Data Storage and Retrieval System

24 July 1991

Glen Nakamoto

MITRE Corporation
Bedford, MA 01730
(617) 271-3032

Data Storage and Retrieval System

Background

The Data Storage and Retrieval System (DSRS) consists of off-the-shelf system components integrated as a file server supporting very large files. These files are on the order of one gigabyte of data per file, although smaller files on the order of one megabyte can be accommodated as well. For instance, one gigabyte of data occupies approximately six 9 track tape reels (recorded at 6250 bpi). Due to this large volume of media, it was desirable to "shrink" the size of the proposed media to a single portable cassette. In addition to large size, a key requirement was that the data needs to be transferred to a (VME based) workstation at very high data rates. One gigabyte (GB) of data needed to be transferred from an archiveable media on a file server to a workstation in less than 5 minutes . Equivalent size, on-line data needed to be transferred in less than 3 minutes. These requirements imply effective transfer rates on the order of four to eight megabytes per second (4-8 MB/s). The DSRS also needed to be able to send and receive data from a variety of other sources accessible from an Ethernet local area network.

System Configuration

In order to meet these requirements, a system was configured using Aptec's Input/Output Computer (IOC-24) with Storage Concepts C51 disk array and Honeywell's Very Large Data Store (VLDS) tape drive (dual channel unit) as the basic components for this file server. The IOC-24 has eight megabytes of shared memory and was hosted on a VAX 11/750 which, in turn, was connected to an Ethernet local area network (LAN). The interface to the VME based workstation was accomplished via Aptec's VME Gateway Controller. The specific (and initial) VME based workstation that needed to be interfaced for this project was the Sun 3/260 workstation containing a Vicom II-9 image computer and the Vicom Fast Disk (Maximum Strategy's parallel disk array). The Sun workstation also contained a 16 megabyte high speed (32 MB/s) memory card from Micro Memory. This memory card was used as a high speed receiver (or transmitter) of data during the initial period (prior to the Vicom Fast Disk becoming available) for debugging purposes. Data needed to be transferred to/from the DSRS file server (C51 disk or VLDS tape) to the parallel disk array under the control of the Sun workstation operating at the data rates previously discussed. The specific configuration is illustrated in Figure 1.

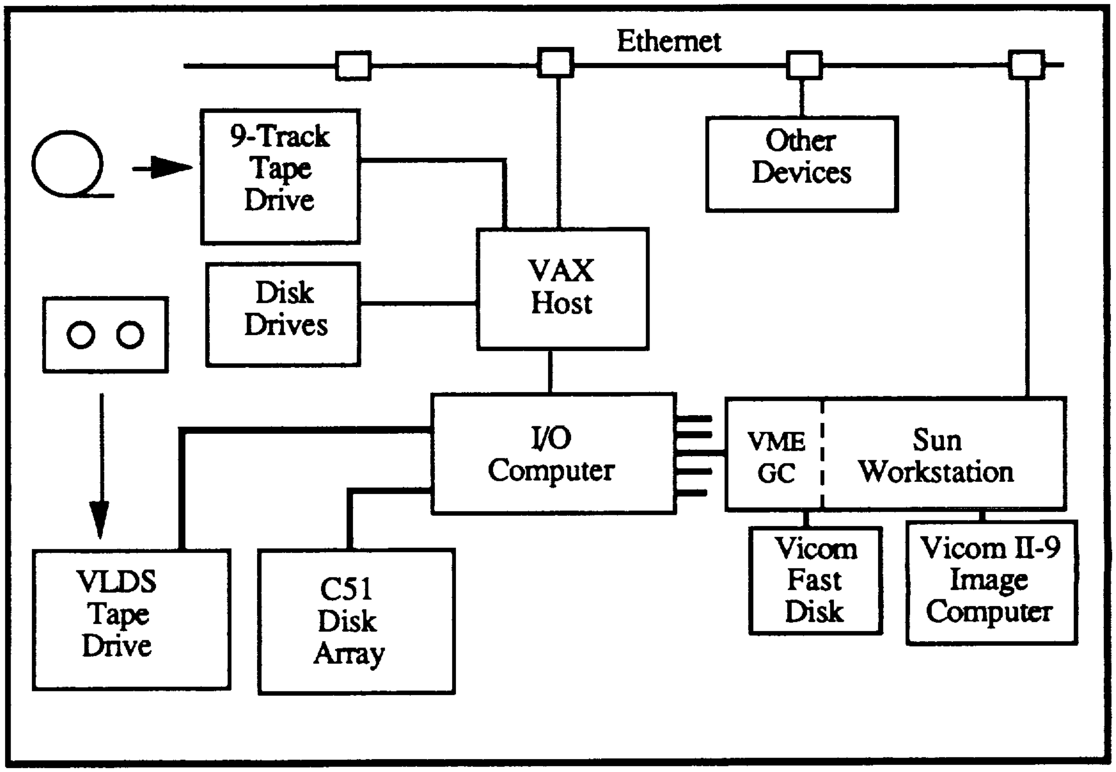


Figure 1. Data Storage and Retrieval System

Component Performance

From a performance standpoint, the C51 disk is the fastest peripheral on the DSRS. It has been measured as transferring (across the VMEbus to high speed memory) in excess of 9 MB/s using transfer block size of one megabyte. The transfer block size is critical in determining the effective transfer rate. Figure 2 shows how the effective transfer rate varies with the block size used when transferring data from the C51 to the Vicom Fast Disk (C512FD) or vice versa (FD2C51). These transfers were done using files on the order of 500 M bytes of data. For the DSRS application, a block size of two megabytes (512 K words) was chosen. This final size was dictated more by the Vicom Fast Disk rather than the C51 disk. For this configuration, the C51 disk array has 2.5 gigabytes of formatted disk space.

