Appeared in *Proceedings of the 18th Conference on Local Computer Networks*, Minneapolis, Minnesota, September 1993.

Interfacing a High Performance Disk Array File Server to a Gigabit LAN[1]

Srinivasan Seshan, Randy H. Katz

Computer Science Division
University of California, Berkeley
Berkeley, CA 94720

Abstract: **Distributed systems in use today depend heavily on network communications between clients and servers. In this report, we describe the design and implementation of the network architecture (hardware, software and protocols) of the RAID-II system. RAID-II is a high speed file server connected to an UltraNetwork. To support high bandwidth network transfers with the RAID-II server, we divided the networking software among the various processors in the system. With the distributed software, the CPU still limits the RAID-II file server to 21Mbytes/second of data bandwidth on our network.**

I. Introduction

Over the past decade, we have experienced a major shift from centralized computing using mainframes to a distributed model of computing using workstations connected via high-performance networks. In the traditional mainframe-centered view of computer systems, storage devices are tightly coupled to the computation system. In the newer workstation model of computing, storage is now attached to file servers distributed throughout a network. The workstation clients make file requests to a server through a message based protocol over a high speed network.

Centralized file storage has several advantages over tightly coupled disk storage. Users can have access to the file system on different client machines. Also, a centralized file system simplifies administration. However, typical client/server environments have many workstations per file server. Therefore, the speed of the entire system is highly dependent upon the efficiency of the communication between the server and its clients. In this paper we explore the design of a network architecture (hardware, software, and protocols) for RAID-II, a high-speed network file server.

The remainder of this paper is organized as follows. Section II describes the RAID-II hardware and the network interfaces of the system. Section III explains the network software architecture that we chose. Section IV presents some network performance measurements taken of the RAID-II system. We present our summary and conclusions in Section V.

1. This project was supported in part by NASA/ARPA under grant #NAG2-591 and the California State MICRO Program.

II. RAID-II

The RAID-II file server was designed to support applications typical to high-speed workstations of the future. This application workload is composed of a mixture of high bandwidth scientific, engineering and multi-media data, and low latency, high transaction rate UNIX-like I/O patterns.

Our previous prototype, RAID-I, identified several bottlenecks in typical file server architectures [Chervenak91]. The most important bottleneck was the lack of a high-bandwidth path between disk, memory and the network. Workstation servers, such as the Sun-4/280, have very slow access to peripherals on busses far from the CPU. For the RAID-II system, we addressed this problem by designing a crossbar interconnect, XBUS board, that provides a 40MB/s path between disk, memory and the network interfaces. However, this interconnect does not provide the system CPU with low latency access to control the various interfaces. To provide a high data rate to clients on the network, we needed to design the network software carefully and efficiently. A block diagram of the system hardware architecture is shown in Figure 1. In the following subsections, we describe pieces of the RAID-II file server hardware that had a significant impact on the design of the network interface. Other papers, [Lee92, Katz93], describe the architecture and implementation of the RAID-II server in greater detail.

A. VME Link Boards

The remote VME links that connect the host CPU to the other sections of the system are extremely slow: about 2Mbytes/second for most applications. Single word transfers across the link take 2us each. In a few select applications, the link board can DMA data at up to 20Mbytes/second. To meet our performance goal for data transfer to the network, very little data can be transferred between the host CPU and the other parts of the system.

B. TMC I/O Backplane

The TMC I/O backplane consists of two unidirectional busses, HIPPIS bus and HIPPID bus, that move data between the TMC HIPPI boards and the XBUS board. Both busses are addressless and only the TMC boards may be the bus master. The bus uses a simple protocol that allows the TMC board to select a target or source board for the transfer and to do flow control. Since the bus is addressless, any source or target

