semi-transparent shadows, for example.
Tagged Image File Format (TIF or TIFF)	Lossless	Standard format for storing professional quality images. A single TIF file can contain many other embedded image files, including vector images. LZW compression can be used. All bit depths up to 48 bits/pixel are supported.

Within multimedia systems most bitmap images are displayed on screens rather being printed. As a consequence it is important to scale bitmap images to a resolution suited to screen display. Most digital cameras and also scanners are able to collect bitmaps with resolutions far exceeding the resolution of most screens. These images should be scaled down to reduce their resolution to a more appropriate size. Currently screen resolutions exceeding 1900 by 1200 pixels are rare and hence there is little point including higher resolution images within most multimedia presentations.

Consider images for the Thunderbirds system:

The images for the Thunderbirds presentation were collected by the Sony camera as 2304 by 1728 pixel true colour JPEGs that each required approximately 1.7MB of storage. The main island image file (island.jpg) was cropped and then scaled down to a resolution of 922 by 578 pixels and now occupies approximately 140kB of storage. Each of the other images were also cropped and then scaled down to a more suitable screen display resolution. A list of the final JPEG images together with their file sizes is reproduced in *Fig 6.85*.

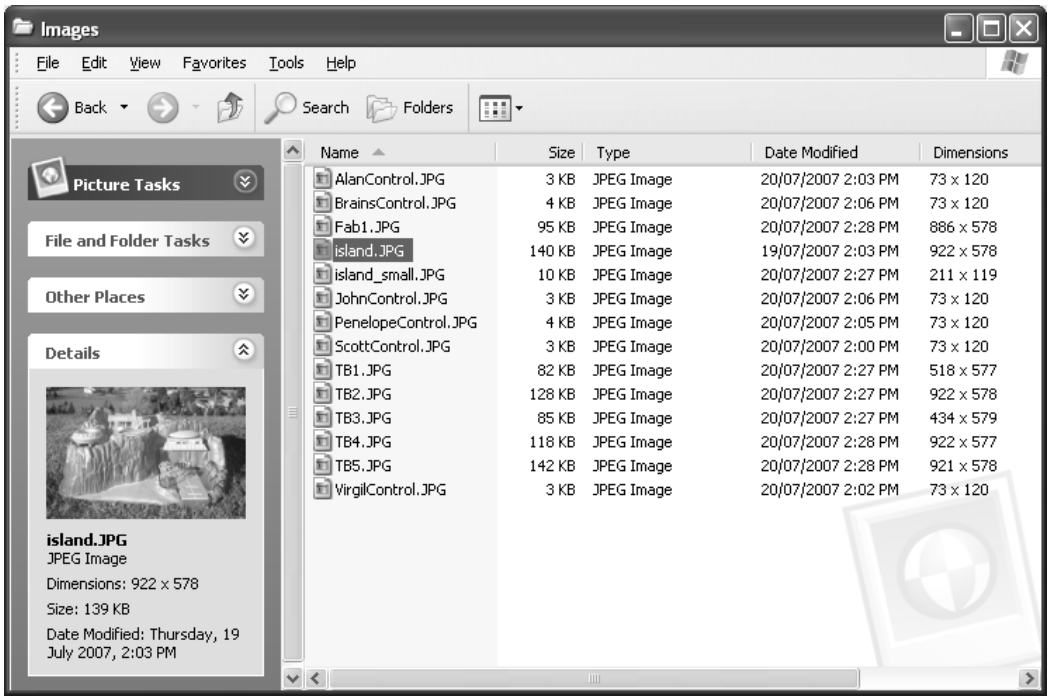


Fig 6.85
Final JPEG images for use within the Thunderbirds presentation.

GROUP TASK Discussion
Explain why JPEG files with identical resolution and colour depth require different amounts of storage.

GROUP TASK Activity
Calculate the storage that would be needed for one of the original files if it had not been compressed. Now calculate the approximate compression ratio used by the Sony camera. Perform the same calculations using some of the final scaled images in *Fig 6.85*.

Vector image file formats

Each of the following formats is more accurately described as metafile format. This means they describe the content using a variety of different text tags much like HTML tags describe the content of a webpage. As a consequence these formats can contain descriptions of the individual lines, shapes, fill patterns and colours within a vector image but they can also include text and even bitmap images. For example a SVG file can include an embedded compressed JPEG image. The SVG tags describe the precise location, orientation and size of the JPEG within the final image.

File format	Compression	Comments
Windows Metafile (WMF, EMF)	None	A Microsoft format that due to its widespread use can be read and written by many other operating systems.
Portable Document Format (PDF)	None	Adobe PDF files are commonly used to distribute electronic versions of printed material. PDF files accurately describe the layout of pages, however they can also include single vector images and also a variety of different interactive elements such as hyperlinks.
Scalable Vector Graphics (SVG)	None	A format developed by the W3C (World Wide Web Consortium). The intention is for SVG to become the predominant format for vector graphics on the web. All web browsers will support the SVG format.
Small Web Format or Shockwave Flash (SWF)	None	A flexible metafile format that can be used for vector images, animation and also video. Most computers with a web browser installed also have a Flash player installed.

Vector image files generally require significantly less storage space compared to similar bitmap images. Furthermore vector images can be resized to any resolution without loss of clarity.

Audio file formats

File format	Compression	Comments
Waveform Audio Format (WAV)	Lossy, lossless or none	Microsoft's WAV format is a metafile format able to include raw or compressed audio data. Various lossless and lossy audio codecs can be used. Common codecs include PCM (lossless) and MPEG-1 Layer 3 (lossy).
Audio Interchange File Format (AIFF)	Lossy, lossless or none	Apple's audio format for the Macintosh. Most AIFF files contain raw sound samples, however they can also include data compressed using either a lossy or lossless codec. AIFF files can also contain note, tempo and pitch data alongside the sound samples.
MPEG-1 Audio Layer 3 (MP3)	Lossy	Popular compressed format for electronic distribution of commercial music files. The lossy compression removes many sounds that would not be noticed by most people.
Windows Media Audio (WMA)	Lossy	Microsoft format designed as a competitor to the popular MP3 format.
Musical Instrument Digital Interface (SMF, KAR, MIDI, MID)	Lossless or none	Specifies each note, tone and perhaps instrument. Primarily used to communicate with synthesisers and digital instruments. Karaoke (KAR) files include lyrics that can be displayed as the music plays.

Sampled audio files are composed of a sequence of sound samples. In terms of storage and retrieval the number of channels, samples per second (sample frequency) and the number of bits used to represent each sample (bit depth) will clearly affect the storage size of audio files. For example a mono (single channel) sound file requires half the

storage of a stereo sound at the same sample frequency and bit depth. Similarly halving the frequency or halving the bit depth will also halve the file size. It is important to determine the raw sample rate, bit depth and number of channels within the raw collected sound. There is little point increasing any of these parameters beyond that of the collected data. For instance if audio is collected using a microphone and sound card at a sample frequency of 24kHz then using software to increase the sample frequency to 48kHz will double the file size, furthermore the added samples are approximations that may actually reduce the quality of the final sound. In general audio should be recorded at the highest sample frequency and bit depth. Audio software can then be used to reduce file size by lowering the sample frequency and bit depth. Such processing is a compromise between sound quality and file size – experimentation is often required to achieve the desired result.

Consider audio for the Thunderbirds system:

The audio for the Thunderbirds presentation was originally collected by the Sony camera within MPEG video files. The software used to extract the audio from the video files created stereo WAV files containing 16-bit samples at a sample frequency of 48kHz. Parameters for these WAV files were then altered using the sound recorder utility included with the Windows operating system.

Details of one of the original WAV files (fab1.wav) extracted from the video and also three altered versions are reproduced in *Fig 6.86*. The original fab1.wav file required 453kB of storage, the altered fab1_V2.wav occupies 114kB, fab1_V3.wav requires just 8kB and fab1_V4.wav only 3kB of storage. After listening to each file it was decided to use the fab1_V3.wav file.

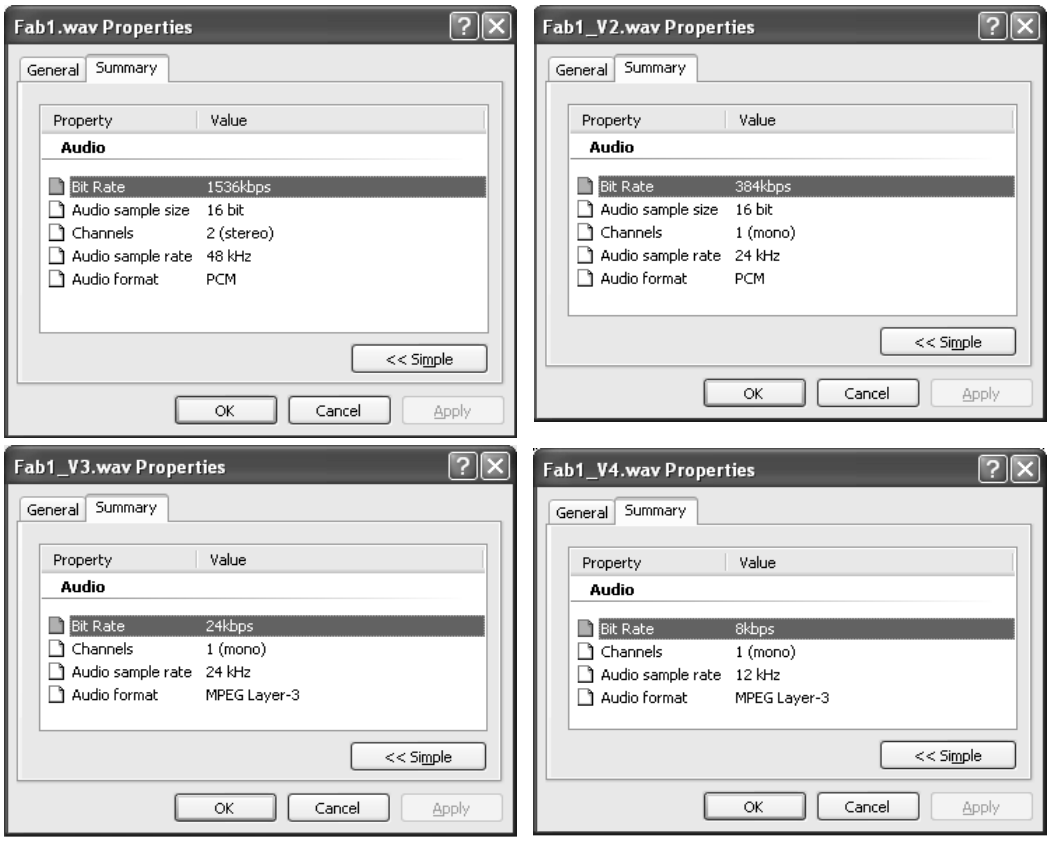


Fig 6.86
Properties of original and altered versions of fab1.wav audio file.

