NASA-TM-107989

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

WORLD DATA CENTER "A" for ROCKETS AND SATELLITES NSSDC/WDC-A-R&S 91-18

Proceedings of the NSSDC Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications

July 23 - 25, 1991

National Space Science Data Center
NASA/Goddard Space Flight Center
Greenbelt, Maryland 20771

NASA
National Aeronautics and Space Administration

Goddard Space Flight Center

NSSDC/WDC-A-R&S 91-18

Proceedings of the NSSDC Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications

Prepared by

Mass Storage Conference

Program Committee

July 23 - 25, 1991

Preface

This report presents the proceedings of the NSSDC Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications held July 23 through 25, 1991 at the NASA/Goddard Space Flight Center. Sponsored by the National Space Science Data Center, the program includes a keynote address, invited technical papers, and selected technical presentations to provide a broad forum for the discussion of a number of important issues in the field of mass storage systems. Topics include magnetic disk and tape technologies, optical disk and tape, software storage and file management systems, and experiences with the use of a large, distributed storage system. The technical presentations describe, among other things, integrated mass storage systems that are expected to be available commercially. Also part of the program is a series of presentations from Federal Government organizations and research institutions covering their mass storage requirements for the 1990s. The three-day conference provides attendees with a unique opportunity to exchange ideas and experiences at a practical and useful level.

All invited papers received prior to the conference appear in this publication in the order in which they were presented, with the exception of those not received in time for printing. In those instances, papers not received in time will be published subsequently in a second document following the conference. This later document will include a transcription of the discussions and question and answer periods of the conference.

This document was prepared by Kim Blackwell with editorial and production assistance from Len Blasso and Ann Lipscomb, all three of ST Systems Corporation, Lanham, Maryland.

CONTENTS

CONTENTS (CONT'D)

ANALYTIC REVIEW FORM OK (Initials) DRD

Primary Record IPS # 121937

___ Document should not receive analytic treatment

SUBSIDIARY ADD

Document page range ____ to ____ New subsidiary # _____
Document page range ____ to ____ New subsidiary # _____
Document page range ____ to ____ New subsidiary # _____

SUBSIDIARY DELETE/CORRECTION

Subsidiary #_____
(IPS#_____)

___ Delete
Reason: ___ limited technical content
___ no separate authorship
___ context dependent

___ Adjust paging
___ Other _____

Subsidiary #_____
(IPS#_____)

___ Delete
Reason: ___ limited technical content
___ no separate authorship
___ context dependent

___ Adjust paging
___ Other _____

Subsidiary #_____
(IPS#_____)

___ Delete
Reason: ___ limited technical content
___ no separate authorship
___ context dependent

___ Adjust paging
___ Other _____

StorageTek®

Enterprise Storage Report for the 1990s

by Fred Moore
Corporate Vice President
Strategic Planning
Storage Technology Corporation

Enterprise Storage Report for the 1990s

Abstract

Data processing has become an increasingly vital function, if not the most vital function, in most businesses today. No longer only a mainframe domain, the data processing enterprise also includes the midrange and workstation platforms, either local or remote. This expanded view of the enterprise has encouraged more and more businesses to take a strategic, long-range view of information management rather than the short-term tactical approaches of the past. This paper will highlight some of the significant aspects of data storage in the enterprise for the 1990s.

StorageTek
PROPRIETARY

ENVIRONMENT - 1990s

- Storage and storage management are most pressing issues
- Networking and connectivity requirements increasing at all levels
- DASD subsystem reliability and availability continue to improve
- Fault tolerant DASD architecture emerging
- Strong acceptance of automated libraries
- New focus and mission for magnetic tape technologies

As the 1990s begin, effective storage management remains possibly the most pressing issue. Poor device utilization and erratic performance are no longer accepted as normal conditions. The cost of ineffective storage management also has received considerable attention as storage costs now exceed the processor costs in mainframe environments.

The definition of the enterprise has moved quickly beyond the world of IBM mainframes to include other mainframes, midrange distributed processors and networks, local and wide area. The need to connect these processing platforms through standard network interfaces and provide access to common storage is increasing rapidly as most users now have mixed-vendor environments (Cray, DEC, IBM, Tandem, etc.) to manage.

The reliability of DASD subsystems continues to improve but even at 99.99 percent availability the only acceptable goal remains 100 percent availability. This trend has encouraged several companies to develop fault-tolerant DASD architectures. Fault-tolerant DASD subsystems provide continuous data availability in the event of any hardware component failure.

The successful introduction of automated tape systems such as StorageTek's 4400 Automated Cartridge System has led to widespread acceptance of automated storage. The 1980's view that library architectures were the least reliable component of a data center is now obsolete. The data processing industry has overcome preconceptions created by various mass storage and rail-type architectures of the past.

The highly successful launch of automated cartridge systems has given new life to tape data storage. Primarily used for low-activity backup, automation has allowed many new applications, not previously considered for tape, to become practical and cost-effective.

ENVIRONMENT - 1990s (Cont.)

- Automated operations becoming a strategic goal for most large installations
- Accelerated new application growth in PC/ workstation segment
- Outsourcing slowing
- Disk growth rates have moderated to 25-30% annually
- Image processing market slowly emerging
- IBM SYSPLEX re-focuses mainframe role

StorageTek
PROPRIETARY

Automated operations is quickly becoming a strategic goal of most large-scale data centers.

For the first time, there is now more storage outside the "glasshouse" mainframe environment than in it. This accelerated growth rate for storage away from the mainframe area will lead to system-managed storage structures, automated tape systems, multi-media libraries and fault-tolerant DASD for the midrange and desktop computing environments.

Outsourcing, a trend that gained considerable visibility in the late 1980s, has lost some of its appeal. Sourcing some or all computer operations, services and development to a source outside the enterprise is intended to save money. Though initial short-term financial gains are possible, many users now are viewing outsourcing as losing control of the most critical component of a business — the information processing function.

Mainframe DASD growth rates have moderated. In the early 1980s, DASD growth rates pushed a 60 percent annual increase in installed gigabytes. As the 1990s unfold, we note that most of the batch-to-online conversions are over. Secondly, data bases are now common, predominant in most organizations, eliminating redundant data. The third reason is that more users are beginning to make more effective use of storage management tools. Finally, many of the new applications that remain to be automated are emerging slowly such as image processing. Even at 25 percent compounded annual growth, the installed base of DASD capacity will double every three years.

It is believed that at the mainframe, midrange and workstation levels, image processing will be the dominant single driving factor for storage demand in the 1990s. This movement, however, has not evolved as quickly as projected due primarily to the lack of an effective enterprisewide image management architecture.

The role of the mainframe in the 1990s was clearly established by IBM's September 5, 1990, announcement of SYSPLEX. This announcement refocused the role of the mainframe in the enterprise as the central server and overseer of the networked enterprise. Mainframe architecture will continually drive many of the standards used for the entire enterprise.

DEMAND TRENDS

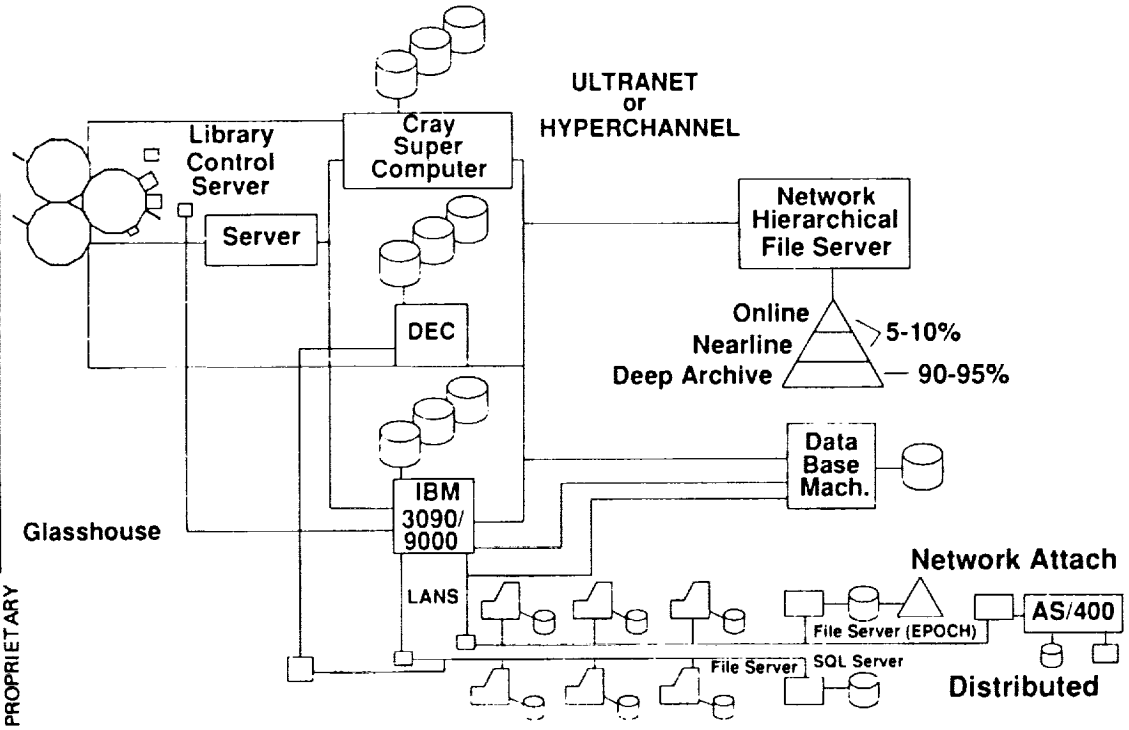
	Processor Demand (MIPS)				Storage Demand (GB)
	1960-1969	1970-1979	1980-1989	1990-1999	1990-1999
Mainframes	15%	20%	40%	25%	25-30%
Midrange	--	35%	50%	40%	40%
Workstation/ Desktop	--	--	150%	60%	55-65%

StorageTek
PROPRIETARY

Let's take a look at some of the growth rates that we have seen in the last three decades and will see in the future. This chart projects CPU or processor demand. Notice that processor demand for mainframes in the 1990s is expected to be around 25 percent, measured in MIPS.

We observe 25 percent growth in the mainframe, 40 percent in the midrange, and the workstation growing overall at about 60 percent annually during the 1990s. Today MIPS demand corresponds almost one-to-one with storage growth. Storage management and the ability to access all data objects from all computing platforms will become both a requirement and a major architectural challenge of the 1990s. The vendor that can resolve this problem best will control the enterprise.

ENTERPRISE DATA STORAGE ENVIRONMENT



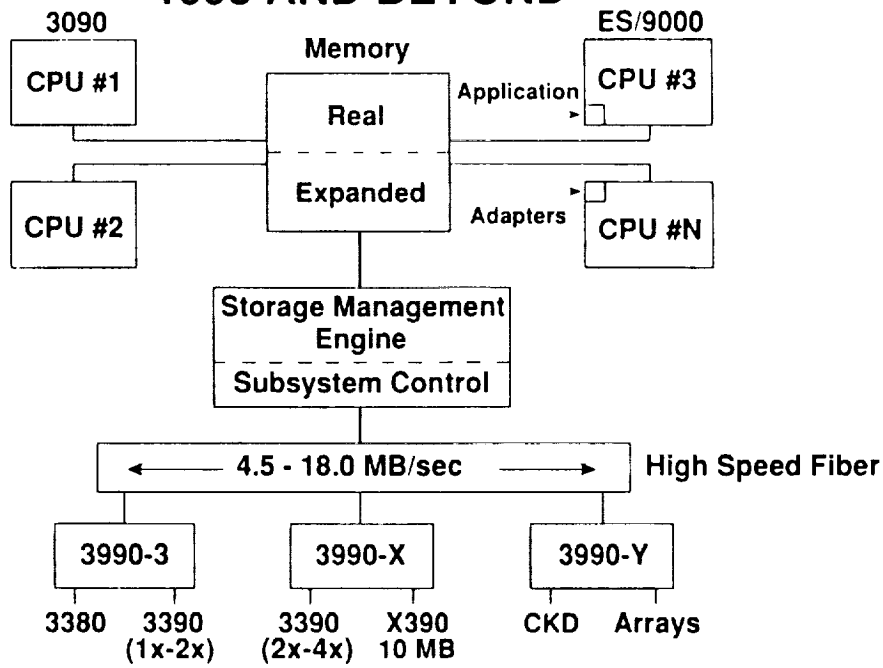
StorageTek
PROPRIETARY

We have had the traditional glasshouse or mainframe view of data processing for a long time. That glasshouse today, dominated by MVS and VM environments, is beginning to share the spotlight with the rapidly emerging midrange and workstation/desktop processing environments. The enterprise now requires management of heterogeneous and complex environments. These three platforms are clearly and distinctly emerging as major areas of processing and storage for the 1990s.

Data processing now becomes distributed at nodes in the enterprise and the objective is to allow transparent access while maintaining the security, integrity and performance of the environment. The role of large systems in the 1990s will become one of management and control for the information enterprise.

We will see the migration of MIPS and storage, and the management issues that go with them, move from mainframe to midrange and desktop. We are not going to be able to limit our views of storage management to MVS and glasshouse and IBM-only for much longer. Storage management solutions must cross those architectural and communication boundaries.

LARGE SYSTEMS ARCHITECTURE
1993 AND BEYOND



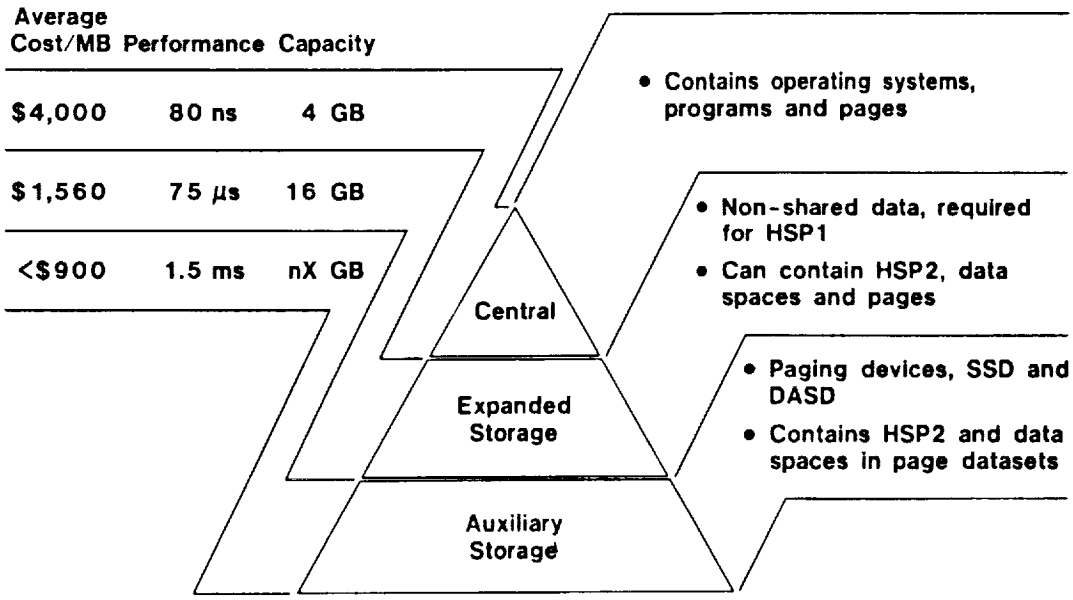
StorageTek PROPRIETARY

IBM's recent SYSPLEX announcement clearly refocused the role of the mainframe in the 1990s. Let's examine a likely scenario for the 1993 time frame and beyond. This has often been referred to as the post-Summit or Future Systems (FS) architecture. It is expected that up to a limit of 16 (CPU #N) ESA-based processors evolving within a SYSPLEX could be connected in the manner shown.

The continued roll-out of this architecture will include a shared expanded storage capability and application-specific adapters implemented via software, licensed internal code and hardware in varying amounts. Hardware assists for DFSORT announced in the September 5, 1990, IBM announcement are using this concept.

A storage management engine or I/O processor will be a new concept used to off-load from the host processor many of the I/O functions such as parts of IOS (I/O Supervisor), VSAM and DF (Data Facility) functions. The storage management engine will attach peripheral devices as we know them today (SSD, DASD, tape, printers and terminals) via the channel subsystem. Attachment of ESCON (Enterprise System Connectivity) serial fiber channels will be preferred though parallel bus and tag channels will need to attach via ESCON converters. The point-to-point limit of ESCON channel transfer rates is 18 MB/sec.

VIRTUAL STORAGE HIERARCHY



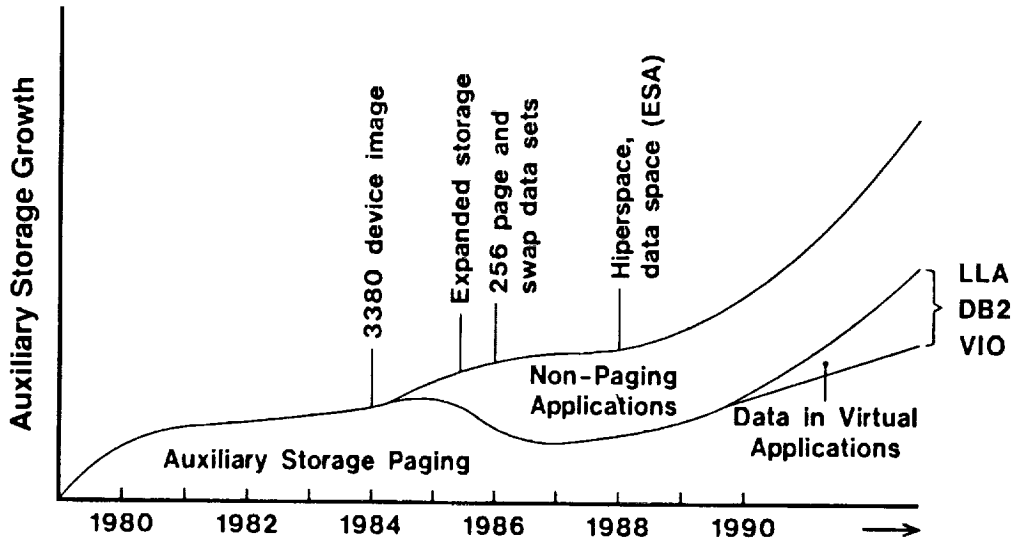
StorageTek
PROPRIETARY

The virtual storage hierarchy of the 1990s will be exploited by ESA/390 architectures and consist of three levels of storage. Central memory, at about $4,000 per megabyte, has an architectural limit of 4 gigabytes and an announced 1 gigabyte limit. Expanded storage has a 75-microsecond access time. The limitation in present ESA/390 architecture is 16 gigabytes of addressable expanded storage, though the announced limit is 8 gigabytes.

Solid-state products are now considerably less than $900 per megabyte and the cost per megabyte is declining quickly. In 1979 when the first SSD was introduced by Storage Technology Corp., the original price was $8,800 per megabyte. We have seen over a 90-percent reduction in pricing on solid-state technology in the 1980 decade. The virtual storage hierarchy contains three levels including auxiliary or paging storage. Careful use of all three technology levels offers the most cost-effective solution to managing the virtual storage hierarchy. It is normally not cost effective for most users to place all performance-critical data in expanded and central storage.

The ES/9000 processor series now permits migrated pages to move directly from expanded storage to the channel subsystem (auxiliary storage paging) improving the synergy and performance between both levels of the virtual storage hierarchy.

DIRECTION OF AUXILIARY STORAGE RAM BASED ARCHITECTURES



StorageTek
PROPRIETARY

By examining RAM-based solid-state products, you will notice their use from 1979 through 1984 was exclusively for paging. When expanded storage appeared in 1985, solid-state devices became viewed as a high-performance disk, and non-paging data such as load libraries, catalogs and indexes were placed on solid-state devices. The 1990s will see the MVS/ESA and VM/ESA hiperspace and data space applications drive up data in virtual requirements and force users to seriously consider using SSD as a cost-effective complement to real and expanded storage. ESA/390 drivers of virtual storage consumption, called methods of I/O avoidance by some, will include linear VSAM, Virtual Lookaside Facility (VLF), hiperspace catalog, hiperspace buffers and DB2. MVS/XA systems previously using 400 to 500 megabytes of auxillary storage will soon identify requirements exceeding 1 gigabyte or more after migrating to ESA.

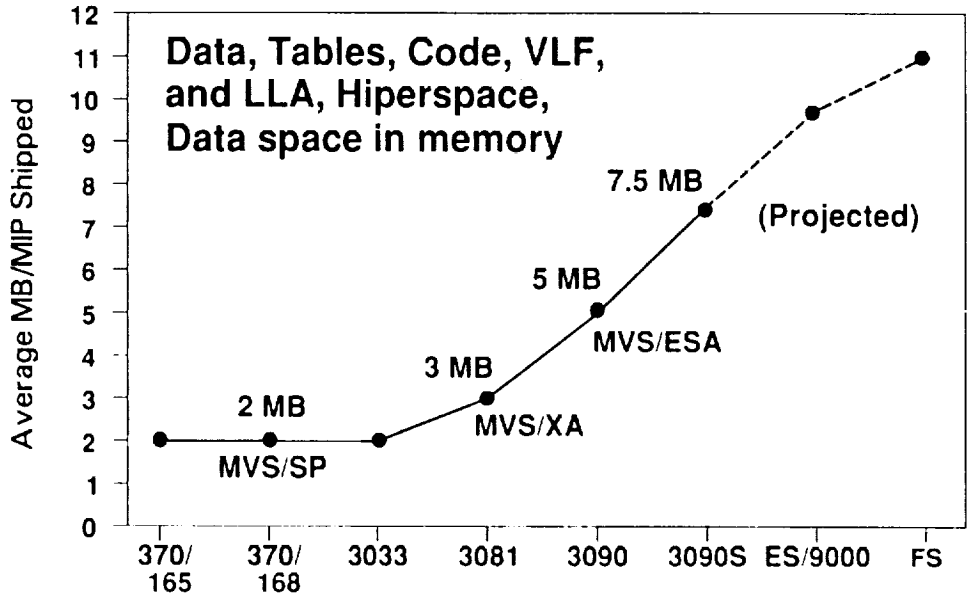
DRAM MEMORY DENSITY PROJECTIONS

Technology (in microns)	DRAM Technology	Development Start	Introduction Date	Peak Output
1.8	256 Kbit	1977	1982-83	1988
1.2	1 Mbit	1980	1985-86	1991
0.8	4 Mbit	1983	1988-89	1994
0.6	16 Mbit	1986	1991-92	1997
0.36	64 Mbit	1989	1994-95	2000
0.25	256 Mbit	1992	1997-98	2003
0.15	1 Gbit	1995	2001-01	2006

Unlike rotating DASD, RAM-based architectures command a very price-elastic market. If the price decreases, the demand increases. You cannot necessarily stimulate the demand for DASD or tape by changing the price. The industry-standard DRAM chip has moved from 1 megabit to 4 megabits. Note that the 4-megabit chip has been under development since 1983. As DRAM densities increase, price decreases along with the physical space required to store information. Thus much higher capacity DRAM storage devices will appear occupying smaller footprints. This trend should continue until the point where DRAM-based storage devices will occupy a large portion of the storage hierarchy currently belonging to rotating and cached DASD.

