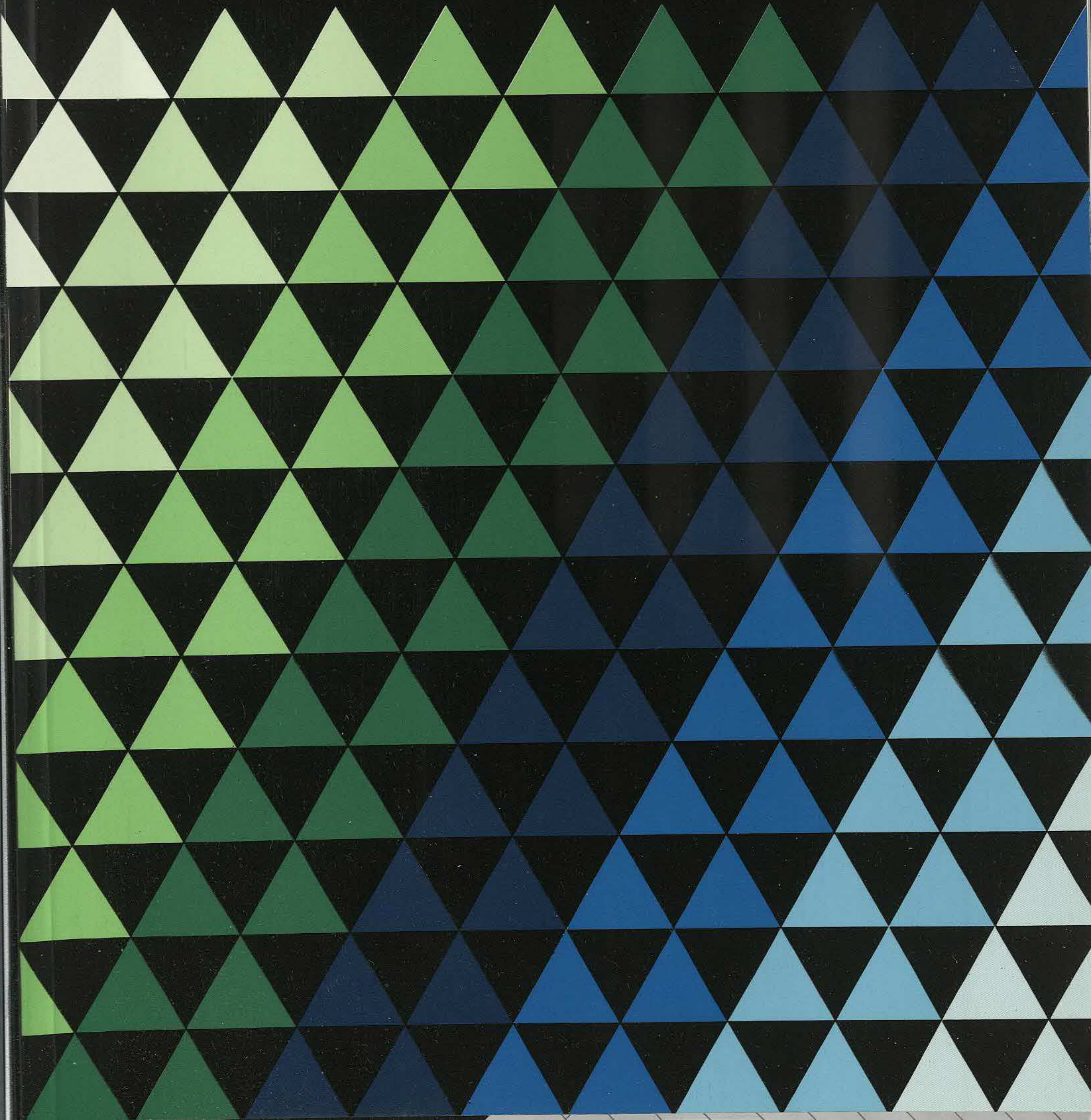
Emerging Technologies

Annual Review for Managers

BUTLER COX
FOUNDATION

Research Report 73, February 1990



BUTLER COX
FOUNDATION

Emerging Technologies

Annual Review for Managers

Research Report 73, February 1990

Butler Cox plc

LONDON
AMSTERDAM MUNICH PARIS

Published by Butler Cox plc
Butler Cox House
12 Bloomsbury Square
London WC1A 2LL
England

Availability of reports
Members of the Butler Cox Foundation receive three copies of each report upon publication; additional copies and copies of earlier reports may be purchased by members from Butler Cox.

Photoset and printed in Great Britain by Flexiprint Ltd., Lancing, Sussex.

BUTLER COX
FOUNDATION

Emerging Technologies

Annual Review for Managers

Research Report 73, February 1990

Contents

BUTLER COX FOUNDATION

Report synopsis

In this report, we identify six emerging technologies that are not yet in widespread use, but will be important to systems managers by 1995. We describe the nature of each technology, the application areas it can address, and the benefits it can provide. We also provide advice on how a systems department can identify opportunities for exploiting these technologies, and how it can design and implement applications that make use of them.

A Management Summary of this report has been published separately and distributed to all Foundation members. Additional copies of the Management Summary are available from Butler Cox.

Chapter 1

Six key technologies for the 1990s

Every systems manager has learned to live with the continuing stream of new products and services available from information technology (IT) suppliers. Every business has become accustomed to applying new types of IT to more and more of its operations and in ever more innovative ways. Minicomputers, high-level languages, and decision-support systems are just a few of the hardware and software developments that, within the last 30 years, have evolved from academic curiosities, through innovative products, to become accepted practice in the application of IT. During the 1980s, personal computers, fourth-generation languages, word processors, facsimile transceivers, expert system shells, local area networks, spreadsheets, and executive-information systems have evolved in similar ways, although not all of them have yet reached the stage of being universally accepted.

As we enter the 1990s, suppliers continue to claim that their latest products are based on significant technical innovations. Busy systems executives know from experience that many of these so-called innovations are likely to be little more than enhancements of existing technology, and their scepticism about suppliers' claims means that they may fail to spot the really significant advances in information technology. The purpose of this report is to identify the emerging technologies that will become important to systems managers within five years (by 1995). It is the first in a series of such reports, which we propose to publish regularly from 1990 on. The annual technology review will constitute one of the six Foundation Research Reports each year.

We have not set out to be exhaustive, but have concentrated, instead, on a few technologies that have particular potential. In choosing the technologies to investigate, we concentrated on hardware and software innovations that:

— Are not currently in commercial use to any great extent.

— Will be developed significantly within the next five years, resulting in products that will come into widespread use.

— Have the potential to make a substantial impact on the way in which Foundation members apply IT to their businesses.

— Are not restricted to a few, specialised, industry sectors.

We identified six technologies that met the criteria, and each forms the subject of one of the chapters of this report. They are parallel computers, groupware, hypermedia systems (sometimes known as multimedia systems), techniques and tools based on artificial-intelligence concepts, neural networks, and speech recognition. We believe that each of the six presents opportunities for the effective use of IT from which most members will be able to benefit within the next five years. The Appendix contains a brief commentary on the other technologies that we considered, but that were not studied in detail. They are data compression; optical-fibre local area networks based on the FDDI standard; image processing; local area wireless networks; metropolitan area networks; object-oriented design, programming, and data management; user interfaces; and very small aperture (satellite) terminals (VSATs).

We also excluded technologies that have been covered in recent Foundation Reports or that will be dealt with in reports in the near future. We do not therefore describe in any detail the developments in mobile communications, document-image processing, or system development tools.

This report, like future reviews of new and emerging information technologies, is different from many other Foundation Reports, because it deals with technologies where there is either very limited experience as yet, or where the experience has been gained in circumstances quite different from those of most Foundation members. This means that there is a lack of conclusive evidence to support the views expressed. The report is therefore based on our considered opinions about the way in which the selected technologies will develop, and the likely implications of those developments. These opinions are based on our own research, the opinions of experts in the various technologies, and our own experience, gained over many years, of forecasting and analysing developments in IT. The research team and the scope of the research carried out for this report are described in Figure 1.1.

We do not suggest that each Foundation member organisation should actively seek to exploit all of the six technologies. Many members will already have committed their IT research and development resources to the investigation of other technologies that are particularly important to them. Indeed, we were pleasantly surprised to find that about half of the 140 members who replied to the questionnaire sent out at the beginning of the research have an IT research and development unit. The average annual budget of these units is about $900,000. The most common new technologies being studied (or implemented) were expert systems, CASE, workstations, document-image processing, electronic data interchange, and system development tools. All of these have been the subject of recent or forthcoming Foundation Reports, and are already being used successfully by many Foundation members.

Figure 1.1 Scope of the research and research team

We began the research by identifying more than 30 areas of IT in which important innovations and developments seemed to be occurring. We sought views on each of these from noted authorities in the area (both from academe and industry), and took account of the published views of many more. We then applied our selection criteria to reduce this list to eight technologies. Each of these was investigated in more detail, with smart cards and very large knowledge bases being excluded from the report. We were unable to find the broad impacts of smart cards and we doubt that very large knowledge bases will have a major impact within the timescale covered by this report.

The research was led by David Flint, Butler Cox's research development manager. He was assisted by David Butler, Chairman of Butler Cox, Simon Forge, a senior consultant based in our Paris office, and Fergal Carton, Kevan Jones, Richard Pawson, and Martin Ray, all consultants based in London.

During the research, we sought the views of leading experts in each of the technology areas we studied. We are very grateful to these and the many other people outside Butler Cox who contributed their insights and experience to this report.

In this report, we concentrate on the potential of the six emerging technologies identified above, none of which is yet in widespread use. To help Foundation members evaluate them, we therefore describe the nature of the technology and the recent advances that have developed the technology to the stage where it is nearly ready for commercial exploitation. We then discuss the application areas that the technologies can address, and the benefits that they can provide. We also provide guidelines on how a systems department might tackle the practical issues of identifying opportunities for exploiting the technologies and for designing and implementing applications that make use of them.

Chapter 2

Parallel computers

Conventional mainframes and minicomputers have several processors and memory units, interlinked via backplanes, buses, and other high-speed links. In most systems, only one of these processors runs the application programs; the others are reserved for specialised functions such as the control of peripherals and networks. These computers are generally known as 'uniprocessors'. Parallel computers, on the other hand, are designed so that an application can run on more than one, sometimes on thousands, of processors. By operating in this way, they can provide much greater computing power than uniprocessors.

Parallel computers therefore include multiprocessor mainframes and supercomputers such as the large IBM 3090s and the Cray YMP, which typically have about four processors. These types of parallel computers have been available for several years and are extensions of well established technologies. The most interesting developments in parallel computers are occurring in other areas, however, where genuinely new types of computers, based on parallel-computing technology, are being developed and are beginning to provide significant business benefits. The world's most powerful computers are now parallel machines that have thousands of processors and provide processing rates of up to 60,000 mips (about 300 times the power of a large IBM mainframe). This means that they can run suitable applications much faster, and at much lower cost, than is possible on a uniprocessor. Early users of these new types of parallel computers have found that the hardware is up to 30 times cheaper than uniprocessors.

An example of the price/performance available from parallel computers is provided by a Transputer-based workstation available from Parsys in London. The minimum configuration costs about $75,000, but it provides processing power equivalent to 40 IBM mips or 80 Vax mips. (MIPS are not directly comparable between different computer architectures. For example, one IBM mainframe instruction is typically equivalent to two Vax instructions or four instructions on a RISC [reduced-instruction set computer] machine. One IBM mips is therefore approximately equivalent to two Vax mips. Throughout this chapter, mips relate to the specific computer architecture being discussed.)

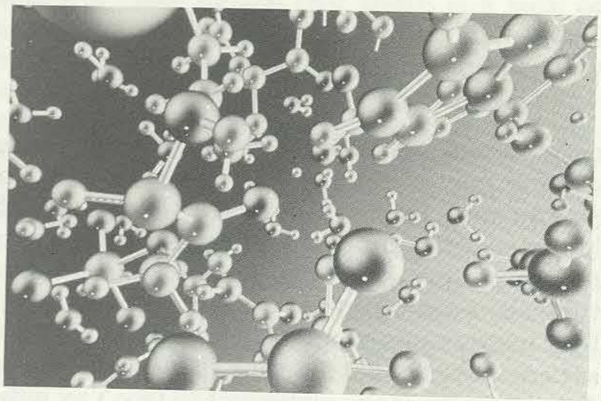
Another advantage of parallel computers is that they can often be enhanced simply by inserting a board containing additional processors. This means that processing power can be added in small increments as demand increases. Users of parallel computers have therefore found that they have been able to defer some of their planned investment in hardware until the extra capacity is required.

The power and cost advantage of parallel computers is particularly useful for new types of science and engineering applications, and for database management. Parallel computers will certainly be required — for example, for research on the decoding of the human genetic code — and they enable holograms to be produced from CAD designs, without the need to construct a model or to use photography. Figure 2.1, overleaf, shows a molecular model produced at Harvard University using Thinking Machines' Connection Machine. This is typical of the types of application that can be run on parallel computers today, because large amounts of processing power are required to handle the complex mathematics of molecular modelling.

In database management, the power of parallel computers can be used to provide flexible and cost-effective access to the largest databases,

Figure 2.1 Parallel computers can provide the processing power required for the complex mathematics of molecular modelling

The photograph is a computer image generated on a Connection Machine, and represents the oxygen, nitrogen, carbon, and hydrogen atoms of a complex enzyme.



(Source: Thinking Machines Corporation)

and to facilitate new approaches to information retrieval. Benchmark tests conducted by Sequent, which provides the Symmetry parallel computer, and Ingres (formerly Relational Technology), the providers of the Ingres relational database management system, have shown than that a four-processor Symmetry, costing about $500,000 and running Ingres, outperforms Digital, Hewlett-Packard, and IBM computers costing up to four times as much. Over the next few years, many organisations will experience huge increases in the volume of data that is stored online. Parallel computers will therefore increasingly be used to manage and provide access to this data.

Parallel computers can also be used instead of uniprocessors for a range of conventional applications, including communications switching, transaction processing, timesharing, and decision support. The use of parallel computers has, however, been held back by the shortage of application packages, a lack of compatibility with existing systems, the need for special skills, and uncertainty about the future viability of some of the suppliers. Each of these barriers is being eroded by current developments, and we expect parallel machines to play a substantial, although partly unrecognised, role in commercial computing by 1995.

Parallel machines will also stimulate the move to open, Unix-based, standards. Unix is already the commonest operating system for parallel machines. Parallel machines will provide the Unix market, which is already growing rapidly because of the open-standards policies of governments and some major users, with the advantages of even higher price/performance and access to kinds of applications that would be impractical with uniprocessor computers. The availability of open-architecture parallel computers will change the IT market in profound ways. In particular, it will make it possible to treat the acquisition of minicomputers and mainframes as commodity purchases, thereby weakening the historical dominance of suppliers of proprietary systems.

The nature of parallel-computer technology

The processor designs used in the most powerful uniprocessor computers are becoming more and more complex and are approaching the limits of current technology. These processors can be made faster only by making them more complex, which often makes them much more expensive. Over the last five years, the power of uniprocessor mainframe computers has, according to IBM, increased by 18 to 20 per cent per year. Such an increase is insufficient to keep up with the growth in demand for processing power, however. Many organisations' processing-power requirements are growing at 50 per cent a year. Parallel computers are the only long-term means available for meeting the increased demand.

In the course of our research, we identified more than 30 commercially available parallel computers, and more are announced every month. They vary greatly in size and design, and hence, in their advantages relative to uniprocessors and to each other, but they can be classified into two main types — shared-memory systems and distributed-memory systems. In shared-memory systems, all the processors have access to the same memory. In distributed-memory systems, each processor has its own memory, and communicates with other processors by sending messages.

Shared-memory systems

Shared-memory parallel computers span a wide range, from dual-processor minicomputers to eight-processor supercomputers (see Figure 2.2).

The largest supercomputers and mainframes, such as the Cray YMP and the IBM 3090, are parallel computers. The use of multiprocessor designs has enabled suppliers to increase the total processing power of their computers faster than they could increase the power of a single processor. The largest Cray system, for example, can now perform at an average of 250 Mflops (million floating point operations per second) and at up to 1,700 Mflops in short bursts for specific types of processing. However, because the memory is shared among all the processors, the power of these systems is limited by the rate at which data can be transferred to and from memory. Once this limit is reached, the addition of more processors produces no further increase in performance.

Programming this type of parallel computer requires an approach known as 'coarse-grained parallelism', which requires a program to be divided into relatively large chunks that are then allocated to processors. These chunks may be the same as programs run on a uniprocessor, which means that uniprocessors and shared-memory parallel computers may be compatible.

The processors in multiprocessor supercomputers and mainframes are extremely complex devices, implemented using state-of-the-art technology. One measure of this is the processor cycle time — 6 nanoseconds in the latest Cray, compared with 30 to 60 nanoseconds in a PC. The processors used in this type of parallel computer are therefore very expensive.

Other categories of shared-memory parallel computer use less powerful processors (these are sometimes known as 'processor farms'). Like distributed-memory systems, which are discussed below, this type of parallel computer can be implemented using the same technology as PCs, sometimes using the same processors. As Figure 2.3 shows, this type of technology results in computers that cost much less for each unit of processing power than large computers. Parallel computers can therefore provide the power of a supercomputer or mainframe, but have a price/performance ratio similar to that of a PC.

Figure 2.3 For each mips of processing power, small computers cost less than large ones

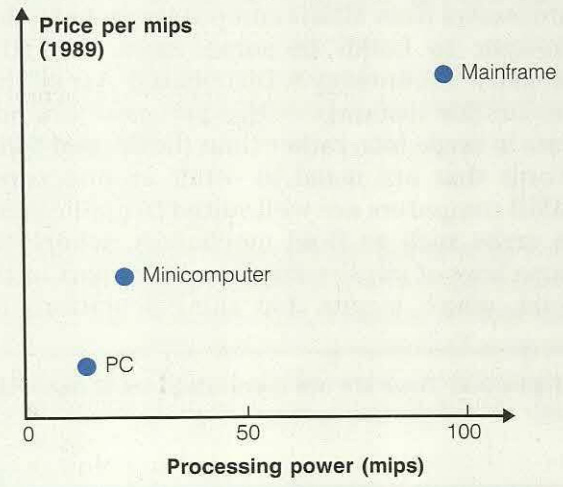


Figure 2.2 Shared-memory parallel computers span a wide range

	Examples		
Category	Product	Type of processor	Maximum number of processors
Supercomputers	Cray YMP	Proprietary	8
Minisupercomputers	Alliant FX/8 Convex C240	Proprietary Proprietary	8 4
Mainframes	IBM 3090	Proprietary	6
Superminicomputers	Pyramid Miserver Encore Multimax 520 Sequent Symmetry	Proprietary Intel 80386 Intel 80386 or 80486	12 20 30
Minicomputers	DEC 6000 Compaq SystemPro	Proprietary Intel 80386	6 2

Distributed-memory systems

Some of the most interesting parallel computers are distributed-memory systems in which each processor has its own dedicated memory. Figure 2.4 shows that there are two main categories of distributed-memory machine — single-instruction, multiple-data (SIMD), and multiple-instruction, multiple-data (MIMD) systems — and gives some examples of each category. In each case, as the terms suggest, each processor has its own dedicated data store, but in an SIMD machine, every processor runs the same program in synchronisation, whereas the processors in an MIMD computer may all run different programs.

Single- and multiple-instruction streams

SIMD parallel computers require simpler processors than MIMD computers and are thus cheaper to build. In some cases — Active Memory Technology's Distributed Array Processor, for instance — the processors handle data in single bits, rather than the 32- and 64-bit words that are usual in other architectures. SIMD computers are well suited to applications in areas such as fluid mechanics, where the same laws of physics apply to every part of the fluid, which means that the calculations for different parts can be allocated to different processors. (This is termed 'fine-grained parallelism'.) Since it is necessary to use all, or at least a large part, of an SIMD computer to solve a single problem, such computers are cost-effective only for large problems.

MIMD computers can be used in a much more flexible way, however. Although they are more expensive to build than SIMD computers, they may be quite small. A Cogent XTM system with 14 processors, for example, has a power of 140 mips but occupies a box only about twice the size of an IBM PC. They can also be easily expanded. The NCUBE 2, a large MIMD system, can be expanded from 64 processors (about 512 mips) to 8,192 processors (over 60,000 mips or 27 Mflops, about 15 times as fast as the largest Cray). MIMD computers are usually programmed by using a technique known as 'medium-grain data-flow programming'.

Communications topology

Processors in distributed-memory systems communicate by sending messages, either directly or via a message-switching system. Since some of these systems can have many thousands of processors, there are many ways in which they can be interconnected, giving rise

Figure 2.4 There are two main categories of distributed-memory parallel computer

Category	Nature of parallelism	Examples		
		Product	Type of processor	Maximum number of processors
SIMD (single instruction, multiple data)	All processors execute the same program	Connection Machine	Proprietary	65,536
		Active Memory Technology DAP 610	Proprietary	4,096
MIMD (multiple instruction, multiple data)	Each processor may run a different program	Intel iPSC	Intel 80386	128
		Teradata DBC		1,024
		Bolt Beranek & Newman TC2000	Motorola 88000	504
		Cogent XTM	Inmos Transputer	400
		Atari ATW		13
		Meiko Computing Surface		128
		Parsys		1,024
		NCUBE 2	Proprietary	8,192

to many different hardware architectures. At present, the most successful approach for large, distributed-memory systems seems to be an architecture called the hypercube, in which each processor is directly connected to others in a predefined pattern. As the size of the computer is increased by adding extra processors, additional connections are added to existing processors.

The advantage of this type of architecture is that the communications capacity increases as the size of the computer increases. This is a considerable advantage because the ultimate limit on the performance of a parallel computer is set by its communications capacity. Scientists at Sandia National Laboratories in New Mexico, which carries out research on nuclear power and weapons, have shown that a 1,024-processor NCUBE 2 machine (which is based on the hypercube architecture) can run 1,000 times faster than a uniprocessor computer. (Their paper reporting this received a Gordon Bell award in 1988 for its 'contribution to parallel processing for practical, full-scale, problems', and the Karp Prize for the first demonstration of a 200-fold or more increase in speed due to the use of a general-purpose parallel computer.)

Recent advances

During the last five years, parallel-computer technology has emerged from the laboratories of suppliers and of universities and is now incorporated in commercially available products that provide unprecedented power. There have also been important advances in data handling and software, and changes in users' perceptions of parallel computers.

Processing power

Laboratories and vendors are working to improve both the power of individual processors and the number that can be included in a single machine. The advances that are being made are still extremely rapid. In September 1989, for instance, SGS-Thomson in France bought the Inmos semiconductor company from the British company, Thorn-EMI. Inmos designed and produced the Transputer, which is still one of the few single-chip computers to have been specifically designed for incorporation in parallel computers. Inmos's new owners immediately announced both price cuts, which will make the cheapest (10 mips) model available in bulk for $20, and a future model with a processing power 'in excess of' 100 mips or 20 Mflops.

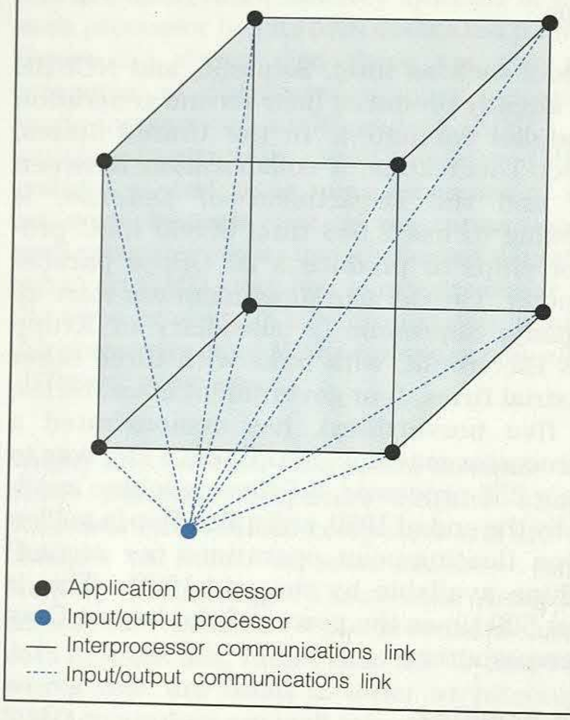
Vendors such as Intel, Sequent, and NCUBE have already produced their second generation of parallel computers. In the United States, Project Touchstone, a collaboration between Intel and the Department of Defense, is intending to use 2,048 Intel 80860 RISC processor chips to produce a 20 Gflops parallel computer. The German supercomputer start-up company, Suprenum (a subsidiary of Krupp Atlas Electronik, with links with three other industrial firms, four government laboratories, and five universities), has demonstrated a 32-processor machine. Suprenum's aim was to have a 256-processor, 5 Gflops machine available by the end of 1989, and a Teraflop (a million million floating-point operations per second) machine available by the mid-1990s. This is about 500 times the power of the largest Cray supercomputer.

Data handling

Initially, parallel computers did not provide very good data-transfer rates, an area critical to both commercial systems and to some scientific and engineering applications — geophysical analysis, for instance. More recently, some suppliers have applied parallel-processing principles to data handling:

— Intel and NCUBE have designed highly parallel input/output systems to complement their processing systems. The 127-processor, concurrent, input/output system for the Intel iPSC/2 has a throughput of 700M bytes per second. (By way of comparison, the largest possible IBM 3090 can have 128 channels running at 4.5M bytes per second, giving a total throughput of 576M bytes per second.) In the NCUBE 2, the input/output system contains one processor for every eight 'application' processors (see Figure 2.5, overleaf). Since each input/output processor can support several devices of varying kinds, the system throughput depends upon the actual configuration. However, the maximum data throughput of the largest possible NCUBE 2 system is 32 gigabytes per second.

Figure 2.5 The NCUBE 2 has one input/output processor for every eight 'application' processors


● Application processor
● Input/output processor
— Interprocessor communications link
----- Input/output communications link

- Thinking Machines Inc uses the technique of 'data spreading' in the DataVault of its Connection Machine. Instead of recording successive bits of a word along a disc track, each bit is written to a different disc drive. The DataVault is an assembly of 42 disc drives, 32 of which are used for data bits and seven for error-correction bits. (The remaining three drives are standbys.) It is therefore possible to remove any one of the drives without reducing the performance of the system. The DataVault can transfer data to the Connection Machine at rates up to 150 million bytes per second.
- IBM is also using the data-spreading technique in its Input/Output Access Method, a 'supercomputer system extension' for 3090 computers.

Software

Operating systems and other system software, compilers, and application packages are being developed for parallel computers. These developments mean that, in some circumstances, it is already possible to transfer applications from uniprocessors to parallel processors without having to rewrite them completely. In our view, it will be relatively simple, by about the end of 1992, to write applications that can be transferred from uniprocessors to parallel computers in a way that enables them to exploit most of the advantages of parallelism.

Operating systems

Most operating systems for parallel computers are based on Unix. Although Unix was designed for uniprocessor computers, the well defined interfaces between its components, and the relatively small size of its kernel (the part that runs in closest contact with the hardware), made it an obvious choice for parallel computing. Its popularity in the academic and engineering markets, and its (mostly) non-proprietary nature, also made it attractive to start-up companies with limited resources, which have done much of the pioneering work in developing parallel computers.

Work is underway at several academic and other laboratories to make Unix more suitable for parallel computers, and to provide a standard approach to parallelism. The most successful 'parallel Unix' so far is the Mach operating system, developed at Carnegie Mellon University, Pittsburgh, and this is emerging as a *de facto* standard. Early in 1989, AT&T, Olivetti, and Prime Computer announced a joint project to integrate the multiprocessor features of Mach into Unix System V. At the end of 1989, the Open Software Foundation, a software house developing system software for major vendors, decided to base its first version of Unix on Mach. (During a visit to the Open Software Foundation in May 1989, Butler Cox Foundation Study Tour delegates were told that the version of Unix to be used in the operating system would be based on IBM's AIX. The Open Software Foundation has, in fact, changed its mind and now favours Mach.)

Other system software

Another approach to parallelism, known as Linda, has been developed at Yale University. Linda is based on a concept called 'tuplespace', which can be thought of as a generalisation of the relational data model. Unlike relational tuples, however, Linda tuples may contain executable code as well as data. The programmer need not — in fact, cannot — consider the location of tuples, but can create, retrieve, and process them.

The Linda tuple operations have been incorporated into several programming languages, including C and Fortran. Linda has been implemented on a variety of parallel computers, including those made by Encore, Intel, Meiko, NCUBE, and Cogent Research. Scientists at Sandia National Laboratories, New Mexico, showed that 14 Vaxes, located across the United States and linked through Linda, were faster than a Cray for certain types of problem. Standards based on Linda are under development.