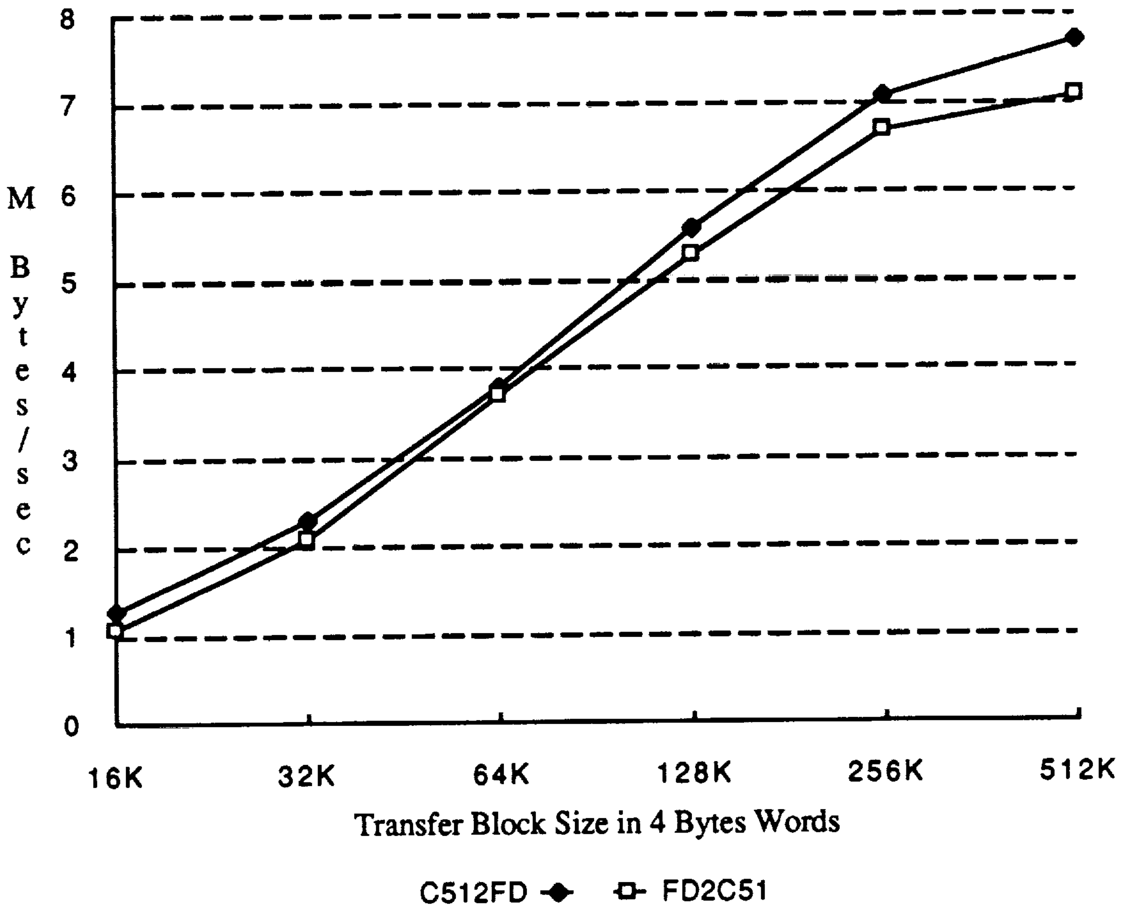


Figure 2. Transfer Rate as a Function of Block Size

The VLDS is a streaming tape drive and streams data at approximately 4 MB/s. Since it cannot start/stop like a conventional tape drive, it is imperative that it operate at its full transfer rate. In the read/playback mode, the VLDS can stop/restart taking approximately eight seconds to restart. Any significant mismatch in transfer rates between the VLDS and another device could slow the overall transfer rate down to eight kilobytes per second (KB/s). On the write/record side of the transfer, the VLDS will write "padding blocks" while continuing to stream. However, there is a maximum number of padding blocks that is (user) specified prior to the system halting (i.e., no restart). VLDS tapes that have padding blocks will have a natural degradation in transfer rate as well as in tape capacity. With no padding blocks, one VLDS cassette (super VHS T-120 cassette) will store approximately 5.2 gigabytes.

The VME gateway controller can effectively move data from the IOC's shared memory to the VMEbus on the workstation at rates in excess of 11 MB/s. The transfer rate also varies as a function of block size as shown in Figure 3. These transfer rates were obtained using the VME gateway controller in the Master mode. In the final configuration of the DSRS, the VME gateway controller is used in the Slave mode. Experience to date indicates that Slave mode transfer performance is similar to Master mode operation. For Sun 3/2XX workstations, the CPU exhibits a 190 microsecond bus timeout. This means that the VME GC cannot hold the bus longer than 190 microseconds after the CPU asks for it else it times out and the Sun operating system (SunOS) crashes. Since the VME GC operates in a release-when-done (RWD) mode versus release-on-request (ROR), it holds

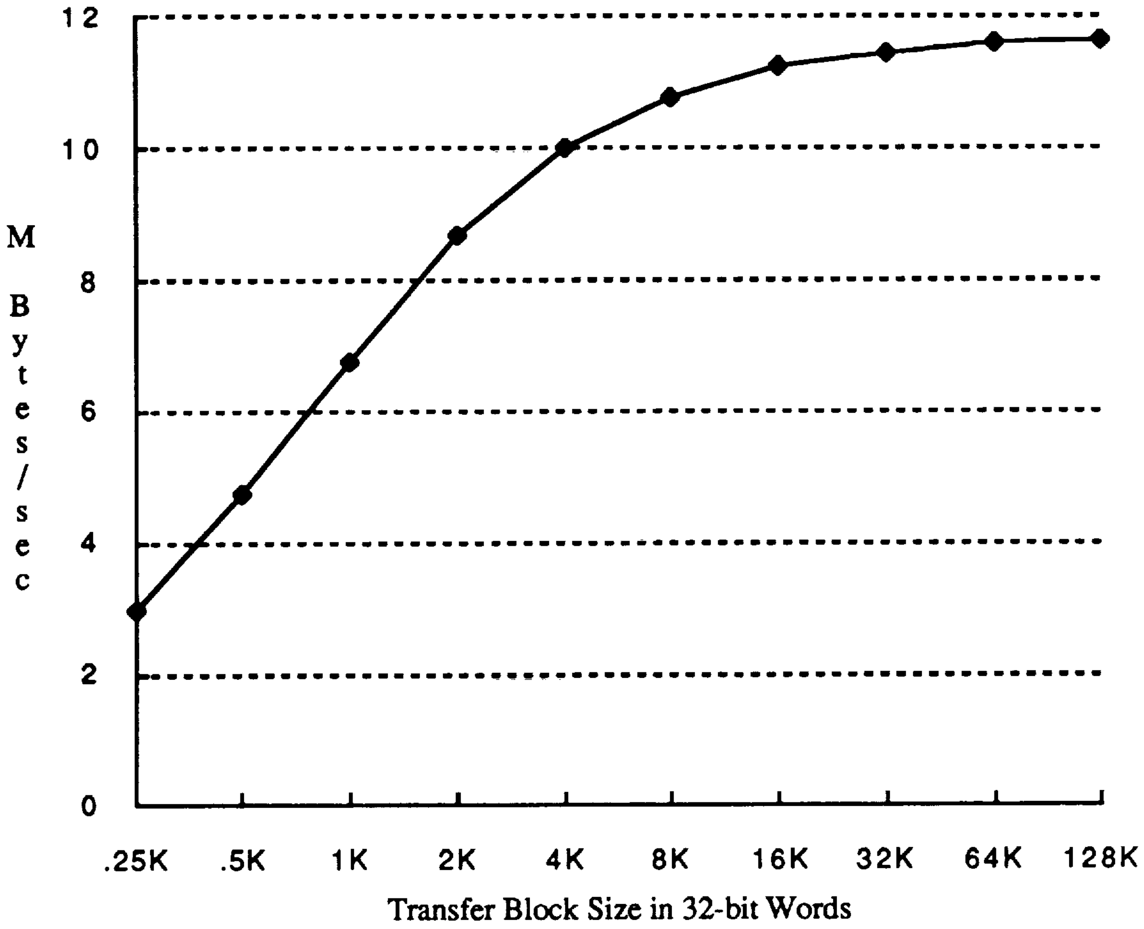


Figure 3. VME Transfer Rate as a Function of Block Size

the bus for the entire transfer. This practically limits the maximum block size to under one kilowords in Master mode operation (in order to stay under the 190 microsecond time limit). This limitation was overcome by using the VME GC in Slave mode and using a fast master controller (Maximum Strategy disk controller) operating in a ROR mode to effect the high speed transfer.