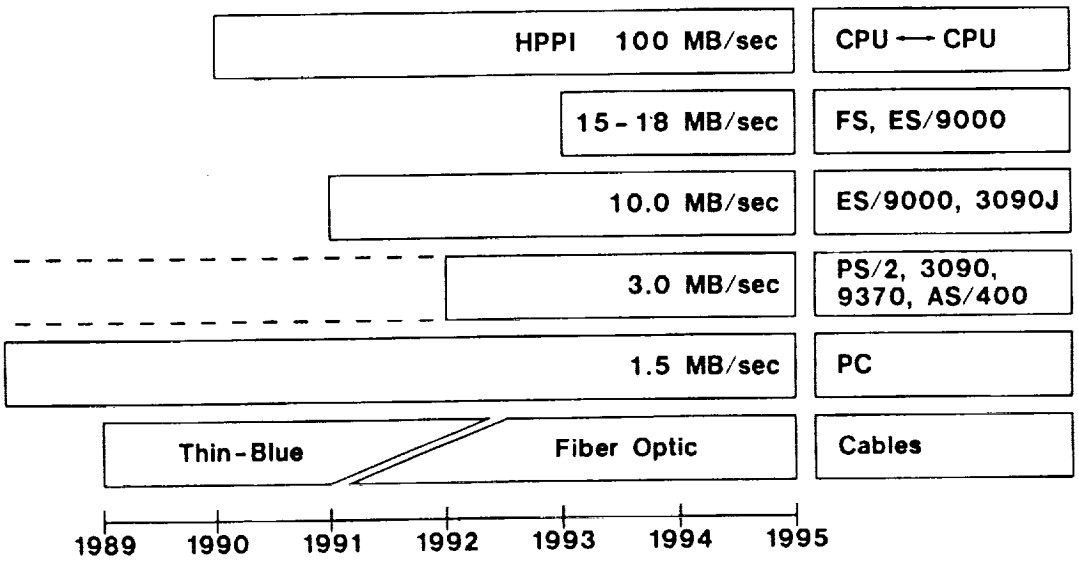
StorageTek
PROPRIETARY

MEMORY/MIPS SHIPPED



Historical tracking of the ratio of memory installed per MIPS installed indicates the trend increasing sharply with MVS/XA (31-bit addressing) and MVS/ESA, effectively 44-bit addressing. The MVS/ESA capability to place data, and program load libraries and other objects into memory, will continue to drive the ratio upward, requiring more and larger RAM-based storage solutions. The announcement of VM/ESA and DOS/VSE/ESA implementing hiperspace and data space concepts into these operating systems will further encourage virtual storage growth. The growth rate grows sharply until shared expanded storage arrives late in the Summit (i.e., future systems) then moderates slightly.

CHANNEL TRANSFER RATE GROWTH PROJECTION



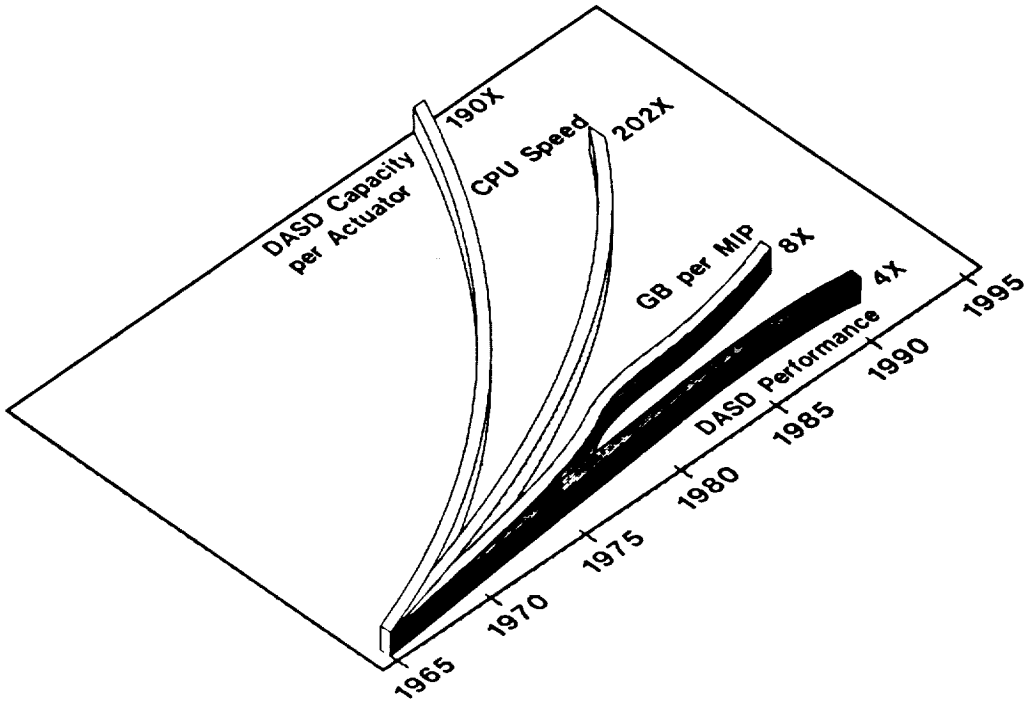
StorageTek
PROPRIETARY

The data transfer rate capabilities at all levels of computing will increase much faster in the first half of the 1990s than they did in the previous 15 years. ESCON channels now offer up to 10 MB/sec data transfer rates. Device exploitation of ESCON channels at 10 MB/sec will come more slowly. The 3990-3 DASD Storage Control will be the first device to exploit ESCON channels at ESCON speeds. Up to 18 MB/sec on the FS series is likely by 1995. The gray and thin blue cables will begin to give way to serial fiber channels providing increased distance (up to 9 km initially) in the early 1990s. The HPPI (High Performance Parallel Interface) channels used by large-scale CPUs for scientific processing will offer attachment of specialized storage devices in the mid-1990s, likely RAM-based, providing high-performance solid-state storage arrays.

Storage Products

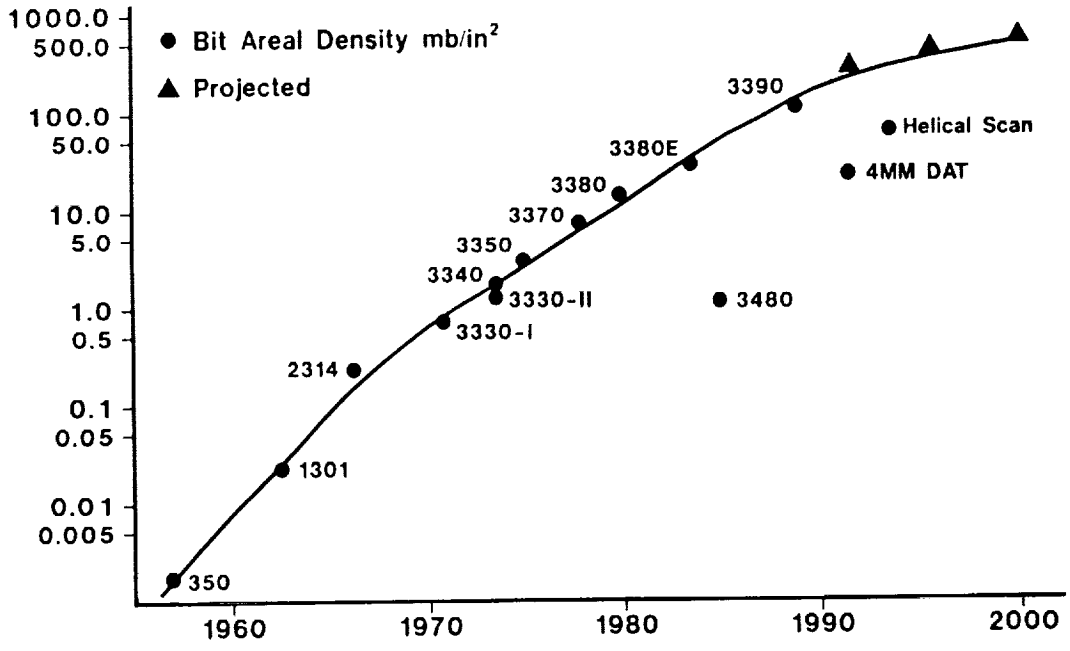
StorageTek
PROPRIETARY

LARGE SYSTEMS TRENDS



In the last 25 years, the processing power of large computers has increased at a much faster rate than the performance capabilities of the I/O subsystem. Since the introduction of System/360 in 1965, we have seen the capacity of a disk actuator increase 190 times. Processor performance has increased over 200 times, but the performance of a disk actuator has improved only 4 times. This divergence of processor speed and the I/O subsystem performance has been the subject of considerable interest, particularly in the 1980s. During this time, we have seen the introduction of a number of technology developments to help bridge the gap. These enhancements include solid-state disk, cached control units, dual port, quad port, actuator level buffers, tape buffers and expanded storage. Despite these advances, the performance gap between processors and I/O subsystems continues to diverge. More solutions will emerge, predominately based on DRAM technology, to place data closer to the processor and remove the performance delays of mechanical devices. The ES/9000 processor announcement by IBM is a good example of this continually diverging trend — processor speed (MIPS) nearly doubled while the speed of the storage subsystem remained unchanged.

TRENDS - AREAL DENSITY



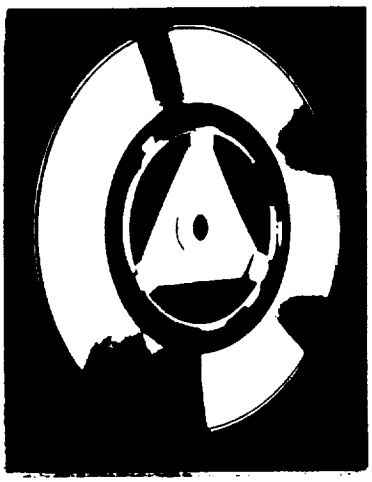
StorageTek
PROPRIETARY

Magnetic recording technology has seen few of the anticipated limitations in the areal density (megabits per square inch) that were expected to occur in the early 1980s. Areal density increases in DASD, once not expected to exceed 30 megabits per square inch, are now over 60 megabits per square inch in 3390-type disk drives and are expected to increase to over 500 megabits per square inch by the end of the 1990s. Laboratory developments of 1 gigabit per square inch have been demonstrated. As the areal density of magnetic recording continues to increase rapidly, the future role of optical disks in the large systems storage hierarchy becomes more questionable.

The helical scan tape format, recording data tracks on a tranverse rather than parallel to the edge of the tape, offers areal densities in the range of 30-50 million bits per square inch and very high data transfer rates. Helical scan technology should enter the hierarchy at the deep archive level and co-exist with 3480 chromium dioxide (Cro2) tape format throughout the remainder of the 1990 decade.

StorageTek PROPRIETARY Storage Products

OPTICAL



WHEN

StorageTek PROPRIETARY

MAGNETIC/OPTICAL STORAGE
5 1/4" OPTICAL PARAMETRICS vs DASD

OPTICAL HDA	1990	1995	2000
Capacity (GB)	1.0	1.5	4.0
Transfer rate (MB/sec)	0.7	1.2	2.5
Cost ($/MB) (OEM)	3-5	1.30	0.50
DASD HDA			
Capacity (GB)	1.5	2.5	5.0
Transfer rate (MB/sec)	1.9	4.0	6.0
Cost ($/MB) (OEM)	2.0	0.60	0.20

9.90 FORUM9022DJ08

Optical disk storage has been hailed as the low-cost mass-storage technology of choice for years. In reality, optical storage has yet to fulfill expectations. Issues such as standardization, throughput, data transfer rates and uncertainty of shelf life (data retention) remain. The areal density of magnetic disk storage is increasing rapidly, while optical disk areal density has remained relatively the same for the past six years. The write-once, read many times (WORM) drives are common in the midrange and desktop markets, but will struggle to find a mainframe niche. Magneto-optic or erasable optical disks offer the large systems market the most benefit, but may face a stiff challenge from large-capacity DASD array storage solutions for the online, large-capacity storage market and advanced automatic cartridge systems for even larger capacity and less costly deep archive storage. Unlike WORM optical, 5.25" magneto optical has the support of formal standards by the International Standards Organization.

TAPE LIBRARIES
MAINSTREAM APPLICATIONS

CURRENT APPLICATIONS

- Tape Management
- DASD Management
  - SMS
  - DF/HSM, DMS/OS
  - DASD savings
- Job scheduling and rerun
  - Improved batch performance
- Software development
- Automated recovery
  - Online data bases
  - Mission-critical data
- Report management/paper/fiche
- Electronic archive
  - Campus
  - Remote vault
- Automated operations
  - Unattended → Lights out
- High speed search applications

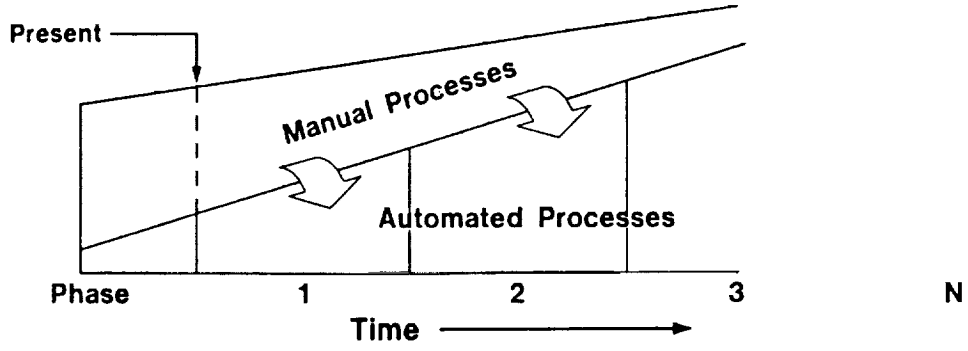
ADVANCED APPLICATIONS

- Anticipatory staging
  - Random access data
  - Data set scheduling
- Network storage
- Deep archive with helical formats
- Scientific data

StorageTek

Automation has opened many new horizons for data storage. Far beyond simply automating what currently exists on tape today, automation has become a primary ingredient in cost-effective storage management and is enabling the promise of systems managed storage to be fulfilled. Several new applications listed above have become areas of opportunity providing cost savings and improved quality of operations beyond prior capabilities. The concept of electronic archiving has been given an increased focus with the announcement of ESCON channel architecture complementing traditional channel extension methods. A form of image storage, report management, has been enhanced with a number of software products that allow computer output microfiche and printed data to be stored on a tape library, viewed at a computer terminal, and printed or sent to fiche only if needed. This new area of library exploitation greatly reduces distribution, copy and filing costs while improving the security aspects associated with printed storage. Applications with much promise for automated tape libraries include anticipatory staging of data and deep archive storage for long term-data storage.

AUTOMATED OPERATIONS



Console Message	Operator	Suppress Auto Reply	Open Architecture
Recovery	Manual	Auto Restart	D/B Recovery Appl. Recovery
SMS	Storage Pools	Space Mgmt. DF/SMS	Performance Management
TMS	Physical Placement	Scratch Retention	ACS 4400
Output Distribution	Burster Trimmer	Roll Feed Bar Code	Online Viewing Selective Print

StorageTek PROPRIETARY

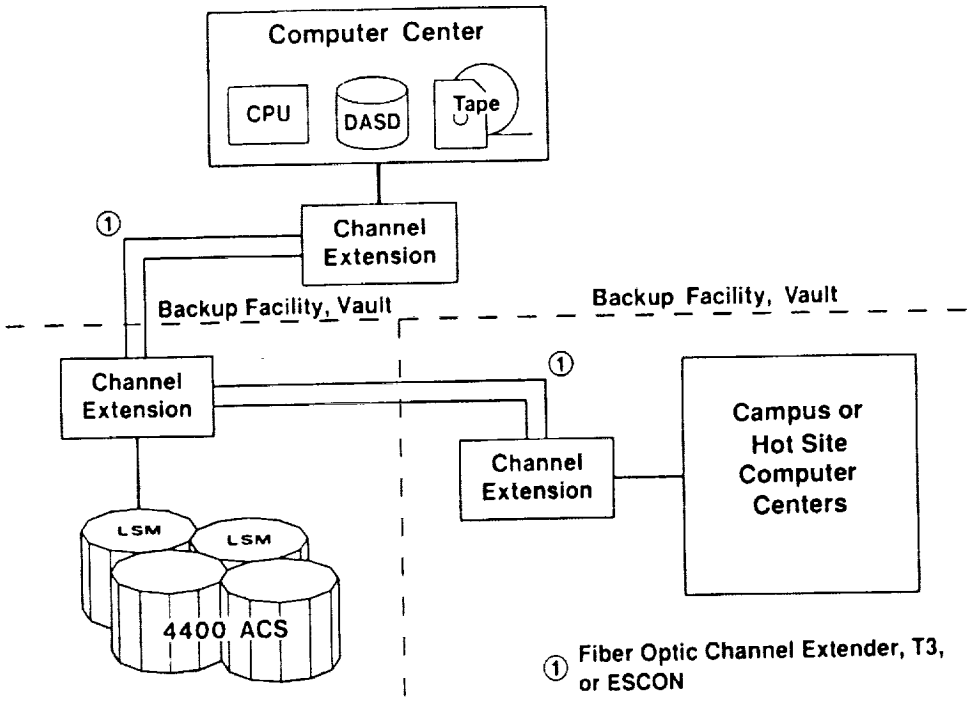
As automated operations becomes a strategic goal for many data processing users, solutions are appearing which are making companies more competitive, more productive and more profitable. Automated operations is usually fully implemented in stages and will evolve to include expert systems solutions to resolve some of the complex, enterprisewide information management issues. The primary reason for automated operations is improved quality of the data processing organization.

In addition to automated operations, business resumption or disaster recovery planning has become a strategic goal for many users. Workshops on these areas of advanced data center operations are available on a worldwide basis from Storage Technology Corp.

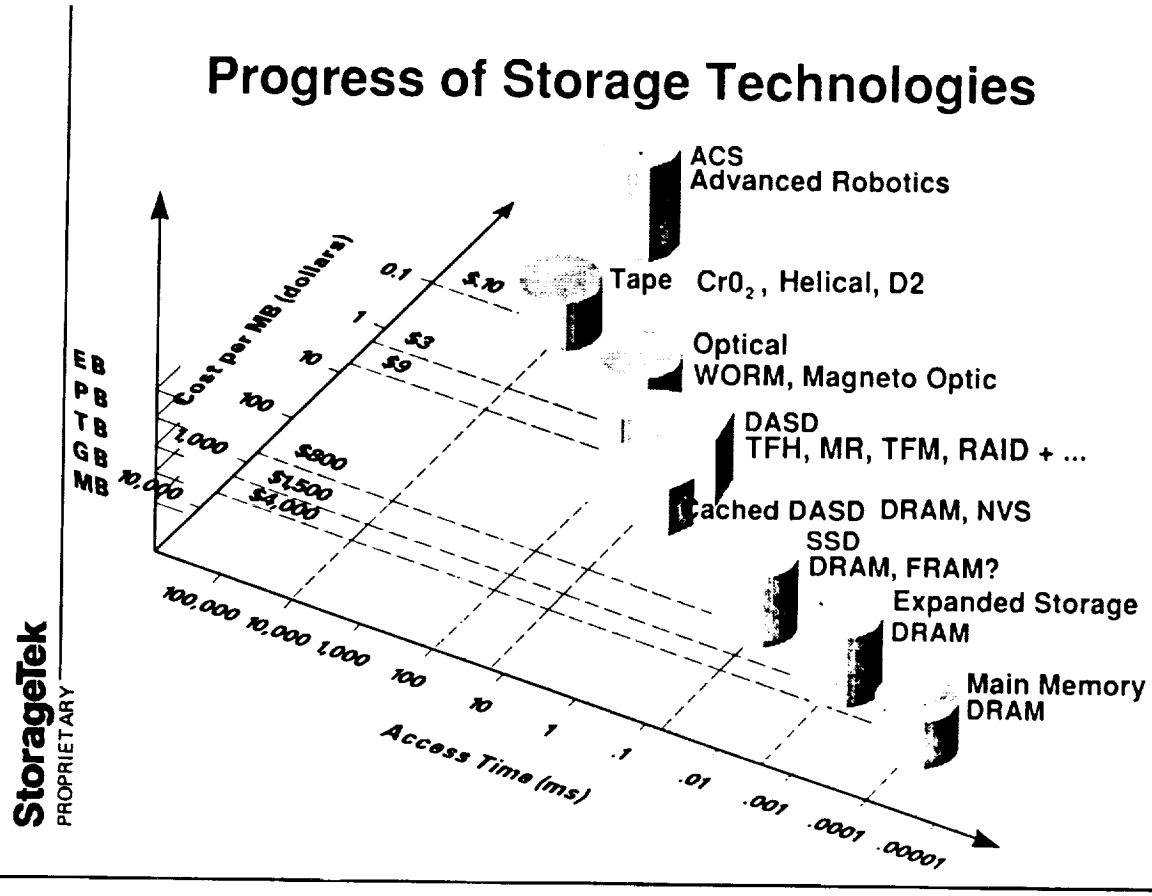
Storage Management

StorageTek
PROPRIETARY

ELECTRONIC ARCHIVING



The rapid acceptance of library storage products such as StorageTek's 4400 Automated Cartridge System (ACS) has provided a means for computer users to archive data electronically to a secure, remote location such as a vault or warehouse. Today, the use of fiber optic channel extensions provides 3 megabyte per second device attachment. Products such as the 4400 ACS can be located at distances well beyond the four walls of the computer center by using fiber optic channel extenders, T3, or ESCON communication lines. The "data vault" provides backup of critical data in a safe location and also can link into a hot site or campus computer facility for a quick recovery in case of a disaster. This trend will expand in the 1990s as the value of information to the corporate enterprise becomes increasingly more important.



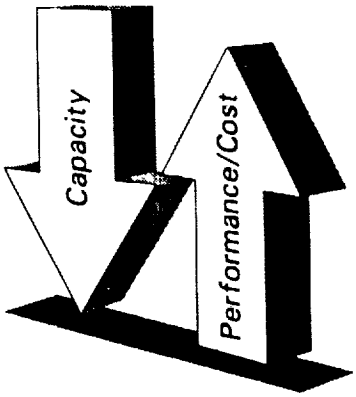
Let's take a look also at a few of the technologies that have merged to help resolve some of the challenges in the 1990s. The 1980s was the the decade of technology. The 1990s will be the decade of how we exploit that technology effectively. In the 1980 timeframe, we clearly remember vendor and customer discussions regarding DASD. Issues centered on such things as the diameter of the disk platter. How thick is the platter lubrication? How high does the read/write head fly? The answers to those questions sometimes influenced buying decisions. Today in DASD acquisitions the issues are gigabytes per square foot; I/Os per second; availability; cost per gigabyte. The size of the platter really doesn't have to make a difference, but gigabytes per square foot should.

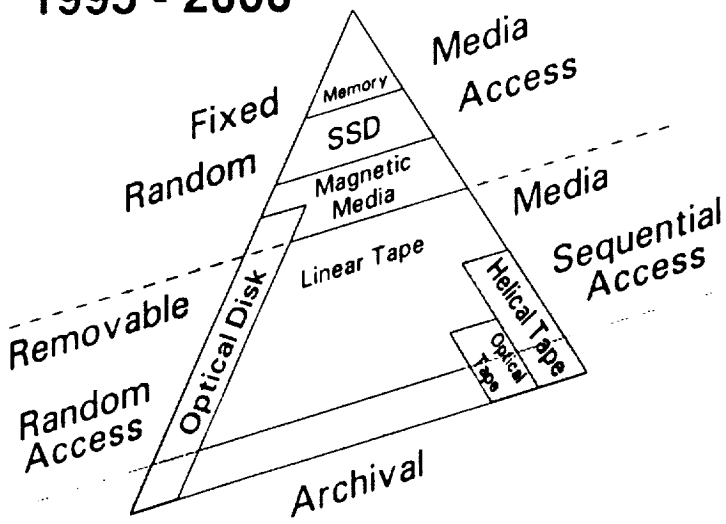
Notice that main memory presently is priced at about $4,000 per megabyte with an access time measured in nanoseconds. Expanded storage has a 75 microsecond access time and is priced at about $1,500 per megabyte. There remains a major performance and price gap between the memory technologies (DRAM) and moving or rotating technologies.

Optical storage and tape provide two interesting comparisons. The areal density of optical storage has witnessed insignificant improvement in the last six years. During this time, magnetic storage has significantly increased in areal density. Interestingly, the IBM Image Plus system, using a write once optical storage device, is priced around $2 to $3 per megabyte. Tape libraries, including compression/ compaction, may realize costs as low as 10 cents per megabyte.

New technologies will evolve to fill this gap such as ferro-electric RAM (FRAM) devices. These are non-volatile RAMs and still under development though expected to be affordable and commercially available in the 1994 time frame.