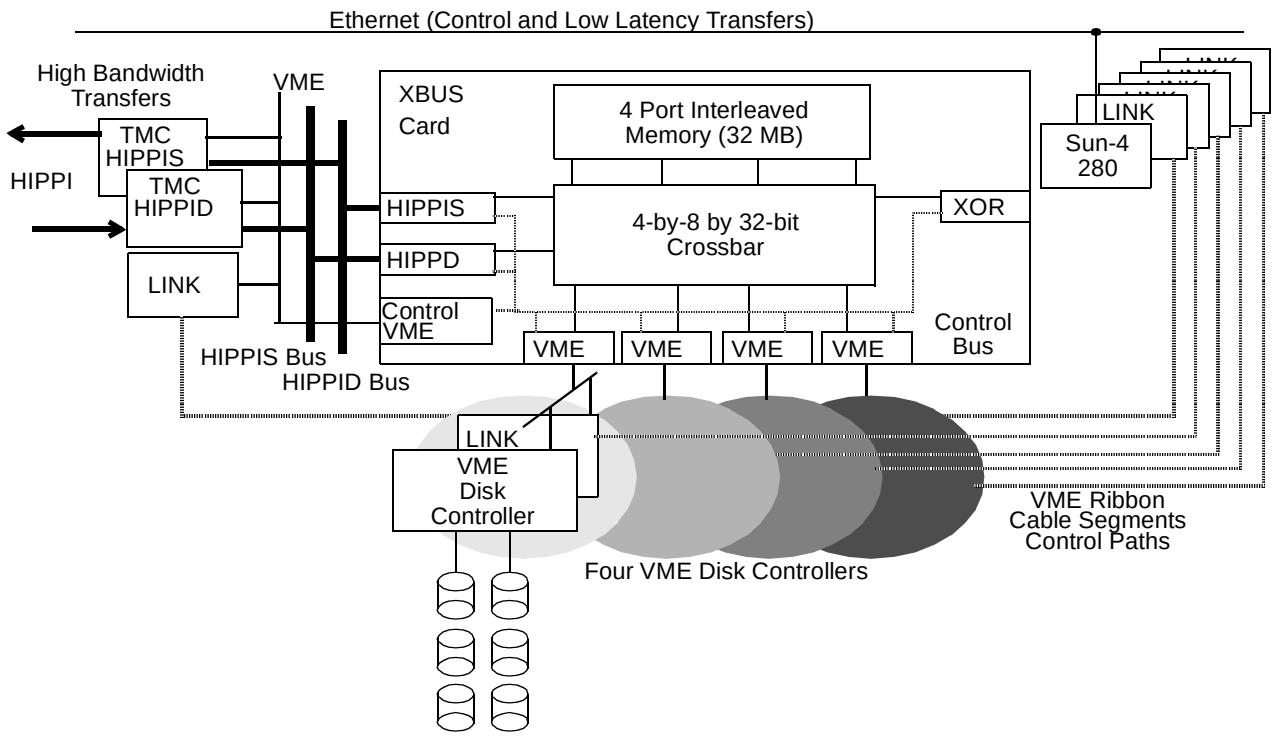

Fig. 1. RAID-II Organization. A high-bandwidth crossbar interconnects the network interface (HIPPI), the disk controllers, a multiported memory system, and a parity computation engine. An internal control bus provides access to the crossbar ports, while external point-to-point VME links provide control paths to the surrounding SCSI and HIPPI interface boards. Up to two VME disk controllers can be attached to each of the four VME interfaces. The design originally had eight memory ports and 128 MB of memory; however, we built a four memory port version to reduce manufacturing time.

device (for example the XBUS board) must be setup before a transfer is initiated.

C. Host

The host CPU in the RAID-II server is a Sun-4/280 single board computer. It has 32MB of VME memory and runs the Sprite operating system. The CPU performance of this machine is approximately 8 SPECMarks. The host is responsible for running most of the code that controls the RAID-II server. The host runs the file system code and controls the drive interfaces, XBUS board and HIPPI boards. The host is slow by today's standards and is likely to be heavily loaded by file system and control tasks. It is important that the network interface not place a significant additional load on the CPU.

D. XBUS Board

The XBUS card implements a 4-by-8 32-bit wide crossbar bus. This board provides a high bandwidth path between the disk controllers, memory and the network interface. Two of the crossbar ports provide connections to the TMC I/O backplane. Since the TMC I/O backplane busses are addressless, the XBUS board must be setup for any transfers across the backplane in advance. These ports can sustain 40Mbytes/second of transfer to and from the TMC HIPPI boards. A control VME connection uses another single port. A set of registers present on this VME interface controls the XBUS board. TMC HIPPI boards or the Sun-4 CPU may write to these control registers. Other ports provide connections to a 32MB memory, a hardware XOR compute engine and four disk controller boards.

E. SCSI controllers

The XBUS board is connected to a set of four VME busses. Each of these VME busses currently contains an Interphase Cougar SCSI controller. Each board is capable of handling approximately 7 Mbytes/second of data traffic from two independent SCSI strings. Physical packaging limits each string to 3 disks. These boards limit the RAID-II system to a maximum of 28Mbytes/second of disk bandwidth and 24 disks. Software and other bottlenecks may limit the performance further. Future SCSI boards will allow the system to use 72 disk drives and to provide up to 32Mbytes/second per XBUS board.

F. TMC HIPPI Boards

The HIPPI interface for RAID is implemented using a two board set built by Thinking Machines Corporation (TMC). The architecture of the boards is shown in Figure 2. Each board contains an interface to a single direction of the HIPPI channel, a unidirectional backplane bus and a control VME

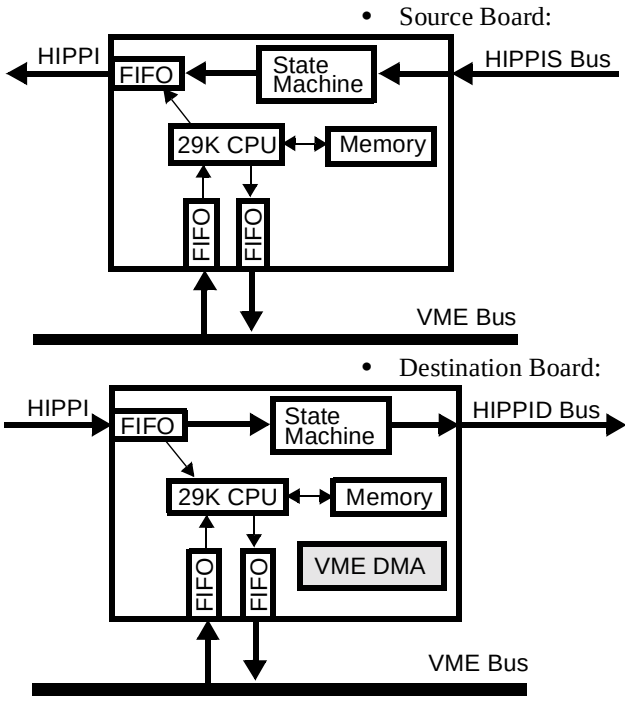

Fig. 2. TMC HIPPI Board Block Diagram

bus. Each board also contains a AMD 29000 (29K) processor and some local memory. Programs and data for the 29K processor can be downloaded from the VME control bus. The 29K processor can setup transfers and run any general purpose code (protocol code). Some significant differences exist between the two boards. The individual boards are described in more detail in the next few subsections.