While not part of the DSRS, the Vicom Fast Disk is the other key device when examining end-to-end file data transfer. The Vicom Fast Disk is actually a disk array made by Maximum Strategy, Inc. (Strategy One Controller). In this configuration, the disk array contained approximately 6 GB of formatted space. The hardware is unchanged from the original product. However, Vicom has written a device driver and a file management system for it. The Vicom Fast Disk was rated at running at eight MB/s. While Vicom literature indicates 12 MB/s burst and eight MB/s sustained, in reality, 12 MB/s burst only occurs during the first megabyte of data (since it comes from memory and not the disk array). The eight MB/s sustained rate applies to the transfer once the disk heads are in position and is "streaming" data. In reality, large block sizes were needed to maintain transfer rates near eight MB/s.

Key Challenges

While all the hardware components were available off-the-shelf, no software was available to allow the components to function as a system. The key integration task was to develop the software and ensure that device performance would not be impeded by this software. A key challenge was to ensure that the VLDS tape unit operated at its full 4 MB/s capacity since it functions as a streaming tape drive. As indicated earlier, running the VLDS at slower than 4 MB/s would significantly slow down the total transfer time. Paramount to the development effort was to keep all overhead to an absolute minimum. Another challenge was to keep all software development at a high level - not develop any assembly language or microcode. Efficient use of library routines was essential. In order to promote a high level of portability, all routines developed on the Sun workstation had to be written in the C language and interfaced to existing Vicom device drivers. No kernel modifications were allowed as well. Since the software effectively controlled key hardware components directly, it was extremely difficult to debug since the typical error message was "bus error" (followed by a system crash). The use of a bus logic analyzer was essential to do problem identification and debugging. During this development effort, several key problems were discovered and fixed on the VME gateway controller. All problems dealt with the use of the VME GC in slave mode operating at high (~12 MB/sec) speeds. Software modifications at the microcode level (in Aptec's software) were also made to get the system operating properly. These changes have now been incorporated as part of Aptec's baseline.

VLDS Tape Unit

The VLDS was the first peripheral to be interfaced to the IOC-24. By using a dual buffering scheme to keep data transferring between the VLDS and the shared memory, the effective transfer rate of 4 MB/s was easily maintained when sending data to the IOC's memory. When the VME gateway controller (VME GC) was added to the IOC, data was then made to flow from the VLDS, to the shared memory, to the VME GC Input/Output Processor (IOP), to the VME GC, and then on to a high speed memory card on Sun 3/260 workstation with no delays. Later when the Vicom Fast Disk was added to the Sun workstation, it became apparent that the Fast Disk needed large (one megabyte or larger) transfer blocks in order to maintain high throughput. Since the VLDS reads and writes in principal block increments of 65,536 bytes (64 KB), a buffer size mismatch needed to be fixed. The dual buffering scheme had to be modified to accommodate 64 KB buffers on the VLDS side and 2 MB buffer size on the VME/Vicom side. This was accomplished by using multiple VLDS buffers adding up to the (two MB) VME side buffer, then taking into account partial buffers and last buffer anomalies. With this approach, it became possible to transfer files from slow devices such as the Sun local disk to/from the VLDS at high data rates (as fast as the slowest device) for files up to 2 megabytes in size with no degradation in performance. Larger files could also be sent but the VLDS start/stop action would cause a degradation in performance.

C51 Disk Array

Interfacing the C51 disk array into the DSRS involved making a key decision regarding its use. The C51 could be used in a dedicated manner, i.e., used by a single process until completed, or shared like a disk server. Used in a dedicated manner, the performance would be optimized and the software would be easier to develop. The major drawbacks were that the disk array would not be shareable between processes and only contiguous files would be supported. This last condition was quite restrictive when there is 2.5 GB of disk space and file sizes of one to two GB could be expected. The contiguous requirement would prevent files from being written even though the space may exist due to disk fragmentation or possibly bad disk blocks/sectors. The DSRS configuration uses the more complex VAX/VMS file management system (QIO) to create and manipulate files on the disk array. This allows files to be written out with multiple extents if needed. Another advantage of this approach is that the disk subsystem is shareable by different processes. Thus, a lengthy transfer that may take 3 minutes can function at the same time that another user is accessing a small file on the same disk, without having to wait for the first transfer to complete. From a VAX user viewpoint, the C51 appears as a standard disk (though non-system bootable) and functions, such as file transfer (ftp) and copy, can be used to transfer data from VAX based peripherals or other peripherals attached via the Ethernet LAN. Transfers involving the C51 also use a similar dual buffering scheme to maximize the transfer to/from the IOC's shared memory then on to the destination device.

VME Gateway Controller

The VME GC was, by far, the most challenging piece of equipment to understand and integrate into the system. A series of protocols were developed to allow the DSRS to communicate with the target workstation. The first protocol (command protocol) involved sending a command and appropriate parameters from the Sun workstation indicating what transfer needed to be done with what files. The second protocol (information protocol) involved communicating information regarding file size, transfer block size, and size of the last transfer. This information was critical to ensure both sides (DSRS and workstation) knew exactly what and how much data was being sent. Finally, the transfer protocol involved the low level "handshaking" needed to keep the data transferred in proper synchronization, i.e., ensure that source and destination devices were ready to receive the appropriate blocks of data. From an ISO networking viewpoint, these protocols fall into the application layer. They effectively establish a means of communications at a high level between a workstation and the file server (DSRS) for the transfer of a file. From a logical viewpoint, the VME GC is set up like a data structure featuring a first-in-first-out (FIFO) location, a mailbox area for small messages, and a set of 16 registers. The base address to this logical structure is user programmable and can exist anywhere within the workstation's memory space. Physically, the VME GC is also user programmable and can exist anywhere within the VMEbus 32-bit address space (barring conflicts with existing devices). The FIFO is used to transfer large blocks of data between the workstation and the DSRS. The mailbox region is used to pass file names, file sizes, and other miscellaneous pieces of information. The registers perform all control functions including setting the direction of transfer, number of bits per word, FIFO full/empty indication, semaphore acknowledgements, etc. The DSRS is set up with a predefined data structure. The Sun workstation also sees this same data structure and communicates with the DSRS by reading and writing to these memory locations as if the DSRS was local to it. By using this approach, different VME based workstations can be integrated with the DSRS with minimal difficulty. All that is needed is an understanding of how to memory map to a specific memory location and a VME memory device driver (both commonly standard with any workstation/operating system). An interface guideline document describing both the hardware interface requirements (for the VME GC board set) as well as a detailed description of the above protocols (for a software interface) has been published.