STORAGE HIERARCHY
1995 - 2000

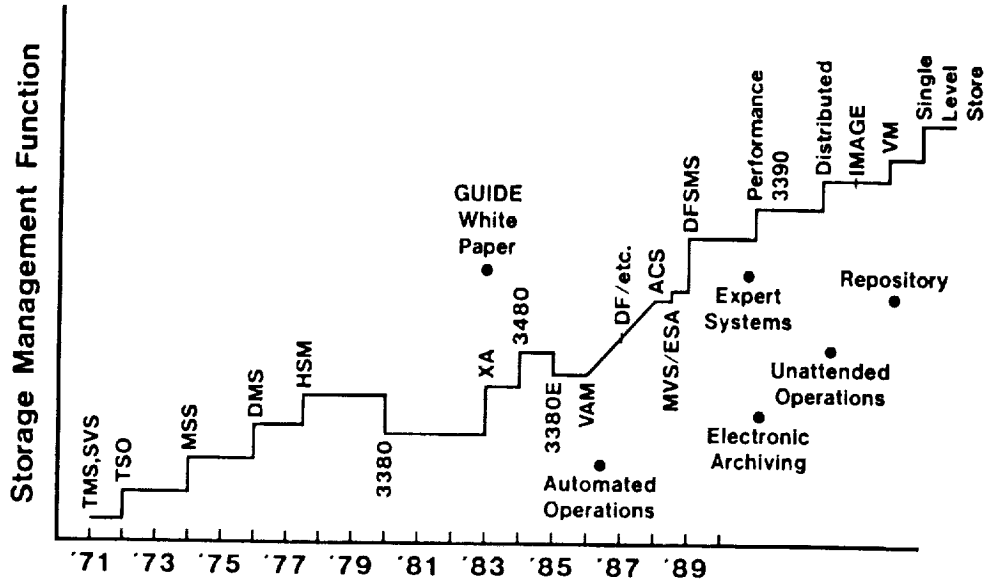




StorageTek
PROPRIETARY

The hierarchy of storage technologies, not devices, in the last half of the 1990s is shown. This hierarchy may be broken down into fixed-media and removable-media segments. Fixed-media storage will consist of RAM and rotating DASD. Optical disk, once viewed as the heir apparent to the removable-media segment, has given way to automated tape systems that are faster and less costly. Further advances in the capacity of 3480 cartridge capacity to 1 gigabyte levels and the increased usage of helical scan formats make magnetic tape (along with magnetic disk) the key technologies of the 1990s.

TRENDS - STORAGE MANAGEMENT EVOLUTION

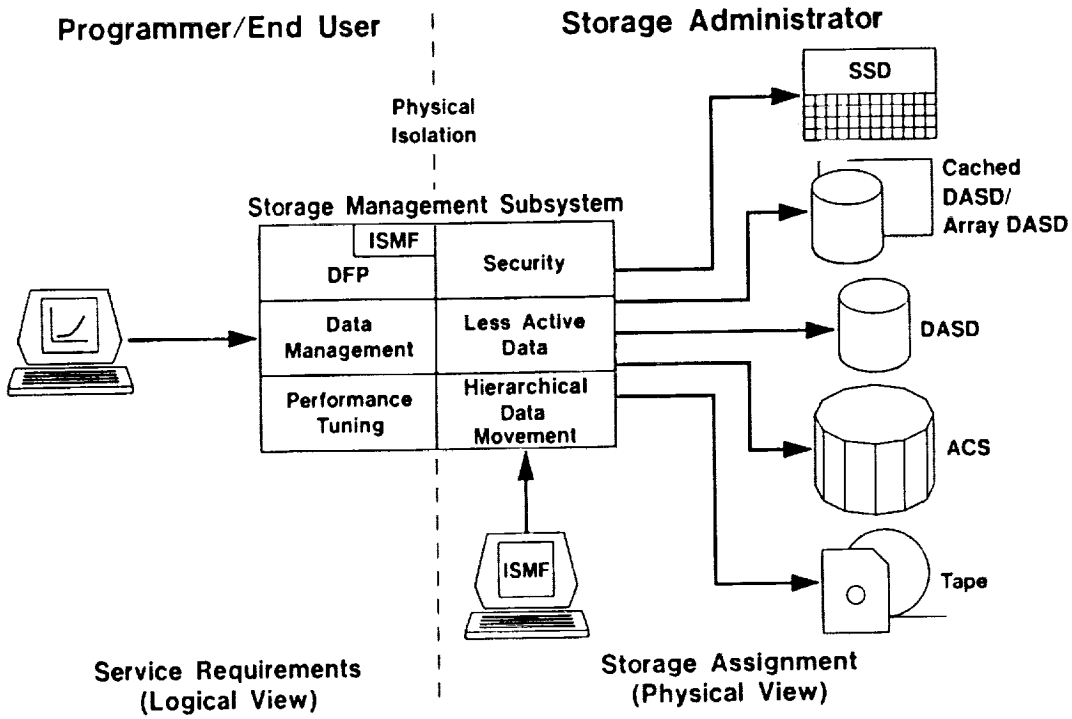


StorageTek
PROPRIETARY

The total cost of storage devices now accounts for more than one half of the total hardware expenses of the typical large data center today. In the early 1970s, storage management meant tape management. With the introduction of the DMS/OS and HSM storage management products in the late 1970s, storage management expanded its scope to include space management for disks. Since that time, storage management had been relegated to improving various facets of space management until the announcement of DFSMS in February 1988. This platform should gradually evolve to include dynamic performance tuning, storage management for distributed processing nodes, networks, workstations, a DFSMS equivalent for VM and development of a repository to identify objects across all computing platforms in the corporate enterprise.

DFSMS is a trademark of IBM Corporation

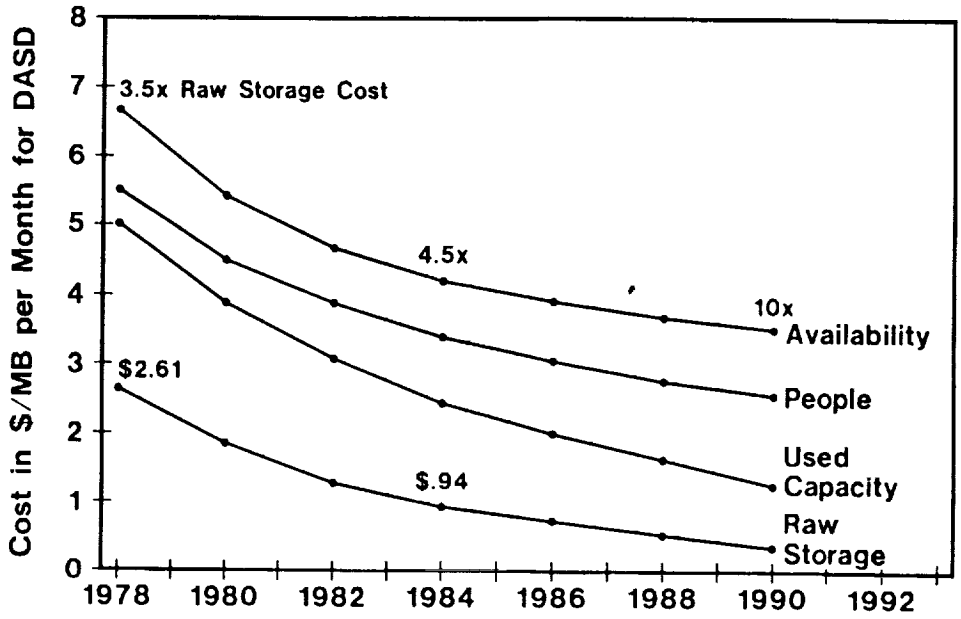
SYSTEMS MANAGED STORAGE



StorageTek PROPRIETARY

Systems managed storage is a concept that allows an operating system to perform many of the human-intensive processes involved with space management, performance tuning, availability management and true hierarchical storage management supporting all tiers of storage. Though in its infancy, systems managed storage must efficiently evolve these processes to provide the platform for single-level storage in the last half of the 1990s and achieve an environment that allows "true systems managed storage." DFSMS is one of several products that make up systems managed storage. Presently the DFSMS product provides no performance tuning capabilities or movement of data vertically throughout the storage hierarchy to optimize performance or space management.

COST OF USABLE, MANAGED STORAGE

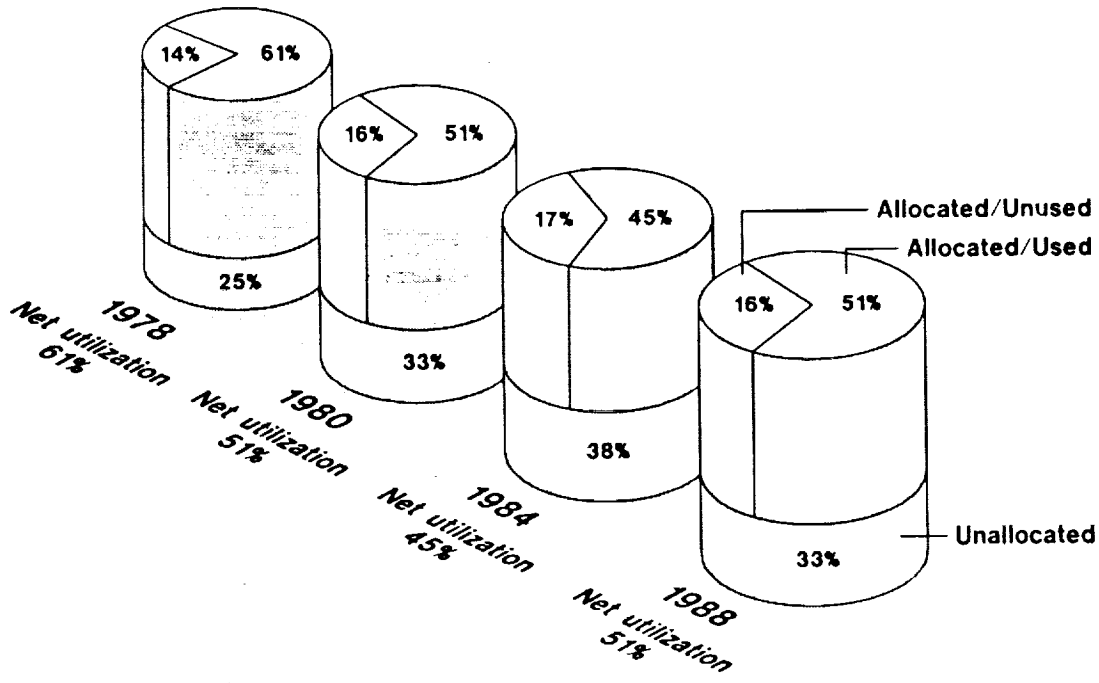


Source: IBM Systems Journal, 1989

StorageTek
PROPRIETARY

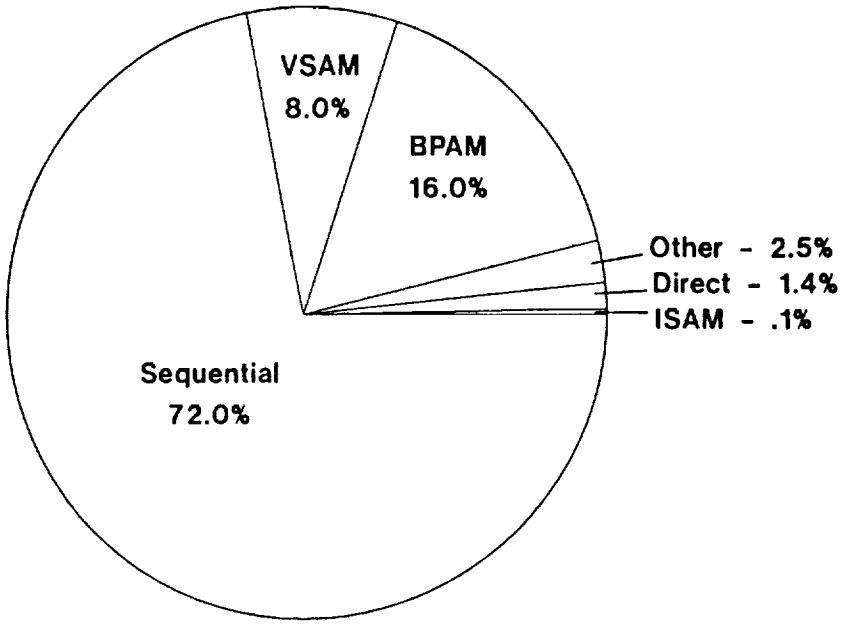
The total cost associated with managing storage has declined since 1978 on a per-megabyte basis. The cost of raw (live) data, plus the costs of unused capacity, people costs to manage the storage system and additional costs of availability such as backing up data, have increased several times since 1978 on a per-megabyte basis. The cost of managing DASD, effectively or ineffectively, has become a primary concern in very large (terabyte plus) data centers and clearly is necessitating the movement toward much improved storage management facilities. It is estimated that the total cost of managing disk storage in the 1990s will be as much as 10 times greater than the cost of actual data stored on DASD. New storage solutions, such as advanced fault tolerant DASD architectures, will dramatically improve these trends.

DASD USAGE TRENDS



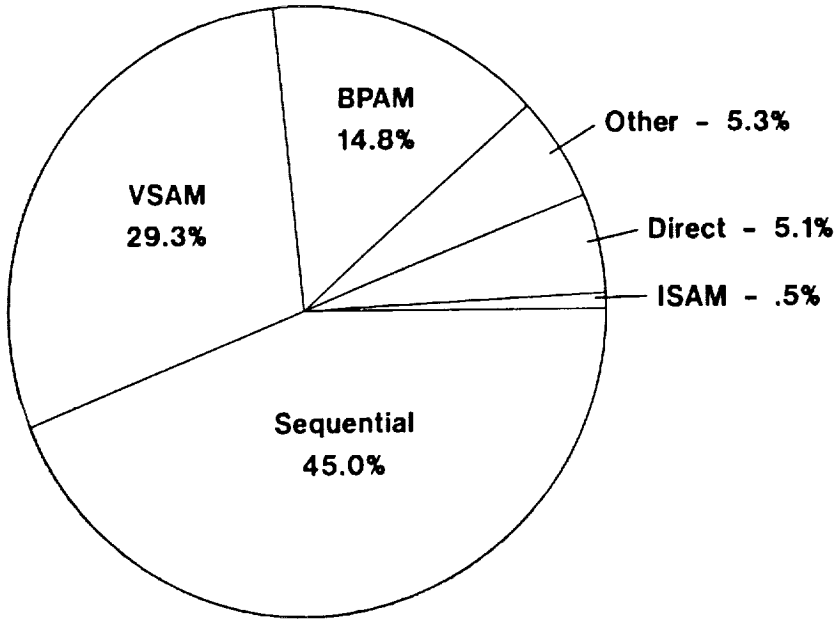
Formal surveys on DASD usage have been conducted since 1978. The net utilization, or the amount of real data on the average device decreased from 61 percent in the 1978 survey to a low of 45 percent in 1984. The 1988 survey indicated an overall increase in net utilization to 51 percent. This survey included single-, double- and triple-capacity 3380-type devices. The single- and double-capacity devices actually increased in utilization; however, the triple-capacity devices continued the downward trend. Utilization of the single- and double-capacity devices increased largely due to the higher percentage of caching permitting increased space allocation on cached DASD. Space utilization figures for 3390-class devices are not presently available though the lack of widespread DFSMS usage to optimize data allocation on 3390-type DASD may initially inhibit more effective utilization.

PERCENT OF DATASETS BY DATASET ORGANIZATION



As a follow-up to the previous chart on DASD space allocation, the percent of DASD data sets by data set organization reveals a correspondingly high percentage of sequential data sets. VSAM and SAM-E (sequential) data set organizations are strategic while partitioned data sets will fold into VSAM format as ESA/390 evolves. Other data sets such as graphics access methods, direct access files and even ISAM will exist as they are today. This profile again reflects the results of extensive tape-to-disk migration activities in the 1980s.

PERCENT OF SPACE BY DATASET ORGANIZATION

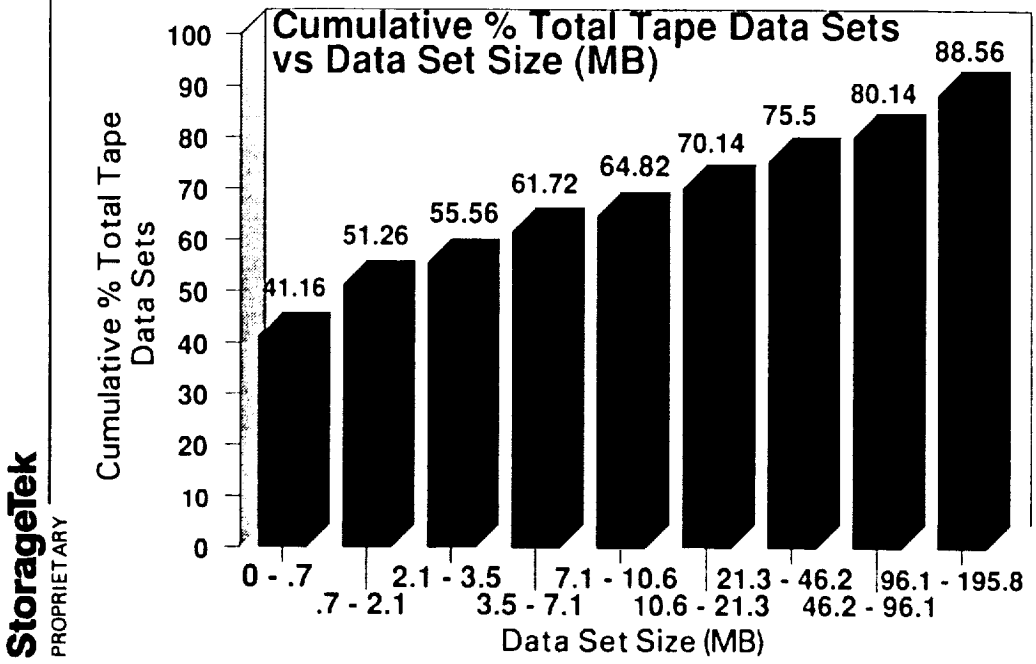


Storage management has increasingly focused on DASD in the last few years, primarily due to the criticality of data residing on DASD. At the end of 1988, sequential data had grown to include 45 percent of allocated disk space. This was primarily a result of the large number of tape-data-set-to-disk-conversions in the early 1980s, occurring for the lack of any successful automated tape library system available to mainframe users. Tape data sets requiring rather quick or frequent access could not often withstand erratic human tape mount times.

This DASD profile, as a result of the 1980s strategy of "put it all on disk," has become a significant cost savings target for automated cartridge systems in the 1990s.

TAPE COMPRESSION/COMPACTION

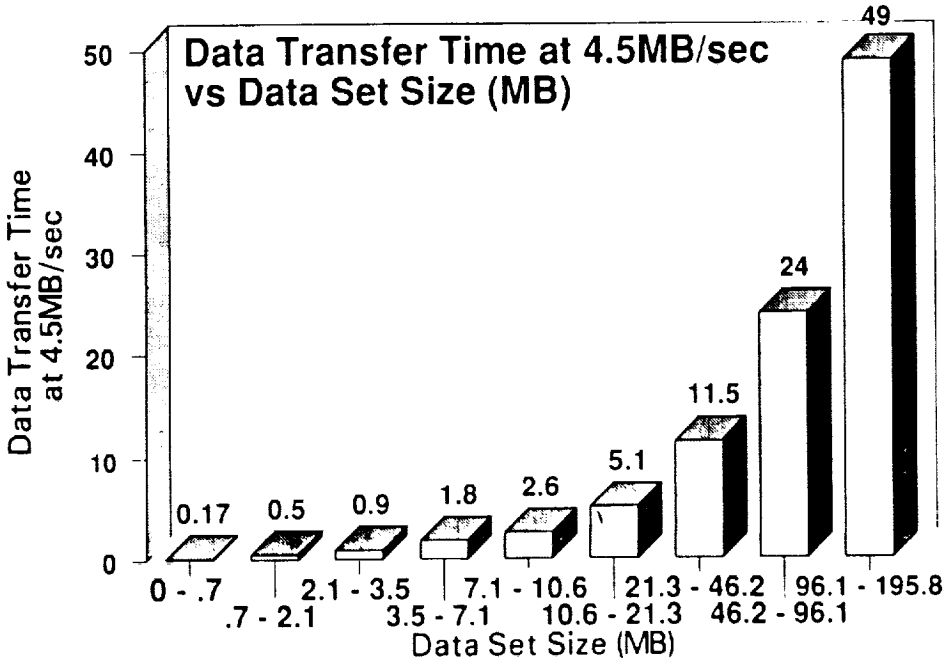
TAPE DATA SET PROFILES



StorageTek
PROPRIETARY

A survey of tape data set sizes indicates that the vast majority of tape data sets are very small. 70.14 percent of the tape data sets in the survey are under 21.3 megabytes in size. Even with the tape-to-disk conversions of the 1980s, many small tape data sets remained on tape and became obvious targets for automation. While 11.44 percent of the tape data sets occupy over one 200MB 3480 tape cartridge, the majority of these data sets are backup applications using processor data-compression functions.

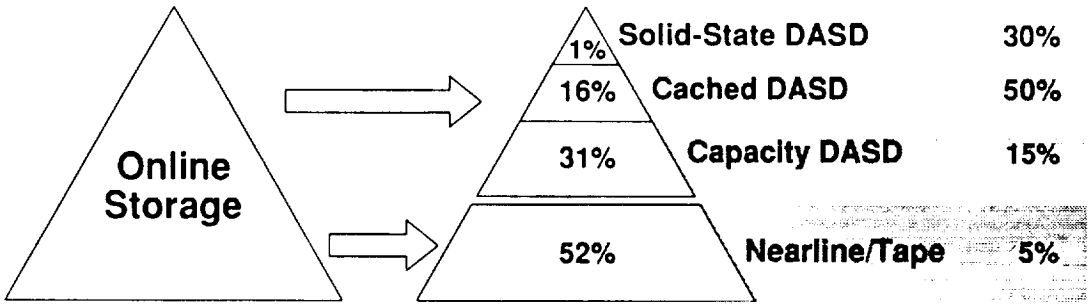
TAPE COMPRESSION/COMPACTION
TAPE DATA SET PROFILES



StorageTek
PROPRIETARY

The tape data set size survey is equated to data transfer time for each classification of data set on a 4.5 MB/sec channel using bus and tag or even ESCON. The 21.3 MB tape data set takes approximately 5.1 seconds to process once mounted. Assuming a 2-to-1 data-compression factor, a total time of 2.55 seconds would be used to transfer the 21.3 MB data set. This minimal gain from compression/compaction is often lost by the variability in manual tape mounts resulting in minimal, if any, throughput benefit for 70 percent of the tape data sets. By the early 1990s, nearly everything written to tape will be compressed. Like 7-track, 9-track, NRZI and GCR in the past, tape data compression will be another format for tape data recording.

COST-EFFECTIVE STORAGE PROFILES



1980s

- Technology decisions
- Get the right data in the right place

1990s

- Business decisions
- Get the right data in the right place at the right time

StorageTek PROPRIETARY

Using advanced data storage placement methodologies such as CPIO (Cost Per I/O) analysis, it is possible to determine the most cost-effective location in the storage hierarchy. Using the size, performance and storage cost per I/O to determine optimal data set placement, studies indicate that online storage users today may not be cost effectively utilizing online storage. Typically, 1 percent of the space allocated in online storage generates 30 percent of the total online IOs. This small but highly active group of data sets is most cost effectively located on solid-state disk. At the other end of the spectrum, a little over half of the data allocated on DASD today is more cost effective on automated cartridge systems or manual tape. Storage management in the 1980s stressed getting the right data in the right place; storage management for the 1990s will stress getting the right data in the right place at the right time.

CPIO is a proprietary software tool from Storage Technology Corp.

notes

52-82

121939

MAGNETIC DISK

N93-15029

John C. Mallinson
MALLINSON MAGNETICS, INC.
7618 Reposada Drive
Carlsbad, CA 92009

Magnetic disk recording was invented in 1953 and has undergone intensive development ever since. As a result of this 38 years of development, the cost per byte and the areal density has halved and doubled respectively every 2-2 1/2 years. Today, the cost per byte is lower than 10^{-6} dollars per byte and area densities exceed 100 10^{+6} bits per square inch.

In this talk, the recent achievements in magnetic disk recording will first be surveyed briefly. Then the principal areas of current technical development will be outlined. Finally, some comments will be made about the future of magnetic disk recording.

PRESENT ACHIEVEMENTS

High end disk drives today operate at areal densities of between 50 and 100 10^{+6} bits per square inch, with, typically, 2500 tracks per inch and 30,000 flux reversals per inch. When "run length limited" coding is used, the effective linear bit density is higher; for example, with 1,7 code, the linear bit density is 40,000 bits per inch. Area densities tend to be higher the smaller the diameter of the disk.

Data rates run as high as 6 Megabytes per second (48 Megabits per second) per single head-disk channel. Parallel access disk systems, with as many as 10 heads in parallel have been manufactured which provide the full CTIR 4:2:2 component digital video output rate (216 Megabits per second).

Since as many as 6 disks can be fitted in the standard 5 1/4"

full height form factor package, 5 1/4" drive data capacities exceeding 2 Gigabytes are now available from several manufacturers.

In summary, it may be said that the magnetic disk products being manufactured today offer access times, data rates and drive bit capacities considerably in excess of those offered by optical disk drives. Areal density is the only parameter which currently falls below that of optical disks, by a factor of 3-4.

