LA-UR -91-2705

-3

LA-UR--91-2705

DE91 018035

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36.

TITLE: PRESENT AND FUTURE SUPERCOMPUTER NETWORK ARCHITECTURES

AUTHOR(S): DON E. TOLMIE

SUBMITTED TO: SPIE (The International Society for Optical Engineering)
OE/FIBERS '91
Boston, MA, Sept. 6, 1991

DISCLAIMER

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

FORM NO 836 R4
ST NO 2629 5/81

Present and future supercomputer network architectures

Don E. Tolmie

Los Alamos National Laboratory, C-5, MS-B255
Los Alamos, New Mexico 87545)

ABSTRACT

Computer networks must provide high data transfer rates to maximize the effectiveness of the interconnected equipment, and especially to maximize the effectiveness of the users, e.g., with visualization. Network speeds are increasing, with the newest systems using 800 Mbit/s data rates. The most common computer networks today use bus and ring architectures. Supercomputer networks are starting to use circuit switching with crossbar switches. Wavelength division multiplexing and all optical networking are research topics today, but hold promise for the future. The architectures, attributes, and problems of these different systems are discussed, with emphasis on their use in the supercomputer environment.

2. WHY DO WE NEED HIGHER SPEEDS ?

When networks were mainly used to carry key strokes between dumb terminals and mainframes, 9600 baud was quite adequate; it was considerably faster than people could read. Today it is more common to pass files and pictures between the workstations, mainframes, and storage systems. The emphasis is on improving the users productivity and avoiding network bottlenecks.

2.1. Visualization

If a picture is worth a thousand words, then remember that it probably also takes a thousand times the bandwidth to transfer that picture. People are not content with just pictures, presenting the computer output data in movie format (called visualization) is the newest craze and offers even higher user productivity increases. The potential bandwidth of the human eye-brain system has been calculated to be on the order of a few gigabits per second, hence gigabit speeds should satisfy the individual user's needs for a while.[1]

The networking factors of importance for visualization are raw speed and non-interference between data streams - if a visualization data stream is interrupted by another packet, then the user sees a glitch which is very distracting. Visualization sessions also tend to last for many seconds, compared to a single packet transfer which may only take a few microseconds. Error control is also unique in that data in error is discarded rather than being retransmitted.

2.2. File Transfers

As the computers become faster, they also increase their appetite for data. A computer that is constipated because of bottlenecks for input or output data is wasting useful compute cycles. A major factor is the bandwidth between the computer and its mass storage system. Mass storage systems used to be limited to single disks attached intimately to individual computer systems; today the trend is for groups of disks to be shared among a group of networked workstations. The networking factors of importance for file transfers are raw speed and fairly large files; latency and interfering data streams are not major concerns.

2.3. Remote procedure calls

An interesting concept that is gaining acceptance is the close coupling of many workstations to achieve the compute power of a supercomputer. Single CPU supercomputers are running out of potential performance gains due to the laws of physics limiting the speed of light and electrons. Performance gains in the future will be achieved by interconnecting many smaller computers and spreading the problem across all of them. This has been termed

"the attack of the killer micros". The networking factors of importance for remote procedure calls (RPC's) are raw speed, low cost (it shouldn't cost more than the workstation), and low latency. The information transferred tends to be mainly short data, control, and synchronizing packets.

2.4. Mixed media

Mixing data, voice, and video in a single session is also a way to improve the user's productivity. The networks to support this mix of traffic have different goals than plain data networks. The transmission of voice and video is time critical. That is, if the data does not arrive in time, then it is useless and should be discarded. This mixing of synchronous, asynchronous, and isochronous data within a single network presents some interesting design constraints.

3. PRESENT COMPUTER NETWORK ARCHITECTURES

The architecture of a network is the sum of the network topology, traffic types supported, speed, access method, etc. Networks can be optimized for many different parameters, and to date there is no one network architecture that has proven to be the best for all applications. This paper does not purport to cover all computer network architectures currently in use.

