Parallel Computer Architecture

Fiber Optics Based

Mr. Sanjay Pathak
Lecturer, Department of Electronics &
Communication Engineering
Amrapali Institute of Technology & Sciences, India.
045sonit@gmail.com

Mr. Mayank
Student, Bachelor of Technology
College of Technology,
G.B. Pant University, India.
max1431992@yahoo.com

***Abstract*—This Computer architecture is the conceptual design and fundamental operational structure of a computer system. It's a blueprint and functional description of requirements and design implementations for the various parts of a computer, focusing largely on the way by which the central processing unit (CPU) performs internally and accesses addresses in memory. In this paper, we present an overview of parallel computer architectures and discuss the use of fiber optics for clustered or coupled processors.**

Presently, a number of computer systems take advantage of commercially available fiber optic technology to interconnect multiple processors, thereby achieving improved performance or reliability for a lower cost than if a single large processor had been used. Optical fiber is also used to distribute the elements of a parallel architecture over large distances; these can range from tens of meters to alleviate packaging problems to tens of kilometers for disaster recovery applications.

Keywords- Network topologies; Transputer; Sky HPC-1; CMU Wrap; Tsukuba CP-PACS/2048; Parallel Sysplex & GDPS; MANs

INTRODUCTION

Not all of the parallel computer architectures discussed in this paper use fiber optic connectivity, but they are presented as a context for those systems that are currently using optics. We will give a few specific examples of parallel computer systems that use optical fiber, in particular the Parallel Sysplex architecture from IBM. Other applications do not currently use optical fiber, but they are presented as candidates for optical interconnect in the near future, such as the Power- Parallel supercomputers which are part of the Advanced Strategic Computing Initiative (ASCI). Many of the current applications for fiber optics in this area use serial optical links to share data between processors, although this is by no means the only option. Other schemes including plastic optics, optical backplanes, and free space optical interconnects Towards the end of the paper, we also provide some speculation concerning machines that have not yet been designed or built but which serve to illustrate potential future applications of optical interconnects. Because this is a rapidly developing area, we will frequently cite Internet references where the latest specifications and descriptions of various parallel computers may be found. Computer engineering often presents developers with a choice between designing a computational device with a single powerful processor (with additional special-purpose coprocessors) or designing a parallel processor device with the computation split among multiple processors that may be cheaper and slower. There are several reasons why a designer may choose a parallel architecture over the simpler single processor design. Before each reason, and other categorizing methods in this paper we will have a letter code, A, which we will use to categorize architectures we describe in other sections of the paper.

1. Speed - There are engineering limits to how fast any single processor can compute using current technology. Parallel architectures can exceed these limits by splitting up the computation among multiple processors.
2. Price - It may be possible but prohibitively expensive to design or purchase a single processor machine to perform a task. Often a parallel processor can be constructed out of offthe- shelf components with sufficient capacity to perform a computing task.
3. Reliability - Multiple processors means that a failure of a processor does not prevent computation from continuing. The load from the failed processor can be redistributed among the remaining ones. If the processors are distributed among multiple sites, then even catastrophic failure at one site (due to natural or man-made disaster, for example) would not prevent computation from continuing.
4. Bandwidth - Multiple processors means that more bus bandwidth can be processed by having each processor simultaneously use parts of the bus bandwidth.
5. Other - Designers may have other reasons for adding parallel Processing not covered above. Current parallel processor designs were motivated by one or more of these needs. For example, the parallel Sysplex family was motivated by reliability and speed, the Cray XMP was primarily motivated by speed, the BBN butterfly was designed with bandwidth considerations in mind, and the transputer family was motivated by price and speed. After a designer has chosen to use multiple processors he must make several other choices like processors. Number of processors, network topology The product of the speed of the processors and the number of processors is the maximal processing power of the machine

(for the most part unachievable in real life). The effect of network topology is subtler.

NETWORK TOPOLOGY

First, Network topologies control communication between machines. While most multiprocessors are connected with ordinary copper-wired buses, we believe that fiber optics will be the bus technology of the future. Topology controls how many computers may be necessary to relay a message from one processor to another. A poor network topology can result in bottlenecks where all the computation is waiting for messages to pass through a few very important machines. Also, a bottleneck can result in unreliability with failure of one or few processors causing either failure or poor performance of the entire system.

Four kinds of topologies have been popular for multiprocessors. They are

□ Full connectivity using a crossbar or bus. The historic C.mmp processor used a crossbar to connect the processors to memory (which allowed them to communicate). Computers with small numbers of processors (like a typical parallel Sysplex system or tandem system) can use this topology but it becomes cumbersome with large (more than 16) processors because every processor must be able to simultaneously directly communicate with every other. This topology requires a fan in and fan out proportional to the number of processors, making large networks difficult.

