NEAR EAST UNIVERSITY

FACULTY OF ENGINEERING

Department of Computer Engineering

GRADUATION PROJECT
COM 400

SUBJECT:Parallel Programming

Supervisor :Besime Erin
Submitted by :Ahmet Binici
Number :980262
Department :Computer Engineer

JUNE 2002

ACKNOWLEDGEMENT

I would like to thank Miss. Besime Erin for accepting to be my supervisor and her support for this project.

I am so grateful to my parents who had always shown patience and understanding to me.

Also, I would like to tkank all the lecturers for helping me see this graduation term.

And finally, I would like to thank all my friends for their support in school and in social life.

ABSTRACT

Ever since conventional serial computers were invented, their speed has steadily increased to match the needs of emerging applications. However, the fundamental physical limitation imposed by the speed of light makes it impossible to achieve further improvements in the speed of such computers indefinitely. Recent trends show that the performance of these computers is beginning to saturate. A natural way to circumvent this saturation is to use an ensemble of processors to solve problems.

The transition point has become sharper with the passage of time, primarily as a result of advances in very large scale integration (VLSI) technology. It is now possible to construct very fast, low-cost processors. This increases the demand for and production of these processors, resulting in lower prices.

Currently, the speed of off-the-shelf microprocessors is within one order of magnitude of the speed of the fastest serial computers. However, microprocessors cost many orders of magnitude less. This implies that, by connecting only a few microprocessors together to form a parallel computer, it is possible to obtain raw computing power comparable to that of the fastest serial computers. Typically, the cost of such a parallel computer is considerably less.

Furthermore, connecting a large number of processors into a parallel computer overcomes the saturation point of the computation rates achievable by serial computers. Thus, parallel computers can provide much higher raw computation rates than the fastest serial computers as long as this power can be translated into high computation rates for actual applications.

TABLE OF CONTENTS

INTRODUCTION

The technological driving force behind parallel computing is VLSI, or very large scale integration-the same technology that created the personal computer and workstation market over the last decade. In 1980, the Intel 8086 used 50,000 transistors; in 1992, the latest Digital alpha RISC chip contains 1,680,000 transistors-a factor of 30 increase. The dramatic improvement in chip density comes together with an increase in clock speed and improved design so that the alpha performs better by a factor of over one thousand on scientific problems than the 8086-8087 chip pair of the early 1980s.

High-performance computers are increasingly in demand in the areas of structural analysis, weather forecasting, petroleum exploration, medical diagnosis, aerodynamics simulation, artificial intelligence, expert systems, genetic engineering, signal and image processing, among many other scientific and engineering applications. Without superpower computers, many of these challenges to advance human civilization cannot be made within a reasonable time period. Achieving high performance depends not only on using faster and more reliable hardware devices but also on major improvements in computer architecture and processing techniques.

There are a number of different ways to characterize the performance of both parallel computers and parallel algorithms. Usually, the peak performance of a machine is expressed in units of millions of instructions executed per second (MIPS) or millions of floating point operations executed per second (MFLOPS). However, in practice, the realizable performance is clearly a function of the match between the algorithms and the architecture.

CHAPTER 1

1 What is Parallel Computing?

Consider the problem of stacking (reshelving) a set of library books A single worker trying to stack all the books in their proper places cannot accomplish the task faster than a certain rate. We can speed up this process, however, by employing more than one worker. Assume that the books are organized into shelves and that the shelves are grouped into bays. One simple way to assign the task to the workers is to divide the books equally among them. Each worker stacks the books one at a time. This division of work may not be the most efficient way to accomplish the task, since the workers must walk all over the library to stack books. An alternate way to divide the work is to assign a fixed and disjoint set of bays to each worker. As before, each worker is assigned an equal number of books arbitrarily. If a worker finds a book that belongs to a bay assigned to him or her, he or she places that book in its assigned spot. Otherwise, he or she passes it on to the responsible for the bay it belongs to. The second approach requires less effort from individual workers.

The preceding example shows how a task can be accomplished faster by dividing it into a set of subtasks assigned to multiple workers. Workers cooperate, pass the books to each other when necessary, and accomplish the task in unison. Parallel processing works on precisely the same principles. Dividing a task among workers by assigning them a set of books is an instance of task partitioning. Passing books to each other is an example of communication between subtasks.

Problems are parallelizable to different degrees. For some problems, assigning partitions to other processors might be more time-consuming than performing the processing locally. Other problems may be completely serial. For example, consider the task of digging a post hole. Although one person can dig a hole in a certain amount of time, employing more people does not reduce this time. Because it is impossible to partition this task, it is poorly suited to parallel processing. Therefore, a problem may have different parallel formulations, which result in varying benefits, and all problems are not equally amenable to parallel processing.

2 The Scope of Parallel Computing

Parallel processing is making a tremendous impact on many areas of computer application. With the high raw computing power of parallel computers, it is now possible to address many applications that were until recently beyond the capability of conventional computing techniques.

Many applications, such as weather prediction, biosphere modeling, and pollution monitoring, are modeled by imposing a grid over the domain being modeled. The entities within grid elements are simulated with respect to the influence of other entities and their surroundings. In many cases, this requires solutions to large systems of differential equations. The granularity of the grid determines the accuracy of the model. Since many such systems are evolving with time, time forms an additional dimension for these computations. Even for a small number of grid points, a three-dimensional coordinate system, and a reasonable discredited time step, this modeling process can involve trillions of operations. Thus, even moderate-sized instances of these problems take an unacceptably long time to solve on serial computers.

Parallel processing makes it possible to predict the weather not only faster but also more accurately. If we have a parallel computer with a thousand workstation-class processors, we can partition the 10^{11} segments of the domain among these processors. Each processor computes the parameters for 10^8 segments. Processors communicate the value of the parameters in their segments to other processors. Assuming that the computing power of this computer is 100 million instructions per second, and this power is efficiently utilized, the problem can be solved in less than 3 hours. The impact of this reduction in processing time is two-fold. First, parallel computers make it possible to solve a previously unsolvable problem. Second, with the availability of even larger parallel computers, it is possible to model weather using finer grids. This enables more accurate weather prediction.

The acquisition and processing of large amounts of data from sources such as satellites and oil wells form another class of computationally expensive problems. Conventional satellites collect billions of bits per second of data relating to parameters such as pollution levels, the thickness of the ozone layer, and weather phenomena. Other applications of satellites that require processing of large amounts of data include remote sensing and telemetry. The computational rates required for handling this data effectively are well beyond the range of conventional serial computers.

Discrete optimization problems include such computationally intensive problems as planning, scheduling, VLSI design, logistics, and control. Discrete optimization problems can be solved by using state-space search techniques. For many of these problems, the size of the state-space increases exponentially with the number of variables. Problems that evaluate trillions of states are fairly commonplace in most such applications. Since processing each state requires a nontrivial amount of computation, finding solutions to large instances of these problems is beyond the scope of conventional sequential computing. Indeed, many practical problems are solved using approximate algorithms that provide suboptimal solutions.

Other applications that can benefit significantly from parallel computing are semi-conductor material modeling, ocean modeling, computer tomography, quantum chromodynamics, vehicle design and dynamics, analysis of protein structures, study of chemical phenomena, imaging, ozone layer monitoring, petroleum exploration, natural language understanding, speech recognition, neural network learning, machine vision, database query processing, and automated discovery of concepts and patterns from large databases. Many of the applications mentioned are considered grand challenge problems. A grand challenge is a fundamental problem in science or engineering that has a broad economic and scientific impact, and whose solution could be advanced by applying high performance computing techniques and resources.

3 Issues in Parallel Computing

To use parallel computing effectively, we need to examine the following issues:

3.1 Design of Parallel Computers

It is important to design parallel computers that can scale up to a large number of processors and are capable of supporting fast communication and data sharing among processors. This is one aspect of parallel computing that has seen the most advances and is the most mature.

3.2 Design of Efficient Algorithms

A parallel computer is of little use unless efficient parallel algorithms are available. The issues in designing parallel algorithms are very different from those in designing their sequential

counterparts. A significant amount of work is being done to develop efficient parallel algorithms for a variety of parallel architectures.

3.3 Methods for Evaluating Parallel Algorithms

Given a parallel computer and a parallel algorithm running on it, we need to evaluate the performance of the resulting system. Performance analysis allows us to answer questions such as How fast can a problem be solved using parallel processing? and How efficiently are the processors used?

3.4 Parallel Computer Languages

Parallel algorithms are implemented on parallel computers using a programming language. This language must be flexible enough to allow efficient implementation and must be easy to program in. New languages and programming paradigms are being developed that try to achieve these goals.

3.5 Parallel Programming Tools

To facilitate the programming of parallel computers, it is important to develop comprehensive programming environments and tools. These must serve the dual purpose of shielding users from low-level machine characteristics and providing them with design and development tools such as debuggers and simulators.

3.6 Portable Parallel Programs

Portability is one of the main problems with current parallel computers. Typically, a program written for one parallel computer requires extensive modification to make it run on another parallel computer. This is an important issue that is receiving considerable attention.

3.7 Automatic Programming of Parallel Computers

Much work is being done on the design of parallelizing compilers, which extract implicit parallelism from programs that have not been parallelized explicitly. Such compilers are expected

to allow us to program a parallel computer like a serial computer. We speculate that this approach has limited potential for exploiting the power of large-scale parallel computers.

CHAPTER 2

1 Parallelism and Computing

A parallel computer is a set of processors that are able to work cooperatively to solve a computational problem. This definition is broad enough to include parallel supercomputers that have hundreds or thousands of processors, networks of workstations, multiple-processor workstations, and embedded systems. Parallel computers are interesting because they offer the potential to concentrate computational resources-whether processors, memory, or I/O bandwidth-on important computational problems.

Parallelism has sometimes been viewed as a rare and exotic sub area of computing, interesting but of little relevance to the average programmer. A study of trends in applications, computer architecture, and networking shows that this view is no longer tenable. Parallelism is becoming ubiquitous, and parallel programming is becoming central to the programming enterprise.

2 The National Vision for Parallel Computation

The technological driving force behind parallel computing is VLSI, or very large scale integration-the same technology that created the personal computer and workstation market over the last decade. In 1980, the Intel 8086 used 50,000 transistors; in 1992, the latest Digital alpha RISC chip contains 1,680,000 transistors-a factor of 30 increase. The dramatic improvement in chip density comes together with an increase in clock speed and improved design so that the alpha performs better by a factor of over one thousand on scientific problems than the 8086-8087 chip pair of the early 1980s.

The increasing density of transistors on a chip follows directly from a decreasing feature size which is now for the alpha. Feature size will continue to decrease and by the year 2000, chips with 50 million transistors are expected to be available. What can we do with all these transistors?

With around a million transistors on a chip, designers were able to move full mainframe functionality to about of a chip. This enabled the personal computing and workstation revolutions. The next factors of ten increase in transistor density must go into some form of parallelism by replicating several CPUs on a single chip.

By the year 2000, parallelism is thus inevitable to all computers, from your children's video game to personal computers, workstations, and supercomputers. Today we see it in the larger machines as we replicate many chips and printed circuit boards to build systems as arrays of nodes, each unit of which is some variant of the microprocessor. An nCUBE parallel supercomputer with 64 identical nodes on each board-each node is a single-chip CPU with additional memory chips. To be useful, these nodes must be linked in some way and this is still a matter of much research and experimentation. Further, we can argue as to the most appropriate node to replicate; is it a ``small" node as in the nCUBE, or more powerful ``fat" nodes such as those offered in CM-5 and Intel Touchstone, where each node is a sophisticated multichip printed circuit board. However, these details should not obscure the basic point: Parallelism allows one to build the world's fastest and most cost-effective supercomputers.

Parallelism may only be critical today for supercomputer vendors and users. By the year 2000, all computers will have to address the hardware, algorithmic, and software issues implied by parallelism. The reward will be amazing performance and the opening up of new fields; the price will be a major rethinking and reimplementation of software, algorithms, and applications. This vision and its consequent issues are now well understood and generally agreed. They provided the motivation in 1981 when CP's first roots were formed. In those days, the vision was blurred and controversial. Many believed that parallel computing would not work.

President Bush instituted, in 1992, the five-year federal High Performance Computing and Communications (HPCC) Program. The activities of several federal agencies have been coordinated in this program. The Advanced Research Projects Agency (ARPA) is developing the basic technologies which is applied to the grand challenges by the Department of Energy (DOE), the National Aeronautics and Space Agency (NASA), the National Science Foundation (NSF), the National Institute of Health (NIH), the Environmental Protection Agency (EPA), and the National Oceanographic and Atmospheric Agency (NOAA). Selected activities include the mapping of the human genome in DOE, climate modeling in DOE and NOAA, coupled structural and airflow simulations of advanced powered lift and a high-speed civil transport by NASA.

More generally, it is clear that parallel computing can only realize its full potential and be commercially successful if it is accepted in the real world of industry and government applications. The clear U.S. leadership over Europe and Japan in high-performance computing offers the rest of the U.S. industry the opportunity of gaining global competitive advantage.

We note some interesting possibilities which include: use in the oil industry for both seismic analysis of new oil fields and the reservoir simulation of existing fields; environmental modeling of past and potential pollution in air and ground; fluid flow simulations of aircraft, and general vehicles, engines, air-conditioners, and other turbomachinery; integration of structural analysis with the computational fluid dynamics of airflow; car crash simulation; integrated design and manufacturing systems; design of new drugs for the pharmaceutical industry by modeling new compounds; simulation of electromagnetic and network properties of electronic systems-from new components to full printed circuit boards; identification of new materials with interesting properties such as superconductivity; simulation of electrical and gas distribution systems to optimize production and response to failures; production of animated films and educational and entertainment uses such as simulation of virtual worlds in theme parks and other virtual reality applications; support of geographic information systems including real-time analysis of data from satellite sensors in NASA's ``Mission to Planet Earth."

A relatively unexplored area is known as ``command and control" in the military area and ``decision support" or ``information processing" in the civilian applications. These combine large databases with extensive computation. In the military, the database could be sensor information and the processing a multitrack Kalman filter. Commercially, the database could be the nation's medicaid records and the processing would aim at cost containment by identifying anomalies and inconsistencies.

Servers in multimedia networks set up by cable and telecommunication companies. These servers will provide video, information, and simulation on demand to home, education, and industrial users. CP did not address such large-scale problems. Rather, we concentrated on major academic applications. This fit the experience of the Caltech faculty who led most of the CP teams, and further academic applications are smaller and cleaner than large-scale industrial problems. One important large-scale CP application was a military simulation and produced by Caltech's Jet Propulsion Laboratory. CP chose the correct and only computations on which to cut its parallel computing teeth. In spite of the focus on different applications, there are many

similarities between the vision and structure of CP and today's national effort. It may even be that today's grand challenge teams can learn from CP's experience.

3 Trends in Applications

As computers become ever faster, it can be tempting to suppose that they will eventually become ``fast enough'' and that appetite for increased computing power will be sated. However, history suggests that as a particular technology satisfies known applications, new applications will arise that are enabled by that technology and that will demand the development of new technology. As an amusing illustration of this phenomenon, a report prepared for the British government in the late 1940s concluded that Great Britain's computational requirements could be met by two or perhaps three computers. In those days, computers were used primarily for computing ballistics tables. The authors of the report did not consider other applications in science and engineering, let alone the commercial applications that would soon come to dominate computing. Similarly, the initial prospectus for Cray Research predicted a market for ten supercomputers; many hundreds have since been sold.