Typical Data Flow (C51 to Vicom Fast Disk)

Once a user on the Sun workstation "launches" a transfer function, the VME IOP spawns a request to the C51 IOP to initiate the file transfer. Four megabytes of shared memory (organized as two 2 MB buffers) are allocated in the IOC. The C51 IOP's task is to fill each 2 MB buffer and mark the empty semaphore flags as "full" upon completion of writing a buffer. At the same, time, the VME IOP reads the same buffers (when the semaphore indicates that the buffer is "full") and flags the buffer "empty" upon completion of the read function. The two processes run concurrently switching buffers as necessary to keep a steady flow of data moving. When the VME IOP reads the 2 MB buffer, the data is simultaneously transferred to the VME GC which has previously initiated "handshaking" with the Sun workstation CPU. In the meantime, the Sun CPU has transferred control to the Vicom Fast Disk controller which masters control of the VMEbus and extracts the 2 MB of data from the (FIFO address location of the) VME GC and transfers the data to its input buffer and subsequently to the disk array.

This single cycle which involves transferring data across three busses (DIB, OPENbus, and VMEbus) with "handshaking" and resource arbitration is executed 500 times (to transfer a gigabyte of data) and runs at 96% of the maximum burst speed of the slowest link.

Software Architecture

The software on the VAX was set up as a single program running as a server listening to commands (via a semaphore register) that it might receive from the Sun workstation. In actuality, the VME IOP has software constantly checking one of the registers to determine if a workstation wants to deposit a command into the mailbox. Once a command is received, the VME IOP is allocated and cannot be used by another device until the command has been completed. The VME IOP "downloads" the appropriate code and executes all transfer routines or spawns appropriate routines to accomplish the commands. Upon completion, the semaphore register is setup for the next request. During the transfer of information, if any substantial delays occur, either side will timeout and revert back to listening mode. This allows the server software to recover from errors that may occur on the workstation. Using this approach with multiple VME IOPs and VME GCs would allow multiple VME based workstations to access the DSRS resources in parallel, limited only by the device speeds, the amount of shared memory, or bus bandwidth within the IOC-24. Over 60 STAPLE routines/procedures were written to support this server software.

The software on the Sun workstation was written as a series of short C routines that effect a transfer from one device to another. A typical routine name was vlds2fd implying (in this case) transfer of data from the VLDS to the Vicom Fast Disk (FD). Following the command, the source and destination file names would be included as parameters. Ten of these routines were developed to transfer data in all conceivable directions. Over 30 routines/procedures were written to support the Sun based software.

Due to the development environment and the tools that were available, the entire system integration and software development effort for the DSRS took less than four months (with a staff of two). Interfacing the Sun/Vicom system took an additional two months (although hardware problems precluded use of the system for almost half of that time).

Conclusion

In the end, the DSRS successfully transferred one gigabyte of data from the VLDS to the Sun/Vicom Fast Disk in 4 min. 13 sec. + 25 sec. for tape setup and rewind. This translates to an effective transfer rate of 4 MB/s during the transfer (which is the streaming rate for the VLDS). The file transfer from the C51 disk array to the Sun/Vicom Fast Disk took 2 min. 15 sec. using 2 MB transfer blocks. This translates to an effective transfer rate of 7.7 MB/s (out of a theoretical maximum rate of 8 MB/s as constrained by the Vicom Fast Disk). File transfer was also performed between the C51 disk array and a device (a local disk attached to another Sun workstation) over Ethernet while, at the same time, transferring a file from the Vicom Fast Disk to the C51. Although the performance was slightly degraded due to the sharing of the C51 disk, it was not noticeable using small files (less than 10 megabytes).

All requirements were fulfilled using commercially available off-the-shelf components with a relatively small software development effort. The system is now operational and is being used to store and retrieve large files on VHS cassette tapes and can load the Sun workstation (Vicom Fast Disk) in minutes versus the many hours it used to take when using 9 track tapes.