□ Pipeline where the processors are linked together in a line and information primarily passes in one direction. The CMU Warp processor was a pipelined multiprocessor and many of the first historical multiprocessors, the vector processors, were pipelined multiprocessors. The simplicity of the connections and the many numerical algorithms that are easily pipelined encourage people to design these multiprocessors. This topology requires a constant fan in and fan out, making it easy to lay out large numbers of processors and add new ones.

□ Torus and Allied topologies where an N processor machine requires √N processors to relay messages. The Goodyear MPP machine was laid out as a torus. Such topologies are easy to layout on silicon so multiple processors can be placed on a single chip and many such chips can be easily placed on a board. Such technology may be particularly appropriate for computations that are spatially organized. This topology also has constant fan in and fan out. Adding new processors is not as easy as in pipelined processors but laying out this topology is relatively easy. Because of the ease of layout sometimes this layout is used on chips and then the chips are connected in a hypercube.

□ Hypercube and Butterfly topologies have several nice properties that have lead to their dominating large-scale multiprocessor designs. They are symmetric so no processor is required to relay more messages than any other is. Every message need only be relayed through log (N) processors in an N processor machine and messages have multiple alternate routes, increasing reliability under processor failure and improving message routing and throughput. Transputer systems and the BBN butterfly were some of the first multiprocessors that adapted this type of topology. This topology has a logarithmic fan out and that can complicate layout when the size of the processor may grow over time.

There is an alternative topology called cube-connected cycles that has the same efficient message passing properties as the hypercube topology but constant fan out, easing layout considerably.

□ Exotic - There are a variety of less popular but still important topologies one can use on their network. The more efficient and fast the bus technology is, the simpler the topology can be. A really fast bus can simply connect all the processors in a machine together by using time multiplexing giving INI slots for every possible connection between any two of the N processors.

COMPUTING TASK

Before The primary computing task for the machine under consideration has a major effect on the network topology. Computing tasks fall into three general categories.

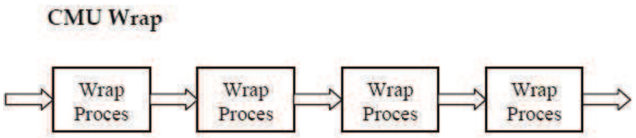
□ Heavy computational tasks - these tasks require much more computation than network communication. Some examples of this task are pattern recognition (SETI), code breaking, inverse problems, and complex simulations such as weather prediction and hydrodynamics.

□ Heavy communication tasks - these tasks involve relatively little computation and massive amounts of communication with other processors and with external devices. Message routing is the classic example of these tasks. Other such tasks are data base operations and search.

□ Intermediate or mixed tasks - these tasks lie between the above or are mixtures of the above. An example of an intermediate task is structured simulation problems, such as battlefield simulation. These simulations require both computation to represent the behavior and properties of the objects (like tanks) and communication to represent interaction between the objects. Some machines may be designed for a mixture of heavy computation and heavy communication tasks. Historically, supercomputers focused on heavy computation tasks, particularly scientific programming, and mainframes focused on heavy communication tasks, particularly business and database applications

DETAILES ARCHITECTURE DESIGN

After In this section we present descriptions of four architectures. Two of these machines are chosen primarily for their historical interest - the CMU Warp and the transputer - and three of these machines represent current architectures of great moment - Sky HPC- 1, Tsukuba CP-PACS/2048, and Parallel Sysplex.



The Warp processor was designed in a project at Carnegie Mellon University (CMU). It was designed for computationally intensive parallel processing on large data sets as is common in signal and image processing. The name of the processor, Warp, refers to an image processing operation. The processors were designed to operate in a pipeline, each processor receiving data from the previous processor and sending processed data on to the next processor in the system.

This kind of processing avoids many synchronization and communication problems that occur in more complex network topologies. Such processing sometimes has problems with latency (the time necessary to fill or exhaust the pipeline) but in the data-intensive processing that the Warp was designed for latency issues are small relative to the size of the data set and the required computation per data point. Each Warp processor had a state-of-the-art floating point coprocessor (pipelined also) and fast integer arithmetic. Each processor was controlled by a VLIW (very long instruction word) microprocessor that allowed maximal internal parallelism because the powerful components of this processor could be addressed and controlled separately. This machine is an early example of a powerful MIMD processor. However, the topology of this machine and the processing expected of it are such that optical components and in particular optical buses are not necessary. It was designed to avoid the issues optical components are designed to address.