AREAS OF TECHNICAL DEVELOPMENT

The overwhelming success of magnetic disk products over the last three or four decades has led to the establishment of a $50 billion per year world wide business in disk drives. This enormous business supports research and development into every conceivable aspect of disk recording technology in order to permit continuing increases in performance. Only the major areas of such research and development can be discussed below.

IMPROVED RECORDING MEDIA

Virtually all modern disk drives now use thin film metallic media with coercivities close to 1000 Oe. It may be expected that coercivities exceeding 2000 Oe will be used in the next few years. Higher coercivities lead to both sharper output pulses of greater amplitude and also to improved signal-to-noise ratios.

IMPROVED WRITING HEADS

As the medium coercivity increase, it is necessary to increase the saturation induction of the writing head pole tip materials. Presently, Alfesil and Permalloy with maximum inductions of 10-12,000 G are used. Materials such as Co-Ru and Fe-N with maximum inductions of 16,000 and 19,000 G may be expected to be introduced.

NARROWER TRACKWIDTHS

It has been realized for two decades that, when seeking higher areal densities, it is better to use narrower trackwidths than higher linear densities. Operation with trackwidths substantially narrower than normal (10^{-4} m) leads to a number of very fundamental questions concerning the operation of the track following servo system. In particular, the outstanding question is "what is the source of the tracking error signal?" In magnetic disks today the source is a previously written magnetic servo track and it is only possible to operate the tracking servo when reading but not when writing. In optical disks, which operate at 5-6 times the track density, the source is always some physical feature (pits, grooves, bumps, etc.) and the tracking servo system is then operable during both reading and writing. This leads to another question: "Will magnetic disks eventually use optical tracking servo systems?"

IMPROVED READING HEADS

As trackwidths decrease, it becomes increasingly difficult to keep the channel signal-to-noise ratio media noise limited because the output voltage of an inductive head falls proportionally with the trackwidth. It is anticipated that the next generation of high end disk drives will use magneto-resistive (m-r) reading heads where the magnetic fields from the medium changes the electrical resistance of a thin film m-r element. Considerably higher output voltages are available with m-r heads and they are independent of head-medium relative velocity.

IN CONTACT OPERATION

Today's disk drives operate with a deliberate head-to-disk

spacing of, typically, 6-8 microinches (0.15μ m). It is known that both the writing and reading processes on magnetic disks improve when the spacing is reduced. All disks today are overcoated (Ag-Sn, amorphous C, ZiO_2 , etc.) in order to control friction and wear and it seems very likely that, together with redesigned heads of significantly lower mass, continuous operation in contact may become possible. This is particularly true at low head-to-disk relative velocities.

SMALLER DISK DIAMETERS

An interesting sequence of design changes becomes possible following a reduction in the head-to-disk spacing. First, a higher linear density may be written. Second, because the data rate has now become too high, the disk diameter or spindle RPM must be reduced. Third, at the reduced head-to-disk velocity, it now becomes possible to reduce the head-to-disk spacing even further because any mechanical impact now transfers less energy. Fourth, if a smaller disk diameter has been chosen, the mechanical tolerances (flatness, areal runout, etc.) are reduced which again permits the head-to-disk spacing to be reduced even further. This sequence has led the drive industry from 5 1/4" to 3 1/2" to 2 1/2" to 1 1/2" diameters with increasing areal density. Still smaller diameters and higher areal densities are anticipated.

As an example of the levels of performance attainable when many of these developments are combined, consider the 1989 IBM 1.1 Gigabit (1100 Megabit) per square inch technology demonstration:

Medium coercivity - Cobalt-Platinum - 1700 Oe

Write Head-thin film-trackwidth 4μ m

Read Head - magneto-resistive - trackwidth 2-3μ m

Head-to-disk spacing - about 1 microinch

Linear density - about 160,000 bpi

Track density - about 7,000 tpi

With this demonstration, IBM showed that magnetic disk recording has the potential to exceed today's optical disk areal densities by about a factor of 2.

THE FUTURE

The IBM 1989 demonstration proved 1.1 Gigabit per square inch feasibility. Today's research papers (see, for example Intermag '91 paper MA-01) discuss demonstrations of 2 Gigabit per square inch (at 17,000 tpi and 120,000 frpi). It seems to be abundantly clear that the magnetic disk technology exists today which will take magnetic disk recorders from the 50-100 Megabit per square inch of today's manufactured hardware to future products with areal densities perhaps as high as 16 times greater.

It used to be said that the great advantage of optical (versus magnetic) recording was that it was not necessary to fabricate anything with dimensions comparable to the wavelength of light in order to achieve very high areal densities because the lens could focus the light down to Lord Rayleigh's diffraction limit.

Nowadays, it seems that a very fundamental change in philosophy has occurred. Indeed, it is frequently stated that the real advantage of magnetic versus optical recording lies in the fact that the only effective limits operating today concern just how small can certain features and objects be made and that their dimensions are not limited by mere physical diffraction of light!

For example, the gap-length in mass-produced 8 mm VCR heads is 10 microinches, which is but one third the wavelength of red light.

The steady increase in areal density, by a factor of 2 every 2-2 1/2 years, has been mentioned already. By this criterion alone, it appears then that magnetic disk recording technology can sustain another 20 years of growth (a factor of $16 = 2^4$; 4 x 2.5 = 20 years) on the basis of demonstrables which exist in the laboratories today.

To move from scientific extrapolation to the realm of technical speculation, it seems to be very likely that 1 Gigabit (10^{+9}) per square inch areal densities will appear in disk (and video tape) drives in considerably less than 20 years. Indeed some industry observers have opined that 5 1/4" full height drives with 100 Giga-byte capacity will appear before the year 2000: this represents a doubling of the historic rate of increase. Given the magnitude of the research and development activities in magnetic disk recording being undertaken worldwide, even such surprising estimates do not appear, to this writer, to be unduly optimistic!

File Servers, Networking, and Supercomputers

Reagan W. Moore

San Diego Supercomputer Center
San Diego, California

Abstract

One of the major tasks of a supercomputer center is managing the massive amount of data generated by application codes. A data flow analysis of the San Diego Supercomputer Center is presented that illustrates the hierarchical data buffering/caching capacity requirements and the associated I/O throughput requirements needed to sustain file service and archival storage. Usage paradigms are examined for both tightly-coupled and loosely-coupled file servers linked to the supercomputer by high-speed networks.

Introduction

The file server capacity requirements are most strongly driven by the CPU power of the central computing engine. The workload that can be sustained on the supercomputer is ultimately limited by the ability to handle the resulting I/O. At the San Diego Supercomputer Center, the central computing resource is a CRAY Y-MP8/864 supercomputer with a peak execution rate of 2.67 Gflops, capable of generating up to

2.67E+9 operations/sec * 8 Bytes/operation * 86,400 sec/day

or 1.8 Petabytes per day. In practice, the actual data generation rate is determined by the workload characteristics. The two major sources of I/O are application disk I/O and job swapping to support interactive use. At SDSC, the batch load averages 8 GBytes of executable job images, while the interactive load peaks at 120 simultaneous users. The batch load is sufficiently large that the idle time on the supercomputer has averaged 1.5% of the wall clock time over the last 6 months. While maximizing CPU utilization has been an explicit goal at SDSC, this has also increased the total amount of data that must be manipulated.

Data Flow Analysis

Not all generated data are archived and not all archived data are saved forever. A data flow analysis is necessary to understand the characteristics of the I/O, including the amount of data actually generated, the length of time over which the data are accessed, and the rate at which the data are moved through multiple caching levels. A simple analysis of the data flow can be used to illustrate the results of changing data access methods, increasing processing power, or improving network bandwidth. In particular, the data flow patterns are expected to be different for loosely-coupled user-initiated file archiving than for automated file servers tightly-coupled to the supercomputer CPU.

Solid State Storage Device Cache

The current archival storage system in use at SDSC is DataTree which supports user-initiated file archiving. This system acts as the archival storage file server for the CRAY supercomputer and is accessed through a 100 Mbits/sec FDDI ring. Data generated on the CRAY Y-MP8/864 are ultimately stored on 3480 cartridge shelf tape. There are five levels of I/O buffering or caching, including a 1 GByte Solid State Storage Device, 42 GBytes of CRAY disk local to the supercomputer, 70 GBytes of archive disks, a 1.2 TByte tape robot, and 2 TBytes of manually mounted shelf tape. Table 1 illustrates the caching hierarchy. As expected, the amount of data moved towards the lowest archival storage level decreases as the required storage life of the data increases. Data resides on the SSD for periods on the order of minutes, on CRAY disk for up to two days, on archival storage disk for up to several weeks, in the tape robot for several months, and finally on shelf tape for years. The amount of data moved per day between each level varies from 1.5 TBytes/day through the SSD to CRAY disk, 14 GBytes/day through archival storage, 9 GBytes/day through the tape robot, and 2 GBytes/day to shelf tape. The residency time at any level may be estimated by dividing the size of the cache by the input I/O rate to the cache. This closely matches measured data residency times.

The SSD serves both as a data cache for the /root file system and the interactive swap space and as a data buffer for the large 42 GByte /usr/tmp file system. Caching versus data buffering depends on the amount of data reuse. The caching of /root to support interactive users is effective since a hit ratio exceeding 99% can be sustained when the cache size is set to 68 MBytes. Data caching for the interactive swap space is effective when about three MBytes of swap space is reserved per user. The actual interactive swap partition at SDSC is 320 MBytes on the SSD and is restricted to supporting job sizes less than 8 MBytes. Since the total SSD size is 1 GByte, there is not enough room to cache the 42 GByte /usr/tmp file system. Instead the /usr/tmp data effectively stream through the SSD with minimal reuse. The net effect is that the SSD buffers 196 kByte disk data reads for 32 kByte accesses by the application codes. This helps minimize the amount of time spent waiting on disk seek latencies. Buffering of /usr/tmp files dominates the I/O rate needed to support swapping of interactive jobs by a factor of 2.5. Although the SSD transfers data at over 1 GByte/sec, the steady state I/O rate needed to support streaming 1.5 TBytes of data per day through the SSD is only 17 MBytes/sec. Replacing with a slower speed communication channel would seriously degrade interactivity. Swapping jobs at the average transfer rate would require up to one/half second to load an interactive job into memory. Thus the dominant I/O support requirements for the SSD are split between providing a large storage area for data buffering and providing very high-speed access for the interactive job swapping data subset.

Local CRAY Disk Cache

The CRAY disks also sustain a total amount of I/O of about 1.5 TBytes per day, or an average of 17 MBytes/sec. Since the total /usr/tmp disk space is only 42 GBytes, the majority of this I/O is to scratch files which disappear at problem termination. This can be calculated using the average batch job execution time of one hour and the write rate to disk being one fourth the read rate. If all the generated data were saved, only about three hours of CRAY execution data could be stored on local CRAY disk before they would have to be migrated elsewhere. In practice, the files reside much longer on disk. Typically 60% of the disk files are up to one day old, and another 25% are up to two days old. The average residency time is about 30 hours, implying that only one tenth of the data written to disk survives application code termination. The CRAY disks therefore are serving as a cache for writing data from the supercomputer.

Archival Storage Disk Cache

The true long-term data generation rate is governed by how fast data are migrated to archival storage. On the DataTree archival storage system in use at SDSC, archiving of files is a user initiated process. Users explicitly choose which files to archive or retrieve. Typically 14 GBytes/day of data are transferred between the CRAY disks and the archival storage system of which one third is data written to storage. This amount of data flow is only 1/7 of that needed to migrate the data that survive on CRAY disk to archival storage. Thus about 1.4% of the total amount of data written to CRAY disk is archived. The archival storage disks form an effective cache between long-term storage on cartridge tape within the tape robot and the CRAY local disks. The hit ratio for archival storage data being retrieved from the archival storage disks is typically 92%.

Archival Storage Tape Caches

The average data transfer rate needed to support archival storage is 0.16 MBytes/sec. This should be compared with the observed sustainable archival storage data rates of 0.6 MBytes/sec supported by DataTree running on an Amdahl 5860 across 4.5 MBytes/sec I/O channels connected to a 12.5 MBytes/sec FDDI backbone network. During periods of heavy usage, the average transfer rate does approach the peak rate.

Long term archival storage to tape occurs both directly from the CRAY disk for large files (sizes greater than 200 MBytes) and by automatic data migration from the archival storage disks. The tape robot serves mainly as a data cache. Data currently reside about 15 months before migrating to shelf cartridges. Data caching attributes can be tracked by the fraction of tape mounts done manually. Typically the 1.2 TByte tape robot processes 85-90% of the tape mounts. The rate at which data are migrated from the tape robot to shelf tape is roughly 2/3 of the rate at which data are written to the robot. This ratio may approach one as data in the robot mature.

This data flow analysis demonstrates some interesting attributes of loosely-coupled user-initiated archival file storage systems.

> 10% of the generated data is stored temporarily on CRAY local disk, 1.4% of the generated data is written to archival storage, and 0.6% of the generated data is eventually transferred to long term shelf tape. Given the need to explicitly save files, users selectively store a fraction of their output.
>
> The multiple levels of the storage hierarchy serve mainly as caches with more data flowing into a given cache than flows out to lower caching levels.
>
> The amount of data read at each caching level is substantially higher than the amount written with the ratio varying from 4:1 for the highest speed cache on the SSD down to 2:1 for archival tape storage.

The above data flow analysis is typical only of user-initiated archival storage. If an automated archival storage scheme is used for supporting the CRAY disks, the amount of data that are archived could grow substantially. This can seriously impact the ability to adequately handle the I/O if the archival storage hardware environment is operating with relatively small safety margins. Pertinent safety factors are:

> cache residency time versus the latency that a data buffer is amortizing,

cache residency time of data files on local CRAY disks versus the time needed for the application to complete, and

sustainable I/O rate versus the peak I/O demand rate.

If any of these factors drop below one, the system will become severely congested and may even fail. At SDSC, all of these safety margins are relatively small. Due to the limited amount of CRAY local disk space, the residency time of files on CRAY disk is comparable to the wall clock time needed to complete an application run for large codes. The weekly average required I/O rate to access files on the archival storage system is 1/4 of the peak observed sustainable rate. Hourly averages of the required I/O rate approach the peak sustainable rate. A usage paradigm shift that increases the I/O load could seriously stress the archival storage system at SDSC.

File Server Paradigm Shifts

Three possible usage paradigm shifts are being investigated at SDSC, two of which are related to file servers tightly coupled to the supercomputer CPU power. The first is a research project funded by the National Science Foundation and DARPA through the Corporation for National Research Initiatives. Prototypes of tightly coupled applications distributed across supercomputers connected by a gigabit/sec network are being developed, including the linkage of an application to the equivalent of a database interface to archived data. The second is a project to investigate the feasibility of incorporating the local CRAY disk and the SSD as caches directly controlled by the archival storage system. The third is the modeling of the impact on the archival storage system of an upgrade to a 100 Gigaflop/sec supercomputer.

High-speed Remote Access

The CASA Testbed is a collaborative effort between the California Institute of Technology, the Jet Propulsion Laboratory, the Los Alamos National Laboratory, and the San Diego Supercomputer Center. One objective is to demonstrate a distributed application efficiently utilizing two supercomputers while simultaneously using a substantial fraction of the gigabit/sec wide area network linking the computers. Simultaneously maximizing bandwidth utilization and CPU utilization requires minimizing the protocol overhead used for the data transmission[1]. The effective bandwidth for the optimal application is given by

$$B / (1 + O * B)$$

where B is the peak bandwidth (bits/sec) and O is the network protocol overhead measured in seconds of overhead per bit transmitted. For high speed networks, network protocol overhead becomes a critical limiting parameter. For present CRAY supercomputers, the network protocol overhead can require the execution power of an entire CPU to support TCP/IP at 700 Mbits/sec.

Given that a suitable file transport protocol is devised with a small enough protocol overhead, the issue of latency across wide area networks may be the next limiting factor. Since the speed of light is finite, data access delays between SDSC and LANL are as great as disk seek times. Efficient access of remote file systems must then cope with buffering data in addition to caching data. The amount of data shipped between an application and a remote database interface to archival storage must be large enough to amortize the data access delay. Depending on the protocol, the amount of data sent may need to be as large as

$$2 L * B$$

where L is the round trip latency measured in seconds. For a LANL/SDSC application running at 800 Mbits/sec, this is still feasible, requiring buffering on the order of 8 MBytes.

Integrated Local and Archival File Systems

Integrating the local file system into the archival storage file system will substantially increase the amount of data that must be processed by the archival storage software. As seen in the SDSC data flow analysis, the amount of data transferred between the supercomputer and the local disks is more than a factor of 1000 larger than the amount transferred to archival storage. Efficiently handling this increase in data rates will require differentiating between "reliable" local file transport and "unreliable" transport across a local network. By scaling the network protocol overhead needed to support TCP/IP at 700 Mbits/sec by the average CRAY local disk bandwidth derived in the data flow analysis, an estimate can be made of the protocol overhead increase. With no protocol enhancements, an additional 20% of a single CPU would be needed to support the archival and local file system integration. This indicates the need for the integrated system to recognize heterogeneous network environments.

An additional complication is that if all of the generated data stored temporarily on CRAY disk is automatically archived, the data flow from local CRAY disk to archival storage could increase by up to a factor of seven. Files written to the scratch /usr/tmp file system require different backup than files written to permanent home directories. An integrated local file system and archival storage file system must allow for a non-uniform usage pattern.

CPU Execution Rate Dependence

A possible ameliorating effect is that as supercomputers become faster, it may become more cost effective to recompute rather than save data. A supercomputer with a sustained execution rate of 100 Gigaflops is expected to be available by 1995. Assuming the data storage patterns remain the same, the I/O generated by such a machine can be estimated by scaling the results of the data flow analysis by the increase in the execution speed, which is roughly a factor of 3000. The cache sizes and I/O communication rates then become:

SSD	3 TBytes	50 GBytes/sec
Local disk	126 TBytes	50 GBytes/sec
Archive disk	210 TBytes	450 MBytes/sec
Shelf tape	6000 TBytes	60 MBytes/sec

The archival storage communication rates need to be decreased by a factor of 10 to become technically feasible. Thus a paradigm shift towards the dynamic regeneration of simulation output may become inevitable.

Acknowledgement

This work was funded in part by the National Science Foundation under Cooperative Agreement Number ASC-8414524 and Grant Number ASC-9020416.

References

1. Moore, Reagan, "Distributing Applications Across Wide Area Networks," General Atomics report GA-A20074, April 1990.

Table 1

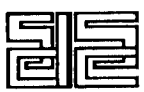
Hierarchical Data Caching Levels

Caching Level	I/O per Day	Data Rate	Capacity	Utilization	Residency Period
SSD	1.5 TB	17 MB/s	1 GB	85-100%	minutes
CRAY Disk	1.5 TB	17 MB/s	42 GB	85-90%	days
Archive Disk	5 GB	0.05 MB/s	70 GB	98%	weeks
Tape Robot	9 GB	0.10 MB/s	1.2 TB	68%	months
Shelf Tape	2 GB	0.02 MB/s	2 TB	70%	years

File Servers, Networking, and Supercomputers

Reagan W. Moore

San Diego Supercomputer Center
San Diego, California



Archival Storage Systems as File Servers

- **Examine Hierarchical Caching systems**
  - Capacity requirements
  - I/O requirements
- **Based on Usage at SDSC**
  - Archiving supercomputer generated data

File System Usage Paradigms

- **Loosely-coupled to CPU**
  - User initiated file transfers to archival storage
- **Tightly-coupled to CPU**
  - NFS access
  - Integrated local and archival file systems

SDSC Archival Storage Environment

- **Data Generated by CRAY Y-MP8/864 Supercomputer**
- **FDDI 100 Mbits/sec backbone**
- **DataTree Archival Storage System on an Amdahl 5860**

Five Levels of Data Caching

- **Solid State Storage Device (SSD)**
  - 1 GB, 1.2 GB/s access from memory
- **CRAY local disk**
  - 42 GB, 10 MB/s access per disk
- **Archive storage disk**
  - 70 GB, 0.6 MB/s access across FDDI
- **STK tape robot**
  - 1.2 TB, 0.6 MB/s access across FDDI
- **Shelf cartridge tape**
  - 2 TB, 0.6 MB/s access across FDDI

SDSC Workload Characteristics

- **Application Disk I/O**
  - Generated by an average batch load of 8 GBs of executable jobs
- **Job Swapping**
  - Generated by up to 120 interactive users
- **User-initiated File Archiving**
  - Partial archiving of supercomputer data

Data Flow Analysis

- **Track Data Through the Multiple Caches**
  - Cache utilization
    - Hit rate
  - I/O throughput
    - Fraction of peak rate
  - File residency time
- **Identify Caching versus Data Buffering**

SDSC Data Flow

Cache Level	Capacity (GB)	Utilization
SSD	1	85%
CRAY disk	42	90%
Archive disk	70	98%
Tape robot	1200	68%
Shelf tape	2000	70%

SDSC Data Flow

Cache Level	Residency Time	Fraction saved of total I/O written from SSD
SSD	(seconds)	100%
CRAY disk	30 hours	10%
Archive disk	4 weeks	1.4%
Tape robot	15 months	1.4%
Shelf tape	5 years	0.6%

SDSC Data Flow

Cache Level	I/O per Day (GBytes)	Data Rate (MBytes/sec)
SSD	1500	17
CRAY disk	1500	17
Archive disk	5	0.05
Tape robot	9	0.10
Shelf tape	2	0.02

Data Caching Versus Data Buffering

- **SSD Cache Used for Both**
  - /root file system and Interactive swap space are cached
    - Hit rate for accesses is 99%
  - /usr/tmp file system is buffered
    - Hit rate for accesses is 75-85%

File Server Safety Factors

- **Cache Residency Time versus Latency Amortization Time**
- **Cache Residency Time versus File Usage Time**
- **Sustainable I/O Rate versus Peak I/O Demand Rate**

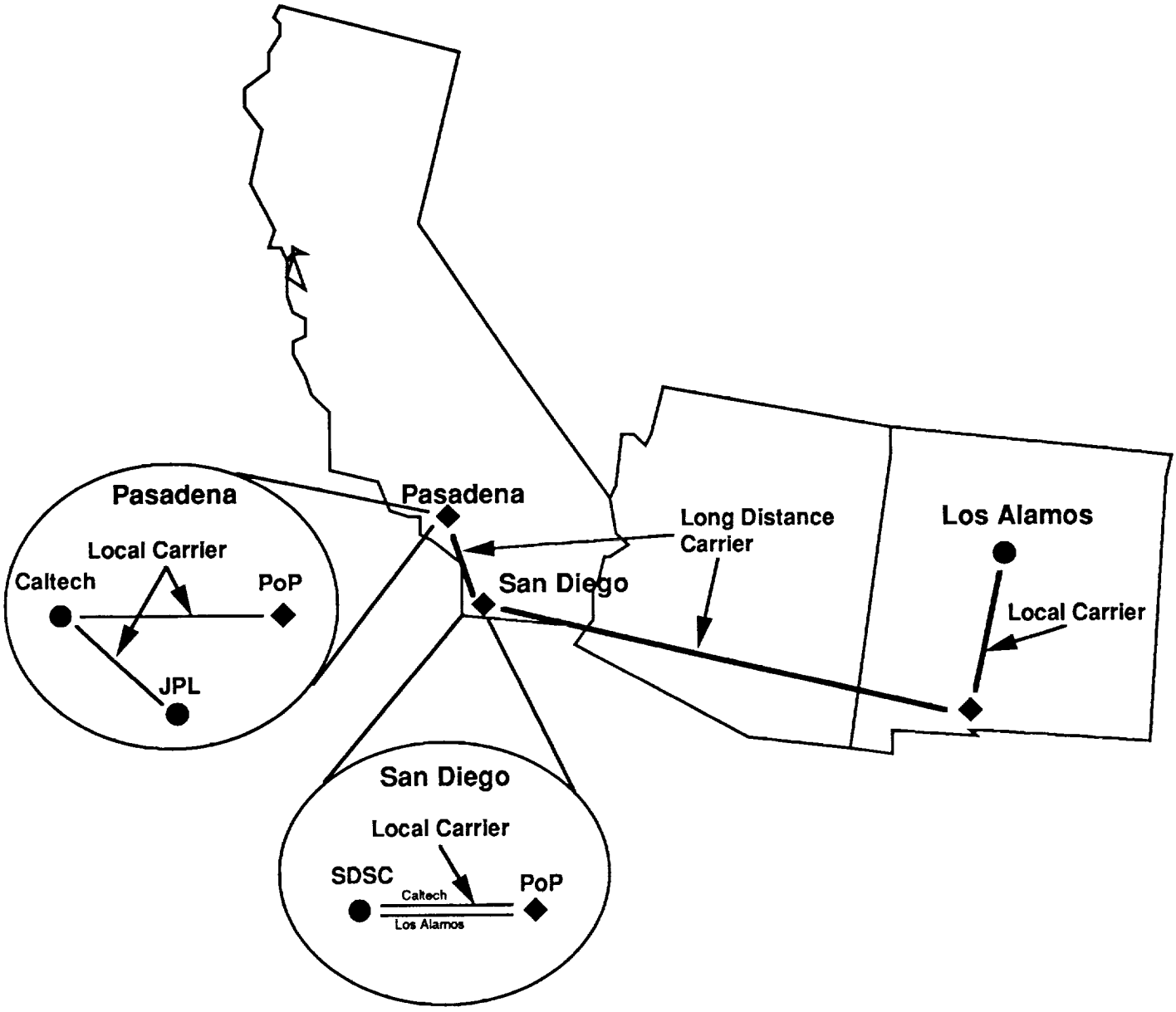
File Server Paradigm Shifts

- **Changes in Funtionality May Require Usage Paradigm Shift**
  - High-speed remote access
  - Integration of local and archival file systems
  - Very high-speed supercomputers