GROUP TASK Discussion
Compare the properties of each of the files in *Fig 6.86*. Perform calculations that explain the differences in file size and compression ratios achieved.

GROUP TASK Discussion
The Bit Rate for each of the files is shown in *Fig 6.86*. Bit Rate is the speed at which the data must be received during play back. Calculate the duration of each version of the audio clip. Your answer for each file should be the same – explain any differences.

GROUP TASK Practical Activity
Record an audio clip and then adjust the sample rate, sample size and others parameters to produce progressively smaller files. Listen to each file and decide where the quality deteriorates below acceptable levels.

Video and animation file formats

Uncompressed raw video files are enormous and lossless compression techniques rarely reduce this size significantly. As a consequence uncompressed video is only used during development whilst for distribution and display virtually all video data is compressed using lossy techniques – most video codecs perform processes similar to the block-based technique described earlier in this chapter.

Animation is usually stored in one of the common video formats for ease of playback – the animation is converted to a sequence of video frames. When the animation is composed of vector type images or includes interactive features then a specialised format such as Flash is often used. Such files are significantly smaller than similar compressed video files but require a specialised player for display.

File format	Compression	Comments
Motion Picture Experts Group (MPG, MPEG)	Lossy	Common file format that usually contains video compressed using one of the earlier MPEG video codecs.
MPEG-4 Layer 14 (MP4, M4A, M4P)	Lossy	Supports a large variety of resolutions and frame rates. Often used on portable devices including PDAs and iPods.
Audio Video Interleave (AVI)	Usually lossy	Older container format created by Microsoft and IBM.
QuickTime (MOV, QT)	Usually lossy	Apple format from which the current MP4 standard was developed. QuickTime can also store interactive media of various types.
Windows Media Video (WMV)	Lossy	Microsoft format often used for streaming video data.
Animated GIF (GIF)	Lossless	Used for small animations on websites. No audio possible. Support is built into web browsers.
Small Web Format or Shockwave Flash (SWF)	None, lossy or lossless	Flexible format that can contain video, animation and many other interactive features. Requires Flash player on end user's computer.
Flash Video (FLV)	Lossy	Popular format for delivering streamed video over the web for display within Flash player.

Most of the above video file formats are known as container formats, this means they can contain data compressed using any of a variety of available video codecs. The end user's computer must have a copy of the approriate codec installed to playback compressed video in any of these formats. Currently the most popular video codecs are defined by the Motion Picture Expert Group (MPEG), however many others, such as DivX, Cinepak and Intel's Indeo codecs are also common. Most, but not all, video files use different codecs for the video and audio tracks.

If a multimedia presentation will be distributed widely, such as on optical disk or over the Internet, then it is advisable to ensure both the video and audio tracks within each video file are compressed using codecs that are installed with all popular operating systems or media players. Furthermore the frame resolution, colour depth and frame rate should be adjusted to suit the devices and screen sizes used for display. For example reducing the resolution of the video from 640 by 480 pixels down to 320 by 240 pixels will reduce file sizes to approximately one quarter of their orginal size.

Consider video for the Thunderbirds system:

The video footage for the Thunderbirds presentation was originally collected by the Sony camera as MPEG files at 25fps, a resolution of 640 by 480 pixels and colour depth of 24-bits. Each video was trimmed to remove excess footage and then converted to a resolution of 320 by 240 pixels and then saved as a WMV file. For example the initial TB1.mpg file contained approximately 16 seconds of footage and required approximately 5.7MB of storage. Using Windows Movie Maker (see *Fig 6.87*) the video was trimmed to 8 seconds of footage and then converted to a 320 by 240 pixel WMV file with a speed of 15fps – the resulting WMV file required just 179kB of storage.

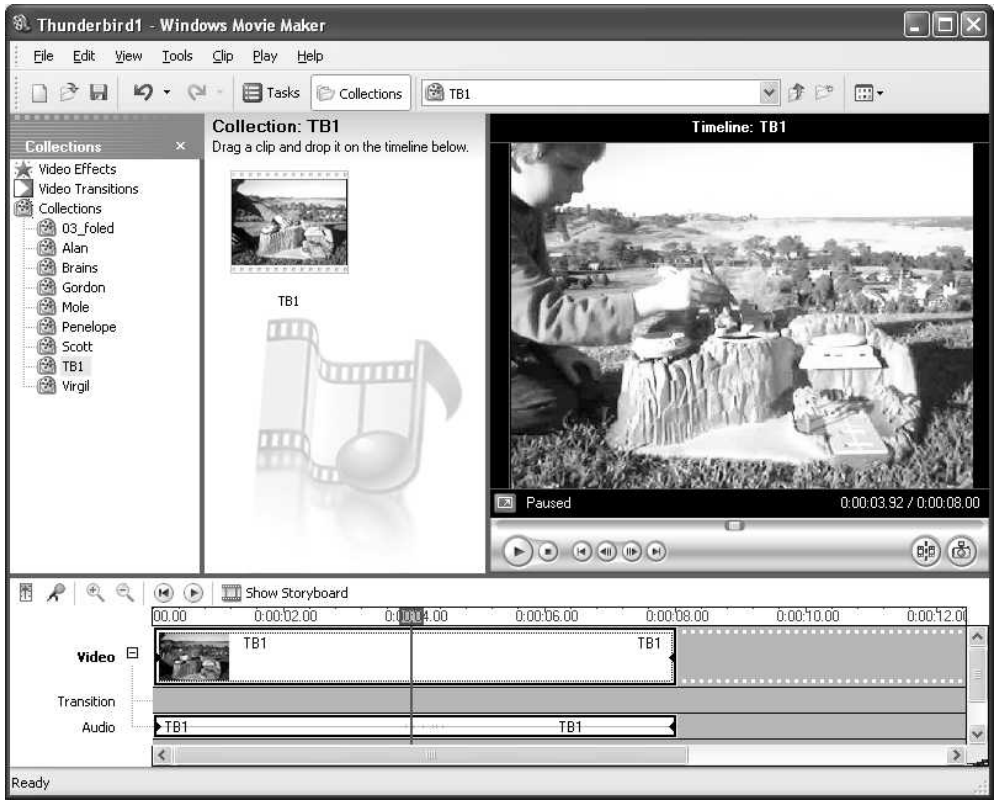


Fig 6.87
Editing video footage using Window's Movie Maker.

GROUP TASK Activity
Calculate the size of the original TB1 video file if it had not been compressed at all. Determine the approximate compression ratio as a result of the MPG codec within the Sony camera. Now determine the compression ratio applied to the final 179kB WMV file.

GROUP TASK Practical Activity
Collect some video footage and then convert the file into various different resolutions and frame rates. Compress the files using different video codecs. Construct a table (and perhaps a graph) comparing the resulting file size, perceived video quality, resolution, frame rate and codecs.

PROCESSING TO INTEGRATE MULTIMEDIA CONTENT

The final multimedia presentation is created using a suitable software application. Presentation software such as Microsoft Powerpoint would be used to produce slideshow style presentations. A word processor could be used to create files with embedded sound, image or video clips. Specialised authoring packages suited to the particulars of the multimedia system can be used, for example Articulate's Quizmaker for creating surveys and quizzes. More general authoring packages such as Adobe Flash are used to combine a variety of media into a single interactive Flash file. If the multimedia system will be distributed over the World Wide Web then an HTML editor or for simple web pages a text editor such as notepad can be used. All these software applications are used to combine and link all the multimedia content into an integrated and interactive multimedia presentation.

The steps and specific tasks vary according to the individual software application used. Some general tasks performed to combine and link multimedia content include:

- Import existing content into the application. Often a library or collection of media files is created by the application. Such libraries can be arranged into a directory structure, for example separate folders for audio, image and video files.
- Create screens, add and format content and create hyperlinks. In many cases textual content, in particular headings, titles, instruction and navigational elements are created directly within the authoring software. The precise location, size and behaviour of each media element is specified. For instance, should sound and video play immediately and should volume and other controls be displayed. Hyperlinks are specified to link screens and media elements. For example hyperlinks to open or play sound and video files.
- Create the final file or files required for distribution and display. In some authoring packages the final presentation must be compiled into a complete integrated package. For instance a single Flash file can contain video, sound and image files together with rich interactive features that link the content together. In other packages a number of separate files are distributed. For instance a web site includes an HTML file for each web page and various directories containing the media files displayed on the web pages. For some multimedia systems it maybe appropriate to compress the entire presentation into a single file, the aim being to further reduce storage size and also to simplify distribution.

GROUP TASK Discussion
List specific tasks performed to integrate multimedia content using an authoring or HTML editor package with which you are familiar.

Consider the Thunderbirds system:

The Thunderbirds system will be developed for display within a web browser where each screen will be implemented as a separate HTML file. We will not produce a screen for Thunderbird 4 as no audio or images of John were available. The navigation map that formed part of the initial storyboard (refer *Fig 6.74*) included a total of eight screens of which we will create seven. Screens will be added to display each of the five launch videos – with a link back to the main island screen. In total twelve HTML files are needed.

We will develop the HTML code for the screens using Windows Notepad to clearly illustrate the HTML tags required. The media files are arranged into separate folders for audio, images and videos. *Fig 6.88* shows listings of all of the final files within the presentation.