3.1. Bus architectures

The most common computer network available today is the Ethernet bus at 10 Mbit/s.[2] Hyperchannel, using a bus architecture, was the supercomputer network of choice in the 1980's. Figure 1 shows a simple bus network. A bus architecture can be characterized as having distributed memory, distributed control, and a single data path. In the figures, the dashed line surrounds what is usually a single physical unit.

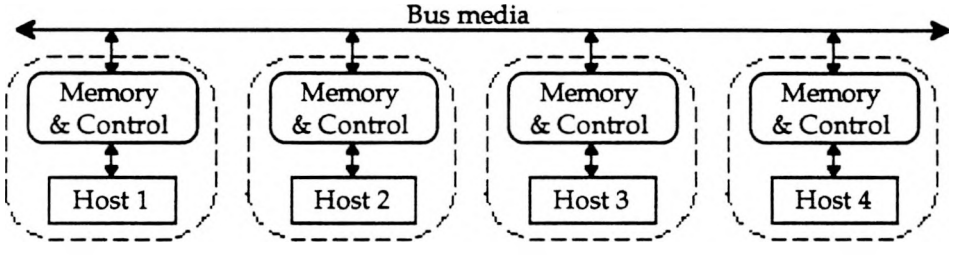

Fig. 1. Bus architecture

A bus is similar in nature to a party-line telephone system. That is, you must wait for the line to be idle before you can talk - resulting in only a single conversation or data transfer at a time. If Host 1 is transferring to Host 2, then Hosts 3 and 4 must wait for this conversation to complete before they can transfer. An advantage of a bus architecture is that it is easy to add one more station. Also, since there is no master station or central administrator, the failure of a single entity usually does not take down the whole bus.

Since everyone sees all of the data, broadcast and multi-cast are easy, but since everyone sees all of the data the security of the system is low. One jabbering station can also take all of the available bandwidth, denying service to the other stations.

A major limitation is that the total bandwidth of a bus based system is limited to the bandwidth of the bus, i.e., there is only one data path. If you need more bandwidth, then you must split the system and run separate busses interconnected by bridges or routers, or upgrade the whole system to a higher speed bus.

The multi-drop nature of busses do not lend themselves well to fiber optic implementations. Hence, the bus media is usually implemented with copper coax or twisted-pair cable. A disadvantage is that the high-speed portion of the system is the bus, and this is also the portion spanning the greatest distance.

3.2. Ring architectures

In some ways rings, as shown in figure 2, are very similar to busses, e.g., only one data path and only one transfer at a time. They too can be characterized as having distributed memory, distributed control, and a single data path. Some rings allow multiple simultaneous conversations, e.g., Host 1 to Host 2 simultaneously with Host 3 to Host 4. Of course the acknowledgements from Host 2 to Host 1 cannot occur simultaneously with the ones from Host 4 to Host 3.

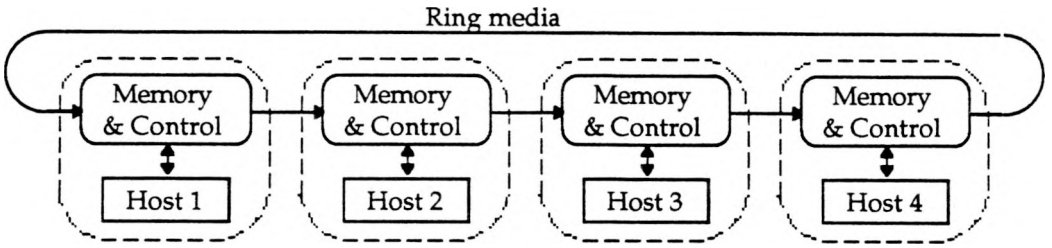

Fig. 2. Ring architecture

Rings have some unique advantages and disadvantages when compared to busses. For example, ring networks often use a token passing scheme for access to the media, allowing each station a fairer access to transmit. A disadvantage is that one dead station can disrupt the whole ring. Ring networks often work around this problem by using dual rings, bypass relays, etc. The Fiber Distributed Data Interface (FDDI) is an example of a ring network using dual counter-rotating rings at 100 Mbit/s speeds.[3]

Since the ring interconnections are point-to-point, the ring media can easily be implemented with copper cabling or fiber optics. Like the bus network, the high-speed portion of the system is the ring media, and this portion also covers the greatest distance. Experience has shown that point-to-point wiring, as compared to multi-drop, will support higher speeds with better reliability. In a bus network, the signal power transmitted by a station is shared by all of the receivers; hence the received power is a function of how many stations are connected - which can range from two to many, giving a very large dynamic range that the receiver must accommodate. On a multi-drop media you also must consider reflections, cross-talk, etc., which are usually less of a problem in point-to-point configurations. Also, shorter links will support higher speeds, and are usually more reliable.

