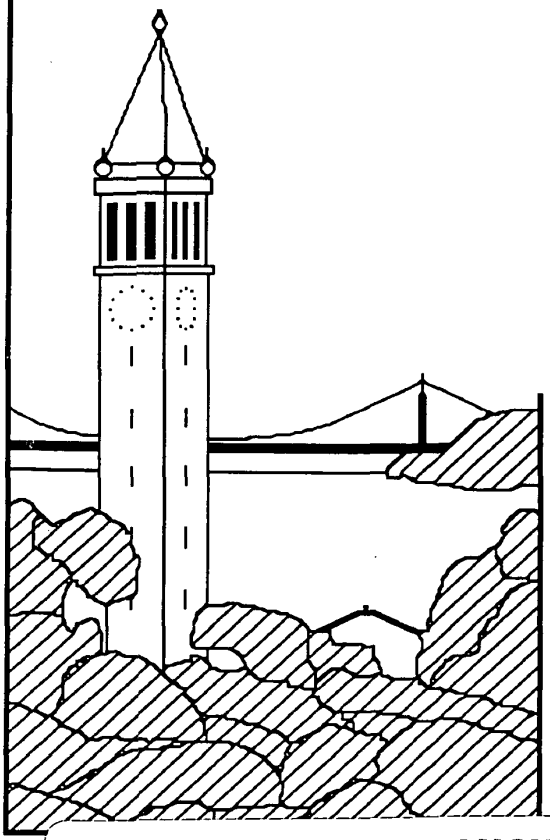
High Performance Network and Channel-Based Storage

Randy H. Katz



Report No. UCB/CSD 91/650

September 1991

Computer Science Division (EECS)
University of California, Berkeley
Berkeley, California 94720

NAS 2.59 1

High Performance Network and Channel-Based Storage

Randy H. Katz

Computer Science Division
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, California 94720

Abstract: In the traditional mainframe-centered view of a computer system, storage devices are coupled to the system through complex hardware subsystems called I/O channels. With the dramatic shift towards workstation-based computing, and its associated client/server model of computation, storage facilities are now found attached to file servers and distributed throughout the network. In this paper, we discuss the underlying technology trends that are leading to high performance network-based storage, namely advances in *networks*, *storage devices*, and *I/O controller* and *server architectures*. We review several commercial systems and research prototypes that are leading to a new approach to high performance computing based on network-attached storage.

Key Words and Phrases: High Performance Computing, Computer Networks, File and Storage Servers, Secondary and Tertiary Storage Device

1. Introduction

The traditional mainframe-centered model of computing can be characterized by small numbers of large-scale mainframe computers, with shared storage devices attached via I/O channel hardware. Today, we are experiencing a major paradigm shift away from centralized mainframes to a distributed model of computation based on workstations and file servers connected via high performance networks.

What makes this new paradigm possible is the rapid development and acceptance of the *client-server model* of computation. The client/server model is a message-based protocol in which *clients* make requests of service providers, which are called *servers*. Perhaps the most successful application of this concept is the widespread use of file servers in networks of computer workstations and personal computers. Even a high-end workstation has rather limited capabilities for data storage. A distinguished machine on the network, customized either by hardware, software, or both, provides a *file service*. It accepts network messages from client machines containing open/close/read/write file requests and processes these, transmitting the requested data back and forth across the network.

This is in contrast to the pure *distributed storage model*, in which the files are dispersed among the storage on workstations rather than centralized in a server. The advantages of a distributed organization are that resources are placed near where they are needed, leading to better performance, and that the environment can be more autonomous because individual machines continue to perform useful work even in the face of network failures. While this has been the more popular approach over the last few years, there has emerged a growing awareness of the advantages of the centralized view. That is, every user sees the same file system, independent of the machine they are currently using. The view of storage is pervasive and transparent. Further, it is much easier to administer a centralized system, to provide software updates and archival backups. The resulting organization combines distributed processing power with a centralized view of storage.

Admittedly, centralized storage also has its weaknesses. A server or network failure renders the client workstations unusable and the network represents the critical performance bottleneck. A highly tuned remote file system on a 10 megabit (Mbit) per second Ethernet can provide perhaps 500K bytes per second to remote client applications. Sixty 8K byte I/Os per second would fully utilize this bandwidth. Obtaining the right balance of workstations to servers depends on their relative processing power, the amount of memory dedicated to file caches on workstations and servers, the available network bandwidth, and the I/O bandwidth of the server. It is interesting to note that today's servers are not I/O limited: the Ethernet bandwidth can be fully utilized by the I/O bandwidth of only two magnetic disks!

