Computer Communications: Ideas and Trends

Dan Winkelstein

MCNC

Abstract

Broadband Integrated Services Digital Network (BISDN) is an emerging high speed telecommunication standard for the next generation of high-speed wide- and local-area networks. This method of information transfer will not only change the way computers communicate but also will change the way we solve problems. Within this paper we detail our vision of the way high-speed communication networks will be organized and will operate. We will then outline some of the effects these high-speed networks will have on the communication protocol stack.

1: Introduction
1.1: Background

Computer communications has entered an era of unprecedented growth and development. We have reached the point where one can literally access knowledge around the globe. The bottleneck today in computational performance is the ability for computers to exchange information quickly and efficiently. Essentially, the communications backbone has failed to keep pace with the growth in processor speed and proliferation of computing platforms.

We see the gigabit-per-second communication rates as the essential technology milestone in order to keep pace with the deployment of ultra-powerful computing platforms. At these rates, the bottleneck in performance will be shifted from the communications infrastructure to the computing platform. This idea of providing a communications infrastructure capacity that exceeds the immediate computer demand will radically alter the way we approach computational problems.

Not since the days of 110-baud Teletypes has the capacity of the communication infrastructure exceeded the processing power of an individual computational platform. Since that time, our ability to generate traffic, either collectively or individually, has far exceeded our ability to transport that traffic to other computing platforms. The ramifications of this bottleneck is that applications tended to be processed entirely on the processor of the local machine. If an application was too large for a particular platform, the entire application would need to be ported to a larger platform with more processing capability. The network on the other hand, provided only a mechanism for moving the display to a convenient location or for moving the data for convenient non-volatile storage (e.g. disks). This approach has an inherent upper bound, at some point a problem may be too complex, too large, or too time consuming to be solved on any one machine.

If on the other hand, applications were designed such that there existed a coherent method for distributed computing in a heterogeneous computing environment, then applications could be processed both in parallel and on the platforms best suited for a particular task. For example, certain computational platforms are well suited for solving problems that involve a large number of sequential operations such as the CRAY Y-MP family of supercomputers. On the other hand, certain platforms are better suited for solving problems that involve a large number of parallel operations such as the CM2\(^1\) or the KSR1\(^2\). Other machines are very specialized for solving specific applications such as graphics (i.e. Pixel Planes 5 at UNC-CH) or for video applications (i.e. Princeton Engine at Sarnoff Research Labs). We see applications of the future being distributed among multiple heterogeneous computing platforms. The computational model will be designed such

---

The work reported in this paper is supported, in part, by BellSouth and GTE Corp. Support is also provided by the National Science Foundation and the Defense Advanced Research Projects Agency under cooperative agreement NCR 8919038 with the Corporation for National Research Initiatives.

1Thinking Machines Inc.

2Kendell Square Research Inc.
that various portions of an application will be executing in parallel on the platform most suited for a particular task and that data and control information will be exchanged over the communications network.

1.2: Motivation

Obviously, in order to achieve distributed applications over a heterogeneous computing platform in a wide area network, the network infrastructure must be able to support both extremely high bit-rates and efficient communications. The present infrastructure, particularly for national wide-area networking is grossly inadequate for this challenge. Much of the nationwide wide-area network referred to as the INTERNET is at T1 rates or lower. While T1 rates is more than adequate for e-mail applications, it is a severe limitation for distributed supercomputing. However, work is in place now to upgrade the network to T3 rates (45Mbps). Plans are taking shape to massively improve this network between the major supercomputing centers to near-gigabit and gigabit rates during this decade.

The local area networks on the other hand, are generally, much faster than the wide area networks. Ethernet and Token Ring at 10Mbps and 16Mbps respectively, provide a good starting ground for experimenting with distributed heterogeneous computing. At these rates, applications such as shared X-windows, low quality packetized video, network file systems, and certain specialized applications (such as applications running PVM) become attractive. At FDDI rates of (100Mbps) applications such as scientific real time visualization, medium quality packetized video, and distributed computing between a small number of heterogeneous platforms become attractive.

However, for grand challenge problems such as real-time radiation therapy planning, large-scale fluid dynamics problems, weather modeling, micro-anatomy, computational chemistry, or any problem involving real-time computational steering of complex physical processes, these data rates are woefully inadequate.