Another significant system software development for parallel computers is the Strand programming tool, developed by Artificial Intelligence Ltd in the United Kingdom. Strand provides a language, but more significantly, it allows modules written in other languages, such as C, Fortran, Cobol, and Pascal, to be integrated with it. Artificial Intelligence Ltd has implemented Strand on several parallel computers, including Sequent's Symmetry, Intel's iPSC/2, and Inmos's Transputer (which forms the basis of several parallel computers). Strand users include the Argonne National Laboratory in Illinois, Case Communications in the United Kingdom, Provincial Management Services in the United Kingdom, Ellemtel Utvecklings AB in Sweden, and the University of Tokyo.

Strand also runs on uniprocessors such as the Sony News workstation, Sun 3 and 4 workstations, and Intel 80386-based PCs. When staff at the Argonne National Laboratory used Strand and the C programming language to write a matching algorithm for DNA sequences, it ran 20 per cent faster than a previous version written in C alone. The main reason for this was that the programmers did not have to concern themselves with controlling the use of the parallel computing resources. Strand allowed them to concentrate on the parallel characteristics of the problem. When running on a 16-processor Encore Multimax, this application ran nine times faster than on a one-processor Multimax.

Compilers

Compilers are available for parallel computers for an increasing range of languages. The compilers available for the Encore Multimax, for example, include Ada, Basic, C, Cobol, Fortran, Lisp, and Pascal. The compilers available for the Sequent Symmetry include Ada, C, Fortran, and Pascal, and these languages have been extended by Sequent to support multitasking. Some compilers for parallel machines (the Fortran compiler, for example, available from Alliant, a US manufacturer of minisupercomputers) are able to recognise some of the parallelism implicit in existing sequential programs, and create a separate piece of code for each task that can be run concurrently.

Applications and development tools

Several organisations have transferred, or are in the process of transferring, technical applications from uniprocessors to parallel computers. Suppliers of software development tools have also begun to create versions of their products that can be used on parallel computers. Software products available (or under development) for the Sequent Symmetry include the Oracle, Ingres, Informix, and Unify database management systems, and development toolkits such as Accell/SQL and Focus. Software products available for the Encore Multimax include the Informix, Ingres, and Oracle database management systems, the WordPerfect word processing software, the 20/20 spreadsheet, and a Pick-compatible interface that enables users to run Pick, as well as Unix, applications. The existence of parallel-computer versions of compilers for commonly used languages, of well known database management systems, and of other system software means that it is relatively easy to transfer existing applications to a parallel computer.

Changes in users' perceptions

Until recently, most business computer users perceived the developments in parallel computers as being of little relevance to them. This perception changed during 1989, when IBM announced that its future top-of-the-line machines will come from parallel-computer developer, Supercomputer Systems Inc, the company founded by Stephen Chen and his design team (Chen was previously employed by Cray to design a massively parallel supercomputer). User organisations are now beginning to recognise that parallel computers can be a much more cost-effective solution than supercomputers, and the commercial fortunes of existing supercomputer companies are beginning to suffer as a result. One parallel-computer supplier told us, ''Many buyers have switched from a prejudice against parallelism to a prejudice in favour of it; we are content to be judged on our performance.''

Applications

Parallel computers are already being used successfully in a range of applications, examples of which are listed in Figure 2.6. We believe that, for the first half of the 1990s, the most important applications of parallel computers will continue to be in the science and engineering areas. However, parallel computers are already being used in commercial data processing as database machines, and they will increasingly be available as general-purpose Unix computers. They will also be used for new types of applications, such as the management of very large knowledge bases, the generation of holographic images, image processors, and realtime, complex decision-support systems.

Science and engineering

Parallel computers are already widely used for science and engineering applications, and there are many examples of applications processing being speeded up by very significant factors. A range of tools (Fortran compilers, for example) is becoming available for developing these types of applications, and some of the tools have 'automatic parallelisation' facilities, which means that the developer does not have to code the application explicitly to make the best use of the parallel-processing facilities. There is also an increasing variety of application packages, aimed at the science and engineering industries, available for parallel computers.

Figure 2.6 There are already examples of successful parallel-computing applications in several areas

The Sikorsky Aircraft Division of United Technologies Corporation has begun to use a 16,000-processor Connection Machine to design the blades of helicoptor rotors. This machine provides the power of a Cray supercomputer at 20 per cent of the cost.

The Bristow Helicopter Group, the United Kingdom's leading civilian helicopter operator, uses a Sequent Symmetry S81 with eight processors and 80 megabytes to run a complex maintenance-scheduling package.

IBM used a 212-processor machine called 'EVE' to simulate the chips used in the AS/400. This system resulted in a substantial reduction in development time because designs could be tested before the chips were available.

Royal Insurance in the United Kingdom installed a Teradata machine in July 1987 to provide decision-support facilities to managers as part of its new, devolved, management structure. The first operational use of the machine was to help the company to manage the follow-up of claims after the severe storm that swept the south of England in October 1987.

Reliance Mutual, a life assurance company, used packages based on the Ingres database management system to help with the introduction of new products in 1988. These packages run on an 8-processor, 32-megabyte, Sequent Symmetry S27 computer.

Database machines

The growth in the size of databases has, for some years, been threatening to outrun the capacity of even the largest mainframes. The management of large databases requires both large amounts of processing power and high data-transfer rates. It is also more efficient if the management of large databases is controlled by specially designed, rather than general-purpose, processors. To meet these needs, parallel computers are increasingly entering commercial use as database machines.

To date, the most successful parallel-processor-based database machine is the DBC/1012 from Teradata Inc. This machine can have from 6 to 1,024 processors, each managing a single disc drive. The database can be accessed by using a variety of facilities, including SQL, Focus, and the AI-based Intellect query language. Database queries are converted to SQL and distributed to all the processors that manage the relevant data. As of October 1989, 55 DBC/1012 systems had been installed, the largest of which was a 168-processor (170 mips) system used by Citibank for its 'relationship banking' system. (Delegates on the Foundation's 1988 Study Tour of the United States visited Teradata and heard at first hand how the DBC/1012 is being used in a range of commercial database applications.)

Vendors of database management systems are also working to modify their products so that they can be used on parallel computers. Oracle has been developing its 'parallel server architecture' since 1986, and in 1989, Ingres announced that it would make its relational database management system, with an optimised 'parallel database query' option, available on Sequent machines.

Parallel-computer-based database machines are also available for smaller databases and as local area network servers. During 1989, a start-up company, Netframe Systems, announced that its new multi-80386 servers would support the Oracle database server.

At present, database machines based on parallel computers have mainly been used to manage structured data stored in conventional databases. Access to unstructured data, such as text, images, and graphics, is becoming more important because of the growing use both of workstations with high-quality image and graphics capabilities and of document image processing systems. Future developments in groupware (which is discussed in Chapter 3) and hypermedia systems (the subject of Chapter 4) will mean that it is essential to be able to access unstructured data. The problems associated with accessing such data, which include slow response times and incomplete retrieval, are more severe than those associated with structured data. Parallel computers have the potential to solve these problems.

The management of structured data is now based on a proven theory — relational analysis. There is no equivalent theory for the management and retrieval of unstructured documents, which impedes the design of suitable systems. Relational databases are not suitable for this purpose, although they are sometimes used for lack of any other available tool. There are, however, promising developments based on the concept of 'objects' and other techniques derived from AI-research, and on hypermedia systems. One early approach, the use of the relevance-feedback technique on a Connection Machine (see Figure 2.7) in the Dow Jones DowQuest retrieval system, was demonstrated to Foundation Study Tour delegates in May 1989. Nevertheless, none of these developments has yet proved itself in large-scale general use.

Whatever technique is used for managing unstructured data, it is likely to require substantial processing power. It is our view that only parallel computers can provide the power required to solve the information-retrieval problem for unstructured documents at an affordable price.

General-purpose commercial computing

To date, the only parallel computers that have been used to any extent in commercial computing applications are MIMD database machines and shared-memory multiprocessors, such as the Sequent range. We believe that, during the first half of the 1990s, MIMD machines will increasingly be used to run commercial data processing applications as well as database management systems. With the exception of shared-memory multiprocessors, such as the IBM 3090 and Vax 8800, these machines will usually run Unix, or parallel operating systems based on it, such as Mach.

Shared-memory processor farms were the first parallel computers to be used for commercial applications, because they have the widest range of commercial software and require the least change in system-design and programming

Figure 2.7 Dow Jones uses a Connection Machine and the relevance-feedback technique to retrieve large quantities of unstructured information

The Dow Jones DowQuest retrieval system was demonstrated during the Foundation's 1989 Study Tour of the United States. In most information-retrieval systems, the user must specify certain words that the retrieval documents must or must not contain. He can usually also specify how close together these words must be. The system then searches predefined indexes to identify the documents. (This is known as a boolean search.) In specifying the document-selection criteria, the user must make a trade-off between failing to retrieve important documents and retrieving too many. This requires considerable skill, so most information-retrieval work is delegated to specialist librarians and information officers.

The technique of relevance feedback has been pioneered by Keith van Rijsbergen at the University of Glasgow and Gerard Salton at Cornell University. In relevance-feedback systems, the user must first identify one document that helps him, possibly using a boolean search. He then asks the retrieval system to rank all other documents stored in the system according to their similarity to the first document. This approach is easier for non-specialists to use, because they can usually easily recognise relevant and irrelevant documents. (In many respects, relevance feedback is analogous to using prototyping for clarifying user requirements.)

Relevance feedback requires very large amounts of processing power, however. Dow Jones uses a 32,768-processor Connection Machine to provide the DowQuest service and has a similar machine as a standby. Even so, the service includes only news items in the year prior to the date of searching.

techniques. They are also the most mature type of parallel computer. Indeed, some smaller organisations have already transferred all of their commercial data processing onto parallel computers. The experience of one early commercial user of parallel computers is described in Figure 2.8.

It will take longer for distributed-memory MIMD machines to come into widespread use. Initially, they will be used to provide a mixture of novel applications and timesharing services. They are unlikely to be widely used for conventional transaction-processing applications before the end of the 1990s. Because of their distinctive nature, distributed-memory MIMD machines will initially be kept separate from mainstream systems, in order to minimise the risks in the event that the vendor does not continue in business.

Distributed-memory SIMD machines will be used in a relatively small number of 'niche' commercial applications, such as the DowQuest system described in Figure 2.7 and some of the novel applications discussed below.

Novel applications

The general nature of some of the new types of application that will be made possible by parallel computers is already clear.

Large knowledge bases: As we explain in Chapter 5, AI-derived techniques will be used to manage the very large knowledge bases that organisations will increasingly find it necessary to construct. Like most techniques derived from AI, this will require huge amounts of processor power. Parallel computers will be the only way of providing the power at an affordable price.

Holograms: Parallel computers can be used to synthesise holographic images from design information, thereby producing realistic three-dimensional images of the object being designed. Such images will be more meaningful than engineering drawings to marketing personnel, for instance. SIMD machines are especially well-suited to this type of application.

Image processing: Parallel computers will also be used to provide the image-processing capabilities that will be needed for factory- and warehouse-automation systems and for security purposes. This type of application may well be implemented on special-purpose machines, such as those being developed to run neural networks. (Neural networks are the subject of Chapter 6.) These special-purpose machines will be embedded in robots and surveillance systems.

Realtime, complex decision-support systems: Parallel computers have the potential to transform decision-support systems, making them much more responsive to the needs of their users. They could, for example, be used to provide AI techniques that can help the user to identify relevant facilities and information, and to interpret the results. Database machines based on parallel computers can be used to

Figure 2.8 By using a parallel computer, a distribution company has been able to reduce delivery times considerably

This major company was recently formed by the merger of several independent companies. Although it is the 'brand leader' in its home market, it was not the market leader in any one major region. It therefore wanted a system that would give it a competitive advantage in the distribution business, although it also wanted to reduce costs. The company did not believe that its existing systems could provide the basis for doing this, so it investigated the opportunities for using new technologies, including fourth-generation languages, relational databases, and parallel computers. The Sequent S81 was selected as the basis for the new distribution system because it provided a price/performance about 30 times better than comparable computers and supported suitable development tools. The system was built using an advanced development tool and a relational database management system.

The distribution system is extremely sophisticated. Orders are collected during the day and are input to the system from handheld devices. The system then decides on the best route for the delivery vehicle, consistent with the delivery requirements of each customer. It allocates drivers and vehicles to the routes decided, according to specified criteria. Vehicle maintenance is included in the application. Use of the system will eventually enable the company to reduce the required number of drivers and trucks, although this will occur progressively as existing trucks are not replaced.

The system allows the company to achieve a very rapid delivery cycle. All orders received by 4.00 pm are delivered by 8.00 am the next day, if required. The system can generate up to 100 optimised routes in three hours. Previously, the various companies that merged to form this company achieved delivery cycles of between 24 and 48 hours.

retrieve relevant information from a variety of structured and unstructured sources, and can then try out many different combinations of the data in rapid succession.

Problems and barriers

Despite their considerable advantages, parallel computers are, at present, being used only to a limited extent in commercial data processing and office systems, and their use is far from universal in scientific and engineering computing. The main barriers inhibiting the wider use of parallel computers are their lack of compatibility with existing systems, a shortage of application packages, the need for special skills, and uncertainties about the long-term future of vendors and hardware architectures.

Lack of compatibility with existing systems

The most important barrier to the wider use of parallel computers is their incompatibility with existing uniprocessor systems. Most organisations have large investments in applications software, and will want to be able to run this software on any future systems that they purchase. Most parallel computers use the Unix operating system (or variants of it), and since Unix is not yet in widespread commercial use, these computers cannot generally be used to run existing system software, packages, or internally written applications.

Suppliers are addressing this problem by implementing widely used software products, such as Oracle and its development tools, on parallel computers. In the case of Oracle, this will provide a software 'virtual machine' that is compatible with existing computers that run Oracle. The priorities for creating parallel-computer versions of popular software packages are, of course, set by the current customer base for parallel computers, which is still largely in the scientific and engineering fields. Nevertheless, the software available for some parallel computers is already adequate to facilitate the rapid construction of business applications.

Applications written for existing Unix systems (and those written for other operating systems, but which make little or no use of features specific to the original operating system, transaction-processing monitor, or database management system) may often be transferred to parallel Unix computers without great difficulty. Although these applications will benefit from the price/performance of parallel computers and the ease with which such computers can be increased in size, they may not run faster. In order for an application to run faster than it would on a uniprocessor, it will often be necessary to make significant changes, which in some cases, will amount to a complete rewrite.

Shortage of application packages

Most of the application packages available today have been written for uniprocessors, since that is where the potential market for such products lies. Sometimes, an application package does not need to be changed to run on some parallel computers, particularly shared-memory systems. Thus, Cray multiprocessors can run the same programs as Cray uniprocessors, the Vax 6000 multiprocessor will run any Vax programs, and IBM applications will, of course, run on multiprocessor 3090 mainframes.

Many Unix packages can be run on parallel computers, although they may not make the best use of the facilities available. Some parallel computers also emulate the Pick operating system, which provides users with a further range of application packages to choose from. In addition, the suppliers of packages for scientific and engineering applications are beginning to create versions of their packages for parallel computers, mainly as a result of pressure from their customers. Progress in the commercial field is much slower, however, although we expect to see a gradual increase over the next few years in the range of application packages available for parallel computers.

Need for special skills

To run very large processing-intensive applications on distributed-memory machines, it is necessary to write them so that they fully exploit the benefits provided by parallel computers. This means that systems designers and programmers will need to use new approaches. Designers will need to consider how applications can be divided into tasks that can run concurrently, and programmers will need to learn the techniques for mapping these tasks onto the available processors and for managing and synchronising communications between the processors.

The design issues and the techniques vary according to the kind of parallel machine being used, and the particular model. Distributed-memory MIMD machines, for example, require the synchronisation process to be explicitly managed. However, parallel software environments such as Linda and Strand (both of which were discussed earlier in this chapter), relieve the programmer of some of this work and provide a method that is independent of particular machine architectures.

There is already evidence to suggest that the task of programming a parallel computer is not as difficult as it may at first appear to be. Craig Fields, Deputy Director for Research at DARPA, the US Defense Advanced Research Projects Agency, reported at the Foundation International Conference in October 1987 that DARPA's programmers have little difficulty writing computer programs for parallel computers. (The full transcript of his presentation was published as a Foundation Position Paper in March 1988.)

Uncertainty about the future of vendors and architectures

The market for parallel computers is at a very early stage of development. It is characterised by a wide variety of technologies and suppliers, competing for a market that is, at present, quite small. Inevitably, some of the specialist suppliers in the market today will go out of business. Already, some products have been withdrawn or are no longer supported, and it is highly likely that this will also happen to some of the products currently available. Prospective users of parallel computers are therefore understandably cautious about investing in systems that they may not be able to upgrade or enhance.

More important, they are unwilling to develop applications and skills that may be made obsolete by changes in technology or market conditions. Software-product suppliers share these reservations, and are not usually prepared to develop special versions of their products for systems that may never be sold in large numbers.

These concerns are being addressed in three main ways:

- 'Parallelising compilers' are being developed, so that programs written for uniprocessors can exploit the power of parallel computers.
- Parallel software environments, such as Linda and Strand, are becoming available. These provide a method of handling parallelism that is independent of particular machine architectures. Programs written for such environments will be able to run on a variety of parallel computers, and even on uniprocessors.
- The implementation of popular database management systems, languages, and development tools will enable programs to be transferred easily to other computers, and will reduce the need for development staff to have machine-specific skills.

Thus, the investment made in software for a parallel computer need not be lost if the supplier goes out of business. The developments outlined above will allow the applications to be transferred to another parallel computer.

Implications for systems departments

Parallel computers are no longer just a research curiosity. Today, shared-memory multiprocessors are a viable alternative for commercial data processing systems. Database machines based on parallel computers are a valid option for large-scale data management. Distributed-memory systems are viable for scientific and engineering applications, and will be so for commercial applications in the next few years. The use of all types of parallel computers in business computing is set to grow, and systems managers should now take four steps to position themselves to exploit this technology. They should evaluate database machines, look for processing-intensive applications, track the development of software for parallel machines, and consider an experimental project.

Evaluate database machines

Some pioneering users of parallel-architecture database machines have already gained considerable benefits, and the price/performance of these machines will improve rapidly as they benefit from further developments in parallel-computing technology. For many Foundation members, this type of application provides the best route into using parallel computers because it allows existing applications to be retained.

The use of database machines is not restricted to large data centres. Local data-handling requirements will grow very rapidly over the next five years, particularly with the introduction of electronic document management, hypermedia systems, and groupware. These requirements can be met by providing a database machine as a shared resource on a local area network, where its data-handling capability will complement the personal autonomy and usability of PC-based or Unix workstations. Developments in cooperative processing, client-server architectures, and standards for data and documents will make this an increasingly feasible approach. Servers based on parallel computers are already available from several vendors, although they do not yet provide all of the benefits they could, because of the immaturity of software environments for local area networks.

Look for processing-intensive applications

In addition to scientific and engineering applications, many organisations have some potential applications that would be impossibly expensive or slow if they were implemented with today's uniprocessor computers. These applications include the processing of complex models, the visual representation of designs, and the retrieval of information from databases and document archives. Parallel computers will bring the computer power required for such applications within the reach of smaller organisations, and will continue to increase the power available to the largest users.

Foundation members should therefore review the case for any such applications that have recently been rejected, in the light of parallel-computer capabilities and costs. They should also look at current applications of parallel computers in other organisations, to see whether they suggest possible applications in their own businesses.

Track the development of software for parallel machines

Although parallel-computer software is advancing rapidly, it is still not possible, in most cases, to transfer existing business applications to parallel computers without substantial rewriting. Even Unix applications cannot usually exploit the full power of parallel architectures unless they are substantially rewritten. However, rapid progress is now being made in transferring packages, tools, and system software to parallel computers, and the more intelligent compilers make it easier to transfer applications. Moreover, generalised parallel-programming environments, such as Strand and Linda, may soon become *de facto* standards.

User organisations should therefore keep track of the developments in software for parallel computers, and watch out for the emergence of standards. The emergence of standards, even *de facto* standards, will indicate that the time has come for non-pioneering organisations to consider installing parallel computers.

Consider an experimental project

To gain experience with parallel-computing technology, we suggest that systems departments carry out an experimental (or pilot) project. The chosen application should have a short life, and can be based on relatively low-cost parallel workstations, or add-on boards for PCs and workstations. Thereafter, and certainly until standards emerge, parallel computers should be used only for applications where:

- The benefits are very great, probably because the applications require substantial processing power and are amenable to parallelisation. These will usually be in the scientific and engineering fields.
- The computers can run relevant applications and development software. Thus, organisations that have already chosen Unix as a standard software environment could consider using shared-memory machines. In the foreseeable future, there will be sufficient software available for distributed-memory systems to allow them to become a real alternative as well.

In view of the specialised nature of parallel computers, any expertise should, for the time being, be developed within a small team. As the technology matures, it should be possible to transfer some of the expertise to mainstream development staff, while restricting the more technical aspects to systems programmers.

Implications for the IT market

Parallel computers will hasten the emergence of a more commodity-like market for IT products. Today, many mainframes and minicomputers can be used only with proprietary operating systems provided by their vendors, and programs written to work with one operating system cannot usually work with any other.

By contrast, a significant degree of operating system standardisation already exists in the PC, engineering-graphics, and supercomputer markets. The emergence of standards for MS-DOS and Unix systems has increased competition amongst the vendors of this type of hardware, with obvious benefits to users. This trend has developed furthest in the market for engineering workstations, where some users have rejected superior technology, such as Apollo's Domain distributed computing system, in favour of Unix and related standards. Suppliers such as Apple, Digital, IBM, and Hewlett-Packard, who have strong commitments to proprietary operating systems in other markets, have all been obliged to offer Unix on their engineering workstations.

As we described in Report 69, *Software Strategy*, open standards based on Unix are becoming increasingly important in the office automation and commercial minicomputer markets. A few major users have already adopted Unix as a software standard for this reason. It is also clear that open standards will become significant for mainframes. These issues will be considered in greater detail in a future Butler Cox Foundation Position Paper, *The Future of the Open Systems Market*.

In the context of these developments, parallel computers, which will mostly be Unix systems, are likely to transform the market for commercial computing products. In particular, they will have price/performance ratios far superior to today's products and will enable new types of applications to be developed. Parallel computers are already allowing their suppliers to win contracts on the basis of their superior price/performance.

Parallel computers will therefore be a driving force in commoditising the market for minicomputers and mainframes, and this will put increasing pressure on the suppliers of proprietary systems. The vendors of such systems will have difficulty competing because, in a commodity market, profit margins have to be reduced to a minimum. In order to compete, all vendors will have to reduce their support costs, either by eliminating support functions or by automating them, and will need to seek additional profits from software. The result of these changes will be increased instability in the IT-supply market. User organisations will be able to obtain remarkable bargains, but from suppliers who may not survive.

Chapter 3
Groupware

Groupware is the generic name for computer software that supports the ways in which groups of people work together. Although groups of people can collaborate by using a variety of IT facilities, including widely available ones such as facsimile, and less common ones such as videoconferencing, groupware is concerned with explicitly supporting the collaboration processes. Groupware is also sometimes known as 'computer support for collaborative work'.

The nature of groupware

Computer users' needs to communicate and collaborate with their colleagues are often met by providing an electronic mail system. Such systems are seldom integrated with the computer applications, and do not usually provide support for the collaboration processes. In order to support group collaboration fully, a system must include functions such as diary management, commitment management, and meeting mediation (which is equivalent to chairing an 'electronic' meeting). It must also support the data structures and message formats related to the collaboration.

In providing such support, groupware represents a natural extension of software for workstations interconnected via a local area network, a development of office automation ideas, and a significant extension to office systems. Most groupware systems can be perceived either as enhanced electronic mail systems or as enhanced conferencing systems (which are themselves closely related to electronic mail).

Groupware can be based on data networks, both local and wide-area, PABXs, and the public telephone network, and they can be provided as public services or as systems for small teams. Staff supported by a groupware system may be in the same room, or they may be spread over several locations, or even time zones. Groupware systems are provided on personal workstations that are usually, but not necessarily, intelligent. However, the most important groupware developments are exploiting the power of intelligent workstations that are interconnected by a local area network.

Groupware can provide many benefits to its users, the greatest of which are improved interpersonal communications, less ambiguity about the work of the group and the roles of its members, and the ability to involve more people in decision-making. These benefits have produced significant improvements in efficiency and effectiveness for early groupware users.

Groupware is not a new idea. The concepts were formulated at least 30 years ago by office systems pioneer, Douglas Englebart, and groupware systems were first built more than 20 years ago by Englebart's team at the Stanford Research Institute. Early groupware systems supported relatively small groups of users, but did not spread beyond these small communities. They failed for two main reasons: they often had complex and cryptic user interfaces, and their developers perceived them as research tools rather than as commercial products. Furthermore, the claims made for groupware systems were threatening to managers who were used to conducting their interpersonal communications without any technology except the telephone. This, too, has slowed down the introduction of groupware systems.

Current trends

Despite the continuing research and the high attendance at academic conferences on the topic, there has been no major breakthrough in the understanding of collaborative work. Nor have there been any notable advances in the ability to develop systems that support collaborative work, although established software suppliers and entrepreneurs are paying close attention to the field, and some interesting products have recently appeared.

Some of the research being carried out in the area of groupware is intended to extend the technology, some is concerned with providing a more sophisticated model of the human collaboration process on which the systems are based, and some is concerned with creating environments for investigating human behaviour. The Object Lens at MIT, for example, is an attempt to extend the power of an electronic mail system to support the exchange of computer 'objects' in general, including data files and programs. The Chaos system at the University of Milan is using expert-system techniques to model the organisational context in which people make commitments to each other.

The main impetus for the growing interest in groupware is, however, the business pressure that is creating the demand for more computerised group-support facilities:

— Managerial and professional work is increasingly done in ad hoc teams, sometimes called task forces, which require more flexible administrative-support arrangements than those provided by established organisational structures.

— In many organisations, most of the routine clerical work has already been computerised. Much of the work that has not been computerised depends upon human communication between professionals and managers.

— Experience of using PCs is continuing to increase among managers and professionals. A significant proportion of senior and middle managers in many Foundation member organisations now have PCs on their desks.

— Experience of using electronic mail systems helps managers and professionals to see computers as a communications medium, not just as a means of processing information.

The growth in the number of PCs, and especially of PCs interconnected via local area networks, provides the technical basis for the provision of groupware. (Most of the more innovative groupware systems are being implemented on PCs and local area networks.) According to IDC, the US market research company, there are six million network-connected PCs in the world today, and this number is expected to rise to nearly 30 million by 1992. Figure 3.1 shows the infrastructure used by most groupware systems. This infrastructure is significant, because:

— The power of the PC, and especially the intuitive nature of the graphical interface now being increasingly used on PCs, enables developers to make groupware systems (and other systems) much easier to use.

— The increasing number of networked PCs reduces the cost of providing groupware facilities, since many of the required PCs will already be in place.

— The transmission rates available with local area networks facilitate rapid online communication with other users.