Meanwhile, other technology developments in processors, networks, and storage systems are affecting the relationship between clients to servers. It is well known that processor performance, as measured in MIPS ratings, is increasing at an astonishing rate, doubling on the order of once every eighteen months to two years. The newest generation of RISC processors have performance in the 50 to 60 MIPS range. For example, a recent workstation announced by Hewlett-Packard Corporation, the HP 9000/730, has been rated at 72 SPECMarks (1 SPECMark is roughly the processing power of a single Digital Equipment Corporation VAX 11/780 on a particular benchmark set). Powerful shared memory multiprocessor systems, now available from companies such as Silicon Graphics and Solborne, provide well over 100 MIPS performance. One of Amdahl's famous laws equated one MIPS of processing power with one megabit of I/O per second. Obviously such processing rates far exceed anything that can be delivered by existing server, network, or storage architectures.

Unlike processor power, network technology evolves at a slower rate, but when it advances, it does so in order of magnitude steps. In the last decade we have advanced from 3 Mbit/second Ethernet to 10 Mbit/second Ethernet. We are now on the verge of a new generation of network technology, based on fiber optic interconnect, called FDDI. This technology promises 100 Mbits per second, and at least initially, it will move the server bottleneck from the network to the server CPU or its storage system. With more powerful processors available on the horizon, the performance challenge is very likely to be in the storage system, where a typical magnetic disk can service thirty 8K byte I/Os per second and can sustain a data rate in the range of 1 to 3 MBytes per second. And even faster networks and interconnects, in the gigabit range, are now commercially available and will become more widespread as their costs begin to drop [UltraNet 90].

To keep up with the advances in processors and networks, storage systems are also experiencing rapid improvements. Magnetic disks have been doubling in storage capacity once every three years. As disk form factors shrink from 14" to 3.5" and below, the disks can be made to spin faster, thus increasing the sequential transfer rate. Unfortunately, the random I/O rate is improving only very slowly, due to mechanically-limited positioning delays. Since I/O and data rates are primarily disk actuator limited, a new storage system approach called *disk arrays* addresses this problem by replacing a small number of large format disks by a very large number of small format disks. Disk arrays maintain the high capacity of the storage system, while enormously increasing the system's disk actuators and thus the aggregate I/O and data rate.

The confluence of developments in processors, networks, and storage offers the possibility of extending the client-server model so effectively used in workstation environments to higher performance environments, which integrate supercomputer, near supercomputers, workstations, and storage services on a very high performance network. The technology is rapidly reaching the point where it is possible to think in terms of *diskless supercomputers* in much the same way as we think about diskless workstations. Thus, the network is emerging as the future "backplane" of high performance systems. The challenge is to develop the new hardware and software architec-

tures that will be suitable for this world of network-based storage.

The emphasis of this paper is on the integration of storage and network services, and the challenges of managing the complex storage hierarchy of the future: file caches, on-line disk storage, near-line data libraries, and off-line archives. We specifically ignore existing mainframe I/O architectures, as these are well described elsewhere (for example, in [Hennessy 90]. The rest of this paper is organized as follows. In the next three sections, we will review the recent advances in interconnect, storage devices, and distributed software, to better understand the underlying changes in network, storage, and software technologies. Section 5 contains detailed case studies of commercially available high performance networks, storage servers, and file servers, as well as a prototype high performance network-attached I/O controller being developed at the University of California, Berkeley. Our summary, conclusions, and suggestions for future research are found in Section 6.

2. Interconnect Trends

2.1. Networks, Channels, and Backplanes

Interconnect is a generic term for the "glue" that interfaces the components of a computer system. Interconnect consist of high speed hardware interfaces and the associated logical protocols. The former consists of physical wires or control registers. The latter may be interpreted by either hardware or software. From the viewpoint of the storage system, interconnect can be classified as high speed networks, processor-to-storage channels, or system backplanes that provide ports to a memory system through direct memory access techniques.