Name	Size	Type
audio		File Folder
images		File Folder
videos		File Folder
control.html	1 KB	HTML Document
fab1.html	2 KB	HTML Document
fab1launch.html	1 KB	HTML Document
island.html	1 KB	HTML Document
tb1.html	3 KB	HTML Document
tb1launch.html	1 KB	HTML Document
tb2.html	3 KB	HTML Document
tb2launch.html	1 KB	HTML Document
tb3.html	3 KB	HTML Document
tb3launch.html	1 KB	HTML Document
tb5.html	3 KB	HTML Document
tb5launch.html	1 KB	HTML Document

Name	Size	Type
Fab1Audio1.wav	10 KB	Wave Sound
Fab1Audio2.wav	14 KB	Wave Sound
Fab1Audio3.wav	12 KB	Wave Sound
TB1Audio1.wav	6 KB	Wave Sound
TB1Audio2.wav	9 KB	Wave Sound
TB1Audio3.wav	6 KB	Wave Sound
TB2Audio1.wav	7 KB	Wave Sound
TB2Audio2.wav	5 KB	Wave Sound
TB2Audio3.wav	7 KB	Wave Sound
TB2Audio4.wav	7 KB	Wave Sound
TB2Audio5.wav	17 KB	Wave Sound
TB3Audio1.wav	9 KB	Wave Sound
TB3Audio2.wav	9 KB	Wave Sound
TB3Audio3.wav	10 KB	Wave Sound
TB5Audio1.wav	5 KB	Wave Sound
TB5Audio2.wav	12 KB	Wave Sound
TB5Audio3.wav	10 KB	Wave Sound

Name	Size	Type
AlanControl.JPG	3 KB	JPEG Image
BrainsControl.JPG	4 KB	JPEG Image
Fab1.JPG	96 KB	JPEG Image
island.JPG	140 KB	JPEG Image
island_small.JPG	10 KB	JPEG Image
JohnControl.JPG	3 KB	JPEG Image
PenelopeControl.JPG	4 KB	JPEG Image
ScottControl.JPG	3 KB	JPEG Image
TB1.JPG	82 KB	JPEG Image
TB2.JPG	128 KB	JPEG Image
TB3.JPG	85 KB	JPEG Image
TB4.JPG	118 KB	JPEG Image
TB5.JPG	142 KB	JPEG Image
VirgilControl.JPG	3 KB	JPEG Image

Name	Size	Type
fab1.wmv	811 KB	Windows Media Audio/Video file
tb1.wmv	1,256 KB	Windows Media Audio/Video file
tb2.wmv	1,123 KB	Windows Media Audio/Video file
tb3.wmv	858 KB	Windows Media Audio/Video file
tb5.wmv	490 KB	Windows Media Audio/Video file

Fig 6.88
Listing of files within the sample Thunderbirds system.

GROUP TASK Discussion
When creating a website it is critical to setup a logical directory structure before creating individual web pages. Why is this? Discuss.

Main Island Screen

The HTML code within island.html is reproduced below. Essentially the single image called island.jpg is used to create an image map. Hyperlinks from rectangular regions within the image link to tb1.html, tb2.html. tb3.html and control.html. To determine the precise coordinate for each hyperlink region simply load the source image into a paint application and write down the coordinates, which usually display in the status bar as you move the mouse over the image. Text within the alt tags is displayed as the user places their mouse over the defined region.

```
<html>
<head>
<title>Tracey Island</title>
</head>
<body>
<img src="images/island.jpg" usemap="#tb" border="0">
<map name="tb">
<area shape="rect" coords="307,119,477,220" href="tb1.html" alt="Thunderbird 1">
<area shape="rect" coords="741,293,600,497" href="tb2.html" alt="Thunderbird 2">
<area shape="rect" coords="264,164,115,267" href="tb3.html" alt="Thunderbird 3">
<area shape="rect" coords="740,226,542,281" href="control.html" alt="Control Room">
</map>
</body>
</html>
```

Thunderbird Screens

The Thunderbird 2 HTML file called tb2.html contains the code that follows and uses the tb2.jpg image shown alongside. Each of the other Thunderbird vehicle screens contains similar code. Notice that the image map for this screen includes irregular regions defined using polygons. Furthermore one of the audio files plays as the user mouses over the associated region. The JavaScript language has been used to code functions that cause the audio to play – one of the three specified audio files is played at random.

```
<html>
<title>Thunderbird 2</title>
<body>
<!-- --><BGSOUND ID="auIEContainer">
<SCRIPT LANGUAGE="JavaScript"><!--
// Embed and play audio files
var aySound = new Array();
// Sound files
aySound[0] = "audio/tb2audio1.wav";
aySound[1] = "audio/tb2audio2.wav";
aySound[2] = "audio/tb2audio3.wav";
```

… (Javascript code not reproduced)

Fig 6.89
The image tb2.jpg contains 922 by 578 pixels.

```
//--></SCRIPT><!-- -->
<img src="images/tb2.jpg" usemap="#tb" border="0">
<map name="tb">
<area shape="rect" coords="209,1,1,120" href="island.html" alt="Tracey Island">
<area shape="polygon" coords="545,105,723,97,865,167,813,201,532,294,342,287,275,
178,227,182,218,252,233,290,146,302,127,143,348,131,395,195,518,177"
href="tb2launch.html" alt="Launch Thunderbird 2">
<area shape="polygon" coords="253,183,286,211,297,277,371,332,337,406,355,556,331,
556,263,496,254,420,208,371,231,301,247,282,225,246,232,201"
onMouseOver="playSound()" onMouseOut="stopSound()" alt="Virgil Tracey Speaking">
</map> </body> </html>
```

Thunderbird Launch Screens

The HTML code within the file tb2launch.html is reproduced below. The code for each of the other launch screens is similar.

```
<html>
<title>Launch Thunderbird 2</title>
<body>
<object id="MediaPlayer" width=640 height=526 classid="CLSID:22D6f312-B0F6-11D0-
94AB-0080C74C7E95" standby="Loading Windows Media Player components..."
type="application/x-oleobject" codebase="http://activex.microsoft.com/activex
/controls/mplayer/en/nsmp2inf.cab#Version=6,4,7,1112">
<param name="filename" value="videos\tb2.wmv">
<param name="Showcontrols" value="True">
<param name="autoStart" value="True">
<embed type="application/x-mplayer2" name="MediaPlayer" width=640 height=526>
</embed> </object>
<a href="island.html"> <img src="images/island_small.jpg" border="0" alt="Tracey
Island"> </a>
</body> </html>
```

GROUP TASK Discussion
Identify alterations so the tb2.html and tb2launch.html files to create HTML files for each of the other Thunderbirds.

Control Room Screen

The HTML code within the file control.html is reproduced below and the final screen within Internet Explorer is reproduced in *Fig 6.90*.

```
<html>
<title>Control Room</title>
<body>
<a href="island.html"> <img src="images/island_small.jpg" alt='Tracey Island'> </a>
<a href="tb1.html"> <img src="images/scottcontrol.jpg" alt="Scott Tracey"> </a>
<a href="tb2.html"> <img src="images/virgilcontrol.jpg" alt="Virgil Tracey"> </a>
<a href="tb3.html"> <img src="images/alancontrol.jpg" alt="Alan Tracey"> </a>
<a href="tb5.html"> <img src="images/johncontrol.jpg" alt="John Tracey"> </a>
<a href="fab1.html"> <img src="images/penelopecontrol.jpg" alt="Lady Penelope">
</a>
</body>
</html>
```

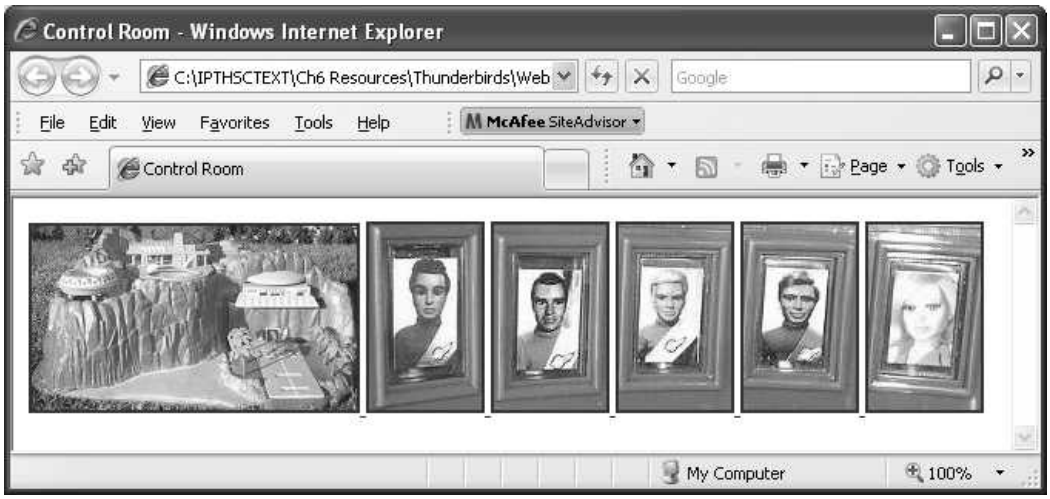


Fig 6.90
Control Room screen displayed within Internet Explorer.

GROUP TASK Practical Activity
Study the above example HTML image map code. Create an HTML image map that contains links to different external media files.

GROUP TASK Discussion
The JavaScript used on each of the Thunderbird screens was copied, pasted and then edited from an existing webpage found on the Internet. Discuss ethical considerations when copying such code from the Internet.

HSC style question:

Most commercial movie titles are now distributed on DVD. These titles include interactive features such as menus and even simple games. It is now possible for individuals to produce DVDs at home that include similar interactive features.

(a) Identify types of software you would use to design and create a DVD containing home movies and an interactive menu. Justify your selection of each type of software.

(b) Discuss developments in hardware that have enabled the production of interactive DVDs at home.

Suggested Solution

(a) Software used to create the DVD menu would be
- A graphics-editing program would be needed to create the background image or images for the menu system.
- Authoring package, with the capabilities to create the interactive menu so the user is able to select the various chapters (or clips) from the menu.
- Audio recording and editing software, so that music or background sound can be recorded or extracted from existing video footage. Such audio plays whilst the DVD menu is being displayed.
- Video editing package to retrieve video clips from the camera and then edit the clips prior to inclusion in the overall presentation.