Transputer: The transputer was one of the more important parallel architectures and many such machines were sold to academic institutions and research groups. This computer actually had a language, Occam, designed for programming it. The transputer was a processor designed to easily connect with other transputers through four bidirectional I/O ports. These ports were serial, making the work of interconnection easy and allowing flexibility in network architectures. This allowed the design of a wide variety of network topologies using transputers. Each transputer was a state-of-the-art RISC (reduced instruction set computation) processor. While the transputer was not as computationally powerful as a typical Warp processor, it was much more flexible in its I/O capabilities. The main idea of the transputer was to make it easy to write parallel software for it. Thus a program could be written that runs on a single processor that simulates a network of transputers. This program could be debugged on the processor and then run on the network. Thus programs generally did not have to worry about the exact size or shape of the multiprocessor. Once again the transputer was designed to deal flexibly with the problems of inter processor communication in a multiprocessor. A high-speed optical bus system can reduce these problems considerably or even render them moot through time multiplexing. Essentially, a sufficiently high bandwidth system can allow all the multiprocessors to communicate with any other multiprocessor in the system and share system resources at their maximum capacity.

Sky HPC-1: Sky's HPC-1 architecture is an example of NUMA (Non- Uniform Memory Access) architecture. This is one of the most popular architectures around because of the simplification of programming tasks. Basically, every processor has access to all the memory of the machine. The processors and their memory are connected together using a high-speed network. Most of the systems previously described, including the HPC-1, use crossbar switches to maximize the network bandwidth. The HPC- 1 network architecture uses packet-based communications to optimize memory operations. In a packet based architecture, such as SKY channel, both the routing information and the data payload travel together in one transmission or packet. Examples of other packet-based communications include the fixed-size packets (or cells) of ATM and the variable length packets of Sun's XDBus. Many of the SKY channel advantages are partly linked to the use of high speed FIFOs (packet queues). Latency issues with FIFOs were solved by providing a mechanism for early cut-through. Some systems use a technique called "worm-hole routing," but other solutions are available.

Split transactions have been added to reduce the contention at the destination node. The actual connectivity of a HPC-1 is a hierarchical crossbar system in which there is a master crossbar connected subsystem whose processors lie on the subsystem crossbar. This connectivity suffices for the machine sizes SKY provides. All of the technological innovations of the SKY architecture can be applied with great effect to a system designed around a high-speed optical bus. Such a system would still be packet-based and still use the FIFOs but would use the bus instead of crossbars to route the packets between processors.

Tsukuba CP-PACS/2048

This machine is an interesting example of modern thought in parallel machine design. It was designed by collaboration between physicists and computer scientists as a high-speed machine devoted to difficult problems in physics. The details of the machine design are particularly available because the machine was designed for academic rather than commercial purposes. This machine is designed as a cube of processors, each processor participating in three crossbar switches. Thus the processor with coordinates 5,4,3 can instantly communicate with any processor that shares any of the coordinates; for example it is connected to 5,7,1 and to 6,4,9 and to 7,2,3. Each processor is an extremely high-speed numerical computation device that at one time would have been called a supercomputer itself. The Tsukuba CP-PACS/2048 has 2048 of these processors tightly coupled together using the network described previously. Data transfer on the network is made through Remote DMA (Remote Direct Memory Access), in which processors exchange data directly between their respective user memories with a minimum of intervention from the operating system. This leads to a significant reduction in the startup latency, and a high throughput. A well-balanced performance of CPU, network, and I/O devices supports the high capability of CP-PACS for massively parallel processing (6/4Gflops). We believe that in the future high speed optical buses can replace the crossbar switches of this architecture, making it more flexible, and easier to maintain.

Parallel Sysplex & GDPS

High-end computer systems running over metropolitan area networks (MANS) are proving to be a near term application for multi-terabit communication networks. Large computer systems require dedicated storage area networks (SANS) to interconnect with various types of direct attach storage devices (DASD), including magnetic disk and tape, optical storage devices, printers, and other equipment. This has led to the emergence of client server- based networks employing either circuit or packet switching, and the development of network-centric computing models. In this approach, a high bandwidth, open protocol network is the most critical resource in a computer system, surpassing even the processor speed in its importance to overall performance. The recent trend toward clustered, parallel computer architectures to enhance performance has also driven the requirement for high bandwidth fiber optic coupling links between computers. For example, large water-cooled main frame computers using bipolar silicon processors are being replaced by smaller, air-cooled servers using complementary metal oxide semiconductor (CMOS) processors. These new processors can far surpass the performance of older systems because of their ability to couple together many central processing units in parallel. One widely adopted architecture for clustered mainframe computing is known as a Geographically Dispersed Parallel Sysplex (GDPS). In this section, we will describe the basic features of a GDPS and show how this architecture is helping to drive the need for highbandwidth dense-wavelength division multiplexing (DWDM) networks.