CASA Gigabit/sec Testbed

- **Collaboration between CalTech, JPL, LANL, SDSC**
- **Demonstrate Tightly Coupled Distributed Applications Linked by Gigabit/sec Wide Area Network**
  - Remote access of archived data through database interface

CASA GIGABIT WAN



◆ Point of Presence

11/20/89

Network Protocol Overhead Impact

- **Simultaneous Optimization of CPU and Bandwidth Utilization**
- **Effective bandwidth is given by**
  - B / (1 + O * B)
  - B = bandwidth (bits/second)
  - O = protocol overhead (seconds/bit transmitted)

Protocol Support Limitations

- **TCP/IP Protocol Can Require Execution Power of Entire CPU of Y-MP8/864 for 700 Mbits/sec Bandwidth**
- **Effective Bandwidth Reduced 45%**

Wide Area Network Latency Can Require Data Buffering in Addition to Data Caching

- **Finite Speed of Light Creates Latency Between SDSC and LANL Comparable to Disk Seek Latencies**
- **Amortize Latency by Shipping Large Files**
  - Size = 2 L * B
    - L = Round-trip latency (seconds)
  - For 800 MBits/sec network, ship 8 MB files

Integration of Local and Archival File Systems

- **Local CRAY Disk Supports 1000 Times as Much Data Transfers as Archival Storage at SDSC**
- **To Minimize Protocol Overhead**
  - Distinguish between
  - "Reliable" local file transport
  - "Unreliable" local network transport
  - Otherwise expect overhead to increase 20%

Integration of Local and Archival File Systems

- **User-initiated File Archival Storage Results In**
  - 1/7 of the data being archived
- **Automatic Migration of Local Files**
  - Allow non-uniform file migration across different file systems
    - /root versus /usr/tmp

Supercomputer I/O Scaling

- **For a 100 Gflops/sec Supercomputer**
  - Scale I/O by ratio of CPU speeds
  - Expect 3000 times as much I/O
- **Massive data generation may require dynamic regeneration of data rather than storage**

CPU Execution Scaling

Cache Level	Capacity (TB)	Data Rate (MB/s)
SSD	3	50,000
CRAY disk	126	50,000
Archive disk	210	450
Tape robot	6000	60
Shelf tape	12000	60

File Server Paradigm Shifts

- **Data Storage Requirements Will Be Increased by**
  - Integration of Local and Archival Systems
  - Higher Speed Supercomputers
- **Possible Shifts**
  - Local regeneration of data
  - Remote File Access

NSSDC Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications

Abstract

Status of Standards for Removable Computer Storage Media and Related Contributions of NIST

by

Fernando L. Podio

Standards for removable computer storage media are needed so that users may reliably interchange data both within and among various computer installations. Furthermore, media interchange standards support competition in industry and prevent sole-source lock-in. NIST participates in magnetic tape and optical disk standards development through Technical Committees X3B5, Digital Magnetic Tapes, X3B11, Optical Digital Data Disk, and the Joint Technical Commission on Data Permanence. NIST also participates in other relevant national and international standards committees for removable computer storage media.

In support of these emerging standards, computer storage magnetic tapes require the use of Standard Reference Materials (SRMs) developed and maintained by NIST. In addition, NIST has been studying care and handling procedures required for digital magnetic tapes.

Optical disks, because of their potential use for the long-term storage of valuable data, require standard testing methods for predicting the life expectancy of the media. NIST is currently developing test methods for determining the life expectancy of optical disks. NIST is also developing care and handling procedures for optical digital data disks.

This presentation reflects the status of emerging magnetic tape and optical disk standards, as well as NIST's contributions in support of these standards.

THE LONG HOLD: STORING DATA AT THE NATIONAL ARCHIVES

Kenneth Thibodeau, Ph.D.
Director, Center for Electronic Records
National Archives and Records Administration

The National Archives is, in many respects, in a unique position. For example, I find people from other organizations describing an archival medium as one which will last for three to five years. At the National Archives, we deal with centuries, not years. From our perspective, there is no archival medium for data storage, and we do not expect there ever will be one. Predicting the long-term future of information technology, beyond a mere five or ten years, approaches the occult arts. But one prediction is probably safe. It is that the technology will continue to change, at least until analysts start talking about the post-information age. If we did have a medium which lasted a hundred years or longer, we probably would not have a device capable of reading it.

The issue of obsolescence, as opposed to media stability, is more complex and more costly. It is especially complex at the National Archives because of two other aspects of our peculiar position. The first aspect is that we deal with incoherent data. The second is that we are charged with satisfying unknown and unknowable requirements.

The data is incoherent because it comes from a wide range of independent sources; it covers unrelated subjects; and it is organized and encoded in ways that not only do we not control but often we do not know until we receive the data.

The sources are potentially any operation of the Federal Government, or its contractors. The National Archives has been in the business of collecting digital data for two decades. The way we get it is through our authority over all Federal records. Under the Federal Records Act, no agency of the Federal Government can destroy or alienate any Federal record without authorization from the Archivist of the United States, who is the head of the National Archives and Records Administration. Simplistically, the way it works is that agencies tell us what records they have, and we tell them which ones they can destroy when they no longer need them, and which ones must be preserved for posterity. (The definition of Federal record in the law explicitly includes machine-readable files.)

Since 1972, we have reached agreements with agencies that provide for them to transfer to us, and for us to preserve, data from 600 data collections. 573 of them are still active. From these agreements, we have received over 10,000 data files. The rate of transfer has increased dramatically in the last two years: In fiscal year 1988, the National Archives received 167 data files. In FY 1989, 645 files came in, and in FY 1990 729. We anticipate a total of 1400 this year. And in each of the next two fiscal years we expect to receive at least 3000 data files. So we are currently operating at eight times the volume of new files we had three years ago, and we expect at least to double that next year.

Those numbers are very encouraging, but the overall picture is rather bleak. If we look at all of the data which was scheduled to arrive in the last twenty years, from those 600 data collections, we have received less than 7% of the transfers which should have been made. We have recently completed development of a system to generate dunning letters to agencies who fail to transfer data as scheduled, and to track each case to completion. But this system creates additional problems. If I implement it as planned, on a governmentwide basis, we would need to increase

our capability to handle new files, not by doubling current capacity, but by increasing it more than six times. And to handle the backlog of data which should have come in before now, I would need at least 10 times our current capacity.

The past gives us pause. But the future is a brave new world. At least it requires a degree of bravura just to glance in that direction. We have underway a study which is looking beyond the 600 data collections we have decided to preserve to see what else is out there. It is a study of major Federal databases being conducted by the National Academy of Public Administration (NAPA). This study has some interesting exclusions. First of all, we told NAPA not to bother with systems used for generic housekeeping functions, such as personnel, payroll, procurement and supply, because there is little likelihood that we would have any interest in preserving data from such a system. Secondly, we told them not to look at big science, because that is such a large and complex area that it deserves separate attention. (We hope to engage in a project with the National Academy of Sciences on the preservation of scientific data.) Thirdly, we told NAPA not to worry too much about databases on PCs, simply because they would never finish the project if they tried to find all the interesting databases sitting on desktops. With those limitations, NAPA has identified over 10,000 databases.

Obviously, that is far too big a number even for us to think about. So we gave NAPA a set of criteria for culling from the total inventory a subset of those databases with some likelihood that the National Archives would be interested in preserving them. We thought we might wind up with a list of the 500 most important databases in the Federal Government, from an archival perspective. That list would pose quite a challenge for us, because it could practically double the total number of data collections generating data that we want to preserve. The subset of 500 currently has about 900 members.

The next phase of this study is to solicit advice from subject area experts about what data we should try to preserve. NAPA has organized five working groups, with a total of 32 experts in a variety of fields. We are bringing these people together at the end of July for a four day meeting where they will try to develop some common opinions on the long term value of the data.

Which brings me back to the basic point here: what we are dealing with is incoherent data. It concerns practically any area in which the United States Government is involved, which is practically anything. The data we already have ranges from data about tektites on the ocean floor to military operations in time of war. In includes census data on population and the economy, data on Japanese-American internees in World War II, detailed data on air traffic and on stock and bond transactions, and on many, many other subjects. The variety of subjects covered is also increasing.

The data is extremely diverse in content, but content is often the only thing we know about the data until it comes in. We know how many transfers are due, but most often we do not know what the volume of data in a transfer will be, or how it will be organized, even at the physical file level. For example, the files which came in during the first six months of this fiscal year ranged in size from 6 K to 1.4 gigabytes. The number of files in a transfer has ranged from one to 400, and we expect some transfers in the next few years will contain thousands of files.

One thing we do know about the data before it arrives is its logical structure: everything we receive is in flat file format, because we require it to come in in that form. However, we realize that this requirement is unreasonable and unrealistic in many cases. We are working to expand the range of formats we will accept to include relational tables. We expect to change

our regulation to that effect by the end of this year. We know that, when we do that, it will be only one of many steps we will have to take in a journey with no foreseeable end.

That is a brief overview of one aspect of the unique situation of the National Archives. The second aspect is that we are charged with satisfying unknown and unknowable requirements.

NARA's mission to preserve and provide access to records with enduring value makes NARA, in effect, the agent of generations yet unborn. What differentiates this agency from other parts of the government is the unique responsibility NARA has to serve the information needs of the distant future. This responsibility is fundamental to the very essence of the National Archives as keeper of the Nation's memory.

NARA's responsibility to the future places us in a perpetual quandary: we must devote ourselves to serving needs which we cannot know. We cannot know the questions the future will ask of its past, nor how future researchers will go about answering these questions. We must assume, however, that the information technology which will be available in the future --- even in the very near future --- will be more powerful and more flexible than what is available today. Information processing problems which today are difficult and costly, if not impossible, to solve will become as simple as getting a computer to print out narrative in paragraph form. (A short 20 years ago that was beyond state of the art.)

Along with the technology, analytic tools will continue to improve: there will be further developments as powerful as the mathematics of chaos which will help researchers to understand things which today appear to defy reason. We can also assume that events will happen in the future, which will be as

threatening as the depletion of atmospheric ozone, or as exciting as Operation Desert Storm, or as commonplace as the passing of generations, which will make future users want to go back to reexamine the records of the past.

S5-82

121942

N93-15032

STEWARDSHIP OF VERY LARGE DIGITAL DATA ARCHIVES

Patric Savage
Shell Development Company

Call an archive a permanent store.

Some of the largest *digital data archives* are operated by oil exploration organizations. The vast bulk of these archives is seismic data. It is kept forever because of its extremely high cost of acquisition, and because it often cannot be re-acquired (due to cultural buildup, political barriers, or difficult logistical/administrative factors). Western Geophysical operates a seismic data archive in Houston consisting of more than 725,000 reels/cartridges, typical of the industry. Oil companies fondly refer to their seismic data troves as "family jewels."

There are relatively few "very large" digital data archives in existence. Most business records are (gladly) expired within five or ten years depending on statutes of limitations. And many kinds of business records that do have long lives are embedded in data bases that are continually updated and re-issued cyclically. Also a great deal of "permanent" business records are actually archived as microfilm, fiche, or optical disk images - their digital version being an operational convenience rather than an archive.

So there is not really much widely known about operating *digital data archives*, let alone *very large* ones. Even the oil companies have been in a sense overwhelmed by this somewhat unplanned for hugeness.

This paper addresses the problems foreseen by the author in stewarding the very large digital data archives that will accumulate during the mission of the EOS. It focuses on the function of "shepherding" archived digital data into an endless future.

Stewardship entails a great deal more than storing and protecting the archive. It also includes all aspects of providing meaningful service to the community of users (scientists) who will want to access the data. The complete steward will:

1. Provide against loss due to physical phenomena.

2. Assure that data is not "lost" due to storage technology obsolescence.

3. Maintain data in a current formatting methodology. Also, it may be a requirement to be able to reconstitute data to original as-received format.

4. Secure against loss or pollution of data due to accidental, misguided, or willful software intrusion.

MD117901.PS

5. Prevent unauthorized electronic access to the data, including unauthorized placement of data into the archive.

6. Index the data in a *metadatabase* so that all anticipatable queries can be served without searching through the data itself.

7. Provide responsive access to the metadatabase.

8. Provide appropriately responsive access to the data.

9. Incorporate additions and changes to the archive (and to the metadatabase) in a timely way.

10. Deliver only copies of data to clients - retain physical custody of the "official" data.

Items 4 through 10 are not discussed in this paper. However, the author will answer questions about them at the conference or by email or telephone.

Providing Against Loss Due to Physical Phenomena

Broadly classifying these we have:

1. Site destruction

2. Theft/robbery

3. Sabotage

4. Media unit suffers severe damage

5. Systemic media degradation

The first three can be guarded against, but not absolutely. The fourth is a rare inevitable eventuality (e.g., a mechanically faulty drive "eats" a tape.)

Systemic media degradation is best managed by using only media that are known to have archival properties, by conservatively rewriting media that, when accessed, are found to have an error, by regularly running PM according to vendors' recommended practice (e.g., winding and re-tensioning tape), and copying the entire archive to new-generation media. The last must be planned for, budgeted for, and be resigned to - it is an imperative. Generally speaking, one media generation can be leapfrogged by the copy procedure: for example, when Shell adopted 3480 technology, all of the 1100bpi tapes were copied; when 3490 technology is adopted, all of the 6250bpi tapes will be copied. However, copying can be mandated earlier if media are observed to be systemically degrading faster than anticipated.

Media failure occurs when an uncorrectable bit error is detected. This always causes/implies loss of an entire error correction block - ordinarily a minimum of one kilobyte of archived data. Archivists should be aware that the media vendors' touted "hard error rate" always has a 10,000-fold impact. A badly degraded media unit might have relatively many unreadable error correction blocks; hence even a redundancy array of media units might then (by a little bad luck) have an unrecoverable error correction block.

A practical cost-effective solution to the problem of protecting against physical loss can be tailored around the following concept (which came to me while ruminating about extending the now-familiar RAID idea to striping tape). This is merely the seed of an idea, to which a good deal of systems thought will have to be given.

Some number (in this example, 10) of archiving sites are chosen to participate/cooperate in a redundancy scheme that provides mutual protection against all modes of physical loss.

The sites should be geographically distributed in order to eliminate concern that a calamity (e.g., earthquake/meteor strike) would wipe out multiple archives. Of course, all sites individually should have reasonably good physical security.

All sites must be accessible via state-of-the-art WAN technology.

Each site houses, primarily, its own archive of data. (A variation would have a single archive partitioned and distributed among its own multiple sites.) Clients of an archive would communicate only with the primary site.

Each site also houses either p-parity or q-parity data generated from (in this example) 9 other sites. (Optionally two sites could be dedicated, one for p- and the other for q-parity.)

In the eventuality of a loss at a site (of an error correction block, or a media unit, or the site itself) any 8 of the 9 other sites reconstruct the lost data. This would not be instantaneous, as with RAID, because an extraordinary procedure would have to be executed; but the insurance would be very certain. Clearly, each site should be practicing high quality archiving methodology, so that losses would occur with extreme rarity (say, no oftener than one per month).

The merits of this scheme are first, that the storage overhead for backup can be small (25% for this example); second, that the degree of protection can be high (with both p- and q-parity) or lower (with p-parity only); third, independent archives do not have to create their own backup systems, but can band together in a consortium for mutual protection.

Assuring That Data is Not "Lost" Due to Storage Technology Obsolescence

The 1960 census was archived on the best storage medium known at the time: UNIVAC metal tape. There was a rude awakening some years later when it was discovered that only two drives existed in the world - one in Japan, and the other, dismantled, in the Smithsonian.

We know now that drive technology lifetime, even assuming heroic geriatric care, is scarcely ten years. Vendors drop maintenance after low-level parts' technologies disappear. For a while thereafter, drives can be cannibalized for parts; but ultimately maintenance becomes impossible.

The optical disk vendors, for example, tell of the fine archival qualities of their media. But their technology is evolving quite rapidly - vendors come and go, and recording formats with them. Here we have a single medium that undoubtedly lasts a long time, but the drive and recording technology has a half-life of less than five years. Considering the relatively high cost of optical media, copying an archive every five years seems out of reason.

Archiving demands that digital data on old storage technology be copied to new storage technology periodically. The frequency depends on the media, on how widely the drives were accepted, and on whether the old technology satisfies current access requirements. Keeping too many generations of storage technology in use can cause serious operational problems, even if they are all in good working condition. For example, 556bpi tape would be much too slow for regular use today, so, even though drives are still available, that technology is obsolete.

Maintaining Data in a Current Formatting Methodology

The winds of computation methodology are ever varying. Yesterday there was no C language. Today C-readable records might be a good bet. Tomorrow the fad may be object files. What will come next? The curse of required media copying is really a blessing because it enables us to continually modernize our data language. Cuneiform tablets were certainly archival, but they contain antiquated, almost unreadable language.

Standards for "self-defining" data formats are evolving rapidly and are already very useful. The time has come to abandon schema-less data formats (where programs know implicitly where every field is in a record, and what each field means).

Even fixed (schema'd) formats are passé for scientific data because of the continual change in interest and emphasis in almost every scientific specialty.

Archivists can extract a side benefit when copying to a new media generation. Indeed, the planning for the copy should include deciding which new formatting standard is to be adopted. Migrating from old to new formats is only slightly less important for archiving as migrating from old to new media technology. What's more, it's almost free.

EMASS™: AN EXPANDABLE SOLUTION FOR NASA SPACE DATA STORAGE NEEDS

Anthony L. Peterson
P. Larry Cardwell

E-Systems, Inc. Garland Division
Dallas, Texas

Abstract

The data acquisition, distribution, processing and archiving requirements of NASA and other U. S. Government data centers present significant data management challenges that must be met in the 1990's. The Earth Observing System (EOS) project alone is expected to generate daily data volumes greater than 2 Terabytes (2×10^{12} Bytes). As the scientific community makes use of this data their work product will result in larger, increasingly complex data sets to be further exploited and managed. The challenge for data storage systems is to satisfy the initial data management requirements with cost effective solutions that provide for planned growth. This paper describes the expandable architecture of the E-Systems Modular Automated Storage System (EMASS™), a mass storage system which is designed to support NASA's data capture, storage, distribution and management requirements into the 21st century.

Introduction

We first discuss NASA's requirements for mass storage with a focus on functional and performance specifications. Next, an overview of the EMASS architecture is presented and evaluated with respect to NASA's requirements. The major EMASS architectural components, hardware, software and interfaces, are then explored with emphasis on the data management capabilities of the EMASS software.

NASA Requirements

Requirements for large volume, mass storage systems have been well established in order to meet the storage needs for NASA's space and Earth science information systems. The use of sophisticated data acquisition instrumentation will continue to evolve, providing large, increasingly complex data sets to be processed, distributed and archived. Therefore, data storage requirements

will continue to grow nonlinearly through the 1990's. For example, the Earth Observing System (EOS) project alone, generating daily volumes greater than 2 Terabytes, will require automated storage libraries with capacities greater than 500 Terabytes by the late 1990's. E-Systems is also currently developing storage systems to meet existing U. S. Government and commercial requirements to be delivered in 1993 having automated data storage library capacities greater than 200 Terabytes.

Data management requirements such as these within NASA and other U. S. Government data centers present significant challenges that must be met in the development of new mass storage systems. These systems must meet increasing performance requirements with cost effective solutions while providing for planned growth. E-Systems is developing the EMASS architecture to address these requirements for extremely large, expandable data storage and data management systems.

As we view NASA's supercomputer-based data management systems we see a need for high bandwidth, high density tape recorder systems having the data quality characteristics of a computer peripheral. As scientific data processing requirements move towards open systems environments, the file management software and server should support a UNIX environment. The file management software structure should provide application specific integrated data management solutions. A file server with high I/O bandwidth is required to accommodate simultaneous data transfers from multiple high bandwidth tape recorder systems. Finally, to keep a perspective on hardware and maintenance costs, the use of commercially available equipment is strongly emphasized.

EMASS Architecture Overview

The EMASS architecture is a family of hardware and software modules which are selected and combined to meet these data storage requirements. Figure 1 illustrates the EMASS architecture. EMASS is a UNIX-based hierarchical file management system utilizing both magnetic disk and tape. The storage capacity ranges from one to several thousand Terabytes depending on the type of storage library used. It has the capability to support both a graphical and metadata interface to the user. The system is user driven by standard UNIX and unique EMASS commands and has user configurable automatic file migration. The EMASS system employs standard protocols for user file transfer, communications and network interfaces.

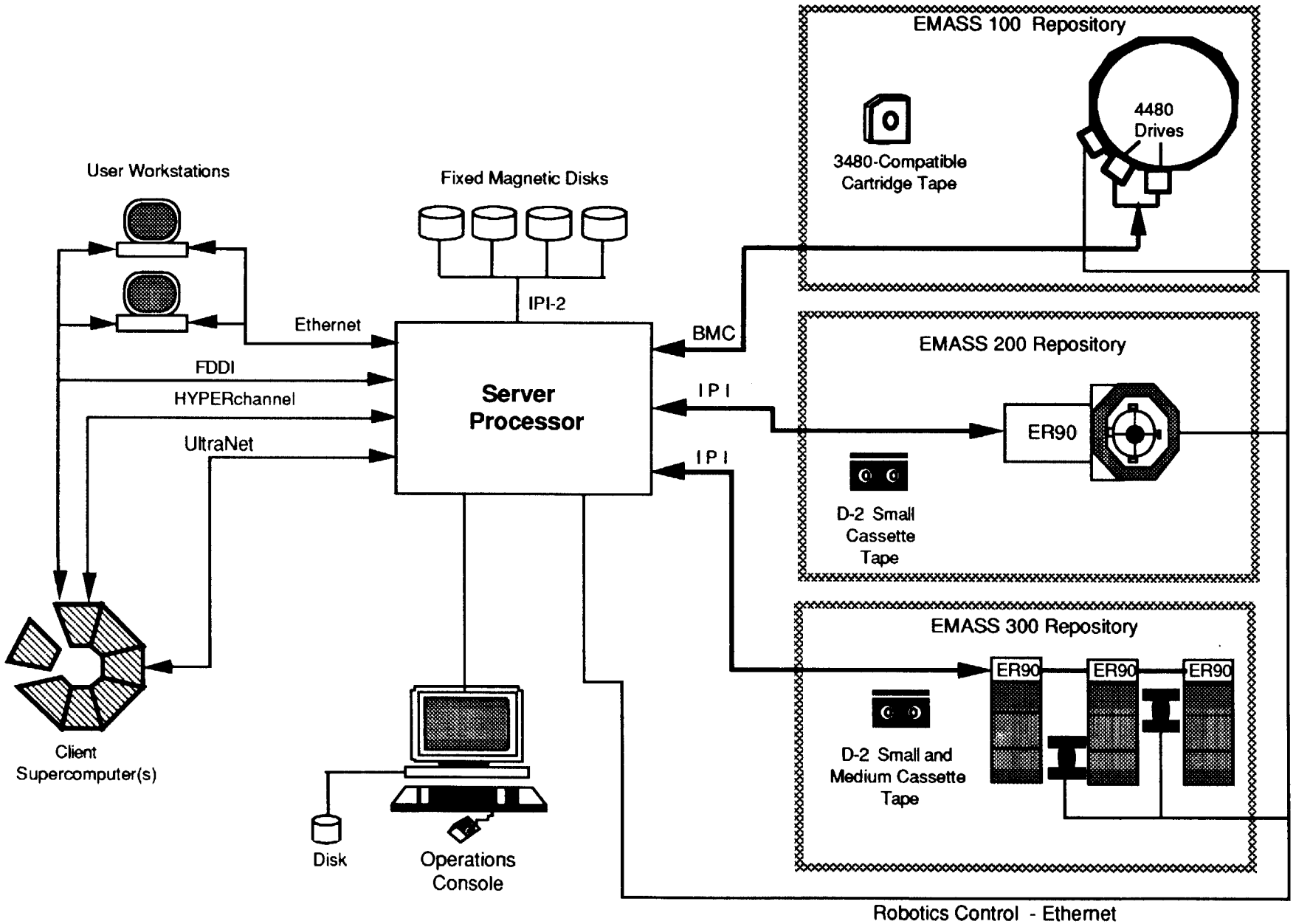


Figure 1. EMASS Architecture

The EMASS system operates as a large data storage node on a network, servicing client requests over a number of standard interfaces, including Ethernet, FDDI, DECnet, HYPERchannel™ and UltraNet™. The system is a two-level hierarchical data storage system. Magnetic disk is the first level of storage, and magnetic tape is the second. Data is managed via a selectable migration policy based on data class, a method of data segregation addressed in a subsequent section. Two alternative types of magnetic tape for data storage are included in the EMASS design: 3480 tape cartridges and D2 digital tape cassettes. Files are migrated to 3480 or D2 tape depending on the migration policy for the data class to which that file belongs. The physical volume repository (Miller[1]) functionality is implemented in three separate types of storage libraries which are selected based on user requirements for performance and expandability.