3.3. Store-and-forward packet switches

Probably the best known examples of store-and-forward packet switches, as shown in figure 3, are the nodes of the Internet (previously known as the ARPA net). Figure 3 is also very similar to the architecture of the Integrated Computing Network at the Los Alamos National Laboratory. A store-and-forward architecture can be characterized as having shared central memory, shared central control, and multiple data paths.

Store-and-forward packet switches can be viewed as being similar to a postal system. That is, a source host writes a message when it wants, at the speed it wants, and sends it to the central post office (the central memory and control). The central control will then forward it to the destination host when the destination is free, and it will send it at the speed compatible with the destination. This is much more decoupled than a bus or ring architecture.

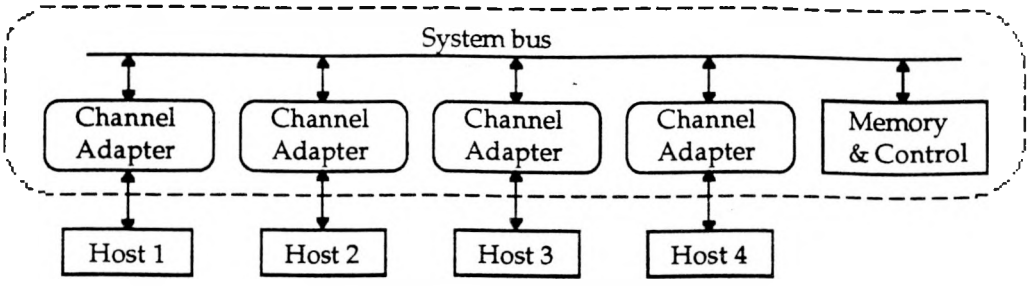

Fig. 3. Store-and-forward packet switch

Security is enhanced since all of the data flows through the central node where security can be checked, and only the source and destination hosts see the data. Of course, this also introduces a single point of failure, but individual host failures do not usually take down the whole system. Expansion is easy until you run out of ports on the central node, and then you may have a major purchase to add the next node.

The high-speed portion of a store-and-forward network is the system bus of the central node. The system bus is usually short (within a cabinet) and provided by one vendor, and hence should be very reliable. The point-to-point links from the individual hosts need to have sufficient bandwidth to handle the traffic of the host, but do not necessarily need to match the speed of the system bus.

3.4. Crosspoint, crossbar, or matrix switches

Crosspoint, crossbar, or matrix switches (they go by these names and other names as well), can be compared to the user's view of a central office telephone system. That is, as shown in figure 4, Host 1 can be connected and talking to Host 3 while Host 2 is simultaneously talking to Host 4, and Host 3 to Host 1. Crossbar switch networks can be characterized as having distributed memory, distributed control, and multiple data paths.