In 1994, IBM announced the Parallel Sysplex architecture for the System/ 390 mainframe computer platforms (note that the S/390 has recently been re branded as the IBM eServer z series). This architecture uses high-speed fiber optic data links to couple processors together in parallel [1-4], thereby increasing capacity and scalability. Processors are interconnected via a coupling facility, which provides data caching, locking, and queuing services; it may be implemented as a logical partition rather than a separate physical device. The gigabit links, known as Intersystem Channel (ISC), HiPerLinks, or Coupling Links, use long-wavelength (1300-nm) lasers and single-mode fiber to operate at distances up to 10 km with a 7 dB link budget (HiPerLinks were originally announced with a maximum distance of 3 km, which was increased to 10 km in May 1998). If good quality fiber is used, the link budget of these channels allows the maximum distance to be increased to 20 km. When HiPerLinks were originally announced, an optional interface at 53 1 Mbit/s was offered using shortwavelength lasers on MM fiber. The 531 Mbit/s HiperLinks were discontinued in May 1998 for the G5 server and its follow-ons. A feature is available to accommodate operation of 1 Gbit/s HiPerLinks adapters on multimode fiber, using a mode conditioning jumper cable at restricted distances (550 meters maximum).

The physical layer design is similar to the ANSI Fibre Channel Standard, operating at a data rate of 1.0625 Gbit/s, except for the use of open fiber control (OFC) laser safety on long-wavelength (1300 nm) laser links (higher order protocols for ISC links are currently IBM proprietary). Open fiber control is a safety interlock implemented in the transceiver hardware; a pair of transceivers connected by a point-to-point link must perform a handshake sequence in order to initialize the link before data transmission occurs. Only after this handshake is complete will the lasers turn on at full optical power. If the link is opened for any reason (such as a broken fiber or unplugged connector) the link detects this and automatically deactivates the lasers on both ends to prevent exposure to hazardous optical power levels. When the link is closed again, the hardware automatically detects this condition and reestablishes the link. The HiPer Links use OFC timing corresponding to a 266 Mbit/s link in the ANSI standard, which allows for longer distances at the higher data rate. Propagating OFC signals over DWDM or optical repeaters is a formidable technical problem, which has limited the availability of optical repeaters for HiPer Links. OFC was initially used as a laser eye safety feature; subsequent changes to the international laser safety standards have made this unnecessary, and it has been discontinued on the most recent version of z series servers. The 1.06 Gbit/s HiPer Links will continue to support OFC in order to interoperate with installed equipment; this is called "compatibility mode." There is also a 2.1 Gbit/s HiPer Link channel, also known as ISC-3, which does not use OFC; this is called "peer mode." There are three possible configurations for a Parallel Sysplex. First, the entire sysplex may reside in a single physical location, within one data center. Second, the sysplex can be extended over multiple locations with remote fiber optic data links. Finally, a multi-site sysplex in which all data is remote, copied from one location to another, is known as a Geographically Dispersed Parallel Sysplex, or GDPS. The GDPS also provides the ability to manage remote copy configurations, automates both planned and unplanned system reconfigurations, and provides rapid failure recovery from a single point of control. There are different configuration options for a GDPS. The single site workload configuration is intended for those enterprises that have production workload in one location (site A) and discretionary workload (system test platforms, application development, etc.) in another location (site B). In the event of a system failure, unplanned site failure, or planned workload shift, the discretionary workload in site B will be terminated to provide processing resources for the production work from site A (the resources are acquired from site B to prepare this environment, and the critical workload is restarted). The multiple site workload configuration is intended for those enterprises that have production and discretionary workload in both site A and site B. In this case, discretionary workload from either site may be terminated to provide processing resources for the production workload from the other site in the event of a planned or unplanned system disruption or site failure.

Multi-site Parallel Sysplex or GDPS configurations may require many links (ESCON, HiPerLinks, and Sysplex Timer) at extended distances; an efficient way to realize this is the use of wavelength division multiplexing technology. Multiplexing wavelengths is a way to take advantage of the high bandwidth of fiber optic cables without requiring extremely high modulation rates at the transceiver. This type of product is a cost effective way to utilize leased fiber optic lines, which are not readily available everywhere and may be very high cost (typically the cost of leased fiber (sometimes known as dark fiber) where available is $3OO/mile/month).

Traditionally, optical wavelength division multiplexing (WDM) has been widely used in telecom applications, but has found limited usage in data com applications. This is changing, and a number of companies are now offering multiplexing alternatives to data com networks that need to make more efficient use of their existing bandwidth. This technology may even be the first step toward development of all-optical networks. For Parallel Sysplex applications, the only currently available WDM channel extender that supports GDPS (Sysplex Timer and HiPerLinks) in addition to ESCON channels is the IBM 2029 Fiber Saver [5-81] (note that the 9729 Optical Wavelength Division Multiplexer also supported GDPS but has been discontinued; other DWDM products are expected to support GDPS in the future, including offerings from Nortel and Cisco). The 2029 allows up to 32 independent wavelengths (channels) to be combined over one pair of optical fibers, and extends the link distance up to 50 km point-to-point or 35 km in ring topologies. Longer distances may be achievable from the DWDM using cascaded networks or optical amplifiers, but currently a GDPS is limited to a maximum distance of 40 km by timing considerations on the ETR and CLO links (the Sysplex timer documents support for distances up to only 26 km, the extension to 40 km requires a special request from IBM via RPQ 8P1955). These timing requirements also make it impractical to use TDM or digital wrappers in combination with DWDM to run ETR and CLO links at extended distances; this implies that at least 4 dedicated wavelengths must be allocated for the Sysplex timer functions. Also note that since the Sysplex timer assumes that the latency of the transmit and receive sides of a duplex ETR and CLO link are approximately equal, the length of these link segments should be within 50 m of each other. For this reason, unidirectional 1+1 protection switching is not supported for DWDM systems using the 2029; only bidirectional protection switching will work properly.