Data Storage

and

Retrieval System

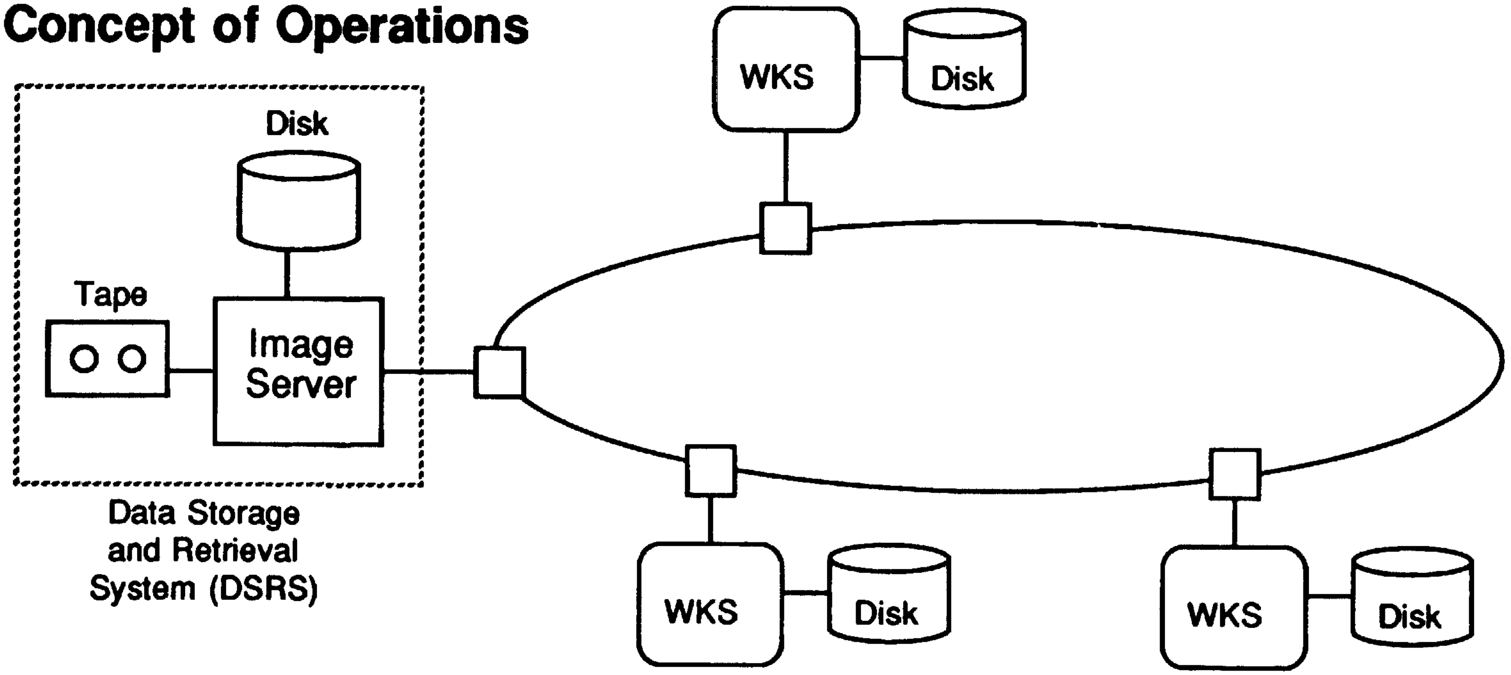
Glen Nakamoto

24 July 1991

MITRE

Background

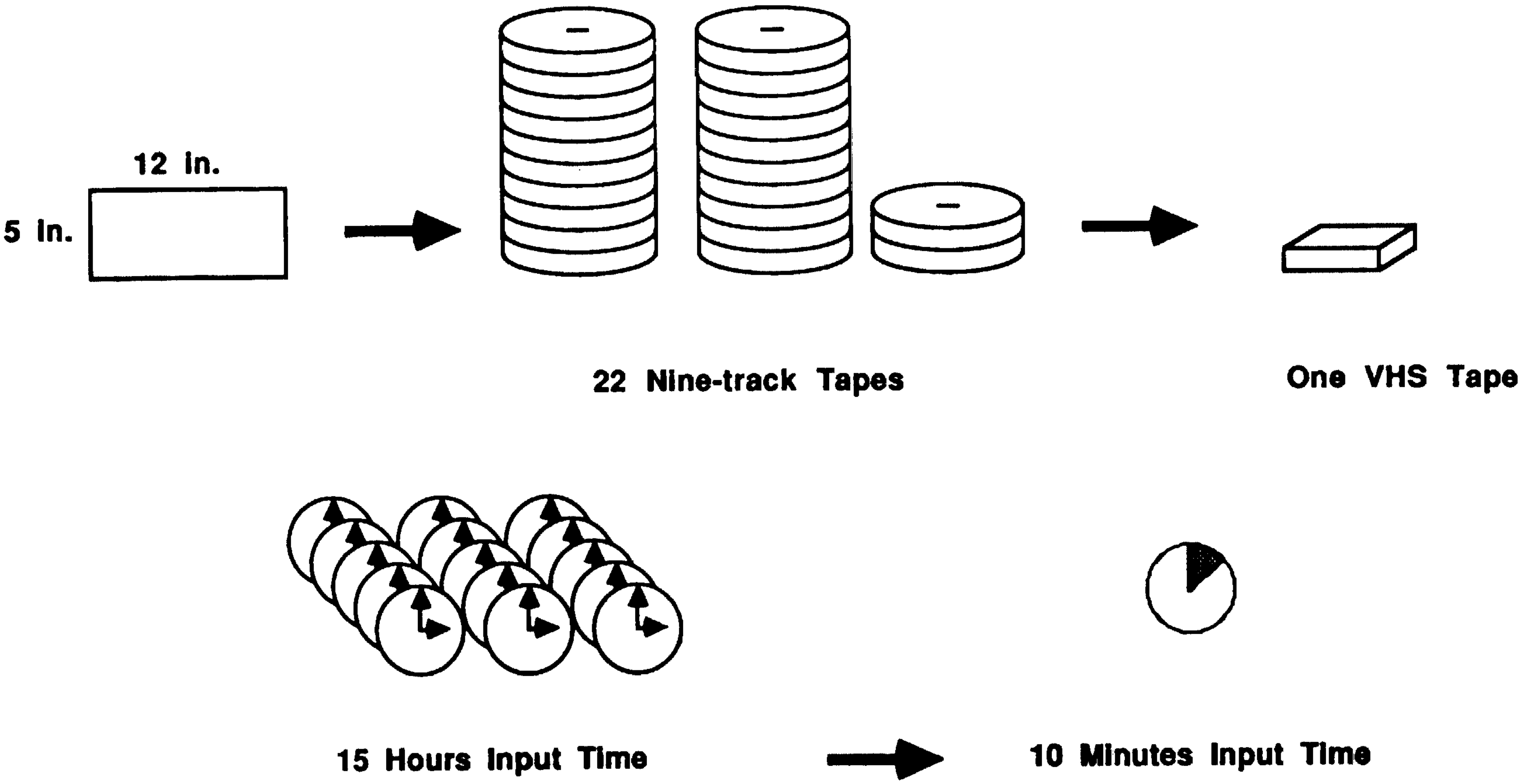
Concept of Operations



- Support evaluation of operationally-oriented softcopy imagery exploitation
- Two sessions per day; four hours per session
- Preload images into workstation prior to session
- Ad Hoc access to any image stored on the server