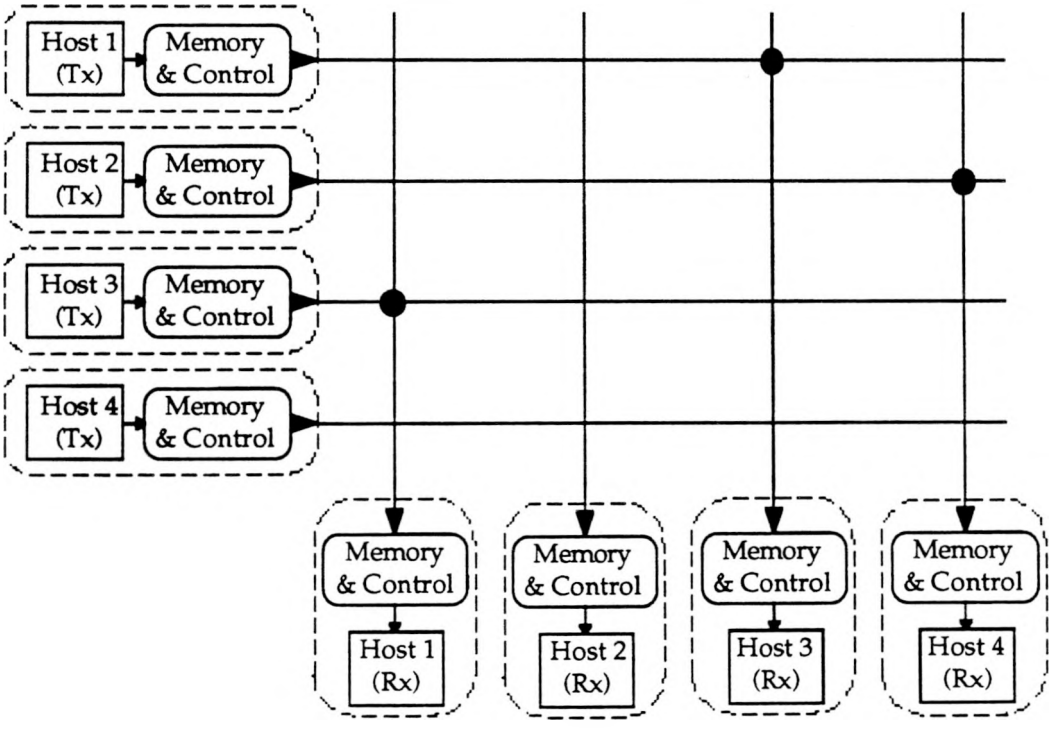

Fig. 4. Crossbar switch

In figure 4 the Host connections are shown as dual-simplex, i.e., they can be sending to one destination while simultaneously receiving from a different source. Dual-simplex is the most general case, but many networks switch in a full-duplex fashion, i.e., a host must send to and receive from a single host at one point in time. Full-duplex is how telephone conversations are switched, and many I/O transfers. Dual-simplex is more common in computer network communications, e.g., Ethernet protocols, TCP/IP, etc. In figure 4, the memory and control for each host is shown as physically located close to the host, where in fact it may be located with the switch.

A major advantage of crossbar switches is that they support simultaneous connections. This is important when using visualization, as the data must be transmitted in a continuous and smooth fashion to the screen; any glitches in the transfer tend to distract the user. Visualization also has high bandwidth demands. For example, using a screen with 1K x 1K pixels, 24 bits of color information per pixel, and 24 frames per second, requires a sustained data rate of about 700 Mbit/s. 50 Mbit/s rates can support 512 x 512 pixels with 8 bits of color per pixel and a 24 frame per second refresh rate. If visualization data were transferred over a shared media, such as a bus or ring, then any other transfers on the media may well cause delays in delivering the visualization data, resulting in glitches.

The aggregate data rate of a crossbar switch is the number of ports times the data rate of each port. For example, a 16 x 16 crossbar switch for the 800 Mbit/s High-Performance Parallel Interface (HIPPI) channel can support 16 separate data streams, with an aggregate data transfer rate of 25.6 Gbit/s.[4,5] The speed of the individual components of the switch and of the interfaces need not be any greater than the speed of the basic link, e.g., 800 Mbit/s. Switches like the one described are commercially available and in production use.[6]

Broadcast and multi-cast services are awkward with a crossbar switch, and these services have been used in bus based networks for such things as learning the address of a particular node, e.g., in the Address Resolution Protocol (ARP). Many of the telephone features may advantageously be applied to crossbar switches, e.g., camp-on. The hosts must also now contend with busy signals if the destination is busy talking to someone else. The use of crossbar switches in computer networks is quite new, but holds great promise, but some new methods must be developed to achieve these promises.[8]

4. FUTURE HIGH-SPEED COMPUTER NETWORKS

4.1. Wavelength division multiplexing

Wavelength division multiplexing (WDM) looks very attractive for future high-speed computer networks.[9] WDM can be compared to tuning your television (TV) set. Many TV channels come into your set over a single cable, and you pick one to view by selecting the frequency of that channel. WDM is similar in that it uses multiple wavelengths of light over a single fiber, and the receiver selects the appropriate wavelength.