Even so, most protection schemes cannot switch fast enough to avoid interrupting the sysplex timer and HiPerLinks operation. HiPerLinks in compatibility mode will be interrupted by their open fiber control, which then takes up to 10 seconds to reestablish the links. Timer channels will also experience loss of light disruptions, as will ESCON and other types of links. Even when all the links reestablish, the application will have been interrupted or disabled and any jobs that had been running on the sysplex will have to be restarted or reinitiated, either manually or by the host's automatic recovery mechanisms depending on the state of the job when the links were broken. For this reason, it is recommended that continuous availability of the applications cannot be ensured without using dual redundant ETR, CLO, and HiPerLinks. Protection switching merely restores the fiber capacity more quickly; it does not ensure continuous operation of the Sysplex in the event of a fiber break. To illustrate the use of DWDM in this environment, consider the construction of a GDPS between two remote locations for disaster recovery, as shown in Fig. 1. There are four building blocks for a Parallel Sysplex; the host processor (or Parallel Enterprise Server), the coupling facility, the ETR (Sysplex Timer), and disk storage. Many different processors may be Inter connected through the coupling facility, which allows them to communicate with each other and with data stored locally. The coupling facility provides data caching, locking, and queuing (message passing) services. By adding more processors to the configuration, the overall processing power of the Sysplex (measured in millions of instructions per second or MIPS) will increase. It is also possible to upgrade to more powerful processors by simply connecting them into the Sysplex via the coupling facility. Special software allows the Sysplex to break down large database applications into smaller ones, which can then be processed separately; the results are combined to arrive at the final query response. The coupling facility may either be implemented as a separate piece of hardware, or as a logical partition of a larger system. The HiPerLinks are used to connect a processor with a coupling facility. Because the operation of a Parallel Sysplex depends on these links, it is highly recommended that redundant links and coupling facilities be used for continuous availability. Thus, in order to build a GDPS, we require at least one processor, coupling facility, ETR, and disk storage at both the primary and secondary locations, shown in Fig.1 as site A and site B. Recall that one processor may be logically partitioned into many different Sysplex system images; the number of system images determines the required number of HiPerLinks. The Sysplex system images at site A must have HiPerLinks to the coupling facilities at both site A and B. Similarly, the Sysplex system images at site B must have

HiPerLinks to the coupling facilities at both site A and B. In this way, failure of one coupling facility or one system image allows the rest of the Sysplex to continue uninterrupted operation.

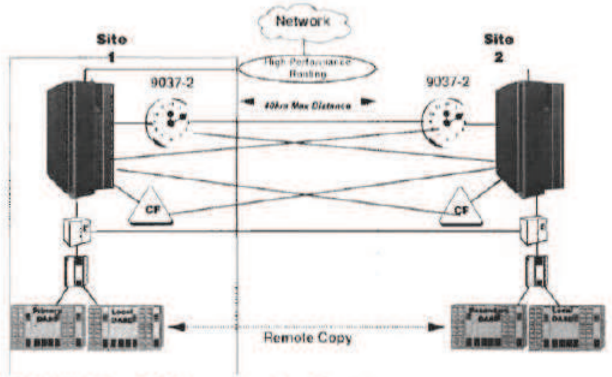


Fig. 1 IBM Parallel Sysplex Architecture

A minimum of two links is recommended between each system image and coupling facility. Assuming there are Sysplex system images running on P processors and C coupling facilities in the GDPS, spread equally between site A and site B, the total number of HiPerLinks required is given by

# HiPerLinks = S * C * 2 (1)

In a GDPS, the total number of inter-site HiPerLinks is given by inter-site

# HiPerLinks = S * C (2)

The Sysplex Timer (9037) at site A must have links to the processors at both site A and B. Similarly, the 9037 at site B must have links to the processors at both site A and B. There must also be two CLO links between the timers at sites A and B. This makes a minimum of four duplex inter-site links, or eight optical fibers without multiplexing. For practical purposes, there should never be a single point of failure in the sysplex implementation; if all the fibers are routed through the same physical path, there is a possibility that a disaster on this path would disrupt operations. For this reason, it is highly recommended that dual physical paths be used for all local and inter-site fiber optic links, including Hyperlinks, ESCON, ETR, and CLO links. If there are P processors spread evenly between site A and site B, then the minimum number of ETR links required is given by

# E m links = (P * 2) + 2 CLO links (3)

In a GDPS, the number of inter-site ETR links is given by Inter-site

# ETR links = P + 2 CLO links (4)

These formulas are valid for CMOS-based hosts only; note that the number of ETR links doubles for ES/9000 Multiprocessor models due to differences in the server architecture. In addition, there are other types of inter-site links such as ESCON channels to allow data access at both locations.