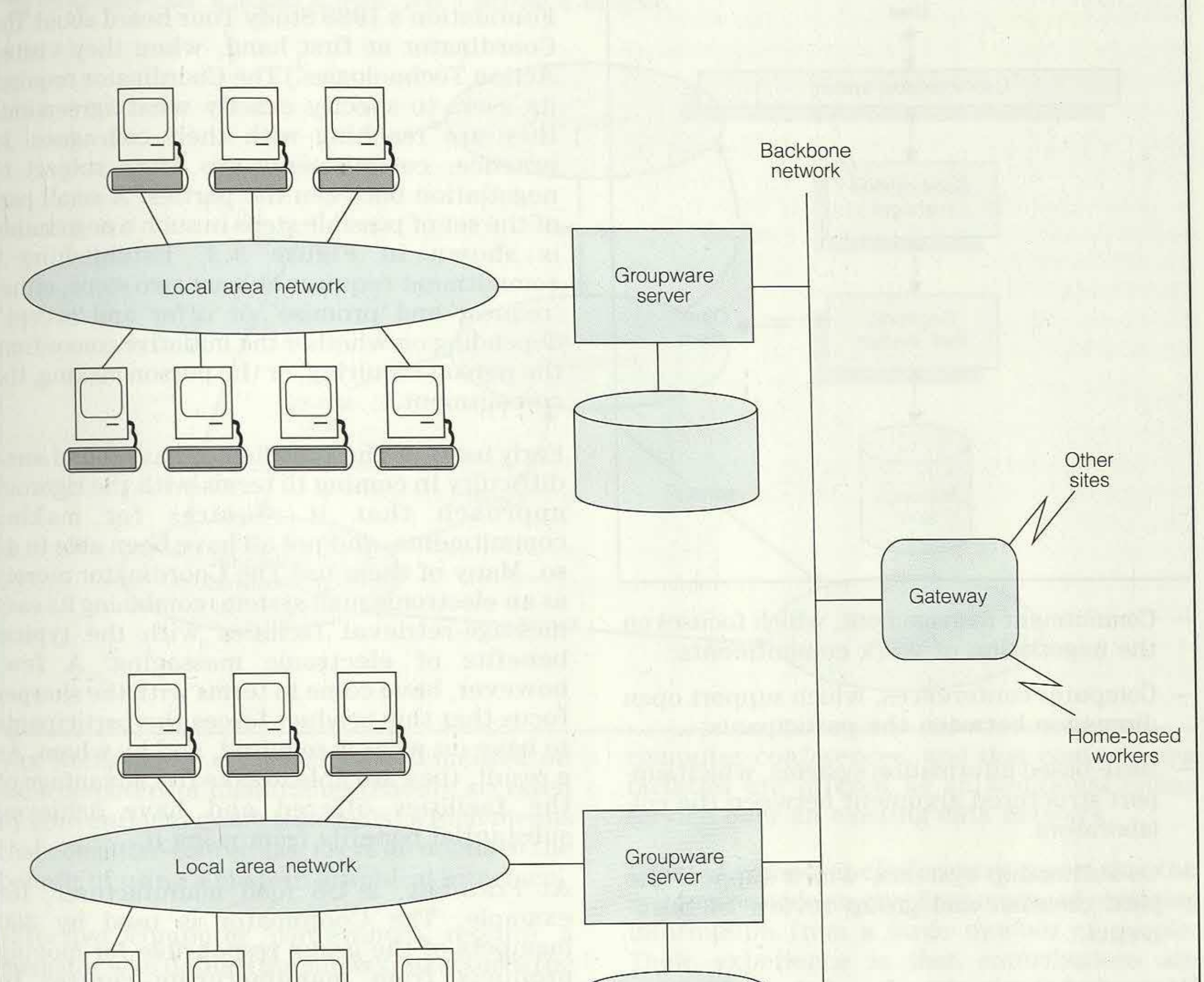
Groupware products have begun to appear in significant numbers only within the last two years, and most of these have sold only in small numbers. This has not, however, prevented a great deal of coverage in the media and at conferences, in an effort to identify those elements of groupware that will be most successful. The search is on for the groupware equivalent of the spreadsheet.

This search is almost certainly misconceived. Groupware, by definition, supports groups of people, and as the working needs of groups vary greatly, there are likely to be many kinds of groupware. There will be both extensions to existing products and new products aimed specifically at group support (such as The Coordinator, and Lotus's Notes product, expected in 1990, both of which are discussed later in the chapter). We believe that, by 1992, all major PC software products will provide groupware facilities. In the longer term, certainly by the end of the 1990s, groupware will have been completely absorbed into office systems.

Groupware systems

The principle of groupware has been applied to a variety of business activities, including

Figure 3.1 Most of the more innovative groupware systems are being implemented on PCs interconnected by local area networks

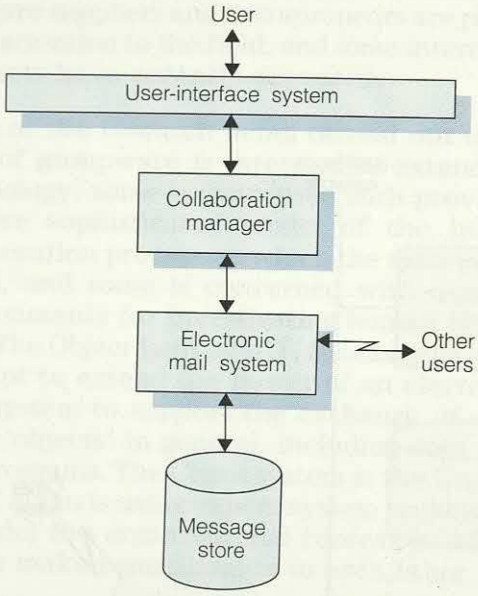


decision-making, document creation, and software development. At present, most groupware exists as discrete applications, each based on a particular perception of the nature of collaboration. We call these systems 'collaboration managers'. As interest in groupware increases, some developers have turned their attention to the production of software systems that can be used to support a variety of approaches to collaboration. We call these 'collaboration-support systems'.

Collaboration managers

Collaboration managers represent groupware in its purest, and least integrated form. Figure 3.2, overleaf, shows how a collaboration manager can be perceived as providing an intelligent front end to an electronic mail system. Different developers of collaboration managers have, however, based their systems on one of the four different models of the collaboration process.

Figure 3.2 Collaboration managers represent groupware in its least integrated form



— Commitment-management, which focuses on the negotiation of work commitments.

— Computer conferences, which support open discussion between the participants.

— Issue-based information systems, which support structured argument between the collaborators.

— Co-authorship systems, which support the joint creation and group review of documents.

Because of the immaturity of the field, most groupware systems currently support only one of these collaboration models, although all of them (and probably others as well) will be appropriate for supporting the interpersonal communications between the individuals in some groups. Future groupware systems may support several of these models, and they will increasingly become available as integrated aspects of more familiar data processing and office systems.

Commitment managers

Commitment managers record the commitments that people make to one another, track progress against these commitments, and remind them as due dates arrive. They may be linked to resource-planning and resource-monitoring systems.

The best known commitment manager is The Coordinator from Action Technologies of Emeryville, California. (Delegates on the Foundation's 1988 Study Tour heard about The Coordinator at first hand, when they visited Action Technologies.) The Coordinator requires its users to specify exactly what agreements they are reaching with their colleagues. In practice, commitments are often subject to negotiation between the parties. A small part of the set of possible steps in such a negotiation is shown in Figure 3.3. Establishing a commitment requires at least two steps, either 'request' and 'promise', or 'offer' and 'accept', depending on whether the initiative comes from the person requiring, or the person making, the commitment.

Early users of The Coordinator have found some difficulty in coming to terms with the rigorous approach that it requires for making commitments, and not all have been able to do so. Many of them use The Coordinator merely as an electronic mail system, combining its easy message-retrieval facilities with the typical benefits of electronic messaging. A few, however, have come to terms with the sharper focus that this product forces the participants to have on what is required, and by whom. As a result, they are able to take full advantage of the facilities offered and have achieved substantial benefits from using it.

At Frito-Lay, a US food manufacturer, for example, The Coordinator is used by 300 members of the group responsible for moving products from manufacturing centres to distribution and retail outlets. Use of The Coordinator avoids the misunderstandings that used to occur frequently among the group. Frito-Lay plans to extend its use to its manufacturing and sales operations.

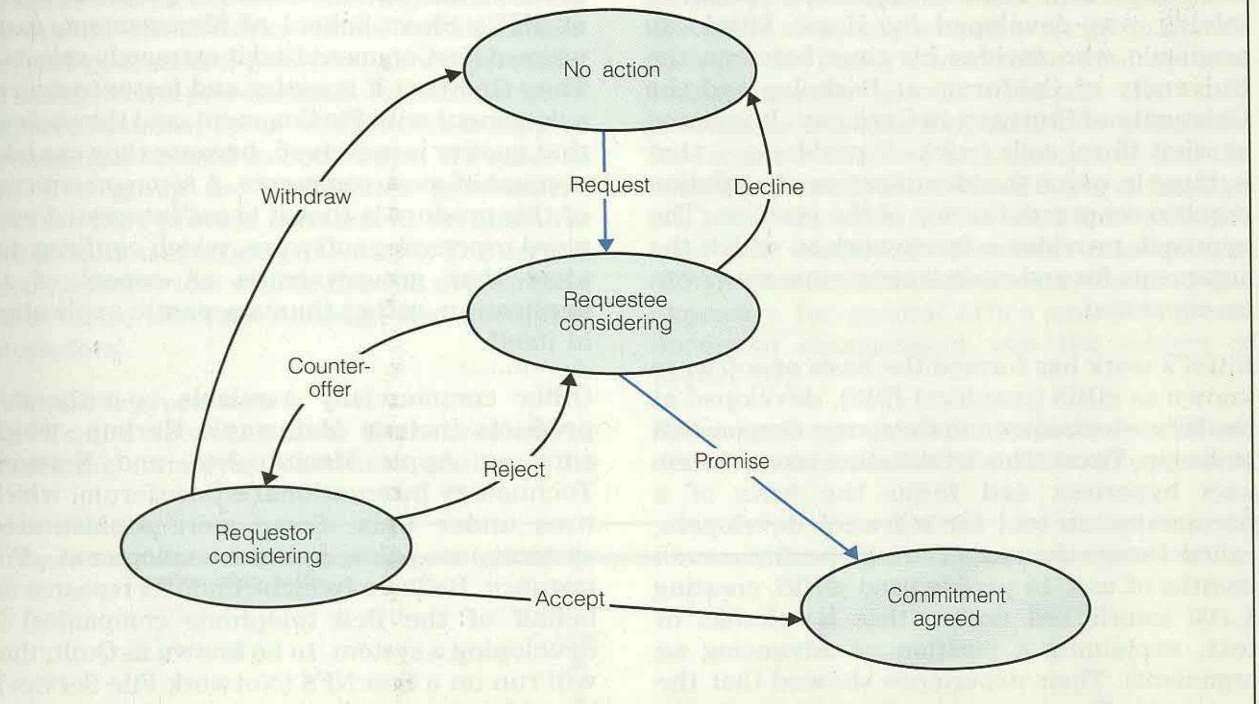
Computer conferencing

Computer conferencing has been used on a limited basis for at least the last 20 years, and is now a well proven approach to decision-making and to the distribution of information within organisations. Conferencing systems are, nevertheless, still immature as commercial products, and there is great scope for improving both the functions and the connectivity they provide.

Computer conferences are based on software such as Hewlett-Packard's Confer and Digital's

Figure 3.3 Negotiating a commitment requires at least two steps, but can be much more complicated

In the simplest case, a request is made, and a promise is extracted.



Vax Notes. They are usually implemented on minicomputers or mainframes that are accessed by conventional dumb terminals, which means that computer-conference users do not have the benefit of using a modern graphical interface.

Effective computer conferences require a mediator who defines the subject, structures the contributions, and directs the discussion. The minimum number of people for a viable conference varies, depending on organisational culture, the significance of the conference topic, and the availability of other sources of information. In some circumstances, a handful of committed participants may well be sufficient; in others, several hundred may not be enough.

To obtain the full benefits from computer conferences, however, requires those who might benefit from participation to be aware of the existence of the facilities. The best way of achieving this is to ensure that a reasonable number of conferences are running concurrently, and to provide simple ways of starting and joining conferences. This implies that, overall, there will be a few hundred users of computer conferences, and that conferencing facilities are offered as an additional online service over an existing data network.

Users of computer conferences report that the systems provide an excellent means of obtaining information from a large number of people. Their experience is that contributions are received from people who would not have participated in face-to-face meetings, and that the contributions are often more thoughtful than they would have been in a face-to-face meeting. This results in considerable savings in effort, a significant speeding-up of work, and often, better decisions or designs. Technical staff in major computer suppliers such as IBM, Hewlett-Packard, and Digital make extensive use of computer conferences, where they are now an integral part of normal business operations. Other organisations have obtained some of the benefits of computer conferencing by using electronic mail systems to create general-access mailboxes for particular topics.

Issue-based information systems

An issue-based information system (IBIS) is a specific type of decision-support system for a group of collaborators. For those involved in making a decision, an IBIS provides a way of expressing different views in a consistent manner and a means of resolving the different views expressed. The IBIS approach to problem-solving was developed by Horst Rittel, an academic who divides his time between the University of California at Berkeley and the University of Stuttgart in Germany. It is aimed at what Rittel calls 'wicked' problems — that is, those in which the identification of a solution requires some redefinition of the problem. The approach provides a framework in which the arguments for and against competing views can be expressed.

Rittel's work has formed the basis of software known as gIBIS (graphical IBIS), developed at the Microelectronics and Computer Corporation in Austin, Texas. This workstation-based system uses hypertext and forms the basis of a documentation tool for software developers, called Design Journal. During the first seven months of use, 16 people used gIBIS, creating 1,100 interlinked nodes (that is, chunks of text, explaining a position or advancing an argument). Their experience showed that the approach allows separate discussions of substantive and procedural matters to take place and helps to expose people who make their points by " hand waving, axe grinding, and clever rhetoric".

One advantage of Design Journal (and also of computer conferencing) is that it is possible to back-track to discover the reasons for making a particular decision and the range of factors that were considered.

Co-authorship systems

Co-authorship systems help people to cooperate on the preparation of documents. The systems may allow multiple versions of documents, or parts of documents, to exist concurrently, and help to organise the flow and resolution of comments on drafts. Anyone who has tried to reconcile comments from half a dozen colleagues will see the advantages.

One of the leading co-authorship products is ForComment from Brøderbund Software Inc, which runs on networked PCs. This product allows an author to circulate a draft, prepared using any of a range of word processors, to several colleagues for review. It then collates the comments according to the part of the document they relate to, making it easier for the author to assimilate and process the comments. According to recent research by Christine Bullen of the Center for Information Systems Research at MIT's Sloan School of Management, early users of ForComment find it extremely valuable. They find that it is easier and faster to deliver a document with ForComment, and they believe that quality is improved, because they can take account of more comments. A common criticism of this product is that it is not integrated with word processing software, which confirms our view that groupware is an aspect of an application, rather than a separate application in itself.

Other commercially available co-authorship products include Mainstay's Markup, which runs on Apple Macintoshes, and Network Technology International's Docuforum, which runs under Unix. Some more sophisticated systems are also under development. For instance, Bellcore (which conducts research on behalf of the Bell telephone companies) is developing a system, to be known as Quilt, that will run on a Sun NFS (Network File Service)-based local area network, and will allow work from several people to be merged into a single document. Office Workstations Ltd, the developer of Guide, an early personal hypertext system, is developing Idex, which allows a group of authors jointly to develop a large hypertext documentation system.

Co-authorship systems can provide some quite sophisticated facilities. Idex, for instance, has been used to support document management at a nuclear power station. Because of the importance of safety to the nuclear industry, and the pressure of regulation, it is vital that technical documentation, such as operating procedures, is kept up to date. Idex not only provides facilities for authors to work collaboratively on documents, but its hypertext structure makes it easy for others to consult the documentation online.

Collaboration-support systems

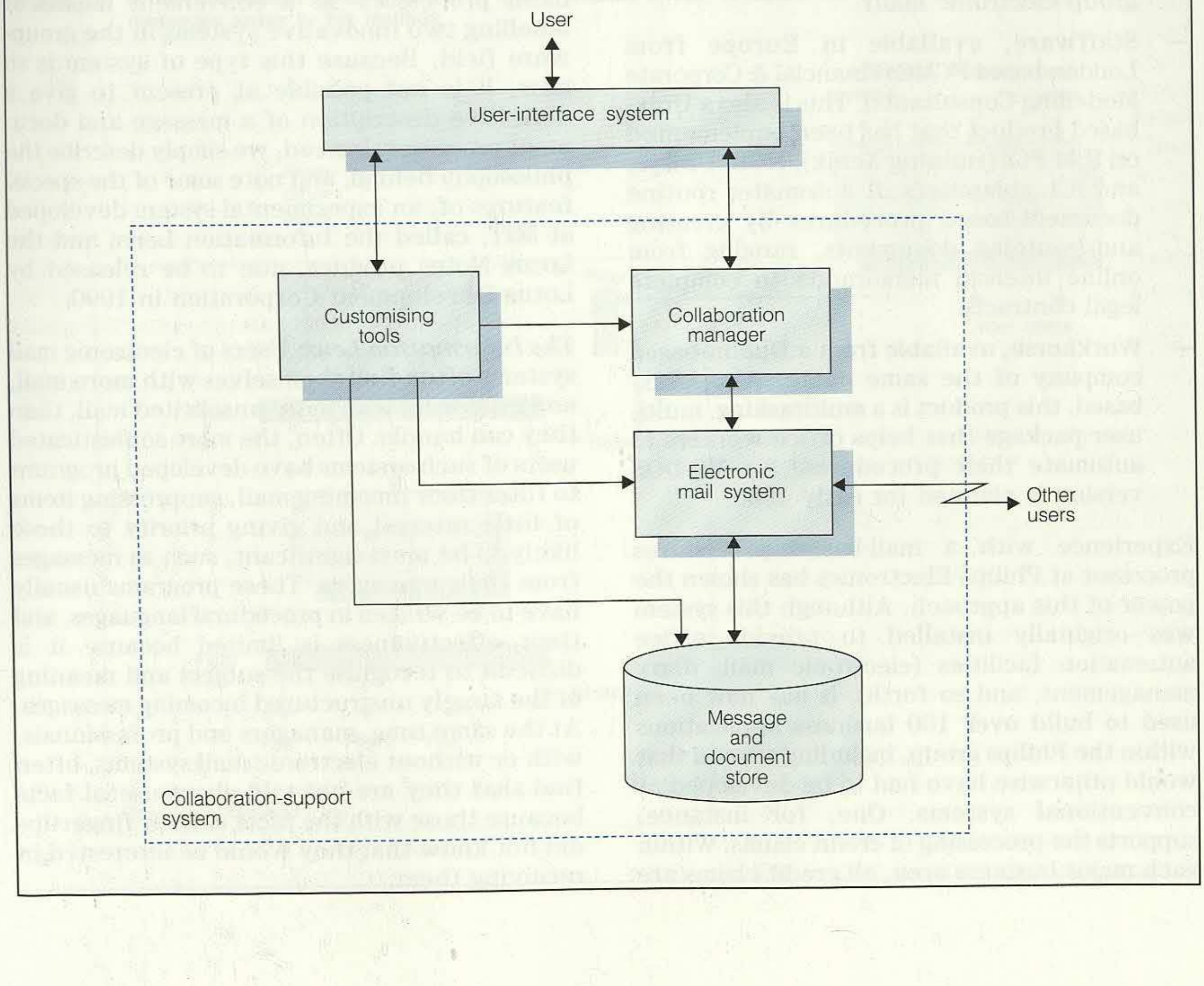
Collaboration-support systems provide the same kind of flexible support for group collaboration as decision-support systems provide for decision-making. They do this by providing facilities for dealing with messages, documents, and procedures, although different systems provide different degrees of support for each. As Figure 3.4 shows, collaboration-support systems usually provide skilled system developers with facilities for tailoring them for use by others, often clerks. Some collaboration-support systems are, however, aimed primarily at managers and professionals. As with decision-support systems, some collaboration-support systems can be used in both ways. We refer to the first group as 'procedures processors', because they provide convenient facilities for the definition of office procedures. The second group is rather less well defined at present and we refer to them as 'message and document processors'.

Procedures processors

As PCs and networks enter the workplace in increasing numbers, electronic mail will be increasingly widely installed and used for routine office tasks. Many of these tasks are carried out by following clear procedures, often formally documented, which lay down the format of a document, when and to whom it should be sent, and what the recipient should do with it. Today's electronic mail systems cannot usually provide support for such procedures, because they do not have facilities for format definition, data validation, and the routeing of documents. These types of facilities do, however, exist in the workflow software used with electronic document management (EDM) systems, although these systems are too expensive for general office use. (Electronic document management was the subject of Report 70.)

Procedures processors provide workflow-definition facilities for PCs and host computers,

Figure 3.4 Collaboration-support systems typically comprise mail and document-management components and customising tools



most of which do not possess the full image capabilities of EDM systems. They therefore make it possible to automate routine and semi-routine office procedures that are based on the movement of documents between people rather than on the capture and analysis of structured data. Since most business processes include both the movement of 'documents', and the storage of information, the difference between an office and a data processing system is largely one of viewpoint. Procedures processors are therefore best seen as a new kind of system development tool, based on messages rather than databases. Commercially available products include:

- LIFE (Linked Information Environment), from Motorola Computer Systems Inc. This Unix-based product runs on Motorola hardware, and has four modules (for producing electronic forms that look like existing paper forms, for providing high-volume data entry for back-office activities, for providing high-speed, high-capacity workgroup spreadsheets, and for workgroup electronic mail).
- Staffware, available in Europe from London-based FCMC (Financial & Corporate Modelling Consultants). This is also a Unix-based product that has been implemented on IBM PCs (running Xenix), NCR, Unisys, and ICL computers. It automates routine document-based procedures by creating and routeing documents, ranging from online internal memoranda to complete legal contracts.
- Workhorse, available from a Dublin-based company of the same name. Also Unix-based, this product is a multitasking, multi-user package that helps office workers to automate their procedures. An MS-DOS version is planned for early 1990.

Experience with a mail-based procedures processor at Philips Electronics has shown the power of this approach. Although this system was originally installed to provide office automation facilities (electronic mail, diary management, and so forth), it has now been used to build over 100 business applications within the Philips group, including several that would otherwise have had to be developed as conventional systems. One, for instance, supports the processing of credit claims. Within each major business area, all credit claims are routed to a single credit controller. Philips has defined and developed a structured 'electronic forms' procedure that automatically routes claims from anywhere in the group to the correct controller, through a series of control and authorisation stages. Because of the inbuilt equivalent of an applications generator, the system took only three man-months to develop and implement, most of which was analysis and liaison with the user departments. Apart from staff savings, the system has reduced the time to process each claim by 20 per cent, and increased control, resulting in significant reductions in outstanding debts and consequent improvements in cash flow.

Since the ideas of document flow and document format are familiar to every manager, procedures processors are likely to be particularly suitable for use by managers who wish to develop their own computer systems.

Message and document processors

We have coined the term 'message and document processors' as a convenient means of labelling two innovative systems in the groupware field. Because this type of system is so new, it is not possible at present to give a definitive description of a message and document processor. Instead, we simply describe the philosophy behind, and note some of the special features of, an experimental system developed at MIT, called the Information Lens, and the Lotus Notes product, due to be released by Lotus Development Corporation in 1990.

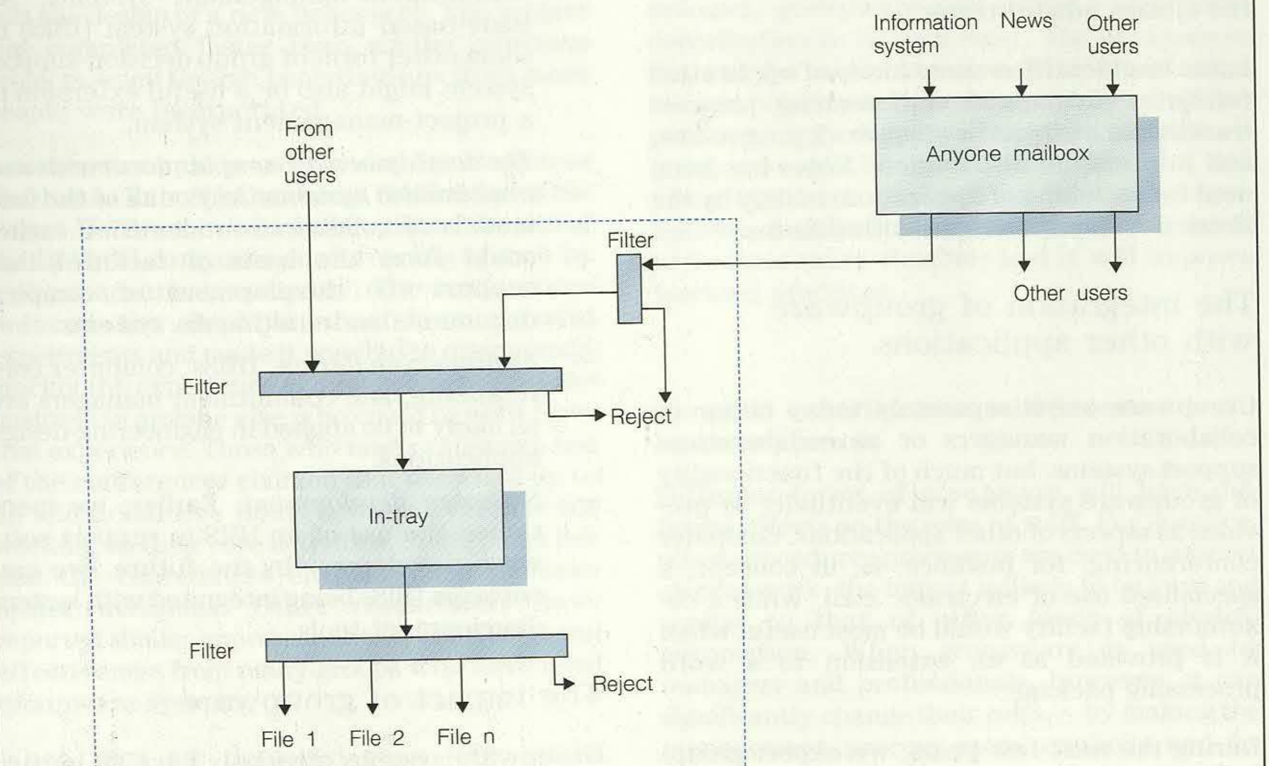
The Information Lens: Users of electronic mail systems often find themselves with more mail, and especially with more unsolicited mail, than they can handle. Often, the more sophisticated users of such systems have developed programs to filter their incoming mail, suppressing items of little interest and giving priority to those likely to be most significant, such as messages from their managers. These programs usually have to be written in procedural languages, and their effectiveness is limited because it is difficult to recognise the subject and meaning of the largely unstructured incoming messages. At the same time, managers and professionals, with or without electronic mail systems, often find that they are not told about useful facts because those with the facts at their fingertips did not know that they would be interested in receiving them.

Professor Tom Malone and his team at MIT have attempted to solve both these problems by developing an enhanced electronic mail system called the Information Lens, which runs on powerful Xerox workstations. (The principle underlying the system is shown in Figure 3.5.) Similar systems are being developed by several vendors. Information Lens users use simple 'if . . then . .' rules, rather than complete programs, to specify the actions to be initiated as the result of an incoming message. These rules may delete, file, or set priorities for the messages. The messages may be structured by using 'message templates' defined by the system administrator or individual users. This means that the type of message can be determined unambiguously, and the rules for dealing with the message can be more effective. Users can also make their outgoing messages available to anyone who might be interested in them by copying them to the 'Anyone' mailbox. Each user can also define rules that·scan both the messages in the Anyone mail box and other public information sources, such as newswires, and pass relevant copies to them. This allows users to regulate the inflow of unsolicited mail.

Lotus Notes: Lotus Notes is an information management system that allows people working on the same, or related, topics to share their documents. Notes users need not work in the same location, but each location will need its own copy of the Notes database. It also provides tools that can be used to construct document-based applications. Since such applications are unfamiliar to most systems staff, Lotus intends to supply a range of complete application templates that users can either install as they stand or modify to meet their business needs.

Notes has two components — a document database server, which typically runs on a

Figure 3.5 In the Information Lens, a series of filters enables a user to define the actions to be taken as messages arrive in his mailbox



local-area-network server, and a workstation program, which runs on an MS-DOS or OS/2 PC. This program provides a colour graphical interface to the Notes system, and development tools for creating Notes applications. Notes includes its own word processor and it can process documents created by other word processors, and graphics and spreadsheet packages. The Notes word processor is intended for short documents and will probably not be appropriate for documents that are either very long or highly structured.

Notes users can define document formats, including validation rules for individual fields. Associated actions can also be defined. For instance, a document type can be defined so that it will automatically be transmitted to a specified person. Whenever an instance of the document type is entered into Notes, a copy will be sent to that person, using the embedded mail system. Summarised document files can be constructed, where certain fields taken from the documents are visible on the screen, but the rest can be seen only if the document is opened. Notes has its own security system and tools for the system administrator.

Lotus has identified three kinds of application for Notes — computer conferencing, progress tracking for a project or other work programme, and information distribution. Notes has been used for each kind of application, mainly by the Notes development team within Lotus.

The integration of groupware with other applications

Groupware exists separately today either as collaboration managers or as collaboration-support systems, but much of the functionality of groupware systems will eventually be provided as aspects of other applications. Computer conferencing, for instance, is, in concept, a specialised use of electronic mail, while a co-authorship facility would be most useful when it is provided as an extension to a word processing package.

During the next few years, we expect groupware functions to be added to PC packages and office systems. We also expect the market for dedicated groupware systems to grow. This pattern of development will be similar to that of desktop publishing. During the past five years, the success of this type of system has led to the provision of desktop publishing functions in the most advanced word processors, although there is still a healthy market for dedicated desktop publishing packages.

