A Progress Report on New Technologies

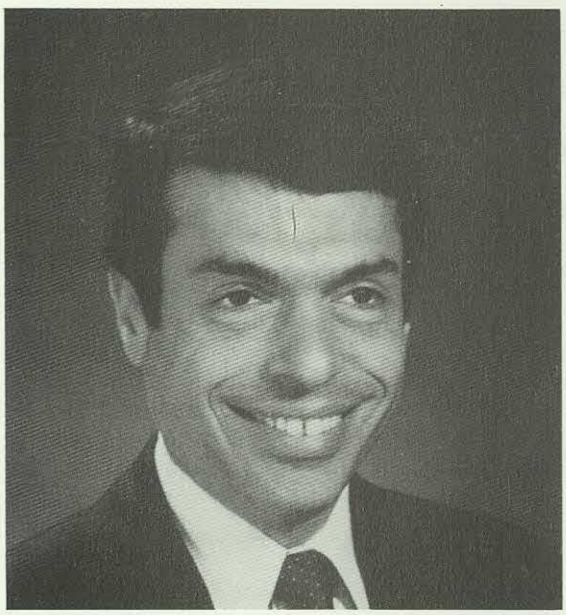
BUTLER COX FOUNDATION

A Paper by Craig Fields
March 1988

BUTLER COX FOUNDATION

A Progress Report on New Technologies

A Paper by Craig Fields
March 1988



Dr Fields is Deputy Director for Research responsible for the direction and management of basic research (including computer science projects) at DARPA, the US Defense Advanced Research Projects Agency. The agency has an annual budget of $850 million, approximately a quarter of which is spent on advanced computer science projects. Dr Fields' responsibilities have included the management of ARPANET, of the Data Computer project, and of programmes in biocybernetics, very large databases, man-machine relations, and image-based systems.

In October 1987, he addressed the International Conference of the Butler Cox Foundation in Munich. His presentation identified five areas in which there are imminent technological discontinuities: multiprocessors, microelectronics packaging, semiconductor manufacturing, superconductors, and lightweight satellites. Developments in each of these areas are allowing new types of computing and communications devices to be produced and are fundamentally changing the cost-performance ratios of the devices. All of these developments are occurring in hardware; Dr Fields does not foresee any discontinuities in software or software development techniques, although coupling the new hardware technologies with AI techniques (particularly expert systems) will have a dramatic impact on the uses to which computers will be put.

His presentation is reproduced in full in this paper.

Published by Butler Cox & Partners Limited
Butler Cox House
12 Bloomsbury Square
London WC1A 2LL
England

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

BUTLER COX FOUNDATION

A Progress Report on New Technologies

A Paper by Craig Fields
March 1988

Contents

BUTLER COX FOUNDATION

A Progress Report on New Technologies

My aim today is to tell you about some important emerging technologies. Some of them will be available in the next two or three years; some are here now, but may not be known to you even though they have been used for the last year or two by hundreds of US companies. I intend to make you slightly uncomfortable because I would like you to think at the end of this session that you actually have to pay attention to these technologies, that perhaps you should do something about them. In many cases, you can take some action today. In fact, not to take action could well be irresponsible.

The focus of my talk is technology, not applications. In particular, I want to highlight discontinuities in technology, places where there is a very large effect in a relatively short period of time. I believe you should be aware of these technologies and should think about how you and your business could make use of them. There are five technologies that I want to bring to your attention:

- The new generation of multiprocessor computers, most of which are supercomputers.
- New work in microelectronics packaging that will make computer systems much smaller than they are today.
- New semiconductor manufacturing techniques that not only will continue to improve the price-performance ratio, but will also provide new opportunities for rapid prototyping of microelectronics systems.
- Developments in superconductors.
- A new generation of lightweight satellites that we are in the process of building.

Note that all of these discontinuities are in the area of hardware, not software. It is not that I do not want to talk about software — I just cannot find any technological discontinuities in the software area. Of course, software technology is improving, but not in any abrupt or startling manner. However, at the end of my session I will say a few words about our experiences with expert system applications.

AREAS WHERE THERE WILL NOT BE A TECHNOLOGICAL DISCONTINUITY

Before discussing the discontinuities I would first like to tell you about a few important areas where I do *not* see any discontinuities in the technology — areas where there is a gradual improvement rather than a major change in a short period of time.

The first area is that of artificial intelligence. It is an area of great excitement for DARPA and for many companies in the United States, and for many of you, but it is not an area where abrupt changes in the technology can be expected. AI technology is steadily getting better year by year; there is no discontinuity. (I will return to the subject of artificial intelligence at the end of my talk.)

The second area is that of software production technology. Although software development tools are getting better and better, I cannot see any discontinuities occurring. If you are hoping for, or expecting, a five-fold or ten-fold or a hundred-fold improvement in software development, the only way you can get it is to hire smarter programmers. There simply are no tools on the horizon that I know of which will give you improvements of those magnitudes.

We can also foresee no discontinuities in the area of data storage. For the last few years DARPA has been searching for new techniques and new technologies to improve data storage and we simply cannot find anything that is significantly better than the best that is being developed.

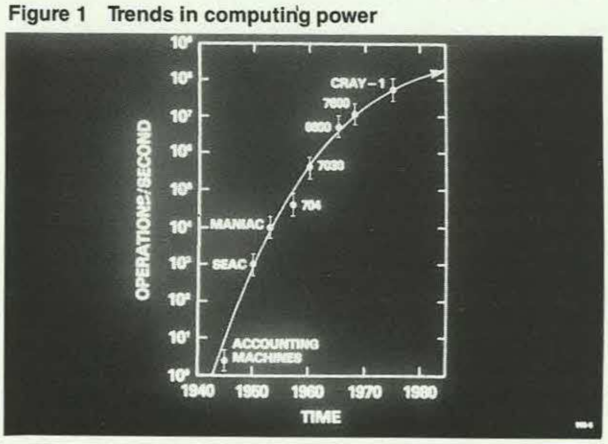
Next on the list of continuities rather than discontinuities is the area of computer security and computer privacy. Again, although security and privacy techniques are improving steadily, we cannot foresee any discontinuities.

Lastly, there is the area of computer networking and computer communications. I am almost embarrassed to have to report that this is an area of continuous improvement rather than discontinuous improvement because I represent the organisation that invented ARPANET and packet switching. But the fact is that there have not been any startling developments since then and we do not see any on the immediate horizon.

