Gigabit Local Area Networks: A Systems Perspective

Their burstiness and large bandwidth mismatches demand new network architectures, but gigabit LANs offer enormous payoffs

H. T. Kung

High-speed local area networks (LANs) represent a key infrastructure ingredient for advanced information technology development. They are critical for various high-performance computing and communications systems. Via high-speed LANs, these systems can become easily accessible and widely used.

Recent advances in network technology have made it feasible to build gigabit LANs. Links in these networks are capable of operating on the order of 1 gigabit per second (1 Gb/s) or higher rates, and thus have at least 100 times more bandwidth than today's 10 Mb/s Ethernets. Generations of LANs, in terms of their speeds, are depicted in Fig. 1.

Gigabit LANs will have a revolutionary impact on applications. With these networks, many important applications (e.g., imaging and distributed computing) will no longer be limited by network speed. More importantly, new applications enabled by these networks will emerge, and will change the fields of computing and communications in a fundamental way.

Despite their importance, the development of high-speed LANs does not appear to be well understood. Unlike many other technology areas, developing any new kind of high-speed network is inherently an involved system issue, in the sense that it is intimately related to a large number of subjects. These include host computer architectures and operating systems, network switching architectures and transport technologies, network protocol standards, applications software, and interoperation with other existing networks. Only when sufficient advances are made in all these areas will the benefits of gigabit LANs to end users be fully realized. Thus, developing gigabit LANs is much more than just pushing bits over some physical media such as fibers at gigabit rates; it is a systems challenge.

This article presents a broad overview of gigabit LANs from a systems perspective. It gives motivations and technical goals of gigabit LANs, describes the challenges of coping with highly bursty traffic and large bandwidth mismatches between network links, discusses major systems issues, and presents some possible solutions.

The article focuses on concerns unique to gigabit LANs, especially issues which differentiate them from gigabit wide area networks (WANs), metropolitan area networks (MANs), and lower-speed LANs. Discussions will be informal and kept at an overview level. The paper will not dwell on issues whose importance is not unique to gigabit LANs, such as network security and management.

H. T. KUNG is Gordon McKay professor of electrical engineering and computer science at Harvard University.

Gigabit LANs and Technology Status

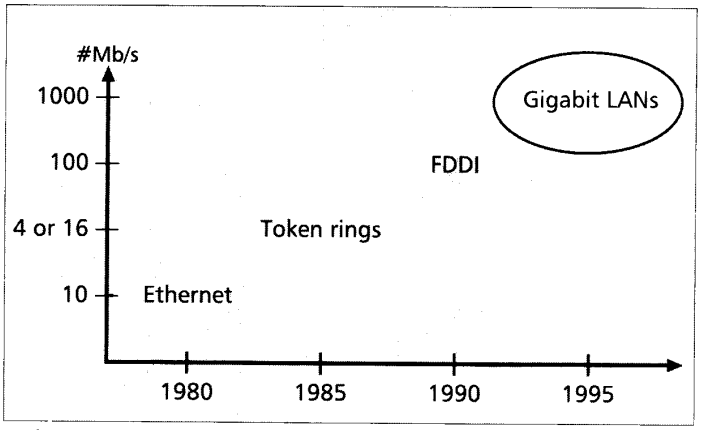
What Is a Gigabit LAN?

A LAN, as defined by the IEEE 802 LAN model in the early 1980s, is a data communication system allowing a number of independent devices to communicate directly with each other, within a moderately sized geographic area over a physical communications channel of moderate data rate. This definition still applies to gigabit LANs provided that moderate data rates is properly interpreted.

LANs differ from WANs or MANs in usage, functions, and economics [20]. Unlike WANs and MANs, which generally use lines and switches provided by carriers, LANs typically use user-installed equipment. LANs have emphasized data communication between computers rather than voice communication. Minimizing cost per host connection has been critical to the acceptance of LANs. To take advantage of the small, propagation delays of media and the simplicity of being within one administrative boundary, special protocol features such as the data link level support of broadcasting have been developed for LANs.

A gigabit LAN is a LAN for which the physical communication medium has a peak bandwidth on the order of 1 Gb/s or higher, and for which an end user is able to realize this gigabit performance. As Fig. 2 depicts, a gigabit LAN connects to hosts, routers, various adapters, etc. These devices connect to the LAN at the 1 Gb/s

 0163-6804/92/$03.00 1992© IEEE824-7504



■ **Figure 1.** *Evolution of LAN speeds.*

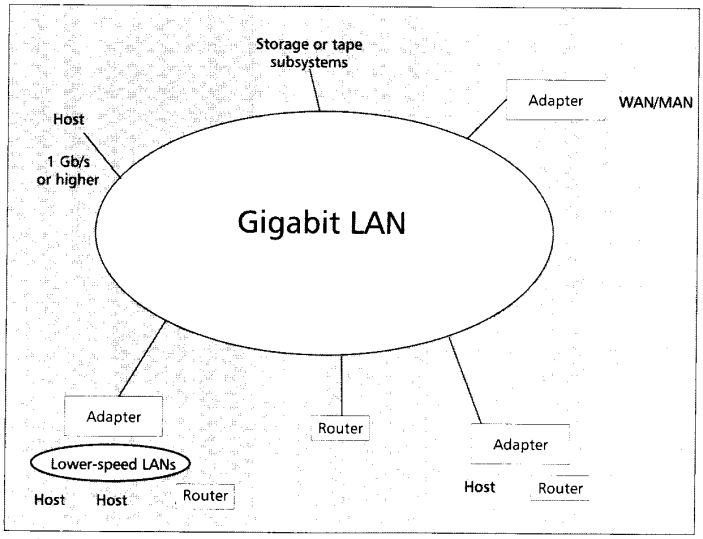
or higher rates, or at a lower speed via adapters. In the second case, the gigabit LAN is able to provide a high-bandwidth backbone to carry the aggregate traffic from a number of devices or other LANs. According to the above definition, there are gigabit LANs already in operation [17]. However, these LANs are based on proprietary protocols and architectures, are quite expensive, and are not easily scalable to include many hosts. They usually are for special applications such as connections for supercomputers.

Some Relevant Standards

Gigabit LANs based on standards are emerging. An example is gigabit LANs based on the high-performance parallel interface (HIPPI) protocol, whose physical layer specification [19] has recently been approved as an ANSI standard. These HIPPI-based LANs, which have become popular in supercomputing centers, support data rates of 800 Mb/s and 1.6 Gb/s. Almost all commercially available supercomputers and parallel machines support the 800 Mb/s HIPPI interface, as do many mainframes.

In addition to HIPPI, there are a number of other high-speed network standards in various stages of development by standards bodies, some of which are briefly described here. The asynchronous transfer mode (ATM) telecommunication standard at rates of 155 and 622 Mb/s has received substantial attention for applications in LAN environments, beyond the public WANs for which ATM was originally targeted. The recently established ATM Forum already has approximately 60 participating companies. The fibre channel standard (FCS), at rates of 100, 200, 400, and 800 Mb/s, has been defined to support high-speed computer I/O, storage systems, and other applications. The FFOL (FDDI Follow-on LAN) effort is developing standards at 622 Mb/s, 1.2 Gb/s, and 2.5 Gb/s.