We also expect to see groupware functions provided as part of specific applications, to create 'collaborative applications' in areas such as executive information, marketing support, computer-aided design, and software development. The structure of such an application is illustrated in Figure 3.6. An early example of a collaborative application is provided by an advanced document-creation system, such as Xerox's Viewpoint, which provides support for co-authorship of documents.

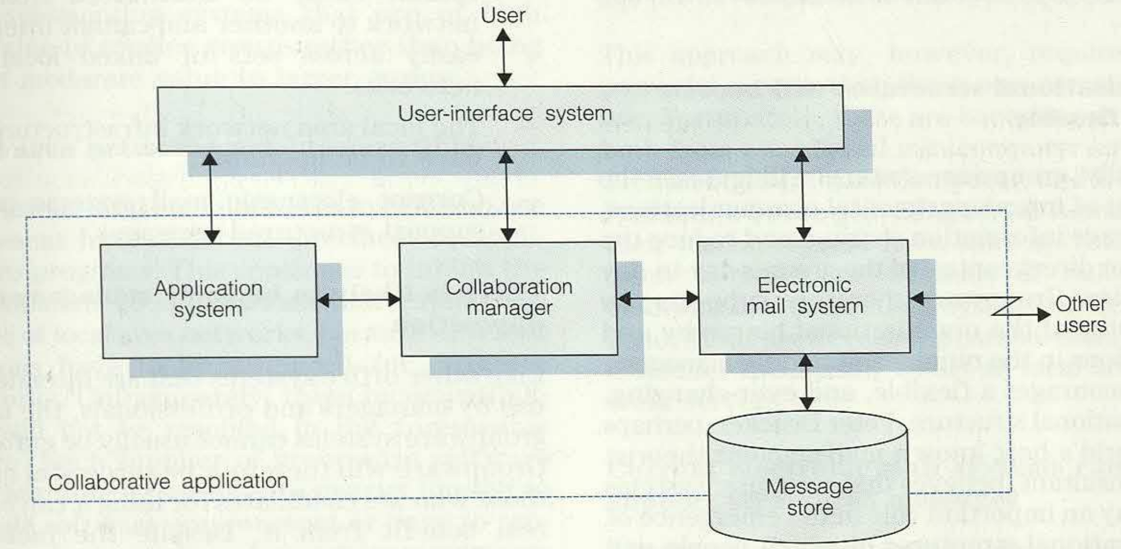
Other areas in which collaborative applications are either under development or could be useful include:

- *Project management*: There will be advantages in integrating resource-allocation and resource-monitoring functions with commitment-management systems. An issue-based information system (IBIS) or some other form of group decision-support system might also be a useful extension to a project-management system.
- *The development of complex documents and multimedia systems*: Any or all of the four models of collaboration identified earlier could form the basis of facilities that support the development of complex documents and multimedia systems.
- *Engineering design*: IBISs, computer conferencing, and commitment managers are all likely to be applied in engineering-design applications.
- *Software development*: Earlier, we mentioned the use of an IBIS to support software developers. In the future, we can envisage IBISs being integrated with system development tools.

The impact of groupware

Groupware systems obviously have an impact on organisational structures and individual roles, but the most significant impact is that they increase organisational efficiency and effectiveness. McDonnell Douglas, for example, used

Figure 3.6 **A collaborative application comprises a specific office or data processing application system that has groupware functions integrated with it**



The Coordinator commitment manager to support the design of a new helicopter. The project was completed faster than similar previous projects, even though contributions from more people were incorporated.

Another example was provided by Professor Shoshanna Zuboff in her book, *In The Age of the Smart Machine*, in which she quotes the use of the DIALOG computer-conferencing system by a pharmaceuticals company. The conferences reduced the duplication of unsuccessful experiments and made it possible to disseminate quickly the experience gained in one part of the business to anyone else who could benefit from that experience. Those who made extensive use of the conferences claimed that they had up to an additional five hours a week available for working on their core activities. Some users felt that the conferences enabled them to make better decisions. Other researchers have reported similar improvements in efficiency and effectiveness from many groups who have used groupware systems.

When they are deployed on a large scale, groupware systems can potentially have at least as great an impact on the work of managers and professionals as data processing has had on the work of clerks. Although the impact will be closely interwoven with that of other technologies, groupware will make a significant contribution in its own right. The evidence so far, and the results of research into the changes in user behaviour brought about by computer conferencing, suggest that the impacts of groupware will be in three main areas: it will place a greater emphasis on personal contributions; it will allow organisational structures to become more flexible; and it will improve working practices.

Personal contributions will become more important

Different forms of groupware will have different effects on the roles of staff. For instance, when procedures processors are used to support clerical work, the impact is likely to be local and similar to that of other kinds of clerical automation. When groupware is used by managers and professionals, however, it can significantly change their roles — by making the management process more explicit, and by reducing the importance of hierarchical management structures. Groupware systems, especially commitment-management systems, make the process of management visible because they model parts of the process. As a result, organisations will eventually be able to recognise inefficiencies in their management processes and managers, and to take corrective action.

Organisational structures will become more flexible

Typically, groupware systems will increase the amount of interdepartmental communications, encourage information sharing, and reduce the need for direct control of the group's day-to-day activities. Groupware therefore supports the flattening of the organisational hierarchy and reductions in the numbers of middle managers, and encourages a flexible, and ever-changing, organisational structure. Peter Drucker, perhaps the world's best known management theorist and consultant, believes that computer systems will play an important role in the emergence of organisational structures in which people will be conducted, like an orchestra, rather than receiving commands through a rigid hierarchy, as in an army.

Working practices can be improved

The implementation of groupware provides an opportunity to identify and rectify defects in working methods, such as an excessive number of steps in an approval process, delays caused by people taking longer than necessary to respond to a request for action, or undue consideration being given to the views of particular individuals. The working practices of the team that the groupware is to support must therefore be analysed before a system is implemented. In some cases, this may cause the team to change its working habits, even before a groupware system is introduced, and this is an important beneficial side-effect of the process.

Problems in implementing groupware systems

There are four main problems with implementing groupware systems, one of which is managerial and three of which are technical:

- The managerial problem, which is much the most significant, is concerned with the likelihood of management opposition to the introduction of groupware systems.
- The different local area networks on which groupware systems will be based are not compatible, which means that systems cannot easily be transferred from one network to another and cannot interwork easily across sets of linked local area networks.
- The local area network infrastructure may have to be upgraded.
- Current electronic mail systems cannot support structured messages.

There is likely to be some management opposition

Like other office systems that are intended for use by managers and professionals, the use of groupware systems cannot usually be enforced. Groupware will therefore be successful only if those who are candidates for using it can obtain real benefit from it. Despite the fact that benefits can be obtained, there is likely to be some managerial opposition to the proposed introduction of groupware systems. This arises both because managers feel threatened by the impacts that the systems are likely to have on them and their roles, and because the use of some kinds of groupware, especially computer conferencing, looks too much like fun to be work. This perception has caused managers to veto conferencing systems in several organisations.

In her book, *In The Age of the Smart Machine*, Professor Shoshanna Zuboff tells of the successful implementation of a large-scale computer conferencing system in a pharmaceuticals company, and of its subsequent termination by senior management. This shows how hostility from senior management can lead to the collapse of a highly successful application with thousands of users. In the example quoted, a significant part of the managerial hostility arose from the apparent waste of time associated with the 'Computer Coffee Break' — an open conference, corresponding to informal meetings, in which users often spoke uninhibitedly about the company.

Management opposition to the introduction of groupware systems can sometimes be overcome by:

- Explaining the nature and likely impacts of the system to management before it is implemented.

- Establishing a code of good manners, so that the system is not used as a medium for disseminating offensive material.
- Implementing systems that are of high value to smaller groups rather than being of moderate value to larger groups.

Local area networks are incompatible

The various local area networks in common use at present have different interfaces to applications programs. This continues to inhibit the development of applications that exploit the power of local area networks, because different versions have to be produced for different networks. Unfortunately, these incompatibilities will not be resolved in the foreseeable future. Each supplier of groupware software will therefore either have to restrict himself to a single software environment or have to produce separate versions for each environment. Given the general trend towards using workstations from several vendors, and more than one architecture, the larger groupware vendors are likely to choose the latter option.

The local area network infrastructure may have to be upgraded

A groupware system consists of at least two parts, one supporting the individual user's work and one managing the interaction between users. (This pattern matches the division between a PC or workstation and a local area network server.) The part of the system that supports interaction between users needs to be available continuously, whereas the part that supports an individual needs to be available on demand. It makes sense, therefore, to run the two parts in separate computers — the former on a server or shared minicomputer, and the latter on the user's workstation.

Although this type of arrangement can easily be implemented on Unix workstations, it has been difficult to achieve on networks of PCs and Macintoshes because of the unsuitable nature of the interface between PC applications and servers. The higher-level standards now emerging, especially those associated with OS/2 and SAA, will make it increasingly possible to implement groupware in this way in the future. The announcement by Novell, a leading supplier of local area network software, that its NetWare 386 product will provide tools to support the splitting of applications into separate client and server components, will also encourage the development of this type of application.

This approach may, however, require more powerful servers than those now commonly in use. Fortunately, these are becoming available both from established minicomputer suppliers such as Digital, Prime Computer, and Hewlett-Packard, who are providing server functions on their machines, and from start-up companies (such as Netframe Systems of Sunnyvale, California and Zenith Data Systems of Glenview, Illinois) that are developing parallel-architecture machines specifically to act as local area network servers.

Current electronic mail systems cannot support structured messages

Ideally, groupware systems would be implemented as applications based on existing electronic mail systems, in the same way that data processing applications are often now based on database management systems. This would provide the groupware systems with the full facilities of the mail system, avoiding the need to include them in the groupware software. Unfortunately, most electronic mail systems are designed to support messages that are analogous to an office memo; the heading is defined, but the format and content of the body of the text is completely unconstrained. Most groupware systems, on the other hand, are based on the notion of a 'structured message'.

Structured messages have a predefined format, content, or routeing, or several of these. Some of their fields can be defined so that messages with certain aspects in common can be grouped together. Thus, a computer-conferencing system groups the messages about the same subject, an IBIS groups the messages that support the same opinion, and a commitment-management system groups the messages about a particular commitment.

To provide effective support for groupware, an electronic mail system should also be able to support structured messages. For the time being, it may therefore be inappropriate for groupware systems to use existing corporate electronic mail systems. In the medium term, we expect the providers of mail systems to introduce support for structured messages, at which point it may be possible to integrate groupware software with these systems.

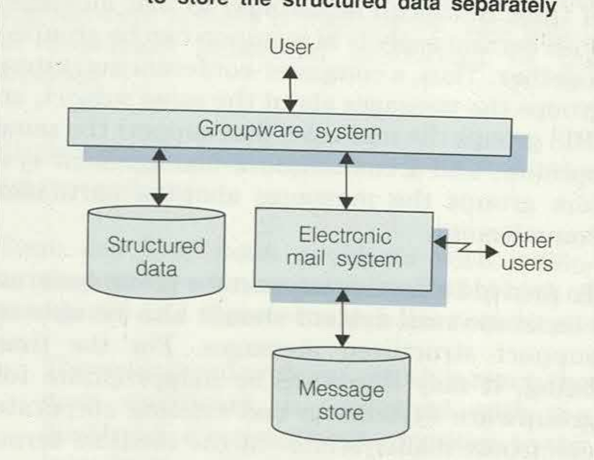
An alternative approach is to use an existing electronic mail system to store the messages, and to store the associated structured data separately, under the direct control of the groupware system, as shown in Figure 3.7. This approach, however, may require the applications to provide their own word processing facilities, and it is difficult to do this well.

Implications for systems departments

The idea of computer support for collaborative work creates opportunities for user organisations, but because these opportunities will not always be appreciated by business managers, it will be necessary for the systems department to explain them. Success with groupware is more than usually dependent on active support from those who will use the system. In taking the initiative to introduce groupware, systems departments should seek areas of the business where they are least likely to encounter staff who are likely to claim that "the PC is a single-function machine", "you can't communicate by computer keyboard", and "I can't possibly change the way I work".

We set out below the specific issues that need to be considered when selecting and implementing groupware applications. We also consider the impact that groupware is likely to have on overall systems planning.

Figure 3.7 An alternative approach to groupware is to store the structured data separately



Selecting applications

When considering appropriate areas of the business where groupware might be used, it is necessary to consider collaboration managers and collaboration-support systems separately. Because there are several kinds of groupware, and others are appearing all the time, there is no single set of criteria that can form a reliable basis for selecting suitable applications. We have, however, compiled a list of criteria applicable to the selection of each kind of groupware system we have identified. In general, it will be easier to justify a groupware application when groupware functions can be added to an existing application, rather than by introducing a complete application from scratch. It is likely to be appropriate to include groupware functions in an application when the intended users will need to discuss or negotiate their contributions with each other, rather than make them independently.

Collaborative applications may be useful in many situations, but since the groupware facilities are only a part of their function, they cannot be selected just on the basis of these facilities. It is just as difficult to make the cost/benefit case for a groupware system as it is for any other office system that provides support for managers and professionals. In view of the varied nature of groupware, the differing circumstances of user organisations, and the immaturity of the field, we are unable to offer detailed guidance. The best approach will usually be to set the likely costs against the probable benefits, and then ask the appropriate business manager to make a decision.

Collaboration managers

Commitment managers have the potential to be used in any organisation. However, their introduction can provoke hostile reactions and their use is likely to be successful only where:

- There is a high penetration of terminals and/or PCs (depending on the system to be installed).
- Staff are already used to communicating by computer.
- The work of managers, professionals, and perhaps, clerks, includes considerable variability, requiring frequent redefinition.

- Managers are rewarded for results achieved, rather than for following established procedures or maintaining personal relationships.

Computer conferencing can be based on any wide-area data network, or on a set of linked local area networks. It can be used to improve the flow of information within the organisation, and especially between physically separated divisions. Once established, a computer-conferencing system may also be used to enable groups of staff to collaborate more closely and to support decision-making. A wide-area conferencing system is likely to be successful in organisations that:

- Employ large numbers of technical/professional staff in multiple locations, especially if these are in several time zones.
- Have 'open' cultures, rather than those that disseminate information on a need-to-know basis.
- Have a high penetration of terminals, online PCs, and workstations among professional staff.

Issue-based information systems (IBISs) can be installed to support any team that performs design and decision-making functions. These include programming teams, design engineers, and product planners. These systems are likely to be successful in departments that already make extensive use of computers and that regularly confront complex and ill-defined problems.

Co-authorship systems can be useful for any group of people who develop large documents. Such groups include technical authors, but also any group that produces documents as an important part of its output — for example, research and development units, marketing staff, and designers.

Collaboration-support systems

Collaboration-support systems should, in principle, be chosen for their suitability for the proposed applications. However, this may be impractical, given the limited range of products and the lack of experience of relevant applications. We expect that products will evolve during the period 1990 to 1993, so that most of the major commercially available products will be able to address the greater part of most departments' requirements for collaboration support. Since this is not the case at present, care is essential in matching requirements to products.

Procedures processors are already an essential part of electronic document management systems. They should also be considered for systems where work procedures are based on processing documents according to predefined principles. Examples include dealing with customer complaints, handling budget requests, and processing engineering change orders. These systems will often be used in conjunction with other data processing systems for purposes such as transaction processing, financial modelling, or engineering designing, and it should be possible to switch rapidly between the different systems and to transfer information between them.

Message and document processors are likely to be useful where a group of staff produce and use large volumes of documentation. In particular:

- Mail-filtering facilities can be useful in any organisation that makes extensive use of electronic mail.
- The more sophisticated facilities of the Information Lens would be worth experimenting with in a technically sophisticated department that also undertakes a variety of non-standard tasks.
- A product such as Lotus Notes might be appropriate where joint authorship and document review are not central concerns. Where they are, a co-authorship system such as Idex might be more appropriate.

Implementing systems

In principle, implementing a groupware system is no different from implementing any other system that supports managers and professionals. However, the immaturity of the field means that there are two areas that require special attention — the need for prototyping, and the shortage of software products.

The need for prototyping

Groupware is still a new concept, and there will be considerable uncertainty about the exact features required and the scale of the benefits that will be obtained from implementing a specific system. It is therefore sensible to start by prototyping the proposed groupware system, providing that:

- The prototype can provide the most distinctive and critical of the features.
- The benefits do not depend on a large population of users (as they do for computer conferencing, for instance), or if they do, they do so in ways that are well understood, or the prototype can involve a large population. In the case of computer conferencing, it might be appropriate for a trial group of staff to participate in conferences provided as public services.

Proprietary products exist in each groupware category and may be used as the basis of the prototypes. If a proprietary product is to be installed, users should be involved in its selection.

The shortage of software products

Because of the immaturity of the market, there are very few groupware products currently available. The limited choice may mean, for example, that appropriate products require hardware, such as PCs, that is not already installed in the organisation, or software, such as Unix, that does not form part of the organisation's software infrastructure.

To resolve these difficulties, it may be necessary to install additional system software for a particular groupware application, or to upgrade staff from dumb terminals to PCs. If this is not an acceptable solution for an operational system, it may still be appropriate if the proprietary product is used as the basis of a prototype system. The operational version of the system can then be developed in-house and in accord with the established technical architecture. According to Bob Johanssen of the Institute for the Future in Menlo Park, California, many of the most innovative groupware systems already in use are in-house developments.

Impact on systems planning

The widespread use of groupware systems has several implications for information systems planning. These were discussed earlier when the problems associated with implementing groupware systems were described and include:

- The need for client-server architectures.
- The need for more local area networks and more powerful network servers.
- Reinforcement of the trend towards greater use of intelligent workstations.
- The possible need to change existing electronic mail systems.

Although the use of groupware is at an early stage of development, we believe that it would be prudent for systems departments to consider what changes to the IT infrastructure would be necessary to accommodate the widespread use of groupware. Many observers of the IT industry predict that groupware will be a major growth area. Without doubt, the growth will occur rapidly, once groupware begins to be introduced, just as the use of electronic mail has grown rapidly in the past few years.

Chapter 4
Hypermedia systems

Like groupware, the field of hypermedia (sometimes known as multimedia) systems has attracted a great deal of attention from the media, from academics, and from suppliers. Various commentators define hypermedia systems in different ways, depending on the particular aspect or feature they wish to emphasise. We use the term hypermedia to refer to computer-based interactive systems that provide flexible access to prerecorded audio, pictures, text, and full-motion video. Hypermedia systems therefore range from the limited use of sound and animation to enhance the interface to a PC application, to interactive videodiscs that allow the user to stop a film and call up explanatory material in the form of text or images. Access to information in a hypermedia system is often controlled by hypertext software. This means that each user is free to access the information in a sequence, or in a combination, that may be unique. We prefer the term hypermedia to multimedia because it highlights the flexible means by which information stored in the different media can be accessed.

Most of the hypermedia systems available today are 'publications', rather than applications. Their contents are assembled by an author (in practice, by a production team) and copies are then distributed to 'readers' (or, more accurately, 'viewers') who can access the information in the sequence, and with the combination of media, that best suits their purposes. A hypermedia application, on the other hand, would enable the user both to store and to access information.

Hypermedia therefore represents a convergence between the IT, publishing, and entertainment industries. This trend was foreseen a decade ago by Professor Nicholas Negroponte, now head of the Media Laboratory at MIT. He first presented this concept to Foundation members at a conference in 1980, and more recently, at the International Foundation Conference in Munich in 1987. The commercial significance of this convergence has also been recognised by major consumer-electronics suppliers, as delegates on the Foundation Study Tour of Japan heard in 1986. During the tour, Mr T Miamoto, Managing Director of Sony, said that Sony intended to specialise in image communication and expected to obtain half its revenue from non-consumer sales by the early 1990s.

Hypermedia systems are already being used successfully in commercial applications. Steelcase Inc, a leading US supplier of office furniture, has developed a highly visual system to support the launch of a new furniture product. Coldwell Banker, a US real-estate firm, has developed a system that allows a prospective purchaser of an industrial building to 'walk' through the building.

As well as generating a lot of excitement among those who use hypermedia systems, the field is also receiving significant interest from major suppliers. Apple is particularly active, perceiving it to be the next 'desktop publishing' — the next application area in which it can excel, and through which it can increase its penetration of the desktop systems market. IBM, with its low-priced Linkway hypertext product, is also active in the hypermedia field, and its backing, through Intel, of video-compression techniques, is also significant.

The wide variety of systems currently under development is an indication of the rapid growth that is imminent in the market for hypermedia systems. We believe that, by the mid-1990s, hypermedia systems will be used by most businesses and will be installed in many homes, providing information, education, and entertainment, as well as commercial opportunities, for those who position themselves to exploit the market for 'hypermedia publications'.

The nature of the technology

The developments that have made hypermedia systems practical today, and those that will occur in the early 1990s, are summarised in Figure 4.1 and described in more detail below. The main technological development is the combination of workstation and optical-disc technologies in products whose prices will encourage the development of a volume market, and the establishment of effective standards. Another important component of hypermedia systems is hypertext software, which will increasingly be used to control and facilitate access to the information stored in a hypermedia system. Developments in image-creation software will also be significant because they will allow photographic-quality images to be created and processed. Providers of public telephone and cable TV services are also enhancing their systems to provide hypermedia-like systems. Each of these developments is described in more detail below.

Workstation technology

To provide the basis for hypermedia systems, workstations require:

— High-quality colour screens.

— Sufficient processing power to drive the hypertext software and to provide acceptable response times.

In the past, these features have been available only for specialised, and therefore expensive, workstations. As shown in Figure 4.2, however, they are now being integrated in modern PCs such as the Apple Macintosh and the IBM PS/2. Bill Gates, chairman of Microsoft, expects that, no later than the end of 1990, a multimedia 'computer' (although 'interactive CD reader' might be a better term) will be available with a high-quality colour display and a CD-ROM drive, and costing about $2,000.

The use of PCs as the technical base for hypermedia systems allows the full creativity

Figure 4.1 Hypermedia technology has been developing since the late 1970s

Year	Workstation developments	Optical-disc developments	Software developments
The early days			
1978		Laservision announced	
1982		Digital audio compact discs available	
The pace increases			
1984		CD-ROM prototype	
1985		CD-I specification	
1986		CD-ROM production CD-I prototype	
1987	PS/2 launched Macintosh II launched	Laservision relaunched as CDV CD-ROM/XA specification DVI demonstrated	Guide (PC/AT) Guide (Macintosh) HyperCard (Macintosh)
1988		CD-I production	Apple CD drive
1989			Linkway (MS-DOS PC) HyperCard CD-audio toolkit Voyager (Laservision) SuperCard
The future			
1990		CD-I consumer shipments	
1991		First DVI chips available	

Figure 4.2 **Wide-ranging capabilities are being integrated on modern PCs to provide a basis for hypermedia systems**

(Photograph courtesy of Videologic, Cambridge, Massachusetts, United States)

of the PC-software industry to be harnessed. Existing skills are being supplemented by the creative skills of many people who have worked in the design, television, and film industries. This is a very powerful combination of skills and guarantees that many exciting new hypermedia systems and interactive 'publications' will appear in the very near future.

Within the next five years (in other words, by the mid 1990s), High Definition Television (HDTV) is likely to make an impact on the workstation displays used for hypermedia systems. HDTV is a group of advanced television technologies that provide much better picture quality by using approximately twice the number of scan lines that are used in today's TV transmission standards.

The development of HDTV has been pioneered in Japan, where a pilot service has already been introduced. Delegates on the Foundation's 1986 Study Tour of Japan saw HDTV demonstrated at Sony's Media World in Tokyo. The overwhelming impression was that HDTV pictures are orders of magnitude better than today's television pictures — the difference is as great as that between listening to music on an old-style AM medium-wave broadcast and listening to a high-fidelity broadcast. Advocates of HDTV claim that the proposed aspect ratio (the relative dimensions of the vertical and horizontal screen size) is particularly good for viewing sporting events and films originally made for the cinema, because like the cinema, the screen is wider than ordinary television.

Apart from its potential impact on the home-entertainment market, HDTV will provide low-cost technology for high-quality computer displays, which could facilitate development of hypermedia systems. HDTV will not, however, change the basic nature of hypermedia systems and it will not supersede existing broadcast TV standards before 1995.

Optical-disc technology

Interactive video, based on analogue Laservision optical discs, has been available for nearly 10 years. The earliest applications of optical disc were in the field of entertainment, although they are now used in several types of IT application, including file archiving and document-image processing. (Report 70, *Electronic Document Management*, described how optical discs form an integral part of EDM systems that can be used to store and manipulate the images of business documents.)

To date, optical-disc technology has provided a read-only storage medium. Several technologies are, however, under development for erasable optical discs, and one advanced workstation, the NeXT Cube (the first product from Steve Jobs's new company, NeXT), includes a 220-megabyte erasable optical disc as standard. Erasable optical discs have lower performance (in terms of access times and data-transfer rates), but higher capacity, than magnetic discs, and they require different approaches to indexing from read-only optical discs. They may therefore require the development of new standards, both for the medium itself and for the way the information is structured and stored.

The success of compact discs in the domestic hi-fi market has both stimulated the development of optical-disc technology and made systems developers more prepared to consider its use. Moreover, the ISO 9660 standard for CD-ROM (compact-disc read-only memory) has made it possible for organisations to publish data, text, and still images in a machine-readable form, and there is a growing variety of publications available on CD-ROM. Further developments in CD technology and standards (such as the CD-I specification) are addressing the need to include moving pictures.

Full-motion video presents a particular challenge for CD technology. About 500k bytes of data is required to describe a full-colour picture for display on a 13-inch screen. Conventional data-compression techniques make it possible to reduce this by a factor of about five, which allows a single CD-ROM disc to store up to 6,500 'frames' of information in compressed form. Full-motion video requires the image to be refreshed on the display screen at least 25 times per second, which requires a data transfer rate of at least 1,700k bytes per second. Since data is transferred from CD-ROM discs at only 150k bytes per second, this medium cannot support full-motion video when conventional data-compression techniques are used. Furthermore, 6,500 images are sufficient for only five minutes of video, which is not long enough for most applications.

At present, therefore, analogue (Laservision) optical-disc technology has to be used for the full-motion video element of a hypermedia system. However, the audio, textual, and static image elements might well be stored on conventional CDs or CD-ROM discs. (Storing the different elements on different media may also be necessary to reduce both response times and the cost of updating the material.)

Two approaches are being taken to overcome the problems of using CDs for full-motion video:

- By restricting the video display to only part of the screen, as in the original CD-ROM/XA and CD-I specifications (although in October 1989, Philips announced that CD-I will be developed to include full-screen moving pictures).
- By using better data-compression methods, as in the DVI specification.

CD-I

CD-I technology was developed by Sony and by Philips, who are now actively promoting it as an industry standard. It provides a digital data stream that can be used for text, audio, graphics, or full-motion video.

Up to 250,000 pages of text can be stored on a single disc, or by using data-compression techniques, up to 13,000 colour-photograph images. Alternatively, up to one hour of full-motion video can be stored, although the original specification of CD-I restricted the video display to one-eighth of the screen area. Several versions of the CD-I standards are available for audio and video, as Figure 4.3 shows.

CD-I requires a special-purpose disc player and proprietary software (the RTOS operating system), and is mainly aimed at home use. Products will be launched in the United States in 1990 and in Europe in 1991. At launch, a CD-I disc player is expected to cost $1,000, falling to $500 as a volume market develops.

CD-ROM/XA

The CD-ROM/XA standard allows industry-standard PCs using CD-ROMs to provide functions similar to those available with CD-I technology. The PC must, however, emulate the RTOS operating system. The main backers of CD-ROM/XA are Philips, Sony, and Microsoft, and it is also 'approved' by IBM.

DVI

DVI (digital video interactive) technology was developed by RCA and is now owned by Intel. The distinguishing feature of DVI is that only the differences between successive frames of a moving image are stored, which allows one hour of full-motion video to be stored on a disc. Storing the data in this way does, however, require very large amounts of processing power — three seconds per frame on a 64-processor parallel computer. Moreover, a special Intel chip set is required to reconstruct the original images

Figure 4.3 CD-I defines standards for audio and video recording*

	Maximum playing time (hours)
Audio	
CD stereo audio	1
Hi-fi stereo	2
'Mid-fi' stereo	4
Stereo speech (comparable to AM radio)	8
Mono speech	19
Video	
Original — one-eighth of the screen area	1
Enhanced — full screen	1

*Several types of recording may be included on the same disc.

as the disc is played back. The initial implementations of DVI will be as add-on boards for MS-DOS PCs. DVI technology is targeted at business users, and a board is expected to cost about $2,000 by 1991.

DVI will provide the kind of interactive-television capability that already exists for analogue Laservision discs, together with the text and data capabilities of CD-ROMs, in a form compatible with CD-ROM drives. It is by no means certain, however, that suppliers of hypermedia systems will adopt DVI as the standard for interactive, full-motion systems.