HIPPI Source Board

The HIPPI source board interfaces to the VME bus through a set of five registers. The functions of these registers are summarized in Table I. The input and output FIFO are the most important of these registers since they provide the only general purpose communication interface between code running on the 29K processor and the host CPU. Since the source board has no VME bus mastering capability, data must be copied into the input FIFO and from the output FIFO. This VME interface has two important consequences. First, the XBUS board must be setup for transfers to the source board by some other part of the system (e.g., host CPU). Second, there is no mechanism to lock access to the FIFOs. This prevents both the Sun-4 CPU and the destination board from communicating with the source board. Therefore, to prevent mixing of data from two sources, we limit access to only the Sun-4 CPU.

TABLE I. HIPPI Source Board VME registers

VME Register	Description
Configuration	Sets up VME functionality of board - enable interrupts, set VME address modifier, etc.
Input FIFO	Receives data from VME into FIFO read by 29K.
Output FIFO	Stores data written by 29K into a FIFO that can be read from the VME.
Status	Stores current status of board
Reset	Controls reset of various sections of board

The source board's interface to the HIPPI output channel and the input backplane bus is controlled by a set of on-board registers. These registers are only accessible by the 29K processor. Data to be sent out on the HIPPI channel is stored in a single FIFO. A simple state machine fills this FIFO with data from the TMC I/O backplane. The 29K initiates this transfer by writing various registers. It must know the total length of the transfer in advance. By not involving the 29K in copying data from the backplane, the system can achieve the full HIPPI bandwidth (100Mbytes/second).

HIPPI Destination Board

The destination board includes several more features than the source board. The VME interface has several new registers that provide general purpose communication with the host. The destination board VME registers are summarized in Table II. In addition to the new registers, the destination board supports VME bus mastering. The board can write data from the output FIFO to any VME location. Similarly, it can read VME locations to the input FIFO. As a result, the HIPPI destination board can setup the XBUS board for transfers.

TABLE II. HIPPI Destination Board VME Registers

Register	Description
Configuration	Sets up VME functionality of board - enable interrupts, set VME address modifier, etc.
Input FIFO	Receives data from VME into FIFO, can be read by 29K.
Output FIFO	Stores data written by 29K into a FIFO that can be read from the VME.
Status	Stores current status of board
Command	Stores data written from VME, can be read by 29K.
Response	Stores data written by 29K, can be read from VME.
Reset	Controls reset of various sections of board

The destination board's interface to the HIPPI input channel is composed of a set of registers accessible by the 29K. Data from the HIPPI channel is automatically placed in a FIFO. Data in this FIFO can be copied by a state machine to the backplane. The 29K must setup the transfer to the backplane by writing the length of the transfer to an on-board register. The 29K must then poll the status of the state machine to identify the end of the transfer.

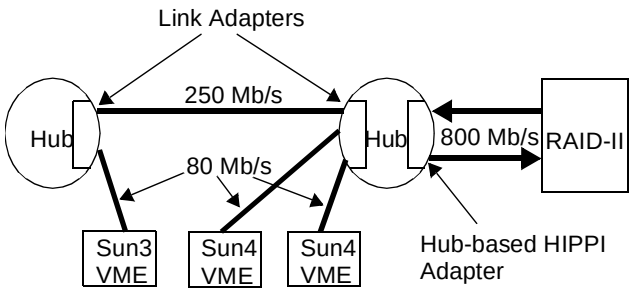

Fig. 3. UC Berkeley UltraNetwork Topology

G. Ultranet

The UltraNetwork is a hub-based store and forward network capable of transmission rates up to 1Gbit/second. Figure 3 shows our Ultranet topology. The hubs create a high speed switching interconnection by routing incoming packets to the proper destination.

Hubs are physically connected by serial links capable of transmission rates of 250Mbits/second. Up to 4 links can be used between a pair of hubs. Data is striped across these links to achieve Gbit/second speed. These links terminate in link adapters in the hubs. Link adapters are also used to connect to machines with Ultranet host adapters. Host adapters are available for machines with industry standard backplanes (e.g., VME). Each host adapter contains an on-board microprocessor and can perform DMA to the host's memory. The on-board microprocessor does all the protocol processing necessary to communicate across the UltraNetwork to remote clients. Computers without standard backplanes, typically mainframes and supercomputers, can connect to the UltraNetwork using standard channel interfaces (e.g., HIPPI, HSX) to a hub-based adapter. This essentially moves the network interface into the hub itself. The processor on the hub-based adapter handles much of the UltraNetwork protocol. However, software must run on channel connected hosts to handle communication to the hub-based adapter. This software is described in more detail below.

VME Ultranet host adapters in a Sun system provide a maximum of about 4Mbytes/second to the network. On the basis of the RAID-II performance goal of 40Mbytes/second, we decided that a HIPPI attachment to the drive array was necessary.