Figure 5 shows one version of a WDM network interconnecting four hosts. WDM is also similar to the crossbar switch in that multiple conversations can occur simultaneously, each using a different wavelength. Like a crossbar, a network based on WDM can be characterized as having distributed memory, distributed control, and multiple data paths.

In figure 5, each host transmits on a fixed wavelength, $\lambda 1$ through $\lambda 4$. At each receiver, the tunable filter selects the appropriate wavelength to listen to a specific transmitter. Another version of a WDM network would have each receiver set to a unique single wavelength, and the transmitters tune to the different wavelengths. Still another version would have both the transmitters and receivers tunable. The network can theoretically have a very large number of channels, e.g., 2500 channels, each 1 GHz with 9 GHz guard bands. This is based on a center wavelength of 1.55 nm and tuning from 1.45 nm to 1.65 nm.[10]

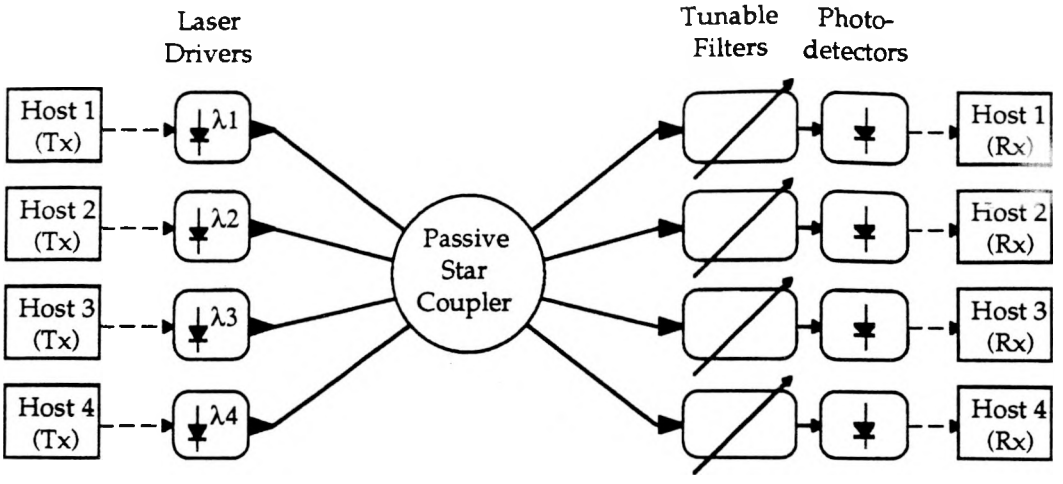

Fig. 5. Wavelength division multiplexing network

There are some basic problems that need to be solved before WDM becomes practical for computer networks. The tuning needs to be fast (less than 1 microsecond) and accurate (to get the maximum number of channels). There also needs to be minimum crosstalk (for the maximum number of channels and adequate bit error rate). There also need to be in-line broadband amplifiers to overcome the losses of the star couplers. Ways to distribute the star coupler to the end points would also help. And of course, the parts need to be inexpensive and mass producible (hand selection of laser wavelengths is not acceptable).

Todays computer networks use what is called "in-band addressing", i.e., the destination address is carried along with the message, not routed on a separate control path. Also, most of todays computer networks use packet switching with datagrams as the underlying transfer mechanism. Here each message is a separate entity with addressing and error control portions. Rather than using packet switching, WDM networks seem to lend themselves more towards circuit switching. With circuit switching a pilot message is sent out to establish a path (circuit) between the source and destination. Once the path is established, the data message can then be transmitted. This circuit set-up adds to the message latency.

With WDM the path must be established, i.e., both the transmitter and receiver must be using the same wavelength, before the message packet can be sent. If you are using tunable receivers and fixed transmitters, then how does the receiver know when a transmitter wants to send something to it so that the receiver can tune to the transmitter's wavelength? Likewise, if the receiver is fixed and the transmitter tunable, then how does the transmitter know that someone else isn't already transmitting on that wavelength? If someone else is transmitting on this wavelength, then the messages will collide resulting in neither message getting through correctly. There are ways to solve this "media access" problem, but most of them require some sort of "out-of-band addressing" at different wavelengths.[10] The problem is not insurmountable, it is just a problem. This media access problem will affect the latency from source to destination.