DEVELOPMENTS IN MULTIPROCESSOR COMPUTERS

I am sure you are all very familiar with the curve shown in Figure 1, which depicts the enormous increases in computing power since the end of World War II. We have witnessed a 10 million-fold increase in speed during the last 40 years or so, but it has been achieved essentially in one way — to build new computers out of faster and faster microelectronic components, using new materials, or by making them smaller and smaller. We believe that this trend will continue and can provide further improvements by a factor of 50 or even 100. But for many applications, improvements of that magnitude are just not good enough. I can quote many examples where computers 100-times faster than today's fastest would not be nearly fast enough — for example, many design applications, image analysis, and modelling simulations require improvements in computing power of a thousand-fold, ten thousand, a hundred thousand, a million. And many artificial intelligence applications (but not all of them) will require huge increases in speed over what is available today.

Figure 1 Trends in computing power



It is getting harder and harder to increase computing speed by using faster components. Experience at the Los Alamos National Laboratory shows that in the early 1950s it took about a year and a half to double performance; in the early 1960s it took just over two years; by the early 1970s it took nearly three years; and by the early 1980s it was taking four-and-a-half years to double performance. Furthermore, some researchers believe that silicon is now within a factor of five of the maximum achievable limits.

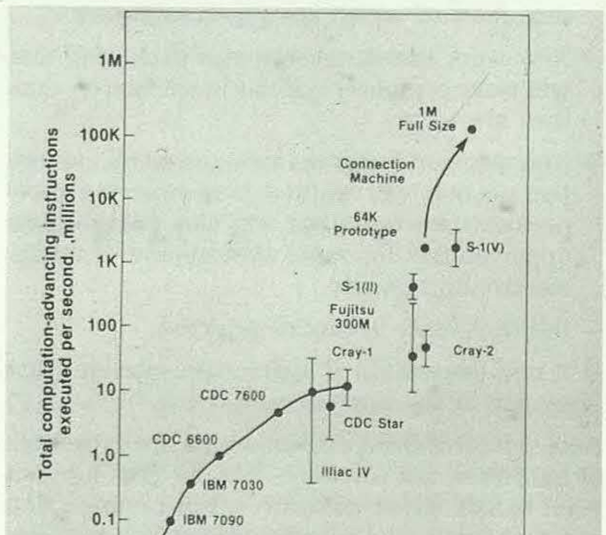
New ways of building ever-faster computers will therefore need to be devised. Much research effort today is focused on building faster computers by combining lots of small slow computers. In other words, to combine large quantities of microprocessors to produce a computing device that performs as a single, fast, and cost-effective computer.

In a state-of-the-art supercomputer today there are many chips each containing a piece of silicon measuring 1 cm x 1 cm or smaller. In total, there is about a square metre of silicon. The question is whether, at the same cost, that square metre of silicon can be used in a different way to provide much higher computing speeds.

The peak performance of today's supercomputers is about 500 megaflops (or half a gigaflop), although we are dealing typically with applications that require 10 megaflops, 100 megaflops, and so on. Today's computing devices range from supercomputers costing tens of millions of dollars to personal computers costing a few thousand dollars and providing the equivalent of 1 mips or less of computing power. I ask you to bear these figures in mind, to give you a reference scale for some of the figures I will present later.

The curves in Figure 2 illustrate the discontinuities I am talking about. The curve in the bottom left corner is equivalent to the curve in Figure 1. Above that are the new multiprocessor computers, ranging from a 64,000 processor prototype to a one-million processor machine we are currently building. In terms of performance, these computers represent a discontinuity. Note that the vertical scale is logarithmic, so I am talking about computers that are ten times or more faster than a Cray.

Figure 2 History and forecast of best available arithmetic performance levels



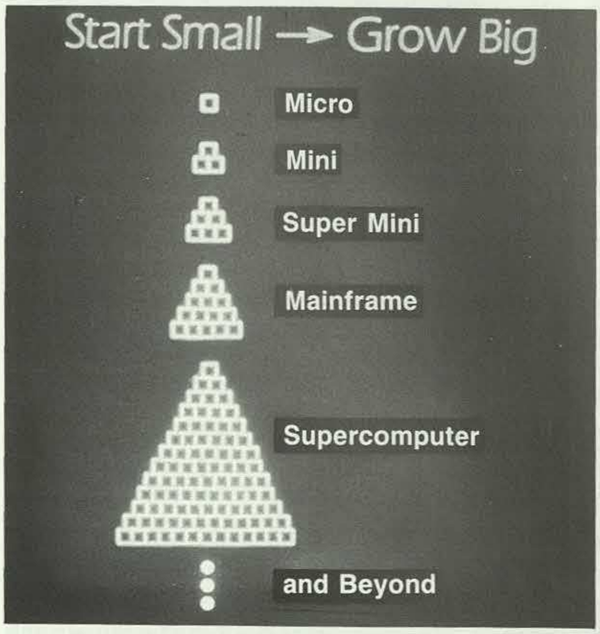
Somewhere between 200 and 300 multiprocessor machines have been sold in the United States and they are being used for a variety of applications. For example, a small company called Deltagraphics (which, as its name suggests, is in the graphics business) has used multiprocessing techniques to produce a graphics display processor that sells for one-hundredth the cost of its competitors.

DARPA has used multiprocessor computers for artificial intelligence applications. One measure of processing speed for an expert system is 'rips' (rule-based instructions per second). We built a very large expert system for managing the flight of aircraft. The system ran at just one rule per second on a conventional computer. Although this was good enough to do some jobs, it is pretty slow. We transferred the application to one of the new multiprocessor machines and it ran at two million rules per second. Thus, we are not talking about performance improvements of a few per cent. We are talking about very large improvements indeed.

The principle of using several processors to obtain faster speed is shown in Figure 3. You can start small with one microprocessor; by adding a few more you get the speed of a minicomputer; add a few more and you get the speed of a superminicomputer, and then a mainframe; by adding more you can produce the equivalent of supercomputer — equivalent to a Cray. The most exciting thing of all, however, is that there does not seem to be any apparent limit — you can keep adding microprocessors to obtain ever-faster machines.

Today, an Apple Macintosh is about eight times more cost effective than a Cray, measured in terms of dollars per mip. If somehow you could simply add together the processing power of 840 Macintoshes you would have a machine with the performance of a Cray but at a fraction of the cost.