(b) Hardware developments enabling interactive home movie production include:
- Digital Video cameras with improved quality and reduced cost have enabled people to film their movies using digital technology and then transfer the video directly to a computer.
- Fire Wire and high-speed USB interfaces have enabled high quality video to be captured directly from video cameras at high speed onto the computer's hard disk.
- DVD's with their large storage capacity mean a feature length movie to will fit on a single DVD. DVD's are direct access devices, which means that interactive features can be included.
- DVD burners are included on many home computers that allow home users to reproduce their movies at low cost from home.
- Increased storage capacity on HDDs allows for the capture of video from the camera and its subsequent editing.
- Increased CPU speed and increases in the amount of RAM means a typical home computer now has the processing power and primary storage needed to display and also edit high-resolution video files.

SET 6E

1. In regard to resolution when collecting image and video data which of the following is true?
   (A) Collect at a resolution lower than required for display.
   (B) Collect at a resolution higher than required for display.
   (C) Collect at a resolution identical to that required for display.
   (D) The resolution of the collected data is of no significance.
2. Which of the following components include the function of an ADC?
   (A) CCD image sensor.
   (B) microphone.
   (C) CMOS image sensor.
   (D) LED.
3. A digital camera takes pictures with a resolution of 2304 by 1728 pixels. The size of each JPEG file is approximately 1.7MB. Which of the following best describes this camera?
   (A) It's a 4 mega pixel camera that uses lossy compression.
   (B) It's a 2 mega pixel camera that uses lossy compression.
   (C) It's a 2 mega pixel camera that uses lossless compression.
   (D) It's a 4 mega pixel camera that uses lossless compression.
4. Which of the following images requires the least storage?
   (A) 640 by 480 pixels, 24 bits per pixel.
   (B) 1024 by 768, 16 bits per pixel.
   (C) 1600 by 1200, 8 bits per pixel.
   (D) 1600 by 900, 1 bit per pixel.
5. Examples of bitmap image file formats includes:
   (A) BMP, JPEG, WMF, WAV.
   (B) JPEG, BMP, TIFF, GIF.
   (C) SVG, WMF, SWF, PDF.
   (D) MP#, MID, WAV, WMA.
6. Lossy compression is inappropriate for vector images because:
   (A) they are small enough already.
   (B) removing any data would destroy a complete shape description.
   (C) the component shapes have already been compressed as they are saved.
   (D) it would be inefficient during decompression to recreate the missing information.
7. Sound waves are a type of:
   (A) electromagnetic wave.
   (B) compression wave.
   (C) transverse wave.
   (D) tidal wave.
8. Which of the following would NOT reduce the storage size of a sampled audio file?
   (A) Decreasing the sample size.
   (B) Decreasing the sample rate.
   (C) Decreasing the number of channels.
   (D) Decreasing the volume.
9. Most digital cameras collect either red, blue or green values for each pixel. What is the name of the process and filter used to determine each of the other colour values for each pixel?
   (A) Interlacing and RGB filter.
   (B) Interpolation and YCrCb filter.
   (C) Demosaicing and RGB filter.
   (D) Demosaicing and Bayer filter.
10. A raw audio 12MB audio file contains stereo sound recorded at 48kHz using 16-bit samples. Audio software is used to reduce the sample frequency to 24kHz and the sample size to 8-bits. The audio is then saved as an MP3 file requiring just 200kB of storage. The MP3 compression ratio for this file is approximately:
    (A) 10:1
    (B) 15:1
    (C) 60:1
    (D) 100:1
11. Describe the organisation of each of the following storyboard layouts. Provide an example of a multimedia system where each layout would be appropriate.
    (a) Linear
    (b) Hierarchical
    (c) Non-linear
    (d) Composite or combination of others.
12. Explain how each of the following devices captures analog data and transforms it into digital files.
    (a) Flatbed scanner
    (b) Digital camera
    (c) Microphone and sound card
    (d) Video camera
13. For each of the following media types, identify a file format and explain how data is compressed using this format.
    (a) Sampled audio
    (b) Bitmap image
    (c) Video
14. Analyse an existing multimedia system. Briefly describe the system and the likely hardware and software used during the development of this system.
15. Based on an image of your own choice, develop an HTML image map that links portions of the image to relevant existing web pages on the World Wide Web.

ISSUES RELATED TO MULTIMEDIA SYSTEMS

The relative ease with which digital content of all types can be copied presents a variety of issues. For those who produce content it is difficult to enforce their copyrights. For those who wish to use content it can often be difficult to determine the source or owner of the content's copyrights. Furthermore individuals are able to create content of all types and distribute it globally at minimal cost. This makes it difficult to verify the correctness or integrity of information. The rapid development and subsequent introduction of new technologies continually changes how multimedia can and is delivered. Those involved in developing multimedia systems must be technologically aware so they are able to make the best use of new and emerging technologies.

COPYRIGHT ISSUES

Copyright laws are used to protect the legal rights of authors of original works. The *Copyright Act 1968*, together with its various amendments, details the laws governing copyright in Australia. Copyright laws are designed to encourage the creation of original works by limiting their copying and distribution rights to the copyright owner. The copyright owner is normally the author of the work, except when the work was created as part of the author's employment; in this case the employing organisation owns the copyrights. Without copyright laws there would be little economic incentive for authors to create new works.

Copyright does not protect the ideas or the information within a work, rather it protects the way in which the idea or information is expressed. For example, there are many software products that perform similar processes, however these processes are performed in different and original ways, hence copyright laws apply. Generally copyright protection continues for the life of the author plus a further fifty years.

All works are automatically covered by copyright law unless the author specifically states that the copyrights for the work have been relinquished. The use of the familiar copyright symbol ©, together with the author's name and publication date is not necessary, however its use is recommended to assist others to establish the owner of a work's copyrights.

When the right to use material is granted (or has been purchased) the copyright holder should always be acknowledged. This confirms compliance and also assists readers to establish the source and integrity of any material within the presentation.

Computer software, data and information is easily copied, and the copy is identical to the original. This is not the case with most other products. As a consequence special amendments to the Copyright Act have been enacted.

In regard to software:

- One copy may be made for backup purposes.
- All copies must be destroyed if the software is sold or otherwise transferred.
- Decompilation and reverse engineering is not permitted. The only exception being to understand the operation of the software in order to interface other software products.

In regard to compilations of information (such as collected statistics, databases of information and multimedia compilations):

- The information itself is not covered.
- There must have been sufficient intellectual effort used to select and arrange the information; or
- The author must have performed sufficient work or incurred sufficient expense to gather the information even though there was no creativity involved.

Consider the Thunderbirds system:

- The video, images and audio used were all collected from SoundTech's Tracey Island toy.
- Javascript to play the random audio files was obtained and modified from a website that performed similar functions.
- Various copyrighted software products were used to create the Thunderbirds system. This included specialised image, video and audio editing software and also a utility to extract audio from video.
- Video and audio files within the presentation were compressed using codecs written by various other companies.

GROUP TASK Discussion
Do you think copyright law applies to each of the above points? How could the legal right to use each of the above be determined? Discuss.

INTEGRITY OF SOURCE DATA

When studying Information Systems and Databases in chapter 2 we discussed the need to acknowledge all data sources and also techniques for assessing the accuracy and reliability of data. With regard to multimedia systems it is common for content to be derived from a variety of different sources. This makes the job of verifying the correctness of the presented information more difficult. When developing multimedia systems, particularly educational systems, it is critical to include references detailing the source of the data. Users of the multimedia presentation should be able to easily determine the source so they can verify its correctness and perform further related research.

Consider the following:

Each of the following situations includes issues with regard to determining the correctness or integrity of information.

- Manufacturer's websites often include links to various external reviews of their products.
- Many multimedia products include excerpts and clips extracted from original source material that do not accurately reflect the original source information.
- Wikipedia is a collaborative online encyclopaedia where most articles can be edited by anyone with Internet access.
- Software is widely available and used that allows audio, in particular MP3 files and video files to be shared between users over the Internet.
- Many web sites and other multimedia do not contain references detailing the author or copyright owner of their content.
- Searching for information to explain a particular topic will often yield conflicting results even when each result is from a reputable and verifiable source.

GROUP TASK Discussion
Identify issues within each of the above situations that cause concern in regard to the integrity of the source data. Suggest strategies to assist in establishing the integrity of the source data.

CURRENT AND EMERGING TRENDS IN MULTIMEDIA SYSTEMS

Consider the following:

Some recent and emerging technological developments related to multimedia systems include:

- RSS (Really Simple Syndication) feeds and also Podcasts where the content provider updates the feed on a regular basis and subscribers computers or mobile devices are automatically updated. Such feeds commonly include audio, however feeds that include video are becoming more popular.
- Delivering pay TV using DSL technology and existing copper phone lines may soon become a reality. Current cable TV and satellite systems transmit all channels and the user's receiver decodes only the channel being viewed. Using DSL technology together with efficient MPEG-4 compression one or two channels could easily be transmitted through a single DSL connection using existing copper telephone lines.
- Many mobile phones, such as the Nokia N95 in *Fig 6.91*, integrate the functions of a digital camera, GPS navigator, MP3 player, Internet browser and a variety of other applications and features. Video calls are now commonplace and network coverage continues to expand and operate at ever increasing speeds.
- Intel, the largest designer and manufacturer of microprocessors, continues to reduce the physical size and increase the processing power of its microprocessors. Currently (2007) Intel prototype chips have been produced that include more than 80 core processors, are smaller than a postage stamp and yet have more processing power than most 1990s supercomputers.

Fig 6.91
Nokia N95 multimedia phone with GPS receiver.

- Seemless integration and communication between a variety of different devices and networks allows content to delivered to a broader audience. For example Skype allows rich multimedia communication between mobile phone, traditional telephone, wireless and Internet networks.
- Software is now likely to be distributed over the web as a service rather than as a product that one must purchase. This is one of the features of Web 2.0 technologies. The software and the data is integrated such that users do not see a clear distinction between the two. Furthermore the software and data that results in displayed multimedia content can be integrated and linked within other multimedia presentations.
- Multimedia systems are created in real time using data within linked databases. For instance different sets of metadata such as cascading style sheets (CSS) can be stored in a database along with the actual content. This allows not just the data but also the look and feel and feel of multimedia systems to change automatically.

GROUP TASK Discussion
Analyse each of the above dot points in terms of their effect on the development and/or display of multimedia systems.