All these standards efforts include corresponding specifications for physical media and framing. A notable one is synchronous optical network (SONET), which can transport ATM and other traffic. With standardized frame formats, SONET networks adhere to a hierarchy of interface rates which are multiples (called OC-N) of a basic signal rate of 51.84 Mb/s (OC-1). OC-3 (155.52 Mb/s) and OC-12 (622.08 Mb/s) have been designated as the customer access rates in future B-ISDN networks. Other important SONET rates are OC-48 (2.488 Gb/s), and possibly OC-192 (9.953 Gb/s) [22].



■ **Figure 2.** *Gigabit LAN connected to various devices.*

Gigabit Technology Status

The status of hardware technology required by gigabit LANs can be illustrated by advances in SONET hardware components (although they were originally developed for WANs). Basically, hardware components operating at gigabit rates are available [22]. Although their current prices are high, they are expected to drop as volume sales develop (such as has been achieved for Compact Disc lasers). If gigabit LANs and WANs can share hardware components rather than just technology, the synergy will help increase volume of the components and thus reduce their costs. The SONET standard will play a key role in helping to reduce the costs as will VLSI implementation of SONET functions including multiplexing and demultiplexing.

Recent Results in Optical Amplifiers

There has been some significant progress in optical amplifiers recently [25]. The new results in optical signal amplification in the 1.3 micron (1300 nm) spectral region could lead to reduced cost for fiber optic systems. Optical technology has enjoyed over 10 years of commercialization of the 1300 nm technology, which is widely used for gigabit transmission over standard single-mode optical fiber, but until these recent results, optical amplifiers operated only at 1500 nm.

The potential system implications of 1300 nm optical fiber amplifier n/s can be especially significant for gigabit LANs. New architectures using these optical amplifiers will substantially lower the cost and increase the performance of both interoffice and distribution fiber networks. For example, gigabit optical star interconnects, which are fundamentally useful in supporting LAN broadcasting, will become inexpensive and easy to implement with these 1300 nm amplifiers. Moreover, these amplifiers have the potential to lead to practical and inexpensive use of wavelength-division multiplexing (WDM) in systems.

Application of Gigabit LANs

Many existing applications require high-speed networks. Gigabit LANs will enable new applications, and will make the implementation of some existing applications much easier. Three areas in which gigabit LANs are essential are high-speed LAN backbones, high-performance computing environments, and distributed computing and workstations; these areas are discussed briefly below. If their requirements can be met, then the high-speed networking needs in many other important areas such as imaging, network multimedia, and distributed manufacturing and engineering also will be satisfied.

High-Speed LAN Backbones

LAN backbones for campus networks typically need to have at least one order of magnitude higher bandwidth than individual stations on the network. Costly T1 or even T3 lines are routinely being installed in industry complexes just to meet a fraction of today's backbone bandwidth requirements. The 100 Mb/s FDDI ring is becoming a major system in the LAN backbone market. In the near future, however, many LAN backbones must have bandwidths much higher than 100 Mb/s, due to emerging high-performance workstations and high-speed host interfaces such as HIPPI mentioned above.

High-Performance Computing Environment

High-performance computing environments of the future will be network-based. As illustrated in Fig. 3, such an environment includes a variety of computing resources. Using high-speed networks, these resources can work together to speed up both computation and I/O, and can incorporate a large amount of memory. Moreover, these environments can take advantage of existing computer architectures, while providing a graceful migration to new ones [3]. Many of today's supercomputing centers already are configured around high-speed networks such as HIPPI-based LANs.

In a more tightly coupled environment, this type of system may be called a multicomputer [4]. Such systems view the network as a backplane. Because of their flexibility in system configuration and relatively low cost, multicomputers built around a "network plane" have developed into an important class of parallel computers [6]. In this case, the network is viewed as a natural extension of the traditional computer "bus backplane".

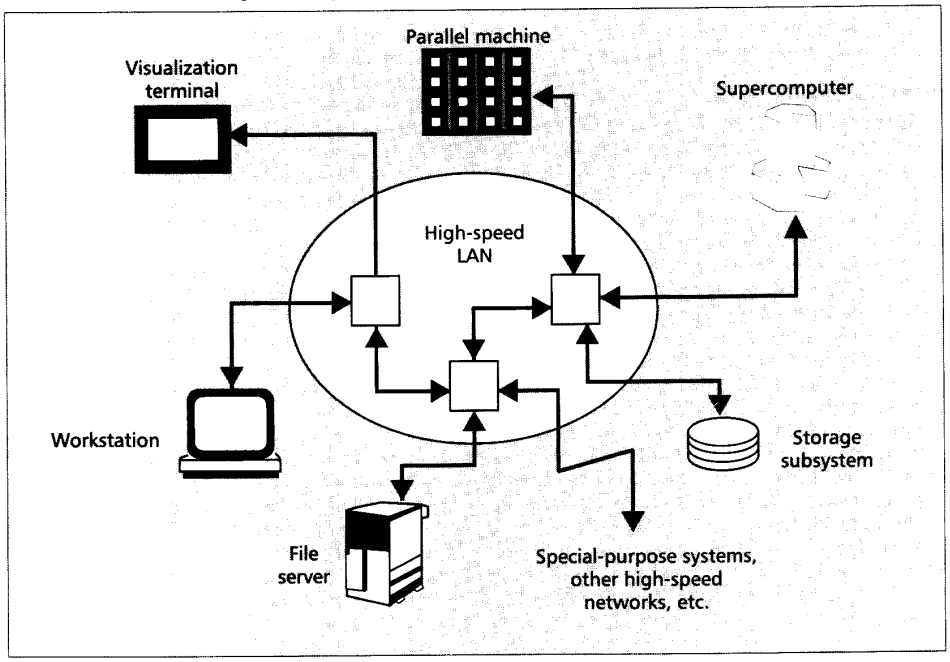
Distributed Computing and Workstations

Distributed computing splits computation among machines on a network. There are several models for distributed computing on a LAN. The client/server model is among the most important ones. This model allows the use of a few file and compute servers and many inexpensive user machines. Distributed computing also supports collaboration over geographically dispersed sites. As networks get faster, more and more data is being shared over networks. Results generated at one site can be displayed at others. In addition, distributed computing has the potential to provide fault tolerance, high availability, and load balancing.

Network performance is a key issue in distributed computing. High-bandwidth communication is needed to supply data to high-performance network hosts. Historically, these high-performance network hosts have achieved performance levels in the 10 to 50 MFLOPS or MIPS range and required approximately 10 megabytes per second network I/O capability to support those performance levels. We expect that supporting a high-performance computer capable of performing 1 billion operations per second will, therefore, generally require a multi-Gb/s network.

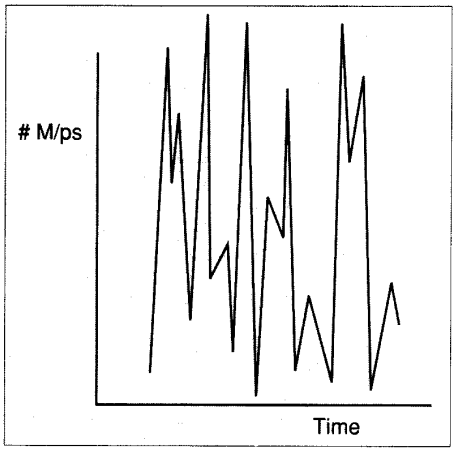
In a tightly coupled environment, low-latency communication is necessary to exploit fine-grain par-

Flow-controlled virtual connections provide an efficient means of directly propagating congestion information back to traffic sources.