Figure 3 Multiprocessors can be added to obtain faster processing speeds



In a multiprocessor machine, each processor does part of the problem, so the problem has to be divided up into pieces, and the pieces assigned to the processors, with the different aspects of the computation being synchronised. People frequently ask how many problems are amenable to being speeded up in this way. There are obvious applications in areas like fluid dynamics simulations, vision systems, weather forecasting, astrophysics, and chemistry. However, we have looked at hundreds, perhaps thousands, of computation examples and to date we have found only one or two that cannot be speeded up in this kind of way. (The exceptions were some very esoteric calculations in number theory.) Sometimes, calculations could be speeded up by only a factor of 25, sometimes it was a factor of a million, but, with very few exceptions, every calculation you look at can be divided up so it can be performed on a multiprocessor machine. We also found that the new machinery was good for the type of symbolic calculations required for AI applications.

You may well be wondering how easy it is to program multiprocessor machines — most programmers have enough difficulty dealing with one processor, let alone hundreds or thousands. Originally, we thought that it would be impossible to program them without AI-based tools that would automatically divide up the programs, assign them to processors, and synchronise the processors. No such tools exist, but it turns out that they are not necessary. Instead, we use a set of editing tools that aid the programmer in dividing up the program. We find that, most of the time, on most of these machines, many programmers have little difficulty in writing computer programs that work. I have programmed several of the machines; most of you could do so as well — it is not that difficult. You might be surprised to hear me say that — but just remember how often within your existing computer programs the system does the same thing over and over again. Many of those repetitive calculations could just as well happen in parallel. It turns out to be very easy to divide them up so this can happen.

We were also concerned with scaling issues — what the effect on performance would be if you doubled the number of processors. It turns out that if you buy a small machine with, say, 128 processors, and you find that it is not fast enough for one of your problems, you can buy another 128 processors and connect them up. You will find that the same software will run (about twice as fast) on the extended machine. Thus, for about twice the price you get twice the performance. This means that it is very easy to begin to use this new technology because it is not necessary to buy a huge computer to begin with. You can buy a small one to find out if multiprocessing technology is useful for you, and then scale up later.

Most of my interest in the new technology is focused on very large, very fast computers. For compatibility, however, you also need smaller and slower machines, so we are now also building smaller, slower multiprocessors that are 'only' as fast as a Cray, not faster. This means that an individual (an engineer, for example) can have the equivalent of a Cray on their desk.

LIMITATIONS OF MULTIPROCESSOR COMPUTERS

In fairness, I should also tell you what these multiprocessor computers are not suitable for. Their limitations fall into four areas:

1. They are not very good for performing extremely simple calculations with extremely large databases. With this type of application, most of the time is taken up with accessing and transferring data — the processor does hardly anything. Multiprocessor computers will not help to speed up this type of application.
2. I have already mentioned that there are a few types of esoteric calculations where the instructions have to be performed sequentially. I doubt if many of you have such requirements.
3. You cannot take your existing programs written in Cobol, Lisp, Fortran, and so on and run them on the new machines without any changes and expect to get huge improvements in performance. For that to happen it is necessary to make modest changes to the software. If it is not possible for you to do this then you cannot use the new technology without writing new code.
4. Multiprocessor computers will not help you to write software if you do not know how to program the application in the first place. I have had people approach me after hearing about the new technology who believe that it could be used to predict, for example, when the government of a 'banana republic' was going to fall and when a new government would take over. No one knows how to write a program to do that on any computer at any speed. All that multiprocessor technology can do is increase the speed and decrease the cost.

MULTIPROCESSOR COMPUTERS ARE COMMERCIALLY AVAILABLE

Despite the above limitations, there are many applications for multiprocessor technology. I shall now describe some applications that are based on commercially available computers that have been sold in some quantity. The first is based on a computer called the Connection Machine. The model we use is a 5-foot cube and costs about $2 million, or perhaps a little more, depending on the discount you get. It is air cooled and contains only two types of board, which make it very easy to build, to expand, and to repair. Our particular model contains 64,000 processors and is about ten-times faster than today's Cray supercomputers, even though it is only about one-tenth the price of a Cray. Applications can be written in Fortran and Lisp, and other languages will be available soon. We have used the Connection Machine for a number of experiments and several organisations have bought them for business purposes.

In one case, we used the Connection Machine for searching through large quantities of text and we made some very careful measurements of its speed in carrying out this task. For this application, the database is large, but the amount of computation is also large, because we are looking for relevant newspaper articles and we have to search the complete text to ensure the right articles are selected. We found that the increase in speed meant that the cost effectiveness of the Connection Machine was 40,000 times better than an IBM mainframe. That is a huge improvement, and it has been replicated many times. Depending on the text being searched, it might only be an increase of 35,000 times, or it might be 45,000 times. That is the scale of improvement this new technology can bring.

The Connection Machine uses conventional wiring, but our next generation of multiprocessor machine, which, I guess will be available in the first quarter of 1989, will make extensive use of fibre optics. This means that the need for heat dissipation will be reduced, the speed will be increased, and the size of the machine will be reduced. Also, because the number of connections and connectors will be reduced, the reliability of the machine will be increased tremendously.

The next machine I want to talk about is BBN's Butterfly Parallel Processor, whose characteristics are shown in Figure 4. This is a much less powerful and less expensive machine than the Connection Machine. It is about as fast as an old Cray (not a new Cray) but it is priced at about one-fifteenth of the cost. The price of a Butterfly is therefore getting near to what could be justified in an engineering department for one or a few people. At the time Figure 4 was made, 65 Butterfly systems had been sold, but many more machines have now been sold. The Butterfly development environment is based on Unix, which is common for machines of this type.

Figure 5 shows that the Butterfly represents just one of several architectural classes for parallel processing. The Butterfly is an MIMD (multiple instruction, multiple data) machine because each

Figure 4 Butterfly parallel processor

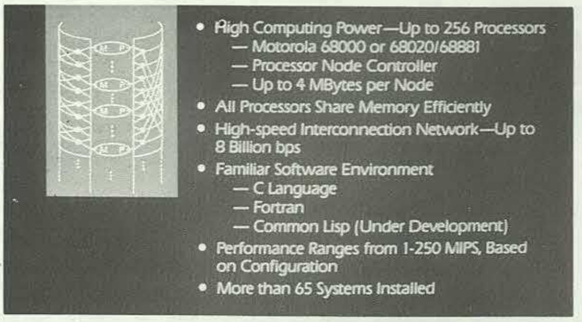


Figure 5 Architectural classes for parallel processing

	Interconnect			
Class	Switch	Cube	Bus	Other
MIMD	Butterfly	Intel NCUBE Ametek Floating Point Systems	Flexible Alliant Sequent Encore Elxsi	WARP (grid)
SIMD		Thinking Machines		Goodyear MPP (grid) NonVon (tree) ICL DAP (grid)

of its 256 processors can perform a different instruction on different data at the same time, and it has a central switch. By contrast, Thinking Machines Corporation's Connection Machine is an SIMD (single instruction, multiple data) machine because it performs the same instruction on lots of different data at the same time and it is interconnected in a cube scheme. The figure shows several other classes as well, and there are now many more not shown in the figure.