GROUP TASK Research
Web 2.0 is not a new set of technologies, rather it is more accurately described as a new way of using web technologies. Research Web 2.0 technology to identify some of its defining features.

GROUP TASK Discussion
Multimedia is commonplace - we take it for granted. Brainstorm common examples of multimedia that simply did not exist 10 years ago.

Consider virtual worlds:

A virtual world is an online simulated environment where people take on another persona using avatars. Some virtual worlds, such as "Second Life" are largely a simulation of the real world. They contain houses, cars, shopping malls, night clubs, and an economy with its own currency and businesses run by the inhabitants. Other virtual worlds are a logical extension of multiplayer online games. Most virtual worlds operate 24 hours a day seven days a week.

Although virtual worlds are developed primarily for entertainment, other uses are now (2007) starting to emerge. For instance real business meetings can take place in a virtual world, which allow 3D interactions despite participants living on different continents. People can overcome their disabilities in a virtual world – most avatars are young, athletic and seem to never get sick. Companies can trial new products without the need to build physical samples. No doubt numerous other applications will soon emerge.

Real Hope in a Virtual World

Online Identities Leave Limitations Behind

After suffering a devastating stroke four years ago, Susan Brown was left in a wheelchair with little hope of walking again. Today, the 57-year-old Richmond woman has regained use of her legs and has begun to reclaim her life, thanks in part to encouragement she says she gets from an online "virtual world" where she can walk, run and even dance.

John Dawley III, who has a form of autism that makes it hard to read social cues, learned how to talk with people more easily by using his computer-generated alter ego to practice with other cyber-personas.

These increasingly sophisticated online worlds enable people to create rich virtual lives through "avatars" -- identities they can tailor to their desires: Old people become young. Infirm people become vibrant. Paralyzed people become agile.

They walk, run, and even fly and "teleport" around vast realms offering shopping malls, bars, homes, parks and myriad other settings with trees swaying in the wind, fog rolling in and an occasional deer prancing past. They schmooze, flirt and comfort one another using lifelike shrugs, slouches, nods and other gestures while they type instant messages or talk directly through headsets.

Because the full-color, multifaceted nature of the experience offers so much more "emotional bandwidth" than traditional Web sites, e-mail lists and discussion groups, users say the experience can feel astonishingly real. Participants develop close relationships and share intimate details even while, paradoxically, remaining anonymous. Some say they open up in ways they never would in face-to-face encounters in real support groups, therapy sessions, or even with family and close friends in their true lives.

Extract of an article By Rob Stein, Washington Post Staff Writer, October 6, 2007

GROUP TASK Discussion
Read the above article and also research other current applications of virtual worlds. Do you think these virtual worlds enhance all people's ability to interact socially? Discuss and debate.

CHAPTER 6 REVIEW

1. Which of the following are coding systems for representing text in binary?
   (A) MPEG, JPEG, MP3.
   (B) ASCII, EBCDIC, Unicode.
   (C) TrueType, Outline, Raster.
   (D) RLE, Huffman, Block-based.
2. The data IIIIIPPPPPPPPTTTT is compressed and stored as 5I8P4T. Which of the following describes the compression used?
   (A) Lossless RLE.
   (B) Lossless Huffman
   (C) Lossy RLE
   (D) Lossy Huffman
3. Creating different mouth shapes to animate a character's speech is an example of:
   (A) cel-based animation.
   (B) path-based animation.
   (C) both cel and path based animation.
   (D) a timeline.
4. <img src="fred.jpg">
   Which of the following best describes the purpose of this HTML code?
   (A) The image me.jpg is displayed as a hyperlink to the image fred.jpg.
   (B) The image fred.jpg is displayed as a hyperlink to the image me.jpg.
   (C) The image fred.jpg is displayed as a hyperlink to the www.me.com website.
   (D) The www.me.com website is displayed with a hyperlink to the image fred.jpg.
5. An image is scaled such that its width is halved but its height remains the same. This is an example of:
   (A) warping.
   (B) morphing.
   (C) cropping.
   (D) distorting.
6. A small animated banner on a website displays a sequence of five images containing a total of 256 colours. The file format used is likely to be which of the following?
   (A) GIF
   (B) JPEG
   (C) SWF
   (D) BMP
7. A 30 second video is collected at 15fps, has a resolution of 320 by 240 pixels and a colour depth of 24 bits. What is the approximate size of the uncompressed file?
   (A) 800kB
   (B) 100kB
   (C) 800MB
   (D) 100MB
8. A fighter jet includes a transparent display overlaying the real view through the windscreen. This display is an example of:
   (A) a head set.
   (B) virtual reality.
   (C) a head-up display.
   (D) a simulation.
9. What is the function of the polarising panels within LCD screens?
   (A) To ensure light passes through unhindered.
   (B) To alter the orientation of the liquid crystals.
   (C) To restrict the light entering and leaving to particular angles.
   (D) To support the TFTs, filter and liquid crystals.
10. Doubling the pixel width and pixel height of a bitmap image and also doubling the bit depth will increase the file size by:
    (A) 2
    (B) 4
    (C) 6
    (D) 8
11. Explain how each of the following media types is represented in digital form.
    (a) Text
    (b) Hypertext
    (c) Audio
    (d) Images
    (e) Video
12. Describe how each of the following hardware devices operate.
    (a) CRT screen
    (b) LCD screen
    (c) Projector
    (d) CD-ROM drive
    (e) Speakers and sound card
    (f) Touch screen
13. Describe compression techniques commonly used for each of the following media types.
    (a) Text
    (b) Bitmap images
    (c) Sampled audio
    (d) Video
14. Discuss effects of the widespread use of digital media on traditional radio, television and telephone communication.
15. Outline the processes and personnel involved during the development of large commercial multimedia systems.

GLOSSARY

1NF	See first Normal Form.
2NF	See second normal form
3NF	See third normal form.
acceptance test	A formal test conducted to verify whether or not a system meets its requirements.
active listening	A strategy involving various feedback techniques that aims to improve the understanding of the intended message from the speaker.
ADC	Analog to Digital Converter
ADSL	Asymmetrical digital subscriber line. A common implementation of DSL.
agile methods	A development approach that places emphasis on the team developing the system rather than following predefined structured development processes.
amplitude	The height of a wave. For audio the amplitude determines the volume or level of the sound.
analog	Continuous. Analog data can take any value within its range.
analysing	The information process that transforms data into information.
anchor tag	An HTML tag that is used to specify all the links within and between web pages.
application software	Software that performs a specific set of tasks to solve specific types of problems.
ASCII	American Standard Code for Information Interchange.
asymmetrical	Not symmetrical. Communication in each direction occurs, or can occur, at a different speed.
asynchronous	Not in time. Communication that does not attempt to synchronise the sender and receivers clock signals. Also called 'start-stop' communication.
audit trail	A system that allows the details of any transaction to be traced back to its origin.
authentication	The process of determining if someone or something is who they claim to be.
backup	To copy files to a separate secondary storage device as a precaution in case the first device fails or data is lost.
bandwidth	The difference between the highest and lowest frequencies in a transmission channel. Expressed in hertz (Hz), usually kilohertz (kHz) or megahertz (MHz).
baud rate	The number of signal events occuring each second along a communication channel. Equivalent to the number of symbols per second.
Bayer filter	A filter used on many CCD based digital cameras. Bayer filters alternate red and green rows with blue and green rows.
bias	An inclination or preference towards an outcome. Bias unfairly influences the outcome.
bit	Binary digit, either 0 or 1.
bitmap image	A method of representing an image as individual picture elements (pixels).
block based encoding	A system for compressing video data.
Boolean operator	An operator that acts upon Boolean variables and values.
boundary	The delineation between a system and its environment.
bps	Bits per second. A measurement of the speed of communication.
break-even point	The point in time when a new system has paid for iteslf and begins to make a profit.

broadband	A transmission medium that carries more than one transmission channel. Each channel occupies a distinct range of frequencies.
browser	A software application that interprets HTML code into text, graphics and other elements seen when viewing a web page from a web server.
buffer	A storage area used to assist the movement of data between two devices operating at different speeds.
byte	8 bits.
cable modem	A modem used to connect to a broadband coaxial network.
cache	A small amount of faster memory that is used to speed up access times to a larger and slower type of memory.
CCD	Charged coupled device.
CCITT	International telegraph and telephone consultative committee. The organsation responsible for maintaining the rules for encoding fax transmissions.
CD-R	Recordable compact disk that can only be written to once.
CD-RW	Rewriteable comapct disk.
cel-based animation	A sequence of cels (images) with small changes between each cel. When played the illusion of movement is created.
cell	The intersection of a row and a column within a spreadsheet.
centralised database	A single database under the control of a single DBMS. All users and client applications connect directly to the DBMS.
centralised processing	A single computer performing all processing for one or more users.
certainty factor	A value, usually in the range of 0 to 1, which describes the level of certainty in a fact or conclusion.
CHS	Cylinder, head, sector. A system for addressing each block on a hard disk.
client-server architecture	Servers provide specific processing services for clients. Clients request a service and wait for a response while the server processes the request.
CMOS	Complementary metal oxide semiconductor
CMTS	Cable modem termination system. The device that connects a number of cable modems to an ISP.
CMYK	Cyan, magenta, yellow and key. Key refers to black ink. CMYK is a system for representing colour on paper, also known as four colour process.
collecting	The information process that gathers data from the environment. It includes knowing what data is required, from where it will come and how it will be gathered.
communication management plan	A project management tool that specifies how communication between all parties involved in a system's development should take place.
confidence variable	An attribute whose value is determined mathematically by combining its assigned values. A measure of the confidence in a respnse or conclusion within an expert system.
context diagram	A systems modelling technique describing the data entering and leaving a system together with its source and sink.
copyright	The sole legal right to produce or reproduce a literary, dramatic, musical or artistic work, now extended to include software.
Copyright Act 1968	A legal document used to protect the legal rights of authors of original works.
CPU	Central Processing Unit