Hypertext software

Access to the different types of information stored in a hypermedia system will increasingly be controlled and facilitated by hypertext software. A hypertext 'document' allows a user to link separate 'chunks' of information in a way that best meets his particular need. Nodes may be linked to nodes in other documents, allowing the reader to pass freely between documents, when appropriate. (Hypertext was described in detail in a Foundation Position Paper published in August 1988.)

Extended hypertext systems, such as Apple's HyperCard, allow images and audio information to be stored as well as text, and provide a programming language that can be used to create special effects (such as animation) and links to other systems. Although Apple's HyperCard was not the first commercially available hypertext software, it was the first to be widely installed. (Apple now provides HyperCard free of charge with every new Macintosh computer.) Despite this, other vendors have also produced hypertext packages for the Macintosh, including some, such as Silicon Beach Software's SuperCard, that provide features superior to those provided by HyperCard, but that can read HyperCard documents. (Alan Kay, one of the founders of Apple, recently described SuperCard as "better than Apple's HyperCard".)

The large installed base of HyperCard software has prompted other suppliers to develop products that supplement and extend HyperCard, and that provide HyperCard interfaces to other products. A good example is Farallon's ScreenRecorder, which records the successive screens that constitute a Macintosh session. A function has been added to the HyperCard scripting language (that is, the language in which HyperCard documents are defined) to allow such screen images to be replayed under the control of HyperCard.

Apple's initiative with HyperCard has, in turn, stimulated the development of products for IBM (and IBM-compatible) PCs. These include HyperPad, a HyperCard clone for the PC, and Linkway, an IBM product now being sold at a remarkably low price.

Hypertext documents, called 'stacks' by Apple, can be stored on diskette, hard disc, or optical disc. Large hypertext documents, such as encyclopaedias and other major reference works, are often now supplied on CD-ROM discs. Those supplied on magnetic media can, of course, be modified by the user. For instance:

— A conference organiser could add his own views on local hotels, possibly including pictures, to a hypertext document that provides information about business conference centres.

— A sales manager could add evaluations of his own advertising campaigns, including artwork, to a hypertext document that describes different advertising media.

The recent appearance of erasable optical discs, first seen commercially with the NeXT Cube, is creating the possibility of large, updatable, hypertext documents.

To read a published hypertext, a PC with between 1 and 2 megabytes of main memory is usually adequate. However, larger and more powerful systems, in which a variety of editors and programming tools can remain in main memory, are likely to be required for creating them. In addition, special hardware may be needed to capture audio information and pictures from microphones, recordings, and cameras, and this type of equipment is available for PS/2s and Macintoshes from a variety of suppliers.

Software specifically designed to support the composition of hypermedia publications has also begun to appear. One of the first such products is Director, from MacroMind of San Francisco. Software tools for animating images are also available. Autodesk's Animator generates animation sequences on an IBM PC, and the HyperAnimator system from Bright Star Technology of Bellevue, Washington, US, supports animation on the Macintosh.

Image-creation software

Software packages are also helping with the creation of images in hypermedia systems. MacroMind's Director, for instance, provides facilities for the automatic animation of cartoons and literally dozens of visual 'special effects' such as fade and overlay. The HyperAnimator package from Bright Star Technology makes the image of a human or an animal face move in time with the speech being generated via the audio output channel.

In the past, computer-created images have almost always been immediately recognisable as having been generated by a computer. More recently, highly skilled graphics programmers, using very expensive hardware, have been able to produce images, even sequences of frames for moving pictures, that are indistinguishable from real photographs. They have solved the 'rendering' problem — the problem of replicating the various effects, such as perspective and reflection, that affect the appearance of real objects.

At present, hardware such as the Pixar image computer (which was seen by delegates on the Foundation's 1989 US Study Tour) is required as a front-end to a powerful workstation such as those supplied by Sun Microsystems. Image-creation software will, however, become more widely available on intelligent workstations as these increase in power and incorporate special parallel-processing graphics subsystems. When used by skilled staff, this software will provide the ability to create realistic images of buildings and machines that are described only by engineering drawings and specifications of the materials. It will even be possible to create these images dynamically in hypermedia systems (as is already done routinely in some engineering graphics applications).

The business implications of systems that allow a wide variety of alternative designs to be shown, in context, without the need for the construction of models, are clearly considerable.

Public network developments

Independently of the developments described above, the providers of telephone and cable TV services are actively developing hypermedia-like services that will be available to the public. France Telecom, for example, is enhancing its Télétel videotex service to make it multimedia. Delivering multimedia information over the telephone network requires ISDN services at a minimum of 64k bit/s. (ISDN provides a digital bit stream that can be used to transmit voice, data, text, and good-quality still images.) The availability of this level of service at a competitive price, together with supplies of multimedia terminals, will vary greatly between countries. ISDN services are already available to 50 per cent of French telephone subscribers, but progress is much slower in other countries. It will not, for example, reach this level in the United Kingdom for some years, and given the lack of enthusiasm for ISDN that is now apparent from many organisations, there may also be a lack of appropriate terminals.

Cable TV systems are better placed to deliver hypermedia services because they can combine analogue and digital transmission over existing cables. However, experimental installations in countries as diverse as Japan, France, Germany, the United States, and the United Kingdom have failed to find a successful formula in this area. We see no reason to expect this to change in the immediate future.

Leading applications

Early users of hypermedia systems are already gaining significant benefits. The most successful systems are in the areas of sales and marketing, training, and corporate communications. Hypermedia is also being used as the basis of a new class of products that we call 'infotainment' publications.

New tools for reaching customers

Several dozen companies, mainly in the United States, have already implemented hypermedia systems designed to provide eye-catching and informative material for customers. These systems are being used in both business and domestic markets. Systems for the business market include those developed by Steelcase Inc and Coldwell Banker (mentioned at the beginning of this chapter). The Steelcase system is described in Figure 4.4.

In the consumer-market area, both B F Goodrich and A C Delco now use hypermedia systems to sell their products in the low-cost retail chain, Pac Stores. Another example is provided by ACME, which sells 300 styles of ladies' boots in many colours. Obviously, retail stores cannot stock the complete range. ACME is therefore installing hypermedia systems in the stores so that a prospective buyer can be shown the complete range, and specify the combination of style and colour that she wants. She then orders the boots automatically and they are delivered to her home.

Also in the United States, the Design and Decorate Corporation has a hypermedia system to help its customers with interior decorating. The system combines photographs of furniture fabrics and wall coverings with 'wire-frame' models of rooms and furniture. It can then show the effects of various combinations of redecoration and renovation, or replacement, of furniture.

Buick (a division of General Motors) now supplies potential car purchasers with an 'interactive brochure' on a floppy disc that can be 'read' on a PC at the dealer's showroom. Potential customers spend more time 'reading' the brochure in this form than they would with a conventional brochure. Twelve per cent of those who subsequently bought a car bought a Buick, twice Buick's usual market share.

Figure 4.4 Steelcase Inc, a leading US designer and manufacturer of office-furniture systems, has implemented a comprehensive hypermedia system

The interactive software system was developed to support the launch of a new furniture product — Context™. The aim of the system is to position Steelcase at the leading edge of the industry, in terms of product and associated information support.

The system provides the first-time user with a wealth of information about the Context™ range of products, and it does it in a manner that is fun and easy to use. It uses animation, sound, and speech, and requires 120 megabytes of hard-disc storage. When it was shown at the 1989 National Furniture Exhibition in Chicago, it was received enthusiastically. As a consequence, Steelcase's competitors have begun to work on similar systems. The system could also be used by Steelcase's clients — who are office designers — to make presentations to their customers.

An ingenious consumer-oriented hypermedia application is being developed by the Computer and Communications Laboratory of the Singapore National Computer Board for a local hairdressing salon. The PC-based hairstyle-matching system will blend a picture of the customer with standard hairstyles stored in the computer, presenting these on screen for the customer's approval. The system also allows the hairdresser to modify the shape and colour of the style, and keeps records of customers.

A new training medium

Computer-based training systems have been available for many years, although they have met with limited success. Until recently, such systems were based on 'drill and practice' methods, because they were, too often, developed by programmers rather than teachers.

More recently, the emphasis has moved to providing more flexible and sophisticated facilities, such as the simulation of laboratory experiments. These developments have been made possible by the availability of optical discs, which, for the first time, have made it possible for a low-cost computer-based training system to include more text and pictures than a textbook. In addition, PC-based tools such as McGraw-Hill's Course Authoring System and Telerobotics International's Course Builder, provide facilities that allow teachers and lecturers easily to construct courses that make the best use of the technology available. Software, such as Video Builder (supplied by Telerobotics International), is also available to facilitate the integration of slides and video sequences into computerised lessons.

An example of successful hypermedia-based training course is described overleaf, in Figure 4.5. This system is used by Codex to train sales staff, systems engineers, and customer staff, to provide them with a basic level of competence in digital voice technology.

A hypermedia-based training system has also been developed by the Pecos River Learning Centre, which has produced an interactive, multimedia version of the SUDS business game developed at MIT. In the game, players run a

Figure 4.5 A hypermedia-based training course is providing substantial benefits

Codex Corporation

The Basics of Digital Voice Technology is a hypermedia training course, developed to train Codex systems engineers, sales representatives, and customer staff (predominantly non-technical managers). The main purpose was to ensure that system users had achieved a basic level of competence in digital voice communications. This basic level of competence was required before sending people on more advanced product courses. Prior to the introduction of the hypermedia system, this had been achieved by sending sales people on a one-day classroom-type course.

The system was sponsored jointly by Codex and Apple. Codex needed the system and Apple wanted a working commercial application to demonstrate its competence in hypermedia systems. The system was developed between May and October 1988. Implementation began in November 1988, and Codex began to sell the system externally in 1989 (which was not part of the original plan).

The system user is presented with a screen depicting the office of a communications manager in a user organisation. Requests arrive in the office by electronic mail from the organisation's Chief Information Officer, and the information needed to deal with them is provided from a library, from a lunch-time discussion with the manager's predecessor, from trade shows, and from technical seminars — all of which are simulated by the system.

Those who use the system are enthusiastic about it — two months before the system was due to be available, there was pressure from sales staff to get access to it, and in several cases, people who had achieved the required level of competence were still using the system a significant time later. Other benefits result from time savings. It takes between two-and-a-half and four hours to complete the course. It would take eight hours to cover the same material with traditional classroom methods. In addition, travel time and travelling expenses are eliminated, and the course is always available for refresher training. It provides a good method of 'just-in-time' training — bringing training closer to the time when the person actually needs the information. The system also makes it possible for managers to collect data on the strengths and weaknesses of users' communications knowledge, and to evaluate the course with a view to making possible updates or improvements in the future.

small business — a retail shop selling alcoholic drinks. Beer passes from brewers, through distribution companies, to the retailer, and thence to the customer. Players can see the effects of their decisions on this process, which is illustrated with graphics and appropriate sounds.

Other early hypermedia training applications include:

- Norsk Hydro, the Norwegian industrial firm, has developed an oil-rig simulation that allows users to take a 'virtual walk' around an oil rig. This system is being used in safety training.
- The British Royal Navy has developed a system that simulates visits to some of its ships — including interviews with personnel — to assist in recruitment.
- GTE North Inc of Westfield, Indiana, is using a hypermedia system to teach workers how to fix telephone cables. Those taught in this way learn much faster than they would with conventional teaching methods.
- Du Pont Co, a major chemicals company, uses a computer-controlled video system to train truck drivers.

Hypermedia systems have also been used in university education for some years. For example, courses on English literature are taught at Brown University, Rhode Island, using a very sophisticated workstation-based hypermedia system called Intermedia.

Hypermedia technology is also being used to create new types of educational material. In the United Kingdom, for example, the BBC's Interactive Television Unit has developed a videodisc-based product called Ecodisc, which is used for the education of environmentalists. The students take a simulated walk around a nature resort, looking at the scenery from different angles and at different seasons. This type of product is an example of an 'infotainment' publication, discussed below.

Improved means of corporate communications

Businesses can use hypermedia systems to improve communications with their workforces or with outside parties. Manufacturers of complex engineering products such as cars and aircraft have already begun to supply interactive documentation systems on CD-ROMs. Examples include Boeing, the Ford Motor Company, and Renault (whose hypertext system was described in the Foundation Position Paper on hypertext). Figure 4.6 shows a car mechanic using the Ford system.

The Texas Utilities Electric Company is using a hypermedia system to satisfy a different type of communications requirement. This company needs to make many presentations to regulatory commissions in order to retain its licence to operate the Comanche Peak nuclear power station. It commissioned The Hypermedia Group to develop an interactive database containing details of the 20-year history of the nuclear industry, the Comanche Peak plant, and relevant world events. The result is a 25-megabyte 'information' base of drawings, charts, pictures, tables, and text that is accessed via SuperCard. The information base is structured in several levels, so that presenters can give an overview, then 'drill down' to answer specific questions in detail.

'Infotainment' publications

There is already a substantial business in using CDs to publish material other than music. The Optical Publishing Association in North America says that, by the end of 1989, there were 425 non-musical CD titles, most of which were reference 'books'. Future non-musical CD publications will increasingly be multimedia.

Some very well-known companies are beginning to use CDs as a publishing medium. Time Warner, ABC News, and the National Geographic Society are all beginning to publish multimedia CD 'books'. Amongst the most interesting titles published to date on CD are:

Figure 4.6 Manufacturers of complex engineering products, such as cars and aircraft, have begun to supply interactive documentation systems on CD-ROM

The photograph shows a mechanic using the Ford Motor Company's service bay diagnostic system.

(Photograph courtesy of Office Workstations Ltd)

— The Domesday Project, a compilation of statistical and local information about the United Kingdom. The discs contain 25,000 maps, a gazetteer of 270,000 place names, 23,000 photographs, and 150,000 pages of text. (The text and photographs were supplied by school students.)

— The Time Table of Science and Innovation, a compilation of 6,000 key events in science and technology.

— Guernica, now in the Prado in Madrid, is a painting by Picasso of the fire-bombing of the Basque town of Guernica during the Spanish Civil War. In Robert Abel's Hypermedia version, the viewer can see not only the painting but also interviews with eye witnesses, newsreel footage of the bombing, an interview with Picasso, Picasso's sketches, and an assessment by an art historian.

— Mozart's The Magic Flute (to be published by Warner New Media). In addition to the music, the set of three CDs will include English and German librettos, scholarly criticism, and a dictionary of musical terms.

— Palenque, a guide to the ruins of the Mayan city. The viewer is given a conducted tour by an Indian boy, but can stop to examine items on the site.

— Compton's Encyclopaedia, which was due to be published on a CD in January 1990 by Encyclopaedia Britannica. Costing $895, it will include 15,000 illustrations and 45 animation sequences.

— A special version of the Ghostbusters film (published by the Voyager Company, California). This disc includes details of the special effects and the full script.

A further selection of hypermedia publications is shown in Figure 4.7, overleaf. These types of products are a rich source of information and can be used as powerful aids, or for private study. However, to those interested in the subject matter, they are also exciting and stimulating in their own right. They are therefore hybrids of information and entertainment — hence, the use of the word 'infotainment' to describe them.

Figure 4.7 There is a wide range of hypermedia publications available today

This list contains a small selection of existing publishers and publications to illustrate the range available.

Purpose of the publication	Publisher	Publication
Reference	Grolier Electronic Publishing	Encyclopædia
	Microsoft	Bookshelf, a volume that includes Roget's Thesaurus, the American Heritage Dictionary, The World Almanac and Book of Facts, Bartlett's Familiar Quotations, The Chicago Manual of Style, and the Houghton-Mifflin Spelling Verifier and Corrector
	Xiphas	The US national telephone directory
	Highlighted Data	Webster's Ninth New Collegiate Dictionary Electronic Map Cabinet
	Brøderbund	Electronic Whole Earth Catalogue
Education and training	Bright Star Technology	Various subjects for pre-school children
	Voyager Company	Master Class Collection — for example, the Brandenburg Concertos and Beethoven's Ninth Symphony
Entertainment	Activision	The Manhole, a new genre of game
Demonstration	Discovery Systems	Macintosh Showcase
	DECa snc di Toninop Ghetti & Co	Technical publishing demonstration

Moreover, the viewer of an 'infotainment' publication can decide on the balance between information and entertainment. The publisher still decides what to include and what to leave out, but he no longer controls the sequence and the flow of the material. The viewer can decide what to read (or watch, or listen to), and can then stop, think, and reread. He can also decide to review the material in summary from, or can explore it in as much depth as he chooses. Publishers will also be able to provide multiple viewpoints on a single disc.

Obstacles inhibiting the corporate use of hypermedia

The corporate use of hypermedia systems is inhibited, at present, by the difficulty and cost of the development process, uncertainties about standards, and the cost of fully featured hypermedia workstations. These last two inhibitors would almost certainly be removed by the development of a volume consumer market for hypermedia systems, and in the next section, we consider the likelihood of this happening.

The development of hypermedia systems and publications is complex. A multidiscipline team, including experts in the subject area, graphic designers, people with audio, film, or video expertise, as well as programmers, will usually be needed to produce a publication that compares favourably with the professional standards achieved by the entertainment industry. (Given their familiarity with television, the audience for hypermedia publications will not tolerate anything of lower quality.) Overall, the process of developing a multimedia application is more like making a film than writing a computer program although, compared with film-making, most of the cost is in 'designing' the application, not in shooting and editing the film itself.

It can also be very difficult to assemble a team of people with the necessary skills. People with all of the required skills are most unlikely to be found in the systems department. Indeed, they may not exist anywhere in the organisation, or at all. Even if an appropriate team can be assembled, the task of project managing the development of a hypermedia system is particularly demanding, given the range of skills and personalities that have to be integrated. The development process may also require substantial computer-processing power, particularly if DVI technology is to be used for recording full-motion video. At present, such processing takes three seconds per frame, which introduces a substantial time delay into the development process, and increases costs.

Several hypermedia-related technical standards have been proposed, but only the most basic, the CD-ROM standard itself, has received general acceptance. Other standards have influential backing, from suppliers such as IBM, Philips, Microsoft, and Intel, but hypermedia is a rapidly developing field and it is possible that future technical developments will supersede some of today's emerging standards. The use of erasable optical discs, for example, may require new standards to deal with the management of indexes.

The more powerful hypermedia systems available today are expensive. In particular, the need for separate optical disc drives for video and data increases the costs. This problem will be reduced as standards such as CD-I and DVI become established, and as more powerful personal workstations, such as IBM's PS/2, Apple's Macintosh II, and the NeXT Cube come into use. By the early 1990s, a hypermedia workstation with video capability should cost only about $500 more than a standard office workstation.

The possible development of a consumer market

The possible development of a consumer market for hypermedia systems and publications is of considerable interest to business users of the technology, because the mass production of hypermedia players will result in lower unit costs. More significantly, the presence of large numbers of hypermedia players in homes would provide both a market for new products, and new business opportunities for product and service promotion and delivery. Although there are favourable indications that a consumer market could develop, there are still several significant barriers that have to be overcome.

The favourable signs include:

- The rapid acceptance of conventional compact discs.
- The growth of non-audio CD 'publications' (although volumes remain small, as yet).
- The existence of an international standard for CDs, and the emergence of further hypermedia-related standards.
- The existence of successful applications and publications.
- The existence of large-scale disc-pressing capacity (there will be sufficient capacity to press at least one billion CD copies per annum by 1990).
- The high level of interest among vendors, particularly Apple, Hewlett-Packard, IBM, Intel, Microsoft, NeXT, Philips, and Sony.

The barriers include the lack of any obvious dominant application, the high cost of developing hypermedia systems, and the immaturity of optical-disc technology.

Except for the lack of a dominant application, all of these favourable indications and barriers have been discussed earlier in this chapter. A dominant application or publication could be the driving force in creating a volume market, in the way that spreadsheets were a prime mover in the growth in the use of PCs in the early 1980s. At present, however, there is no sign of such an application or publication.

Even so, we believe that it is highly likely that a volume market for hypermedia systems will emerge during the next five years, based on PC extensions. Whether hypermedia systems will enter the ordinary household in large numbers within this timescale seems more doubtful, but this is not a prerequisite for significant increases in business uses of the kinds of applications described earlier.

Implications for systems departments

Systems departments should begin to plan for the introduction of hypermedia systems. Three areas need particular attention — assessing the impacts on systems planning, identifying appropriate applications, and developing hypermedia systems.

Assessing the impacts on systems planning

In many organisations, the initial uses of hypermedia systems will be in the sales, marketing, customer-service, and public-relations areas. These initial systems will run on powerful workstations, and other specialised equipment and will not usually need to exchange multimedia documents with existing corporate systems. Thus, their impact on corporate systems and networks will be limited.

The visual appeal and high impact of these initial externally oriented hypermedia systems will inevitably create pressures for similar facilities to be available for in-house use. The workstations used for hypermedia systems will therefore, over time, need to be connected to corporate networks.

Early experience with multimedia communications systems shows that they provide users with additional capabilities and make systems more interesting to use. Thus, network connections, the spread of early experiences, and growth in the use of electronic mail and computer conferencing will create demands for multimedia communications within organisations.

As with the introduction of desktop publishing systems, and the growth in the use of PCs before that, systems departments will not be able to prevent hypermedia systems being installed, and any attempt to slow down their introduction, even if temporarily successful, will certainly be resented. Furthermore, the use of hypertext documents implies that printout will not be an acceptable output medium for hypermedia systems. This implies that, unlike desktop publishing, hypermedia systems will not be able to exist as a 'technological island' within a publication department.

For all these reasons, systems departments must plan for a general migration towards multimedia systems, although in many cases, this will not occur until after 1995. The most important areas to consider are the impacts on systems architecture and on network planning.

Systems architecture

Hypermedia systems require software (such as hypertext) and hardware (such as optical discs) with which many systems staff may be unfamiliar. Moreover, these systems cannot usually be implemented using established document architectures such as IBM's DCA, and will therefore require the use of products that do not conform with the organisation's existing software standards. Any overall review of software standards should certainly include consideration of future needs for hypermedia systems.

Initial hypermedia systems should be based on emerging industry standards, and the best hardware and software for the job should be chosen for each development, regardless of existing corporate standards. (The only exception would be the case where hypermedia systems are to be integrated with other corporate systems in the short term.) In the longer term, we expect to see established systems architectures expanded to encompass hypermedia documents, and it will then be appropriate to consider the alignment of corporate practice with these architectures.

Network planning

The communications volumes involved with hypermedia systems will depend very much on the nature of the individual organisation and the policies of its systems department. However, once hypermedia systems begin to be used, the growth in multimedia communications traffic is likely to be very rapid. Since multimedia communications can increase the communications load by a factor of 10, or even 100, compared with text-only communications, the long-term implications for network planning are considerable.

Networks with the capacity for large volumes of multimedia traffic are likely to be based on 2.048M bit/s lines between sites, FDDI (Fibre Distributed Data Interface) local area networks for on-site backbone networks, and slower local area networks, such as Ethernets and IBM Token Rings, for smaller sites and general office areas. The need for high-capacity networks will also stimulate the installation of optical fibres to individual work locations, as discussed in Report 62, *Communications Infrastructure for Buildings*.

Identifying appropriate applications

The responsibility for identifying appropriate hypermedia applications rests firmly with business managers. The responsibility of the systems department is to ensure that it is sufficiently familiar with the technology and current applications to be able to explain their potential to business managers. In these discussions, the systems department should particularly emphasise the business problems that hypermedia systems have already proved able to address — sales and marketing, training, and complex technical documentation. For many businesses, sales and marketing applications are likely to be the most attractive initial applications.

As the 1990s progress, however, and more and more routine business functions are automated, organisations are likely to find that information delivery becomes more important than 'applications processing' in their IT environment. In following this trend, IT will simply be following the examples of the publishing, film, and television industries, the impact of which today lies in their creative products, not in the technology used to process the products. As this trend develops, there will be more and more opportunities for using hypermedia technology.

For example, many organisations have existing information and presentation material that they could repackage in hypermedia systems and sell, or use as the basis for new services. At present, those best placed to exploit these opportunities are the video, film, and television companies, and major broadcasting companies such as the ABC and the BBC are now entering this market. In the future, other companies may find that photograph archives, built up for such utilitarian purposes as building maintenance, could have a commercial value if they are made available as part of a hypermedia publication.

These developments imply that, in the future, the market for PC 'software' will increasingly be dominated by the suppliers of 'content' rather than tools, and that the content will not consist solely of information.

Developing hypermedia systems

We have already highlighted the fact that the development and the project management of a hypermedia system presents particular difficulties in assembling the project team, in managing the development process, and in designing the system. There are also novel difficulties concerning intellectual property rights of any existing material that may be used.

The project team

A multidiscipline project team will usually be needed to produce a high-quality professional hypermedia system. In particular, new types of skills will be required, and these are unlikely to be found in the systems department. The leadership of a hypermedia project will often come from the business unit for which the system is being developed, or even from a consultancy. The skills that the systems department should contribute are:

— Experience in requirements definition.

— Project-management skills, although these may need to be modified to deal with non-systems personnel from very different backgrounds.

— Knowledge of the organisation's current information, network, and systems resources, and of alternative, possibly simpler, means of meeting the requirements.

The development process

Once an appropriate application has been identified, the first, and most critical, stage in developing a hypermedia system is to define the requirements in business, rather than in technical, terms. Without this, it is not possible to choose the best combination of media, and there will not be a sound basis for estimating the cost and timescale for the project.

As with other novel areas of IT, it is possible to compensate for a lack of familiarity with the technology by building prototypes and trying them out both on the internal sponsors and on prospective users of the system. As a general rule, hypermedia systems should be developed iteratively, with new ideas and modifications being added as the development progresses. The whole team should be involved continuously in the process, so that everyone can make creative contributions throughout the project.

System design

Although clever graphics and special effects are good ways of capturing people's attention, they are not a substitute for useful information and functions. The entertainment aspects of hypermedia systems are insufficient for business users; the applications must also do something useful. It is easy to over-use the special effects made possible by the technology, especially the use of audio. Over time, experience will help to curb this tendency, but in the short term, it is sensible to insist that designers of hypermedia systems exercise restraint, perhaps even rationing their use of the more exotic facilities.

When designing any interactive system, it is always tempting to include functions that gather data about the users and the performance of the system. Hypermedia systems often include such functions but, in practice, the data gathered is rarely used. Such functions should be included only if it is clear how the data will be analysed and for what purposes it will be used.

Intellectual property rights

If material originally created by someone else is to be included in a hypermedia system, it is usually necessary to pay the holder(s) of the intellectual property rights to the material. In an extreme case, this could mean negotiations with a studio, several actors and musicians, a composer, a lyricist, and the author of the screenplay for permission to include a short sequence of film in the system.

Despite these difficulties, we recommend that Foundation members begin to gain experience with developing hypermedia systems. By the mid 1990s, hypermedia systems will be well established for sales and marketing applications and for corporate communications, and they will be beginning to be used in decision-support and office systems.

Chapter 5

Tools and techniques derived from artificial intelligence research

For the foreseeable future, the significance of artificial intelligence (AI) for commercial computer users will lie not in the ability to use 'artificial intelligence', as such, to solve difficult problems, but rather to use the tools and techniques that are created as a byproduct of AI research to solve more modest problems. Some of these tools and techniques will, however, expand the range of problems that can be tackled, and this is the reason that there is currently such interest in them.

Many systems departments, for example, are now using expert systems (or intelligent knowledge-based systems), which are based on rule-based systems and the associated methods of determining the appropriate rules (known as 'knowledge elucidation'). By the mid-1990s, other techniques, notably object-oriented programming, will also move out of AI research laboratories and into commercial use in systems departments. These techniques will be embedded in applications and system software as application-specific knowledge bases, and as 'fifth-generation' development toolkits.

Knowledge bases embedded in system software will improve the performance of computers and networks and will provide advice that will enable them to be managed better. Knowledge bases embedded in the support systems used by managers and professionals will improve the consistency and quality of their work and increase their productivity. Figure 5.1, overleaf, shows a design drawing generated by just such an AI-based engineering-support system.

Fifth-generation toolkits will increase the range of applications that systems departments can tackle. They will, in particular, make it possible to develop systems that will be able to interpret complex sets of rules — for example, the rules governing the entitlement to benefit from a pension or social security fund, or the front-end procedures for a weekly payroll system. At present, interpretation of these types of rulebooks cannot easily be computerised because they are so complicated and change at frequent intervals.

