1988

A White Paper on High-Speed Network Architecture

Douglas E. Comer
Purdue University, comer@cs.purdue.edu

John M. Steele
Purdue University, jms@cs.purdue.edu

Raj Yavatkar

Report Number:
88-829

A WHITE PAPER ON HIGH-SPEED NETWORK ARCHITECTURE

Douglas Comer
John Steele
Raj Yavatkar

CSD-TR-829
November 1988

A White Paper on High-Speed Network Architecture

Douglas Comer
John Steele
Raj Yavatkar

Computer Science Department
Purdue University
West Lafayette, IN 47907

CSD-TR 829
November 1988

Summary

Research and development of high speed communication networks is a national priority. Such networks serve two purposes, providing backbone connections that are used simultaneously by many pairs of communicating machines and providing high-throughput connections for an individual pair of machines. It is the thesis of this paper that it is possible to construct a single high-speed network technology that accommodates both needs. We argue a communication and switching fabric for such a technology can be built from existing electronic parts that will operate two orders of magnitude faster than existing systems. Faster electronic parts available in the future will enable the same technology to be extended to even higher speeds.

1. Introduction

The National Research Council report on network communication support for scientific research community [15] eloquently argues the need for research and development of high speed national computer communication networks. In addition, networks like the NSFnet demonstrate the need for high-bandwidth systems. There are two primary motivations for high speed. First, high speed network hardware is necessary to accommodate the aggregate traffic generated when many pairs of data sources and sinks communicate simultaneously. Second, high speed hardware is needed to attain high throughput needed when a single pair of communicating machines conduct bulk data transfers.

The National Science Foundation's NSFnet [14] and the Defense Advanced Research Projects Agency's ARPANET [13] provide examples of the first case. Both networks form national backbones that interconnect many sites. Although the traffic generated by a pair of communicating machines is less than the total capacity of the communication technologies used, the aggregate traffic generated by hundreds of machines can easily exceed the network capacity.

Image transport illustrates the second motivation for high speed. A single image typically contains data for 10^6 pixels, where each pixel is represented by a gray scale or by three colors per pixel, where each color uses 8 bits (the printing industry uses 4 colors for many of its images). Thus, the file for single image typically contains between 5 and 32 megabytes of data. To transport such a file from one machine to another quickly (quick in terms of human response time is under 10 seconds) means throughput rates of between 4 and 25 Mbps, excluding the cost of protocols. When the data is a sequence of images to be displayed in real time, not as much data is kept per image, but the network rates needed are close to 100 Mbps.

2. Related Work

Currently, researchers are exploring several different approaches to high speed networking. Most of the work has concentrated in the areas of improved switching architectures and communication protocol design.

Recent advances in fiber optics and VLSI technology have led to improved packet switch architectures. The innovative packet switch designs include photonic switches [7,8,16], self-routing, nonblocking packet switches based on Batcher-Banyan sorting networks [9,18], and Knockout Switch [20].

The advent of high speed networking and associated high performance applications has indicated need for low latency internetwork transport services. Traditional implementations of network and transport protocols based on TCP/IP or OSI suite are not well-matched for high speed operations. Recently, researchers have focused attention on designing communication protocols that minimize protocol processing in switches and reduce protocol overhead in flow and error control. Greg Chesson's XTP [2,19] is an example of a lightweight transport protocol designed for an implementation in a VLSI architecture. XTP provides connection oriented real-time datagram service in an Internet environment. VMTP [1] provides a transaction oriented transport protocol with rate based flow control and also has a VLSI implementation [11]. Internet Engineering task force [10] is also considering a connection oriented Internet protocol.

3. Design Goals For A High-Speed Network

We assert that it is possible to build a high-speed network technology that accommodates both aggregate traffic as well as high throughput on a pair of connections. We have in mind a technology that will support a wide variety of interconnections, exhibit high throughput and low delay, and provide congestion avoidance as well as quick response to hardware failure (e.g., link failure). This section reviews our fundamental assumptions and outlines design goals for such a technology.

Small packets.

While circuit-style networks can handle bulk data transport from one point to another, a high speed network can only be used for multiplexing communication among vast numbers of slower machines if it supports packet switching. The key lies in finding ways to accommodate both styles in a single network without excluding small packets.

Protocol Independence.

Families of network and internet protocols continue to evolve [4,17]. Moreover, the availability of high-speed network technology will encourage new protocols and new applications. Thus, the goal is to build high speed technologies that are independent of underlying protocols. In practical terms, it means we view the network as a universal packet delivery system that can accept and deliver packets without understanding their contents (i.e. the network should not have a static notion of "protocol type" in its design).

Flexible Topology.

The key to building a versatile high-speed network technology is to allow end users to configure the network to meet their needs. Some sites will prefer ring structures, others will opt for trees, and still others will build many redundant links or use application specific topologies [12]. As a general rule, it should be possible to study network use and configure the speed of point-to-point connections between switches accordingly.

Accommodate Growth.