EMASS Hardware

The storage system file server function is implemented in a CONVEX C3200 series computer. The CONVEX was selected after an extensive survey of available computers. The major evaluation factor leading to the selection of the CONVEX machine is its high I/O throughput performance. The CONVEX supports four channels, each having I/O bandwidths of 80 Megabytes per second peak and 60 Megabytes per second average. Other key evaluation factors included cost, compatibility with a UNIX environment, modularity, expandability, upgradability, reliability, and support.

The file server interfaces with three types of tape libraries, the STORAGETEK (STK) 4400 Automated Cartridge System, the EMASS DataTower™ and the EMASS DataLibrary™. The STK tape cartridge library data interface is implemented through ANSI standard Block Multiplexor Channel interfaces which connect to the STK 4480 drives through a SUN Library Server . The DataTower™ and DataLibrary™ data interfaces are implemented with enhanced ANSI standard IPI-3 tape controllers within the file server connected to E-Systems ER90 digital D2 recorders.

The DataTower and DataLibrary robotic systems provide data archive expandability. The DataTower, with dimensions illustrated in Figure 2, serves as a medium scale storage device, with a capacity of 6 Terabytes on 227 small D2 cassette tapes. This device was implemented by

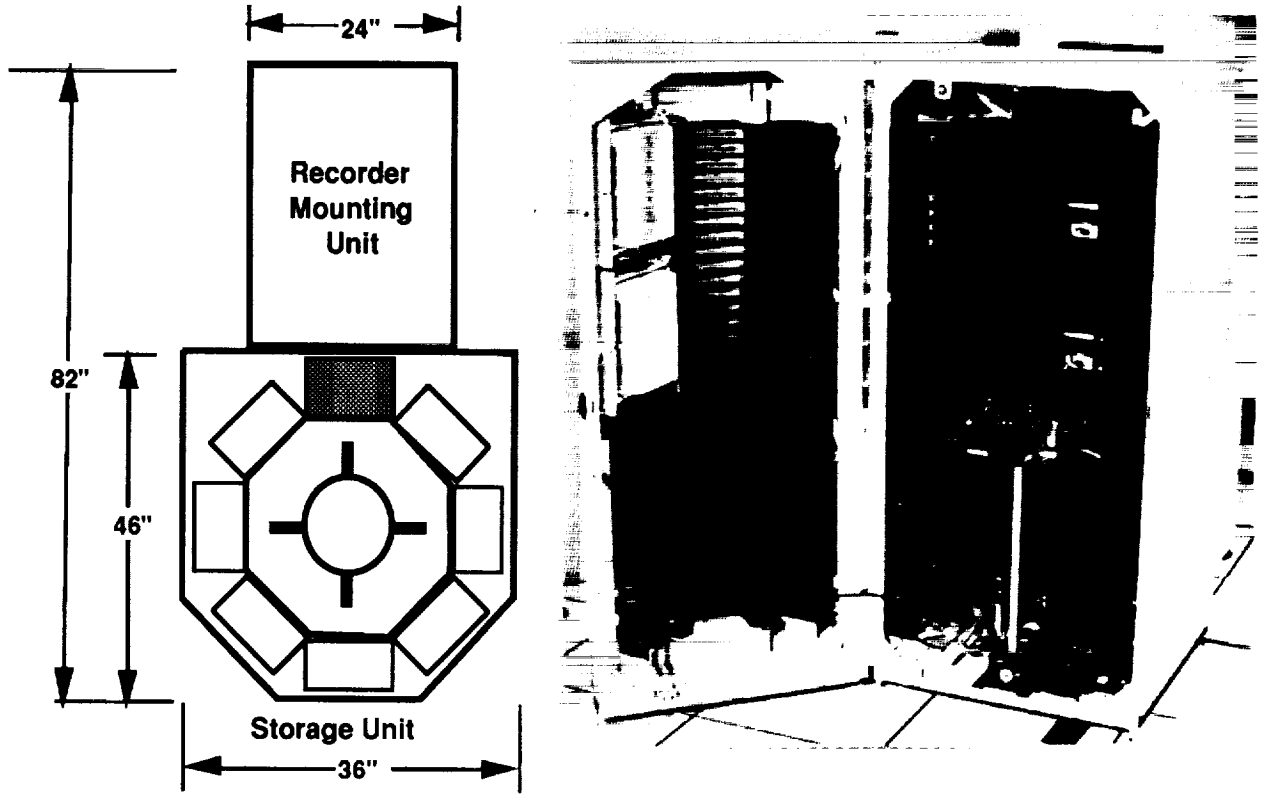


Figure 2. EMASS DataTower

modifying an existing automated robotics tower currently in volume production for the broadcast industry. The device may be expanded by adding up to three additional expansion storage units, for a total capacity of 25 Terabytes.

The DataLibrary, illustrated in Figure 3, is a modular aisle architecture comprised of a series of modules each four feet in length. This design specifically addresses the needs for a modular, expandable data storage solution required for NASA's large data archives from EOS and Space Station Freedom. Each shelf module contains up to 207 small, or 192 medium, D2 tape cassettes, for a maximum capacity of 14 Terabytes. Shelf units reside in rows on either side of a self-propelled robot and can be added incrementally as the library grows. The row of modules may be expanded to lengths of 80 feet, providing a maximum of 288 Terabytes per row. Further expansion is accomplished by adding additional rows and robots. Cassette access times are specified at 45 seconds maximum for robot travel spanning an 80 foot aisle for cassette retrieval. The DataLibrary configuration will be housed within a sealed watertight structure with interior fire protection using CO_2 supplied on demand.

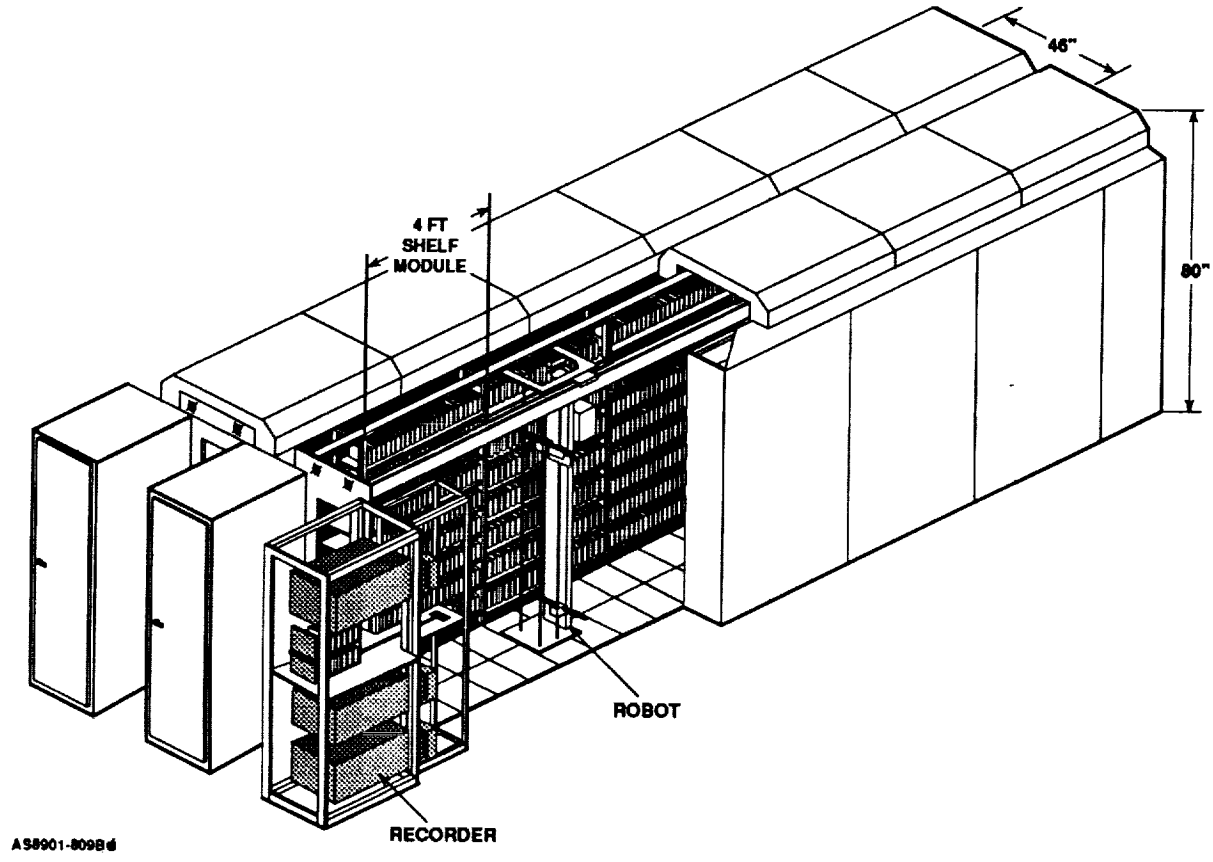


Figure 3. EMASS DataLibrary

E-Systems selected D2 helical scan tape and recorder system technology to meet high density and bandwidth requirements. As shown in Table 1, the 19mm D2 tape cassette is available in three form factors: small, with a capacity of 25 Gigabytes of user data, medium, with a capacity of 75 Gigabytes and large, with a capacity of 165 Gigabytes.

The suitability of the 19mm D2 helical scan media and recorder for use as a computer peripheral has been reviewed by Wood[2]. The D2 recorder provides a format already in wide use within the broadcast industry. Ampex, SONY, and Hitachi have delivered over 2000 D2 units to the broadcast industry since 1988. The D2 video broadcast recorder has been modified to develop the ER90 digital recorder peripheral. Key features of the ER90 recorder include air guides to minimize tape wear, azimuth recording and automatic scan tracking.

The ER90 provides a sustained data rate of 15 Megabytes per second, with burst rates up to 20 Megabytes per second. Additional error detection and correction coding has been implemented using a three-level interleaved Reed-Solomon code. Resulting error event rates of 1 in 10^{13} bits are being achieved. The recorder absolute positioning velocity is 300 inches per second and logical positioning velocity is 150 inches per second. This results in an average file access time for mounted media of 10 seconds for D2 small cassettes and 30 seconds for D2 medium cassettes. To provide compatibility with existing file management systems the ER90 provides ANSI 9-Track file labeling compatibility.

PARAMETER	SPECIFICATION
Tape Media	19mm - D2
Tape Cassette Capacities	25 GB (S), 75 GB (M), 165 GB (L)
Data Rate	15 MB/sec - Sustained 20 MB/sec - Burst
Error Detection/Correction	3-Level Interleaved R-S Code
Error Event Rate	1 in 10^{13} bits
Tape Positioning Velocity	300 in/sec - Absolute Address
Average File Access Time (Mounted Media)	10 sec - D2 Small 30 sec - D2 Medium
Data Format	Compatible With ANSI 9-Track File Labeling
Peripheral Interface	Enhanced IPI Physical (ANSI X3T9/88-82) IPI-3 Logical (ANSI X3.147-1988)

Table 1. Recorder System Performance

The ER90 drive uses the enhanced IPI physical interface (ANSI X3T9/88-82) and the IPI-3 Magnetic Tape Command Set (ANSI X3.147-1988)at the logical interface level. The enhanced IPI physical interface can sustain transfer rates commensurate with the basic transport performance. A second enhanced IPI interface port can be added to allow a separate master-slave path to another server. A large internal buffer (approximately 60 Megabytes) has been incorporated for rate smoothing to minimize recorder start-stop sequences.

EMASS Software

The EMASS server stores files in an extended UNIX File System (UFS). EMASS software is divided into separate components as depicted in Figure 4. These components are the user interface, the event daemon, the migration manager, the file mover, and the physical device manager. All EMASS software executes as UNIX processes at the application level. All UNIX kernel enhancements/modifications were accomplished by CONVEX and are included in ConvexOS™ 9.0.

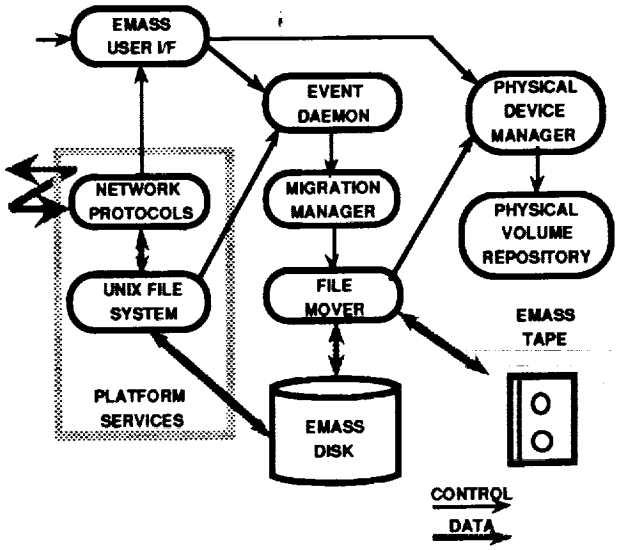


Figure 4. Overview of EMASS Software.

Users have two methods to gain access of EMASS migration services. One method is through a user interface front-end which provides migration override control to end-users. The other method is through direct access of the CONVEX UFS. This second method provides transparent access to EMASS file migration services to both local and networked users.

In order to provide the ability to transparently migrate files, CONVEX has upgraded their ConvexOS to allow the UFS to provide notification of selected critical file system events to a user-level event daemon. This modification is similar to those made by the BRL/USNA Migration Project (BUMP)[3], a joint development of the US Army Ballistic Research Laboratory and the US Naval Academy. The EMASS event daemon receives file events and forwards them to the migration manager. The migration manager collects this information. When migration policy is triggered, the migration manager will select files for migration and forward the list of selected files to the file mover.

The movement of files from an EMASS server disk to magnetic tape and from magnetic tape to an EMASS server disk is controlled by the file mover. For each list of files to migrate, a file mover process is created to perform the read and write operations. The file mover design provides for the addition of software routines to support new media types.

The final major EMASS software component is the physical device manager. The physical device manager provides a standard interface for tape movement services to all other EMASS software components. The physical device manager will translate a generic tape movement request into the format required by the target physical volume repository (PVR). The translated request is then sent to the PVR for processing. The physical device manager will later receive the results from the commanded PVR. The results are then placed into a generic format and sent to the process that requested the tape movement. Additional PVRs will be supported by the addition of software modules to the physical device manager.

DataClass™

Before describing EMASS software in more detail, a discussion of the abstraction known as DataClass™ is required. The file systems that are to be provided EMASS migration services are subdivided at specific points in the file system tree structure by identifying those directory point(s) beneath which all files are to be managed alike. These directory point(s), which are referred to as migration directories, are what define each DataClass. When a new directory point is added to a DataClass, the event daemon will request the UFS to associate the directory and all files below it (both present and future) with the event daemon.

Figure 5 depicts a DataClass to migration directory relationship. The directory /test/dick/special is the only migration directory in DataClass SPECIAL. All files beneath /test/dick/special will be managed together. The DataClass PURPLE contains all files under the directories /prod/blue and /prod/red, but none under /prod/green, showing that some directories at a certain level may be excluded from a DataClass. All files under /test/jane and /test/dick/public belong to DataClass TESTERS. This illustrates that the assignment of migration directories to DataClass is not restricted to a certain level in the tree structure. In fact, migration directories from different file systems may be in the same DataClass. Also, a file system can be mounted onto a mount point underneath a migration directory, for example /test/jane/dir1.

The definition of DataClass is key to site administration. Migration policy parameters are configurable on a DataClass basis, thus providing the EMASS administrator with a great deal of control over the behavior of the EMASS system. Time interval between policy application, time required on disk prior to migration, and desired time for migrated files to remain cached on disk are examples of DataClass based migration policy parameters. Quotas for tape utilization (both

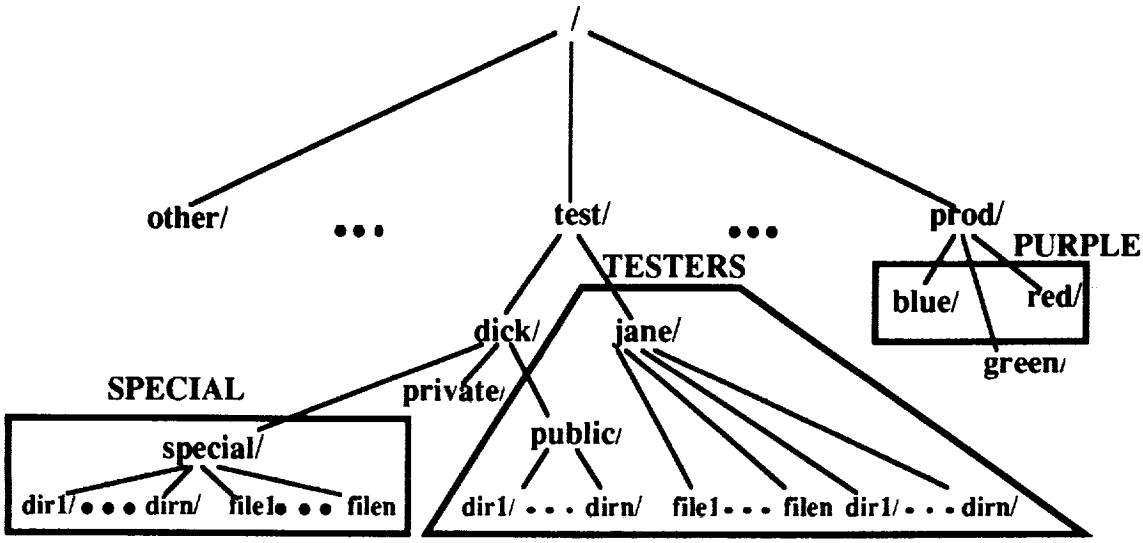


Figure 5. DataClass Example.

a warning limit and a hard limit) are also kept on a DataClass basis. DataClass based parameters are kept in the INGRES database, so tuning can be easily done while the EMASS system is active.

EMASS software also uses DataClass as the means to segregate files on tape. All files on a tape will be of the same DataClass. This provides a level of physical security for those sites which might require it. This segregation also ensures that retrieval of files from different user groups (as defined by DataClass) will not collide trying to access the same physical tape.

EMASS Interfaces

A key concept of the EMASS architecture is that it will provide multiple archival file storage choices to the users of various networked client systems. EMASS supports connectivity over industry standard interfaces including Ethernet, DECNet, FDDI, HYPERchannel, and UltraNet. This provides connectivity from the smallest workstations to the largest supercomputers. The industry standard transfer protocols available to the user include the File Transfer Protocol (FTP)

from the TCP/IP family, the Network File System (NFS™) as defined by Sun Microsystems, and the UNIX utilities UNIX to UNIX copy (UUCP) and remote copy (rcp). This support is provided by placing the EMASS interface under the UFS. The EMASS system will receive notification of all managed file system events for files in every DataClass. Migration services are therefore provided for any connectivity available on the EMASS server system to the UFS. Thus, as new connectivity options are offered by the EMASS server vendor, the EMASS system will automatically also provide support.

Hierarchical Data Storage

EMASS provides three levels of data storage. These three levels of hierarchical storage are EMASS server disk, robotically managed tape, and human managed tape. When files are placed in a DataClass, the residency is on server disk. The EMASS migration policy will schedule placing the files onto tape based on the migration rules defined for its DataClass. The user can preempt the migration policy by giving the EMASS system a directive to migrate specific files to tape immediately.

When the migration policy is executed for a DataClass, all files in that DataClass which are not solely on tape are examined. If the time since last modification (or the time since retrieval from tape if unmodified) of a file is greater than that specified in the policy for that DataClass, then that file's data is placed in the staging directory. If an up-to-date copy of the file is not on tape, the file is added to a list of files to be migrated. When all files in the DataClass have been examined, the list of files to migrate is forwarded to the file mover.

To store files on tape, the file mover first allocates tape(s) through the physical device manager. The files in the list are next migrated to tape. The file mover will record the location of the tape copy of each file in the INGRES database as they are successfully written to tape. The disk copy is left intact as a cached copy which will later be removed from disk when either 1) the length of time since migration has exceeded its DataClass defined limit or 2) the file system requires additional disk data blocks and it is the oldest staged file on disk. When the file mover has completed migrating files to tape, the tape and drive are released back to the physical device manager.

Not removing disk data blocks from the staging directory on strictly a first-in-first-out basis provides additional flexibility. A file system can be divided into DataClasses which can have different access requirements. Each DataClass will have its own disk retention period, so one

portion of a file system can have data that is not cached on disk as long as another portion. One DataClass thus can be set up to never free disk data blocks except when required to provide needed free disk space for the file system.

When the UFS receives a request for data blocks for a file which is not currently on disk, the requesting process is suspended and the EMASS event daemon is notified. The event daemon forwards that notification to the migration manager which immediately instructs the file mover to retrieve the requested file to disk. The file mover requests the tape containing the file be mounted and copies the file to disk. The requesting process is now allowed to continue processing. The EMASS system maintains knowledge of the tape copy. If the disk copy is unchanged, the file will not be re-migrated by the migration policy.

The retrieval of a user-directed range of bytes from a file is also available to the EMASS user. This is accomplished much like the retrieval of a complete file. However, the file mover will copy the specified range of bytes into a UNIX file of a different name as specified by the user. Thus, the user can retrieve only the portion of the file of interest, reducing the amount of data brought back.

The EMASS system will also manage tapes that are not under robotic control. This will allow sites to have EMASS management of many more tapes than the robotic system can support. When access is requested for a file that is only on a tape that is not under robotic control, the EMASS system will request the operator to return the tape to active service so that the file may be copied onto disk. The effect to the user is only a longer delay waiting for a tape mount.

Infinite File Life

A mass storage system must provide for the integrity of its client's data. In order to insure that the client's data is always available, the EMASS system has several features to provide safeguards against data loss. These features include automatic Error Detection and Correction (EDAC) monitoring and secondary file copy maintenance.

For every file segment written by an ER90 drive, the drive will automatically perform a read while write comparison. If the data written is not recoverable or to successfully recover the data written required more than a minimum threshold of correction, that segment is automatically re-written to tape by the drive without any action required by the host system. This provides positive assurance that the data written is retrievable without much stress on the EDAC at the time it is recorded.

For every file read by an ER90 drive, the EMASS system will request the drive to return EDAC statistics and if the level of correction was excessive, that tape will be placed in a "suspect" list for system administrator action. The system administrator can then at a later time request the EMASS system to move all files off of the old tape. This provides for refreshing the EDAC encoding for all files that were on the old tape.

To ensure the health of D2 tapes that have not been accessed for a while, the EMASS system provides a tape sniffing service. Tape sniffing is the process of periodic monitoring of tapes that have not been accessed for a length of time defined by the data center. The EMASS system will schedule the reading of sample files from the tape and then examine the EDAC statistics to determine if the level of correction was excessive. If excessive, the tape is placed in the "suspect" list for system administrator action.