4.2. Optical crossbars

Optical crossbars function like the electrical crossbars described in 3.4. The advantage of optical crossbars over their electrical counterparts is that most of the components can be passive (improving reliability), and packaging may be simplified (especially at higher speeds). A problem that needs to be addressed is how the packet address is extracted from the packet and used to control the switch. In electrical networks this is done with integrated circuits examining 48-bit addresses and doing look-ups to find the appropriate path to forward the packet, and in the meantime storing the packet in a buffer memory while the decision making is being done.

5. COMPUTER AND COMMUNICATIONS "CULTURES"

The telcom networks and computer networks have traditionally used different techniques. The telcom networks have effectively used circuit switching and time division multiplexing of many slow channels to a single fast channel. The computer networks have used packet switching with datagrams, where each packet takes the total bandwidth of the media. The telcom networks have been very concerned with guaranteed bandwidth so that the data is not delayed, for example causing uneven time delays in speech traffic. The computer networks were less worried about incremental delay, and were more concerned with making use of all of the available bandwidth.

Now we are seeing the two "cultures" starting to merge. The computer networks need some of the guaranteed bandwidth circuit switching techniques to transmit video and voice among the end nodes. Likewise, the telcom networks are becoming digital and using small packets, e.g., 53-byte cells in SONET, for carrying multiple traffic streams.[11,12] The telcom networks still need a call set-up to load the address translation look-up tables in the route.

Even within the computer "culture" there are subcultures, e.g., communications people who are comfortable with datagrams that can disappear in transit, and channel people who do not anticipate anything dropping on the floor between the host computer and the peripherals.[13]

6. SUPERCOMPUTER NETWORK NEEDS

Supercomputers are the leading edge of the computer industry. The performance of a supercomputer of ten years ago is now available in a desktop workstation. This performance growth does not seem to be tapering off much. Hence, what is required for supercomputers today will probably be needed for the next generation of workstations. The computer networks of today have offered cost effective solutions to varying problems. Each architecture has its own bottlenecks and drawbacks, as well as benefits. The users must consider their own applications and needs to select the most appropriate architecture for their systems.

6.1. Latency

Latency is the time from when the source transmits the first byte of information until the destination sees that first byte. Latency is especially important when the computers are closely coupled working on a common problem. The next generation of supercomputers are going to be multiprocessors - we are running into speed of light limitations with single processors. Not only will single cabinet supercomputers use multiple processors, but pseudo supercomputers will be formed by networking together workstations and mainframes. Hence, latency is becoming a more important issue in computer networking.

Media access affects latency, especially if it is a shared media like a bus. With a ring, you need to wait for the token before you can transmit. In a store-and-forward switch the source can transmit quickly, but the central control may delay delivery while it is handling packets from other sources. Crossbar switches can be quick, but will bottleneck if everyone is trying to send to the same destination.

Time of flight was not a major problem before, but is becoming a problem as higher speeds and longer distances are used. For example, the National Research and Education Network (NREN) being funded by NSF and DARPA aims to have a 10 Gbit/s backbone network across the United States in less than ten years. If you are transmitting at 800 Mbit/s (HIPPI speeds today) from New Mexico to California (as we will be doing in the Casa NREN testbed) then there can be over 1 MByte in transit just due to time of flight. One of the goals of the NREN is to closely couple machines at different sites to work on a common problem.

6.2. Total bandwidth

A high total bandwidth is necessary so that the processors do not get backed up waiting for more data, or trying to get rid of data they have generated. The total bandwidth needs also grow as more and more systems are interconnected. Multiple paths, either with crossbars or WDM seem to be answer. Not only must the bandwidth be available, but the data must be transferred with few errors - errors cause retransmissions that further eat into the total bandwidth, and delays for the computers involved.