Each transfer between the UltraNetwork hub and the hub-adapter attached host is composed of a DMA word followed by either a request block or data. The Ultranet adapter limits the maximum size of the data segment of each transfer to 32KB. The DMA word accompanying each transfer describes the contents of the transfers. Analyzing the DMA word provides sufficient information to identify the correct memory destination for the transfer. Request blocks are commands that pass between the hub-based adapter and the host. Each request block roughly has an analogue in BSD 4.2 network socket calls. This made it easy to provide the file system with a socket interface to the network. Several of the most important request blocks are summarized in Table III. Only a few standard data formats are used to transmit the various request blocks. As a result each request block requires sending significantly more data than is necessary.

TABLE III. UltraNetwork Request Blocks

Request Block	BSD Equivalent
OPEN	socket()
ADAPTER LISTEN	combination of bind(), listen() and accept()
CONNECT	connect()
CLOSE	close()
SEND	send()
RECEIVE	recv()

III. Software Architecture/ Implementation

Both TMC and Ultranet provided software to support the original uses of their systems. After examining the provided code, we decided that completely new software was needed for several reasons. First, the RAID-II file server runs the Sprite Operating System. Both the TMC and Ultranet software were developed for Sun-OS and needed a significant amount of work to port to Sprite. Second, the software was developed to support the more standard machine interconnection. As a result, it could not provide the high performance we needed on the RAID system. In this section, we describe the organization of the networking software we developed for the RAID-II file server. We examine the decisions made during the software implementation and the reasoning behind these decisions.

A. Architecture

The interface provided to the file system code and the division of code between the 29K and Sun-4 CPUs were two basic issues of the software architecture. On the basis of the Ultranet request block format, we decided to provide the file system code with a socket interface to the network, making both the networking and the file system code easier to implement. Also, we decided to implement most of the software in the 29K for a variety of reasons. First, we estimated that the Sun-4 CPU would be heavily loaded by running the file system software and controlling the hardware of the RAID-II system. Second, the connection of the Sun-4 CPU to the rest of the system is through slow VME link boards. The involvement of the Sun-4 CPU in data transfers would reduce the bandwidth of the RAID-II server significantly. To support a high bandwidth between the network and memory, the 29K CPUs must control as much of the data transfer as possible. The 29K CPUs were programmed to understand the Ultranet request block interface and handle incoming data transfers. However, since access to the source board cannot be shared, the Sun-4 CPU must setup the outgoing data transfers. The software architecture is shown in Figure 4. An example transfer is described in the next section to clarify the software architecture.

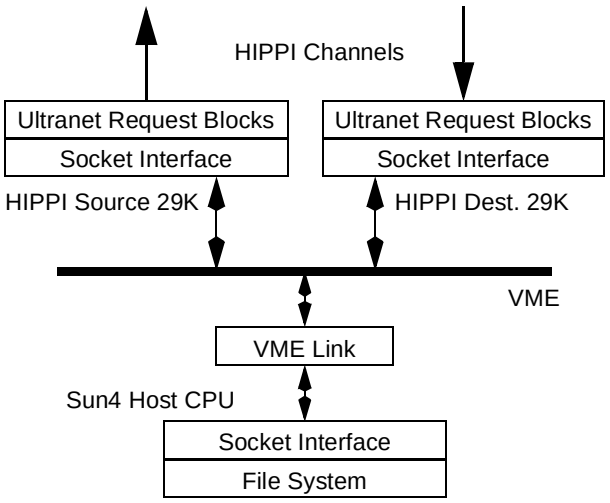

Fig. 4. Software Architecture Division

Sample Transaction

This section describes a sample network transaction that may occur between the RAID server and a client on the Ultranet. In this example, the client creates a connection to the server, sends some data and receives a reply. This communication is graphically shown in Figure 5.

1. The file server will start by issuing an `open()` of a socket. This will result in the HIPPI source board sending out an OPEN request block. The HIPPI destination board will receive the completed OPEN request block from the Ultranet hub. The destination 29K interprets the request block and returns a new socket id to the file server and the `open()` call completes.
2. The file server issues a `listen()` on the socket id. This is accomplished by the source board sending an ADAPTER LISTEN request block to the Ultranet hub. `listen()` is a combination of BSD bind(), listen() and accept(). The completed ADAPTER LISTEN request block returns to the destination board when a client creates a connection to the file server. The destination board sends information about the established connection to the file server and the `listen()` call completes.
3. The file server does a `recv()` on the connected socket id. A pointer to an empty host buffer and a tag that uniquely identifies the transfer are passed to the destination board. The source board sends a RECEIVE request block to the Ultranet hub. The Ultranet matches up the request block with a client's `send()` of data and transfers the data to the HIPPI destination board. The destination board uses the unique tag to identify the transfer. The destination board then sets the XBUS board up for the transfer and begins the transfer to the backplane. The Ultranet sends a completed RECEIVE request block after the transfer completes. The destination board sends the request block status to the file server and the `recv()` completes.