In the Butterfly, each of the processors is connected to a central switch and each can send requests for information or send replies. Physically, each processor has its own memory but from the point of view of the programmer it looks as if there are many processors sharing a single memory. Figure 6 illustrates how the switch, which is the important part of the machine, works. In the upper left, there is the binary address of a message. Its route through the switch is determined by a series of decisions based on whether the next digit in the address is zero or one. Zero means up, and one means down. By following this simple logic, the message is delivered to the right place. The simple switching logic also means that the switch operates quickly and is inexpensive.

Sometimes, switch contention occurs where two different processors send messages that arrive simultaneously at the same point in the switch. This problem is easy to solve by adding an extra column in the switch so that there are multiple paths.

Figure 7 shows the development configuration for a Butterfly where it is connected to a host computer,

Figure 6 Butterfly switch

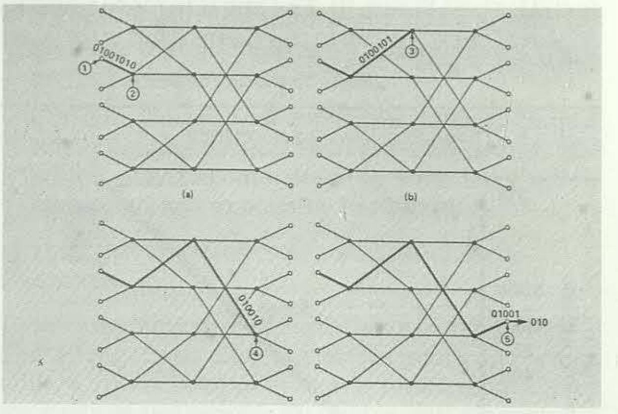


Figure 7 Development configuration

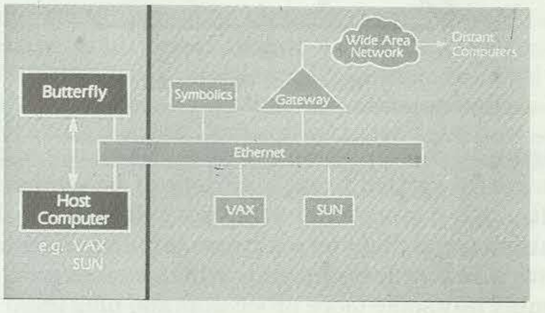


Figure 8 BBN Advanced Computers Inc. partial list of customers

Organization	Application
Advanced Decision Systems	Expert Systems
Ballistic Research Laboratory	Simulation, Battle Management
Boeing Computer Services	AI Research
DuPont	Chemical Plant Simulation
GTE Government Systems	Parallel Processing Research
Government Agencies	Classified
Hughes Aircraft	Image Understanding
Martin Marietta	Expert Systems, Vision
Naval Research Laboratory	Parallel Processing Research
Northrop Corporation	Real-Time Simulation
RCA	Parallel Processing Research
Rockwell International	Vision, Robotics
Science Applications International	Expert Systems
University of California	Circuit Simulation
University of Maryland	Image Understanding
University of Rochester	Vision, AI Research, Languages
University of Utah	Simulation

which could be a Vax, or a Sun, or a Symbolics machine, or whatever. Figure 8 shows a partial list of organisations that have purchased a Butterfly machine and the types of application they are using it for.

PERFORMANCE OF PARALLEL PROCESSORS

Developments in multiprocessors have been going on for 20 years or so. Up till a few years ago, we did not know how to interconnect and program the different processors to make effective use of additional processors. What happened was that as a few processors were added the machine got faster, but adding more processors caused the performance to drop off almost to zero as the different processors competed with each other for communication facilities (see Figure 9). The aim is to move towards the ideal curve shown in the figure. That is the cost-effective curve.

Figure 9 Theoretical linear performance

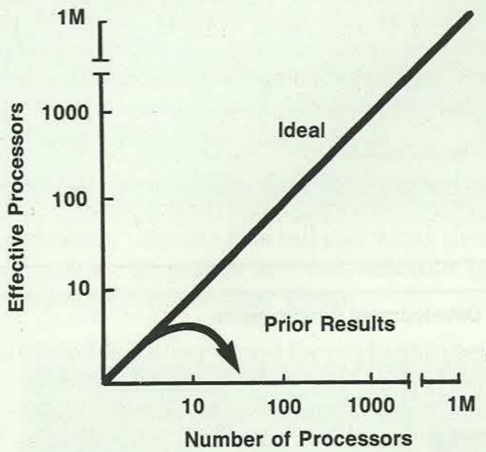
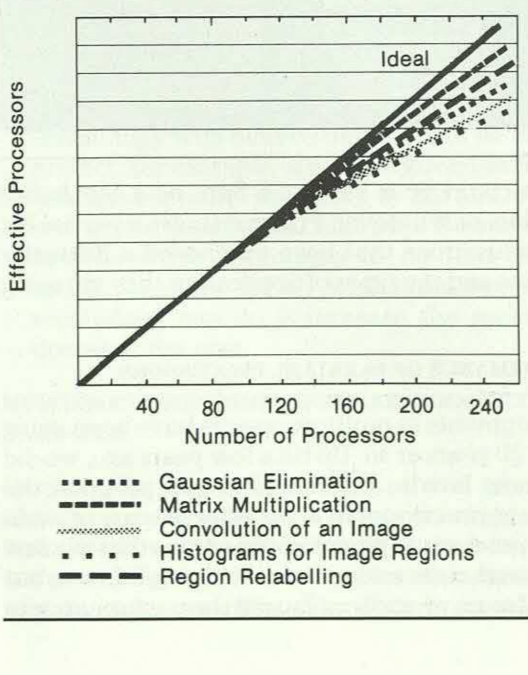


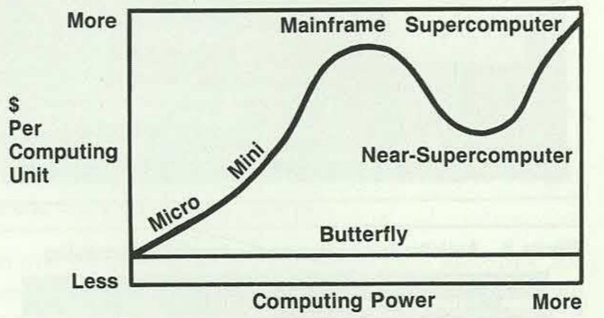
Figure 10 shows the measured performance of the Butterfly system using a variety of benchmarks. Although it is not ideal, it is pretty close. What this data shows is that you really can buy a small system with a few processors and, when you outgrow it, simply buy some more processors and plug them in, and you will speed up the system in proportion to the number of additional processors. It also means that you have a very cost-effective investment. In terms of dollars per mips, the Butterfly costs much less than large minicomputers (such as 4300s, VAXs, Prime Series 50, Data General's MV Series, and

Figure 10 Measured performance of Butterfly system



Wang's VS Series). On the same basis, it also costs much less than large mainframes and today's supercomputers. The superior price-performance of the Butterfly is illustrated graphically in Figure 11.