6.3. Security

Security concerns include someone getting the data that they shouldn't, or someone clogging up the system so that the network is unavailable to deliver any data. This implies robust networks, networks with good management capabilities, data encryption, and good password schemes. At the higher speeds these are all difficult.

6.4. Standards

The computing industry has become aware that hardware and software standards are necessary for future growth. No single company can provide all of the solutions, and interoperation with other vendors requires agreed upon interfaces. The users are also demanding conformance to standards so that they can purchase from multiple vendors, and minimize their training costs.

Some years ago some people thought that standards stifled creativity. It is our observation that standards allow a company to invest a larger amount in their own areas of special expertise, with a smaller investment required to interface to multiple other vendors that conform to the standard. Otherwise, the cost of separate interfaces to every other vendors products may well outweigh the cost of the main business.

We have also seen that the standards process usually brings together the best and brightest people of many companies to work collectively on a problem. Design by committee really does work; the output of a standards committee is usually considerably more thorough and of higher quality than if one person or one company had done the complete job.

In the gigabit computer networking arena, the High-Performance Parallel Interface (HIPPI)[4,5] and Fibre Channel (FC)[13,14] are examples of lower level interfaces currently in the standards process. Higher layers are being addressed by other standards bodies, with most of them following the Open Systems Interconnect (OSI) model developed in the International Organization for Standardization (ISO). Synchronous Optical Network (SONET) is an example of standardization of higher speeds in the telcom industry.

7. Conclusions

Computer networks are a growth industry, and are becoming faster and more important. The trends seen in today's supercomputer networks will be used in the mainstream computer networks of tomorrow. Computer network architectures are changing to accommodate new needs as computer I/O channels approach gigabit per second speeds. Future networks will utilize more fiber optic components for longer distances and improved reliability and error characteristics. New architectures based on WDM and other advanced optical techniques will appear when the components become available. It is unclear how ATM will influence future computer networks.

8. ACKNOWLEDGEMENTS

The Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36. This work was performed under auspices of the U.S. Department of Energy. This paper is LA-UR 91-????.

9. REFERENCES

1. K. H. Winkler, M. L. Norman, and J. L. Norton, "On the Characteristics of a Numerical Fluid Dynamics Simulator," *Austin Symposium on Algorithms, Architectures, and the Future of Scientific Computation,* March, 1985.
2. IEEE 802.3 Standard.
3. F. E. Ross, "FDDI - A Tutorial," *IEEE Commun. Mag.*, vol. 24, November 1986, pp. 10-17.
4. High-Performance Parallel Interface - Mechanical, Electrical, and Signalling Protocol Specification (HIPPI-PH), *American National Standards Institute,* X3.183-1991.
5. D. E. Tolmie, "The High-Speed Channel (HSC) Standard," *Proc. of COMPCON Spring '89*, pp. 314-317.
6. R. L. Hobelheinrich, and R. G. Thomsen, "Multiple Crossbar Network: A Switched High-Speed Local Network,", *Proc. of 14th Conference on Local Computer Networks*, Minneapolis, October, 1989, pp. 285-291.
7. J. Shandle,, "Gigabit nets get ready for lift-off," *Electronics*, December 1989, pp. 66-70.
8. High-Performance Parallel Interface - Switch Control (HIPPI-SC), *X3T9.3 Preliminary Draft American National Standard.*
9. P. E. Green, "The Future of Fiber Optic Computer Networks," *COMPUTER*, September, 1991.
10. N. R. Dono, P. E. Green, K. Liu, R. Ramaswarni, and F. F. Tong, "Wavelength division multiple access networks for computer communication," *IEEE Jour. Sel. Areas in Comm.*, vol 8, no. 6, 1990.
11. P.E. White, "The Role of the Broadband Integrated Services Digital Network," *IEEE Communications Mgazine*, vol 29, no. 3, March, 1991.
12. D Delisle, L Pelamourgues, "B-ISDN and how it works," *IEEE Spectrum*, vol28, no. 8, August 1991.
13. R. Cummings, "New era dawns for peripheral channels," *Laser Focus World*, September 1990, pp. 165-174.
14. Fibre Channel - Physical Level (FC-PH), *X3T9.3 Preliminary Draft American National Standard.*