MITRE

A Key Problem



22 Nine-track Tapes

One VHS Tape

15 Hours Input Time

10 Minutes Input Time

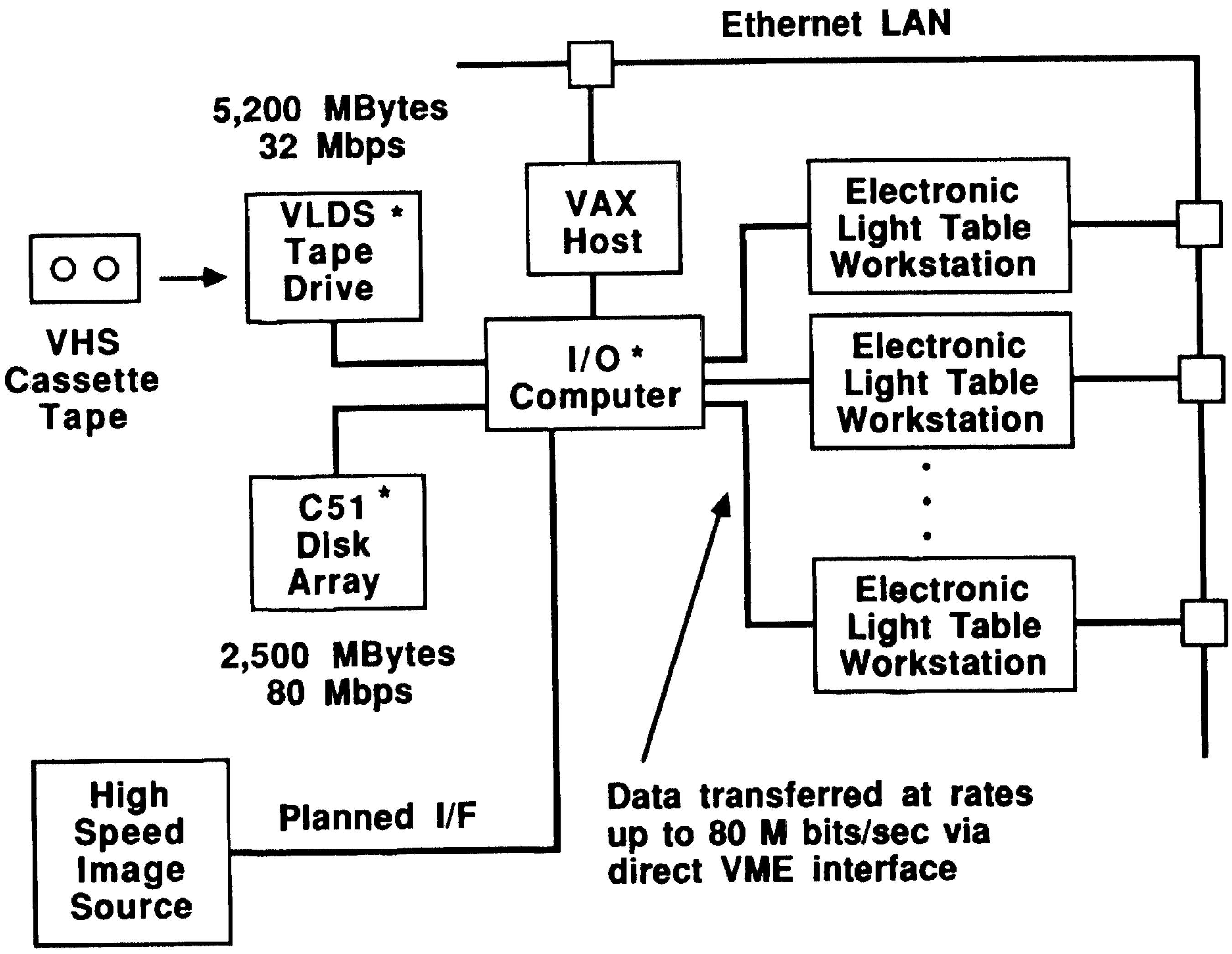
MITRE

FY90 Goal

- Develop a prototype data storage and retrieval system
  - Support image files 20 - 30 times larger using single portable media
  - Transfer files at rates 90 times current capability
- Establish interface guidelines for future workstations
  - Multiple standards (Physical, networking, application)
  - Portable media (Tape, format)
- Interface initial commercial workstation

MITRE

DSRS Architecture

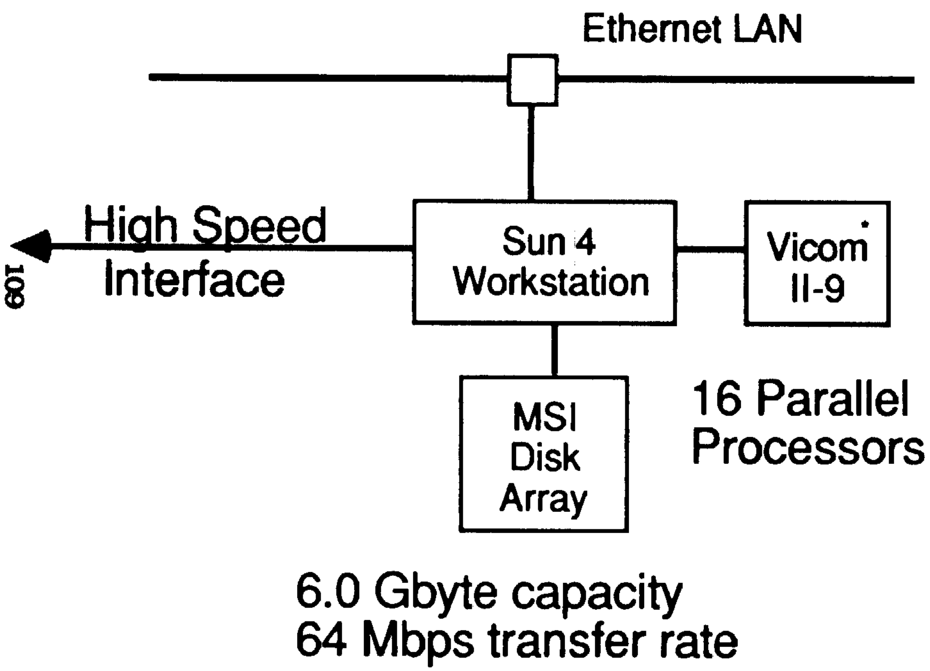


Data transferred at rates up to 80 M bits/sec via direct VME interface

* These items fit in a rack space of 27 inches.

MITRE

Electronic Light Table Workstation



- Electronic Light Table Functions
- Integrated Database -

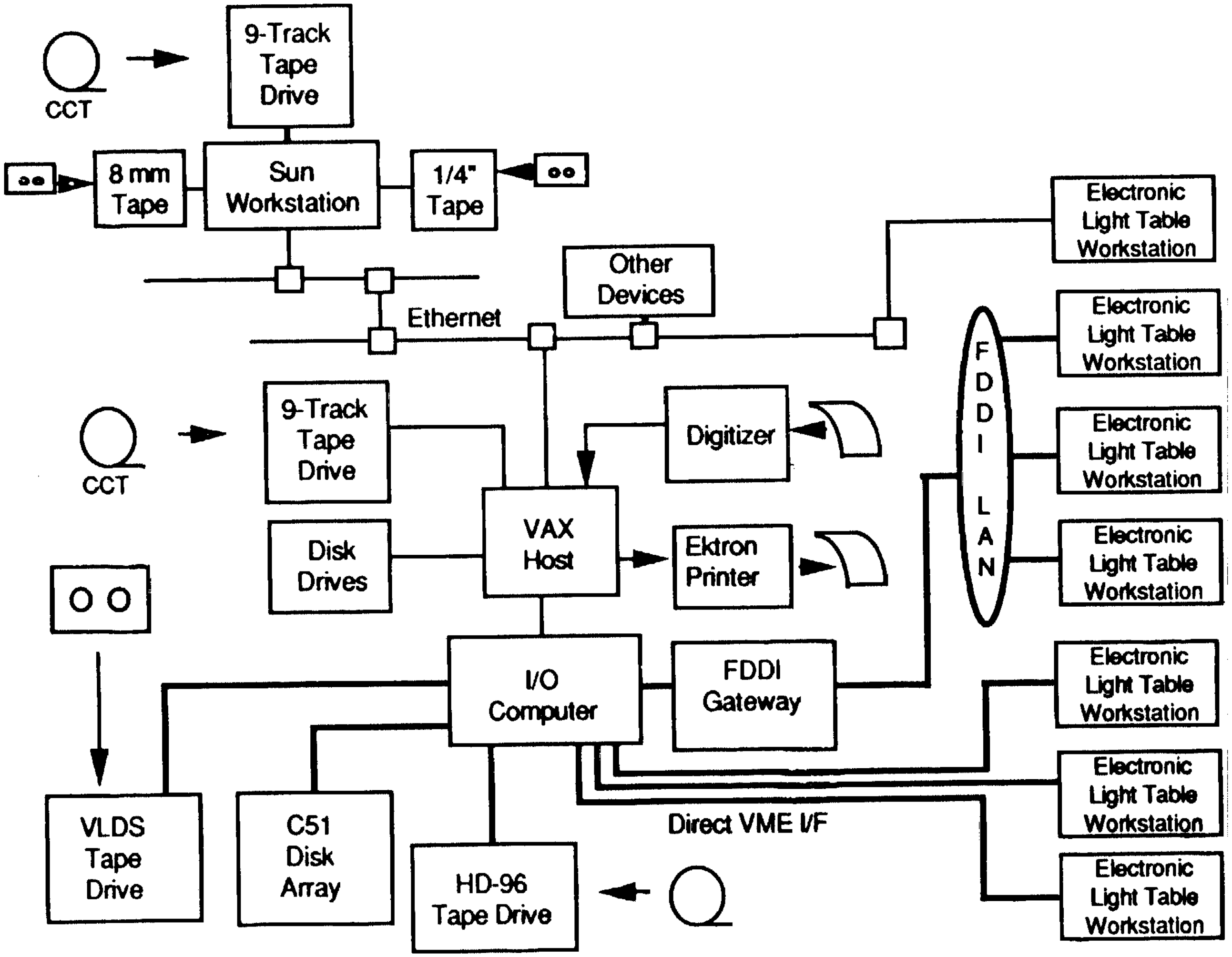
  Text, Graphics, Imagery
- Large Image File Manipulation
- Real-time Decompression
- 2K by 2K display size (search)
- High Speed Image Transfer