A high-speed network technology should be designed to accommodate growth. Growth is inevitable; any technology that cannot expand will have a short lifetime. Growth occurs in two ways: in size as the network expands to accommodate more sites/machines, and in capacity to accommodate more traffic. In the latter case, increases in traffic can come from existing machines moving more data or from the addition of pairs of machines that introduce high load. The key is making a single technology sufficient for all needs, so managers can choose to upgrade an existing network instead of adding new ones.

Ease of Management.

Existing networks are difficult to manage. New technologies must have network monitoring and control facilities that allow managers to detect and correct problems. Furthermore, high speed technologies need global policy-controlled resource allocation schemes that allow

managers to understand and control resource allocation.

Cost Effectiveness.

While it is easy to envision complex hardware and software for high-speed networking, the key to solving the problem lies in finding inexpensive solutions. We assert that it is possible to build networking technologies (using off-the-shelf electronic parts) that operate at least two orders of magnitude faster than current systems while keeping costs close to those for current local area networks.

4. Design Criteria

This section discusses specific design criteria that will meet the goals outlined above.

High speed electronic switching.

Switching small packets means making more decisions per unit of time. We assert that it is possible to create electronic switches that can switch multiple fiber lines at speeds of 100 Mbps per line (i.e., two orders of magnitude faster than existing switches) by using multiple processors; it will not be possible using uniprocessors.

Hybrid Circuit/Packet Switching.

To support guaranteed throughput, next generation network hardware will need a concept of reserved bandwidth paths. Borrowing terminology from Clark [3], we call them *flows* [6]. A flow is a communication channel that has specific performance characteristics associated with its traffic. Two endpoints establish a flow by prereserving resources along a path and then send data down the path at the agreed rate. Flows also allow datagram traffic (that has no resources explicitly allocated in advance) to coexist in the network. Although flows resemble virtual circuits, it will not be possible to achieve high speed with conventional virtual circuits because conventional virtual circuit protocols include extra overhead that guarantees reliable, flow-controlled delivery. Also, virtual circuits guarantee bandwidth, but make no provisions for specifying data rates or limits on traffic delays. By contrast, the concept of flow includes only resource reservation with no overhead of flow and error control, making it possible to switch at high speed. Prespecification and reservation of resources reduces *per-packet* protocol processing overhead. Judicious congestion avoidance along with pre-reservation of resources will allow us to maximize the network utilization.

Congestion Avoidance And Control.

Flows handle the case where two endpoints of a network need guaranteed bandwidth. To accommodate conventional packet switching, the network must also accept and switch individual datagrams. The chief liability of a datagram facility lies in congestion. Because each datagram travels independently, bursts of traffic over some part of the network can lead to congestion. It is impossible to handle congestion at the point of occurrence because that point is only switching traffic, not generating it. To avoid congestion, the network must have a control scheme that allows interior nodes to monitor traffic and report to exterior nodes, which then

reduce the rate at which new traffic enters the network. Traditional adaptive congestion control schemes react after increased traffic (congestion) is detected at an interior node by sending information back along the paths from which the traffic arises. However, datagram networks that use adaptive congestion control information are subject to potential stability problems [20]; we assert that such schemes cannot work in a high speed network. For example, backpressure flow control used in virtual-circuit networks is unacceptable for networks that carry real time traffic because such a flow control would introduce unacceptable delays [21]. Instead, a scheme is needed that allows exterior nodes to understand the network status and take action to reduce the rate of packet acceptance based on the current network state and packet destination [5]. Thus, interior modes must propagate changes in link status/utilization quickly to all nodes. Such control traffic must have highest priority.

Global Topology Store.

To make the propagation of link status efficient and reliable, each switching node must have knowledge of the network topology. It will be important that high-speed network technologies allow a variety of topologies (tree, ring, etc), and that the network system include an mechanism that provides for automated topology discovery and dissemination of topology information.

Precomputation Of Failure Recovery.

In a high-speed network, failure of an individual link or switch can result in dramatic data loss. Thus, high-speed networks must respond quickly to failures. Quick response is possible if the switches can precompute new routes for all possible failures, minimizing the latency between discovering a failure and responding by routing around it.

Intelligent High-Speed Interfaces.

To make high-speed networks accessible to various computers, they should support a variety of computer interfaces. For applications where the high-speed network is used to handle traffic from many small machines, the key is versatility -- having many possible interfaces means accommodating many machines. For example, the network should at least support an Ethernet interface and a direct bus interface (e.g., Multibus II or VME bus). To accommodate such interfaces, a general-purpose link-level protocol is needed that will allow any machine to communicate with the network. The protocol must allow the machine to send and receive packets, allocate and deallocate flows, ask for the network maximum packet size, and receive information on maximum packet rates (datagram flow control rate). In cases where the network supports high-speed connections between a pair of communicating machines, it is important that the machine interface be fast enough to transfer data between the computer and the network. For most systems, a high-speed interface implies special purpose hardware because conventional interface hardware will not support speeds above a few tens of megabits per second (unless the transfer stops the CPU).