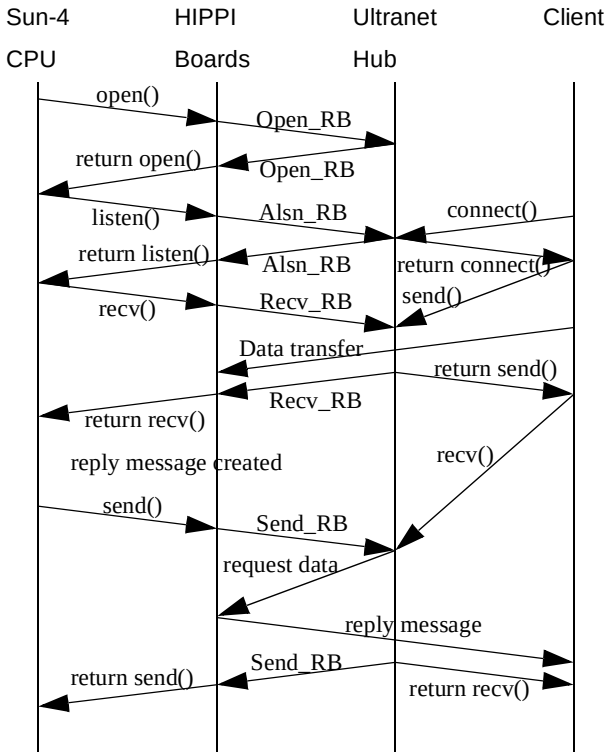

Fig. 5. Communication between Sun-4 CPU, TMC HIPPI Boards, Ultranet Hub

4. The host will issue a `send()` of the reply data on the socket id. A pointer to the host data buffer and a tag that uniquely identifies the transfer are passed to the destination board. The source board sends a SEND request block to the Ultranet hub. The Ultranet matches up the request block with a client's receive of data. The Ultranet hub sends a request to the HIPPI destination board to begin transfer of the data. The destination board uses the unique tag to identify the transfer request and determine the data to be sent. The destination board requests the Sun-4 to setup the XBUS board and source board to transfer the desired data to the Ultranet hub. The Ultranet sends a completed SEND request block to the destination board after the transfer completes. The destination board sends the request block status to the file server and the `send()` completes.

B. Source Board Code

From the example transfer, it should be evident that the HIPPI source board must send both data and request blocks to the Ultranet hub. The commands to perform these actions are summarized in Table IV. These commands are performed by the Sun-4 CPU writing to the source boards VME FIFO. In general, the Ultranet request block formats contain many data fields that can be eliminated. Commands between the 29K and Sun-4 CPU contain only the essential fields of the associated request blocks. The effectiveness of this "compression" is discussed in Section B.

TABLE IV. Commands between Source 29K and host CPU

Command	Description
UltraOpen()	Sends request to UltraNetwork to open a socket.
UltraListen()	Sends request to bind Socket ID to a port. Then listens for a connection and accepts it.
UltraClose()	Sends request to close connection active on a socket ID.
UltraSend()	Sends request to send data on the connection associated with a socket ID. Each send request block is given a unique 8 bit tag that identifies it.
UltraRecv()	Sends request to receive data on the connection associated with a Socket ID. Each receive request block is given a unique 8 bit tag that identifies it.
SendData()	Sends a requested number of bytes from the Sun-4 and the XBUS on the HIPPI channel.

C. Destination Board Code

To support the example transfer, the destination board needs to interpret the incoming Ultranet request blocks and scatter-gather Ultranet data requests. The Sun-4 uses the command described in Table IV to notify the destination board of buffers allocated for both incoming and outgoing transfers.

TABLE V. Commands Between Host CPU and Destination 29K

Command	Description
ScatterGather()	Allocates buffers in both Sun-4 and XBUS Memory for a transfer associated with a specific `tag`

The destination board must also read and interpret all transfers from the Ultranet. Every HIPPI transfer sent from the Ultranet hub to the destination board starts with the following DMA word structure.

31....24	23....16	15....8	7....0
Content Description			
Transfer Offset			
Tag			
Transfer Length			

`Content Description` identifies if the transfer contains a request block, a data request or data. If the current transfer is part of a larger multipart data transfer (larger than 32KByte transfer), `Transfer Offset` provides the byte offset of the data being sent into the entire transfer. `Tag` is the unique identifier for every send or receive of data. `Transfer Length` is the byte length of the current transfer. **Due to buffering limitations in the Ultranet hub,** `Transfer Length` **is never more than 32KBytes**.

Completed Request Blocks

The destination board 29K must notify the Sun-4 of any completed request blocks it receives. When a request block arrives at the destination board, the VME DMA engine is used to copy the essential fields (same fields that are used by the source board to send a request block) of the request block into the Sun-4 CPU's main memory. Next, the destination board interrupts the Sun-4 CPU to notify it of the completion of a Ultranet request. The host CPU may then examine the completed request block for either status or returned values

Incoming Data

When incoming data arrives at the destination board, the 29K processor uses the `tag, transfer offset` and `transfer length` fields of the DMA word and previously processed ScatterGather() commands to determine the destination of the data. If the destination address of the data is in host memory, the 29K removes the data from the HIPPI channel and DMA copies the data to the proper VME location. However, if the data should be placed in XBUS board memory, the destination board sets the XBUS board up for the transfer by writing to the XBUS VME registers. Next, the 29K enables the state machine to copy data from the HIPPI channel to the XBUS board. The data transfer is complete when the state machine finishes.