To illustrate this point, consider the grand challenge problem of micro-anatomy. The problem behind micro-anatomy is to evaluate the interaction at the cellular level of a tissue sample. In this problem, tissue samples are analyzed at a granularity of roughly a micron (a typical cell is a little larger than four microns). In order, to analyze a 1 x 1 x 1 cm tissue sample at this granularity would require at least a trillion ($10^{12}$) samples. If each sample was only 8-bits, this corresponds to a Terabyte of data. In order to do anything meaningful with this data, one would need to do massive quantities of image recognition and graphical rendering. In order to solve this type of problem in any reasonable time period would require Teraflops of computing power.

Clearly no single computing platform has Teraflops of computing power and Terabytes of disk space; however, if the problem could be partitioned and executed in parallel among platforms all over the country and information between peer or hierarchical processes could be exchanged, this type of problem could be solved using existing computing platforms. The difficulty of course in exchanging this information, is that gigabytes of data will need to be transferred between platforms in real-time. This is where the next generation of computer communication networks is required. For these reasons, we contend that the solution to some of these grand challenge problems rests squarely on the ability of the underlying computer communication network to achieve data rates of gigabits-per-second or higher.

1.3: Scope

Within this paper we will detail our vision of the direction computer communication networks will probably take in the next decade. We will not only look at the requirements for computer networks for grand challenge problems but also for applications that provide improved human-to-human interaction as well as typically mundane applications.

We begin by outlining the technology used for fast packet switching in the local-and wide-area networks. We then will explore the impact fast packet switching will have on the communication protocol stack.

Throughout this paper, it is important to remember that the communications network is not just the physical cable that interconnects multiple computing platforms. The network also refers to the host software and routers necessary for useful application-to-application communication. Thus when we speak of communications rates, we must remember that the raw bit rate on the cable is meaningless. It is the data rate delivered to the application that is important.

2: Computer Networks of the Future

The next generation of computer networks that will support the data rates and features necessary for wide area distributed heterogeneous supercomputing will be based on a fast packet technology referred to as Asynchronous Transfer Mode (ATM). ATM will be carried over a Synchronous Optical Network (SONET). ATM and SONET are part of the emerging telecommunications standards referred to as Broadband Integrated Services Digital Network.
SONET is a very high speed, optical based, communication standard designed to permit seamless interconnectivity throughout the world.

ATM is a technology for very fast packet switching over a virtual connection. The idea behind ATM is to provide a homogeneous communication media for voice, data, video, still images, or any other type of information. The ATM standard calls for very small packets called cells. These cells are then transported over a virtual connection across the ATM network. In essence ATM is a hybrid between packet switching and circuit switching.

We envision ATM networks evolving on a nationwide scale as illustrated in figure 1.

In the local environment, heterogeneous computing platforms will either hook directly to the local ATM network or via gateways from more traditional LANs. The local ATM network then will be interconnected to a high speed nationwide area network. The existing packet switching network (INTERNET) will then become a subset of the nationwide ATM network.

3: Network Protocols

3.1: Physical Layer Issues

The physical layer is the protocol layer used to transmit and receive the actual information over the physical medium. For the local area network this layer is probably the easiest layer to work with, the only major issue associated with this layer is money.

The issue of cost to implement the physical layer can
be subdivided into two major problems: the cost of installing and maintaining a fiber network and the cost of providing interfaces to host computing platforms.

The most expensive part of establishing a fiber optic network is putting the fiber in the ground. In fact this represents the single largest expense to a telephone operating company, long distance provider, or private network provider. This expense not only includes digging the hole for which to bury the fiber, but also establishing or purchasing right-of-way for new facilities. The second major cost of establishing ATM over SONET LANs is providing a mechanism to communicate between local hosts and the ATM network. At present, there are no ATM over SONET interfaces for any of the major computing platforms. One solution to this problem is either for the computer manufactures or some third party to provide a native mode ATM interface to their platform. A second approach is to provide external gateways that interface between one standard and ATM over SONET.

3.2: Link Layer Issues

For ATM over SONET networks, the ISO-OSI link layer is primarily characterized by the ATM layer. In the ISO-OSI model, the link layer is responsible for moving data between intermediate hops in the data network.

In an ATM network, a virtual connection is established at call setup, a user is guaranteed that the cells will arrive in the order that they are sent. However, the user is not guaranteed that the cells will not be delayed or that some cells will not be dropped due to congestion. Unlike a packet switching network; where there is no guarantee of delay or packet loss, the BISDN network providers envisioned that customers may negotiate (and pay for) various grades-of-service (GOS). These grades-of-service set upper bounds on delay and lost cells.

3.3: Network Layer Issues

For ATM over SONET networks, the network layer is partially mapped to what is referred to as the ATM Adaptation Layer (AAL). However, in order to communicate with the existing INTERNET, the IP network layer must be encapsulated into the AAL.