Traditionally, developments at the high end of computing have been motivated by numerical simulations of complex systems such as weather, climate, mechanical devices, electronic circuits, manufacturing processes, and chemical reactions. However, the most significant forces driving the development of faster computers today are emerging commercial applications that require a computer to be able to process large amounts of data in sophisticated ways. These applications include video conferencing, collaborative work environments, computer-aided diagnosis in medicine, parallel databases used for decision support, and advanced graphics and virtual reality, particularly in the entertainment industry. For example, the integration of parallel computation, high-performance networking, and multimedia technologies is leading to the development of video servers, computers designed to serve hundreds or thousands of simultaneous requests for real-time video. Each video stream can involve both data transfer rates of many megabytes per second and large amounts of processing for encoding and decoding. In graphics, three-dimensional data sets are now approaching volume elements (1024 on a side). At 200 operations per element, a display updated 30 times per second requires a computer capable of 6.4 operations per second.

Although commercial applications may define the architecture of most future parallel computers, traditional scientific applications will remain important users of parallel computing technology. Indeed, as nonlinear effects place limits on the insights offered by purely theoretical investigations and as experimentation becomes more costly or impractical, computational studies of complex systems are becoming ever more important. Computational costs typically increase as the fourth power or more of the "resolution" that determines accuracy, so these studies have a seemingly insatiable demand for more computer power. They are also often characterized by large memory and input/output requirements. For example, a ten-year simulation of the earth's climate using a state-of-the-art model may involve floating-point operations, ten days at an execution speed of floating-point operations per second (10 gigaflops). This same simulation can easily generate a hundred gigabytes (bytes) or more of data. Yet scientists can easily imagine refinements to these models that would increase these computational requirements 10,000 times.

In summary, the need for faster computers is driven by the demands of both data-intensive applications in commerce and computation-intensive applications in science and engineering. Increasingly, the requirements of these fields are merging, as scientific and engineering applications become more data intensive and commercial applications perform more sophisticated computations.

4 Trends in Computer Design

The performance of the fastest computers has grown exponentially from 1945 to the present, averaging a factor of 10 every five years. While the first computers performed a few tens of floating-point operations per second, the parallel computers of the mid-1990s achieve tens of billions of operations per second. Similar trends can be observed in the low-end computers of different eras: the calculators, personal computers, and workstations. There is little to suggest that this growth will not continue. However, the computer architectures used to sustain this growth are changing radically from sequential to parallel.

The performance of a computer depends directly on the time required to perform a basic operation and the number of these basic operations that can be performed concurrently. The time to perform a basic operation is ultimately limited by the ``clock cycle'' of the processor, that is, the time required to perform the most primitive operation. However, clock cycle times are

decreasing slowly and appear to be approaching physical limits such as the speed of light. We cannot depend on faster processors to provide increased computational performance.

To circumvent these limitations, the designer may attempt to utilize internal concurrency in a chip, for example, by operating simultaneously on all 64 bits of two numbers that are to be multiplied. However, a fundamental result in Very Large Scale Integration (VLSI) complexity theory says that this strategy is expensive. This result states that for certain transitive computations (in which any output may depend on any input), the chip area A and the time T required to perform this computation are related so that must exceed some problem-dependent function of problem size. This result can be explained informally by assuming that a computation must move a certain amount of information from one side of a square chip to the other. The amount of information that can be moved in a time unit is limited by the cross section of the chip. This gives a transfer rate of , from which the relation is obtained. To decrease the time required to move the information by a certain factor, the cross section must be increased by the same factor, and hence the total area must be increased by the square of that factor.

This result means that not only is it difficult to build individual components that operate faster, it may not even be desirable to do so. It may be cheaper to use more, slower components. For example, if we have an area of silicon to use in a computer, we can either build components, each of size A and able to perform an operation in time T, or build a single component able to perform the same operation in time T/n. The multicomponent system is potentially n times faster. Computer designers use a variety of techniques to overcome these limitations on single computer performance, including pipelining (different stages of several instructions execute concurrently) and multiple function units (several multipliers, adders, etc., are controlled by a single instruction stream). Increasingly, designers are incorporating multiple ``computers,'' each with its own processor, memory, and associated interconnection logic. This approach is facilitated by advances in VLSI technology that continue to decrease the number of components required to implement a computer. As the cost of a computer is (very approximately) proportional to the number of components that it contains, increased integration also increases the number of processors that can be included in a computer for a particular cost. The result is continued growth in processor counts.

5 Trends in Networking

Another important trend changing the face of computing is an enormous increase in the capabilities of the networks that connect computers. Not long ago, high-speed networks ran at 1.5 Mbits per second; by the end of the 1990s, bandwidths in excess of 1000 Mbits per second will be commonplace. Significant improvements in reliability are also expected. These trends make it feasible to develop applications that use physically distributed resources as if they were part of the same computer. A typical application of this sort may utilize processors on multiple remote computers, access a selection of remote databases, perform rendering on one or more graphics computers, and provide real-time output and control on a workstation.

We emphasize that computing on networked computers ("distributed computing") is not just a subfield of parallel computing. Distributed computing is deeply concerned with problems such as reliability, security, and heterogeneity that are generally regarded as tangential in parallel computing. (As Leslie Lamport has observed, "A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable.") Yet the basic task of developing programs that can run on many computers at once is a parallel computing problem. In this respect, the previously distinct worlds of parallel and distributed computing are converging.

6 Summary of Trends

This brief survey of trends in applications, computer architecture, and networking suggests a future in which parallelism pervades not only supercomputers but also workstations, personal computers, and networks. In this future, programs will be required to exploit the multiple processors located inside each computer and the additional processors available across a network. Because most existing algorithms are specialized for a single processor, this situation implies a need for new algorithms and program structures able to perform many operations at once. Concurrency becomes a fundamental requirement for algorithms and programs.

This survey also suggests a second fundamental lesson. It appears likely that processor counts will continue to increase perhaps, as they do in some environments at present, by doubling each

year or two. Hence, software systems can be expected to experience substantial increases in processor count over their lifetime. In this environment, scalability resilience to increasing processor counts is as important as portability for protecting software investments. A program able to use only a fixed number of processors is a bad program, as is a program able to execute on only a single computer. Scalability is a major theme that will be stressed throughout this book.

CHAPTER 3

1 Flynn's Taxonomy

In general, digital computers may be classified into four categories, according to the multiplicity of instruction and data streams. This scheme for classifying computer organizations was introduced by Michael J. Flynn. The essential computing process is the execution of a sequence of instructions on a set of data. The term stream is used here to denote a sequence of items (instructions or data) as executed or operated upon by a single processor. Instructions or data are defined with respect to a referenced machine. An instruction stream is a sequence of instructions as executed by the machine; a data stream is a sequence of data including input, partial, or temporary results, called for the instruction stream.
Computer organizations are characterized by the multiplicity of the hardware provided to service the instruction and data streams. Listed below are Flynn's four machine organizations:

1. Single instruction stream single data stream (SISD)
2. Single instruction stream multiple data stream (SIMD)
3. Multiple instruction stream single data stream (MISD)
4. Multiple instruction stream multiple data stream (MIMD)

1.1 SISD computer organization

This organization represents most serial computers available today. Instructions are executed sequentially but may be overlapped in their execution stages.

1.2 SIMD computer organization

In this organization, there are multiple processing elements supervised by the same control unit. All PE's receive the same instruction broadcast from the control unit but operate on different data sets from distinct data streams.

1.3 MISD computer organization

There are n processor units, each receiving distinct instructions operating over the same data stream and its derivatives. The results (output) of one processor become the input (operands) of the next processor in the macropipe.

1.4 MIMD computer organization

Most multiprocessor systems and multiple computer systems can be classified in this category. MIMD computer implies interactions among the n processors because all memory streams are derived from the same data space shared by all processors. If the n data streams were from disjointed subspaces of the shared memories, then we would have the so-called multiple SISD (MSISD) operation, which is nothing but a set of n independent SISD uniprocessor systems.

The last three classes of computer organization are the classes of parallel computers.

2 A Taxonomy of Parallel Architectures

There are many ways in which parallel computers can be constructed. These computers differ along various dimensions.

2.1 Control Mechanism

Processing units in parallel computers either operate under the centralized control of a single control unit or work independently. In architectures referred to as stream, multiple data stream (SIMD), a single control unit dispatches instructions to each processing unit. Figure 2.2(a) illustrates a typical SIMD architecture. In an SIMD parallel computer, the same instruction is executed synchronously by all processing units. Processing units can be selectively switched off

during an instruction cycle. Examples of SIMD parallel computers include the Illiac IV, MPP, DAP, CM-2, MasPar MP- I, and MasPar MP-2.

Computers in which each processor is capable of executing a different program independent of the other processors are called multiple instruction stream, multiple data stream (MIMD) computers. Figure 2.2(b) depicts a typical MIMD computer. Examples of MIMD computers include the Cosmic Cube, nCUBE 2, iPSC, Symmetry, FX-8, FX-2800, TC-2000, CM-5, KSR-1, and Paragon XP/S.

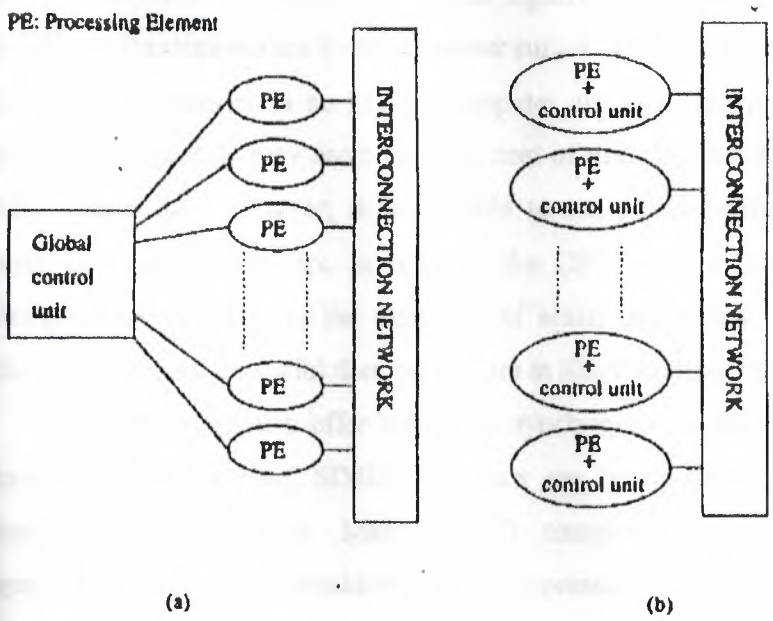


Figure 2.2 A typical SIMD architecture (a) and a typical MIMD architecture (b).

SIMD computers require less hardware than MIMD computers because they have only one global control unit. Furthermore, SIMD computers require less memory because only one copy of the program needs to be stored. In contrast, MIMD computers store the program and operating system at each processor. SIMD computers are naturally suited for data-parallel programs; that is, programs in which the same set of instructions are executed on a large data set. Furthermore, SIMD computers require less startup time for communicating with neighboring

processors. This is because the communication of a word of data is just like a register transfer (due to the presence of a global clock) with the destination register in the neighboring processor.

A drawback of SIMD computers is that different processors cannot execute different instructions in the same clock cycle. For instance, in a conditional statement, the code for each condition must be executed sequentially. This is illustrated in Figure 2.3. The conditional statement in Figure 2.3(a) is executed in two steps. In the first step, all processors that have B equal to zero execute the instruction C = A. All other processors are idle. In the second step, the 'else' part of the instruction (C = A/B) is executed. The processors that were active in the first step now become idle. Data-parallel programs in which significant parts of the computation are contained in conditional statements are therefore better suited to MIMD computers than to SIMD computers.

Individual processors in an MIMD computer are more complex, because each processor has its own control unit. It may seem that the cost of each processor must be higher than the cost of a SIMD processor. However, it is possible to use general-purpose microprocessors as processing units in MIMD computers. In contrast, the CPU used in SIMD computers has to be specially designed. Hence, due to the economy of scale, processors in MIMD computers may be both cheaper and more powerful than processors in SIMD computers.

SIMD computers offer automatic synchronization among processors after each instuction execution cycle. Hence, SIMD computers are better suited to parallel programs that require frequent synchronization. Many MIMD computers have extra hardware to provide fast synchronization, which enables them to operate in SIMD mode as well. Examples of such computers are the DADO and CM-5.

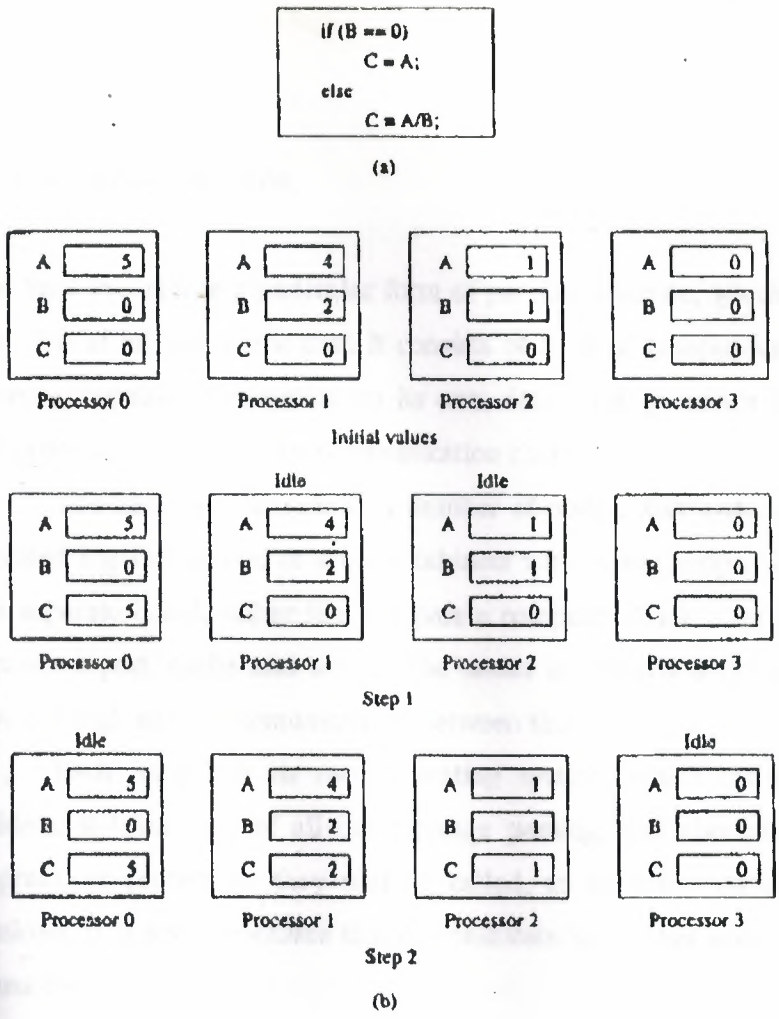


Figure 2.3 executing a conditional statement on an SIMD computer with four processors: (a) The conditional statement; (b) The execution of the statement in two steps.

3. A Parallel Machine

The Intel Paragon is a particular form of parallel machine, which makes concurrent computation available at relatively low cost. It consists of a set of independent processors, each with its own memory, capable of operating on its own data. Each processor has its own program to execute and processors are linked by communication channels.

The hardware consists of a number of nodes, disk systems, communications networks all mounted together in one or several cabinets with power supply for the whole system. Each node is a separate board, rather like a separate computer. Each node has memory, network interface, expansion port, cache and so on. The nodes are linked together through a back plane, which provides high-speed communications between them.

Each node has its own operating system, which can be considered as permanently resident. It takes care of all the message passing, and also allows more than one executable program, or process as they will be called, to be active on each node at any time. Strictly speaking, it is node processes that communicate with other node processes rather than the nodes themselves.