  (64 Mbps)
- Off-the-Shelf Hardware

* Previously Pixar

MITRE

Image Server Configuration



CCT

9-Track Tape Drive

8 mm Tape

Sun Workstation

1/4" Tape

Ethernet

Other Devices

CCT

9-Track Tape Drive

Disk Drives

VAX Host

Digitizer

Ektron Printer

I/O Computer

FDDI Gateway

FDDI LAN

VLDS Tape Drive

C51 Disk Array

HD-96 Tape Drive

Direct VME I/F

Electronic Light Table Workstation

Electronic Light Table Workstation

Electronic Light Table Workstation

Electronic Light Table Workstation

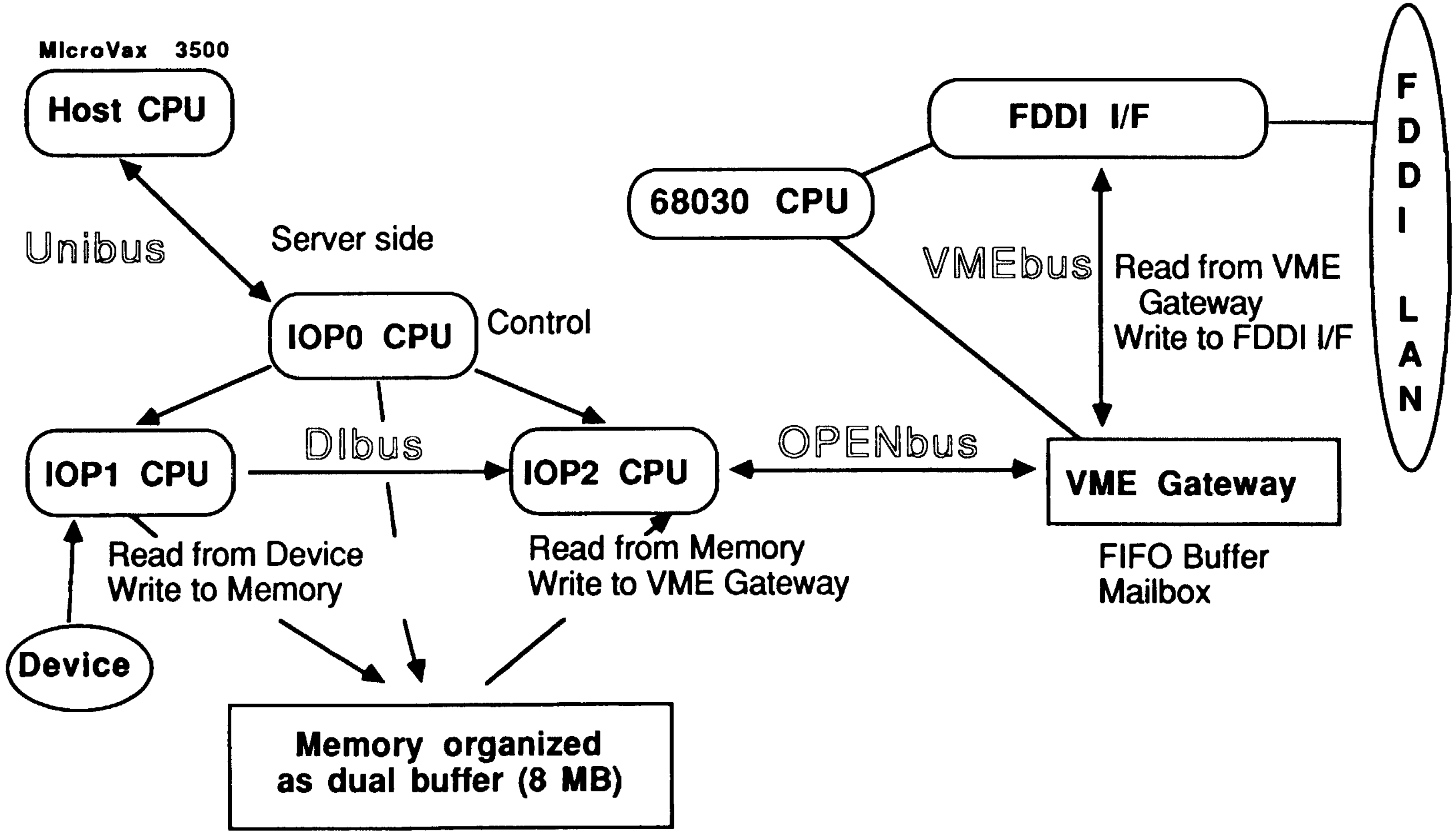
Electronic Light Table Workstation

Electronic Light Table Workstation

Electronic Light Table Workstation

MITRE

Software Architecture



MicroVax 3500

Host CPU

Unibus

Server side

IOP0 CPU

Control

IOP1 CPU

DIbus

IOP2 CPU

OPENbus

VME Gateway

FIFO Buffer
Mailbox

Device

Read from Device
Write to Memory

Read from Memory
Write to VME Gateway

Memory organized
as dual buffer (8 MB)