The ATM Adaptation Layer is subdivided into two sublayers, the Convergence Sublayer (CS) and the Segmentation and Reassembly Sublayer (SAR). The Convergence Sublayer is responsible for taking the user protocol data unit, which in our case is an IP packet and prepends and appends it such that it is in a format that is appropriate for the SAR sublayer. The SAR sublayer actually breaks the data into very small chunks for insertion into the payload of a cell. Depending on the AAL chosen, either the SAR sublayer or the CS sublayer may perform error detection.

One of the issues that is important for gigabit communications and supercomputing is the issue of seamless multicast at the network layer. By multicast, we mean that a single source can transmit to multiple destinations without explicitly knowing the number and particular addresses of these destinations. In essence, a source transmits data to a group address and assumes the network will deliver the message to all parties that are part of that group.

To implement multicast, a message must be duplicated as it transverses the network. This function is performed at both the CS level for an ATM network and at the IP network layer for the INTERNET.

Another area of interest for the network layer is the concept of bandwidth reservation. Before an ATM connection can be established, the network must know \textit{apriori} the traffic characteristics to expect for a particular connection. This is a new concept in the computer communications realm. In the telecommunications realm, this concept is well understood for voice traffic, however there is still a considerable amount of research necessary before we can reliably predict traffic patterns for most computer applications.

3.4: Transport Layer Issues

The transport layer is responsible for reliable (or unreliable) end-to-end data transfer between hosts and may represent a bottleneck in the communications network.

Typically the transport layer is responsible for flow control, buffer management, reliable data transfer, and recovery from lost, misdirected, or out of sequence packets.

The related issues of flow control and buffer management are handled today in TCP by use of a window based system. We believe that a rate based flow control mechanism may provide superior throughput and simpler implementation when compared to a window based system. An optimum approach to flow control and buffer management may be a layered approach consisting of both rate based and window based flow control.

3.5: Session Layer Issues

In the 4.3BSD unix environment, the session layer is entirely encompassed in the sockets abstraction. It is through this abstraction, that an application initiates the necessary handshaking required to establish a virtual connection for reliable packet delivery.
In high speed ATM networks, the session layer becomes more important because establishing a virtual connection across an ATM network is more complicated. Since ATM networks employ a preventive rate-based flow control mechanism, all virtual connections must be established through a process of bandwidth negotiations.

Therefore, it is the responsibility of the session layer to negotiate the bandwidth limitations on a virtual connection during the process of call setup. This issue is particularly difficult at the receiving end of a virtual connection, since the session layer software must ensure that sufficient processing capability exists to handle expected peak traffic conditions on all virtual connections bound for that host. This expected peak traffic conditions are based on the statistical probability of traffic arrivals and balanced against the required quality of service for that virtual connection. If the session layer is too conservative in its estimates of peak traffic, then available bandwidth to the host could be wasted. If the session layer is too liberal in its estimates of peak traffic then the quality of service for all connections is compromised.

3.6: Presentation Layer Issues

One of the main functions of the Presentation layer is format conversion. Different computers store datasets in different formats. A floating point number on one architecture may be stored very differently on another machine - the bytes may be swapped, the word sizes may be different, even the bits may be ordered completely backwards. The conversion of datasets into a form that another machine can understand is often called presentation-layer processing. Presentation processing can either convert the data into a common "network canonical" format (i.e. XDR format), requiring possible conversion on each end. Alternatively if the destination is known and static, it may simply do the conversion once, sending the data in a format that will be recognized immediately upon receipt at the destination.

The conversion processes of the presentation layer may introduce the greatest bottleneck for heterogeneous computation. For example, conversion to an External Data Representation (XDR) can slow the output of a 800-Mbps HIPPI link from a CRAY Y-MP8/432 to 20-Mbps. This problem, one of the most difficult issues to solve in gigabit networking, is exacerbated by the fact that supercomputers often achieve performance by highly optimizing the way data is stored internally in the machine.

4: Conclusion

In this paper we have presented a brief overview of some of the issues associated with high speed networks of the future. We envision the high speed local- and wide-area network will be based on ATM over SONET. We also envision that the future networks will continue to use a layered approach to data protocol; however, the lines and functionality for each of the layers will probably be redrawn. In closing, high speed communication networks will not only change the way we communicate, but the way we solve complex problems.

5: References


6) E. Minzer, Broadband ISDN and Asynchronous Transfer Mode (ATM)," IEEE Communications Magazine, Sept 1989 pp 17-24


8) Dan Stevenson, "VISTAnet Communications Research," Cray Users Group, October 1990, Austin, TX.