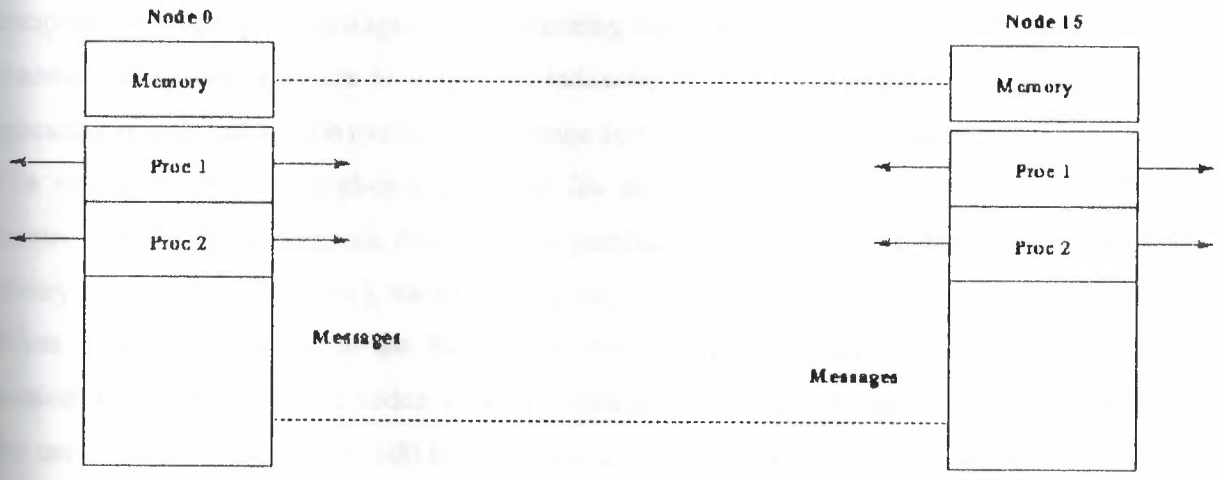


Figure 1.1: Process communication on the nodes

Remember, nodes use their own copy of the program and have their own memory allocation. No variables are shared between nodes or even between processes on the same node. Data can only be shared by sending them as messages between processes.

The Paragon supercomputer is a distributed-memory multicomputer. The system can accommodate more than a thousand heterogeneous nodes connected in a two-dimensional rectangular mesh. A lightweight MACH 3.0 based microkernel is resident on each node, which provides core operating system functions. Transparent access to file systems is also provided. Nodes communicate by passing messages over a high-speed internal interconnect network. A general-purpose MIMD (Multiple Instruction, Multiple Data) architecture supports a choice of programming styles and paradigms, including true MIMD and Single Program Multiple Data (SPMD).

We will adopt the SPMD programming paradigm (Single Program Multiple Data) i.e. each process is the same program executing on different processors. Each program executes essentially the same algorithms, but different branches of the code may be active in different processors. The general architecture of the machine is illustrated in figure 1.2.In the illustration, nodes are arranged in a 2D mesh. Each compute node consists of two i860XP processors. One of these is an application processor and the other a dedicated communication processor. User applications will normally run using the application processor. The figure illustrates that each compute node may pass messages to neighbouring nodes through a bi-directional communication channel. When messages are to be passed indirectly between non-neighbouring processors, the operating system will handle routing the message between intermediate processors.

File system support and high-speed parallel file access is provided through the nodes labelled service and I/O in the diagram. Access to the parallel file system is made through standard OSF library routines (open(), close(), read(), write(), etc.,).

When a user is logged in to the Paragon system, the operating system will allocate the login session to one of the service nodes. Exactly which service node is in use is totally transparent to the user. The user will usually edit files, and compile, link and run applications while logged in to one of the service nodes. Note also that most sites will have available a so-called cross-environment which allows most of the program development stages - editing, compiling, linking and debugging - to be carried out on a workstation away from the paragon system. Using the cross-environment is highly recommended, as the available capacity for such operations is

usually greater on a workstation than on the service nodes. Consult your local system administrator to find out how to use this facility.

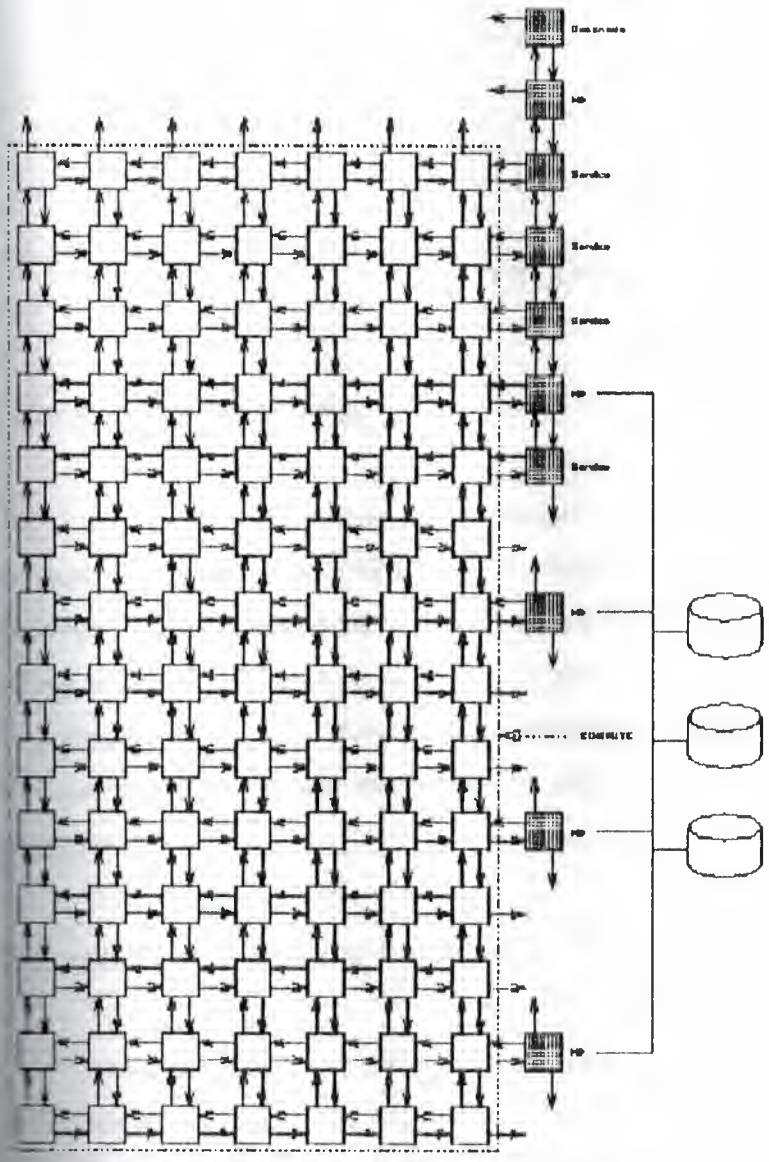


Figure 1.2: Overview of the Paragon system

CHAPTER 4

1 Parallel Programming

To run the algorithms on a parallel computer, we need to implement them in a programming language. In addition to providing all the functionality of a sequential language, a language for programming parallel computers must provide mechanisms for sharing information among processors. It must do so in a way that is clear, concise, and readily accessible to the programmer. A variety of parallel programming paradigms have been developed. This chapter discusses the strengths and weaknesses of some of these paradigms, and illustrates them with examples.

2 Parallel Programming Paradigms

Different parallel programming languages enforce different programming paradigms. The variations among paradigms are motivated by several factors. First, there is a difference in the amount of effort invested in writing parallel programs. Some languages require more work from the programmer, while others require less work but yield less efficient code. Second, one programming paradigm may be more efficient than others for programming on certain parallel computer architectures. Third, various applications have different types of parallelism, so different programming languages have been developed to exploit them. This section discusses these factors in greater detail.

2.1 Explicit versus Implicit Parallel Programming

One way to develop a parallel program is to code an explicitly parallel algorithm. This approach, called explicit parallel programming , requires a parallel algorithm to explicitly specify how the processors will cooperate in order to solve a specific problem. The compiler's task is straightforward. It simply generates code for the instructions specified by the programmer. The programmer's task, however, is quite difficult.

Another way to develop parallel programs is to use a sequential programming language and have the compiler insert the constructs necessary to run the program on a parallel computer. This approach, called implicit parallel programming, is easier for the programmer because it places a majority of the burden of parallelization on the compiler.

Unfortunately, the automatic conversion of sequential programs to efficient parallel ones is very difficult because the compiler must analyze and understand the dependencies in different parts of the sequential code to ensure an efficient mapping onto a parallel computer. The compiler must partition the sequential program into blocks and analyze dependencies between the blocks. The blocks are then converted into independent tasks that are executed on separate processors. Dependency analysis is complicated by control structures such as loops, branches, and procedure calls. Furthermore, there are often many ways to write a sequential program for a given application. Some sequential programs make it easier than others for the compiler to generate efficient parallel code. Therefore, the success of automatic parallelization also depends on the structure of the sequential code. Some recent languages, such as Fortran D, allow the programmer to specify the decomposition and placement of data among processors. This makes the job performed by parallelizing compilers somewhat simpler.

2.2 Shared-Address-Space versus Message-Passing

In the shared-address-space programming paradigm, programmers view their programs as a collection of processes accessing a central pool of shared variables. The shared-address-space programming style is naturally suited to shared-address-space computers. A parallel program on a shared-address-space computer shares data by storing it in globally accessible memory. Each

processor accesses the shared data by reading from or writing to shared variables. However, more than one processor might access the same shared variable at a time, leading to unpredictable and undesirable results. For example, assume that x initially contains the value 5 and that processor P_1 increases the value of x by one while processor P_2 decreases it by one. Depending on the sequence in which the instructions are executed, the value of x can become 4, 5, or 6. For example, if P_1 reads the value of x before P_2 decreases it, and stores the increased value after P_2 stores the decreased value, x will become 6. We can conrect the situation by preventing the second processor from decreasing x while it is being increased by the first processor. Shared-address-space programming languages must provide primitives to resolve such mutual-exclusion problems.

In the message-passing programming paradigm, programmers view their programs as a collection of processes with private local variables and the ability to send and receive data between processes by passing messages. In this paradigm, there are no shared variables among processors. Each processor uses its local variables, and occasionally sends or receives data from other processors. The message-passing programming style is naturally suited to message-passing computers.

Shared-address-space computers can also be programmed using the message-passing paradigm. Since most practical shared-address-space computers are no uniform memory access architectures, such emulation exploits data locality better and leads to improved performance for tnany applications. On shared-address-space computers, in which the local memory of each processor is globally accessible to all other processors (Figure 2.5(a)), this emulation is done as follows. Part of the local memory of each processor is designated as a communication buffer, and the processors read from or write to it when they exchange data. On shared-address-space computers in which each processor has local memory in addition to global memory, message passing can be done as follows. The local memory becomes the logical local memory, and a designated area of the global memory becomes the communication buffer for message passing.

Many parallel programming languages for shared-address-space or message-passing MIMD computers are essentially sequential languages augmented by a set of special system calls. These calls provide low-level primitives for message passing, process synchronization, process creation, mutual exclusion, and other necessary functions. Extensions to C, Fortran, and C++ have been developed for various parallel computers including nCUBE2, iPSC 860, Paragon XP/S

CM-5, TC 2000, KSR- 1, and Sequent Symmetry. In order for these programming languages to be used on a parallel computer, information stored on different processors must be explicitly shared using these primitives. As a result, programs may be efficient, but tend to be difficult to understand, debug, and maintain. Moreover, the lack of standards in many of the languages makes programs difficult to port between architectures. Parallel programming libraries, such as PVM, Parasoft EXPRESS, P4, and PICL, try to address some of these problems by offering vendor-independent low-level primitives. These libraries offer better code portability compared to earlier vendor-supplied programming languages. However, programs are usually still difficult to understand, debug, and maintain.

2.3 Data Parallelism versus Control Parallelism

In some problems, many data items are subject to identical processing. Such problems can be parallelized by assigning data elements to various processors, each of which performs identical computations on its data. This type of parallelism is called data parallelism. An example of a problem that exhibits data parallelism is matrix multiplication. When multiplying two n x n matrices A and B to obtain matrix $C = (c_{i,j})$, each element $c_{i,j}$ is computed by performing a dot product of the i^{th} row of A with the j^{th} column of B. Therefore, each element $c_{i,j}$ is computed by performing identical operations on different data, which is data parallel.

Several programming languages have been developed that make it easy to exploit data parallelism. Such languages are called data-parallel programming languages and programs written in these languages are called data-parallel programs. A data-parallel program contains a single sequence of instructions, each of which is applied to the data elements in lockstep. Data-parallel programs are naturally suited to SIMD computers.

A global control unit broadcasts the instructions to the processors, which contain the data. Processors execute the instruction stream synchronously. Data-parallel programs can also be executed on MINID computers. However, the strict synchronous execution of a data-parallel program on an MIMD computer results in inefficient code since it requires global synchronization after each instructions. One solution to this problem is to relax the synchronous execution of instructions. In this programming model, called single program, multiple data or

SPMD, each processor executes the same program asynchronously. Synchronization takes place only when processors need to exchange data. Thus, data parallelism can be exploited on an MINID computer even without using an explicit data-parallel programming language.

Control parallelism refers to the simultaneous execution of different instruction streams. Instructions can be applied to the same data stream, but more typically they are applied to different data streams. An example of control parallelism is pipelining. In pipelining, computation is parallelized by executing a different program at each processor and sending intermediate results to the next processor. The result is a pipeline of data owing between processors. Algorithms for problems requiring control parallelism usually map well onto MIMD parallel computers because control parallelism requires multiple instruction streams. In contrast, SIMD computers support only a single instruction stream and are not able to exploit control parallelism efficiently.

Many problems exhibit a certain amount of both data parallelism and control parallelism. The amount of control parallelism available in a problem is usually independent of the size of the problem and is thus limited. In contrast, the amount of data parallelism in a problem increases with the size of the problem. Therefore, in order to use a large umber of processors efficiently, it is necessary to exploit the data parallelism inherent in an application.

Note that not all data-parallel applications can be implemented using data-parallel programming languages nor can all data-parallel applications be executed on SIMD computers. In fact, many of them are more suited for MIMD computers. For example, the search problem has data parallelism, since successors must eventually be generated for all the nodes in the tree. However, the actual code for generating successor nodes contains many conditional statements. Thus, depending upon the code being generated, different instructions are executed. As shown in Figure 2.3, such programs perform poorly on SIMD computers. In some data-parallel applications, the data elements are generated dynamically in an unstructured manner, and distribution of data to processors must be done dynamically. For example, in the tree-search problem, nodes in the tree are generated during the execution of the search algorithm, and the tree grows unpredictably. To obtain a good load balance, the search space must be divided dynamically among processors. Data-parallel programs can perform data redistribution only on a global scale; that is, they do not allow some processors to continue working while other

processors redistribute data among themselves. Hence, problems requiring dynamic distribution are harder to program in the data-parallel paradigm.

Data-parallel languages offer the programmer high-level constructs for sharing information and managing concurrency. Programs using these high-level constructs are easier to write and understand. Some examples of languages in this category are Dataparallel C and C*. However, code generated by these high-level constructs is generally not as efficient as handcrafted code that uses low-level primitives. In general, if the communication patterns required by the parallel algorithm are not supported by the data-parallel language, then the data-parallel program will be less efficient.

3 Primitives for the Message-Passing Programming Paradigm

Existing sequential languages can easily be augmented with library calls to provide message-passing services. This section presents the basic extensions that a sequential language must have in order to support the message-passing programming paradigm.

Message passing is often associated with MIMD computers, but SIMD computers can be programmed using explicit message passing as well. However, due to the synchronous execution of a single instruction stream by SIMD computers, the explicit use of message passing sometimes results in inefficient programs.

3.1 Basic Extensions

The message-passing paradigm is based on just two primitives: SEND and RECEIVE. SEND transmits a message from one processor to another, and RECEIVE reads a message from another processor.