CRT	Cathode ray tube.
DAC	Digital to Analog Converter
data dictionary	A table identifying and describing the nature of each data item. Data dictionaries are used in many areas of system design, including the design of databases.
data flow	A labelled arrow on context and data flow diagrams describing the nature and direction of data movement.
data flow diagram	A diagram that shows the logical flow of data through a system or subsystem.
data independence	The separation of data and its management from the software applications that process the data.
data integrity	A measure of how correct and accurately data reflects its source.
data mart	Reorganised summary of specific data extracted from a larger database. Data marts are designed to meet the needs of an individual system or department in an organisation.
data mining	The process of discovering non-obvious patterns within large collections of data.
data store	Where data is maintained prior to or after it has been processed. Data stores are represented as open rectangles on data flow diagrams.
data validation	A check, at the time of data collection, to ensure the data is reasonable and meets certain criteria.
data verification	A check to ensure the data collected and stored matches and continues to match the source of the data.
data warehouse	A large separate combined copy of different databases used by an organisation. It includes historical data, which is used to analyse the activities of the organisation.
database schemas	A technique for modelling the relationships within a relational database. Also known as Entity Relationship Diagrams (ERDs).
DBMS	Database management system.
DDBMS	Distibuted Database Management System.
decision	A decision between two or more alternatives. Committing to one alternative over other alternatives.
decision table	A tool for documenting the logic upon which decisions are made. They represent the rules, conditions and actions as a two-dimensional table.
decision tree	A tool for documenting the logic upon which decisions are made. They represent the rules, conditions and actions as a diagram.
decryption	The process of decoding encrpted data using a key.
demodulation	The process of decoding a modulated analog wave back into its original digital signal. The opposite of modulation.
device driver	A program that provides the interface between the operating system and a peripheral device.
DFD	See Data flow diagram.
dial-up modem	A modem used to transfer data over a traditional voice telephone line.
diary	A project management tool for recording the day-to-day progress and detail of completed tasks. Diaries tend to be used to record future appointments and factual information.
digital	Discrete. Digital data is coded and represented as distinct numbers. Computers use binary digital data.
direct conversion	Completely replacing an old system with a complete new system at a particular point in time. Also called direct-cutover.
display adapter	Synonym for video card.

displaying	The information process that outputs information from an information system.
distributed processing	Multiple CPUs used to perform processing tasks, often over a network and transparent to the user.
DMD	Digital micromirror device. Used within DLP projectors.
DMT	Discrete multitone. A modulation standard used by ADSL to dynamically assign frequencies.
DNS	Domain name server. A server that determines the IP address associated with a domain name.
DOCSIS	Data over cable service interface specifications. The standards specifying communication over a cable network.
dot pitch	The width of each pixel in mm. Commonly used to describe the resolution of screens.
downloading distributed database	A type of distributed database whereby each server download copies of data as it is required from remote databases and stores the data within its local database.
dpi	Dots per inch. A measure of screen or printer output resolution.
draw software applications	A software application for manipulating vector images.
DSL	Digital subsriber line.
DSLAM	DSL access multiplexor. A device at the telephone exchange that combines multiple signals from ADSL customers onto a single line to ISPs, and extracts individual customer signals from a single line.
DSP	Digital signal processor
DVI	Digital video interface. Used to connect digital monitors to video cards.
EFM	Eight to fourteen modulation. A system that converts each byte into 14 fourteen bits such that all bit patterns include at least two but less than 10 consecutive zeros.
email	Electronic mail.
embedding	Importing a source file into a destinationr file. The source file becomes part of the destination file.
encryption	The process of making data unreadable by those who do not possess the decryption code.
environment	The circumstances and conditions that surround an information system. Everything that influences or is influenced by the system.
ERD	Entity Relationship Diagram. See database schemas.
ergonomics	The study of the relationship between human workers and their work environment.
ethical	Dealing with morals or the principles of morality. The rules and standards for right conduct or practice.
evaluation	The process of examining a system to determine the extent to which it is meeting its requirements.
external entity	A source or sink for data entering or leaving the system. External entities are not part of the system.
fault tolerance	The ability of a system to continue operating despite the failure of one or more of its components.
feasibilty study	A study that analyses possible solutions and recommends sutiable solutions. Used to determine if the development should commence (or not).
feasible	Capable of being achieved using the available resources and meeting the identified requirements.
fibre optic link	A transmission medium that uses light to represent digital data.

file	A block of data comprised of a related set of data items that may be written to a storage device. May be made up of records, files, words, bytes, characters or bits.
file server	A computer (including software and hardware) dedicated to the function of storing and retrieving files on a network.
first normal form	The organisation of the database after the first stage of the normalisation process is complete. Also known as 1NF.
flash memory	Electronic solid-state non-volatile memory.
flat-file database	A single table of data stored as a single file. All rows (records) are composed of the same fields (attributes).
floating-point	A binary system for representing real numbers. Floating point does not represent all numbers exactly.
flow control	A system that controls when data can be transmitted and when it can be received.
font	A specific example of a particular typeface. For example Time New Roman Italic 12 point.
foreign keys	Fields that contain data that must match data from the primary key of another table.
fragmentation distributed database	A type of distributed database that utilises both vertical and horizontal fragmentation whereby individual data items are physically stored once only at one single location.
FTP	File transfer protocol. A set of rules for transfering files across a network.
full duplex	Communication in both directions at the same time.
funding management plan	A project management tool for ensuring a project is developed within the allocated budget.
Gantt chart	A project management tool for scheduling and assigning tasks.
GB	Gigabyte.
Gb	Gigabit
GDSS	Group Decision Support System.
GIF	Graphics interchange format.
GIS	Geographic Information System.
GLV	Grating light valve. Used within digital projectors.
group information system	An information system with a number of participants who work together to achieve the system's purpose.
hacker	People who aim to overcome the security mechanisms used by computer systems.
half duplex	Communcation in either direction but not at the same time.
handshaking	The process of negiotiating and establishing the rules of communication between two or more devices.
hard copy	A copy of text or image based information produced on paper.
hard disk	A random access magnetic secondary storage device. A type of disk in which the platters are made from metal and the mechanism is sealed inside a container.
hardware	The physical units that make up a computer or any device working with the computer.
helical	A type of magnetic tape system where multiple tracks are written at an angle to each other. Helical technology is also used within VCRs.
heuristic	A rule of thumb considered true, usually with an attached probability or level of certainty.
HID	Human interface device. A standard that forms part of the USB standard. HID

	drivers are included as part of most operating system.
hot swap	The ability to connect and disconnect devices whilst the system is operating.
HSL	Hue, saturation and luminance. A system for representing colour.
HTML	Hypertext markup language.
HTTP	Hypertext transfer protocol.
hub	A device for connecting nodes on a LAN. Messages are repeated to all attached nodes.
huffman compression	An example of lossless compression. Huffman compression looks for the most commonly occurring bit patterns within the data and replaces these with shorter symbols.
hypermedia	An extension of hypertext to include non-sequential links with other media types such as image, audio and video.
hypertext	Bodies of text that are linked in a non-sequential manner. Each block of text contains links to other blocks of text.
I/O	Input/Output.
IDE	Integrated drive electronics. An interface used to transfer data between the system bus and secondary storage devices. A term used to describe storage devices that contain their own controller, rather than it being on the motherboard.
IMAP	Internet message access protocol. A protocol used to download email messages from an email server to an email client.
inference engine	The part of an expert system that contains the logic processing functions. Used to draw conclusions from stated facts and relevant rules.
information	The meaning that a human assigns to data. Knowledge is acquired when information is received.
information processes	What needs to be done to transform the data into useful information. These actions coordinate and direct the system's resources to achieve the system's purpose.
information technology	The hardware and software used by an information system to carry out its information processes.
integers	Whole numbers. Includes negative and positive whole numbers and zero.
IP	Internet Protocol
ISP	Internet service provider. A connection point to the Internet. An ISP provides connection to the Interent for many customers.
IX	Internet exchange. Another name for a NAP.
journal	A project management tool for recording the day-to-day progress and detail of completed tasks. Journals often include detailed analysis and reflection on recent events.
Kb	Kilobit.
KB	Kilobyte.
knowledge engineer	A person who translates the knowledge of an expert into rules within a knowledge base.
LAN	Local area network. A network connecting devices over small physical distances and using the same rules of communciation.
laser	Light amplification by stimulated emission of radiation.
LBA	Logical block addressing. An addressing system where each block of data on a hard disk is assigned a sequential number.
LCD	Liquid crystal display.
LCOS	Liquid crystal on silicon.

LED	Light emitting diode.
linking	Establishing a connection between a source and destination file. Alterations to the source file will be reflected in the destination file.
liquid crystal	A substance in a state between a liquid and a solid.
live data	Real data that is processed by the operational system. Live testing using live data takes place once the system has been installed to ensure it is operating as expected.
logical topology	How data is transmitted and received between devices on a network regardless of their physical connections.
MAC address	Media Access Controller address that is hardwired into each device. A hardware address that uniquely identifies each node on the network.
macro	A short user defined command that executes a series of predefined commands.
mail-merge	A process where information from a database or other list is inserted into a standard document to produce multiple personalised copies.
MB	Megabyte.
Mb	Megabit.
MEM device	Micro-electromechanical device.
META tag	An HTML tag that is used to store information that describes the data within a web page. Intended for use by search engines.
metadata	Data that defines or describes other data.
MICR	Magnetic Ink Character Recognition.
microwave	High frequency electromagnetic waves that travle in straight lines.
MIDI	Musical Instrument Digital Interface
MIME	Multipurpose Internet Mail Extensions
mirroring	A process performed by various RAID implementations where the same data is simultaneously stored on multiple hard drives. Mirroring improves read access times but not write times.
MIS	Management Information System.
mixing software application	A software application used to manipulate and combine sampled audio data.
model	A representation of something. Computer models are mathematical representations of systems and objects.
modem	Shortened form of the terms modulation and demodulation. A device whose primary function is to modulate and demodulate signals.
modulation	The process of encoding digital information onto an analog wave by changing its amplitude, frequency or phase.
MPEG	Moving Pictures Expert Group.
MR effect	Magneto-resistance effect. A soft magnetic material that conducts electricity well when in the presence of a magnetic field but is otherwise a poor conductor.
NAP	Network access point. A NAP connects many ISPs to high speed lines to other NAPs. Also called an Internet exchange (IX).
narrowband	A transmission medium that supports a single transmission channel. Compare with broadband.
NIC	Network interface card. The interface between a computer and a LAN.
normalisation	The process of modifying the design of a database to exclude redundant data. Progressively decomposing the design into a sequence of normal forms.
NOS	Network Operating System.