In a GDPS with a total of N storage subsystems (also known as Direct Access Storage Devices or DASD), it is recommended that there be at least four or more paths from each processor to each storage control unit (based on the use of ESCON Directors at each site); thus, the number of inter-site links is given by Inter-site

# storage (ESCON) links = N * 4 (5)

In addition, the sysplex requires direct connections between systems for cross-system coupling facility (XCF) communication. These connections may be provided by either ESCON channel-to-channel links or HiPerLinks. If coupling links are used for XCF signaling, then no additional HiPerLinks are required beyond those given by equations (1) and (2). If ESCON links are used for XCF signaling, at least two inbound and two outbound links between each system are required, in addition to the ESCON links for data storage discussed previously. The minimum number of channel -to-channel (CTC) ESCON links is given by

# CTC links = S * (S - 1) * 2 (6)

For a GDPS with SA Sysplex systems at site A and SB Sysplex systems at site B, the minimum number of inter-site channel-tochannel links is given by Inter-site

# CTC links = SA * SB * 4 (7)

Because some processors also have direct local area network (LAN) connectivity via FDDI or ATM/SONET links, it may be desirable to run some additional inter-site links for remote LAN operation as well. As an example of applying these equations, consider a GDPS consisting of two system images executing on the same processor and a coupling facility at site A, and the same configuration at site B. Each site also contains one primary and one secondary DASD subsystem. Sysplex connectivity for XCF signaling is provided by ESCON CTC links, and all GDPS recommendations for dual redundancy and continuous availability in the event of a single failure have been implemented. From eq. (1-7), the total number of inter-site links required is given by # of inter-site links:

# CTC links = SA * SB * 4 = 2 * 2 * 4 = 16 (8)
#timer links = P +2 = 2+2 = 4
# HiPerLinks = S * C = 4 * 2 = 8
# Storage (DASD) links = N * 4 = 8 * 4 = 32

or a total of 60 inter-site links. Note that currently, only ESCON links may be used for the direct connection between local and remote DASD via the Peer-to-Peer Remote Copy (PPRC) protocols. Other types of storage protocols such as Fibre Channel or FICON may be used for the DASD connections. Note that any synchronous remote copy technology will increase the I/O response time, because it will take longer to complete a writing operation with synchronous remote copy than without it (this effect can be offset to some degree by using other approaches, such as parallel access to storage volumes).The tradeoff for longer response times is that no data will be lost or corrupted if there is a single point of failure in the optical network. PPRC makes it possible to maintain synchronous copies of data at distances up to 103 km; however, these distances can only be reached using either DWDM with optical amplifiers or by using some other form of channel extender technology. The performance and response time of PPRC links depends on many factors, including the number of volumes of storage being accessed, the number of logical subsystems across which the data is spread, the speed of the processors in the storage control units and processors, and the intensity of the concurrent application workload. In general, the performance of DASD and processors has increased significantly over the past decade, to the point where storage control units and processors developed within the past two years have their response time limited mainly by the distance and the available bandwidth. Many typical workloads perform several read operations for each write operation; in this case the effect of PPRC on response time is not expected to be significant at common access densities. Similar considerations will apply to any distributed synchronous architecture such as Parallel Sysplex. In some cases, such as disaster recovery applications, where large amounts of data must be remotely backed up to a redundant storage facility, an asynchronous approach is practical. This eliminates the need for sysplex timers, and trades off continuous real-time data backup for intermittent backup; if the backup interval is sufficiently small, then the impact can be minimized. One example of this approach is the extended Remote Copy (XRC) protocols supported by FICON channels on a z series server. This approach interconnects servers and DASD between a primary and a backup location, and periodically initiates a remote copy of data from the primary to the secondary DASD. This approach requires fewer fiber optic links, and because it does not use a sysplex timer the distances can be extended to 100 km or more. The tradeoff with data integrity must be assessed on a case-by-case basis; some users prefer to implement XRC as a first step toward a complete GDPS solution. The use of a parallel computing architecture over extended distances is a particularly good match with fiber optic technology. Channel extension is well known in other computer applications, such as storage area networks; today, mainframes are commonly connected to remote storage devices housed tens of kilometers away. This approach, first adopted in the early 1990s, fundamentally changed the way in which most people planned their computer centers, and the amount of data they could safely process; it also led many industry pundits to declare “the death of distance.” Of course, unlike relatively low bandwidth telephone signals, performance of many data communication protocols begins to suffer with increasing latency (the time delay incurred to complete transfer of data from storage to the processor). While it is easy to place a long-distance phone call from New York to San Francisco (about 42 milliseconds round trip latency in a straight line, longer for a more realistic route), it is impossible to run a synchronous computer architecture over such distances. Further compounding the problem, many data communication protocols were never designed to work efficiently over long distances. They required the computer to send overhead messages to perform functions such as initializing the communication path, verifying it was secure, and confirming error-free transmission for every byte of data. This meant that perhaps a half dozen control messages had to pass back and forth between the computer and storage unit for every block of data, while the computer processor sat idle. The performance of any duplex data link begins to fall off when the time required for the optical signal to make one round trip equals the time required to transmit all the data in the transceiver memory buffer. Beyond this point, the attached processors and storage need to wait for the arrival of data in transit on the link, and this latency reduces the overall system performance and the effective data rate. As an example, consider a typical fiber optic link with a latency of about 10 microseconds per kilometer round trip. A mainframe available in 1995 capable of executing 500 million instructions per second (MIPS) needs to wait not only for the data to arrive, but also for 6 or more handshakes of the overhead protocols to make the round trip from the computer to the storage devices. The computer could be asting 100 MIPS of work, or 20% of its maximum capacity, while it waits for data to be retrieved from a remote location 20 kilometers away. Although there are other contributing factors, such as the software applications and workload, this problem generally becomes worse as computers get faster, because more and more processor cycles are wasted waiting for the data. As this became a serious problem, various efforts were made to design lower latency communication links. For example, new protocols were introduced that required fewer handshakes to transfer data efficiently, and the raw bandwidth of the fiber optic links was increased from ESCON rates (about 17 Mbyte/s) to nearly 100 Mbyte/s for FICON links. But for very large distributed applications, the latency of signals in the optical fiber remains a fundamental limitation; DASD read and writes times, which are significantly longer, will also show a more pronounced effect at extended distances.