■ **Figure 3.** *Network-based high-performance computing environment.*

Important goals for gigabit LANs include low-latency communication, and guaranteed and robust performance.



■ **Figure 4.** *Highly bursty traffic expected over gigabit LANs*

allelism, and hence a higher degree of parallelism, across different network hosts. With smaller latency and finer-grain parallelism it is possible to parallelize more applications efficiently and create a high-performance computing environment from a distributed set of processors [5]. As will be shown later, gigabit LANs make it feasible to have low-latency communication.

Many current workstations already have the raw hardware capacity to keep up with high-speed LANs, and thus support high-performance distributed computing. Examples of existing workstation I/O buses, with sustained bandwidths of 100 Mb/s or more, include DEC Turbo Channel, HP SGC I/O bus, IBM SIO and MicroChannel, and SUN S-Bus.

Systems Challenges

The high speeds of gigabit LANs impose a new set of challenges, not shared by gigabit WANs and MANs, or by lower-speed LANs. New solutions are required to meet these challenges.

Highly Bursty Traffic

Traffic on gigabit LANs is expected to be highly bursty. Data traffic, which will constitute most of the load on gigabit LANs, is intrinsically more bursty than voice traffic. As network speeds increase, the peak rates will increase faster than the average, thus making traffic become even more bursty.

A single traffic source (e.g., a supercomputer with an 800 Mb/s HIPPI interface) will be able to pump data into a network at a very high speed and consume a large fraction of the peak bandwidth of links throughout the network. On the other hand, the traffic source can complete its data transmission in a short time because data is transmitted at such high rate. Once the transmission is complete, the network load will suddenly drop sharply. Therefore, with the presence of very high-bandwidth traffic sources, the network must be prepared for the large increase in network load fluctuations. In general, traffic over any network resource (such as network link, access port, or switch port) is expected to be highly bursty over a wide range of time scales [2]. For each time scale, the bursty traffic is expected to resemble the pattern in Fig. 4.

Increased Mismatches in Bandwidth

When the peak speed of links increases in a network, so may bandwidth mismatches in the network. For example, when a 1 Gb/s link is added to a network which includes a 10 Mb/s Ethernet, there will be two orders of magnitude difference in their speeds. When data flows from the high-speed link to the low-speed one, congestion will occur quickly.

Similarly, there will be an increased mismatch between the network's speed and the speed at which a destination host or device can receive data. In a general, high-speed network environment it is common to have some relatively slow hosts which cannot keep up with the peak network speed. Even a fast host may not be able to keep up with the network, as it may be occupied with processing data received previously from the network or with some other computing tasks. For example, a printer on the network may take longer to process and print more complicated pages (e.g., those containing images). As networks become faster, this kind of mismatch will become even more severe.

A Network Pitfall Scenario

Despite their high bandwidths, gigabit LANs still can collapse quite easily due to congestion if the network is not properly protected. Consider, for example, a network file system [11] based on a simple transport protocol such as user datagram protocol (UDP) [1], which does not have a built-in congestion control mechanism. Whenever network congestion causes packet loss, the file system will retransmit the lost packets regardless of the degree of congestion. The retransmitted packets actually increase network load and, consequently, congestion will persist and can only become worse until the network stops working completely. Note that even transient congestion can trigger this catastrophe.

The highly bursty traffic and increased bandwidth mismatches expected in gigabit LANs, as discussed previously, will increase the chance of transient congestion. It therefore becomes absolutely imperative for gigabit LANs to ensure that transient congestion does not persist and evolve into permanent network collapse.

For the network file system example above, a possible solution could be replacing UDP by a more sophisticated transport protocol such as TCP which, by establishing per connection state at both end hosts, will perform application-independent congestion control and end-to-end flow control [18]. This can substantially reduce the chance of congesting the network with retransmitted packets. This solution may work only if all applications on the network use such network-friendly protocols.

These protocols, however, may not be suited to some applications, which need to use simple protocols such as UDP for efficiency and flexibility. For example, the common case for Internet name servers is short queries, short responses, and no packet loss; this is common especially on LANs. Under these conditions, a name can be resolved with only two UDP packets, whereas TCP would require seven or more packets to establish the connection, transfer data, and close the connection. Moreover, a simple protocol allows the application itself to have the flexibility to choose

proper packet retransmission policies. For instance, depending on the application's state and the type of lost packet, the application may decide the best time to retransmit the packet, or even discard the packet (as in some video applications). Therefore a challenge in network architecture and protocol design is protecting the network from irrecoverable congestion while realizing high network utilization and providing applications sufficient efficiency and flexibility.

Additional Design Goals for Gigabit LANs

Besides supporting high-bandwidth data transfers, other important goals for gigabit LANs include low-latency communication, and guaranteed and robust performance.

A vital concern in the design of a gigabit LAN is the minimization of latency. If the elapsed time to send and receive a message is short, it becomes possible to divide computations into finer grains, thus making it profitable to use more network hosts for a single application and to fruitfully parallelize a wider variety of computations.

The communication latency of delivering a packet to a destination host is the sum of two quantities: 1) the packet header delay, which is the time to deliver the first bit of the packet; and 2) the packet data delay, which is the time to deliver the remainder of the packet. The packet header delay is bounded below by the medium propagation delay, which is approximately 5 μs per km over a fiber. For a network with a relatively large span (say 10 km), the propagation delay is about 50 μs. Actual delays, however, can be hundreds of μs or more because of protocol and host software overheads at the two ends [9, 16]. On the other hand, the packet data delay is determined by the packet size and network bandwidth. On a slow network, the packet data delay is large even for small packets. For example, on a 10 Mb/s network, the packet data delay for a packet as small as 100 bytes is about 100 μs. Thus, for networks with a large geographic span or with a small bandwidth, a relatively high communication latency is inherent.

However, for gigabit LANs, which operate in a local area and have a large bandwidth, the communication latency can be made small. The packet data delay can be as small as 1 μs or lower for 100-byte packets. As stated above, the medium propagation delay for a 1 km fiber optic LAN is only about 5 μs. Thus, reducing the packet header delays, which are hundreds of μs or more with today's LANs and hosts, makes a huge difference in minimizing the overall communication latency.

Many applications, such as network multimedia, require guaranteed bandwidth and latency even in the presence of unexpected transient behavior, and need to work around failures of individual network components.

The mechanism used to achieve these goals should not itself create significant side effects such as excessive transmission retries or long delays while messages wait for resources. The traditional approach of simulating the network under assumed typical loads such as Poisson, although providing valuable insight, cannot accurately reflect highly bursty workloads and cannot verify performance guarantees. Thus, we will have to use new types of analysis to prove certain properties of the network. Moreover, as to be discussed later, new network architectures with built-in mechanisms to support guaranteed performance (e.g., virtual connections which can be scheduled according to their priorities) may be useful.