Storage-to-Storage Transfer.

Bulk data transfer is of special interest in high speed networking because it provides strong motivation. Researchers at MIT have designed the NETBLT [22] protocol to facilitate bulk data transfer over a network with a large delay-bandwidth product. Although NETBLT improves throughput when conventional computers communicate, we need to look for new solutions for the specific problem of transferring huge volumes of data. In particular, when moving large images or other bulk data, the emphasis is often on quick storage-to-storage transfer rather than on processor-to-processor transfer. It will be possible to accommodate such applications with intelligent storage devices that attach directly to the network and use low-overhead bulk transfer protocols to move data quickly. If done well, a storage device attached to a network can handle both bulk transfer to remote storage devices as well as page-level transfer to local computing systems.

5. Conclusions

We have argued that it is possible to create a high speed network technology that supports both guaranteed throughput flows as well as independent datagram delivery. The network can be independent of high level protocols, support storage-to-storage transfers, and accommodate a wide variety of topologies.

6. References

[1] D.R. Cheriton, "VMTP: a transport protocol for the next generation of communication systems," *Proceedings of SIGCOMM 1986 Symposium*, pp. 406-415, August 1986.

[2] G. Chesson, "Protocol Engine Design," *USENIX Conference Proceedings*, pp. 209-215, June 1987.

[3] D. Clark, "Options for Research in Networking," Unpublished Note, M.I.T. Laboratory for Computer Science, January 1988.

[4] D.E. Comer, *Internetworking with TCP/IP: Principles, Protocols, and Architecture*, Prentice Hall, 1988.

[5] D.E. Comer and R.S. Yavatkar, "A Congestion Filtering Scheme for Packet Switched Networks," CSD-TR-758, Computer Science Department, Purdue University.

[6] D.E. Comer and R.S. Yavatkar, "FLOWS: Performance Guarantees in Best Effort Delivery Systems," *To appear in Proceedings of IEEE INFOCOM '89*, April 1989.

[7] K.Y. Eng, "A Photonic Knockout Switch for High-Speed Packet Networks," *IEEE Journal on Selected areas in Communications*, Vol. 6, No. 7, August 1988.

[8] Z. Haas, "Packet Switching in Future Fiber-Optic Wide Area Networks," Ph.D. Dissertation, Stanford University, May 1988.

[9] A. Huang and S. Knauer, "STARLITE: A Wideband Digital Switch," *Proceedings of IEEE GLOBECOM Telecommunications Conference*, Atlanta, November 1984.

[10] *Internet Enginnering Task Force Meeting*, Ann Arbor, MI, October 1988. Proceedings available from DDN Network Information Center, SRI International, Menlo Park, CA.

[11] H. Kanakiya and D.R. Cheriton, "The VMP Network Adaptor Board: High Performance Network Communication for Multiprocessors," *Proceedings of SIGCOMM 1988 Symposium*, pp. 175-187, August 1988.

[12] N.F. Maxemchuck, "The Manhattan Street Network," *Proceedings of GLOBECOM Conference*, New Orleans, December 1986.

[13] J. McQuillan and D. Walden, "The ARPA Network Design Decisions," *Computer Networks*, Vol. 1, No. 5, pp. 243-289, August 1977.

[14] D.L. Mills and H. Braun, "The NSFNET Backbone Network," *Proceedings of ACM SIGCOMM 87 Symposium*, pp. 191-196, August 1987.

[15] National Research Network Review Committee, "Toward a National Research Network," National Academy Press, Washington D.C., 1988.

[16] E. Nussbaum, "Communication Network Needs and Technologies - A place for Photonic Switching," *IEEE Journal on Selected Areas in Communications*, Vol. 6, No. 7, pp. 1036-1043, August 1988.

[17] Tannenbaum A., *Computer Networks*, Prentice Hall, Inc., 1981.

[18] J. Turner, "Design of an Integrated Service Packet Network," *IEEE Journal on Selected Areas in Communication* pp 1370-1380, November 1986.

[19] "XTP Definition, Revision 3.1," Protocol Engines Inc., March 1988.

[20] Y.S. Yeh, M.G. Hluchyj, and A.S. Acampora, "The Knockout Switch: A Simple, Modular Architecture for High-Performance Packet Switching," *Proceedings of International Switching Symposium*, pp. B10.2.1-B.10.2.8. IEEE, 1987.

[20] J.P. Fernow and M.L. El-Sayed. "Stability of Adaptive Congestion Controls in Packet Networks," *Proceedings of IEEE INFOCOM '83*, pp. 107-113

[21] M. Gerla and L. Kleinrock, "Congestion Control in Interconnected LANs," *IEEE Network*, Vol. 2, No.1, pp.72-76, Jan. 1988.

[22] D. Clark, M. Lambert, and L. Zhang, "NETBLT: A High Throughput Transport Protocol," *Proceedings of SIGCOMM '87 Workshop*, pp. 343-352, August 1987.