The general form of the SEND primitive is

SEND(message, messagesize, target, type, flag)

Message contains the data to be sent, and message size is its size in bytes. Target is the label of the destination processor. Sometimes, target can also specify a set of processors as the recipient of the message. For example, in a hypercube-connected computer, target may specify certain sub cubes, and in a mesh-connected computer it may specify certain sub meshes, rows, or columns of processors.

The parameter type is a user-specified constant that distinguishes various types of messages. For example, in the matrix multiplication algorithm described in Section there are at least two distinct types of messages.

Usually there are two forms of SEND. One allows processing to continue immediately after a message is dispatched, whereas the other suspends processing until the message is received by the target processor. The latter is called a blocking SEND, and the former a no blocking SEND. The flag parameter is sometimes used to indicate whether the SEND operation is blocking or no blocking.

When a SEND operation is executed, the operating system performs the following steps. It copies the data stored in message to a separate area in the memory, called the communication buffer. It adds an operating-system-specific header to the message that includes type, flag, and possibly some routing information. Finally, it sends the message. In newer parallel computers, these operations are performed by specialized routing hardware. When the message arrives at the destination processor, it is copied into this processor's communication buffer and a system variable is set indicating that a message has arrived. In some systems, however, the actual transfer of data does not occur until the receiving processor executes the corresponding RECEIVE operation.

The RECEIVE operation reads a message from the communication buffer into user memory. The general form of the RECEIVE primitive is

RECEIVE(message, message size, source, type, flag)

There is a great deal of similarity between the RECEIVE and SEND operations because they perform complementary operations. The message parameter specifies the location at which the data will be stored and message size indicates the maximum number of bytes to be put into message. At any time, more than one message may be stored in the communication buffer. These

messages may be from the same processor or different processors. The source parameter specifies the label of the processor whose message is to be read. The source parameter can also be set to special values, indicating that a message can be read from any processor or a set of processors. After successfully completing the RECEIVE operation, source holds the actual label of the processor that sent the message.

The type parameter specifies the type of the message to be received. There may be more than one message in the communication buffer from the source processor(s). The type parameter selects a particular message to read. It can also take on a special value to indicate that any type of message can be read. After the successful completion of the RECEIVE operation, type will store the actual type of the message read.

As with SEND, the RECEIVE operation can be either blocking or nonblocking. In a blocking RECEIVE, the processor suspends execution until a desired message arrives and is read from the communication buffer. In contrast, nonblocking RECEIVE returns control to the program even if the requested message is not in the communication buffer. The flag parameter can be used to specify the type of RECEIVE operation desired.

Both blocking and nonblocking RECEIVE operations are useful. If a specific piece of data from a specific processor is needed before the computation can proceed, a blocking RECEIVE is used. Otherwise, it is preferable to use a nonblocking receive. For example, if a processor must receive data from several processors, and the order in which these data arrive is not predetermined, nonblocking RECEIVE is usually better.

Most message-passing extensions provide other functions in addition to SEND and RECEIVE. These functions include system status querying, global synchronization, and setting mode for communication. Another important function is WHOAMI. The WHOAMI function returns information about the system and the processor itself. The general form of the WHOAMI function is:

WHOAMI (processorid, numofprocessor s)

Processorid returns the label of the processor, and numofprocessor s returns the total number of processors in the parallel computer. The processarid is the value used for the target and source parameters of the RECEIVE and SEND operations. The total number of processors helps

determine certain characteristics of the topology of the parallel computer (such as the number of dimensions in a hypercube or the number of rows and columns in a mesh).

Most message-passing parallel computers are programmed using either a host--node model or a hostless model. In the host-node model, the host is a dedicated processor in charge of loading the program onto the remaining processors (the nodes). The host also performs housekeeping tasks such as interactive input and output, termination detection, and process termination. In contrast, the hostless model has no processor designated for such housekeeping tasks. However, the programmer can program one of the processors to perf,orm these tasks as required.

The following sections present the actual functions used by message passing for some commercially-available parallel computers.

3.2 nCUBE 2

The nCUBE 2 is an MIMD parallel computer developed by nCUBE Corporation. Its processors are connected by a hypercube interconnection network. A fully configured nCUBE 2 can have up to 8192 processors. Each processor is a 32-bit RISC processor with up to 64MB of local memory. Early versions of the nCUBE 2's system software supported the host-node programming model. A recent release of the system software primarily supports the hostless model.

The nCUBE 2's message-passing primitives are available for both the C and Fortran languages. The nCUBE 2 provides nonblocking SEND with the use of the nwrite function.

C
```
int nwrite (char *message, int messagesize, int target, int type, int *fiag)
```

Fortran
```
integer function nwrite(message, messagesize, target, type, flag)
dimension message (*)
integer messagesize, target, type, flag
```

The functions of nwrite's parameters are similar to those of the SEND operation. The main difference is that the flag parameter is unused. The nCUBE 2 does not provide a blocking SEND operation.

The blocking RECEIVE operation is performed by the nread function.

C int nread(char *message, int messagesize, int *source, int *type, int *flag)

Fortran integer function nread (message, messsgesize, source, type, flag)
dimension reasage (*)
integer messagesize, source, type, flag

The nread function's parameters are similar to those of RECEIVE with the exception of the flag parameter, which is unused. The nCUBE 2 emulates a nonblocking RECEIVE by calling a function to test for the existence of a message in the communication buffer. If the message is present, nread can be called to read it. The ntest function tests for the presence of messages in the communication buffer.

C int ntest (int *source, int *type)

Fortran integer function ntest (source, type)
integer source, type

The ntest function checks to see if there is a message in the communication buffer from processor source of type type. If such a message is present, ntest returns a positive value, indicating success; otherwise it returns a negative value. When the value of source or type (or both) is set
to-1, ntest checks for the presence of a message from any processor or of any type. After the function is executed, type and source contain the actual source and type of the message in the communication buffer.

The functions npid and ncubesize implement the WHOAMI function.

C int npid()
int ncubesize()

Fortran integer function npid()

integer function ncubesize()

The npid function returns the processor's label, and ncubesize returns the number of processors in the hypercube.

3.3 iPSC 860

Intel's iPSC 860 is an MIMD message-passing computer with a hypercube interconnection network. A fully configured iPSC 860 can have up to 128 processors. Each processor is a 32-bit i860 RISC processor with up to 16MB of local memory. One can program the iPSC using either the host-node or the hostless programming model. The iPSC provides message-passing extensions for the C and Fortran languages. The same message-passing extensions are also available for Intel Paragon XP/S, which is a mesh-connected computer.

The iPSC's nonblocking SEND operation is called csend.

C
```
csend (long type, char *message, long messagesize, long target,
long flag)
```

Fortran
```
subroutine csend (type, message, messagesize, target, flag)
integer type
integer message (*)
integer messagesize, target, flag
```

The parameters of csend are similar to those of SEND. The flag parameter holds the process identification number of the process receiving the message. This is useful when there are multiple processes running on the target processor. The IPSC does not provide a blocking SEND operation. We can perform blocking RECEIVE by using the crecv function.

```
C         crecv (long type, char  *rnessage, long messagesize)

Fortran   subroutine crecv (type, message, messagesize)
            integer type
            integer message (*)
            integer messagesize
```

Comparing the crecv function with the RECEIVE operation, we see that the source and flag parameters are not available in crecv. However, crecv allows information about the source processor to be encoded in the type parameter. The iPSC provides nonblocking RECEIVE by using a function called irecv. The arguments of irecv are similar to crecv, with the exception that irecv returns a number that is used to check the status of the receive operation. The program can wait for a nonblocking receive to complete by calling the msgwait function. It takes the number returned by irecv as its argument and waits until the nonblocking RECEIVE operation has completed.

The iPSC functions mynade and numnodes are similar to WHOAMI. They return the label of the calling processor and the number of processors in the hypercube, respectively.

```
C         long mynode()
          long numnodes()
Fortran   integer function mynode()
          integer function numnodes()
```

3.4 CM-5

The CM-5, developed by Thinking Machines Corporation, supports both the MIMD and SIMD models of computation. A fully configured CM-5 can have up to 16384 processors connected by a fat tree interconnection network. The CM-5 also has a control network, used for operations involving many or all processors. Each CM-5 node has a SPARC RISC processor and four vector

units with up to 32MB of local memory. One can program the CM-5 using either the host-node or hostless programming models.

When the CM-5 is used in MIMD mode, it is programmed with the use of message-passing primitives that are available for the C, Fortran, and C++ languages.

The CM-5's blocking SEND function is CMMD_send_lack.

```
C           int CMMD_send_block (int target, int type, void *message,
            int messagesize)

Fortran     integer function CMMD_send_block (target, type, message, messagesize)
            integer target, type
            integer message (*)
            integer messagesize
```

The parameters of CMMD_send_block are similar to those for the generic SEND primitive. The CM-5's nonblocking SEND operation is CMMD_send_async.

```
C           CMMD_mcb CMMD_send_async (int target, int type, void *message,
            int messagasize, void (*handler) (CMMD_mcb))

Fortran     integer funotion CMMD_send_asyno (target, type, message,
            messagesize, handler)
            integer target, type
            integer message (*)
            integer messagesize, handler
```

Most of the parameters required by CMMD_send_async are similar to those required by the SEND operation. The CMMD_send_async function returns a pointer to a message control block (CMMD_mcb) after it has queued the message for transmission. The programmer is responsible for preserving the data in the buffer pointed to by message, and for freeing the

CMMD_mcb when the message has been sent. The parameter handler allows the programmer to define a handler routine that is invoked automatically when the message has been sent.

The CM-5 provides blocking RECEIVE with the CMMD_receive_block function

```
C           int CMMD_receive_blook(int source, int type, void *message,
            int messagesize)

Fortran     integer  function CMMD_receive_block (source, type, message,
            messagesize)
            integer source, type
            integer message (*)
            integer messagesize
```

A nonblocking RECEIVE operation is provided by the function CMMD_receive_async.

```
C           CMMD_mcb CMMD_receive_async (int source, int type, void *message,
            int messagesize, void (*handler) (CMMD_mcb))

Fortran     integer     function CMMD_receive_async (source, type, message,
            messagesize, handler)
            integer source, type
            integer message (*)
            integer messagesize, handler
```

The parameters of the CMMD_receive_lock and CMMD_receive_async operations are similar to those for the corresponding CMMD_send_lock and CMMD_send_async operations.

On the CM-5, the send function does not actually send the message until the destination node invokes a receive function, indicating that it is ready to receive a message. Furthermore, the CMMD send functions send no more data than the receiver has signaled it can accept. Thus, the number of bytes sent is the smaller of the number of bytes requested (that is, the messagesize of

the send function) and the number of bytes the receive function allows (that is, the messagesize of the receive function).

The CM-5 provides the functionality of WHOAMI with the functions CMMD_self_address and CMMD_partition_size. These functions return the label of the calling processor and the total number of processors.

C int CMMD_self_address()
int CMMD_partition_size()

Fortran int function CMMD_self_address
int function CMMD_partition_size

4 Data-Parallel Languages

The main emphasis of data-parallel languages is to make it easier for the programmer to express the data parallelism available within a program in a manner that is independent of the architectural characteristics of a given parallel computer. A data-parallel language has the following characteristics:

(1) It generates only a single instruction stream.

(2) It implies the synchronous execution of instructions. Hence, it is much easier to write and debug data-parallel programs, since race conditions and deadlocks are impossible.
It requires the programmer to develop code that explicitly specifies parallelism.

(3) It associates a virtual processor with the fundamental unit of parallelism. The programmer expresses computation in terms of operations performed by virtual processors. The advantage of virtual processors is that programmers need not be concerned with the number of physical processors available on a parallel computer. They simply specify how many processors they need. However, using virtual processors inappropriately may result in inefficient parallel programs.

(4) It allows each processor to access memory locations in any other processor. This characteristic creates the illusion of a shared address-space and simplifies programming since programmers do not have to perform explicit message passing.

Since data-parallel languages hide many architectural characteristics from the programmer, writing data-parallel programs is generally easier than writing programs for explicit message passing. However, the ease of programming comes at the expense of increased compiler complexity. Compilers for data-parallel languages must map virtual processors onto physical processors, generate code to communicate data, and enforce synchronous instruction execution.

4.1 Data Partitioning and Virtual Processors

In a data-parallel language, data are distributed among virtual processors. The virtual processors must be mapped onto the physical processors at some point. If the number of virtual processors is greater than the number of physical processors, then several virtual processors are emulated by each physical processor. In that case, each physical processor partitions its memory into blocks-one for each virtual processor it emulates-and executes each instruction in the program once for each of the virtual processors. For example, assume that an instruction increments the value of a variable by one and that three virtual processors are emulated by each physical processor. The physical processors execute the instruction by performing three consecutive increment operations, one for each virtual processor. These operations affect the memory blocks of each virtual processor.

The amount of work done by each physical processor depends on the number of virtual pmcessors it emulates. If VPR is the ratio of virtual to physical processors. then the work performed by each physical processor for each program instruction is greater by a factor of VPR. This is because each physical processor has to execute VPR instructions for each program instruction. However, the amount of communication performed may be smaller or larger than VPR. For instance, if the virtual processors are mapped so that neighboring virtual processors

reside on physical processors that are farther away, the communication requirements will be higher than VPR. In most cases, however, it is possible to map virtual processors onto physical processors so that nearest-neighbor communication is preserved. If this is the case, some virtual processors may need to communicate with virtual processors mapped onto the same physical processor. Depending on how smart the emulation is, this may lead to lower communication requirements.

Some data-parallel languages contain primitives that allow the programmer to specify the desired mapping of virtual processors onto physical processors. This is essential in developing efficient parallel programs. The efficiency of a mapping depends on both the data communication patterns of the algorithm, and the interconnection network of the target computer. For example, a mapping suited to a hypercube-connected parallel computer may not be suited to a mesh-connected parallel computer.

4.2 C*

C* is a data-parallel programming language that is an extension of the C programming language. C* was designed by Thinking Machines Corporation for the CM-2 parallel computer. The CM-2 is a fine-grain SIMD computer with up to 65,536 processors. Each CM-2 processor is one bit wide, and supports up to 1 Mbit of memory. C* is also available for the CM-5.

C* adheres to the ANSI standard for C, so programs written in ANSI C compile and run correctly under C*. In addition, C* provides new features for specifying data parallelism. The features of C* include the following

(1) A method to describe the size and the shape of parallel data and to create parallel variables.

(2) Operators and expressions for parallel data that provide functionality such as data broadcasting and reduction. Some of these operators require communication.

(3) Methods to specify data points within selected parallel variables on which C* code is to operate.

4.2.1 Parallel Variables

C* has two types of variables. A scalar variable is identical to an ordinary C variable; scalar variables are allocated in the host processor. A parallel variable is allocated on all node processors. A parallel variable has as many elements as the number of processors.

A parallel variable has a shape in addition to a type. A shape is a template for parallel data-a way to configure data logically. It defines how many parallel elements exist and how they are organized. A shape has a specific number of dimensions, referred to as its rank, with a given number of processors or positions in each dimension. A dimension is called an axis. For example, the following statement declares a shape called mesh, of rank two and having 1,048,576 positions:

```
shape [1024] [lO24] mesh;
```

Similarly, the following statement declares a shape of rank four with two positions along each axis:

```
shape [ 2 ] [ 2 ] [ 2 ] [ 2 ] fourcube;
```

The fourcube shape declaration declares a template containing a total of 2 x 2 x 2 x 2 = 16 positions. A shape should reflect the most logical organization of the problem's data. For example, a graphics program might use the mesh shape to represent the two-dimensional images that it is going to process. However, not all possible configurations can be declared using the shape primitive. For example, shape does not allow us to declare a triangular-shaped or a diamond-shaped mesh. However, we can do this by declaring a larger shape and using only a portion of it. For example, we can obtain a triangular shape by declaring a square shape and using only half of it.