Old Solutions No Longer Applicable

It is likely that traditional congestion control methods will no longer work well for gigabit LANs. Some of these traditional methods are described below, with reasons why they may no longer be applicable.

Statistical Load Prediction. This method has worked well in some traditional WAN trunks carrying many lower-speed traffic streams such as voice traffic. Operating in a local area, however, a LAN inherently carries a smaller number of streams and hence its traffic cannot accurately be modeled as a statistical average. Moreover, in a gigabit LAN the variation in network load caused by just one or few traffic sources can increase tremendously when the speed of access links increases by a large amount, say from 10 Mb/s to 1 Gb/s.

Generally, it is inappropriate to use the law of large numbers to model loads for gigabit LANs. This makes useful network load predictions almost impossible. As mentioned above, a single supercomputer computation, or even an application on a high-performance workstation, can drastically alter traffic patterns. Thus, methods such as statistical multiplexing and steady state modeling, which work well in some existing WANs for aggregate low-speed traffic, will not necessarily work in the gigabit LAN environment.

Over-Designed Network Capacity. This approach is based on the premise that solutions to network congestion are difficult only if they also attempt to maintain a high network utilization. Thus, by deliberately over-designing the network capacity (e.g., by 300 percent), network utilization is sacrificed so that simple congestion control mechanisms can be employed.

This approach has worked well for constant bit-rate traffic at moderate speeds. However, for highly bursty traffic produced by very high-bandwidth traffic sources, it would be impractical to over-design networks. For example, just to accommodate a single 800 Mb/s HIPPI source, an over-designed network would require multi-Gb/s bandwidth. Today's workstations already can pump data into the network at hundreds of Mb/s, and this rate will increase as LANs speed increases. We expect that, for the foreseeable future, a single high-performance workstation host, or a small collection of them, always will be able to saturate the highest bandwidth that networks can practically support.

Admission Control. This method avoids overloading a network by applying admission control to traffic sources. Sessions can be established only if the method determines that all the required network resources will have the budget to handle the load associated with the sessions, or at least have a high probability of being able to do so.

For a high-bandwidth traffic source such as an 800 Mb/s HIPPI host, it is unlikely that the method can admit any large fraction of the traffic source's peak bandwidth, if other traffic of significant bandwidths are also to be accomodated. As a result, the traffic source is most likely given only a small fraction of its bandwidth at the

It is likely that traditional congestion control methods will no longer work well for gigabit LANs.

One reason for using switch-based architectures is that the limited capacity of a shared-medium architecture will be inadequate for gigabit LANs.

admission time. Without fast feedback from the network to the traffic source, network bandwidth may be wasted because the traffic source cannot be notified in time about releases of network resource by other traffic. Therefore, this admission control method can work well only if new network architectures capable of fast feedback (such as those to be described later) are employed.

Network Buffering. Buffering is another typical method for dealing with network congestion problems. In this approach, buffers are placed at various places in the network to hold excessive traffic. When bandwidth mismatches increase, however, network buffering becomes less effective or even harmful. First, since buffers can now be filled quickly, many large and high-speed buffers will be needed. Second, once traffic is buffered, the network is committed to handle it somehow. In particular, the buffered traffic eventually needs to be cleared out from the network, and this requires time. Thus, the more data that the network buffers, the more delays the network will cause for future traffic, which may actually have a higher priority. Generally speaking, large network buffers designed to hold excessive traffic are undesirable because they cause side effects that significantly compromise high-speed network performance.

We see that the above methods by themselves, even if they are used together, are insufficient to meet the systems challenges of gigabit LANs.

New Systems Solutions Required

The deployment of gigabit LANs, which fulfill the application needs and address the systems challenges discussed above, will require significant technological advances. It is important to realize that high-performance network communication requires a system solution, not just a hardware or software solution.

For example, improvements in individual network protocols, although necessary, are not sufficient. Many actions which take place on hosts are invisible at the protocol level (e.g., data placement and formatting, memory allocation, bus protocols, processor context switches, and scheduling). All present obstacles to the ideal of passing information between distant user processes. It is the systems environment (in which protocols are executed) that makes the major performance difference.

The remainder of this paper will address systems issues, and some solutions will be described. The discussions will be centered around architecture and interface aspects of gigabit LANs.

Gigabit LAN Architectures

As previously discussed, new architectures are required for gigabit LANs. This section describes some new directions in gigabit LAN architectures.

Use of Switches

Although the telephone network typically has used a physical and logical switched star architecture, today's LANs generally use a shared medium, multiple-access architecture (ring or bus) as typified by Ethernet or FDDI. The situation is expected to change with gigabit LANs, which likely will use switch-based architectures instead of traditional shared media architectures. The recent movement toward physical stars, usually called hubs, has been motivated by several factors, including the need to easily and centrally isolate faulty or misbehaving links. The next step is a move from the hub structure to a switch.

One reason for using switch-based architectures is that the limited capacity of a shared-medium architecture will be inadequate for gigabit LANs. This is especially true as the supported access rate approaches the speed of the medium. In the foreseeable future, it is unlikely that shared-medium architectures will be able to support a gigabit LAN with a large number of users each having 1 Gb/s access. Even optimistic projections of advanced electronics limit shared media electronic components to the order of 10 Gb/s, which is still only 10 times the basic access rate of 1 Gb/s. Recently proposed multiple-wavelength optical networking architectures have the potential of offering the required bandwidth. However, these will be costly until the price of multiple wavelength optoelectronics technology is substantially reduced. Fundamentally, it is not cost effective for a user to have a gigabit access to a shared medium while receiving only a fraction of the bandwidth whenever there are other competing loads on the network.

Switch-based LAN architectures provide some additional advantages. Multiple switches can be mesh-connected to form a large network supporting many hosts. Various switch interconnection topologies can be used to fit the traffic patterns in applications. Moreover, a mesh-connection network is highly flexible. A pair of source and destination can be connected with any one of many possible physical routes provided by the mesh. This flexibility in selecting the route can be used to balance the load on the network and avoid faulty subsystems.

Fast Congestion Notification

Any solution to avoid or recover from congestion must ultimately rely on the fact that one or more traffic sources which contribute to the congestion will be notified by the network, and will then take actions to reduce their loads on the network. Buffering inside the network to hold excessive traffic is not a solution, because when the traffic is sufficiently bursty and when the bandwidth mismatches are sufficiently large, buffers can overflow quickly. In addition, large buffers have negative side effects.

For high-speed networks, congestion information must reach appropriate traffic sources quickly, because network overflows can occur within a short time. A tight and direct feedback loop from congested points to these traffic sources is required. Ideally this flow control mechanism should be so effective that excessive traffic can never enter the network, even under highly unpredictable network loads. Since excessive traffic is blocked outside of the network, rather than being accepted and consequently discarded, network resources are not wasted in handling the excessive traffic. Moreover, because excessive traffic never enters the network, it cannot block other traffic and cause delays.

Commonly used indirect congestion feedback methods may no longer be adequate. In these methods, congestion status is indirectly derived by communicating hosts using an end-to-end

transport protocol such as TCP. Network congestion is assumed if the protocol discovers lost packets or experiences increased round-trip delays. The problem is that indirectly derived congestion information may not be sufficiently timely or accurate.