Outgoing Data

When a Ultranet request for data arrives at the destination board, the 29K processor uses the `tag, transfer offset` and `transfer length` fields of the DMA word and previously received ScatterGather() commands to determine the source of the data. The destination board cannot use the VME DMA engine to setup the transfer for several reasons. First, the host and the destination board cannot share access to the source board VME FIFO. Second, the destination board's VME DMA engine reads data into the destination board's VME input FIFO. However, the host CPU must also access this input VME FIFO. Access to this FIFO cannot be shared. As a result the host CPU must setup the transfer of data. The destination board copies the length and source address of the transfer to its VME output FIFO and interrupts the host CPU. The host CPU uses the length and source address to setup the XBUS and HIPPI source board for the transfer. This is done by writing to the XBUS control registers and issuing the SendData() command to the HIPPI source board.

D. Implementation

The approximate size of code running on the TMC HIPPI boards is summarized in Table VI.

Almost 7000 additional lines of code were written for the Sun-4 host CPU to support the UltraNetwork and HIPPI boards. Much of the code and time can be attributed to the lack of documentation for the Ultranet and the poor match between the architecture and the Ultranet protocol.

TABLE VI. HIPPI Board Code Statistics

Section	Lines of Code	Estimated Man Hours
Destination Board C Code	3500	900
Source Board C Code	3500	
Shared TMC Boards C Code	1500	
Shared TMC Boards Assembly	700	

IV. Performance Measurements

In this section, we examine the end-to-end network performance of the RAID-II file server. We analyze measurements of network bandwidth, CPU load of the RAID-II system and system hardware bandwidths to identify the bottlenecks that limit the network performance of the system.

A. RAID-II Hardware Performance

The RAID-II system was designed to support a 40MBytes/second data path between disk, memory and network. The performance of the system is carefully analyzed in [Chen93]. Measurements show that transfers between the XBUS board memory and the TMC HIPPI boards have a latency of 1.1ms and a maximum throughput of 38.5MBytes/second. The majority of this latency is attributed to the configuring of the XBUS board and the handling of the HIPPI channel by software on the TMC board. These measurements were taken on a system with minimal software on the host CPU and on the 29K CPUs. They indicate the maximum achievable performance from the RAID-II hardware.

B. Reduction of VME Link Traffic

To improve network performance of the RAID-II system, we include only the essential fields of Ultranet request blocks in the messages between the Sun-4 and the HIPPI boards. This was done to reduce the utilization of the slow VME link between the Sun-4 and HIPPI boards. This link is capable of handling 2MBytes/second. The link must carry messages between the Sun-4 and HIPPI boards and file system metadata. The "compression" achieved is summarized in Table VII. On the average, messages are reduced in size by 50%.

TABLE VII. Ultranet Request Block Size in Bytes

Request Block	Normal Size	Compressed Size
OPEN	44	20
LISTEN	92	56
CLOSE	92	20
SEND	44	24
RECEIVE	44	24

Unfortunately, the utilization of the link is not easily measurable. As a result, we cannot identify if it is near saturation.

C. Network Performance

The UltraNetwork currently installed at UC Berkeley supports three Sun VME workstations. Each Sun workstation can produce or consume approximately 3.5MBytes/second [Clinger89]. This provides a maximum aggregate bandwidth of 10.5MBytes/second. RAID-II is capable of completely satisfying this network load (and more). Under the current maximum load, all clients receive data at their full desired bandwidth. Therefore, bandwidth limitations of the RAID-II network interface can currently only be estimated from scaling arguments. The performance numbers reported are based on a thousand packets of a fixed size sent over a single connection between the XBUS memory in RAID-II and a client machine on the Ultranet. The time to complete these transfers was used to obtain both average bandwidth and latency measurements for various packet sizes.

Figure 6 shows the bandwidth of data for different sized packets being sent between RAID-II and individual clients. The bandwidth of a Cray supercomputer communicating with a single Sun-3 client is shown for comparison. SunOS 3.5 operates approximately 10-15% faster than SunOS4.1. The maximum bandwidth for the Sun-3 clients is 3.5MBytes/second reading data from RAID-II and 3.7MBytes/second writing data to RAID-II. The maximum bandwidth for the Sun-4 clients is 3.0MBytes/second reading data from RAID-II and 3.8MBytes/second writing data to RAID-II. This large performance gap reading and writing data from a Sun-4 is due to cache conflicts in the Sun-4 memory system. When data is being written to the Sun-4 memory from the network, the virtually addressed cache in the Sun-4 must be updated. This results in a lower bandwidth writing to the Sun-4 memory.

Figure 7 shows the latency to send different sized packets between RAID-II and individual clients. The performance of a Cray supercomputer communicating with a single Sun-3 client is shown for comparison. The minimum latency of packets for a Sun-3 is 6.0ms reading from RAID-II and 4.8ms writing to RAID-II. The minimum latency of packets for a Sun-4 is 2.2ms reading from RAID-II and 1.3ms writing to RAID-II. Measurements of the RAID-II hardware [Chen93] indicate that approximately 1.1ms of this latency is due to delays in the file server. These numbers indicate that it is the processing speed of the clients that limits the end-to-end latency of communication.

Applications on the clients are unable to consume the 3.5Mbytes/second of data delivered. For example, video stored on the RAID-II file server can be played back on the Sun-4 clients at a rate of 5 frames/second. This corresponds to a transfer rate of 1.5Mbytes/second. The video data is copied across the clients VME backplane twice, first from the network interface to memory and then from memory to the frame buffer. This contention for the VME backplane reduces the available bandwidth in half.

D. CPU Utilization

The network software for RAID-II splits the workload of network communication across three processors, the Sun-4 host CPU and the two AMD 29K CPUs on the HIPPI boards.