Figure 11 Price performance



We have continued to develop the Butterfly concept in Project Monarch where we are building machines with even more processors. We will have a 'small' machine with 1,000 processors and a large one with 8,000 processors. The larger machine operates at up to 8,000 mips, which is probably about the limit for this kind of machine.

Figure 12 lists some facts about another multiprocessor computer — the T Series made by Floating Point Systems, a modest-sized company in Oregon. The interesting thing about this machine is that its peak performance is equivalent to about 200 Crays, or 262 gigaflops.

Figure 12 T Series facts

- Parallel supercomputer based on Caltech hypercube approach
- Announcement: 2nd quarter, 1986
- First Customer Ship: 2nd quarter, 1986
- Air Cooled
- Mass Producible: 2 board types, 3 major chip types
- Modular Expandability from:
  Work station: 2'x2'x4', 256 MFLOPS
  Building: Peak performance equal to 200 Cray 2's, 262 GFLOPS

BUSINESS APPLICATIONS FOR MULTIPROCESSOR MACHINES

Let me now give you some examples of business applications for multiprocessor machines. A company in Los Angeles called Digital Productions has purchased some of these machines in order to produce television commercials and computer generated movies. Previously, it used a Cray but has found that multiprocessor machines are faster and cheaper. Other companies in the same business are now also using multiprocessor machines. A company called MRJ, a small US firm that is part of Perkin Elmer, has purchased two Connection

Machines. They are being used to design optical systems and microelectronic circuits, for searching through text documents, and for analysing images. The Northrop Corporation has been a pioneer in purchasing and using a lot of multiprocessor machines. It is probably ahead of any other aerospace manufacturer and is using the machines for aerodynamic design.

Aerodynamic design is a very important issue for DARPA because we are responsible for building the very fast aircraft that President Reagan called the 'Orient Express' — the so-called national aerospace plane that is supposed to be able to fly from Washington to Tokyo in a couple of hours. That plane will fly at such a high speed that there are no wind tunnels available for testing the design. We therefore have to rely completely on numerical simulation. Today, we are using Crays for those calculations, but we are in the process of changing to Connection Machines. For a typical aircraft design we need to make calculations at 10,000 points on the surface, and the calculations for each point take 72 hours on a Cray — hence the need for us to switch to the new machinery.

Other business applications for multiprocessor machines include:

— Producing training devices and training simulators.

— Controlling the design of new materials and new polymers.

— Controlling the processes in chemical factories.

— Robotics applications, which require a great deal of computation to obtain the precise control of complex processes, which requires more computational power than you can get economically in the conventional way.

— Designing very large scale integrated circuits.

— Providing information for aircraft pilots.

— Various artificial intelligence applications such as speech understanding, natural-language understanding, and computerised vision systems.

IMPLICATIONS FOR INFORMATION SYSTEMS DIRECTORS

If you, as information systems directors, decide to buy a multiprocessor machine, you can do so in the knowledge that you will not be pioneers — several hundreds of such machines have already been installed. If you are not yet ready to install a multiprocessor machine you should at least begin to consider them as you develop systems for conventional hardware. I believe that it is inevitable that you will be using multiprocessor machines in the future, so the earlier you start to think about the implications of using them the easier the transition will be. Specific actions include:

— Use operating systems like Unix wherever you can.

— Divide up problems into logical pieces that later can be mapped easily onto different processors.

— As you design and implement applications, remember that you may well re-implement them on multiprocessor machines.

— Consider the types of applications that today are too expensive or that take too long to run, but which may become possible with very much faster and less expensive hardware.

FUTURE DEVELOPMENTS IN MULTIPROCESSOR MACHINES

Today's multiprocessor machines run at between 1 and 10 gigaflops. The norm five years ago was a million instructions per second. Today, it is a billion instructions per second. By 1990 or 1991 we expect it to be a teraflop — a trillion instructions per second.

When I set our engineers to work on designing teraflop machinery I thought it was a great challenge, that it would be difficult, that it would tax their ingenuity, and that it would require great creativity. I was disappointed, however, because it turned out to be so easy that they came up with four competing designs. Consequently, we have now set them a second much harder challenge — to design a 'petaflop' machine capable of processing a thousand trillion instructions per second. It is probably going to take us to the mid-1990s to achieve that.

MICROELECTRONICS PACKAGING

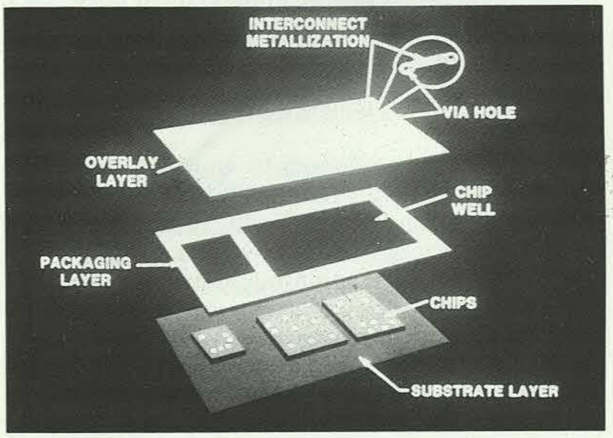
The second technology in which discontinuities are occurring is that of microelectronics packaging. I am going to mention just two new techniques — high-density interconnect, which has been developed by General Electric, and a process for literally gluing chips together, invented by a small firm called Irvine Sensors. These two are representative of several other techniques that are all aimed at making devices that are about a thousand times smaller than the best technology available today.

There are two main advantages from reducing the size of electronic components. The most obvious one is that it enables devices to be small enough to become portable. Second, smaller components lead to much faster devices. Decreasing the volume by a factor of about a thousand means on average that the distance signals have to travel is decreased by about a factor of ten.