NPP	National privacy principle. There are 10 NPPs contained within the Privacy Act 1988.
NPV	Net present value. A measure of the predicted real cost benefits of an investment.
OCR	Optical character recognition.
OLAP	Online Analytical Processing.
OLTP	Online Transaction Processing.
operation manual	A manual that describes the procedures participants follow as they use the system.
organising	The information process that determines the format in which data will be arranged and represented in preparation for other information processes.
OSI model	Open systems interconnection model. A set of standards developed by the International Standards Organisation (ISO). The OSI model is a seven layer model of communication ranging from the application layer down to the physical layer.
outsourcing	The contracting of services to external companies specialising in particular tasks.
paint software application	A software application for manipulating bitmap images.
parallel conversion	A method of converting to a new system where both the old and new systems operate together for a period of time.
parallel port	A port that transfers bytes of data using 8 parallel wires.
parallel processing	A form of distributed processing where multiple CPUs operate simultaneously to execute a single program or application.
parallel transmission	Method of communication where bits are transferred side by side down multiple communication channels.
participant development	A development approach whereby the same people that will use and operate the system are also the developers of the system.
participants	People who carry out or initiate information processes within an information system. An integral part of the system during information processing.
password	A secret code used to confirm that a user is who they claim to be.
path-based animation	A line (path) is drawn for each character to follow. When played each character moves along their line infront of the background.
PDA	Personal digital assistant.
phased conversion	A gradual conversion from an old system to a new system.
physical topology	The physical layout of devices on a network and how the cables and wires connect these devices.
Piezo crystal	A crystal that expands and contracts as electrical current is increased and descreased.
pilot conversion	A method of conversion where the new system is installed for a small amount of users. The users learn, use and evaluate the new system and when it is deemed satisfactory, then the system is installed and used by everyone.
pixel	Picture element. The smallest element of a bitmap image.
polarizing panel	A panel that only allows light to enter at a particular angle.
POP	Post office protocol. A protocol used to download email messages from an email server to an email client.
PoP	Point of presence. The devices at an ISP that connect individual users to the Internet.
primary key	A field or combination of fields that uniquely identifies each record in a table.

privacy	An indivudual's right to feel safe from observation or intrusion into their personal lives. Consequently individual's have a right to know who holds their personal information and for what purpose it can be used.
Privacy Act 1988	The legal document specifying requirements in regard to the collection and use of personal and sensitive information in Australia.
procedure	A series of steps required to complete a process successfully.
processing	The information process that manipulates data by updating and editing it. Processing alters the actual data present in the system.
project management	A methodical, planned and ongoing process that guides all the development tasks and resources throughout a project's development.
protocol	A formal set of rules and procedures that must be observed for two devices to transfer data efficiently and successfully.
prototyping	A limited model of the system used to demonstrate the system to users/customers/particiapnts. Used to determine needs and requirements.
public key encryption	An encryption system where one key (the public key) is used to encrypt the data and a second key (the private key) is used to decrypt the data. Also known as asymmetrical encryption.
punched card	Cards used for both input and output during the 1950s and 1960s.
purpose	The aim or objective of the system and the reason the system exists. The purpose fulfils the needs of those for whom the system is created.
QAM	Quadrature amplitude modulation. A common modulation technique where the amplitude and phase of the wave are altered.
QBE	Query by example. A visual technique for specifying a database query.
RAID	Redundant Array of Independent Disks
RAM	Random access memory.
random access	Data can be stored and retrieved in any order.
raster scan	A technique for drawing or refreshing a screen row by row.
RDBMS	Relational Database Management System.
record	A collection of facts about an entity. A record comprises of one or more related data items. Also known as a tuple.
redundant data	Unnecessary duplicate data. Reducing or preferably eliminating data redundancy is the aim of normalisation.
reflective projector	A projector that reflects light off a smaller reflective image.
refresh rate	The number of times per second that a screen is redrawn.
relational database	A collection of two-dimensional tables joined by relationships.
relationships	How tables are linked together. A relationship creates a join between the primary key in one table and a foreign key in another.
replication distributed database	A type of distributed database whereby the aim is for all local databases to hold copies of all the data all of the time.
requirements	Features, properties or behaviours a system must have to achieve its purpose. Each requirement must be verifiable.
requirements prototype	A working model of an information system, built in order to understand the requirements of the system.
requirements report	The requirements for a system. A 'blue print' of what the system will do.

RFID	Radio Frequency Identification.
RGB	Red, green and blue. A system for representing the colour of light. Cpmpare with CMYK.
RLE	Run Length Encoding. An example of lossless compression. RLE looks for repeating patterns within binary data and replaces them with smaller symbols.
ROI	Return on investment. A measure of the percentage increase in an investment over time.
router	A device that directs messages to the intended receiver over the most efficient path. Routers communicate between many networks that may use different protocols.
RSI	Repetitive strain injury.
sampling (Audio)	The level, or instantaneous amplitude, of an audio signal recorded at precise intervals.
sans serif	Without serifs. Refers to a font that does not include serifs.
SAR	Successive approximation register. A component within an ADC that repeatedly produces digital numbers.
SATA	Serial advanced technology attachment. A serial version of the ATA standard.
satellite	A transponder in orbit above the earth.
schematic diagrams	See database schemas.
screen resolution	The number of horizontal pixels by the number of vertical pixels on a screen. Screen resolution can also be measured in dots per inch (dpi) or dot pitch (width of each pixel in mm).
SDLC	System development life cycle. Sometimes abbreviated to SDC.
search	To look through a collection of data in order to locate required data.
search engine	A program that builds an index of website content. Users can search the indexed content to locate relevant website content.
second normal form	The organisation of the database after the second stage of the normalisation process is complete. Also known as 2NF.
secondary storage	Non-volatile storage. Examples include hard disks, CD-ROMs, DVDs, tapes and floppy disks.
secret key encryption	An encryption system where a single key is used to both encrypt and decrypt data. Also known as symmetrical encryption.
sequential access	Data must be stored and retrieved in a linear manner.
sequential file	Files that can only be accessed from start to finish. Data within a sequential file is stored as a continuous stream.
serial transmission	Method of communication where a bits are transferred one after the other.
serif	Small strokes present on the extremities of characters in serif typefaces.
simplex	Communcation in a single direction only.
simulated data	Test data designed to test the performance of systems under simulated operational conditions.
simulation	The process of imitating the behaviour of a system or object. A specific application of a model.
sink	An external entity that is the recipient of output from an information system.
SMTP	Simple mail transfer protocol. A protocol used to send email from an email client to an SMTP server and also to transfer email between SMTP servers.

social	Friendly companionship. Living together in harmony rather than isolation.
software	The instructions that control the hardware and direct its operation.
sort	To arrange a collection of items in some specified order.
sound card	A device that converts digital audio to analog and viceversa.
source	An external entity that provides data (input) to an information system.
speech synthesis	The process of producing speech from text using a computer.
spot colour	A printing system that uses one or more inks of a predetermined colour. Compare with four colour process.
spreadsheet software application	A software application for manipulating numeric data. Spreadsheets combine input, processing and output within a single screen.
SQL	Structured query language.
SSL	Secure Sockets Layer.
SSML	Speech synthesis markup language.
start-stop communication	See asynchronous.
stepper motor	A motor that repeatedly turns a precise distance then stops for a precise period of time.
storyboard	An annotated sequence of drawings representing the screen designs and possible sequence of navigation in a proposed application, animation or motion picture.
streaming	The process of delivering data at a constant and continuous rate. Streaming is necessary when delivering audio and video data.
striping	A process performed by various RAID implementations where data is split into chunks and each chunk is simultaneously stored (and retrieved) across multiple hard drives. Striping improves data access times.
switch	An intelligent device for connecting nodes on a LAN. Messages are directed to the intended receiver.
synchronous	Communciation where data is received precisely in time with when it was sent.
system	Any organised assembly of resources and processes united and regulated by interaction or interdependence to accomplish a common purpose.
systems analyst	A person who analyses systems, determines requirements and designs new information systems.
systems flowchart.	A systems modelling technique describing the logic and flow of data, together with the general nature of the hardware tools.
TCP	Transmission Control Protocol.
TCP/IP	Transport control protocol internet protocol. A set of protocols used for communciation across networks, inclduing the Internet.
teleconference	A multi-location, multi-person conference where audio, video and/or other data is communicated in real time to all participants.
TFT	Thin film transistor.
third normal form	The organisation of the database after the third stage of the normalisation process is complete. Also known as 3NF.
TPM	Transaction Processing Monitor.
traditional approach	An approach to development that involves very structured, step-by-step stages. Each stage of the cycle must be completed before progressing to the next stage. Also known as the 'Structured Approach' or the 'Waterfall Approach'

transaction	A unit of work composed of multiple events that must all succeed or all fail. Events perform actions that create and/or modify data.
transmissive projector	A projector that directs light through a smaller transparent image.
transmitting and receiving	The information process that transfers data and information within and between information systems.
transponder	A device that receives and transmits microwaves. A contraction of the words transmitter and responder.
TTS	Text to speech.
tweeter	A speaker designed to reproduce high frequency sound waves.
UPS	Uninterruptible power supply.
URL	Universal resource locator used to identify individual files and resources on the Internet.
USB	Universal serial bus. A popular serial bus standard where up to 127 peripheral devices share a single communcation channel.
user interface	Part of a software application that displays information for the user. The user interface provides the means by which users interact with software.
users	People who use the information produced by an information system either directly (direct users) or indirectly (indirect users). An information system exists to provide information to its users.
vector image	A method of representing images using a mathematical description of each shape.
video card	An interface between the system bus and a screen. It contains its own processing and storage chips. Also called a display adapter.
view	The restricted portion of a database made available to a user or client application. Views select particular data but have no affect on the underlying organisation of the database.
virus	Software that deliberately produces some undesired or unwanted result.
VoIP	Voice over Internet Protocol.
volume data	Test data designed designed to ensure the system performs within its requirements when processes are subjected to large volumes of data.
VRAM	Video random access memory.
W3C	Wolrd wide web consortium.
WAN	Wide area network. A network connecting devices over large physical distances.
woofer	A speaker designed to reproduce low fequency sound waves.
WWW	World wide web.