As an added measure of protection for tape-based files, secondary file copying is provided. If enabled for its DataClass, files will automatically have a secondary copy maintained on a separate tape. This DataClass feature can be overridden on a file basis, thus allowing the user to request a secondary copy be created when the DataClass default is to not maintain a copy. The user can conversely request the EMASS system not to maintain a secondary copy of a file when its DataClass default is to maintain a copy.

Through the use of automatic EDAC monitoring and secondary copies, the EMASS system provides for the integrity of its client's data. The life of the EMASS client's data can in fact be prolonged well beyond that of any one type of storage media, as the file sniffing service will promote data from one media onto another.

Summary

EMASS software provides a UNIX-based data storage solution with automatic and transparent file migration and retrieval. Data archive centers can be provided with very large (up to Petabytes), expandable, automated data storage systems. These data storage systems connect to high speed networks, providing 24 hour per day accessibility for rapid delivery of requested data in the Space Station and EOS era. File access can be provided to networked users through standard file transfer protocols. A graphical user interface can also be provided. Thus, the client is not required to have special networking software. The implementation of DataClass provides a flexible method for tuning the behavior of the system at each installed center.

Trademarks

EMASS, DataTower, DataLibrary and DataClass are trademarks of E-Systems, Inc.

CONVEX and ConvexOS are trademarks of CONVEX Computer Corporation.

UNIX is a trademark of AT&T.

INGRES is a trademark of ASK Computer Systems.

NFS is a trademark of Sun Microsystems, Inc.

HYPERchannel is a trademark of Network Systems Corporation.

UltraNet is a trademark of Ultra Network Technologies, Inc.

References

[1] S.W. Miller, "A Reference Model for Mass Storage Systems", *Advances in Computers,* Vol. 27, Academic Press, 1988, pp. 157-210.

[2] Tracy G. Wood, "A Survey of DCRSi and D2 Technology", *Digest of Papers,* Proc. Tenth IEEE Symposium on Mass Storage Systems, May 1990, p. 46 (1990).

[3] Michael John Muuss, Terry Slattery, and Donald F. Merritt, "BUMP, the BRL/USNA Migration Project", *Unix and Supercomputers*, 1988.

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.

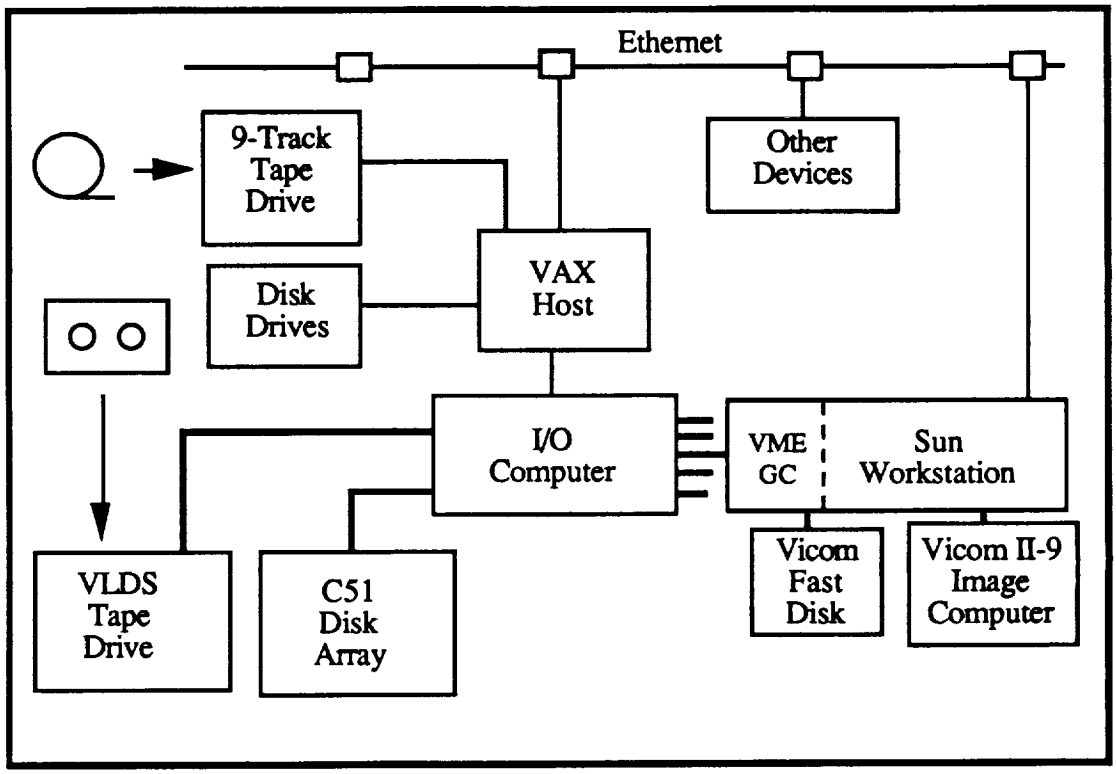


Figure 1. Data Storage and Retrieval System

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.

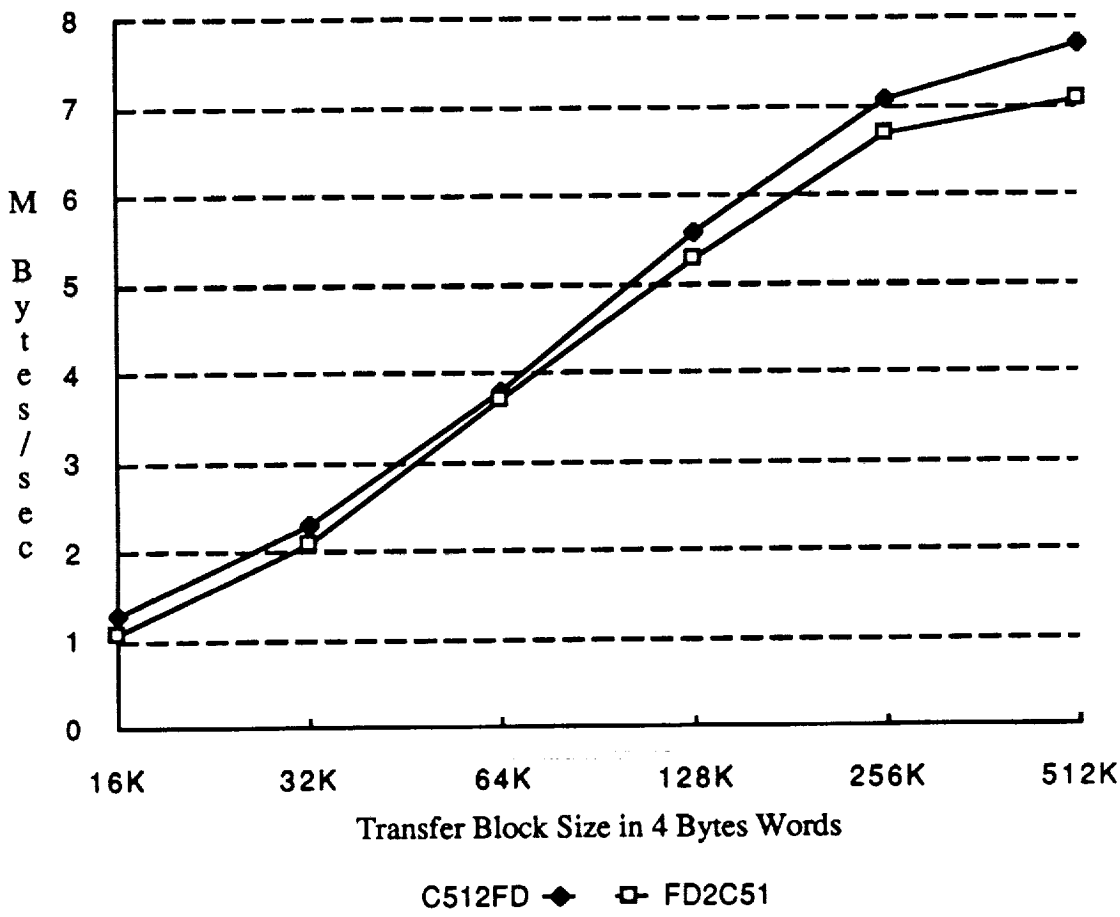


Figure 2. Transfer Rate as a Function of Block Size

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

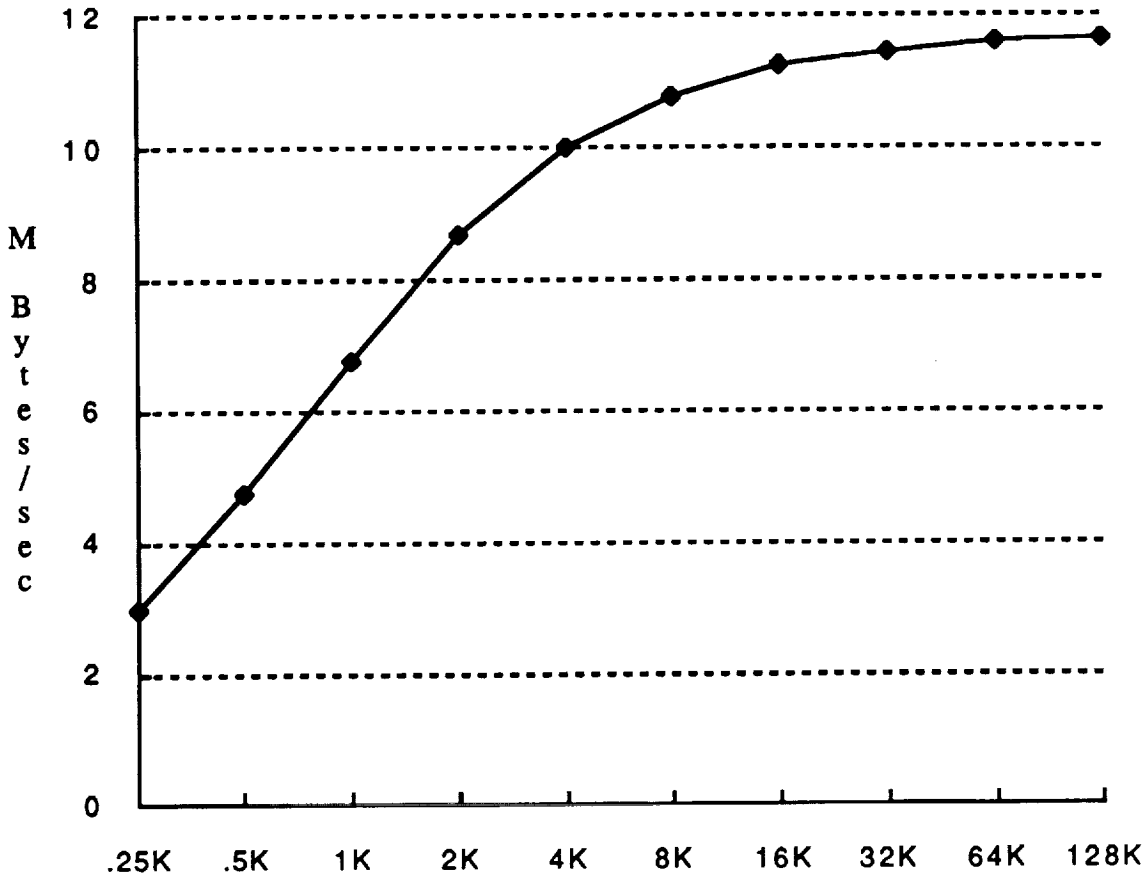


Figure 3. VME Transfer Rate as a Function of Block Size

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

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

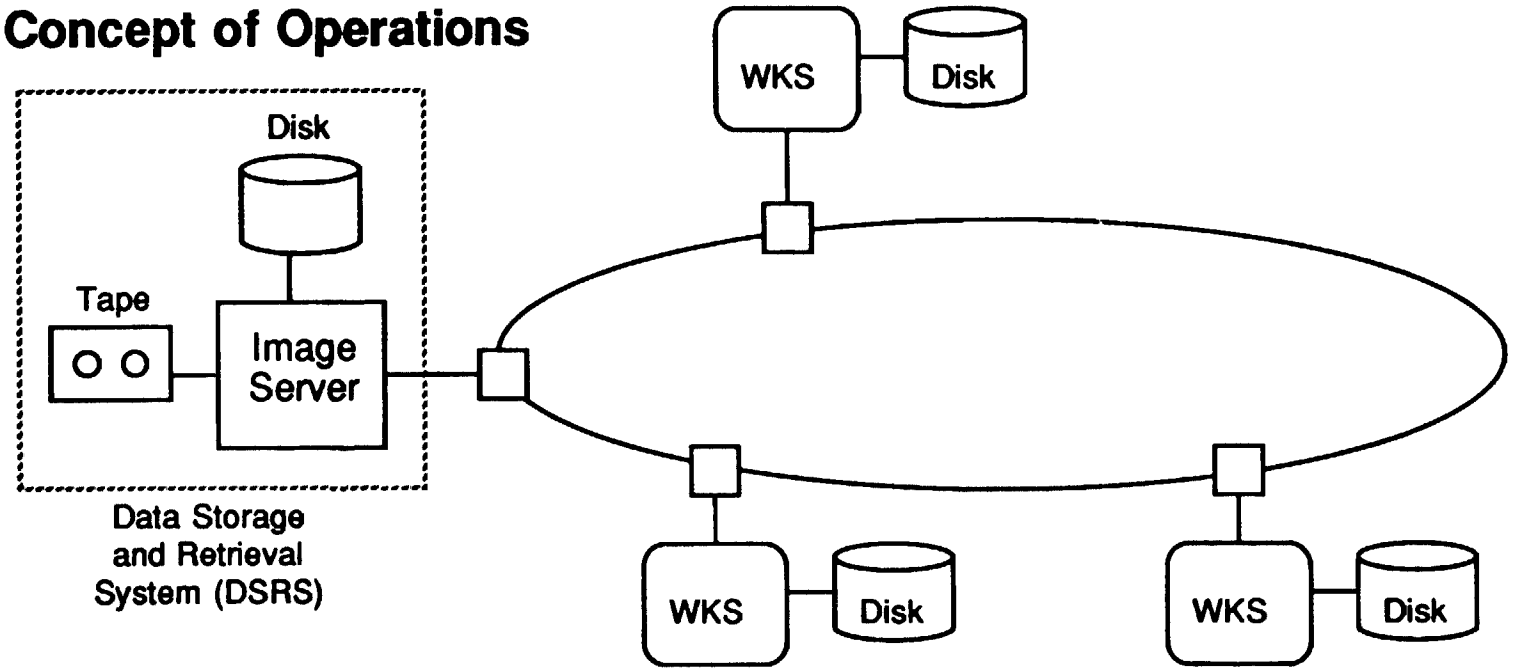
Glen Nakamoto

24 July 1991

MITRE

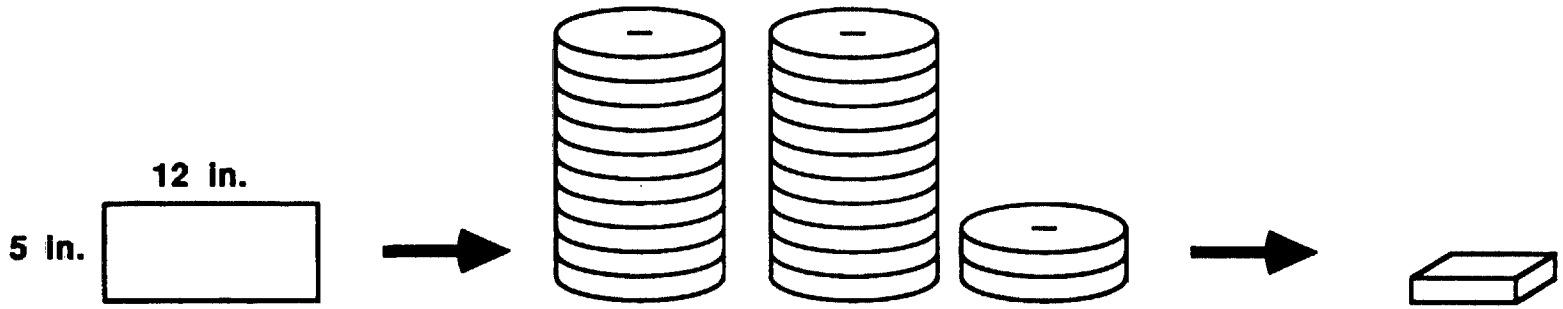
Background

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

A Key Problem



22 Nine-track Tapes

One VHS Tape

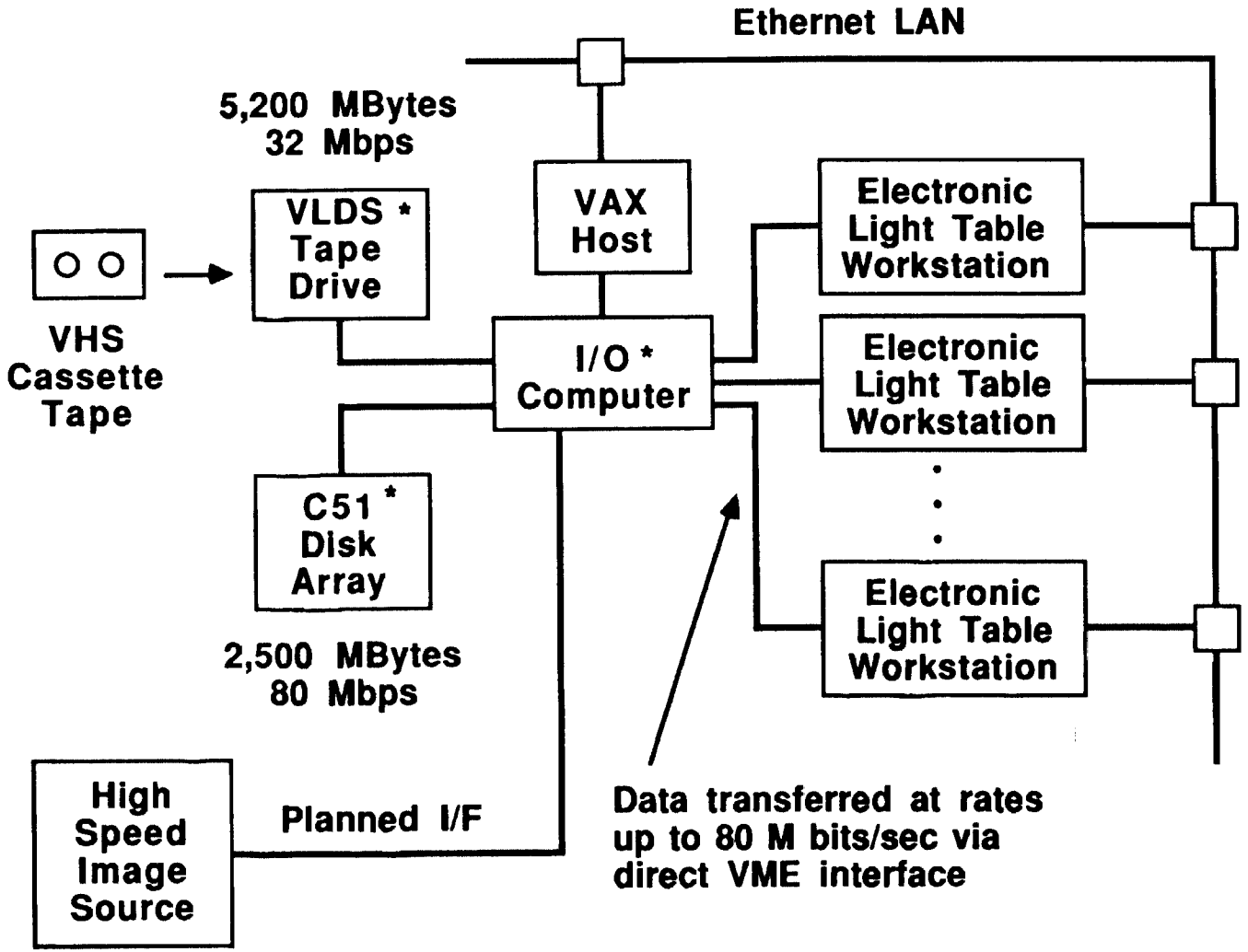
15 Hours Input Time → 10 Minutes Input Time

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

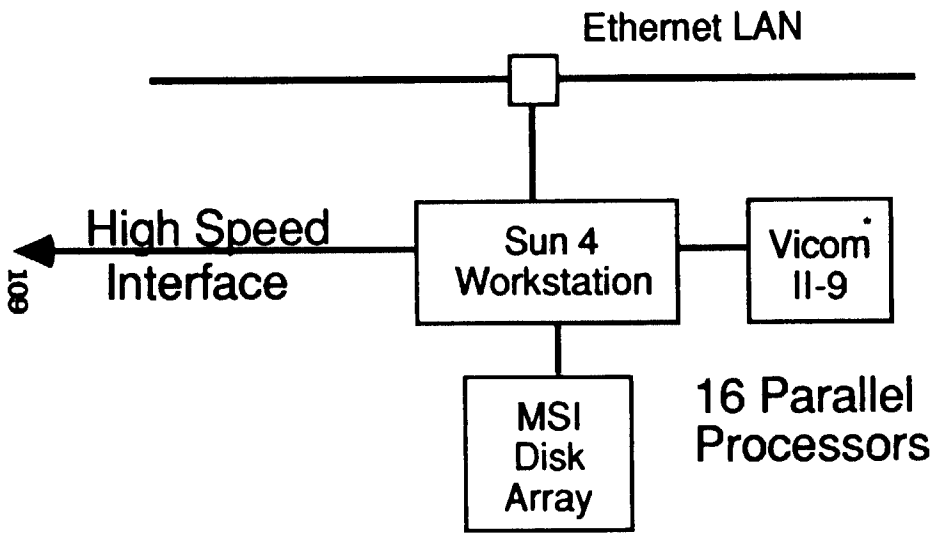
DSRS Architecture



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

MITRE

Electronic Light Table Workstation



6.0 Gbyte capacity
64 Mbps transfer rate

- Electronic Light Table Functions
- Integrated Database -

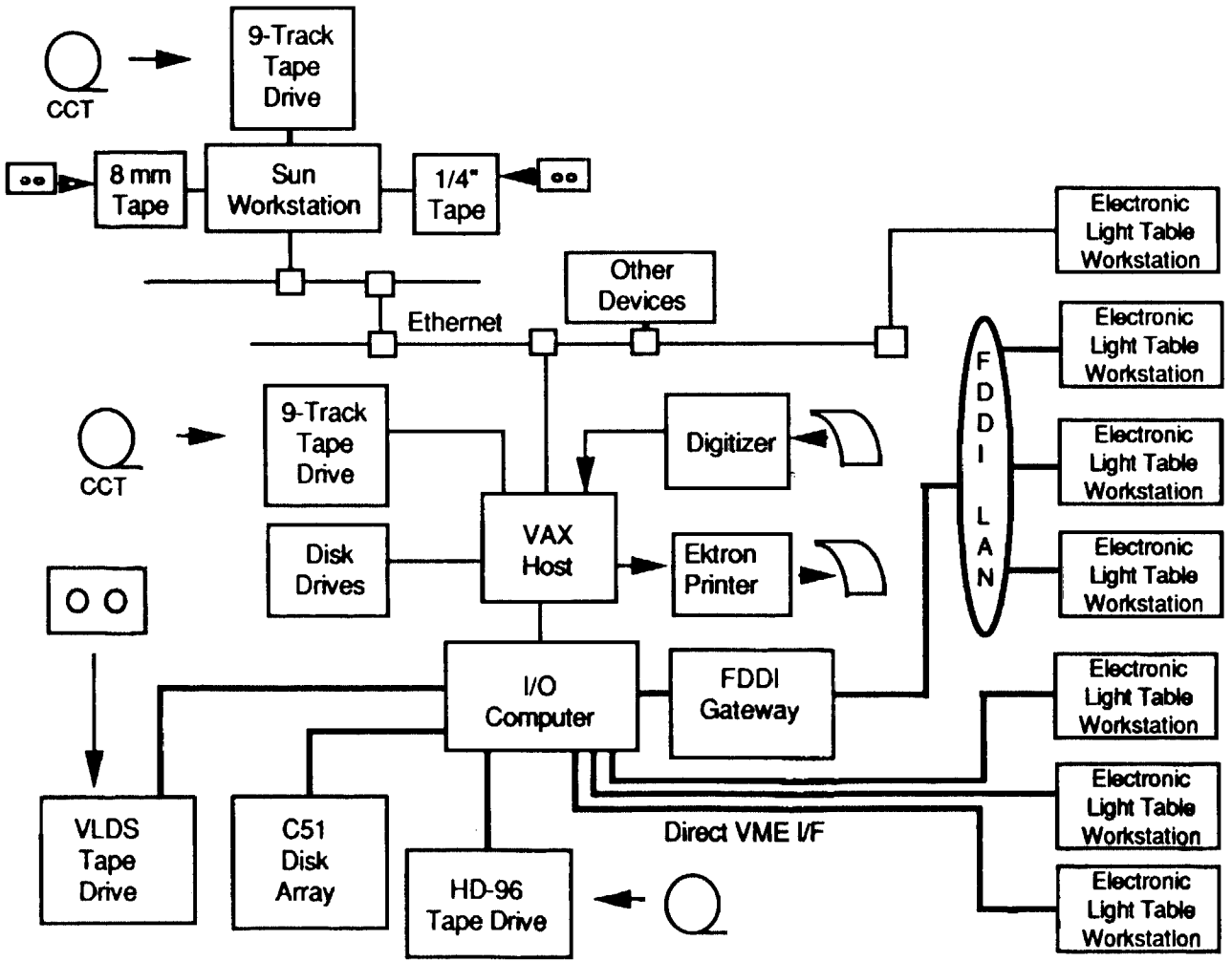
  Text, Graphics, Imagery
- Large Image File Manipulation
- Real-time Decompression
- 2K by 2K display size (search)
- High Speed Image Transfer