OPTICALLY INTERCONNECTED PARALLEL SUPERCOMPUTERS

Latency is not only a problem for processor-to-storage interconnections, but also a fundamental limit in the internal design of very large computer systems. Today, many supercomputers are being designed to solve so-called “Grand Challenge” problems, such as advanced genetics research, modeling global weather patterns or financial portfolio risks, studying astronomical data such as models of the Big Bang and black holes, design of aircraft and spacecraft, or controlling air traffic on a global scale. This class of high risk high reward problems is also known as “Deep Computing.” A common approach to building very powerful processors is to take a large number of smaller processors and interconnect them in parallel. In some cases, a computational problem can be subdivided into many smaller parts, which are then distributed to the individual processors; the results are then recombined as they are completed to form the final answer. This is one form of asynchronous processing, and there are many problems that fall into this classification; one of the best examples is SETI@home, free software which can be downloaded over the Internet to any home personal computer. Part of the former NASA program, SETI (Search for Extra-Terrestrial Intelligence) uses spare processing cycles when a computer is idle to analyze extraterrestrial signals from the Arecibo Radio Telescope, searching for signs of intelligent life. There are currently over 1.6 million SETI@home subscribers in 224 countries, averaging 10 teraflops (10 trillion floating point operations performed per second) and having contributed the equivalent of over 165,000 years of computer time to the project. Taken together, this is arguably the world‘s largest distributed supercomputer, mostly interconnected with optical fiber via the Internet backbone. More conventional approaches rely on large numbers of processors interconnected within a single package. In this case, optical interconnects offer bandwidth and scalability advantages, as well as immunity from electromagnetic noise, which can be a problem on high-speed copper interconnects. For these reasons, fiber optic links or ribbons are being considered as a next-generation inter connect technology for many parallel computer architectures, such as the Power Parallel and NUMA-Q designs. The use of optical backplanes and related technologies is also being studied for other aspects of computer design. To minimize latency, it is desirable to locate processors as close together as possible, but this is sometimes not possible due to other considerations, such as the physical size of the packages needed for power and cooling. Reliability of individual computer components is also a factor in how large we can scale parallel processor architectures. As an example, consider the first electronic calculator built at the University of Pennsylvania in 1946, ENIAC (Electronic Numeric Integrator and Computer), which was limited by the reliability of its 18,000 vacuum tubes; the machine couldn’t scale beyond filling a mom about 10 by 13 meters, because tubes would blow out faster than people could run from one end of the machine to the other replacing them. Although the reliability of individual components has improved considerably, modem-day supercomputers still require some level of modularity, which comes with an associated size and cost penalty. A well-known example of Deep Computing is the famous chess computer, Deep Blue, that defeated grand master Gary Kasparov in May of 1997. As a more practical example, the world’s largest supercomputer is currently owned and operated by the U.S. Department of Energy, to simulate the effects of nuclear explosions (such testing having been banned by international treaty). This problem requires a parallel computer about fifty times faster than Deep Blue (although it uses basically the same internal architecture). To accomplish this requires a machine capable of 12 teraflops, a level computer scientists once thought impossible to reach. Computers with this level of performance have been developed gradually over the years, as part of the Accelerated Strategic Computing Initiative (ASCI) roadmap; but the current generation, called ASCI White, has more than tripled the previous world record for computing power. This single supercomputer consists of hundreds of equipment cabinets housing a total of 8,192 processors, interconnected with a mix of copper and optical fiber cables through two layers of switching fabric. Because the cabinets can’t be pressed flat against each other, the total footprint of this machine covers 922 square meters, the equivalent of 2 basketball courts. This single computer weighs 106 tons (as much as 17 full-size elephants) and had to be shipped to Lawrence Livermore National Labs in California on 28 tractor trailers. It’s not feasible today to put the two farthest cabinets closer together than about 43 meters, and this latency limits the performance of the parallel computer system. Furthermore, ASCI White requires over 75 terabytes of storage (enough to hold all 17 million books in the Library of Congress), which may also need to be backed up remotely for disaster recovery; so, the effects of latency on the processor-tostorage connections are also critically important. Future ASCI programs call for building a 100 Teraflop machine by 2004.