The limiting factor on the speed of a computer is the 'slow' speed of light — it takes a nanosecond to travel one foot. The smaller you can build a computer, the faster it will operate. Unfortunately, small computers lead to heat dissipation problems. The odd shape of a modern Cray is determined by the need to keep it small whilst preventing it from melting down during the first couple of minutes it is turned on. The new packaging techniques are concerned with building very compact devices that do not have heat-dissipation problems.

What General Electric has done is to take a set of chips and mount them very close together on any kind of substrate (see Figure 13). There is then an intermediate packaging layer, on top of which an overlay (or decal) is placed. The overlay contains connectors for joining the different pieces of the chips together, both for purposes of data communication and for power.

Figure 13 High density interconnect (exploded view)



The clever part of GE's technique is in the production of the decal. It is not too difficult to produce the decal that, in theory, should match up with the chips. In practice, however, for reasons to do with manufacturing economy the chips on the substrate are slightly at angles to each other — they are slightly out of alignment. This means that a standard decal will not precisely match the chips as they are laid out on the substrate. GE's process produces a custom-made decal that exactly matches the slight misalignments of the chips. The process uses an adaptive lithography system that produces a custom-made mask. It is rather like making made-to-measure suits.

Using this technique means that the layers can be stacked quite close together, giving large reductions in volume. Furthermore, you do not have the heat-dissipation problems that you have with older technology because the chips face upwards, not downwards. We are in the process of building a computer with components produced by this technique. It will be ready in a year or two, and will be the size of a can of soup. However it will be a one gigaflop machine with four gigabytes of memory.

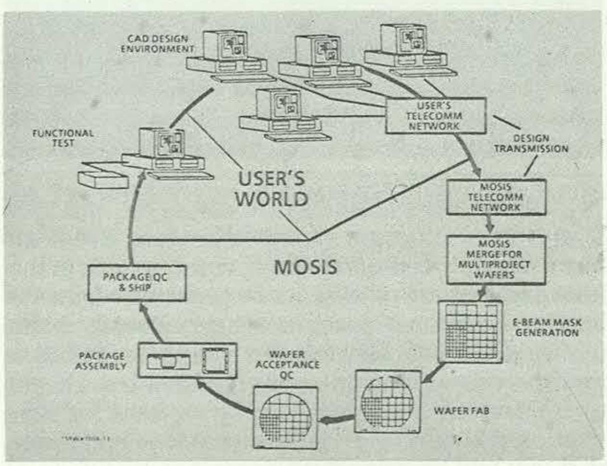
The Irvine Sensors' technology is even more exotic. This company has worked out how to take 128 chips and glue them together into a stack and then lithographically lay down a wiring backplane on the side of the stack. The end result is a 1 cm cube that is almost solid silicon but contains 128 chips. We are using this type of component to build a trillion-bit memory that will occupy one cubic foot. Trillion-bit memories for storing large databases have been available for ten years or so but they require the space of an average hotel room. There will also be corresponding increases in speed and reliability. Because the memory is basically solid state, there are no connectors that can be loosened by vibrations.

SEMICONDUCTOR MANUFACTURING

The next area in which we see technological discontinuities occurring is semiconductor manufacturing. In the past, if you wanted to produce a microelectronics circuit for a special application, you designed the chip, sent the design to a factory, waited for anything up to a year and eventually received 50,000 chips at a tremendous cost. For many applications that type of process, and cost, are just not acceptable. Supposing you only want one chip or a few chips? Or you want to get the chip in a week or two so that you can try out a new design? Today, with MOSIS (metal oxide semiconductor implementation service) that is exactly what you can do. It is possible to get just one chip in two weeks for a cost of $1,000.

The way work flows through the MOSIS system is shown in Figure 14. The designer sits at his computer-aided design terminal and the design specification is sent off over the telecommunications network. The specification is merged with several others so that many designs can be included on the same wafer. An electron beam mask is then generated, a wafer is fabricated, the individual chips are assembled into packages and tested, before being shipped back to the designer two weeks after the specification was transmitted.

Figure14 Project flow through MOSIS



Prior to the advent of the MOSIS system it was impossible to do small-volume manufacturing. It just was not possible to try out an idea or to do rapid prototyping. MOSIS has been in operation for a couple of years, and it is used regularly by 300 or 400 designers.

Progress is also being made in improving the quality of the basic raw material from which chips are manufactured. The problem is that the basic wafers have too many defects, which means that a high percentage of the chips have to be thrown away. For a mature process, anything up to 40 per cent might have to be thrown away. For a leading-edge process, using the most advanced technology, over 99 per cent of the chips might have to be thrown away. That sort of yield is just not acceptable.

We discovered that there was a professor who routinely made extremely pure wafers, but only in small quantities. When his wafers were used to produce chips the yield was very high. One of DARPA's scientists went to the professor's laboratory to try and work out how he managed to produce high-quality wafers, and to see if his techniques could be applied to the mass production of wafers.

We found that, first, the professor was smarter than the average production-line manager in a factory and he controlled the process better. He used better heuristics, better rules of thumb, to produce a higher quality product. Second, he used different sensors to control the process than those available in the factory. In particular, he used his eyes to observe the crystal on a second-by-second basis as it grew and made continuous judgements about exactly how to control the process. That kind of 'eyeball' sensor is not usually available on the factory floor. However, we set about determining if it was possible to produce industrial-quality sensors that would do some of the tasks being done in the university laboratory, and whether it was possible to capture the heuristics of the university professor in an expert system. The answer was 'yes'. Using the expert system has resulted in gallium arsenide wafers being mass produced with almost the same purity as those achieved in the laboratory.

As an aside, the principles we have devised for manufacturing gallium arsenide and other semiconductor materials can be applied to producing advanced materials in general. Anybody associated with advanced composites and polymers and ceramics knows that there are many great ideas in the laboratory and that it is possible to produce small quantities of a fantastic new material. The problem is that it is extremely difficult to scale-up the fabrication process to produce usable quantities. We are using this technique for carbon-carbon composites and for a variety of advanced materials.

One of the major problems with semiconductor manufacturing is to create a 'clean' environment in which to carry out the production process. Market forces push manufacturers to smaller and smaller dimensions, smaller and smaller devices, so that more devices can be packed onto a chip. Because the features are so small, very small dust particles can cause an individual chip to be worthless and cause the number of rejects in a batch of chips to be high. The cost of setting up a clean factory is very high — anything up to $2,000 per square foot, and that is before the costs of people and equipment. Moreover, the factory will probably be obsolete within two to three years.