C* does not allow the programmer to specify virtual-to-physical processor mappings explicitly. C* maps virtual processors onto physical processors so that neighboring virtual processors are mapped onto neighboring physical processors. However, C* allows us to specify across which dimensions of the shape communication will. be performed more frequently. The compiler uses this information to reduce communication cost.

After a shape is specified, parallel variables of that shape can be declared. Parallel variables have a type, a storage class, and a shape. The following statement declares the parallel variable count of type int and shape ring:

shape [8192] ring;
int: ring count;

This declaration creates a parallel variable count with 8192 positions each of which is allocated to a different processor. We can access individual elements of the parallel variable count by using left indexing. For example, [1] count accesses the value of the count that resides on the second processor (numbering is from 0 to 8191). Figure 13.1 illustrates the differences between scalar and parallel variables.

Any standard or user-defined data type can be used with parallel variables. For example, an entire C structure can be a parallel variable. As another example, int: fourcube a [1000] declares the 16-position parallel variable a, in which each element is an array of 1000 integers.

4.2.2 Parallel Operations

C* supports all standard C operations and a few new operations for data-parallel programming. In addition, C* defines additional semantics for standard C operations when they are used with parallel variables.

If the operands of an operation are scalar, then C* code behaves exactly like standard C code and the operation is performed on the host computer. The situation is different when one or more operands are parallel variables. For example, consider a simple assignment statement of the form x + = y, where both x and y are parallel variables. This assignment adds the value of y at each shape position to the value of x at the corresponding shape position. All additions take place in parallel. Note that an expression that evaluates to a parallel variable must contain parallel variables of the same shape as the resulting parallel variable. Hence, in this example, x and y must be of the same shape. In a statement of the form x = a, where a is a scalar variable, the value of a is stored in each position of x. This is similar to a broadcast operation.

A more interesting situation arises when the left side of an assignment operation is a scalar variable and the right side is a parallel variable. There are two cases in which this assignment makes sense. In the first case, the parallel variable is fully left indexed. For instance, if a is a scalar variable and x is a parallel variable of rank one, then a = [4]x is a valid statement and assigns to a the value of x at the fifth position of the shape. In the second case, the operation is one of those shown in Table 13.1. The result of these operations is a reduction. For instance, a + = x sums all the values of x and stores the result in a.

shape [1024] ring

shape [1024] [1024] mesh

int ring : a

int mesh : b

int flag

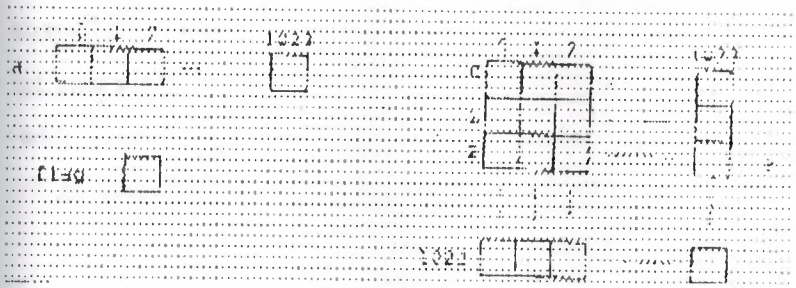


Figure 13.1 Examples of parallel and scalar variables. a and b are parallel variables of different shapes, and flag is a scalar variable. Courtesy of Thinking Machines Corporation.

Table 13.1 C* reduction operations.

Operator	Meaning
+ =	Sum of values of parallel variable elements

- =	Negative of the sum of values
&=	Bitwise AND of values
^=	Bitwise XOR of values
|=	Bitwise OR of values
<?=	Minimum of values
>?=	Maximum of values

4.2.3 Choosing a Shape

The with statement enables operations on parallel data by setting the current shape. Operations are performed on parallel variables of the current shape. In the following example, the With statement is required for performing the parallel addition:

```
shape [8192] ring;
int: ring x, y,z
 with (ring)
    x= y+z;
```

4.2.4 Setting the Context

C* has a where statement that restricts the positions of a parallel variable on which operations are performed. The positions to be operated on are called active positions. Selecting the active positions of a shape is called setting the context. For example, the where statement in the following code avoids division by zero:

```
with (ring) {
   where (z != 0)
     x  = y / z;
}
```

The where statement can include an else clause. The else clause complements the set of active positions. Specifically, the positions that were active when the where statement was executed are deactivated, and the inactive positions are activated. For example,

```
with (ring) {
  where (z != 0)
    x = y / z;
  else
    x = y;
}
```

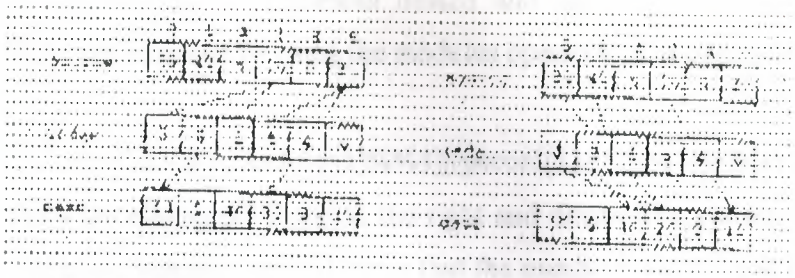
On the CM-2 (since it is an SIMD machine) the where and else clauses are executed serially. One should limit the use of the where-else clause because multiple context setfings degrade performance substantially.

4.2.5 Communication

C* supports two methods of interprocessor communication. The first is called grid communication, in which parallel variables of the same type can communicate in regular patterns. The second method is called general communication, in which the value of any element of a parallel variable can be sent to any other element, whether or not the parallel variables are of the same shape. The regularity of grid communication makes it considerably faster than general communication on many architectures. In particular, on CM-2, grid communication can be mapped onto the underlying interconnection network quite efficiently.

Data communication in C* uses left indexing, but instead of using a scalar value to left-index a parallel variable, a parallel variable is used. This operation is called parallel left indexing. A parallel left index rearranges the elements of the parallel variable based on the values stored in the elements of the parallel index. The index must be of the current shape.

```
dest = [index] source          [index] dest = source
```



A get operation A send operation

Figure 13.2 Examples of the send and get general communication operations. Courtesy of Thinking Machines Corporation.

C* allows both send and get operations. If index, dest, and source are parallel variables of rank one, the general form of the send operation is

```
[index]dest = source;
```

and the general form of the get operation is

```
dest = [index]source;
```

These operations are illustrated in Figure 13.2.

For general communication, the values of the index variable can he arbitrary. For grid communication, C* uses a new function called pcoord to provide a self-index for a parallel variable along a specified axis. In grid communication, data can be sent only a fixed distance along each dimension. For example,

```
destid = [pcoord(0)+1] source1d;
```

shifts the elements stored in sourceld by one to the right,

```
destid = [pcoord(0)-2]source1d;
```

shifts the elements by two to the left, and

```
dest2d = [pcoord(0)+1] [pcoord(1)+1]source2d;
```

shifts the elements of source2d by one to the left and up. Note that dest1d and source1d are one-dimensional shapes, whereas dest2d and source2d are two-dimensional shapes. Wraparound shifts are achieved by using the modulus operation. For example,

dest2d = [(pcoord(0)+1)%%4][(pcoord(1)+1)%%3]source2d;

shifts the elements by one to the right and down. The elements that fall off the two-dimensional shape are wrapped around. Note that the numbers 4 and 3 used in the modulus operation, are the number of positions along the corresponding axis. The operator % %' is similar to C's %' operator but works with negative values as well.

To summarize, in general we can say that data-parallel programs tend to be smaller than explicit message-passing programs. Furthermore, programs that use the virtual-processor paradigm tend to be simpler to implement.

CHEAPTER 5

NETWORK COMPUTING

Network Structure and the Remote Procedure Call Concept

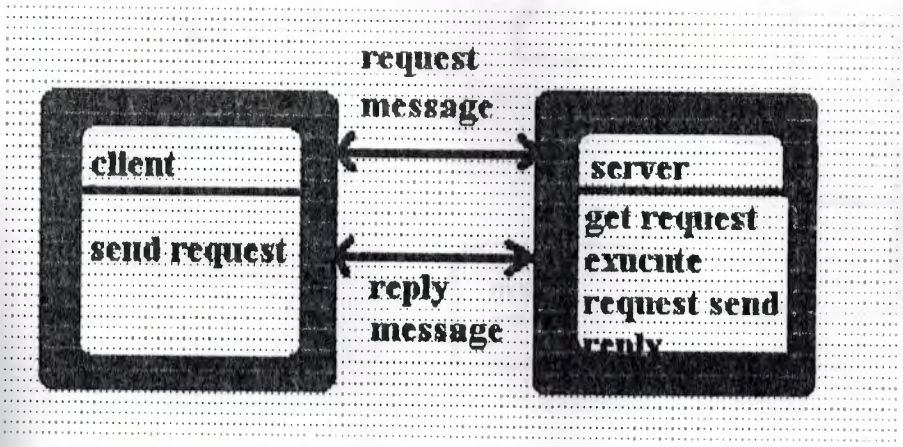
Networked computing is characterized by several sequences of jobs, which arrive independently at various nodes. The jobs are designed and implemented more or less independently of each other and are only loosely coupled. The distributed sys- tem serves primarily as a resource-sharing network.

A very common example of resource sharing is the file server. All files are located on a dedicated node in a distributed system. Software components running on other nodes send their file access requests to the file server software. The file server executes these requests and returns the results (to the clients).

In addition to file servers many other kinds of servers such as print servers, compute servers, data base servers, and mail servers have been implemented As with the file server, clients send their requests to the appropriate server and receive the results for further processing. Servers process the requests from the various clients more or less independently of each other. The programs

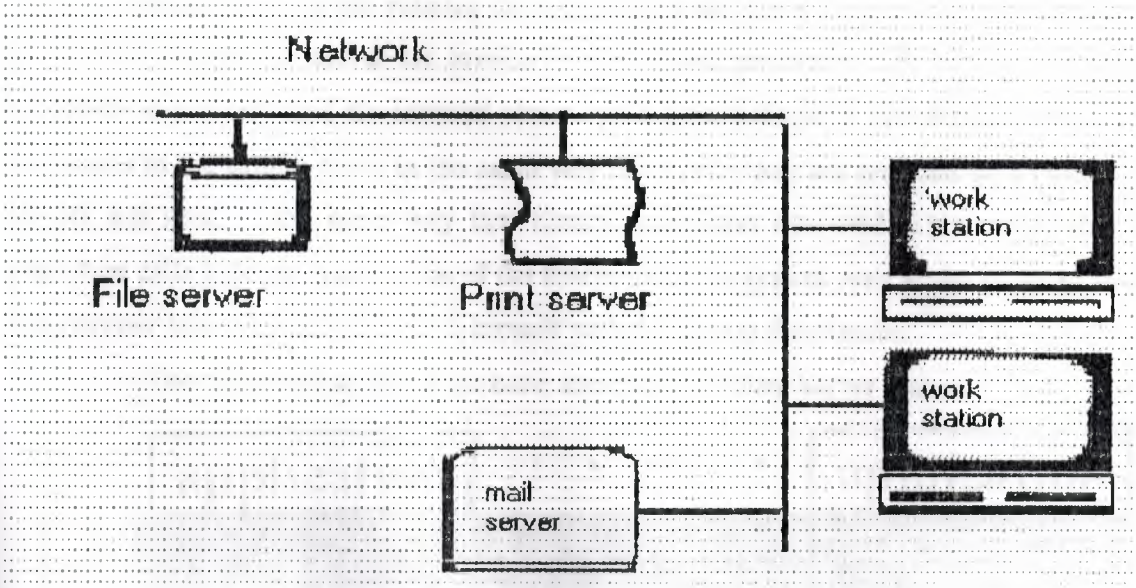
running on the clients can be viewed as being designed and developed independently of each other.

The following figure shows the concept of client server systems.



In client server system, the clients represent the users of a distributed system and servers represent different operating system functions or a commonly used application.

The following figure shows a simple example of a client server system.



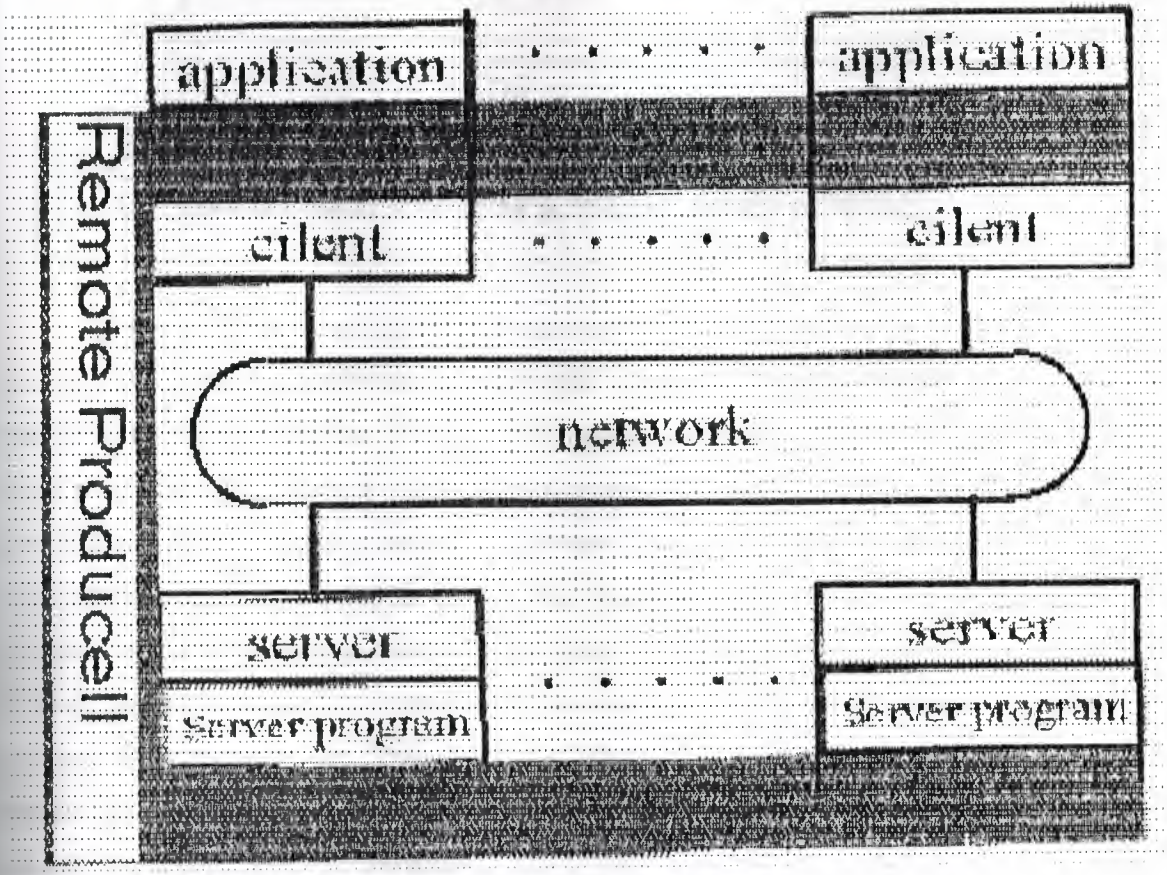
This system has a print server, a file server and the users which run on workstations and personal computers. The server software and the client software can run on the same type of computer. The different nodes are connected by a local area network.