PARALLEL FUTURE

Current Parallel Sysplex systems have been benchmarked at over 2.5 billion instructions per second, and are likely to continue to significantly increase in performance each year. The ASCI program has also set aggressive goals for future optically interconnected supercomputers. However, even these are not the most ambitious parallel computers being designed for future applications. There is a project currently under way; led by LBM fellow Monty Denneau, to construct a mammoth computer nicknamed “Blue Gene” which will be dedicated to unlocking the secrets of protein folding. Without going into the details of this biotechnology problem, we note that it could lead to innumerable benefits, including a range of designer drugs, whole new branches of pharmacology, and gene therapy treatments that could revolutionize health care, not to mention lending fundamental insights into how the human body works. This is a massive computational problem, and Blue Gene is being designed for the task; when completed, it will be 500 times more powerful that ASCI White, a 12.3 petaflop machine - well over a quadrillion (operations per second, forty times faster than today’s top 40 supercomputers combined. The design point proposes 32 microprocessors on a chip, 64 chips on a circuit board, 8 boards in a 6- foot-high tower, and 64 interconnected towers for a total of over 1 million processors. Because of improvements in packaging technology, Blue Gene will occupy somewhat less space than required by simply extrapolating the size of its predecessors; about 11 x 24 meters (about the size of a tennis court), with a worst-case diagonal distance of about 26 meters. However, the fast processors proposed for this design can magnify the effect of even this much latency to the point where Blue Gene will be wasting about 1.6 billion operations in the time required for a diagonal interconnect using conventional optical fiber. Further, a machine of this scale is expected to have around 10 terabytes of storage requirements, easily enough to fill another tennis court, and give a processor-to-storage latency double that of the processor-to-processor latency. Because of the highly complex nature of the protein folding problem, a typical simulation on Blue Gene could take years to complete, and even then may yield just one piece of the answer to a complex protein-folding problem. While designs such as this have yet to be realized, they illustrate the increasing interest in parallel computer architectures as an economical means to achieving higher performance. Both serial and parallel optical links are expected to play an increasing role in this area, serving as both processor- to- processor and processor-to-storage interconnects.

REFERENCES

DeCusatis, C.,D. Stigliani, W. Mostowy, M. Lewis,D. Petersen, andN. Dhondy. Sept./Nov. 1999. "Fiber optic interconnects for the IBM W390 parallel enterprise server." IBM Journal of Research and Development. 43(5/6):807-828; also see ZBM Journal of Research & Development, special issue on "IBM System/390 Architecture and Design," 36:4 (1992).

J. Clerk "IBM Corporation 9729 optical wavelength division multiplexer."Photonics Spectra (Special Issue: The 1996 Phonetics Circle of Excellence Awards), Vol. 30 June 1996.

'9729 Operators Manual" (IBM document number GA27- 4172), IBM Corporation, Mechanicsburg, Pa. (1996).

Bona, G.L., et al. December 1998. "Wavelength division multiplexed add dropping technology in corporate backbone networks." Optical Engineering, special issue on Optical Data Communication.

DeCusatis, C., D. Petersen, E. Hall, and E Janniello. December 1998. "Geographically distributed parallel sysplex architecture using optical wavelength division multiplexing." Oprical Engineering, special issue on Optical Data Communication.

Thiebaut, D., Parallel Programming in C for the Transputer, 1995,http://www.cs.smith.edu/-thiebaut/transputer/descript.html?clkd=iwm.