INDEX

3G mobile networks 359
acceptance test 90
ACID properties 377-379
active listening 5
ADSL modem 342-343
agile methods 58-59
amplitude 552
analog data to analog signal 320-321
analog data to digital signal 324
analysing
  charts and graphs 492-493
  what-if scenarios 497
anchor tag 156-157
appropriate field data types 121-122
artificial neural networks 476, 527-533
audio 628-629
authentication 306
backup and recovery 170-171
  differential backup 415
  full backup 415
  incremental backup 415
  transaction logs, mirroring and rollback 416
backup media
  hard disks 418
  magnetic tape 417
  online systems 419
  optical media 418
backup procedures
  grandfather, father, son 420-421
  round robin 421
  towers of hanoi 421-422
backward chaining 514-516
bandwidth 248-249
barcode readers 426-427
baud rate 246
Bayer filter 621-622
bias 442
bitmap image 554, 626-627
bitmaps 626-627
blogs 359
Bluetooth 334
break-even point 49
bridge 340
broadband 248
cable modem 344
cartridge and tape 167
CCD 619
cel-based animation 558
centralised database 192
certainty factor 510
changing nature of work 96, 98
charts and graphs 492-493
checksums 251-253
client-server architecture 238, 305-306
coaxial cable 327-328
collecting and displaying
  consistency of design 204-205
  data validation 210-211
  grouping of information 205-206
  text 208-209
  white space, colour and graphics 207-208
collection
  forms 429-431
  online 431-433
collection hardware
  barcode readers 426-427
  magnetic stripe readers 427-428
  MICR 425-426
communication management plan 17
communication systems
  the IPT framework 229
communications control and addressing level 231-232
components of transaction processing
  data/information 372-373
  hardware 373-374
  software 374-375
  participants 371-372
compression and decompression
  huffman 549-550
  lossless 549, 555-556, 626, 628,
  lossy 553-555, 622, 626, 628
  RLE 549-550
confidence variable 509
conflict resolution 7-8
consistency of design 204-205
context diagram 65-66
copyright 638
Copyright Act 1968 638
copyright laws 18
cost-benefit analysis 48-49
CRT 565
current and emerging trends
  3G mobile networks 359
  blogs 359
  online radio, TV and VOD 359
  podcasts 359
  RSS feeds 359
  virtual world 641
  wikis 359
customisation 56
cyclic redundancy check 253-256
DAC 320
data 139
data cube 224
data dictionary 66-67
data flow diagram 68-69
data independence 162
data integrity 375, 443
data mart 468
data mining 469-470
data quality 443-444

data security 443
data validation 210-211, 376
data verification 376-377
data visualisation 472-473
data warehouse 435, 468
database management systems 162
database schemas 131
database servers 348
DBMS 175, 467
DDBMS 192-197
decision 449
decision support systems 538-542
  semi-structured 452-457
  unstructured 457-461
decision table 71-72
decision tree 71-72
decision tree algorithms 470
demodulation 342
design principles 204-210
DFD 68-69
diary 16
differential backup 415
digital camera 620-622
digital data to analog signal 323
digital data to digital signal 321-322
direct and sequential access 164
direct conversion 85
distributed database systems 193-197
DMD 572-573
downloading distributed database 195-196
drill downs 473-474
economic feasibilty 47-49
encoding and decoding
  analog data to analog signal 320-321
  analog data to digital signal 324
  digital data to analog signal 323
  digital data to digital signal 321-322
encryption and decryption 172-174
enterprise systems 439
environment 109
ERD 131
ergonomics 97
error checking methods
  checksums 251-253
  cyclic redundancy check 253-256
  parity bit check 249-250
Ethernet 243-244
evaluation 95
expert systems 466, 506-507
external entity 65
feasibilty study 46-47
file formats
  audio 628-629
  bitmaps 626-627
  vector image 628
  video and animation 630-631
file servers 346-347
first normal form 140
flash (SWF file format) 559-561
flat-bed scanner 618-620
flat-file database 119-120
  non-computer 125
  organising 120
floating-point 121-122
foreign keys 131
forms 429-431
forward chaining 516-517
fragmentation distributed database 193-195
full backup 415
funding management plan 16-17
Gantt chart 15
gateway 340-341
GDSS 475
GIF 558-559
GIS 477-478
GLV 573
goal seeking 498-500
grandfather, father, son 420-421
grouping of information 205-206
hard disks 166, 418
health and safety 18, 97
heuristic 511
HTML 154-156
HTTP 238-239
hub 339-340
huffman compression 549-550
hypermedia 150
hypertext 150
hypertext and hypermedia 151, 197
incremental backup 415
inference engine 513
infrared 335
intelligent agents 476
internet fraud 355-356
interpersonal 357-358
Interview 27-28
Inteview techniques 10-11
IP 241-243
IPT framework
  communications control and addressing level 231-232
  presentation level 231
  transmission level 232
issues related to
  data integrity 443
  data quality 443-444
  data security 443
  decision support systems 538-542
  internet fraud 355-356
  interpersonal 357-358
  power and control 356
  removal of physical boundaries 357
  work and employment 96-98, 358, 441-442
journal 16
K-nearest neighbour 471-472
knowledge base 508-519
knowledge engineer 508
LCD 566-568

LCOS 572
LED 619
live data 92
logical topologies
  logical bus 311-314
  logical ring 314-316
  logical star 316
lossless 549, 555-556, 626, 628,
lossy 553-555, 622, 626, 628
MAC address 232
macro 494
magnetic storage 165-166
magnetic stripe readers 427-428
magnetic tape 417
mail servers 348
maintaining 98-99
META tag 156
metadata 156
MICR 425-426
microwave 330
MIME 288
MIS 436, 479
mobile phones 335
modems
  ADSL modem 342-343
  cable modem 344
modulation 246-247, 342
network connection devices
  bridge 340
  gateway 340-341
  hub 339-340
  network interface card 339
  repeater 339
  router 345
  switch 340
  wireless access points 341
network interface card 339
NIC 339
non-linear regression 471
normalisation 139
NPP 218-219
NPV 48-49
OLAP 224, 472
  data cube 224
  data visualisation 472-473
  drill downs 473-474
OLTP 224, 475
online 88, 431-433
on-line and off-line storage 165
online radio, TV and VOD 359
online systems 419
operation manual 93-94
operational feasibility 50
optic fibre cable 328-329
optical media 418
optical storage 169, 578-580
organising 119, 120
  flat-file database 119-120
  hypertext and hypermedia 151
  non-computer 125
  relational databases 127-130
OSI model 231-232
outsourcing 54
parallel conversion 86
parity bit check 249-250
participant development 57
path-based animation 558
phased conversion 86
physical bus 307-308
physical measures 171-172
physical security measures 171-172
physical topologies
  physical bus 307-308
  physical hybrid 309-310
  physical mesh 310-311
  physical ring 309
  physical star 308
pilot conversion 86
plasma screens 569-570
podcasts 359
point-to-point terrestrial microwave 331
power and control 356
presentation level 231
primary key 129
print servers 347
Privacy Act 1988 218-219
privacy of the individual 18, 96
procedure 93
project management 3, 27-28
project triangle 3
protocols
  Ethernet 243-244
  HTTP 238-239
  IP 241-243
  MIME 288-289
  SMTP 284-287
  SSL 298
  TCP 239-241
  token ring 315
prototype 79
prototyping 55
proxy servers 348
purpose 107
QAM 322
query by example (QBE) 183
RAID 166-167
redundant data 139
referential integrity 377
relational database 127
  organising of 128-130
repeater 339
replication distributed database 196-197
requirements 27
requirements prototype 33-34
requirements report 26, 36-40
restricting access 174-175
RLE 549-550
round robin 421

router 236, 242, 345
RSS feeds 359
rule induction 470-471
satellite 330-333
schedule feasibility 49
screen design principles 204-210
SDLC 22-24
search databases 199-200
search engine
  operation of 198-199
  process user searches 200-201
  search databases 199-200
searching and retrieval
  query by example (QBE) 183
  tools for 179-189
  multiple tables 184-188
  single tables 179-183
second normal form 141
securing data 170
  backup and recovery 170-171
  encryption and decryption 172-174
  physical measures 171-172
  restricting access 174-175
security of data and information 18
semi-structured 452-457
servers
  database servers 348
  file servers 346-347
  mail servers 348
  print servers 347
  proxy servers 348
  web servers 348
simulated data 91
single tables 179-183
SMTP 284-288
social and ethical issues
  changing nature of work 96, 98
  copyright laws 18
  ergonomics 97
  health and safety 18, 97
  privacy of the individual 18, 96
  security of data and information 18
spreadsheets 466, 479-489
SSL 298
statistical analysis 501
storage and retrieval
  cartridge and tape 167
  database management systems 162
  direct and sequential access 164
  hard disks 166
  hypertext and hypermedia 197
  magnetic storage 165-166
  on-line and off-line storage 165
  optical storage 169
  RAID 166-167
  securing data 170
  tape libraries 168
storyboard 73-74, 151-153, 616-617
survey 27-28
switch 340
system design tools 65-72
system development
  introduction to 21-23
  approaches 53-54
systems analyst 26
tape libraries 168
TCP 239-241
team building 11-14
technical feasibility 47
testing the system 90-92
text 208-209
third normal form 144-145
token ring 315
touch screens 570-571
towers of hanoi 421-422
traditional approach 53-54
training
  implementing 87
  online 88
  peer 88
  traditional group 88
  traditional printed manuals 88
transaction logs, mirroring and rollback 416
transmission level 232
transponder 331
twisted pair 326-327
unstructured 457-461
URL 157-159
vector image 556, 628
video and animation 630-631
virtual world 641
VoIP 282-284
volume data 91
web servers 348
what-if analysis 497-498
what-if scenarios 497
white space, colour and graphics 207-208
wikis 359
wired transmission media
  coaxial cable 327-328
  optic fibre cable 328-329
  twisted pair 326-327
wireless access points 341
wireless LANS 333
wireless transmission media
  Bluetooth 334
  infrared 335
  mobile phones 335
  point-to-point terrestrial microwave 331
  satellite 331-333
  wireless LANS 333
work and employment 96-98, 358, 441-442