First, discovering congestion only after packets are lost due to congestion is obviously too late to take action to avoid the present congestion. Moreover, the end-to-end feedback is longer than necessary for relaying congestion information from the middle of the network to the traffic source.

Second, end-to-end round-trip delays may not give accurate network congestion information for reasons such as 1) delays may include other overheads, like host delays, not caused by network congestion, 2) delays may vary due to use of different physical routes rather than different degrees of network congestion, and 3) delays reflect the sum of the delays in all the switches and links that the packet has traversed, and thus do not necessarily reflect existence of any single congestion hot spot in the network.

Link-by-Link Flow Control

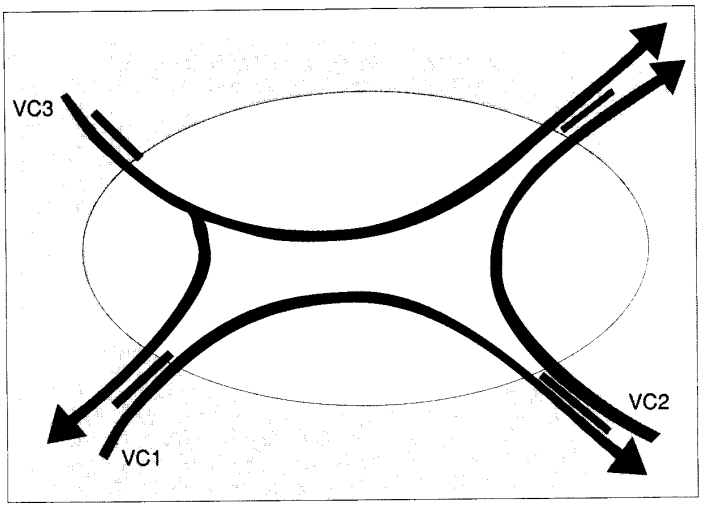
A mechanism which is capable of providing direct congestion feedback is link-by-link flow control. The receiver of each link will send its buffer status information to the sender. Using this information, the sender ensures that it will never send more data than the receiver can hold. (Note that these flow-control buffers need only be sufficiently large to sustain the link bandwidth. They are different from those buffers used to hold excessive traffic as described previously.)

Through link-by-link flow control, back pressure can build up along a connection spanning multiple links. When a traffic source, the starting point of a connection, encounters back pressure, it will stop sending data into the connection until the back pressure is off. For LANs, such link-by-link flow control can be implemented efficiently and inexpensively, because of small propagation delays. We expect that the same method is applicable to wider area networks, but the details need further evaluation.

The link-by-link flow control mechanism described herein represents a new kind of network control, for which new designs are needed. The following are some desirable design goals: 1) the peak link bandwidth can be achieved, 2) the flow control overhead is relatively small, and 3)the flow control is tolerant to transient link failures. Note that, unlike data, the link-by-link flow control information (e.g., the receiver buffer status) is not protected by high-level protocols to ensure reliable delivery, and its loss can have fatal impact on the network's operation. Thus, property 3) deserves special attention.

Cell-Level Multiplexing

A necessary condition for low-latency communication, as discussed previously, is that fine-grain multiplexing be used in time-multiplexing network resources. A typical method of achieving fine-grain multiplexing is to split each packet into small fixed-size data units, called cells, which are multiplexed over network resources. This is the case with ATM networks where packets are transmitted and switched in 53-byte cells.



■ **Figure 5.** *Virtual-connection network (VCN).*

The reason cell-level multiplexing can support low-latency communication is that it allows a high-priority packet to bypass a low-priority packet over the same physical link. That is, the low-priority packet yields the link to the high-priority one at the cell boundary, rather than only at the packet boundary. This ensures that a long packet being sent over a link cannot cause long delays for other packets (i.e., head-of-line blocking cannot occur).

Virtual Connections

A systematic way of viewing the cell-level multiplexing is that it implements a set of virtual connections. Each packet is transmitted over a single virtual connection, and those virtual connections sharing the same network resource are time-multiplexed at the cell level. So packets traveling on these virtual connections are multiplexed at the cell level. The connections are virtual in the sense that there can be a large number of them in the network at any given time, without being bounded by physical bandwidth of links or switches. The cell-level multiplexing of virtual connections is transparent to the user.

VCNs — Putting It Together

In view of the reasons given above for having switches, link-by-link flow control, cell-level multiplexing, and virtual connections, it is natural to consider an ideal network with all these desired properties. This is referred to as the virtual-connection network (VCN) in this paper. A VCN is a switch-based network capable of implementing cell-based, link-by-link flow-controlled, single- or multi-destination virtual connections. A VCN implementing three virtual connections, where VC1 and VC2 are single-destination, and VC3 is multi-destination is illustrated in Fig. 5.

Virtual connections described herein differ from virtual channels in the ATM literature in that virtual connections have link-by-link flow control and can be multi-destination. Virtual connections are most effective for networks, mainly LANs and MANs, where medium propagation delays are relatively small. These networks allow fast, reliable and low-cost implementation of the cell-

level handshaking protocols required by the link-by-link flow control. This link-by-link flow control property distinguishes VCNs from wide-area ATM networks.

VCNs support both connection-oriented and connectionless data link layer traffic. In the connection-oriented case, a virtual connection is maintained over time to transfer multiple packets in sequence. In the connectionless case, a virtual connection is set up solely for the transfer of a single packet. Both connection-oriented and connectionless traffic can co-exist in the network simultaneously.

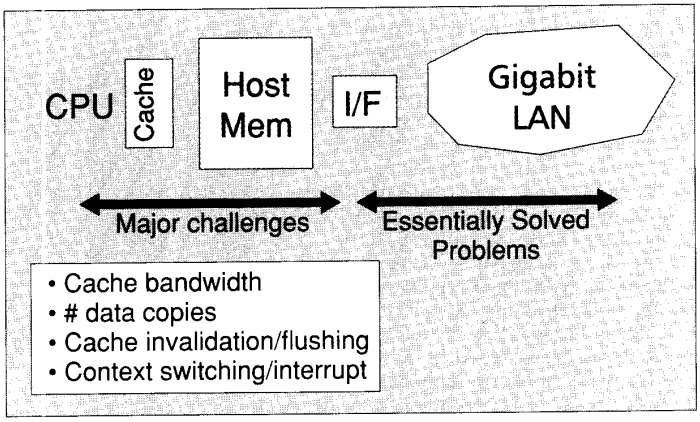
The VCN architecture is an example of a gigabit LAN architecture with the potential of successfully meeting the systems challenges described previously. Some of the advantages and architectural features of VCNs are summarized as follows:

A high-bandwidth traffic source can pump data into the network over a VC at the peak rate of the access link as long as there is no other competing traffic. If there is competition, the rate automatically will decrease to the value determined by the link-by-link flow control over the VC. This allows the traffic source to use the maximal bandwidth available at any given time, and thereby solves the problem related to the admission control scheme previously described.