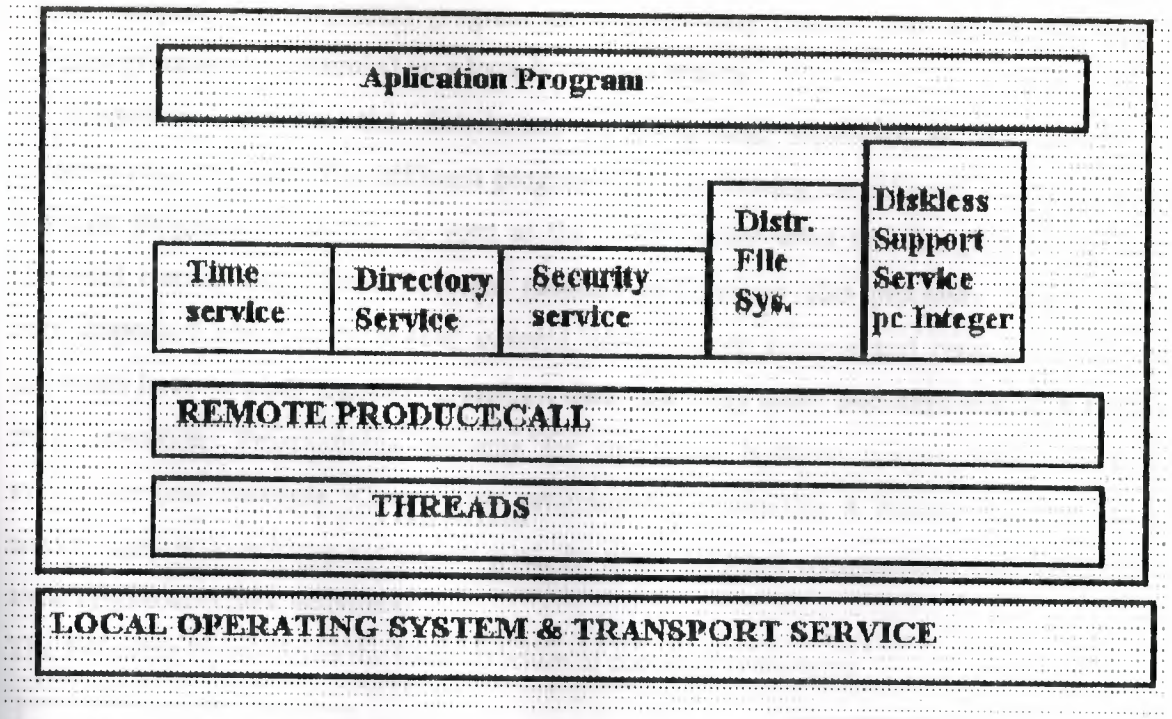
From a user's point of view a client/server system can hardly be distinguished from a central system. e.g. a user cannot see whether a file is located on his local system or on a remote file server node. For the user the client/server system appears to be a very convenient and flexible central computing system. Mostly the user does not know whether a file is stored on his PC or on a file server. To the user the storage capacity of the server appears to be part of the PC storage capacity.

Client/server systems are also very flexible. For a new application a specialized new server can be added e.g. data base systems run on specialized data base servers, which have short access times. The local client primarily controls data base applications; all the data is stored at the data base server and special computations are executed by a compute server (also called number cruncher). The application program running on the client, calls the required functions provided by the servers. This is done mainly by way of remote procedure calls (RPC). An RPC resembles a procedure call except that it is used in distributed systems. The following is a description of how

the RPC works. The program running on the client looks like a normal sequential program. The services of a particular server are invoked via a remote procedure call. The caller of a remote procedure is stopped until the invoked remote procedure is finished and the server has provided the results to the calling client in the same way that parameters are returned by a procedure. The servers are used in the same way that library procedures are used. This means that remote procedure calls hide the distribution of the functions of the system even at the program level. The programmer does not need to concern himself with the system distribution.

The figure below shows the basic structure of a client/server system.





In the DCB client and server programs are executed by threads i.e. processes. Threads use an RPC in order to communicate with each other and binary semaphores and conditional variables for synchronization. In the DCE remote procedure calls are supported by directory services (DCE Call Directory Service) and security services (DCE Security Service). Directory services map logical names to physical addresses. If a client calls a particular service provided by a server, the directory service is used to find the appropriate server. The DCE security service provides features for secure communication and controlled access to resources. Distribute Time Service provides precise clock synchronization in a distributed system. This is required for event logging, error recovery, etc. The distributed file service allows the sharing of files across the whole system. Finally the diskless support service allows workstations to use background disk files on file servers as if they were local disks

/SCHILL93/, /05F92/.

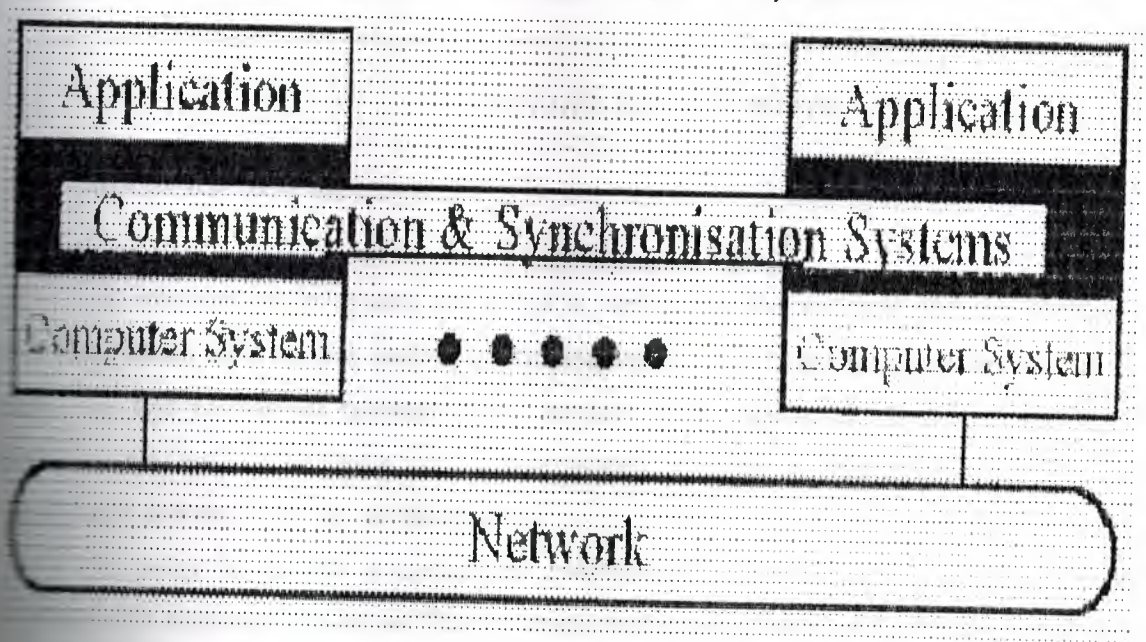
Cooperative Computing

In cooperative computing a set of processes runs on several processing nodes. These processes cooperate to reach a common goal and together they form a distributed program. This is different from the client/server systems described above. In cooperative systems the processes

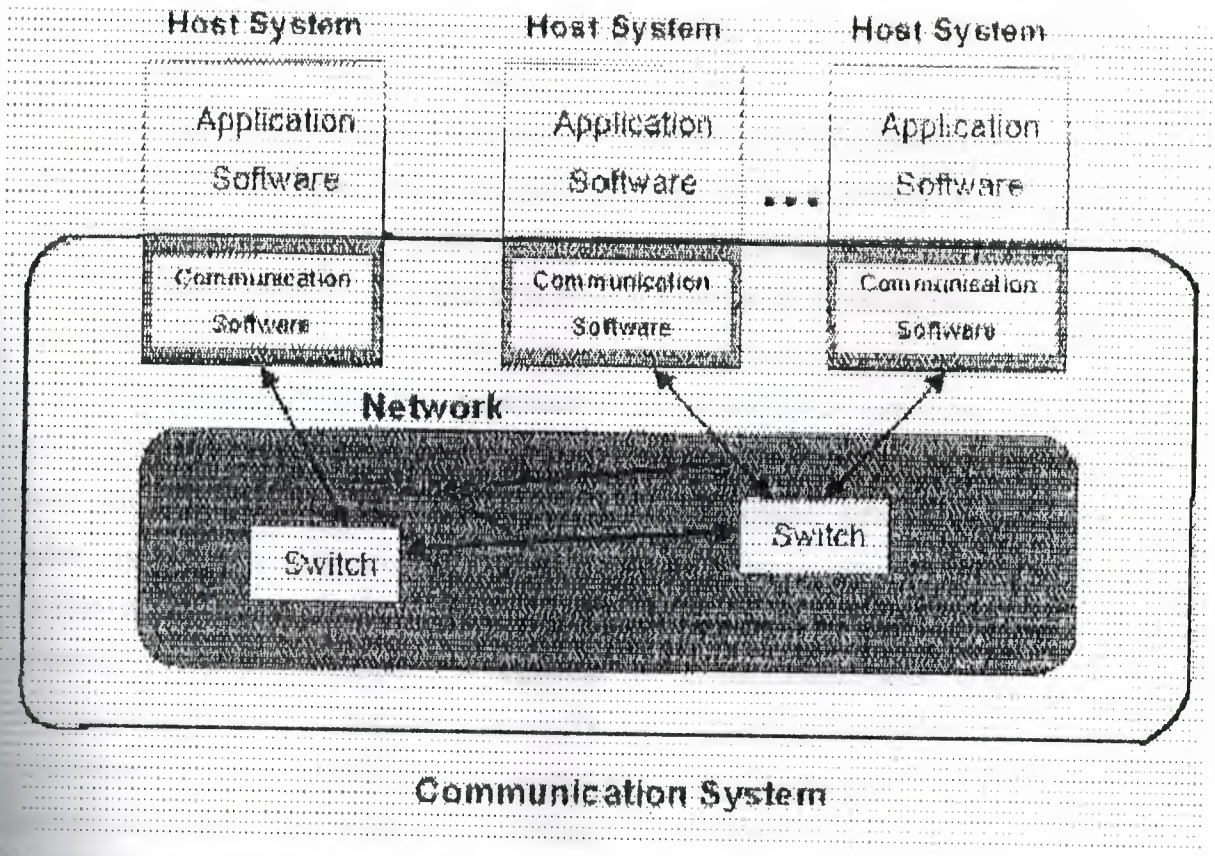
which comprise the distributed program are coupled very closely. This means that the closely coupled processes are executed on a loosely coupled system.

In cooperative systems, the distribution of computing capability is not hidden behind programming concepts. The different program sections running on different computers comprise a single program; but it can be seen at the programming level that the program sections are executed concurrently. These different program sections are also processes. Processes form a very important concept for central systems, client server systems and cooperative systems. If processes have to work together to perform their task, they must exchange data and synchronize their execution. Programming Systems for concurrent Systems contain communication and synchronization concepts. Cooperative programming resembles a human organization which works together to achieve a common goal. Its members must communicate with each other and must synchronize their activities.

The following figure shows the basic structure of cooperative Systems



Cooperative systems are mainly used for the automation of technical process and the implementation of communication software. Technical process in the mostly part consists of several parallel activities. This means that several processes, which can be implemented in different ways, work together to perform their task.



Communication Software Systems

A communication system consist of a communication network and the communication software which runs on the various processing nodes. The communication software provides a more are less convenient communication service for the application software. The application software on each node uses the communication service to exchange messages with the application software running on other nodes.

In order to provide a convenient communication service the communication software systems also exchange messages. This message exchange is based on the simpler communication mechanism provided directly by the network. For example the network provides a communication service which only allows the transfer of a single byte. The communication service provided by the communication software allows byte strings of a fixed or even an unlimited length to be sent or received. This can be implemented in the following way: The application software of a host system A wants to send a sequence of bytes to the application software of a host system B. The sequence of bytes is given to the communication system but the

application system. The communication system on host system A sends a byte with the length of the byte string (the number of bytes) to the communication system on host system B. The communication system on host system B sends back an acknowledgement. This is a byte with a certain value. After the communication software on host system A has received the acknowledgement it starts to transfer the bytes of the byte string. then system B has received the number of bytes indicated in the first byte it again sends an acknowledgement. After sending the acknowledgement. the communication software on host system B gives the received byte string to the application software.. This communication sequence which implements the transfer of a byte string just a simplistic illustration of what communication software can do. As the example above shows. the communication between the communication software systems follows well defined rules. These rules are called protocols the need to provide convenient communication services for the application software leads to software communication protocols which can he extremely complex and must be organized in layers Each layer offers an improved communication service to the layer above. The widely used reference model for open system interconnection (OSI) defined by the International Standard Organization (ISO) pro- poses seven protocol layers /IS07498/. Each layer provides a certain service to the layer above. The service provided by a layer is implemented by the protocol specific to its layer and byte services of the layer below. In a host system the services specific to the layer are realized by protocol entities. The layer protocol is defined between protocol entities of the same layer. These exchange information by using the service of the layer below. In each lost system there must he at least one entity per layer. The set of entities of different layers in a host system is called a protocol stack. The implementation of these protocol stacks is called communication software. Communication software has the following execution properties /DROB 86/:

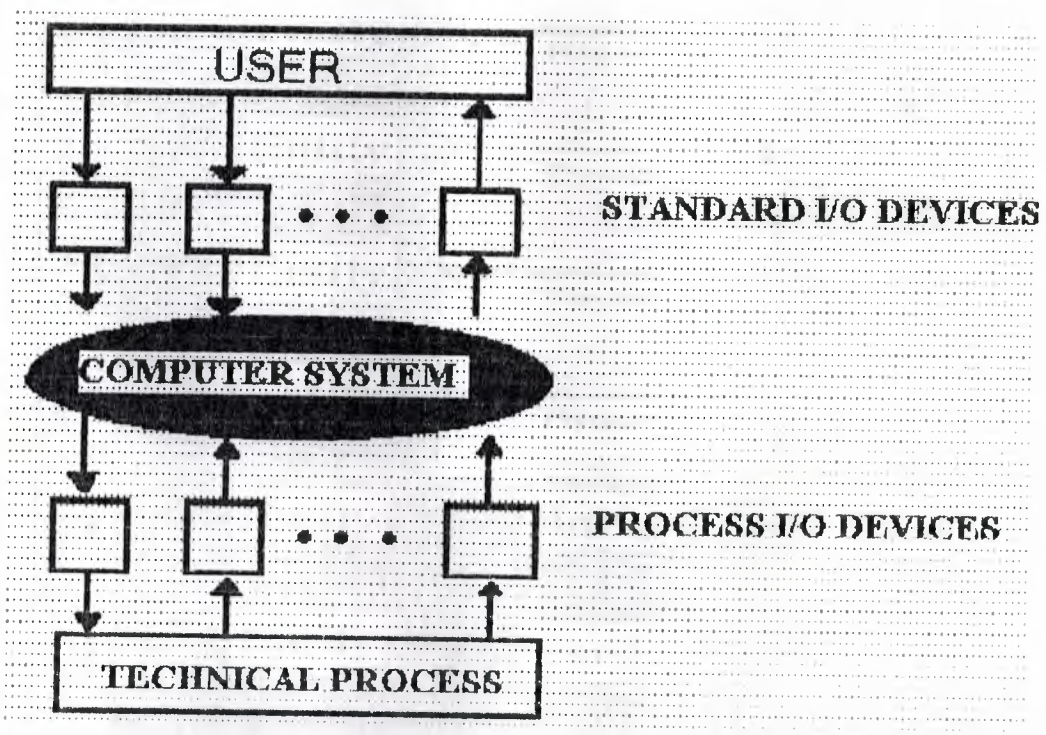
· Interleaved execution of several entities on the same system

· Distributed execution of entities of the same layer on different systems. Interleaved and distributed computations are usually' modeled as systems of parallel processes.

Processes executing in parallel normally have to exchange information if they are to by one cooperate in solving a common task. One processes model entities. Representing or providing a service means exchanging information with processes representing entities of the layer below or above. The figure above shows.