A US company came up with an idea to solve this problem that, with hindsight, is obvious but is very hard to implement in practice. They reasoned that you only need the immediate area around the wafer to be clean — you do not need the whole building to be clean. They produced vacuum cocoons within which there is an almost completely clean environment where it is possible to produce chips with a much higher yield. More importantly, the capital investment required for a new chip-making facility has been reduced by a factor of between two and four. That is a huge decrease in an industry that works on margins of one or two per cent.

There are also other advantages to the new technology — one being that you can intermix different materials within the same factory.

The last advance in semiconductor manufacturing I want to highlight stems from work done at the Lincoln Laboratory, which is part of MIT. The conventional process of producing microelectronic circuits involves some 200 separate steps. A small mistake at any one of those steps can mean that the entire batch has to be thrown away. Researchers at the Lincoln Laboratory devised a new manufacturing technique that reduced the number of steps from 200 to an average of between 40 and 50. This technique can lead to a huge increase in yield. For a mature process producing a 40 per cent yield with the conventional techniques, we have evidence that the yield can increase to 80 per cent. More importantly, for a process that uses leading-edge technology where the yield might only be one per cent with conventional techniques, we have evidence that the yield can be increased to 20 per cent.

Advances in semiconductor manufacturing such as those I have just mentioned will allow much higher performance circuits to be produced at much lower cost. Those circuits will be needed to produce machines like the teraflop and the petaflop computers that I talked about earlier.

SUPERCONDUCTORS

The topic of superconductors made from new ceramic materials is receiving much attention at present. The potential of the new materials is tremendous, although they do have one major drawback in that they are very brittle and are therefore hard to form into precise shapes, and it is difficult to control their quality. We have begun to speculate about the types of computer that could be produced as a consequence of using these new materials. First, they could be made much smaller because there would be no heat to dissipate. We have begun to think in terms of 'Crays per cubic centimetre', and we have several designs on the drawing board. These designs will become reality once some of the practical problems of dealing with the new materials have been solved.

Substantial progress is being made in solving the problems. The temperature at which the materials now reliably operate has been raised to 150°K from about 90°K. Many people have reported superconductor materials that will operate at room temperature, but usually for just a few minutes. The superconductivity effect has then disappeared and no one has yet worked out how to reproduce it. I believe that, someday, superconductors operating at room temperature will be available, but I am not prepared to speculate when.

Progress is also being made in making the materials much less brittle. At Pennsylvania State University, a technique for inserting polymers, plastic strands, into the ceramics is being used to strengthen the materials so they become ductile (flexible) even at operating temperatures as low as liquid nitrogen. Thus, it is now possible to make coils of wire from the new materials.

Some of the hydroscopic problems have also been solved. It is not widely known that these materials absorb water from the air and dissolve themselves (we call it 'ceramic suicide'). We have also shown that it is possible to lay down thin films reliably, and it is possible to do thin-line lithography, although not at the feature sizes that will be required in the future. We can reliably show Josephson Junction effects, creating switching times of a few nanoseconds. However, although progress is being made, superconductor technology will not be available for commercial use in the next three years or so.

LIGHTWEIGHT SATELLITES

Last year, we decided to take a whole new approach to the satellite business (although, in effect, we were returning to the techniques and technology of the early 1960s). At present, it takes up to seven years to design and build a satellite, because they are designed to be incredibly reliable and to operate for many years. As a consequence, they are extraordinarily expensive. Hundreds of millions of dollars per satellite is not at all unusual. Because they are so expensive, very few satellites are launched. They provide tremendous functionality when they are in space, but because they are large and heavy, they are difficult to launch. If there is a problem on the launching pad and the satellite does not get into orbit, you lose a great deal. All of these constraints mean that fewer and fewer satellites are launched each year.

We set about finding out if it was possible to produce satellites quickly using commercial-construction practices, rather than space-construction practices. We wanted to reverse the situation where 20 pounds of satellite was accompanied by 150 pounds of documentation. We wanted to build inexpensive satellites, costing less than $1 million a satellite. We were prepared to accept the fact that cheaper satellites are less reliable — perhaps nine out of ten will work. However, it turns out, in many cases, to be a better way of doing business. The result was an experiment called GLOMR (Global Low Orbiting Message Relay). Some of the specifications for the GLOMR satellite are shown in Figure 15.

Figure 15 GLOMR orbital characteristics

Launch:	Shuttle Mission 61A
GLOMR Launch Date:	31 Oct 85
Launch Method:	GAS Can Spring Ejection
Seperation Velocity:	4 Ft/Second
Inclination:	56.98 Degrees
Launch Altitude:	175 NM
Period:	90 Minutes
Single Pass Visibility:	2 - 10 Minutes
	3-4 Consecutive Revolutions
	6 Hour Gap
Re-Entry:	Mid-Sep 86

It took us just 11 months to build GLOMR. As a consequence, we were able to use the most modern technology rather than technology that is five, six, seven years old. We made no effort to keep the satellite in orbit for a long period of time, but it did work perfectly for 14 months. We have now set up a production line to manufacture these communications satellites.

One of the other reasons we are interested in lightweight satellites is that it is a lot easier to launch lightweight satellites than heavy satellites. There are many more launch vehicles available, many more opportunities and options.

We are now building and launching a system called the multiple satellite system (see Figure 16) that consists of 240 small satellites like the GLOMR satellite. This system will provide very reliable communications for up to 10,000 users around the globe. It will be reliable because of the large number of satellites — up to 50 per cent of them could fail before a typical user will notice any degradation in the service.

EXPERT SYSTEMS APPLICATIONS

I would like now to share with you some of our experiences of building expert systems. At present, there are 1,500 expert systems in use in American

Page 16 General MSSP system parameters

NUMBER OF SATELLITES	240
INCLINATIONS	90 at 27.5 deg
	90 at 57.5 deg
	60 at 90 deg
ALTITUDE	340 – 400 nmi
ORBIT PERIOD	92 – 98 min
VELOCITY	4 nmi/sec
EARTH COVERAGE PER SAT.	2 %
ALTITUDE STABILIZATION	GRAVITY GRADIENT BOOM
ROTATIONAL MOTION	0 –1 rpm (slowly varing)
STATION KEEPING	NONE
NUMBER OF TERMINALS	200 – 1000

BASED ON ESL-TM1632

businesses and about 3,000 more are under development. The first lesson is that it is impossible to predict costs and computational requirements. DARPA is building about a dozen quite large expert systems, and all of them will end up costing much more than was expected.