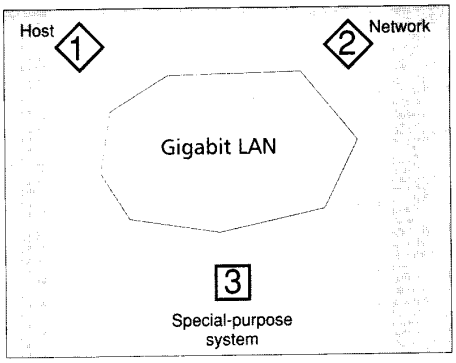
By properly scheduling the traffic of individual VCs, the network guarantees performance in bandwidth and latency for selected VCs.

Through back pressure propagation, virtual connections provide an efficient means of directly relaying congestion information back to traffic sources, and support simple and effective network congestion control schemes. As discussed, the link-by-link VC flow control mechanism will prevent excessive traffic from entering the network. That is, excessive traffic will be blocked at the network's boundary, instead of being allowed to enter the network, and cause difficult congestion problems. The problem of using network buffers to hold excessive traffic, as described before, does not exist here.

Multi-destination VCs provide a natural framework for supporting reliable multicasting or broadcasting. For example, in Fig. 5, VC3 can implement a two-destination multicasting. The multicasting/broadcasting capability is important to many high-performance applications such as distributed computing and network multimedia.



■ **Figure 7.** *Host interface issues.*



■ **Figure 6.** *Three interface types.*

Gigabit Interfaces

Beyond the LAN architectures themselves, many systems issues of gigabit LANs are related to their interfacing with other systems. Note that whenever the sustained bandwidth of an interface is increased, the allowed time for the interface to process a packet must decrease proportionally. For example, to sustain the 1 Gb/s throughput, the interface must process packets at the rate of approximately one packet every 1 μs for 100 byte packets, or every 10 μs for 1 kilobyte packets. This implies that if a processor capable of executing one instruction every 50 ns is used, only 20 or 200 instructions, respectively, are allowed for each packet.

There are several solutions to this stringent processing requirement. The traditional approach is to use a single fast processor. Alternatively, a number of relatively slow processors could be used in parallel. Another approach could use special-purpose circuits to handle those parts of the processing that are simple to implement in hardware, and use the host processor for the remaining processing. Independent of the above options, the interface architecture and the protocol can be streamlined to reduce the number of required instructions per packet.

The performance bottleneck in the long term, however, is expected to be related to the memory speed rather than processing speed, since the former will be unlikely to increase as rapidly as the latter. Interface architectures should be aimed at achieving bandwidths near the peak bandwidth of the memory employed. Basically, there are three types of interfaces for gigabit LANs which are of interest, as depicted in Fig. 6. Special issues in each of these interfaces will be discussed.

Interfacing with Hosts

The primary function of host interfaces is to move data between the attached network and the host memory accessible to the host application. It is well-known that most common transport protocols, in particular transmission control protocol/internet protocol (TCP/IP) [1], do not form performance bottlenecks [8, 16] provided that the protocols are properly implemented.

For the efficient implementation of a protocol, it is essential to minimize the number of accesses to the host memory because memory bandwidth represents a major limitation as pointed out above. In traditional host interface imple-

mentations for Ethernets, there can be as many as six accesses to or from the host memory for each word received from or transmitted to the network. This excessive use of memory bandwidth will be a severe bottleneck when interfacing with a gigabit network. Fortunately, schemes exist which reduce the number of accesses to only one or two for each transmitted or received word [6, 21].

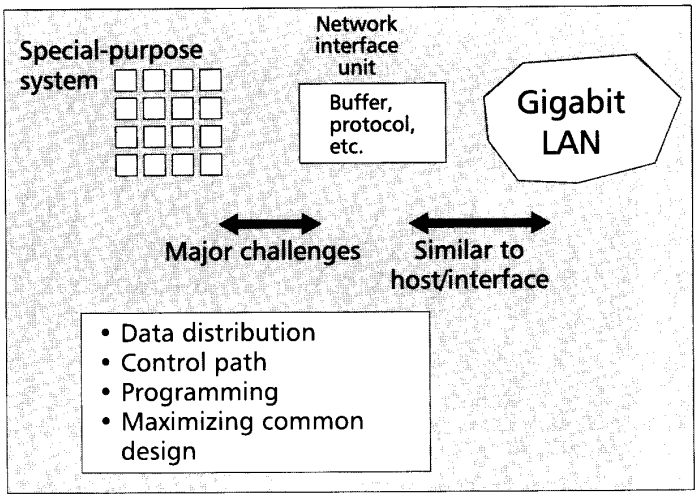
Generally, the host system can support streamlined implementation of protocols by having an environment which reduces copy and buffer operations, pipelines data transfers, providing high-bandwidth DMA, and using efficient implementations of checksumming, error handling, compression, encryption, etc. The current status in host interfaces with gigabit LANs is summarized in Fig. 7. With some minor hardware assistance and a high-bandwidth host memory, today's high-performance workstations can interface to networks at hundreds of Mb/s [13, 14, 21]. In preparing packets for transmission and in acting on received packets, however, the CPU incurs considerable overheads related to the cache architecture, page move/copy, page locking/unlocking, context switching, interrupts, etc. Network researchers need to work closely with workstation vendors to remove these bottlenecks.

Interfacing with Special-Purpose Systems

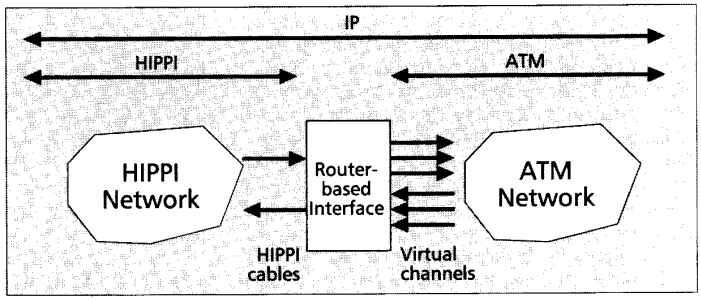
Gigabit LANs can encourage the development of specialized systems optimized for a specific set of applications. Examples of these special-purpose systems include high-resolution printers, high-bandwidth storage systems, high-performance visualization stations, and highly parallel computing engines. With gigabit LANs, these systems can become easily shared resources and thus can afford to be highly sophisticated since their costs can be amortized across many users. The high-performance computing environment described previously and depicted by Fig. 3 is an example of this scenario.

Interfacing special-purpose systems to gigabit LANs, however, imposes some unique issues. Typically, a special-purpose system itself is not designed for high-speed network interface. An additional network interface unit, as shown in Fig. 8, needs to be used to provide required capabilities such as high-performance buffering and protocol processing. An example of such an interface unit is the communication accelerator block (CAB) system under development by Carnegie Mellon University and Network Systems Corporation for connecting Intel's iWarp parallel machine to a HIPPI network.

The issues related to the network half of such an interface are similar to those host-network interface issues discussed previously. However, interfacing with the special-purpose system represents major challenges (some of which are listed in Fig. 8). For example, when interfacing with a parallel processor array, the data distribution from the interface to individual processors on the array usually is done in an ad-hoc manner. There is a lack of high-level protocol support for primitives such as distributing a data array evenly among multiple processors. Identifying common network I/O requirements for different types of special-purpose systems would be a useful first step in streamlining network interfaces with these systems.