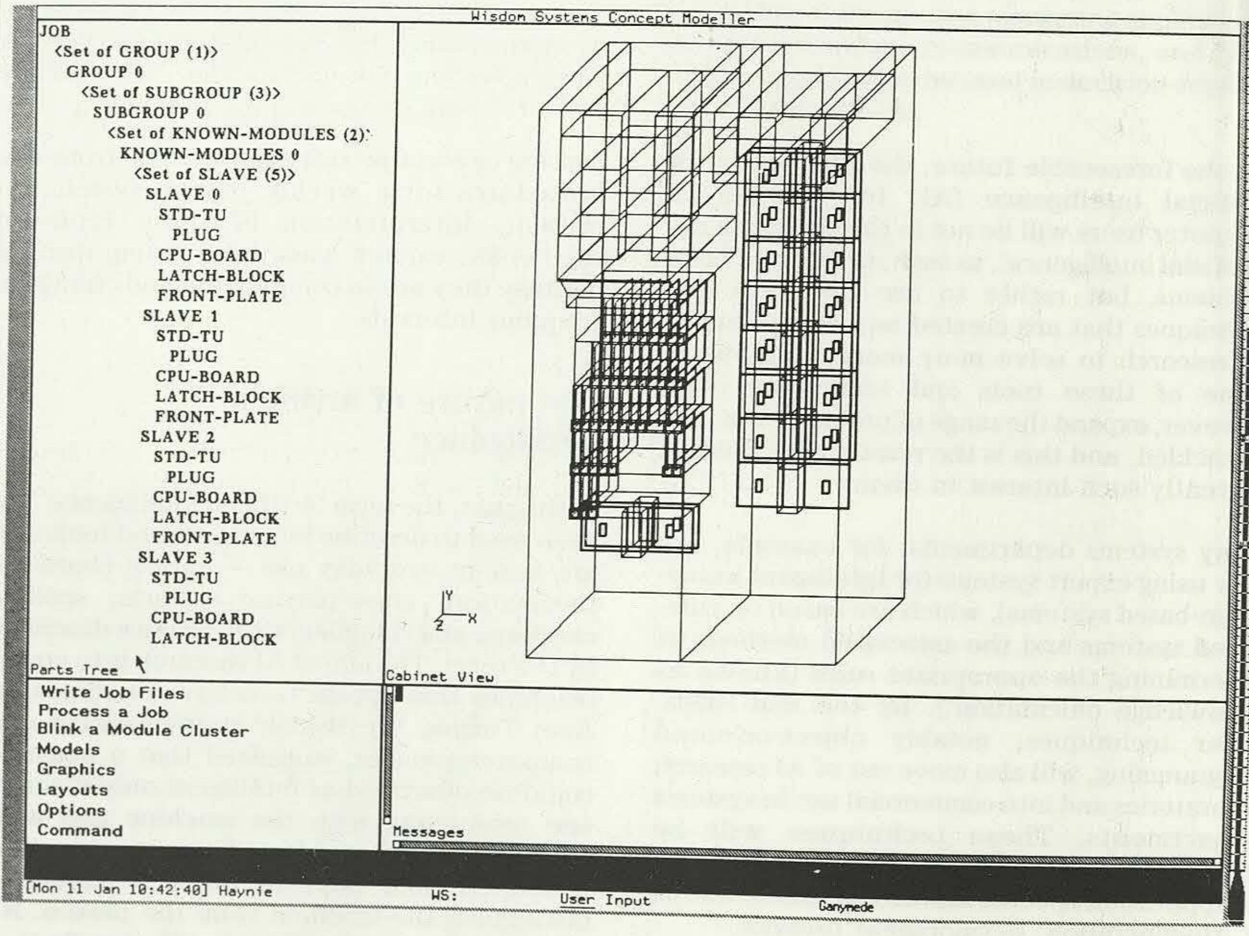
The nature of artificial intelligence

In the past, the term 'artificial intelligence' has been used to describe techniques and tools that are now in everyday use — optical character recognition, chess-playing systems, spelling checkers, and compilers were all once described by the term. The aim of AI research is to create machines that appear to behave intelligently. Alan Turing, the British mathematician and computer pioneer, suggested that a machine could be described as intelligent only if someone interacting with the machine and with another person, under circumstances that concealed their physical form, is unable to distinguish the machine from the person. No machine has yet satisfied the 'Turing Test'.

AI research is conducted by two main groups—cognitive psychologists, who seek to understand the mechanisms and processes of human thought, and AI scientists, who seek to produce results that would indicate intelligence if a person produced them. Since both of these groups deal with very complex problems, they have had to invent new analytical methods and programming tools in order to make progress. When these are packaged for commercial use and employed in business applications, they can be of considerable practical value, both in the form of complete technologies, such as optical character recognition, and as tools and techniques for use by programmers, professionals, and managers. However, commercial

Figure 5.1 AI-based techniques can be used to generate complex engineering designs

The picture is an example of the output produced by The Concept Modeller, a design-automation tool, which is based on an object-oriented approach. The example shown is the arrangement of equipment in a set of cabinets for a microprocessor-based distributed process-control system.



```
JOB
  <Set of GROUP (1)>
    GROUP 0
      <Set of SUBGROUP (3)>
        SUBGROUP 0
          <Set of KNOWN-MODULES (2):
            KNOWN-MODULES 0
              <Set of SLAVE (5)>
                SLAVE 0
                  STD-TU
                    PLUG
                  CPU-BOARD
                  LATCH-BLOCK
                  FRONT-PLATE
                SLAVE 1
                  STD-TU
                    PLUG
                  CPU-BOARD
                  LATCH-BLOCK
                  FRONT-PLATE
                SLAVE 2
                  STD-TU
                    PLUG
                  CPU-BOARD
                  LATCH-BLOCK
                  FRONT-PLATE
                SLAVE 3
                  STD-TU
                    PLUG
                  CPU-BOARD
                  LATCH-BLOCK
```

```
[Mon 11 Jan 10:42:40] Haynie
```

(Source: Wisdom Systems (UK) Ltd)

systems designers and users will not be particularly interested in whether their systems really are intelligent — they wish to know only that the tools and techniques can be used to solve a business problem. These different perceptions about the nature of AI have not been understood by most vendors or consultants, who have tended to equate AI with a particular technique or method.

Indeed, many systems departments at present equate AI with expert systems. Since 1987, when we last reported on expert systems (Report 60, *Expert Systems in Business*), many Foundation members have begun to use expert system shells — that is, programming environments for rule-based languages. When these shells were first introduced, great emphasis was placed on their ability to capture human expertise, since this was what AI researchers had used rule-based languages for. A survey of actual applications, however, shows many in which there is either no expertise, or in which it is present only as mathematical algorithms or as displayed text.

A major oil company, for instance, has used an expert system shell to develop a support system for its North Sea Emergency Centre. This system displays information and instructions to the emergency crew, the choice of display depending on the nature of the emergency. However, the expert system shell was chosen for this application not because of its reasoning power, but because the development staff were already familiar with the shell, and could use it easily to develop the required system.

Some commercial 'expert systems' do not even have an inference engine — the software that provides automated reasoning, the substance of which is expressed as rules in a knowledge base. An inference engine is usually regarded as one of the main features of an expert system.

Thus, there is now a clear distinction between the idea of capturing human expertise in an expert system, which might be done in many ways, and the technique of rule-based programming. Rule-based programming is an advanced programming technique that offers substantial productivity gains because it provides a convenient notation for solving certain types of problems and because of its nonprocedural nature. Expert systems are just one class of business application to which rule-based programming may be applied. Other applications are found in cognitive-psychology and AI research, although these applications are of no practical interest to commercial systems managers.

Recent advances

Rule-based programming is an example of an AI-derived technique that is beginning to be used in commercial computing. Figure 5.2 lists some of the other AI-derived techniques that have the potential to be used in a similar way. Of these, only neural networks and object-oriented programming systems seem to be developing rapidly. Neural networks form the subject of the next chapter. Object-oriented data management systems were discussed in Report 64, *Managing the Evolution of Corporate Databases*, and object-oriented programming systems will be dealt with in Report 74, *The Future of System Development Tools*.

Another interesting area of development concerns the extension of expert systems to encompass very large knowledge bases, although the work is unlikely to result in commercially usable tools and techniques before 1995. Researchers at the Imperial Cancer Research Fund in London and at the Microelectronics and Computing Corporation in Austin, Texas, are developing the approaches and tools that will be needed to construct very large knowledge bases, containing from 10 to 100 million items of information.

The Imperial Cancer Research Fund team is laying the foundations for an expert system, the Oxford System of Medicine, that could advise a general medical practitioner in his work. The prototype knowledge base contains 5,000 items (where items might be rules, text, or images) and the complete system is expected to need between one and ten million items.

The Microelectronics and Computing Corporation was founded by the US electronics and computing industry as its response to the Japanese Fifth-Generation Computer Project. The Project CYC team is creating an information base of general, commonsense, knowledge. The aim is to create a system that will be able to understand newspaper articles and

Figure 5.2 There are many AI-derived techniques that have not yet moved into large-scale commercial use

Blackboard architecture	A systems architecture based on a shared storage area through which a number of subsystems exchange data.
Fuzzy logic	A form of logic that allows for uncertainty, and that therefore produces only probable results.
Rule induction	A technique for deriving logic rules from actual or hypothetical data.
Neural networks	A computer architecture based on simple elements that, like the neurons in biological brains and nervous systems, have many inputs and a single output.
Frames	A method of structuring a knowledge base that recognises the existence of common structures.
Object-oriented systems	Systems based on the notion of 'objects' — structures that include both data and programs — and that represent things existing in the external world.
Quantified reasoning	A form of reasoning that makes explicit use of the probabilities that certain alleged facts are true or that certain implications can be correctly made.

encyclopaedia entries, and to answer questions based on this understanding. By the middle of 1989, the CYC knowledge base contained about one million data elements out of the 100 million that the team now believes will be needed. If successful, Project CYC's sponsors will use it as the basis for a future generation of expert systems that will be able to assimilate knowledge from textbooks and similar sources, use general knowledge to fill in the gaps in their formal, task-specific, knowledge, and create new knowledge and insights by generating analogies between different subject areas.

Emerging applications

Expert system tools and techniques (particularly expert system shells) are already being used to build significant commercial applications. Over the next five years, knowledge-based techniques will be applied in computer-aided engineering systems, and in other application-specific areas. AI-derived language-processing techniques will be used increasingly to simplify the task of accessing databases. Other AI-derived tools and techniques will be packaged in a new kind of system development tool, which we call fifth-generation toolkits.

Computer-aided engineering systems

Computer-aided engineering (CAE) systems provide support for engineering design processes. They go beyond established CAD systems by including either information about the properties of components and materials, or support for one or more design methodologies, or both. The developers of CAE systems are finding it increasingly practical to represent parts of the information and methodologies as knowledge bases, rather than by procedural programming, and CAE tools will become increasingly dependent on knowledge bases over the next five years.

These knowledge bases can be inspected by users, which means that they can be used for training and education purposes. As well as enforcing the rules of the methodology, they can be tailored to meet specific requirements, and extended to provide advice about good practice. Figure 5.3 gives examples of the benefits achieved by an engineering group from using expert system techniques to support customised engineering design.

Application-specific knowledge bases

CAE systems may be seen as applications that support a particular business process, whose core is engineering design. Other business processes, such as auditing and assessment for promotion, are also conducted in accordance with a written, or possibly implicit, methodology. Some of these methodologies are used by a wide range of organisations because they are necessary to meet legal requirements or are published as codes of practice. These methodologies can therefore be embodied in application systems or represented as application-specific knowledge bases.

Commercial computer users will increasingly be able to buy systems and tools that have application-specific knowledge built into them. Moreover, it will be possible to customise the knowledge to match the needs of a particular organisation. Such systems and tools will include factual knowledge and will enforce a particular method. This is already happening in the areas of CASE and network management, and several laboratories are researching the possibility of using this type of technique in the areas of decision-support and office systems.

Application-specific knowledge bases are a particularly interesting development because they have the potential to be integrated with existing systems in the application area. In the absence of standards, however, the publication of application-specific knowledge bases will not become a significant business area for some years. Nevertheless, we expect that this type of knowledge base will eventually become an important means of publishing information.

Database access

People often have considerable difficulties when they try to access a database with which they are unfamiliar. These difficulties include unfamiliar terminology and concepts (in both the database system and the access mechanism), and implicit definitions. The continuing rapid growth in the number and variety of online files and databases, and the increasing requirement for staff to consider a wider range of factors in their work, make these difficulties more significant. Language-processing techniques derived from AI research have been applied to this problem in products such as Intellect,

Figure 5.3 An engineering group successfully uses expert system techniques to support customised engineering design

McDermott

This US-based group of engineering companies builds a wide range of products. All the companies customise their products to meet individual customer requirements and expect to win business by competitive bidding. In 1982, it became clear that proposals for custom-engineering work were causing problems. The prices quoted were often based on inaccurate cost estimates, which either resulted in loss of the business or unprofitable contracts. Furthermore, it was not possible to estimate how much professional effort went into the preparation of proposals.

The group decided that it needed to build more component and design knowledge into the CAD system used by designers, especially for use at the quotation stage of a project. Early attempts to support a commercial CAD package with Fortran design-calculation programs, spreadsheets, and databases failed because of the special requirements associated with engineering customisation. McDermott therefore decided to investigate the use of an expert system, but commercially available expert system shells were found to be inadequate.

Thus, in 1983, McDermott started to develop its own shell, Concept Modeller, using object-oriented techniques, inheritance, back-tracking, and demand-driven calculations, in conjunction with an integral CAD modeller. Concept Modeller has been used by McDermott companies to design central utility boiler plant, offshore oil rigs, centrifugal fans, and sootblowers, and to configure cabinets for process-control systems.

Designers of the company's NETWORK 90 distributed process-control system access Concept Modeller via PCs and use it to configure processor cabinets. Using Concept Modeller in this way is estimated to reduce design labour costs by $160,000 a year, while producing more consistent designs, and more complete documentation. The company also benefits from better quality control and faster turnaround.

The Concept Modeller 'front ends' traditional geometry-based CAD systems and has been adopted by Prime Computer Inc as its recommended knowledge (or rule-based) front end for the Computer Vision CADDS 4X CAD system, which has one of the largest installed bases of some 40,000 workstations.

Concept Modeller is now available as a commercial product from Wisdom Systems (UK) Ltd, Wembley, Middlesex, and has been installed by a wide variety of engineering companies in Japan, the United Kingdom, and the United States, in the aerospace and automotive industries, and in other areas of mechanical or electromechanical engineering.

Symantec's Q&A, and TI's Natural Access. We expect the use of these products to increase and their effectiveness to improve as further advances are made in language comprehension.

The development of commonsense knowledge bases, such as the one being developed by the Project CYC team, will also help people to access databases. By 1994, they should have developed to the stage where they can help users to understand what is available in a database and to frame their queries in a way that ensures that they receive all the data they require and can interpret it correctly.

Fifth-generation toolkits

Fifth-generation toolkits form a new class of system development tool that supports a variety of programming techniques derived from AI research. They represent a significant advance on existing fourth-generation languages and expert system shells, and each product includes several tools. Examples include:

- The Aion Development System (ADS), which supports rule-based induction and deduction, quantified reasoning, and object-oriented development.
- Nexpert Object, from Neuron Data of Paris and California, which supports objects (structures that include both data and programs and that often represent things in the real world) and rule-based programming.
- Keris, from Integral Solutions Limited, Basingstoke, United Kingdom, which supports objects, frames, and rules.

These toolkits differ from earlier specialised tools used to develop knowledge-based systems, in that they are designed for use by ordinary systems analysts and programmers, not specialist 'knowledge engineers'. Programmers will not need to learn 'AI languages', such as Lisp and Prolog, in order to use these toolkits, because the facilities they provide embody the techniques developed by AI researchers in the past.

A further significant difference is that most of these toolkits run on mainframes and minicomputers as well as on PCs and workstations.

Thus, ADS runs on IBM mainframes, PCs, and Digital Vaxes, and Nexpert Object runs on Vaxes, IBM mainframes, PCs, Macintoshes, and workstations. Older expert system shells, such as Software A&E's Knowledge Engineering System (KES) are also being made available on mainframes.

Fifth-generation toolkits are already being used to build both expert systems and systems that include no elements of expertise. For example, staff at Yale University receive payment from many sources of funds and there are many different rules for allocating the money. The university has used ADS to build a payroll system that takes account of these differences.

In the United Kingdom, P&O Properties used TOP-ONE (a rule-based shell supplied by Telecomputing of Oxford) to build a system that apportions the operating costs of its Arndale shopping centres among the tenants. Each centre is occupied by many tenants and the costs have to be apportioned among them in ways that depend both on the degree to which they benefit from the various services provided, and on the specific terms of their leases. Many of these leases were 'inherited' by P&O Properties, having been negotiated originally with previous landlords.

In fields as diverse as the law, physics, and payroll calculations and deductions, the accumulated wisdom and knowledge are often expressed as rules. Fifth-generation toolkits will enable computerised systems for applying and interpreting these rules to be built much more easily than they could have been in the past. Our view is that leading user organisations will begin to use fifth-generation toolkits for mainstream systems development before the end of 1991.

The availability of fifth-generation toolkits will increase as suppliers of expert system tools add other AI-derived tools to their rule-based shells, and as suppliers of commercial system development tools add AI-derived tools to their product range, and then, more slowly, integrate them with their existing tools. For example, Information Builders Inc of New York, supplier of the Focus development toolkit, has bought the software house that developed Level5, a rule-based shell. By early 1989, Level5 could access Focus databases and could call Focus procedures as subroutines. During 1990, Information Builders expects to release a new version of Level5 that supports objects and frames, and that is more closely integrated with Focus. This version will include an object-oriented database management system — a development that we foresaw in Report 64, *The Evolution of Corporate Databases*. (Foundation members will have the opportunity to hear about these developments at first hand during the 1990 Study Tour.)

Software AG of Germany, supplier of the Adabas database management system and Natural (a fourth-generation language), will release Natural Expert in 1990. Natural Expert combines a sophisticated non-procedural language (technically a 'functional' language), an entity-relationship database, and a documentation and development environment called Semantic Associator. These features are based on concepts very similar to objects and rule-based programming.

The benefits of fifth-generation toolkits

The main benefit of fifth-generation toolkits is that they enable systems departments to develop and maintain complex systems where the requirements change rapidly. Today, third-generation languages are suitable for constructing very complex systems, provided that the requirements are relatively stable, or are likely to change in foreseeable ways. Fourth-generation languages enable systems to be built and changed quickly, but are not the most suitable development tool for systems that are very complex. Fifth-generation toolkits overcome both of these limitations.

At present, however, applications developed with a fifth-generation toolkit will cost much more to run than equivalent applications developed in third- or fourth-generation languages. Fifth-generation toolkits should therefore be used only where there is either a high degree of complexity and high levels of requirements changes, or where the problem to be solved is already expressed in a form similar to that required by the toolkit.

The classic example of fifth-generation toolkits allowing organisations to tackle applications that could not have been contemplated in the past is the XCON system used by Digital to generate Vax configurations. Before tackling this application using a rule-based programming system (an expert system tool), Digital had made several unsuccessful and expensive attempts to solve the problem by using conventional systems development approaches and tools. Digital turned to AI-derived techniques as a last resort. The wider availability of fifth-generation toolkits will make it easier to use AI-derived tools and techniques, and will ultimately lead to their becoming a normal part of systems development.

Barriers to the use of fifth-generation toolkits

The main barriers to the use of fifth-generation toolkits are:

— *Cost*: The most comprehensive toolkits (ADS, for example) cost as much as $250,000.

— *Lack of familiarity*: Fifth-generation toolkits are based on 'objects' and 'rules', concepts with which neither programmers nor their managers are familiar.

— *Lack of examples of successful use*: There are, as yet, relatively few examples of the successful use of fifth-generation toolkits.

— *Requirements for computer power*: AI-derived techniques are often complex and may require large amounts of processing power. These requirements are often difficult to estimate.

Over time, the significance of each of these barriers will be reduced. Prices will fall, the concepts will become more familiar, more successful implementations will become known, and the spread of parallel computers will provide much cheaper processing power for applications based on AI-derived techniques.

Implications for systems departments

Systems departments should expect future application packages of all kinds to include knowledge bases. This will increase both the scope for customisation and the degree to which such systems impinge on existing working practices, and will decrease the requirements for professional knowledge and skills among systems developers and users. This process will occur progressively over the next five years. To exploit these packages, systems departments will need:

— The technical skills to understand, modify, and integrate the knowledge bases with existing applications and development tools.

— The organisational skills to bring about the changes in staffing, workflow, and standards that will be needed to exploit the new opportunities arising from AI-derived tools and techniques.

— The business skills to translate the opportunities into a valid business case.

In the longer term, knowledge bases, like databases, will be critical corporate resources, with an important place in the systems infrastructure. Although the implications of this are likely to be considerable, the field is much too immature at present to identify either the appropriate strategies or the necessary infrastructure for long-term success. Nevertheless, we are convinced that Foundation members need to keep a close eye on the rapid developments that are occurring.

Chapter 6

Neural networks

A neural network (more precisely an artificial neural network) is a network of simple processing units, each modelled loosely on the functioning of a biological neuron, whose combined behaviour mimics, to some extent, that of an animal brain or nervous system.

Neural networks mark a new phase of artificial intelligence, based on low-level biological behaviour rather than high-level reasoning. They represent a decisive break with the data processing, decision-support, and expert systems approaches, because they are 'trained', not programmed. Moreover, unlike most expert systems, they do not require the trainer to specify how the problem is solved — only to supply sufficient examples of correct solutions.

Neural networks are not a new approach to computing, since the first theoretical work was published in 1943. However, their development and exploitation has suffered several set-backs, both technical and political. Figure 6.1 gives some of the history.

Furthermore, the field of neural networks shares common ground with several other disciplines. Researchers in signal processing and pattern recognition argue that neural networks are simply a new way of implementing algorithms that they have been using for years for identifying patterns in speech, video, and data communications. Neural networks also share much ground with research into parallel processing and there is considerable overlap with neuro-biology and cognitive psychology, with each field contributing to the understanding of the others. It is important, however, not to confuse these terms since the term 'neural networks' does refer to a very specific set of approaches to computing, which we explain in this chapter.

Suddenly, neural networks seem to be very much in vogue. The European Commission has announced funding of 10 million ECUs for two major neural network projects, one to develop tools and standards, the other to investigate possible applications. The US Defense Advanced Research Projects Agency (DARPA) has provisionally earmarked $400 million to spend on neural network research in the next few years, $30 million in 1989/90. From 1991, the Japanese Ministry of Trade and Industry will be sponsoring neural network research as the successor to its widely publicised Fifth-Generation Computer Project.

However, a major problem is that the field is becoming so hyped that it is difficult to identify the real developments. In this chapter, we attempt to distinguish truth from hype, and we look at the possible consequences for Foundation members.

The nature of the technology

All neural networks are based on simple elements that, like the neurons in biological brains and nervous systems, have many inputs and a single output. The output of each element is a simple function of the inputs, and this function is generated and adapted by training. The output is in turn connected to the inputs of many other neurons to form a complete network.

Thus, depending on the particular architecture used, a small network containing just 100 elements could have up to 10,000 connections (if each one was connected to every other). It is this 'rich connectivity' that is the source of neural networks' intelligence and robustness, rather than the combined raw processing power of their elements. For this reason, neural network research is sometimes referred to as 'connectionism'. A conventional computer typically has far greater total processing power,

Figure 6.1 The development and exploitation of neural networks have suffered several setbacks

1943 – McCullough and Pitts at MIT develop the first mathematical model of a neural cell in the brain or nervous system.

1947 – D Hebb proposes a simple model for neural learning, whereby the strength of any connection is increased whenever the neurons at both ends are active.

1959 – F Rosenblatt at Cornell University develops a computer simulation of the McCullough and Pitts model, called the Perceptron.

1969 – M Minsky and S Papert at MIT, two eminent professors of mainstream artificial intelligence, publish 'Perceptrons', demonstrating that the Perceptron model is unable to solve certain very simple problems (including the 'exclusive-or gate'). Funding for neural network research evaporates because of this paper.

1982 – J Hopfield at the California Institute of Technology proposes an alternative neural network architecture – the Hopfield Net. He also claims that such networks may succeed where rule-based systems have failed.

1986 – Rumelhart, Hinton, and Williams of Carnegie Mellon University create a multilayer version of the Perceptron model, together with rules for adjusting the connection weights. They demonstrate that this new architecture overcomes Minsky and Papert's objections.

1986 – T Sejnowski and C Rosenberg at John Hopkins University develop NETtalk – a neural network that learns to read aloud.

1989 – Several researchers and electronics manufacturers announce working prototype versions of neural network chips.

1989 – US Defense Advanced Research Projects Agency allocates $400 million budget for neural network research and development.

but far less 'connectivity' between the data items. As well as being the source of their power, the large number of connections is also the prime cause of difficulty associated with neural network development. Implementing 10,000 adjustable connections requires a lot of computational power in a simulation, or a lot of silicon in a hardware implementation.

There are many different architectures for neural networks although they have some common features. The most common, and the most representative, architecture is called a 'multi-layer perceptron' and this is illustrated in Figure 6.2, overleaf. The figure also shows how such a network might be applied to a share-trading application. The inputs are quantities such as share price and price/earnings ratio, while the outputs represent the possible strategies — increasing holding, maintain holding, and so forth.

Typically, there are three layers of elements, or neurons. Data is fed into the input layer via input lines. The data may be taken directly from the 'environment' by means of electronic sensors, or it may come from another computer program or database. In the case of an image-processing system, there might be one input neuron for each pixel — that is, each dot in the picture. In the case of a data processing application, there might be one input neuron for each data field in a database record. Indeed, it is quite common for a neural network to be embedded within a conventional computer system, with conventional programs being used to prepare the input data, and then to present the output from the neural network to the user or another program.

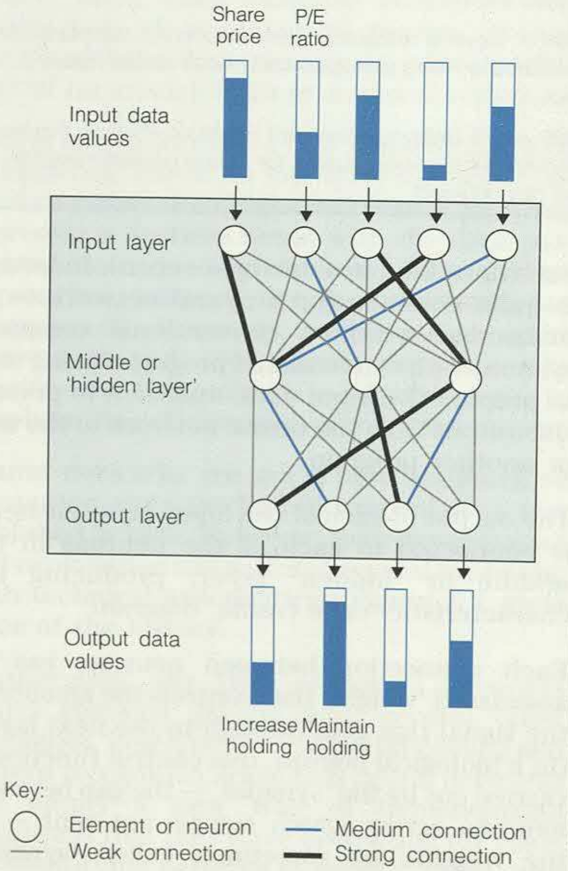
The output of each of the input-layer elements is connected to each of the neurons in the middle or 'hidden' layer, producing the characteristic 'cat's cradle' diagram.

Each connection between neurons has an associated 'weight' that controls the amount of the signal that gets through to the next layer. (In a biological neuron, this control function is carried out by the 'synapse' — the gap between neurons across which signals are sent — so the weights are sometimes called 'synaptic weights'.) The process of learning in a neural network is carried out by repeatedly adjusting the hundreds or millions of weights in the system until the inputs produce the required outputs.

In essence, an individual neuron merely adds together the inputs that it receives, each one multiplied by a synaptic weight to enhance or reduce its importance. The neuron 'fires' — that is to say, generates an output signal — if, and only if, the total 'weighted sum' of the inputs exceeds a threshold level. In some neural

Figure 6.2 The elements of a neural network are arranged in three layers

Every neuron (or element) in a layer is connected to every neuron in the next layer. The strength of each connection (also called the 'weight') can, however, vary. The weights of all the connections are adjusted until the network produces the required output from a given set of inputs. In this simplified example, the network is designed to assist in share-trading decisions. Input data includes quantitative information such as share price, plus certain value judgements. The output values correspond to actions that can result from the input data — the highest value indicates the recommended action. In the example, the recommended action is to continue holding the share in question.



network implementations, the output is binary—changing from '0' to '1' whenever the neuron fires. In other implementations, the output is analogue and the strength of the output signal is a function of the weighted inputs.