68030 CPU

FDDI I/F

VMEbus

Read from VME
Gateway
Write to FDDI I/F

FDDI LAN

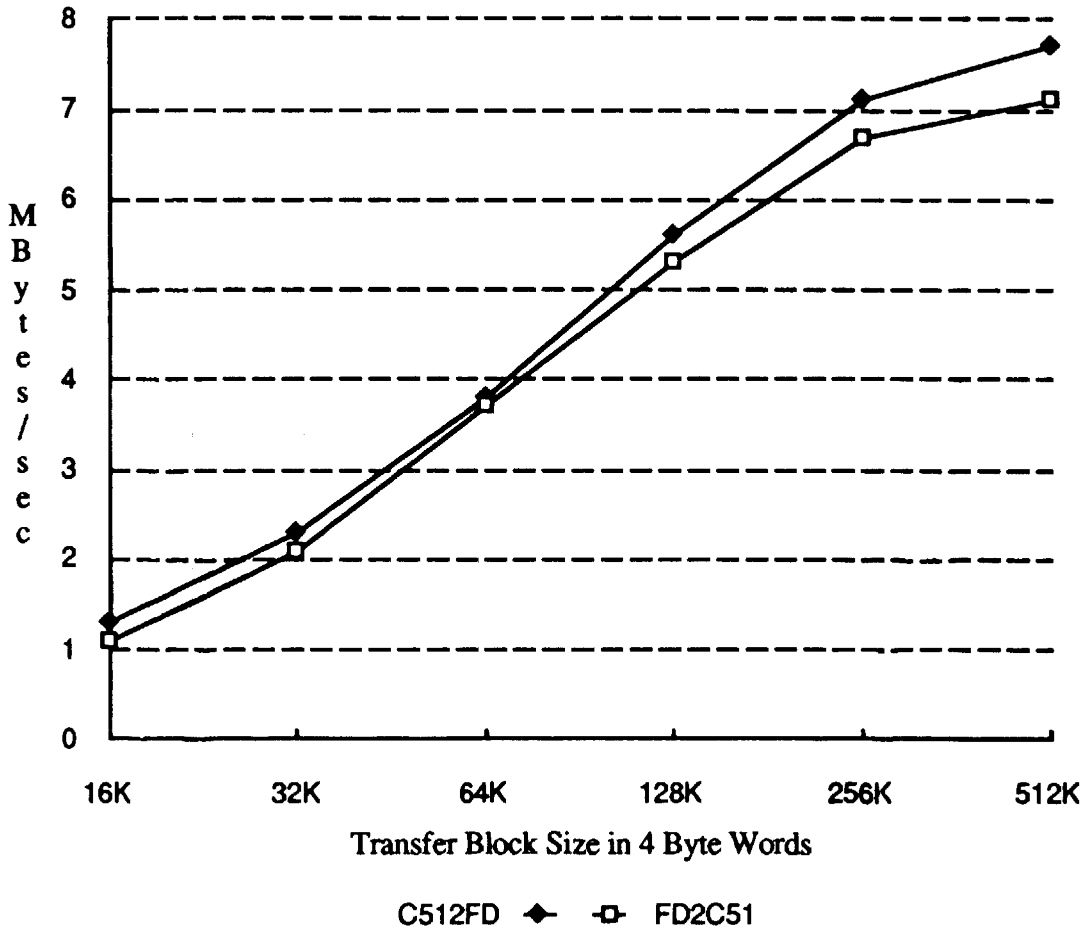
Measured Transfer Rates

Network	Burst Rate	Effective Throughput
Ethernet	10 Mbps	2.4 Mbps
FDDI	100 Mbps	2.5 Mbps
UltraNet	125 Mbps	20.0 Mbps
DSRS	96 Mbps	93.6 Mbps
IOC based FDDI	100 Mbps	48.0 Mbps (not measured)

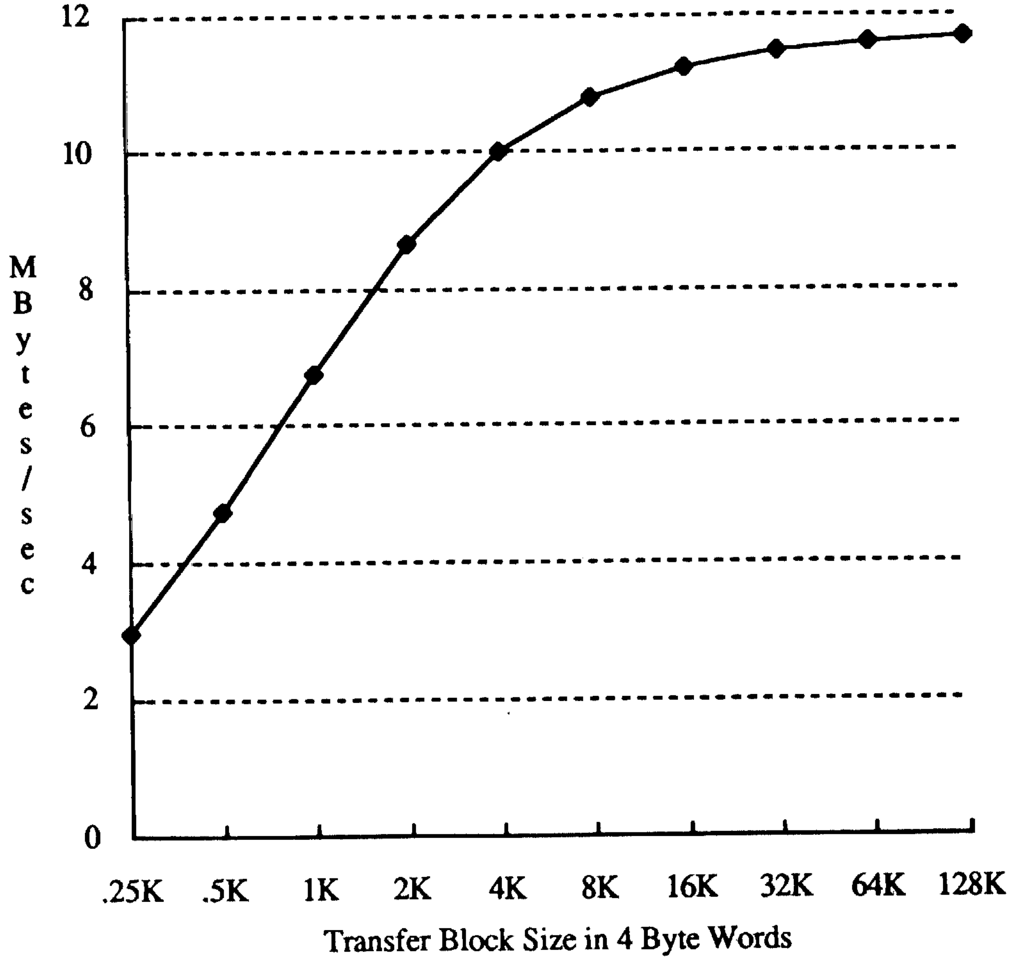
Notes:

Sun memory to Sun memory transfer with no network contention.
SunOS 4.1 (Sun 3/260 with 8 MB memory).
Sun provided TCP/IP software.
UltraNet used TP4/IP with board level protocol processing.

C51-PFD: Transfer Rate vs. Block Size



VME-MM Transfer Rate



MITRE

Summary

- Requirements for DSRS finalized in September 1989
- IOC based system purchased in October 1989
- System delivered on 18 January 1990
- The DSRS was developed and delivered 18 June 1990
- Integration of the Sun/Pixar workstation was completed by 9 August 1990
- Improves transfer times 90:1
- Improves storage approximately 100:1
- Allows search oriented experiments to be conducted
- Improves the management of an image library
- Promote standards and interface guidelines

MITRE

FY91 Goal

- Integrate an FDDI network into the DSRS
  - Develop an FDDI gateway for the DSRS
  - Initially support TCP/IP protocols
  - Provide capability to install other protocols
  - Provide capability to support multiple gateways per IOC
  - Maintain maximum performance end-to-end
- Upgrade IOC-24 to IOC-100
- Upgrade 2.5 GB Disk Array to 7.5 GB capacity
  - Provide means to address greater than 32 bits