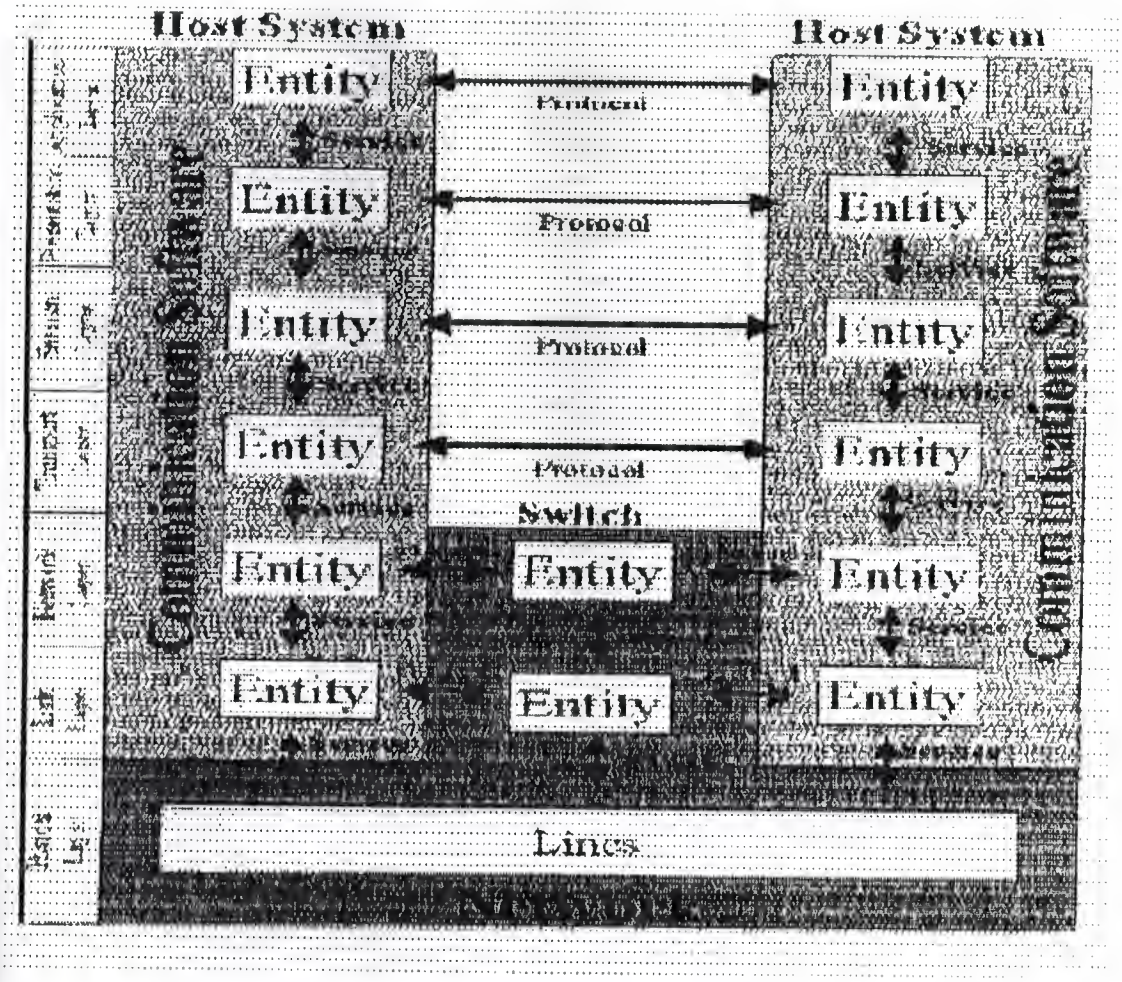
Technical Process Control Software Systems

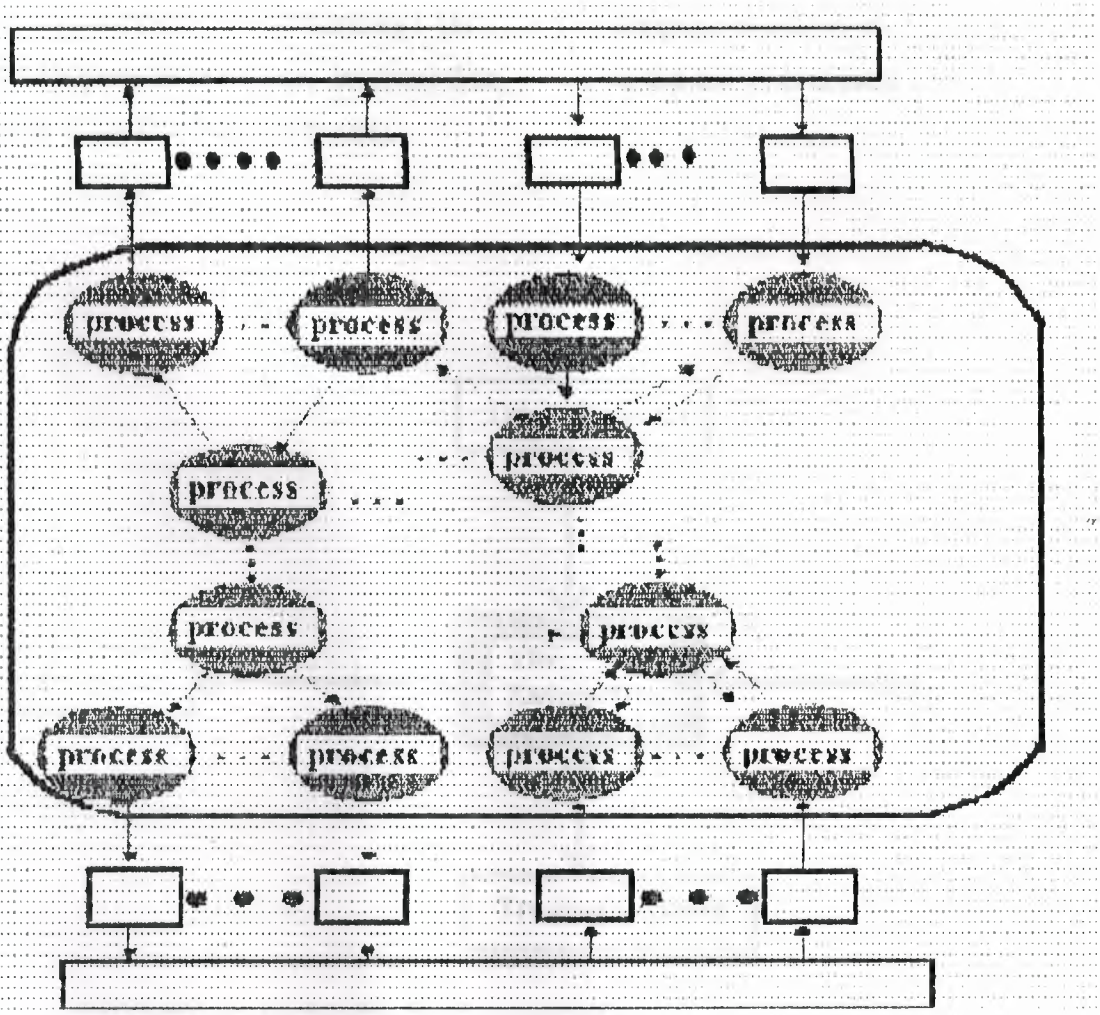
Another important example of cooperative computing is a distributed technical process control system. The basic structure of technical systems controlled systems is shown in the following figure /NEHM84/.



The communication between computer systems and technical systems must meet hard real-time requirements, whereas the communication with the user is more or less dialogue-

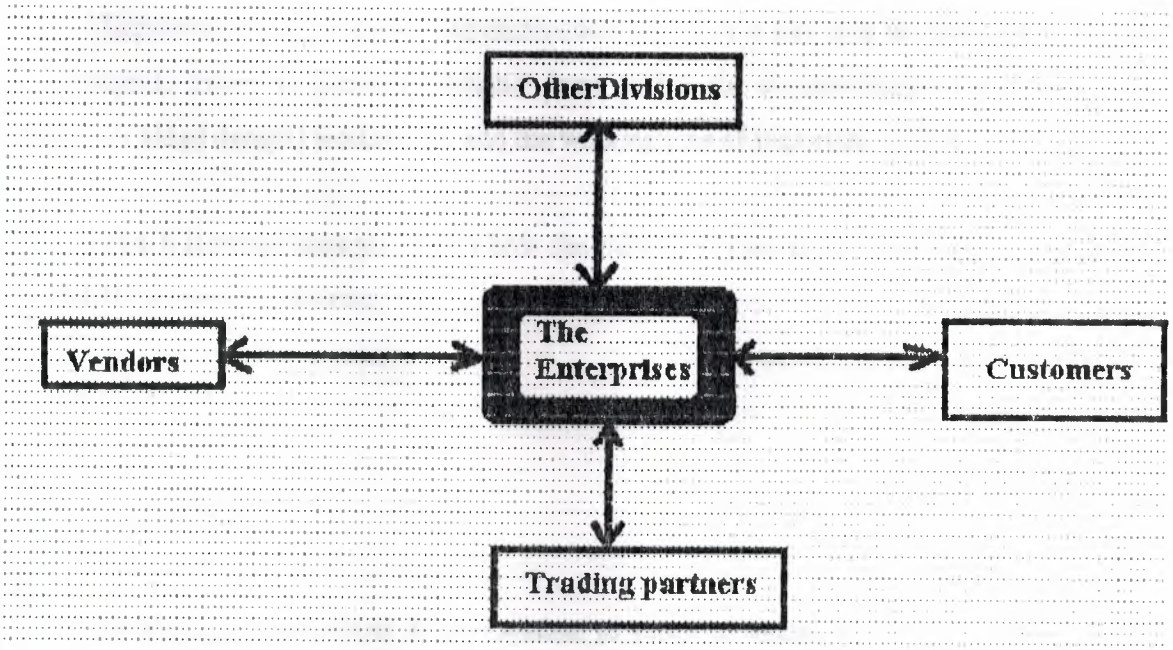
oriented with less emphasis on time conditions (except in the case emergency signals such as fire alarms). For the sake of simplicity, we will focus on the relationship between technical Systems and real-time computer systems. A technical system consists of several mutually independent functional units, which communicate via appropriate interfaces with the computer System. Therefore the real time program must react to several simultanous inputs. This implies the structuring of a process control software system that takes into account a number of processes. Each process handles a certain group of signals. The basic requirement for a process control software system is the capability to follow the changes of the technical system as fast as possible. The information in the process control software must be as close as possible to the state of the technical system. The easiest way to achieve this is to design a process for each interface element. This leads to the software system structure shown in the following figure INEHM84/.





1.3.5 ELECTONIC DATA INTERCHANGE(EDI)

Electronic Data Interchange (EDI) is the computer-to-computer exchange of inter- and intracompany technical and business data, based on the use of standards /DIGIT90/ (see figure below of the EDI business model).



These data can be structured or unstructured. Exchanging unstructured data follows specific communication standards although the data content is not in a structured format. More important is the exchange of structured data. Examples of structured data exchange are:

-Trade Data Interchange

This type of EDI document exchange is mainly used to automate business processes. Examples of trade data interchanges include a request for quotation (RfQ), purchase orders, purchase order acknowledgements, etc. Each company and industry has its own requirements for the structure and contents of these documents. A number of specific industry and national bodies have been formed with the intention of standardising the format and content of messages. For the chemical industry CEFIC is the EDI standard and for the auto industry the related EDI standard is

called ODETTE. The standard defined by CCITT is called EDIFACT. In order to exchange EDIFACT documents very ofien the CCITT E-Mail standard X.400 is recommended /LilLL9O/.
- Electronic Funds Transfer Payment against invoices, electronic point of sale (EPOS) and clearing systems are examples of electronic funds transfer.

- Technical Data Interchange

Improvement in technical communication can play a key role in determining the success of a project. There is growing demand from trades for communication between their CAD (computer aided design) workstation and the workstations of important vendors.

The following example shows how the different types of EDI interactions are used to handle a business process.

Groupware

In organizations people work together to reach a common goal. The formal interaction between members of an organization is described by structures and procedures. Additionally there exist informal interactions, which are very important. Both types of interactions can and should be supported by computers. Computer Supported Cooperative Work (CSCW) deals with the study and development of computer systems called groupware, which purpose it is to facilitate these formal and informal interactions /ENGLEH88/.

CSCW projects can be classified into four types /ENGLEHB8/ namely:

1. Groups which are not geographically distributed and require common access in realtime
Examples: presentation software, group decision systems
2. Groups which are geographically distributed and require common access in realtime
Examples: video conferencing, screen sharing

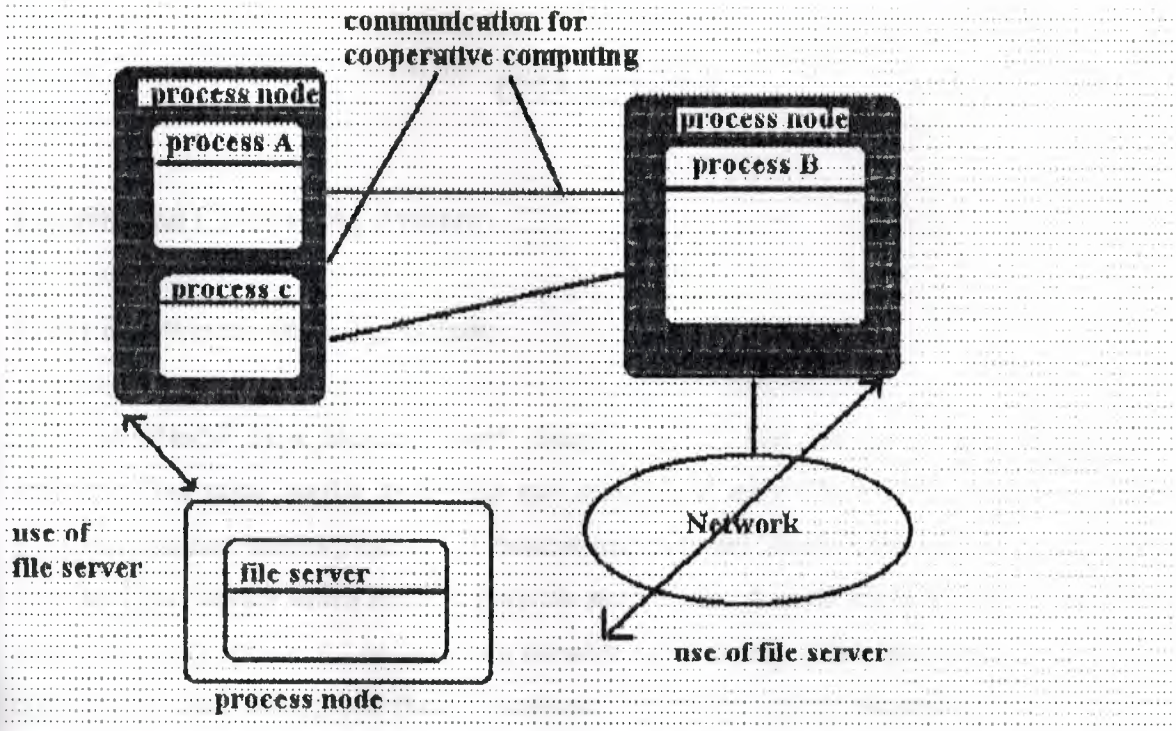
3. Asynchronous collaboration among people who are geographically distributed. Examples; notes conferences, joint editing

4. Asynchronous collaboration among people who are not geographically distributed Examples: project management, personal time schedule management

Groupware requires computers connected by a network. Thus groupware systems are distributed systems. Members of a group share data and exchange messages. Therefore groupware software systems are combinations of network and cooperative computing.

Combination of Network Computing and Cooperative Computing

Cooperative computing can be combined with client server systems. Processes in a distributed system can have access to servers. From the standpoint of a client server system the processes of a cooperative system can be considered as client processes. In a technical process control software system a process can collect data from the technical process. This data is stored in a file located on a file server node. The following figure shows an example of a combination of a cooperative and a client/ server system. Process A. Process 13 and Process C form a cooperative software system. Process B and Process C use the server. This means that process B and process C are clients of the file server.