Each neuron in the middle layer is connected to each neuron in the output layer, which presents its signals to the user, or to another computer program. In a pattern-recognition application, there would usually be one output neuron for each possible class of patterns. Thus, in a handwritten input system, there would be one output neuron for each recognisable character A . . Z, 0 . . 9, and so on. The number of neurons required in the middle layer for a particular application, however, usually has to be determined by trial and error.

The knowledge learned by a neural network is effectively 'stored' in the strengths or weights of the many interconnections, rather than in any explicit form. However, researchers have occasionally found that the neurons in the hidden layer come to represent certain intrinsic properties of the data they are handling. In some speech-recognition applications, for example, it has been found that individual neurons in the hidden layer fire only when the word is a verb, even though the concept of a verb has never been presented to the network. This property seldom has any direct value, but it does provide insights into the way that neural networks learn.

To train a network, all the weights are initially set to random values. A series of known examples is then repeatedly presented to the network. Each time a known example is presented, the answer generated by the network is compared with the correct answer — the difference being called the 'error value'. In a process known as 'error back-propagation', this error value is used to adjust the values of all weights in the system that could have contributed to that error. For the first few examples, the network's output will be almost random, but as training continues, it will tend to produce the correct answers to the sample problems.

When the weights finally stabilise, the network is 'trained' and it is possible to move onto real problems. Sometimes, the weights will not stabilise, meaning that the network is unable to learn even the training examples. Occasionally, this can be rectified by restarting the training with a different set of random weights.

The state of the art

Most of today's neural networks are software simulations rather than physical networks. Particular applications are generally built using neural network packages such as the Decision Learning System from Nestor Computing. There are now several dozen such simulation packages on the market. Most are written in standard programming languages such as C, and many will run on PCs.

The performance (principally the speed of operation) of a simulated neural network can be enhanced by means of an add-on board. One such board can perform 22 million floating-point instructions per second. This magnitude of processing power is needed to recalculate thousands of weight values with each learning example. Parallel-processing technology will soon result in even more powerful boards that can be used to support neural networks.

A very specialised hardware implementation of a neural network that uses off-the-shelf memory chips to simulate neurons, and known as WISARD, has already gone into commercial application. Figure 6.3 provides further details.

A wider range of neural network applications will become feasible when purpose-designed chips reach the market. Several of the major semiconductor manufacturers are developing prototype neural network chips. Most are still extremely limited. Bell Labs has achieved 200 neurons on one chip, and that is among the densest built so far. The main problem has been designing a suitable circuit to implement the weights — storing them as digital numbers requires considerable space. The most promising devices store the weights as analogue voltages, but these are at a very early stage of development. Oxford University in the United Kingdom, one of the leaders in developing this technology, has achieved 12 analogue neurons on one chip.

Figure 6.3 WISARD is a working neural network implemented in hardware

WISARD was the brainchild of Professor Igor Aleksander and co-workers at Brunel University in the United Kingdom. (Professor Aleksander is now head of the electrical engineering department at Imperial College, London.) This team created a neural network element based on standard random access memories, and found that a large network of such elements operates in a manner analogous to a neural network. Even though it does not require special chips, such a network operates very much faster than a software simulation of a neural network. WISARD can recognise an image in 0.04 seconds, fast enough to examine every frame from a TV camera. One of the most celebrated features of the system is its ability to recognise human faces, and even to recognise certain facial expressions. A commercial version of WISARD has been on sale for several years now, and has been applied to a variety of problems, from site security to banknote sorting.

Another exciting development is the incorporation of simple neural networks into solid-state electronic sensors. The best example is the 'silicon retina', developed by Carver Mead at the California Institute of Technology, and soon to be commercially marketed. The chip combines a conventional CCD array (the electronic sensor at the heart of any modern video camera) with a simple neural network that processes the raw image by smoothing out noise, and enhancing lines and edges in the picture. The process is directly analogous to that of the retina in the human eye, which preprocesses the observed image before transmitting it to the brain via the optic nerve. Mead has since applied the same concept to auditory signals by developing a 'silicon cochlea'.

In the medium term, neural networks for embedded and high-performance applications (such as speech recognition and radar target tracking) will be developed as software simulations before being implemented as special-purpose hardware. Silicon-based neural networks will eventually appear in a vast array of products, from workstations to robots, and from military aircraft to microwave ovens.

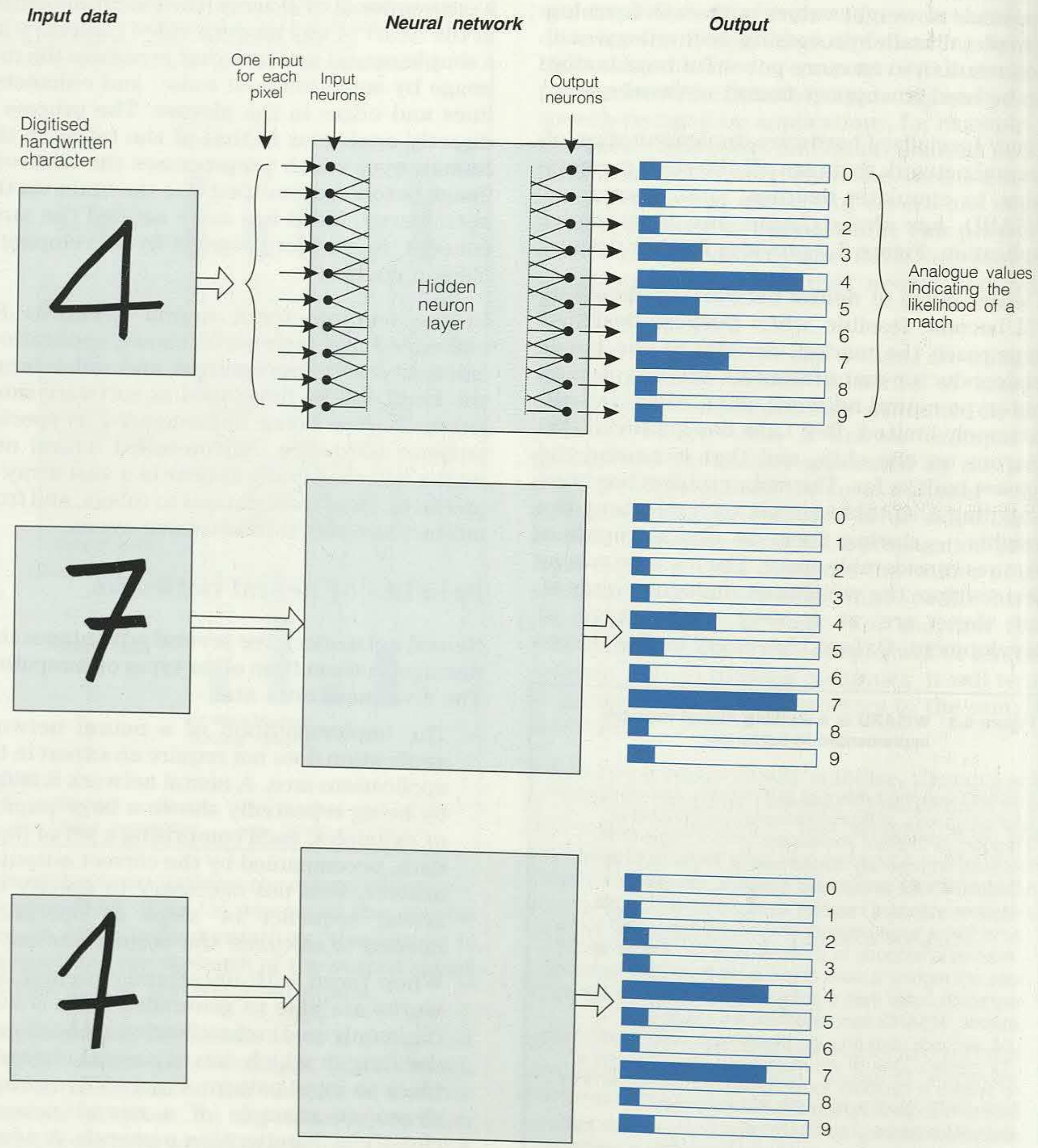
Benefits of neural networks

Neural networks have several advantages that distinguish them from other types of computing. The five main ones are:

- The implementation of a neural network application does not require an expert in the applications area. A neural network is taught by being repeatedly shown a large number of examples, each comprising a set of input data, accompanied by the correct output or answer. It is not necessary to specify the actual sequence of steps or operations needed to calculate the correct answer.
- When faced with uncertainty, neural networks are able to generalise. This is most commonly used in classification applications—deciding in which one of several classes to place an input pattern. Figure 6.4, overleaf, shows an example of a neural network classifying handwritten numerals. A neural network seldom produces an answer that

Figure 6.4 Neural networks can cope with uncertainty

Neural networks are very good at coping with uncertainty — such as in recognising handwritten characters. In this example, the input character is digitised onto a grid, and each pixel is fed into the neural network. There are 10 outputs from the network, each representing the likelihood that the input data corresponds to each of the 10 numerals that it has previously been taught. Note that, in the third example, the network's output indicates that the input character has an equal likelihood of being a '4' or a '7'.



is 100 per cent certain — the answer is taken as being the class whose output line has the highest value. In the case of the third example in Figure 6.4, the network is indicating that there is an equal probability of the input being a '4' or a '7'.

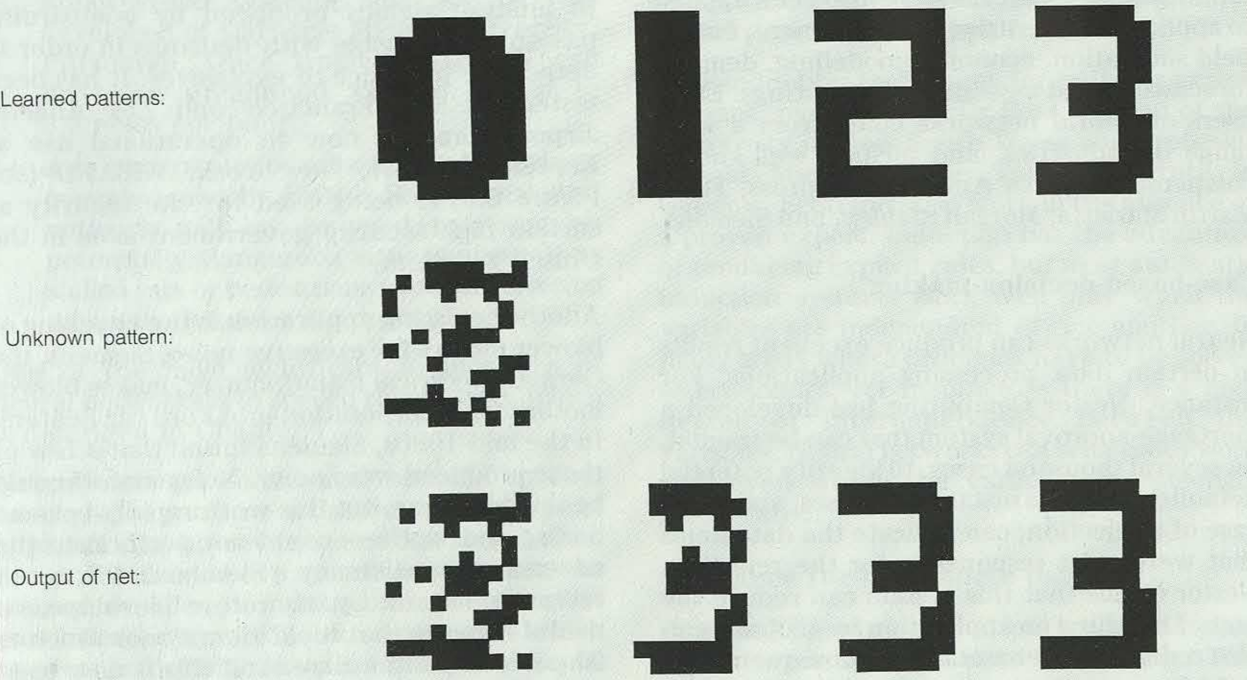
- Neural networks can operate with incomplete or distorted information. Figure 6.5 shows a specific kind of neural network (called a Hopfield Net) using an iterative process to restore a corrupted input pattern to one of the taught examples. This aspect of neural networks shares much with the field of error correction in data communications, and with the concept of 'contents addressable memory', where a database record can be retrieved on the basis of incomplete information.
- When neural networks are eventually implemented in purpose-designed hardware, especially those implementations that use analogue circuitry rather then digital, they will be extremely fast compared with even the fastest digital computers. This will make them applicable to certain complex realtime problems, such as speech processing or identifying moving targets.
- Neural network systems are extremely robust. In a large network, the failure of any one element or link may degrade the performance or accuracy, but it will not usually cause the system to fail.

These advantages produce two main business benefits — they extend the range of problems to which computers can be applied, and in favourable cases, they reduce the costs of developing certain types of applications.

Neural networks are particularly relevant to the processing of data from sensors, and they are increasingly seen as the optimum computing technique in this field. Neural networks are also

Figure 6.5 Neural networks can restore corrupted data

All neural networks can cope with incomplete or partially corrupted data. One particular type (called a Hopfield Net) is specifically designed for this purpose. In the example below, the net has been taught four patterns (0, 1, 2, and 3). When presented with a highly corrupted pattern, the network iterates towards the most likely solution. The technique has applications for error correction in data communications and for retrieving information from contents-addressable memory, when the input is complete.



beginning to be applied to more conventional data processing and decision-support problems, where they are sometimes able to solve problems that cannot be solved by other means. Finally, their learning capability, and their ability to generalise, allow neural networks to reduce the costs of building certain types of applications.

Leading applications

When dedicated neural-network chips become available (and are in turn incorporated into purpose-designed neural computers), there may be benefit in converting some conventional computer applications to take advantage of the speed of neural networks. For the most part, however, and certainly for the time being, there is nothing to be gained by converting existing successful applications so that they can be run on neural networks. The main benefits of neural networks arise from new types of applications — either applications where other techniques such as expert systems have failed, or applications that have previously been considered to be outside the scope of information technology.

The principal applications for neural networks may be divided into case-based decision-making, planning, and pattern recognition in sensor input. Neural networks have also been applied to applications as varied as recruitment, battle-field simulation, economic modelling, demand forecasting, and assessing credit ratings. Early users of neural networks come from a wide range of industries, and include well-known companies such as American Express, Ford, Martin Marietta, Morgan Stanley, and Siemens.

Case-based decision-making

Neural networks can produce excellent results in certain data processing applications. For instance, Nestor Computing has developed a mortgage-approval system that can be trained, on several thousand cases, to identify potential defaulters. It gives instant responses, and in the case of a rejection, can indicate the data items that were most responsible for the rejection. Nestor claims that this system can reduce the cost of handling an application by $50, and can also reduce the percentage of subsequent loan defaults.

A similar type of system could be used for assessing risks in insurance underwriting, in personnel recruitment, and possibly for valuing shares and bonds.

Planning

The best known of the few proven planning applications of neural networks is probably the Airline Marketing Tactician (AMT), developed by BehavHeuristics. AMT uses two neural networks to control an airline's pre-flight allocation of seats between classes. One network predicts the demand for seats in each fare class, while the other predicts the number of 'no shows' — booked passengers who do not check in for the flight.

Pattern recognition in sensor input

Neural networks are very valuable in the analysis of speech, images, and other signals. The images may be produced by cameras of various kinds, or radar, or sonar sets. Other signals may be produced by scientific apparatus such as spectrometers.

These capabilities allow neural networks to be used in a range of security applications. Among the systems that exploit neural networks' signal analysis capabilities is the SNOOPE airport security system. SNOOPE uses a neural network to analyse signals produced by bombarding passengers' luggage with neutrons in order to detect the presence of explosives. It has been tested at San Francisco and Los Angeles airports, and is now in operational use at Kennedy airport, New York. WISARD (see Figure 6.3) is being used for site security at certain high-security government sites in the United Kingdom.

Another existing application is the checking of blower motors for excessive noise. Siemens, the German electrical manufacturer, makes blower motors for incorporation into Ford car heaters. In the mid-1980s, Siemens found that a few of those produced were noisy. Noisy motors could be tested by ear, but this work rapidly became boring and workers could not perform to the necessary consistency. Siemens' Princeton research laboratory therefore developed a neural network that could identify noisy motors 90 per cent of the time, and this is now used to check every motor produced.

Even more extraordinary is a research project being conducted at Warwick University in the United Kingdom. This project is being sponsored by a major brewery, and its aim is to develop a 'robot nose' for monitoring the quality of beer production. One half of the project is associated with the sensing side — using a solid-state electronic device to detect trace elements. However, the problem is that no formulae exist to translate measured trace elements into 'taste' parameters. Many industries (from winemaking to perfumery) are totally reliant on human 'tasters', whose subjective judgements may vary over a long time period. Using a software simulation, Warwick has managed to train a neural network to recognise 16 different brands of beer from the detected trace elements.

Other systems exploiting a neural network's pattern-recognition capability include:

- A banknote sorter, based on WISARD, and used by the De La Rue Co to sort banknotes. (De La Rue is a United Kingdom-based international corporation that is a world leader in the printing of banknotes and travellers' cheques.)
- A signature-verification system that is 96 per cent accurate. It will ultimately allow a bank to check the signatures on all cheques, the cost of which would otherwise be prohibitive.
- Robot vision systems, such as ALVINN developed at Carnegie Mellon University, Pittsburgh. Once trained, ALVINN can drive an unmanned vehicle at up to 3.5 miles per hour.
- A safety system for railway level crossings, commissioned by British Rail. The neural network will be designed to recognise potentially dangerous situations such as stalled cars or pedestrians using the crossing when the barriers are down.

One of the most significant applications of neural networks is to speech recognition, which is described in Chapter 7.

Neural networks may also be able to find unsuspected patterns in databases. This type of application is, however, somewhat speculative. As we discussed above, neural networks can be trained on files of existing cases to predict particular outcomes — for example, which applicants for home loans are likely to default. They cannot usually 'explain' their reasoning by citing patterns in the original data, but once the neural network has established the existence of such patterns, statistical or other analyses may serve to identify them.

Advantages of neural networks to business applications

The strengths and weaknesses of neural networks and other approaches to applying computers to the solution of business problems are compared in Figure 6.6, overleaf. The main advantages of neural networks are that they can be applied to poorly understood problems, they can operate with incomplete data, and they can be cheap to build.

Application to poorly understood problems

The development of either a conventional or an expert system for decision-making usually requires one or more experts who know how to make good decisions. They then have to provide (or at least agree on) the rules for decision-making. In neural network development, it is not necessary to start with decision-making rules, but only with a set of cases with known outcomes (measured or assessed) and some sense of what data is relevant. Neural networks therefore work well in situations in which it is known, or suspected, that a pattern exists but where the pattern is not known. This situation exists in the analysis of images in factory, security, and military applications.

Rule-induction techniques for building expert systems, pioneered by Dr Ross Quinlan of the University of New South Wales and Professor Donald Michie of the Turing Institute in Glasgow, Scotland, also work in the absence of a proven expert. Induction has the advantage of producing explicit rules, but most automatic induction systems work well only when the variables are independent of one another. In many areas where problems have to be solved — economics, for instance — there is a high degree of interdependence between the variables. Neural networks can sometimes work well even where the variables are strongly interrelated.

Tolerance for incomplete data

A neural network can often produce a good result even if some data is missing. This tolerance occurs because the response of the

Figure 6.6 Each approach to the application of computers to the solution of business problems has its strengths and weaknesses

Approach	Strengths	Weaknesses
Traditional systems development	Gives exact answers Source of answers is clear	Lengthy Expensive
Decision-support system	Supports management analysis Often cheap to build	No result guaranteed High user costs
Expert system with human knowledge elicitation	Captures expert heuristics	Requires one or more experts Adaptation often troublesome
Expert system with rule induction	Expert does not have to specify the rules explicitly	Requires a set of classified cases Adaptation often troublesome
Neural network	No expert needed Robust Often cheap to build Adapts 'rules' during use	Requires large set of relevant, classified cases Rationale for results obscure Network needs careful design

network arises from the high connectivity between all the elements, rather than from specific rules or procedures. Other approaches do not have this tolerance as an inherent feature, and adding it will usually require considerable programming effort.

Low cost

Because neural networks are usually run as software simulations on personal computers, the hardware is inexpensive. Software is available for as little as $300, although more complex packages that combine neural networks with other programming techniques can cost as much as $50,000.

The development of a neural network application can be very rapid, because it does not require explicit rules to be identified and programmed. With suitable software, a network can be defined in minutes and trained in hours. Raytheon, a US developer and producer of sophisticated electronic systems and components, spent one year developing a sonar target-recognition system using expert-systems technology. It was 93 per cent accurate. A neural network was trained to recognise targets with greater accuracy in just one day.

Training a neural network does, however, require a large set of relevant, classified, cases. In the Raytheon example, this set already existed. If a large number of training examples are not readily available, and need to be researched or created, the cost of developing a neural network system may not be so favourable.

Problems with neural networks

There are several problems associated with the development and application of neural networks. The largest of these stems from the fact that neural networks are still an immature technology — there is still little formal methodology governing their design. The other main problems are that they have limited applicability, sometimes they do not converge on the right answers, they generally cannot explain the 'reasoning' behind an answer, and the data presented to a neural network may need to be carefully conditioned or preprocessed.

There is little formal methodology governing neural network design

Despite nearly 40 years of research, there are still no hard and fast rules for designing a neural network. There is no rule, for example, about how many neurons should be built into the middle or 'hidden' layer of a three-layer network to solve a particular problem. Yet this is critical: too few neurons and the network will not even learn the training examples; too many and the network will learn the training examples exactly but will be unable to apply what it has learned to new examples. As a result, most neural network design relies on trial and error, which can be very time-consuming.

They have limited applicability

Popular press reports have suggested that neural networks could render conventional computers obsolete within 10 years. This is nonsense. Almost all applications to date can really be described as pattern-recognition or classification problems. Only a small proportion of business computing falls into that category. Interestingly, though, quite a high proportion of managerial decisions can be thought of in this way — whether to hire a particular person, deciding on the right selling price for a product, forecasting if the exchange rate will go up or down, and so on. It will be many years, however, before neural networks can be applied to other types of problems, and there may never be any benefit in so doing.

Neural networks do not always give the right answers

Like most optimisation and planning systems, neural networks can also arrive at a local optimum rather than the true optimal solution. Usually, this problem can be overcome by altering the design or the starting conditions for a neural network, although, as stated above, there may not be any rules for doing this.

Neural networks cannot explain the reasons for their decisions

Neural network systems can produce clear diagnoses, identifications, or recommendations, and can even indicate the degree of uncertainty. However, they have not, so far, been able to explain the reasons for the results in the way that a good expert system can. Inspection of the state of the neural network by an expert can verify that the arithmetic and logical operations have been performed correctly, but cannot provide an explanation in terms intelligible to the user, except in the rare cases where neurons in the hidden layer come to represent specific properties of the data.

Recent research by Professor Stephen Gallant of Northeastern University, Massachusetts, has shown that a set of rules can be derived from neural networks of certain kinds. These rules could be used either to provide justification for the network's results, or to generate, automatically, an equivalent expert system. This research has been applied by Hecht-Nielsen Neurocomputer of San Diego, California, which expects to have commercial products based on it available in the first half of 1990.

The absence of any explicit justification for a result does not prevent effective use of neural networks in applications where it is only the results that count — the majority of business decisions, for example. However, it severely limits their application to critical applications involving finance, safety, or legal liability.

Data presented to a neural network may need to be carefully preconditioned

Experience has shown that the success of a neural network application can depend upon how the input data is coded. An excellent example of this concerns the much celebrated NetTalk system.

NetTalk, developed at John Hopkins University in Baltimore, Maryland, is a speech synthesiser that converts plain English text into speech. The problem of any text-to-speech system lies in the complex rules required to convert characters into one of the 46 phonemes, or basic sounds, used in English, since pronunciation of any character depends on its context. (The most sophisticated system developed at the Royal Signals and Radar Establishment in the United Kingdom, for example, has 20,000 rules. To operate in realtime, this would require phenomenal computing power.)

NetTalk uses a neural network of approximately 200 neurons. The text characters are represented in binary code and are presented in turn to the neural network inputs, which, in addition to looking at the binary code for a particular character, also looks at the three characters on either side. The output layer of the network features one neuron for each of the 46 possible phonemes.

The developers found that the network learned very quickly, just by giving it plenty of 'reading matter' accompanied by the correct pronunciation. However, they were not able to achieve better than 95 per cent accuracy, which is not good enough for most applications.

Researchers at Oxford University in the United Kingdom have now found that accuracy can be improved significantly by using several alternatives to ASCII coding.

Implications for systems departments

Neural networks are, as yet, a very immature technology, and it will be several years before they become an integral part of information technology. The majority of applications implemented to date are still only at a prototype or experimental stage. Nevertheless, it is unquestionably an emerging technology, and organisations that wish to remain at the forefront can and should now be evaluating its potential for real applications.

Rather than being applied to traditional systems applications, neural networks are best considered for applications that have previously been beyond the scope of information technology, or those where alternative methods (such as expert systems) have failed to produce the desired results. In particular, systems managers should look out for pattern-recognition, classification, and decision-support applications in areas where the organisation is totally reliant on regular human judgement or classification, especially where the expertise is having to be learned rather than taught. This can apply to quality assurance (go/no-go decisions), pricing policy, credit control, personnel evaluation, product design, and market research, to name but a few.

Having identified a potential application, it is tempting to start experimenting, given the wide availability of low-cost neural network simulation packages and appropriate hardware add-ons. Our experience is that this approach seldom achieves results, unless the developers have had some formal training in neural network design. The following guidelines should help to avoid the more common pitfalls:

— Use a specialised consultant both to identify applications and to specify the problem in neural network terms. This is generally quite a short stage, and once it has been completed, the actual implementation and testing falls well within the scope of a typical systems department. Indeed, the greater part of the task will be extracting the input data from existing systems, and presenting the data output from the neural network in the most useful format.

— Ensure that the inability of a neural network system to provide justifications for its results will be acceptable to its intended users and, where appropriate, to those affected by the results.

— Keep a small proportion, perhaps 10 per cent, of cases for human assessment, and retrain the system periodically on these cases. This is necessary because, like other systems, neural networks need maintenance. In most applications, the network's original training may become inappropriate.

— Pay particular attention to the user interface. As with expert systems in the past, the user interface of neural networks, and their integration with existing systems, will be critical issues. The development of interfaces for data acquisition and training should present few problems. Functionally integrating a neural network with, for instance, a transaction-processing system, is likely to prove considerably more difficult, and should be carefully considered when selecting neural network tools.

— Develop a pilot neural network system that addresses a real, but not business-critical, requirement. Success with the pilot is likely to lead to further applications, while failure should teach valuable lessons. Once the in-house skills have been built up, it may be possible to make the tools and techniques more widely available within the systems department.

Neural network technology is the least advanced of the technologies described in this report. Nevertheless, we believe that by the mid-1990s it will be in use for certain types of commercial computing applications. The guidelines set out above will help those Foundation members who wish to experiment with neural networks to do so in a controlled manner.

Chapter 7
Speech recognition

If we had written a Foundation Report about emerging technologies in the mid-1970s, it would have included a chapter on speech recognition, and we would have predicted then that the technology was nearly ready for wide-scale commercial use. Fifteen years later, the situation is not very different. Although there are now an increasing number of speech-recognition systems on the market, some of which have achieved modest success, decisive breakthroughs in performance have proved to be elusive, and speech-recognition technology is not yet in widespread commercial use. However, we remain convinced that speech recognition will become a significant technology for many organisations before 1995.

This optimism is based on recent advances in the use of low-cost signal-processing chips and the application of neural networks and sophisticated mathematical techniques to speech-recognition problems. In addition, as hardware costs continue to fall, vendors of speech-recognition systems should be able to reduce product prices to levels that will make them available to a mass market.

The main application of speech-recognition technology will be to provide users of word processing packages with the ability to dictate, rather than type, their text. The technology will also promote the increased use of computers for word processing, electronic mail, and other text-based applications, and will make all kinds of computer applications easier to use. There are also some interesting opportunities for synergy between speech recognition and other AI techniques and tools.

The nature of speech recognition

A speech-recognition system processes the sounds that are picked up by a microphone in order to identify the words that were spoken. The words are then processed or stored as if they had been typed. (Speech recognition should be distinguished from voice-recognition, or voiceprint, technology, which identifies the voice of the speaker rather than the words, and is used for different types of applications.) The central problem in speech recognition is the variability of the spoken word. People pronounce words differently in different contexts and on different occasions, and different people use different pronunciations. In addition, some sounds may correspond to several different words ('two' and 'too', for example). An added complexity is caused by the very large number of words in any natural language.