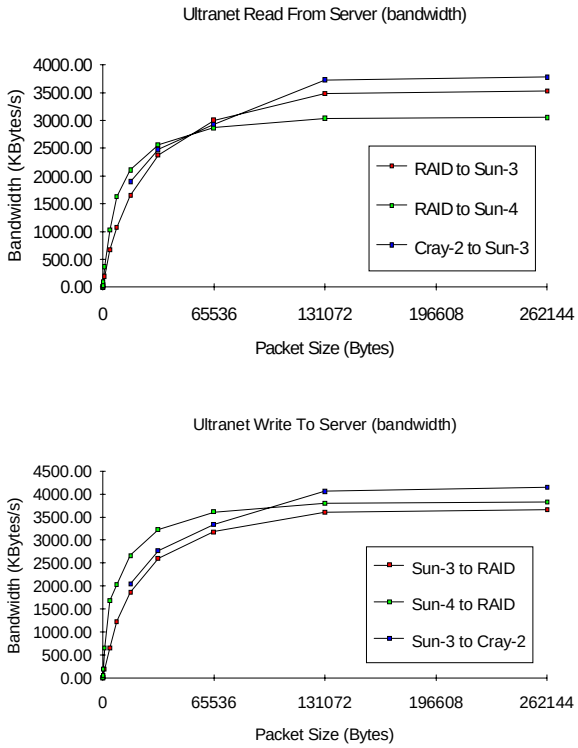

Fig. 6. Bandwidth vs. Packet Size for transfers between RAID-II and a single client

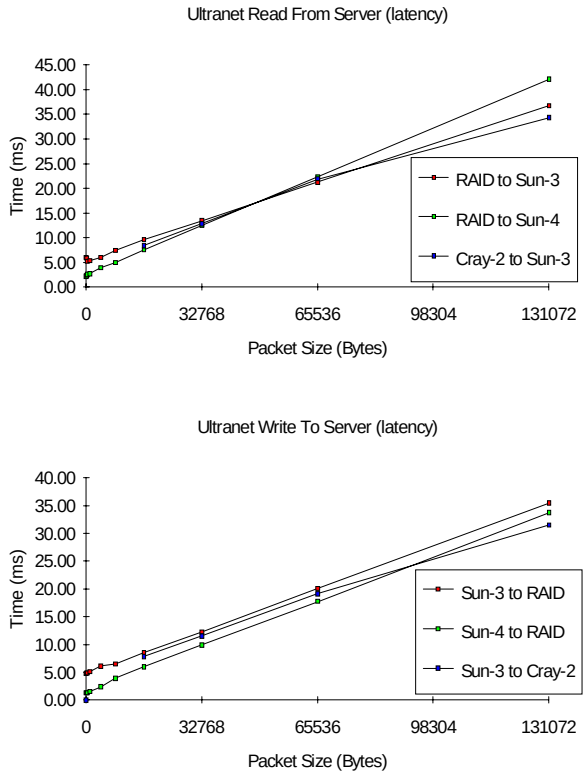

Fig. 7. Latency vs. Packet Size for transfers between RAID-II and a single client.

In this section, we examine the CPU utilization of these processors during transfers.

The utilization of the Sun-4 CPU is highly dependent on the packet size of the transfers occurring. Figure 8 shows the utilization of the Sun-4 CPU when all three clients transferring data. The three clients consume or create approximately 10.5MBytes/second of data traffic. When the clients are writing data to RAID-II the host CPU must take a single interrupt per packet. As a result the load on the host CPU is inversely proportional to the packet size. When clients are reading data from RAID-II, the host CPU must be interrupted for every outgoing data fragment transfer requested by the Ultranet hub. All packets are fragmented into 32Kbyte transfers across the HIPPI channel. As a result, the host CPU utilization has a minimum of 48%. CPU utilization remains almost constant for packets larger than 32Kbytes.

The host CPU utilization limits the network performance of clients reading from RAID-II to 21MBytes/second, about twice the currently available performance. Since packets on the UltraNetwork can be several megabytes, the host utilization places no limits on the bandwidth of clients writing data to the RAID-II system

The utilization of the 29K CPUs on the HIPPI boards depends mostly on the bandwidth of data being transferred. This is due to the fact that the 29K processors have a fixed computation overhead per 32KByte fragment transferred on the HIPPI channel. Their utilization is, therefore, not dependent on packet size but only on the actual bandwidth of data.

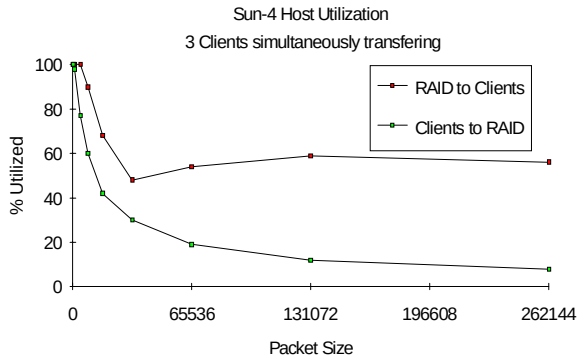

Fig. 8. Sun-4 Host CPU utilization vs. Packet Size for 3 clients communicating with RAID-II

Table VIII shows the utilization of the 29K CPUs for different bandwidths of data. When writing data to RAID, the destination board is highly utilized since it must setup and perform the data transfers. The source board only processes outgoing request blocks for these transfers. During reads from the RAID system, the source board must perform the overhead of transferring the data. The destination must still setup the transfers of data.