■ **Figure 8.** *Special-purpose systems interface issues.*



■ **Figure 9.** *HIPPI-ATM network interface based on IP-router.*

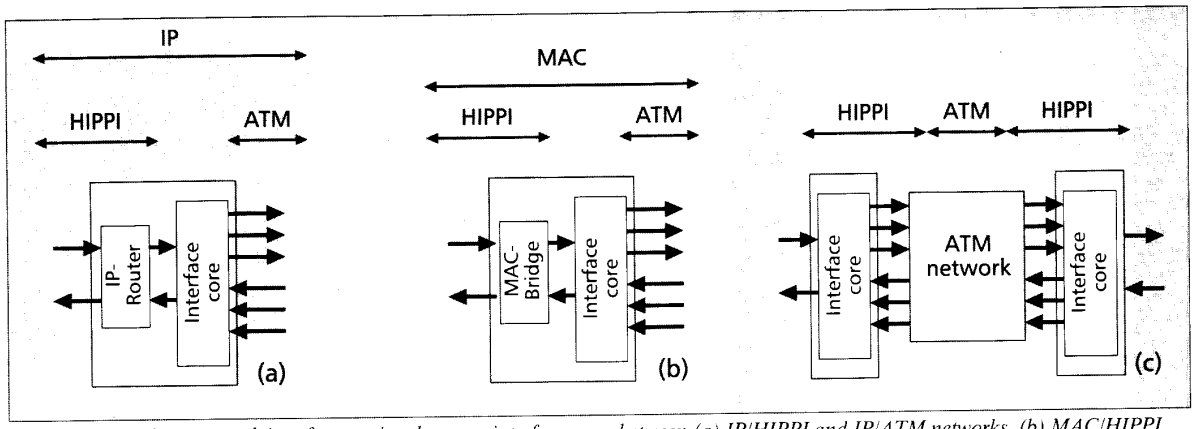
Interfacing with Other Networks

To enter the mainstream, gigabit LANs must interoperate efficiently with various present and future networks. Because they are new networks with many new features, gigabit LANs present many choices concerning their interfaces with other networks, and these options must be carefully studied. However, one thing that is certain is that these interfaces need to support multiple network protocols, as demanded by today's world of networks.

Let us consider the interface between a HIPPI network and an ATM network as an example. This HIPPI-ATM interface will illustrate some key network interface issues for gigabit LANs.

A straightforward HIPPI-ATM interface, as illustrated in Fig. 9, could be a network layer router, say an IP router. This router-based interface works as follows. In the HIPPI-to-ATM direction, the IP packet encapsulated in each arriving HIPPI packet is converted into ATM cells. Then the cells are transmitted over the virtual channel associated with the IP address of the packet. In the ATM-to-HIPPI direction, the arriving ATM cells corresponding to an IP packet are reassembled into a packet according to some ATM adaptation standard. The resulting packet, encapsulated in a HIPPI packet, is transmitted to the proper HIPPI port associated with the packet's IP address.

In contrast to the router-based interface above which supports only the IP protocol, a multiprotocol interface approach is depicted in Fig. 10. This approach is attractive in that it allows the same interface core to be used in a variety of network interface configurations. The interface core



■ **Figure 10.** *Three network interfaces, using the same interface core, between (a) IP/HIPPI and IP/ATM networks, (b) MAC/HIPPI and MAC/ATM networks, and (c) HIPPI networks.*

basically performs the following functions: The first is the translation between HIPPI packets and ATM cells. The second is the mapping of each arriving HIPPI packet to the associated virtual channel in the ATM network, based on the switching address (the I-field in HIPPI terminology [19]) of the packet. Conversely, the interface tags each reassembled packet arriving from the ATM network with the associated HIPPI I-field, based on the identifier of the virtual channel on which the packet arrives. When interfacing with a link-by-link flow controlled network, such as the one described previously, it is useful that the interface be able to propagate back pressure for individual virtual connections.

The interface core can be used as a building block for various network interfaces. Figures 10a and 10b show that, by combining the interface core with two full-duplex pairs of HIPPI channels, an interface between IP/HIPPI and IP/ATM networks can be formed. Figure 10b shows that, by combining the interface core with a MAC-bridge, an interface between MAC/HIPPI and MAC/ATM networks can be formed. (MAC refers to the media access control sublayer of the data link layer in the IEEE 802 standard for LANs [1].) Figure 10 c shows that, by using a pair of the interface cores, two HIPPI networks can be connected via an ATM network.

Note that the interface of Figure 10c allows HIPPI packets to move uninterpreted from one HIPPI network to the other. Unlike the router-based interface of Fig. 9, the interface does not need to perform translations based on network layer addresses, and can transport raw HIPPI packets without having to assume that some higher-layer protocols are used.

On the other hand, when the interface core is connected to IP-routers and MAC-bridges as shown in Figures 10a and 10b, the resulting interfaces can still perform translations based on IP and MAC addresses. Therefore, the approach is fundamentally flexible and not necessarily expensive, because the common interface core can be shared between different network interface configurations. As part of the Nectar WAN testbed research [12, 23], an experimental version of the HIPPI-ATM interface core is being jointly developed by Bellcore and Carnegie Mellon.

This interface core, called HAS (HIPPI-ATM-SONET) [7, 22], uses eight 155 Mb/s STS-3c SONET channels to carry ATM virtual channels.

The HAS experiment has revealed another general issue in the design of gigabit network interfaces: how to design a high-speed interface made of many lower-speed ones. For example, the HAS interface with the ATM network is composed of eight 155-Mb/s channels as mentioned above. Such an interface requires traffic to be distributed and collected between the multiple lower-speed channels. At the destination, buffers must be provided to collect data arriving from the different channels. Ideally, one would provide a shared buffer space for all the channels instead of a fixed buffer for each channel, so that buffer space not used by one channel can be used by the others.

There also is a question of whether to stripe large packets or small cells over the channels. Cell-striping allows use of the aggregate bandwidth for the transfer of even a single packet, but requires a relatively large buffer at the destination to absorb possible time skews between cells arriving from the different channels. Packet-striping does not require as large a buffer. In addition, it can support a scalable, modular design: a high-speed packet interface can be formed by simply combining a number of low-speed packet interfaces with a packet striper. For these reasons, packet striping is used in the HAS.

Conclusion

Gigabit LANs provide both high bandwidth (multi-Gb/s) and low latency (tens of μs or less) end-to-end communication. LANs with this performance will pave the way for many next-generation, high-performance computing and communications systems and new applications. Demands for gigabit LANs are widely recognized, and hardware component technology at gigabit speeds is basically available.

The performance trend of LANs, as shown in Fig. 1, clearly indicates that gigabit LANs represent the next performance milestone. The question is how soon the widespread availability of gigabit LANs can become a reality. It has taken FDDI approximately 10 years to mature. It is important that the gigabit LANs' development not experience a similar delay.

Gigabit LANs differ from other networks in many ways. Because of their high speeds, gigabit LANs face new challenges related to issues such as highly bursty traffic and large bandwidth mismatches in network links. To successfully meet these chal-

lenges, new LAN architectures, such as the virtual-connection architecture, are needed. In addition, because they are new networks with many new features, gigabit LANs need to resolve many architectural issues related to interfacing with hosts, special-purpose systems, and other networks. The solutions and ideas presented in this paper are mainly for the purpose of illustrating concepts and concerns. Much work is needed to verify the approaches proposed herein.