There is also a significant problem in deciding which group of sounds constitute a word, since people often run words together (by saying, for example, ''whatsthetime?'' rather than ''what is the time?''). It is this problem that makes the recognition of continuous speech much more difficult than the recognition of isolated words.

All speech-recognition systems may be characterised by five parameters:

- Whether the system is speaker-independent, which means that it must be able to recognise the words spoken by many people, rather than just one person whom it has been trained to recognise.
- Whether the speaker speaks normally or inserts distinct pauses between words ('discrete word recognition').
- The amount and nature of any background noise and distortion that can be tolerated (the sound quality).
- The size of the vocabulary that can be recognised.

— The percentage of words correctly recognised by a system (the recognition rate) in a particular application. Research shows that, for most purposes, the recognition rate must be at least 97 per cent. If a system recognises less than 97 per cent of the words, people will refuse to use it.

Different types of speech-recognition applications require different combinations of the first four parameters. Thus, as Figure 7.1 shows:

- A speech interface to a machine might require speaker-dependent recognition of up to 20 discrete words, spoken into a good-quality microphone, although possibly with background noise.
- A public telephone-information system, such as an interactive timetable, would probably require speaker-independent recognition of up to 50 discrete words, spoken over a telephone line.
- Audio input to a word processor would require speaker-dependent recognition of at least 5,000 words, presented as continuous speech spoken into a good microphone, but with limited background noise.
- An automatic shop assistant might require speaker-independent recognition of more than 5,000 words, spoken continuously in a moderately noisy environment. To be really useful, however, such a system would also need to understand what was being said.

One example of a successful speech-recognition application is provided by Jaguar, the UK-based manufacturer of luxury cars. Jaguar's factory-floor quality inspectors are provided with a speech-recognition facility. They receive instructions and information either through headphones or from display screens, and reply through a microphone, indicating whether the component inspected is acceptable. Jaguar has found that this system enables faults to be fixed more quickly and with less disruption than the previous paper-based system.

Other organisations have been less successful at implementing this type of speech-recognition application, because the background noise level is too high or the intended users are not willing to take the time to train the system to recognise their voices. Speech-recognition systems usually need to be trained to recognise the words that will be presented to them, and speaker-dependent systems need to be trained by the intended user.

The telephone-based information systems (often called audiotex systems) installed to date have met with limited success. However, organisations such as banks and telephone companies are actively developing new systems and some of these will undoubtedly be successful. Significant advances have also been made towards providing audio input to word processors. The fourth type of application shown in Figure 7.1, the automated shop assistant, is unlikely to become a reality within the timescale covered by this report, mainly because of the difficulty of understanding what is said.

Recent advances

Recent advances in speech recognition result from the availability of signal-processing chips

Figure 7.1 The characteristics of a speech-recognition system depend on the particular application

	System characteristics			
Application	Speaker-independence	Continuous speech	Sound quality	Vocabulary (words)
Machine-tool interface	No	No	Good	About 20
Timetable enquiries	Yes	No	Telephone	About 50
Audio word processor	No	Yes	Good	5,000
Automatic shop assistant	Yes	Yes	Fair	More than 5,000

and the use of neural networks and related techniques.

A digital signal processor such as AT&T's DSP32, TI's TMS320, and Motorola's DS56000 consists of a single chip that provides a fast analogue-to-digital convertor and specialised logic for floating-point arithmetic. A digital signal-processor chip processes analogue audio signals, extracting features such as the frequencies of the tones that make up the sound, ready for further processing. In recent years, digital signal processors have become available off-the-shelf at modest prices, and are being included as standard components in advanced workstations such as the NeXT Cube.

During the past two years, researchers have begun to apply neural networks and mathematical forms, known as hidden Markoff chains, to the problems of speech recognition. The great advantage of these techniques is that they do not require the researchers to start with detailed knowledge of the exact differences between phonemes (the basic building blocks of human speech), because neural networks acquire this knowledge as they are trained.

Perhaps the most successful of these systems to date is that developed by Professor Teuvo Kohonen at Helsinki University, which is described in Figure 7.2. Kohonen's system works in English, Finnish, and Japanese, and requires ten minutes to train for a new speaker. According to Ian Croall of the UK Atomic Energy Authority, who has used the system, it is "98 per cent accurate at dictation speed, or rather better than the average secretary". Unfortunately, this system is no longer available for public inspection because the industrial sponsor of one of Kohonen's students, Osahi Chemicals, has bought all the rights to this system. Osahi Chemicals says it will be offering products based on this technology by 1991.

Figure 7.2 The Kohonen speech-recognition system is probably the most successful to date

The Kohonen speech-recognition system operates in three stages. The sound signal is first converted to tones, the tones are then converted to phonemes, and the phonemes are finally converted to words.

The sound signal is processed by a digital-signal processor, the output of which is converted to tones by applying a Fourier transform (a mathematical technique that, approximately, describes the speech waveform as a set of pure tones). The output from the Fourier transform is passed to a self-organising neural network, implemented on two M68000 chips, which classifies the tones as phonemes. Although the system has been implemented only in Finnish, the phonemes are the same in English and Japanese. Finally, the phonemic transcription is passed to a second neural network, implemented on a PC/AT, which has been trained to convert sequences of phonemes into words.

Other researchers have, however, replicated parts of Kohonen's work. The key to success appears to be the starting conditions for the network, after which training achieves the final state. Kohonen has since produced another speech-recognition system that, he says, has even better performance.

Another interesting and significant speech-recognition system is that developed by Dr Fred Jelinek of IBM's Thomas G Watson Laboratory, Yorktown Heights, New York. This system is implemented as add-on boards for a PC/AT. In the first stage, the system recognises 'diphones' — that is, transitions between phonemes — in the input sounds. In the second stage, the system resolves ambiguities, such as that between 'to', 'too', and 'two', through the use of statistical tables.

This system is trained by speaking 100 sentences, a process that takes 20 to 25 minutes, after which the system takes four hours to construct its recognition model for a vocabulary of 20,000 words. It is claimed to have a recognition rate of more than 95 per cent, providing that the speaker leaves a pause of 0.2 seconds between words. This system was seen by delegates who attended the Industry Tour that followed the International Foundation Conference in Cannes in October 1989. It is a laboratory system, however, and IBM has made no commitment concerning any future product based on this system.

We believe that the application of neural networks, hidden Markoff chains, or other related techniques, will produce a usable, speaker-dependent, continuous speech-recognition system with a vocabulary of at least 5,000 words certainly by 1995, and probably by 1993. Such a system would need to be trained for each separate speaker, but this training would take minutes rather than hours.

Applications

The advances in speech-recognition technology described above will result in audio-input facilities being provided for word processors, electronic mail, and groupware systems. Used in these ways, speech recognition will have a significant impact on the user interface to computer systems in general.

The audio word processor

Speech-recognition facilities will first be provided on office workstations in the form of add-on boards accessed via special software. We expect that, once the technical feasibility and commercial viability of this approach are established, major suppliers of word processing software will add support for these boards to their products. The suppliers of user-interface managers will then move to provide support for speech input, including correction functions, in their products.

Most business and professional documents require a vocabulary of only a few thousand words, excluding proper names. Speech-recognition technology will be able to provide an acceptable level of performance for such a vocabulary, assuming that the speaker speaks as slowly and clearly as he or she would into a dictaphone. Since an intelligent personal workstation is used by only one person at a time, an audio-input system for word processing needs to be able to recognise only one user at a time.

An audio-input system will inevitably not be able to recognise every word that is spoken, and system designers will need to give careful thought to ways of detecting and correcting mistakes. Possible means of detecting mistakes include:

- Programming the recognition system to question the user when its level of confidence in having correctly identified a word falls below a certain threshold.
- Using syntactical and grammatical analysis to determine when a string of words does not make sense.
- Allowing the user to review the text generated from his or her spoken inputs.

In some cases, it may be possible to correct the misrecognised word automatically. Usually, however, it will be necessary for the user to repeat, spell, or key in the word in question. The initial implementations of audio input to word processors will require their users to decide between alternative words and spellings, to enter proper names, and to correct misspelt words using a keyboard and/or pointing device.

The corrections may be made either by interrupting the dictation at the point where the system is unable to recognise a word, or by reviewing the whole of the generated text and correcting any words that have been flagged. The optimum mix of techniques for correcting text generated by speech-recognition systems, and the design of the human interface to such systems, are matters that require further research and development work. An audio word processor will, for example, also need to provide features similar to those available with an ordinary word processor — cut-and-paste, page formatting, control of font and type size, and so on.

The widespread use of word processing packages on PCs has shown that many managers and professionals are prepared to type their own documents if they are provided with suitable facilities. Few of them are, however, competent touch typists, and because of this, many of them enter their shorter documents themselves but ask a typist or secretary to key in the longer ones. Given a choice, most of them would prefer to dictate their larger documents to an audio word processor, rather than key them in themselves or have someone else key them in on their behalf.

These managers and professionals are major originators of business documents, and they comprise an important market for the makers of audio word processors. The audio word processor will also expand the use of word processing to people who are at present unwilling to type.

Speech recognition as part of the user interface

The increasing use of hypermedia systems will make people more used to combining speech input with other means of interacting with a computer. The technology already exists to allow users to speak their commands to operating systems and software packages, but its use is largely restricted to circumstances in which either the user cannot use a keyboard because his hands are being used for another purpose, such as holding a measuring instrument, or the interaction must take place over a telephone connection.

In many circumstances, however, speech-recognition facilities are not a particularly attractive option for existing users of PCs and office systems because:

— PC users do not see the need for speech-input facilities, believing that the existing (and familiar) keyboard-input functions are perfectly adequate.
— A good interface requires the integration of speech and keyboard functions, and software suppliers have generally not provided this.
— In some cases, using speech-input functions would be slower than keyboarding.
— The technology is expensive.
— Some people would be embarrassed to be seen talking to a machine.

We therefore conclude that today's speech-recognition technology will not replace the use of keyboard and mouse. However, there seems no reason why a speech-recognition subsystem that has been installed to facilitate audio input to a word processor should not also be used for input to a database, spreadsheet, or other office system.

Despite the drawbacks listed above, we believe that speech recognition will increasingly be used as part of the user interface. In particular, interfaces that combine windows, pointing, and speech recognition will increase the range of people who use computers, because such interfaces will overcome the 'keyboard phobia' that some people still have. (Most of the reluctance to use keyboards is, however, due to a realisation that the computer system on offer provides little benefit to the person concerned.)

The most significant group whose use of speech recognition will increase is senior managers. Some senior managers in large organisations have already shown that they are willing to use computer systems to retrieve information that is important to their work. Indeed, specially designed executive information systems are an important growth area. However, users of executive information systems are usually unwilling to input much of the information themselves, although some of them have become extensive users of electronic mail systems. By making the entry of information, especially of text, easy, speech recognition will increase the number of senior managers who use IT systems, especially electronic mail and groupware systems.

Barriers to the use of speech recognition in the office

The use of speech-recognition systems has been restricted in the past by many factors including their small vocabularies, the need to train systems to recognise individual users, the high cost of the technology, the unsuitability of office acoustics, and the possible embarrassment of prospective users.

Advances in speech-recognition technology will increase the vocabularies of available systems and reduce the training time (because they will be trained on a set of phonemes rather than on every word in the vocabulary). Technical advances will not directly reduce the cost of speech-recognition technology, but as a volume market for speech-recognition systems develops, prices will fall.

In many offices, and especially open-plan offices, little thought has been given to the acoustic environment. As a result, some staff find even the use of telephones and keyboards by other staff intrusive and disruptive. In such office environments, speech input to, and speech output from, computers may be impractical. It is often possible to improve office acoustics by installing sound-absorbing furniture and wallcoverings, or by injecting a low level of 'white noise'. The effective exploitation of speech-recognition technology may require the use of these techniques in many offices.

Future developments

In the past, some researchers have attempted to implement speech-recognition systems by using AI-derived techniques such as natural-language processing, and even expert systems, to compensate for the limitations of the basic speech-recognition technology. The application of neural network and related techniques seems likely to remove the need to enhance the basic recognition technology, although AI-derived techniques will certainly be necessary if the system is to 'understand' enough of the words spoken to take appropriate action. There are already systems that can automatically analyse the contents of texts that are restricted to a small subset of natural language (banking telexes, for example), and database systems that allow queries to be expressed in a wide variety of syntactic and verbal forms have been available for some years.

Once speech-recognition technology has reached a sufficiently high level of performance, there is clearly scope for using it as a front-end to a natural-language processing system. Such a system could provide a senior manager with a powerful and easy-to-use interface to corporate databases, or enable a travelling representative to retrieve the information without having to use a computer terminal.

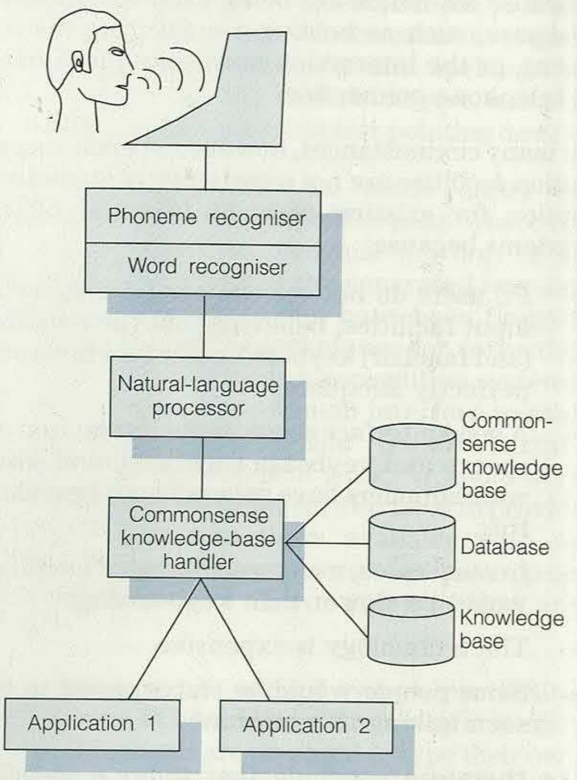
The combination of speech recognition with other AI techniques could be taken a step further to provide systems that appear to have intelligence. In Chapter 5, we said that commonsense knowledge bases, like the one being developed by Project CYC at the Microelectronics and Computing Corporation, may be available by about 1994. The goal of Project CYC is to develop a comprehensive commonsense knowledge base that can be used to understand the contents of texts such as newspaper articles. Such a knowledge base would also be able to understand the meaning of words spoken to it, and to generate appropriate responses and actions based on that understanding.

A system with the architecture shown in Figure 7.3 would seem to have all the elements necessary to pass the Turing Test (which was described in Chapter 5), when it is used as a reference source for information included in the knowledge bases and databases to which it had access. We do not, however, expect such systems to be available by 1995.

Implications for systems departments

Speech-recognition technology can be used today for applications that require only small vocabularies and that operate in reasonably favourable acoustic environments. Commercial products are available from vendors such as Articulate Systems (based in Switzerland), Kurzweil (of Massachusetts), and Speech Systems (of California); their use should be considered for applications where screen and keyboard access is, for some reason, impossible or undesirable.

Figure 7.3 Combining speech recognition with a commonsense knowledge base will produce a very sophisticated 'intelligent' system


Phoneme recogniser
Word recogniser
Natural-language processor
Commonsense knowledge base
Commonsense knowledge-base handler
Database
Knowledge base
Application 1
Application 2

There is little that user organisations can do to bring forward the availability of better speech-recognition technology and, in our view, not much that they can do to prepare for using it. In view of the uncertainties, we recommend only that members should track the development of the new speech-recognition technology and products.

In planning new buildings and office refurbishments, systems managers should, however, ensure that future acoustic requirements are fully considered. This, and many other aspects of the impact of IT on buildings, are discussed in *Information Technology and Buildings*, published by Butler Cox in 1989 following a major study conducted on behalf of four leading IT vendors.

Appendix

Other technologies not investigated in detail

In this appendix, we briefly describe some of the emerging technologies that did not meet our criteria for inclusion in this report, and that were therefore not studied in detail.

Data compression

Data-compression technology is developing rapidly as powerful but inexpensive single-chip computers become available to run the powerful compression algorithms that have been developed over the last 15 years. The approaches to compression range from those that work for all kinds of data to those that are particularly appropriate for a particular kind, such as digitised speech. Data compression can be used to reduce both the bandwidth required on communications lines and the amount of storage required on discs, tapes, and other media. Compression is particularly important in voice and video applications where large amounts of data need to be stored or transmitted.

The most basic approach to compression replaces frequently occurring patterns in a stream of bits or bytes with symbols representing the patterns. This form of data compression is used in modern facsimile transceivers and can reduce the amount of information that has to be transmitted by up to 90 per cent. Usually, the data reconstructed from the compressed form is exactly the same as the original data.

A second approach, which is appropriate to the digital representations of continuous analogue signals, such as sounds and images, is based on transmitting changes rather than complete data. This approach may be applied to the signals produced by scanning consecutive lines in preparation for facsimile transmission or to the changes between two successive frames of a video transmission. It is also used in some telephony systems, in digital videoconferencing systems, and in compressing moving images for storage on optical discs. In some situations, the data can be reduced by up to 98 per cent. However, there may be some distortion in the reconstructed data where there is a very high rate of change in the data, such as rapid movement in a television picture.

The most sophisticated approach to data compression identifies the phoneme represented by a sound, or the character or geometric shape represented by an image, and transmits a description of the phoneme, character, or shape. This approach requires even more processing power than the one based on transmitting changes, but it can reduce that volume of data by up to 99.8 per cent. It is used in Encapsulated PostScript to describe complex line art. However, some loss of detail is inevitable when the sound or image is reconstructed.

The most significant consequence of the more advanced kinds of data compression is the ability to increase the playing time of moving images stored in digital form on an optical disc.

Fibre Distributed Data Interface (FDDI)

FDDI is an emerging standard for optical-fibre local area networks operating at 100M bit/s. Products are already available from a few specialist suppliers and are under development in the research and development departments of many suppliers. IBM and Digital have both indicated that FDDI technology has a place in their strategic plans. However, the products available today are much more expensive than Ethernet and Token Ring products.

For at least the next five years, we expect FDDI networks to be used mainly as backbone networks in large buildings and on campuses. They will also be used to connect very-high-performance workstations to each other and to supercomputers, a requirement that will become more common as parallel computers and hypermedia systems come into more general use.

Image processing

The term 'image processing' covers two distinct technologies. One is the capture and management of document images (usually business documents), and is sometimes known as document-image processing. Report 70 (*Electronic Document Management*) described how document-image processing provides major opportunities for cost savings and service improvements when combined with other technologies to form an electronic document management system.

The second type of image-processing technology is concerned with the processing of an image to determine what it shows. Examples include recognising an intruder by a security system, identifying characters in a document image, and recognising a particular kind of aircraft in a radar image. This type of image processing can also be used to analyse photographs and the images produced by various scientific instruments, and can be applied in factory automation, security systems, and military surveillance. The technology is based on advances in parallel computers and neural networks and is increasingly being used in conjunction with document-image processing technology. However, with the possible exception of security systems, these types of applications are probably relevant to only a minority of Foundation members, so we have not examined this technology in detail.

Local area wireless networks

The costs and problems associated with wiring buildings for IT systems have encouraged several suppliers to develop wireless systems based either on radio links, typically in the 900 MHz range, or on infra-red transmission. Most of the wireless systems currently available provide data transmission at rates up to 9600 bit/s and have acceptable error rates over distances up to 30 metres. There are also some more powerful systems able to transmit over longer distances.

A well known example of this type of technology is the cordless telephones that are widely available for domestic use. However, the small number of frequencies available for this purpose in most countries has made them unsuitable for offices. The introduction of second generation cordless (CT2) systems late in 1989 heralded a significant change. Several suppliers are developing PABXs and key systems based on CT2 technology, and we expect these products to become an attractive alternative to conventional wiring for telephony. Nonetheless, the business implications of this will probably not be very great, because almost every established office is already wired for telephony.

Another disadvantage of office communication systems based on CT2 technology is that the data transmission rates are likely to fall far short of those required by today's personal computers, and to be totally inadequate for those that will be needed by future hypermedia workstations. In view of the small amount of data transmission that is handled by PABXs today, we do not expect local wireless data networks to be a significant technology for Foundation members by 1995.

Metropolitan area networks

Metropolitan area networks operate at the speeds typical of local area networks, but over much wider areas, typically a city. They are expected to operate at between 35M and 150M bit/s and will be able to carry digital communications of all kinds, although their use for full-motion video may not often be economical.

The concept of the metropolitan area network was formulated several years ago, but significant progress was not made until the relevant IEEE committee aligned its standard (802.6) with emerging standards for broadband ISDN. This attracted the support of the regional Bell telephone operating companies, and it now seems likely that metropolitan area network products and services will be available in a very few years. Services in some countries will be offered by telephone and cable operators.

At present, however, it is far from clear what metropolitan area networks will be used for:

- Initial experience with ISDN services has been concentrated on conventional voice and data communications.

- In general, businesses do not see much need for the high data transmission rates that metropolitan area networks can provide.
- The services are likely to be too expensive for domestic subscribers for many years.

In view of the uncertainties about the timescales and implications of metropolitan area networks, we did not study them in detail for this report.

Object-oriented design, programming, and data management

Object-oriented design, programming, and data management are all based on the concept that information systems should model the world in which they exist, and that the world is composed of objects, which have certain things in common.

Object-oriented design may be seen as an extension of the entity-relationship approach to data modelling that underlies most of today's structured techniques and CASE tools. Object-oriented programming can be thought of as an approach to programming in which:

- Data and program code to process the data are 'packaged' as an 'object'. Each object is separately designed, coded, and maintained.
- Objects inter-react with each other by transferring messages, not by inspecting each other's data and program structures.
- New objects may be created from existing ones in a controlled manner, so that they retain (or inherit) the properties of the existing objects.

Object-oriented data management is the natural complement to object-oriented programming. However, it may also be seen as an extension to the relational database approach that enables data structures that cannot be 'normalised' (such as those found in office systems, CAD, and multimedia systems) to be managed. In Report 64, *Managing the Evolution of Corporate Databases*, we stated that object-oriented databases will be important in the future both for structured databases and for knowledge bases. Hypertext and hypermedia systems (discussed in Chapter 4) use a restricted form of object-oriented data management.

In Report 74 (*The Future of System Development Tools*), we shall discuss object-oriented design, programming, and data management further.

User interfaces

The recent introduction by many suppliers of graphical user interfaces, and their endorsement by IBM, has focused attention on user-interface technology. Most computer vendors are now developing toolkits that will allow graphical user interfaces to be built into application systems.

Recent advances in the nature of the user interface include the introduction of animation, greater use of sound, and hypertext, all discussed in Chapter 4. In the next five years, much greater use will be made of these techniques and of voice input. (Developments in speech recognition are discussed in Chapter 7.)

One disadvantage of almost all current user interfaces is that they provide every user with the same interface, even though different users may have different needs at different times. Unfortunately, recent research indicates that the achievement of a highly adaptable user interface requires considerable work for each application, or class of application, rather than the use of a few specific techniques. No breakthrough seems imminent in the development of adaptable 'customised' user interfaces.

VSAT technology

Very small aperture (satellite) terminals (VSATs) communicate with satellites in geosynchronous orbit using dishes that are only 1.2 or 1.8 metres in diameter. A typical VSAT network provides two-way data communications between a head-office site, which is equipped with a large satellite dish, and a large number of widely dispersed locations, each of which is equipped with a small dish. The market for VSAT networks is growing rapidly in north America, where large distances, a liberal regulatory regime, and the existence of many organisations operating on a continental scale, can often make VSATs cheaper than terrestrial communications facilities.

These conditions seldom exist outside North America, so VSAT networks are less likely to be feasible elsewhere. VSAT technology was therefore not investigated in detail in our research.

BUTLER COX
FOUNDATION

The Butler Cox Foundation

The Butler Cox Foundation is a service for senior managers responsible for information management in major enterprises. It provides insight and guidance to help them to manage information systems and technology more effectively for the benefit of their organisations.

The Foundation carries out a programme of syndicated research that focuses on the business implications of information systems, and on the management of the information systems function, rather than on the technology itself. It distributes a range of publications to its members that includes Research Reports, Management Summaries, Directors' Briefings, and Position Papers. It also arranges events at which members can meet and exchange views, such as conferences, management briefings, research reviews, study tours, and specialist forums.

Membership of the Foundation

The Foundation is the world's leading programme of its type. The majority of subscribers are large organisations seeking to exploit to the full the most recent developments in information technology. The membership is international, with more than 400 organisations from over 20 countries, drawn from all sectors of commerce, industry, and government. This gives the Foundation a unique capability to identify and communicate 'best practice' between industry sectors, between countries, and between IT suppliers and users.

Benefits of membership

The list of members establishes the Foundation as the largest and most prestigious 'club' for systems managers anywhere in the world. Members have commented on the following benefits:

— The publications are terse, thought-provoking, informative, and easy to read. They deliver a lot of message in a minimum of precious reading time.

— The events combine access to the world's leading thinkers and practitioners with the opportunity to meet and exchange views with professional counterparts from different industries and countries.

— The Foundation represents a network of systems practitioners, with the power to connect individuals with common concerns.

Combined with the manager's own creativity and business knowledge, Foundation membership contributes to managerial success.

Recent Research Reports

56 The Impact of Information Technology on Corporate Organisation Structure
57 Using System Development Methods
58 Senior Management IT Education
59 Electronic Data Interchange
60 Expert Systems in Business
61 Competitive-Edge Applications: Myths and Reality
62 Communications Infrastructure for Buildings
63 The Future of the Personal Workstation
64 Managing the Evolution of Corporate Databases
65 Network Management
66 Marketing the Systems Department
67 Computer-Aided Software Engineering (CASE)
68 Mobile Communications
69 Software Strategy
70 Electronic Document Management
71 Staffing the Systems Function
72 Managing Multivendor Environments
73 Emerging Technologies: Annual Review for Managers

Recent Position Papers and Directors' Briefings

Information Technology and Realpolitik
The Changing Information Industry: An Investment Banker's View
A Progress Report on New Technologies
Hypertext
1992: An Avoidable Crisis
Managing Information Systems in a Decentralised Business
Pan-European Communications: Threats and Opportunities

Forthcoming Research Reports

The Future of System Development Tools
Assessing the Value from IT
Systems Security
New Telecommunications Services
Using IT to Transform the Business

Butler Cox

The Butler Cox Foundation is one of the services provided by the Butler Cox group. Butler Cox is an independent management consultancy and research company. It specialises in the application of information technology in industry, commerce, and government throughout Europe and the rest of the world. It offers a wide range of services to both users and suppliers of information technology. For more information about our services, please contact your nearest office at the address shown on the back cover.

Butler Cox plc
Butler Cox House, 12 Bloomsbury Square,
London WC1A 2LL, England
☎ (01) 831 0101, Telex 8813717 BUTCOX G
Fax (01) 831 6250

Belgium and the Netherlands
Butler Cox BV
Burg Hogguerstraat 791,
1064 EB Amsterdam, The Netherlands
☎ (020) 139955, Fax (020) 131157

France
Butler Cox SARL
Tour Akzo, 164 Rue Ambroise Croizat,
93204 St Denis-Cédex 1, France
☎ (1) 48.20.61.64, Télécopieur (1) 48.20.72.58

Germany (FR), Austria, and Switzerland
Butler Cox GmbH
Richard-Wagner-Str. 13, 8000 München 2, West Germany
☎ (089) 5 23 40 01, Fax (089) 5 23 35 15

Australia and New Zealand
Mr J Cooper
Butler Cox Foundation
Level 10, 70 Pitt Street, Sydney, NSW 2000, Australia
☎ (02) 223 6922, Fax (02) 223 6997

Finland
TT-Innovation Oy
Meritullinkatu 33, SF-00170 Helsinki, Finland
☎ (90) 135 1533, Fax (90) 135 2985

Ireland
SD Consulting
72 Merrion Square, Dublin 2, Ireland
☎ (01) 766088/762501, Telex 31077 EI,
Fax (01) 767945

Italy
RSO Futura Srl
Via Leopardi 1, 20123 Milano, Italy
☎ (02) 720 00 583, Fax (02) 806 800

Scandinavia
Butler Cox Foundation Scandinavia AB
Jungfrudansen 21, Box 4040, 171 04 Solna, Sweden
☎ (08) 730 03 00, Fax (08) 730 15 67

Spain and Portugal
T Network SA
Núñez Morgado 3-6°b, 28036 Madrid, Spain
☎ (91) 733 9866, Fax (91) 733 9910