These numbers indicate that the 29K CPUs would limit the network to approximately 32Mbytes/second for both reads

TABLE VIII. Utilization of 29K Processors during Network Transfers

Bandwidth (Mbytes/ second)	Read From RAID		Write To RAID	
	Source 29K	Dest. 29K	Source 29K	Dest. 29K
3.5	21%	7%	18%	14%
7.0	N/A	18%	16%	23%
10.5	35%	20%	18%	27%

and writes to RAID-II.

V. Conclusions

The two basic goals of the RAID-II network software were to provide high bandwidth to clients on the UltraNetwork and reduce the load on the host CPU. [Chen93] measurements indicate that the RAID-II system hardware can support a raw bandwidth of 38.5MBytes/second between memory and the network. On the basis of our scaling estimates, the RAID-II server can source approximately 21Mbytes/second to the Ultranet (limited by the host CPU) and sink 32MBytes/second (limited by the destination board 29K CPU) from the network. Upgrading the host CPU to more modern hardware would allow the RAID-II system to source 32MBytes/second to the network. This bandwidth is significantly higher than that of Ethernet-based file servers in our environment. For comparison, our Sprite OS file server supports a bandwidth of about 1MByte/second to the network [Welch90]. These results show that the RAID-II network interface was effective at providing a high bandwidth to clients on the Ultranet. Although the software design did reduce the load on the host CPU by effectively using the 29K CPUs, we could not prevent the host CPU from being a critical resource for sourcing data. We feel that the network performance of the RAID-II server with Ultranet clients cannot be improved significantly.

With some minor hardware changes, there are a number of mechanisms to improve the performance of the system to the maximum 38.5Mbytes/second. First, the limiting CPU utilizations could be reduced by sharing access to the HIPPI source board by the host CPU and the HIPPI destination board. The sharing would make it unnecessary to interrupt the Sun-4 host every 32Kbytes. However, this sharing is impossible to achieve efficiently without an improved VME interface on the HIPPI boards. Another possibility would be using larger packets to communicate to and from the TMC HIPPI boards. The Ultranet hub architecture currently limits us to 32Kbyte transfers. The utilization of both the 29K CPUs and the Sun-4 CPU would greatly be reduced by the use of larger packets. This would allow us to scale to much higher bandwidths. To increase the packet size, we plan on replacing the Ultranet with a HIPPI switch network. Using the HIPPI switch network we hope to support transfers at over 70Mbytes/second to a pair of XBUS boards.

VI. References

[AnonA] *Network Operations Manual,* Ultra Network Technologies, Part Number 06-0001-001, Revision A, (1990). Chapter 2: UltraNet Architecture; Chapter 3: UltraNet Hardware.

[AnonB] *HPPI Destination Module (HPPID) Hardware Specification.* Thinking Machine Corp. October 1990.

[AnonC] *HIPPI Source Interface Hardware Register Specification.* Thinking Machine Corp. September 1990.

[ANSI91] *High-Performance Parallel Interface - Framing Protocol (HIPPI-FP)*, American National Standard for Information Systems X3T9.3/89-013 Rev 4.2. June 1991.

[Chen93] Peter M. Chen, Edward K. Lee, Ann L. Drapeau, Ken Lutz, Ethan L. Miller, Srinivasan Seshan, Ken Shirriff, David A. Patterson, Randy H. Katz. Performance and Design Evaluation of the RAID-II Storage Server. to appear in *International Parallel Processing Symposium 1993 Workshop on I/O*.

[Chervenak91] Ann L. Chervenak and Randy H. Katz. Performance of a Disk Array Prototype. *Proceedings of the 1991 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, volume 19, pages 188-197, May 1991.

[Clinger89] Marke Clinger. Very High Speed Network Prototype Development; Task 2.1: Measurement of Effective Transfer Rates. Ultra Network Technologies. October 1989.

[Katz91] Randy H. Katz. High Performance Network and Channel-Based Storage. *Proceedings of the IEEE, Vol 80, No. 8.* pages 1238-1260. August 1992.

[Katz93] Randy H. Katz, Peter M. Chen, Ann L. Drapeau, Edward K. Lee, Ken Lutz, Ethan L. Miller, Srinivasan Seshan, and David A. Patterson. RAID-II: Design and Implementation of a Large Scale Disk Array Controller. *1993 Symposium on Integrated Systems,* 1993. University of California at Berkeley UCB/CSD 92/705.

[Lee92] Edward K. Lee, Peter M. Chen, John H. Hartman, Ann L. Chervenak Drapeau, Ethan L. Miller, Randy H. Katz, Garth A. Gibson, and David A. Patterson. RAID-II: A Scalable Storage Architecture for High-Bandwidth Network File Service. Technical Report UCB/CSD 92/672, University of California at Berkeley, February 1992.

[Patterson88] David A. Patterson, Garth Gibson, and Randy H. Katz. A Case for Redundant Arrays of Inexpensive Disks (RAID). In *International Conference on Management of Data (SIGMOD),* pages 109-116, June 1988.

[Welch90] Brent B. Welch. Naming, State Management, and User-Level Extensions in the Sprite Distributed File System. University of California at Berkeley UCB/CSD 90/567. April 1993