To ensure rapid progress, there should be extensive collaborative efforts between industry, university, and government. These efforts should complement other existing efforts such as gigabit WAN testbeds [12, 23] and research in network-based multicomputer systems. Cross-industry commitment in interoperability between networks should be fostered. Early standardization in areas such as physical media and framing, where agreements can be reached relatively easily, should be encouraged. Example research areas include switch and transport architectures; lightweight protocols; low-latency host interfaces; integrating architectural development of hosts, LANs, MANs, and WANs; and new applications and usage models. Hopefully, there will be a tight coupling with industry to achieve the goal of one to two years of experimentation immediately followed by the initial commercial products.

Acknowledgments

This work was supported in part by the Defense Advanced Research Projects Agency (DOD) monitored by DARPA/CMO under Contract MDA972-90-C-0035, and in part by the National Science Foundation and the Defense Advanced Research Projects Agency under Cooperative Agreement NCR-8919038 with the Corporation for National Research Initiatives. The author thanks Ira Richer (Mitre Corporation) and David Eckhardt (Carnegie Mellon) for their helpful comments on this manuscript.

References

[1] A. S. Tanenbaum, *Computer Networks*, 2nd Ed. (Prentice Hall, 1988).

[2] H. J. Fowler and W. E. Leland, "Local Area Network Traffic Characteristics with Implications for Broadband Network Congestion Management," *IEEE J. Sel. Areas in Commun.*, Vol. 9, No. 7, pp. 1139-49, Sept. 1991.

[3] H. T. Kung, "Network-Based Multicomputers: Redefining High Performance Computing in the 1990s," Proc. Decennial Caltech Conf. on VLSI, Pasadena, CA, March 1989.

[4] H. T. Kung, *et al.*, "Network-Based Multicomputers: An Emerging Parallel Architecture," Supercomputing Conf., Nov. 1991.

[5] H. T. Kung, *et al.*, "Parallelizing a New Class of Large Applications Over High-Speed Networks," Proc. Third ACM SIGPLAN Symp. on Principles and Practice of Parallel Programming (PPoPP), Williamsburgh, VA, April 1991.

[6] E. A. Arnould, *et al.*, "The Design of Nectar: A Network Backplane for Heterogeneous Multicomputers," Proc. Third Int'l. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS III), Boston, MA, April 1989.

[7] H. J. Chao, *et al.*, "Transport of Gigabit/sec Data Packets Over SONET/ATM Networks," Proc. IEEE Globecom '91, Phoenix, AZ, Dec. 1991.

[8] E. C. Cooper, *et al.*, "Protocol Implementation on the Nectar Communication Processor," Proc. SIGCOMM '90 Symp. on Communications Architectures and Protocols, Philadelphia, PA, Sept. 1990.

[9] P. A. Steenkiste, "Analysis of the Nectar Communication Processor," Proc. IEEE Workshop on the Architecture and Implementation of High Performance Communication Subsystems, Tucson, AZ, Feb. 1992.

[10] A. Birrell and B. Nelson, "Implementing Remote Procedure Calls," *ACM Trans. on Computer Systems*, Vol. 2, No. 1, Feb. 1984.

[11] NFS: Network File System Protocol Specification, RFC 1094, SRI Network Information Center, Menlo Park, CA, March 1989.

[12] "Gigabit Network Testbeds," *IEEE Computer*, Vol. 23, No. 9, pp. 77-80, Sept. 1990.

[13] B. S. Davie, "A Host-Network Interface Architecture for ATM," Proc. SIGCOMM '91 Symp. on Communications Architectures and Protocols, Zurich, Switzerland, Sept. 1991.

[14] C. Brendan, S. Traw, and J. M. Smith, "A High-Performance Host Interface for ATM Networks," Proc. SIGCOMM '91 Symp. on Communications Architectures and Protocols, Zurich, Switzerland, Sept. 1991.

[15] "Preliminary report on Broadband ISDN Transfer Protocols," *Bellcore Special Report*, SR-NWT-001763, Issue 1, Dec. 1990.

[16] D. D. Clark, *et al.*, "An Analysis of TCP Processing Overhead," *IEEE Commun. Mag.*, pp. 23-29, June 1989.

[17] M. Gerla and J. A. Bannister, "High-Speed Local-Area Networks," *IEEE Spectrum*, pp. 26-31, Aug. 1991.

[18] V. Jacobson, "Congestion Avoidance and Control," Proc. SIGCOMM '88 Symp. on Communications Architectures and Protocols, Aug. 1988.

[19] High-Performance Parallel Interface — Mechanical Electrical and Signalling Protocol Specification (HIPPI-PH), ANSI X3.183-1991.

[20] B. W. Abeysundara and A. E. Kamal, "High-Speed Local Area Networks and Their Performance: A Survey," *ACM Comput. Surveys*, Vol. 23, No. 2, pp. 221-64, June 1991.

[21] J. Lumley, "A High-Throughput Network Interface to a RISC Workstation," Proc. IEEE Workshop on the Architecture and Implementation of High Performance Communication Subsystems, Tucson, AZ, Feb. 1992.

[22] N. K. Cheung, "SONET/ATM — The Infrastructure for Gigabit Computer Networks," *IEEE Commun. Mag.*, April 1992.

[23] R. Binder, "Networking Testbeds at Gigabit/second Speeds," *Opt. Fiber Commun. Conf. Dig.*, Paper TuE1, p. 27, San Jose, CA, Feb. 1992.

[24] Digital Hierarchy — Optical Interface Rates and Formats Specifications, American Nat'l. Standard for Telecommunications ANSI T1.105-1988, American National Standards Institute, Inc., Sept. 1988.

[25] Special Issue on Optical Amplifiers, *J. of Lightwave Tech.*, Vol. 9, No. 2, Feb. 1991.

Biography

H. T. KUNG joined the faculty of Carnegie Mellon University in 1974 after receiving his Ph.D. degree there. Since January 1992 he has been Gordon McKay Professor of Electrical Engineering and Computer Science at Harvard University. During a transition period, he continues his involvements with projects underway at Carnegie Mellon. He was Guggenheim Fellow in 1983-84, and a full time Architecture Consultant to ESL, Inc., a subsidiary of TRW, in1981. In 1991, while on a sabbatical leave, he worked on high-speed network architectures at Bell-Northern Research Ltd. He has led a research team at Carnegie Mellon on the design and building of novel parallel computers and switch-based networks. Together with this students, he pioneered the concept of systolic array processing. This effort recently culminated in the commercial announcement by Intel of the iWarp product line of parallel computers. In the area of networks, his team has developed the Nectar system which uses fiber-optic links, large crossbar switches, and dedicated network coprocessors. A prototype system employing 100 megabits/second links and more than 20 hosts has been operational since early 1989. The team is currently working with industry on the next-generation Nectar, which will employ fibers operating at gigabits/second rates. The gigabit Nectar is one of the five testbeds in a national effort to develop gigabits/second wide-area networks. His current network research is directed towards gigabit cell-based LAN architectures capable of guaranteeing performance.

Because of their high speeds, gigabit LANs face new challenges related to issues such as highly bursty traffic and large bandwidth mismatches in network links.