Distributed Computing System

A distributed computing system is not yet Noema. Niany of the component are present but some are still missing or not fully integrated. The network would be the communication mechanism for the distributed computing Noema supporting message passing, protocols, and asynchronous communication. The languages of communication are the protocols built up with bytes of data. Replication and groups of services could be made available with special name space management services available on the network. Some information may be kept in a data warehouse for analysis. Some information could be locally cached. Some functions could be pre-evaluated and stored in anticipation of usage. Both code and data may have a common representation. Thus programs are to be treated as data in some cases and programs in other cases. Not all data can be interpreted as a program. The distributed computing Noema would need a security system with authentication, authorization, and data privacy. The next chapters define how to build a distributed computing Noema.

CHEAPTER 6

Distributed Computing System

In our distributed computing system:

A "Node" is a Network-User* Interface (NUI) that provides network access to the WWW*. This node maybe as simple and economical as a "JavaTerm", which has a decent processor, limited memory/cache, I/O devices and optional pheripherals such as CD ROM, hard disk, an input device which handles portable storage etc.. A node could also be a terminal, such as a UNIX workstation, PC or Mac with network capabilities*. Their processing storage and local applications may differ, but their operations should be mostly dependent on their network bandwidth (which network service providers, such as PacTel, MCI provide) and the pipe of the servers (end-service providers).

A ''Server'' is a computer that provides services interactively. Services include providing executables (e.g. we may remotely load Word and run it in our network interface), database or search engine (e.g Component library of TI), banks, stockbroker firms or any entity that handles and processes requests.

A "Site" is a network destination that provides non-interactive information. For example, most people/organization's home page nowadays which contains visual display only and does not accept/require user input is merely a site.

What differentiates a Server from and a Site is: a server is interactive "active" while a site is Serena (aka wleung) argues that the above two could/should be grouped together and called sites, while another definition of Server should be formulated.

During the last group meeting (10/12), Professor Newton mentioned that there could/should be something between a node and servers. This intermediary could be:

1) State Manager

State Manager manages things that doesn't fit into the cache, it could be handled by a central "Service Provider"* which interacts with other servers/sites. However, this would present a major security problem; that's to believe that a "Service Provider" would ensure security of clients' data from internal and external access. (Maybe digital signatures would be required to access and retrieve unscripted data, or maybe inscription could be done at the clients or over the network) There would also be a durability problem. What happens when a State Manager goes down? If we have mirror images, then consistency and security problems arise and this all leads us to the ultimate debate of how distributed systems should be architect. As for network main memory and mirror sites, administration problems immediately come up to my mind. How can they be administered, monitored and by whom? How can data security be provided for this virtual object?

My argument is that none of these intermediate objects should exist, i.e. nodes should interact directly with servers (present model of WWW). At today's price and technology curve, pockets-sized DRAM or hard disk at an acceptable price, performance and capacity (>=500MB) is imminent. One might argue that 500MB is not a lot of storage. That's because in today's standards, people store executables in their hard disks, but in the future, all people need is their personal documents (e.g. word-processing files, database, spread-sheets, etc.) that they (regularly) edit as executables will be run off the Net. As for large audio, video files and graphically intense operations such as CAD or games, they should stay at their respective servers where an adequate bandwidth and special transmission mechanisms are provided.

State management in this case is done either on a local storage (cache or hard disk) and/or at the server. Less consistency concerns is achieved at the expense of a higher response time for applications (updates need to go as far as the server instead of an intermediate node).

The Future

Microsoft's dominance of local processing will be displaced by major database and database tools (e.g. Oracle, Informix) companies together with software vendors that develop

network-based applications that run at the servers, aimed at providing high throughput, scalability, etc.

Hardware vendors, such as Cisco and Bay Networks will be a force as well in helping clients design and implement the appropriate network/WAN strategies.

Footnote

1) A User may be a human being, processes or other computers.

2) WWW may include or be a part of the Information Superhighway.

3) If "Everything" (from mail to Word, Quicken) is run within a network interface, Would CPU processing power and speed be relevant in the future, or this will be a Hardware issue that primarily interests "Server" side of the operations. Primary end-user Concern would be network bandwidth and display capabilities.

4)"Service Provider" could be network services providers such as Pastels or software vendors such as Oracle.

HORUS: A Flexible Group Communications System

Computing represents a promising step towards robustness for mission-critical distributed applications. Process replicated for availability or as part of a coherent cache. They can been used to support highly available security domains. And group mechanisms fit well an emerging generation of intelligent network and collaborative work applications.

Yet there is little agreement concerning how process groups should look or behave. The requirements that applications place on a group infrastructure can vary tremendously, and there may be fundamental tradeoffs between semantics and performance. Even the most appropriate way to present the group abstraction to the application depends on the setting.

This paper reports on the Horns system, which provides an unusually flexible group communication model to application-developers. This flexibility extends to system interface the

properties provided by a protocol stack, and even the configuration of Horus itself{which can run in user space, in an operating system kernel or micro kernel or be split between them.

Horns can be used through any of several application interfaces. These include too I Kit styled interfaces, but also interfaces that hide group functionality behind Unix communication system-calls, the TK/TCL programming language, and other distributed computing constructs. The intent is that it be possible to be slide Horus beneath an existing system as transparently as possible, for example to introduce fault-tolerance or security without requiring substantial changes to the system being hardened.

Horns provides efficient support for the virtually synchronous execution model. This model was introduced by the Isis Toolkit, and has been adopted with some changes by such systems as Tran sis, Synch, Trans/Total, and Rampant Rampart. The model is based on group membership and communication primitives, and can support a variety of faculty-tolerant tools, such as for load-balanced request execution, fault tolerant computation, coherently replicated data and security.

Although often-desirable properties like virtual synchrony may sometimes be unwanted, introduce unnecessary overheads, or conflict with other objectives such as real-time guarantees. Moreover, the optimal implementation of a desired group communication property sometimes depends on the runtime environment. In an insecure environment, one might accept the overhead of data encryption, but wish to avoid this cost when running inside a firewall. On a platform like the IBM SP2, which has reliable message transmission, protocols for message retransmission would be superfluous.

Accordingly, Horus provides an architecture whereby the protocol supporting a group can be varied, at runtime, to match the specific requirements of its application and environment.

It does this using a structured framework for protocol composition, which incorporates ideas from systems such as the Unix "streams" framework and the x-kernel, but replaces point-to point communication with group communication as the fundamental abstraction. In horns group communication stacking protocol modules that have a regular architecture and in which each module has a separate responsibility provides support. Dynamically including or excluding particular modules from its protocol stack can optimize a process group.

Horus also innovates by introducing run-time configuration, group communication interfaces full thread-safety, and supporting messages that may span multiple address spaces.

Since horus does not provide control operations and has one single address format, protocol layers can be mixed and matched freely. In both streams and the x-kernel, the different protocol modules supply many different control operations, and design their own address format, both severely limiting such configuration flexibility.

1- A LAYERED PROCESS GROUP ARCHITECTURE

We find it useful to think of horus central protocol abstraction as resembling a Lego block, the hours "system" is thus like a "box" of Lego blocks. Each type of block implements a micro protocol that provides a different communication feature. To promote the combination of these blocks into macro protocols with desired properties, the blocks have standardized top and bottom interfaces that allows them to stacked on top of each other at run time in a variety of ways. Obviously, not every sort of protocol block makes sense above or below every other sort. But the conceptual value of the architecture is that where it makes sense to create a new protocol by restacking existing blocks in a new way, doing so is straightforward.

Technically, each horus protocol block is a software module with a set of entry points for down call and up call procedures. For example there is a down call to send a message and an up call to receive a message. Bach layer is identified by an ASCII name and registers its up call and down call handlers at initialization time. There is a strong similarity between horus protocol blocks and object classes in an object-oriented inheritance scheme and readers may wish to think of protocol blocks as members of a class hierarchy.

To see how this works, consider the horus message-send operation. It looks up the message send entry in the topmost block and invokes that function. This function may add a header to the message and will then typically invoke message-send again. This time control passes to the message send function in the layer below it. This repeats itself recursively until the bottom most block is reached and invokes a driver to actually send the message.

The specific layers currently supported by horus solve such problems as interfacing the systems to varied communication transport mechanisms overcoming lost packets encryption and decryption ,maintaining group membership helping a process that joins a group obtain the state of the group merging a group that has partitioned, flow control, e.tc. Horus also includes tools to assist in the development and debugging of new layers.

the group merging a group that has partitioned, flow control, e.tc. Horus also includes tools to assist in the development and debugging of new layers.

Bach stack of block is carefully shielded from other stacks. It has its own prioritized threads, and has controlled access to available memory through a mechanism called memory channels. Horus has a memory scheduler that dynamically assigns the rate at which each stack can allocate memory depending on availability and priority so that no stack can monopolize the available memory. This is particularly important inside a kernel, or if one of the stacks has safe real-time requirements.

Besides threads and memory channels each stack deals with three other types of objects: end points, groups, and messages. The endpoint object models the communicating entity Depending on the application it may correspond to a machine, a process, a thread ,a socket, a port ,and so forth. An endpoint has an address and can send receive messages. However as we will see later messages are not addressed to endpoint but to groups. The endpoint address is used for membership purposes.

It does this using a structured framework for protocol composition, which incorporates idea from systems such as the Unix "streams" framework and the x-kernel, but replaces point-to point communication with group communication as the fundamental abstraction. In horus group communication support is provided by stacking protocol modules that have a regular architecture and in which each module has a separate responsibility. Dynamically including or excluding particular modules from its protocol stack can optimize a process group.

Horus also innovates by introducing run-time configuration, group communication interfaces full thread-safety, and supporting messages that may span multiple address spaces. Since horus does not provide control operations and has one single address format, protocol layers can be mixed and matched freely. In both streams and the x-kernel, the different protocol modules supply many different control operations, and design their own address format, both severely limiting such configuration flexibility..

A group object is used to maintain the local protocol state on an endpoint. Associated with each group object is the group address to which messages are sent and a view a list of destination endpoint addresses that are believed to be accessible group members. Since a group object is purely local , horus technically allows different views of the same group. An endpoint

may have multiple group objects allowing it to communicate with different groups and views. A user can install new views when processes crash or recover and can use one of several membership protocols to reach some form of agreement on views between multiple group objects in the same group.

Horus provides a large collection of micro protocols. Some of the most important ones are:

Proposed Sidebar

Com The COM layer provides the horus group interface to such low-level protocols as IP,UDP, and some ATM interface.

NAK- This layer implements a negative acknowledgement based message retransmission protocol.

CYCLE-Multimedia message dissemination

PARCLD- Hierarchical message dissemination

FRAG-Fragmentation/reassembly.

MBRSHIP- This layer provides each member with a list of end points that are believed to be accessible. It runs a consensus protocol to provide it users with a virtually synchronous execution model.

EC- Flow Control

TOTAL- Totally ordered message delivery.

STABLE- This layer detect when a message has been delivered to all destination endpoints, and can be garbage collected.

CRYPT- Encryptions/ denyption

MERGE- Location and merging of multiple group instance.

The message object is a local storage structure. It is interface includes operations to push and pop protocol headers. Message are passed from layer by passing a pointer and never need be copied.

A thread at the bottom most layers waits for message arriving on the network interface. When a message on to the layer above it. This repeat itself recursively. If necessary a layer may

drop a message or buffer it for delayed delivery. When multiple messages. However since each message is delivered using its own thread, this ordering may be lost depending on the scheduling policies used by the thread scheduler. Therefore, horus numbers the message and uses event count synchronization variables to reconstruct the order where necessary.

2-Protocol Stacks

The micro protocol architecture of horus would not be of great value unless the various classes of process group protocols that we might wish to support can be significant functionality. Our experience in this regard has been very positive.

The layers FRAG,NAK and COM respectively break large messages into smaller ones, overcome packet loss using negative acknowledgements and interface .Hour to the underlying transport protocols. The adjacent stack is similar, but provide weaker ordering and includes a layer that supports "state transfer "to a process joining a group or when groups merge after a network partition To the right is a stack that supports scaling through a hierarchical structure in which each parent process is responsible for a set of "child" processes. The dual stack illustrated in this case represents a feature whereby a message can be routed down one of several stacks, depending on the type of processing required. Additional protocol blocks provide functionality such as data encryption packing small messages for efficient communication, isochronously communication.

Layered protocol architectures sometimes perform poorly. Traditional layered systems impose an order on which protocols process messages limiting opportunities for optimization and imposing excessive overhead. Clack and Tennenhouse have suggested that the key to good performance rests. Systems based on the ILP principle avoid inter-layer ordering constraints and can perform as well as monolithically structure system.

3-Using Horus to build a robust groupware application

Earlier we commented that horus can be hidden behind standard application programmer interfaces. A good illustration of how this done arose when we interfaced the graphical programming language to horus.

A challenge posed by running systems like horus side with a package like windows.

That such packages are rarely designed with threads or horus communication stacks in mind .To avoid a complex integration task.

Architecturally, CMT consists of a multi-media server process that multicasts video and audio to a set of clients. We decided to replicate the server using a primary -backup .approach. Where the backup servers stand by to back up failed or slow primaries.

4-Electra

The information of process groups into CMT required sophistication with horus and its intercept proxies. Many potential users would lack the sophistication and knowledge required to do this hence we recognized a need for a way to introduce horus functionality in a more transparent way. This goal evokes an image of "plug and plug" robustness, and leads one to think in terms of an object-oriented approach computing.

The common object request broker architecture (CORBA) is emerging as a major standard for supporting object-oriented distributed environments. Object-oriented distributed applications that comply with CORBA can invoke one-another methods with relative ease. Our work resulted in a CORBA compliant interface to horus which we call Electra can be used without horus, and vice versa ,but the combination represents a more complete system.

CONCLUSION

The increasing density of transistors on a chip follows directly from a decreasing feature size, which is now for the alpha. Feature size will continue to decrease and by the year 2000, chips with 50 million transistors are expected to be available. What can we do with all these transistors? With around a million transistors on a chip, designers were able to move full mainframe functionality to about of a chip. This enabled the personal computing and workstation revolutions. The next factors of ten increase in transistor density must go into some form of parallelism by replicating several CPUs on a single chip.

By the year 2000, parallelism is thus inevitable to all computers, from your children's video game to personal computers, workstations, and supercomputers. Today we see it in the larger machines as we replicate many chips and printed circuit boards to build systems as arrays of nodes, each unit of which is some variant of the microprocessor. Parallelism allows one to build the world's fastest and most cost-effective supercomputers.

Parallelism may only be critical today for supercomputer vendors and users. By the year 2000, all computers will have to address the hardware, algorithmic, and software issues implied by parallelism. The reward will be amazing performance and the opening up of new fields; the price will be a major rethinking and re-implementation of software, algorithms, and applications.

REFERENCES

NETWORK COMPUTING

(Joel M. Crichlow/Springer-Verlag)

" INTRODUCTION TO PARALLEL COMPUTING "

[(Vipin Kumar, Ananth Grama, Anshul Gupta, George Karypis / Univ. of Minnesota)

The Benjamin/Cummings Publishing Company Inc.

Copyright© 1994 by The Benjamin/Cummings Publishing Company Inc.]

" PARALLEL PROCESSING "

[(M.E.C. Hull / Univ. of Ulster, D. Crookes / The Queen's Univ., Belfast, P.J. Sweeney / Univ. of Ulster) Addison-Wesley Publishing Company

Copyright© 1994 Addison-Wesley Publishers Ltd.

Copyright© Addison-Wesley Publishing Company Inc.

" LECTURE NOTES ON PARALLEL PROCESSING "

(Rza E. Bashirov/Eastern Mediterranean Univ.)

"www.parallel+programming.com

Internet