Networks, channels, and backplanes differ in terms of the interconnection distances they can support, the bandwidth and latencies they can achieve, and the fundamental assumptions about the inherent unreliability of data transmission. While no statement we can make is universally true, in general, backplanes can be characterized by parallel wide data paths, centralized arbitration, and are oriented towards read/write "memory mapped" operations. That is, access to control registers is treated identically to memory word access. Networks, on the other hand, provide serial data, distributed arbitration, and support more message-oriented protocols. The latter require a more complex handshake, usually involving the exchange of high-level request and acknowledgment messages. Channels fall between the two extremes, consisting of wide datapaths of medium distance and often incorporating simplified versions of network-like protocols.

These considerations are summarized in Table 2.1. Networks typically span more than 1 km, sustain 10 Mbit/second (Ethernet) to 100 Mbit/second (FDDI) and beyond, experience latencies measured in several ms, and the network medium itself is considered to be inherently unreliable. Networks include extensive data integrity features within their protocols, including CRC checksums at the packet and message levels, and the explicit acknowledgment of received packets.

Channels span small 10's of meters, transmit at anywhere from 4.5 MBytes/s (IBM channel interfaces) to 100 MBytes/second (HiPPI channels), incur latencies of under 100 µs per transfer, and have medium reliability. Byte parity at the individual transfer word is usually supported, although packet-level checksumming might also be supported.

Backplanes are about 1 m in length, transfer from 40 (VME) to over 100 (FutureBus) MBytes/second, incur sub µs latencies, and the interconnect is considered to be highly reliable. Backplanes typically support byte parity, although some backplanes (unfortunately) dispense with parity altogether.

	Network	Channel	Backplane
Distance	>1000 m	10 - 100 m	1 m
Bandwidth	10 - 100 Mb/s	40 - 1000 Mb/s	320 - 1000+ Mb/s
Latency	high (>ms)	medium	low (<μs)
Reliability	low Extensive CRC	medium Byte Parity	high Byte Parity

Table 2.1: *Comparison of Network, Channel, and Backplane Attributes*
The comparison is based upon the interconnection distance, transmission bandwidth, transmission latency, inherent reliability, and typical techniques for improving data integrity.

In the remainder of this section, we will look at each of the three kinds of interconnect, network, channel, and backplane, in more detail.

2.2. Communications Networks and Network Controllers

An excellent overview of networking technology can be found in [Cerf 91]. For a futuristic view, see [Tesla 91] and [Negraponte 91]. The decade of the 1980's has seen a slow maturation of network technology, but the 1990's promise much more rapid developments. 10 Mbit/second Ethernets are pervasive today, with many environments advancing to the next generation of 100 Mbit/second networks based on the FDDI (Fiber Distributed Data Interface) standard [Joshi 86]. FDDI provides higher bandwidth, longer distances, and reduced error rates, due largely to the introduction of fiber optics for data transmission. Unfortunately cost, especially for replacing the existing copper wire network with fiber, coupled with disappointing transmission latencies, have slowed the acceptance of these higher speed networks. The latency problems have more to do with FDDI's protocols, which are based on a token passing arbitration scheme, than anything intrinsic in fiber optic technology.

A network system is decomposed into multiple protocol layers, from the application interface down to the method of physical communication of bits on the network. Figure 2.1 summarizes the popular seven layer ISO protocol model. The *physical* and *link* levels are closely tied to the underlying transport medium, and deal with the physical attachment to the network and the method of acquiring access to it. The *network*, *transport*, and *session* levels focus on the detailed formats of communications packets and the methods for transmitting them from one program to another. The *presentation* and *applications* layers define the formats of the data embedded within the packets and the application-specific semantics of that data.

A number of performance measurements of network transmission services all point out that the significant overhead is not protocol interpretation (approximately 10% of instructions are spent in interpreting the network headers). The culprits are memory system overheads due to data movement and operating system overheads related to context switches and data copying [Clark 89, Heatly 89, Kanakia 90, Watson 87]. We will see this again and again in the sections to follow.

The network controller is the collection of hardware and firmware that implements the interface between the network and the host processor. It is typically implemented on a small printed circuit board, and contains its own processor, memory mapped control registers, interface to the network, and small memory to hold messages being transmitted and received. The on-board processor, usually in conjunction with VLSI components within the network interface, implements

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.