We found that you do not need to hire a specialist knowledge engineer to build expert systems — a scientist, or an engineer, or an accountant, or businessman can do the job. It is also not necessary always to produce big systems. We found that very small systems implemented on a personal computer such as an IBM PC or Macintosh can perform intelligently. Hitherto, the premise has been that only large systems could act intelligently. Our experience is that this is a false premise.

We found that expert systems do not have to be written in specialised AI languages such as Lisp. We have systems written in at least a dozen different languages, most of which will be familiar to you. We found that, increasingly, expert systems are not written in isolation and are not used in isolation. Instead, they are combined with other more conventional applications. For example, an oil company might couple a simulation system to an expert system that contains heuristic guidelines for pricing. The implication is that system developers have to have a broad range of skills ranging from conventional programming to AI programming.

However, although a lot of expert systems can run on small computers, we find that there are more and more applications that require extremely high speed computers to obtain a reasonable operational performance. Thus we are now, almost as a matter of routine, putting expert systems onto multiprocessors. The advantage is that, as the system develops and we find that it requires more computing power than we anticipated, we just buy more processors and add them on.

We have one large expert system in operation in Hawaii at the moment. It helps to 'control' the American Pacific Fleet by providing advice, by making tentative decisions, for American Commanders. The Commanders then decide whether or not to take the advice offered by the system. The system makes plans based on a set of high-level goals provided by human operators. In providing its advice, the system takes account of the conditions of the ships, the national needs, weather conditions, the history of equipment break downs, and unexpected needs. The system is monitoring hundreds of different ships, not only from the United States but also from many other countries, and provides a continuous stream of operational suggestions — where to send ships, what repairs to do, what preventive maintenance to do, and so on and so forth.

I am giving you a military example because that is what I am familiar with. However, I am sure you can all extrapolate from this example to your own business environments and find examples of situations where there are a large number of interacting factors, where you have to make plans, where you have goals, where unexpected events take place, and where sometimes you have to be rather clever and creative to use your resources in the best way possible.

We built this system by installing a new multiprocessor computer in some borrowed floor space in the command centre where the work was usually done. We created an optical link between the new computer and the existing database management computer system so that data could flow into the new computer system, but so that there was no easy way that information could accidentally be sent back to the operational system. We developed the expert system by mercilessly interrogating the people who currently perform the job to understand how they did it, what they cared about, and why they did the things they did. Gradually, we built up a rather large expert system — the equivalent of several thousand rules, although it is more complicated than that because it is not all rule-based. But it is a very large system that requires considerable computational power to handle the many thousands of messages an hour about things that are happening to the fleet.

We were delighted to find that people stopped using the old system and started using the new system. No one told them to do this. In fact, if they had been ordered to use the new system, they would probably have resisted the change.

Occasionally, though, people would disagree with the advice provided by the system. Remember that these were the same people who both wrote the system and who previously did that job. However, after the system had explained its line of reasoning in arriving at its recommended action, and had described the other possibilities that had been considered and rejected, we found that the system's advice was accepted.

For me, that is the real significance of the technological advances I was talking about earlier — the fact that you can combine the technology and AI techniques in clever and unusual ways to produce systems that can perform a task better than people can. If you run conventional systems on faster hardware you get the results faster. With expert systems (and other AI applications) you get *better* answers by using faster hardware, because additional computational power allows the software to try out more combinations, to look at more data, and to examine more ways of solving the problem. There may be limits to how far this process can be carried, but at present I do not know where those limits are. Thus, by investing in today's expert systems technology, by investing in today's multiprocessor technology, you may create systems that can sometimes behave better than your smartest employees.

Butler Cox

Butler Cox is an independent management consultancy and research organisation, specialising in the application of information technology within commerce, government and industry. The company offers a wide range of services both to suppliers and users of this technology. The Butler Cox Foundation is a service operated by Butler Cox on behalf of subscribing members.

Objectives of the Foundation

The Butler Cox Foundation sets out to study on behalf of subscribing members the opportunities and possible threats arising from developments in the field of information systems.

New developments in technology offer exciting opportunities — and also pose certain threats — for all organisations, whether in industry, commerce or government. New types of systems, combining computers, telecommunications and automated office equipment, are becoming not only possible, but also economically feasible.

As a result, any manager who is responsible for introducing new systems is confronted with the crucial question of how best to fit these elements together in ways that are effective, practical and economic.

While the equipment is becoming cheaper, the reverse is true of people — and this applies both to the people who design systems and those who make use of them. At the same time, human considerations become even more important as people's attitudes towards their working environment change.

These developments raise new questions for the manager of the information systems function as he seeks to determine and achieve the best economic mix from this technology.

Membership of the Foundation

The majority of organisations participating in the Butler Cox Foundation are large organisations seeking to exploit to the full the most recent developments in information systems technology. An important minority of the membership is formed by suppliers of the technology. The membership is international with participants from Australia, Belgium, France, Italy, the Netherlands, Sweden, Switzerland, the United Kingdom and elsewhere.

The Foundation Research Programme

The research programme is planned jointly by Butler Cox and by the member organisations. Each year Butler Cox draws up a short-list of topics that reflects the Foundation's view of the important issues in information systems technology and its application. Member organisations rank the topics according to their own requirements and as a result of this process members' preferences are determined.

Before each research project starts there is a further opportunity for members to influence the direction of the research. A detailed description of the project defining its scope and the issues to be addressed is sent to all members for comment.

The Report Series

The Foundation publishes six research reports each year. The reports are intended to be read primarily by senior and middle managers who are concerned with the planning of information systems. They are, however, written in a style that makes them suitable to be read both by line managers and functional managers. The reports concentrate on defining key management issues and on offering advice and guidance on how and when to address those issues.

Butler Cox & Partners Limited
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
☎ (020) 139955, Fax (020) 131157

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

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

United States of America
Butler Cox Inc.
150 East 58th Street, New York, NY 10155, USA
☎ (212) 486 1760

Australia/New Zealand
Mr J Cooper
Butler Cox Foundation
3rd Floor, 275 George Street, Sydney 2000, Australia
☎ (02) 236 6161, Fax (02) 236 6999

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

Italy
SISDO
20123 Milano, Via Caradosso 7, Italy
☎ (02) 498 4651, Telex 350309, Fax (02) 481 8842

The Nordic Region
Statskonsult AB
Stora Varvsgatan 1, 21120 Malmo, Sweden
☎ (040) 1030 40, Telex 12754 SINTABS

Spain
Associated Management Consultants Spain SA
Rosalía de Castro, 84-2°D, 28035 Madrid, Spain
☎ (91) 723 0995