  (64 Mbps)
- Off-the-Shelf Hardware

• Previously Pixar

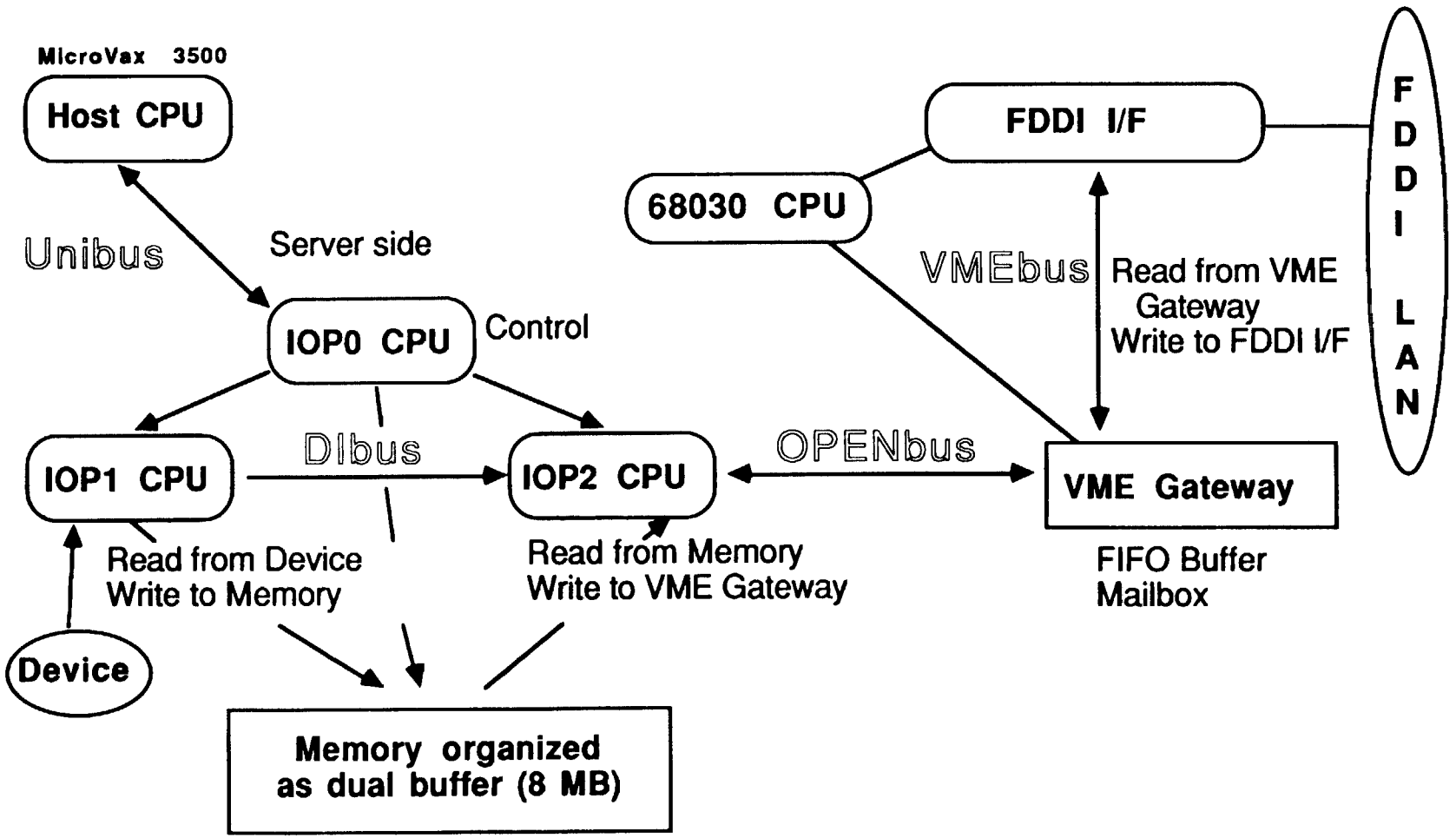
MITRE

Image Server Configuration



MITRE

Software Architecture



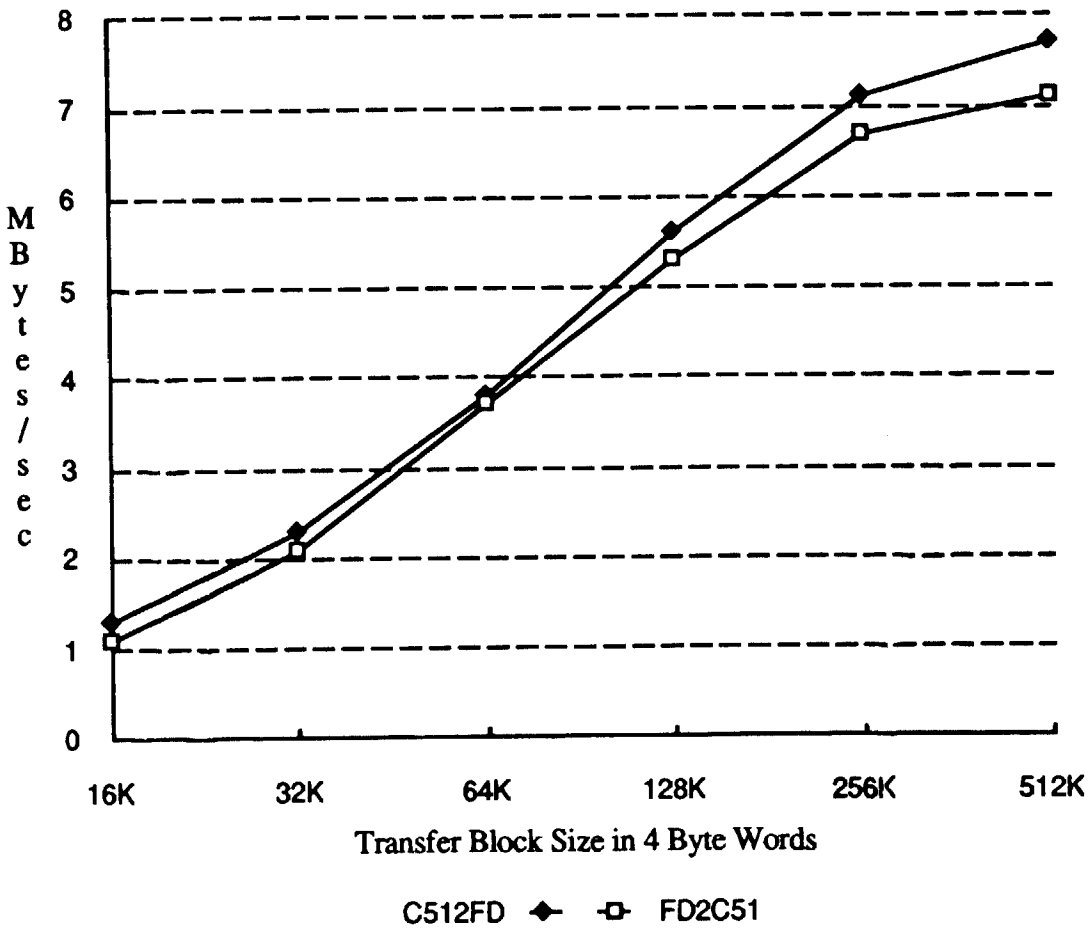
Measured Transfer Rates

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:

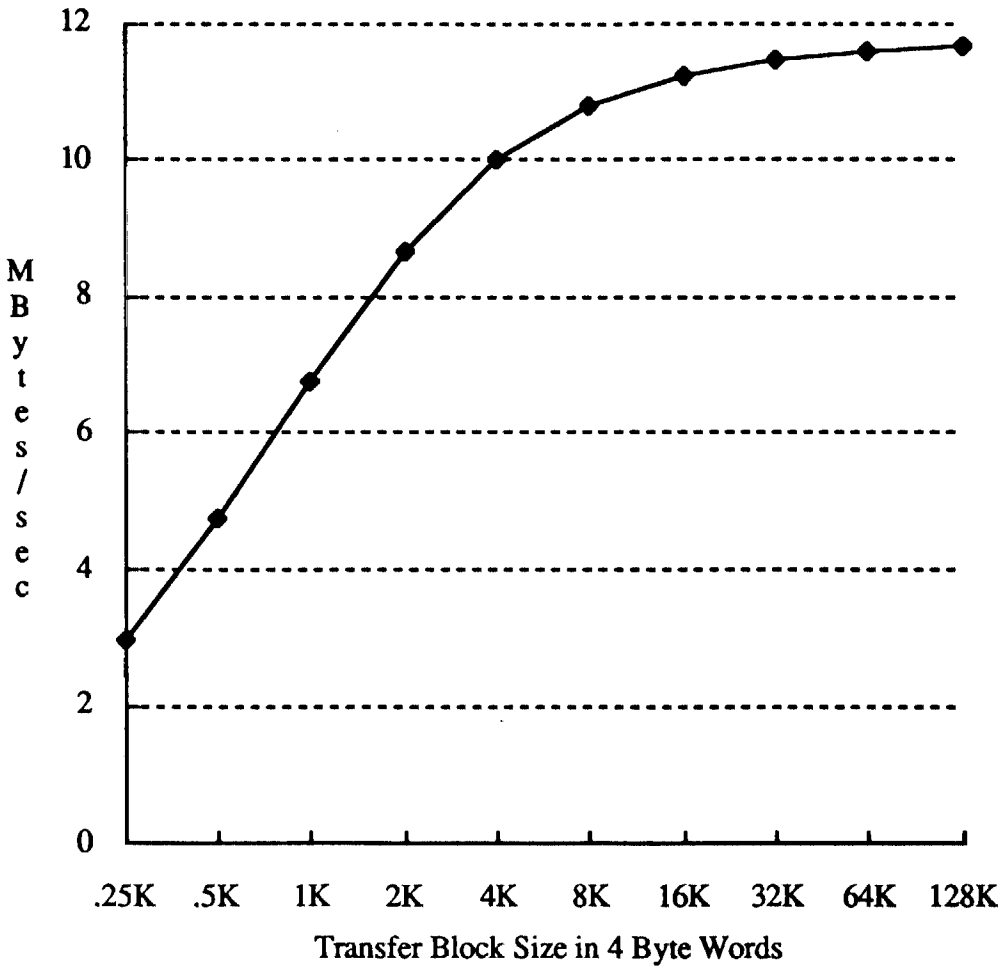
Sun memory to Sun memory transfer with no network contention.
SunOS 4.1 (Sun 3/260 with 8 MB memory).
Sun provided TCP/IP software.
UltraNet used TP4/IP with board level protocol processing.

C51-PFD: Transfer Rate vs. Block Size



MITRE

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

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

NSSDC Mass Storage Workshop

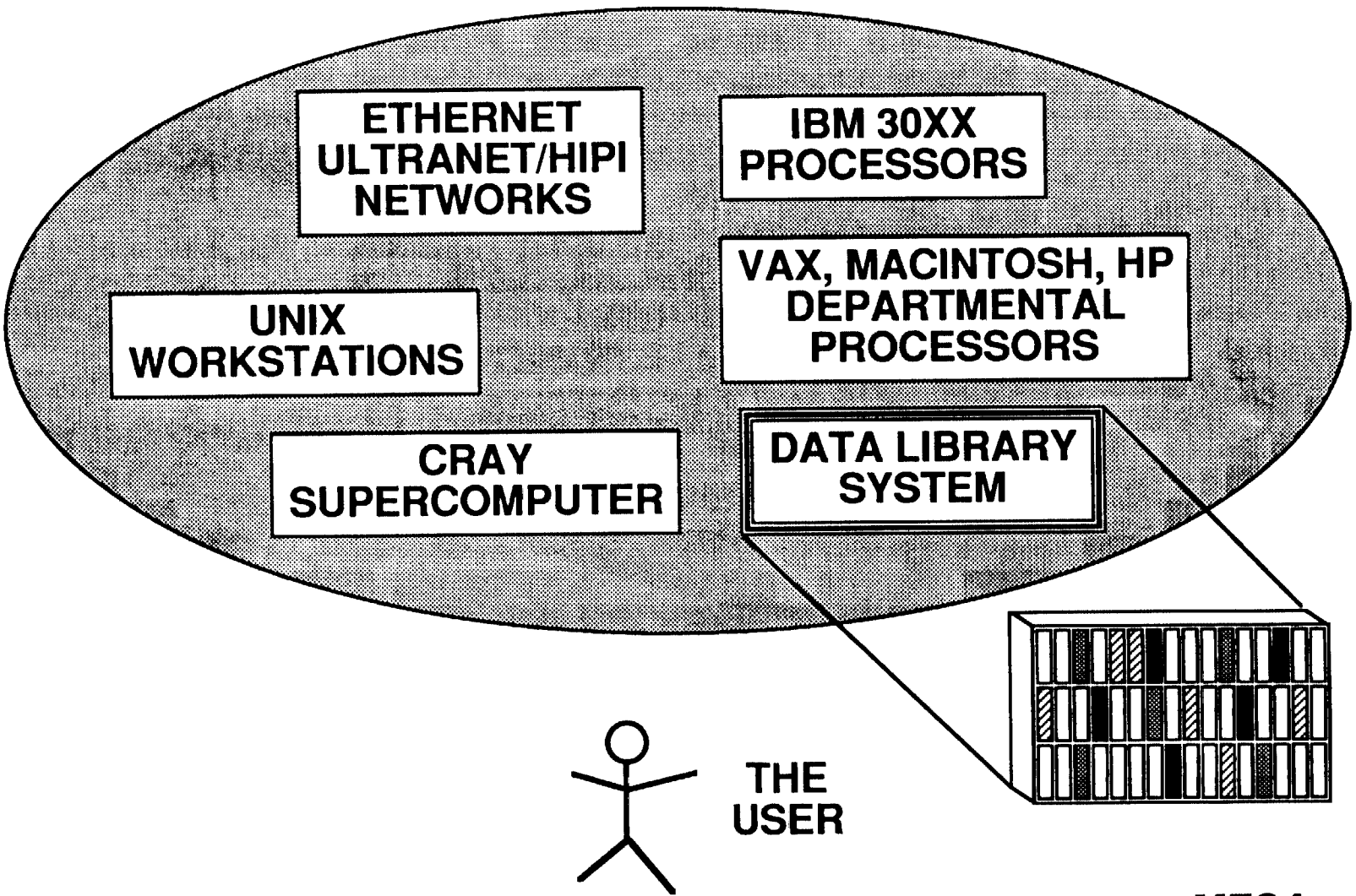
July 23-25, 1991

Data Archiving

Mesa Archival Systems, Inc.

MESA
Archival Systems, Inc.

Computing Environment



MESA
Archival Systems, Inc.

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH (NCAR)

Atmospheric and oceanographic research

Inititiator of IEEE Storage Model

Status

- Operational since 1986
- 4,000 users
- 82,000 3480 cartridges 11/90
- ~15 TB, growing at 6 TB/year with Y/MP

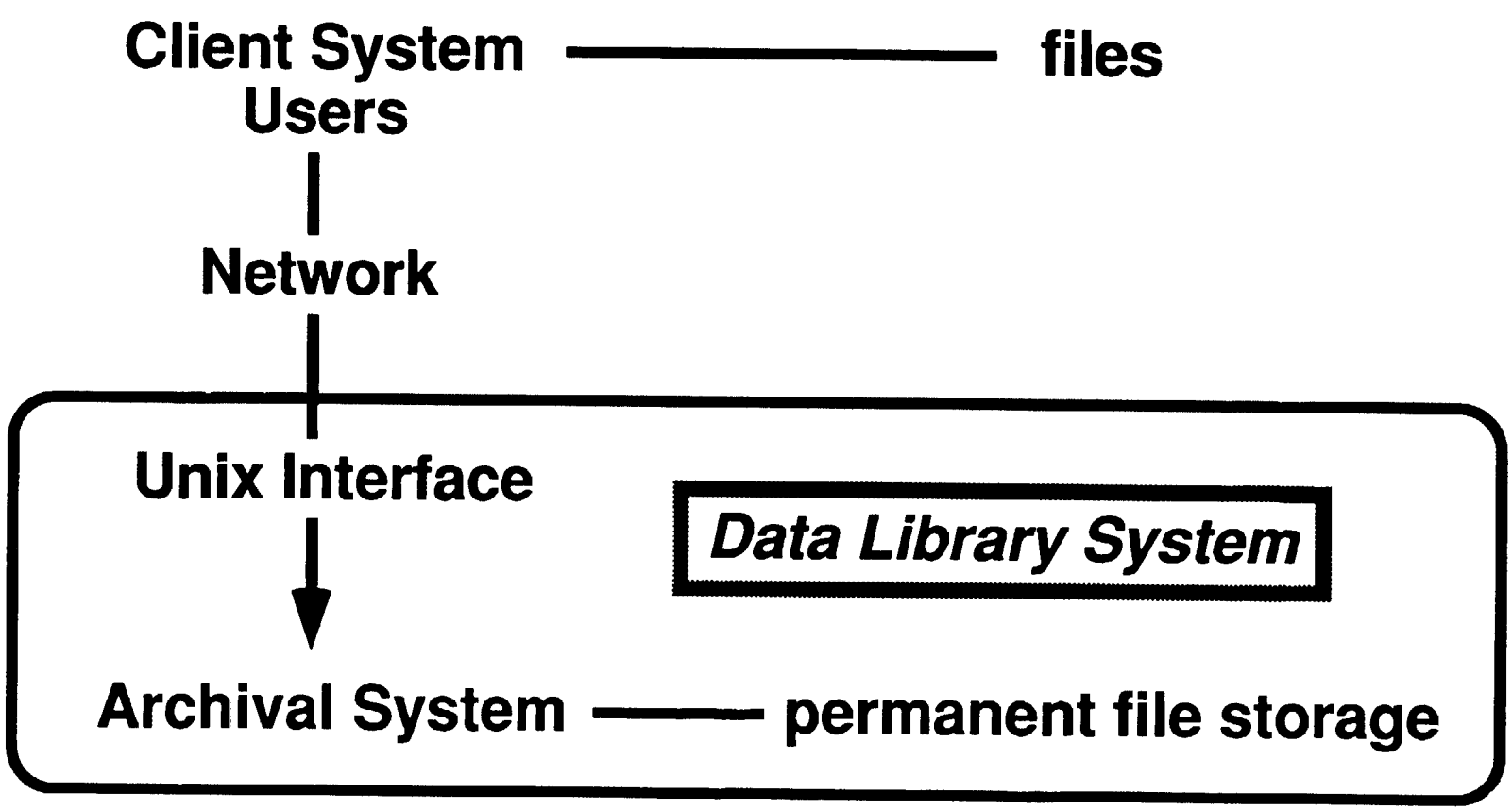
Product Goals

Designed for change

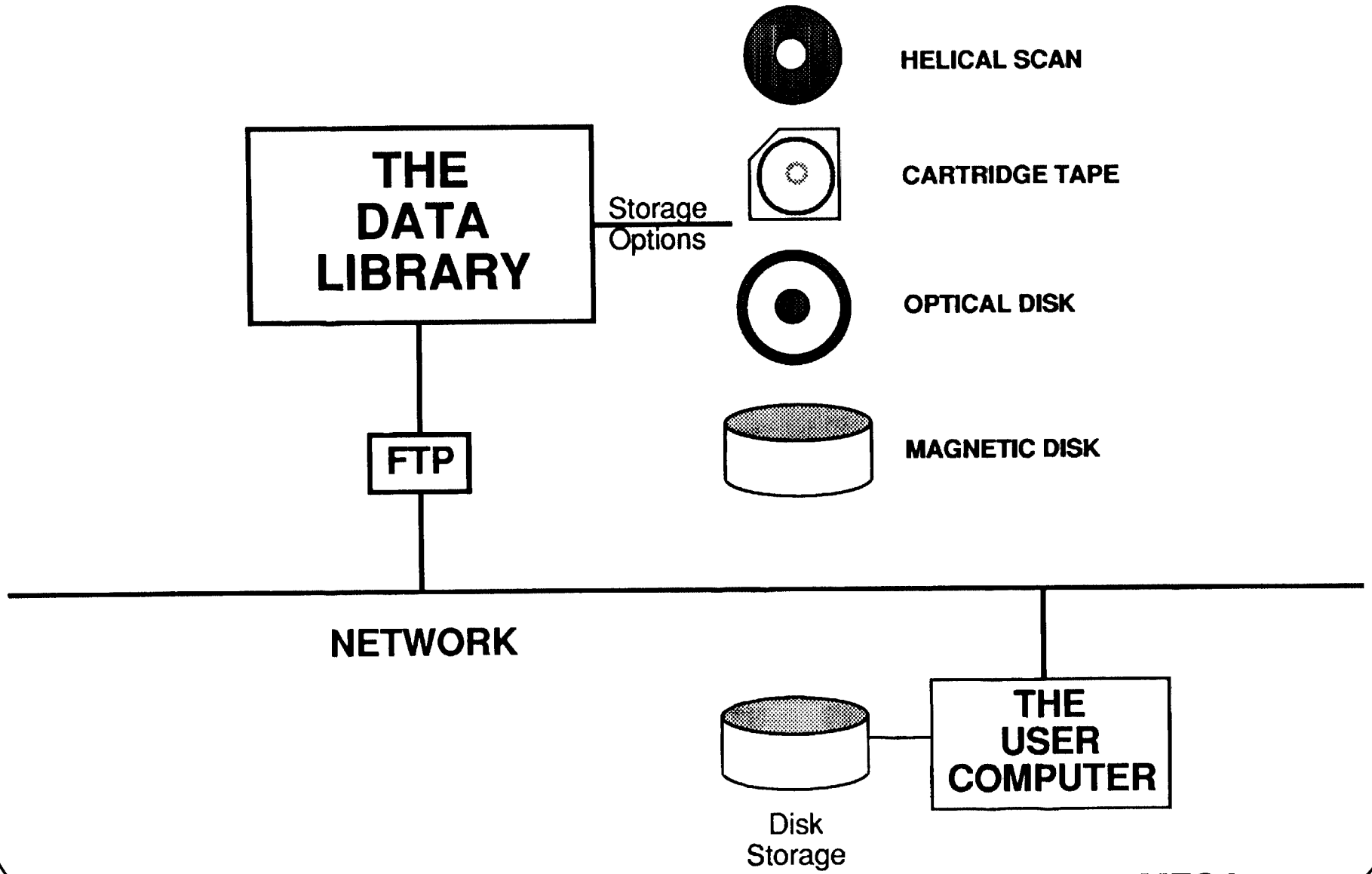
System availability

Commercial support

Standard products


Client System
Users
files
Network
Unix Interface
Data Library System
Archival System
permanent file storage

DLS Environment



MESA
Archival Systems, Inc.

Client System Commands

FTP Interface

- *GET*
- *PUT*
- *DIR*
- *LS*
- *RENAME*

Other

- *USER ACCESS*
- *IMPORT / EXPORT*

Networks

Protocols

- FTP, User Access
- TCP/IP, NETEX

Networks

- Ethernet
- Ultranet
- HYPERchannel

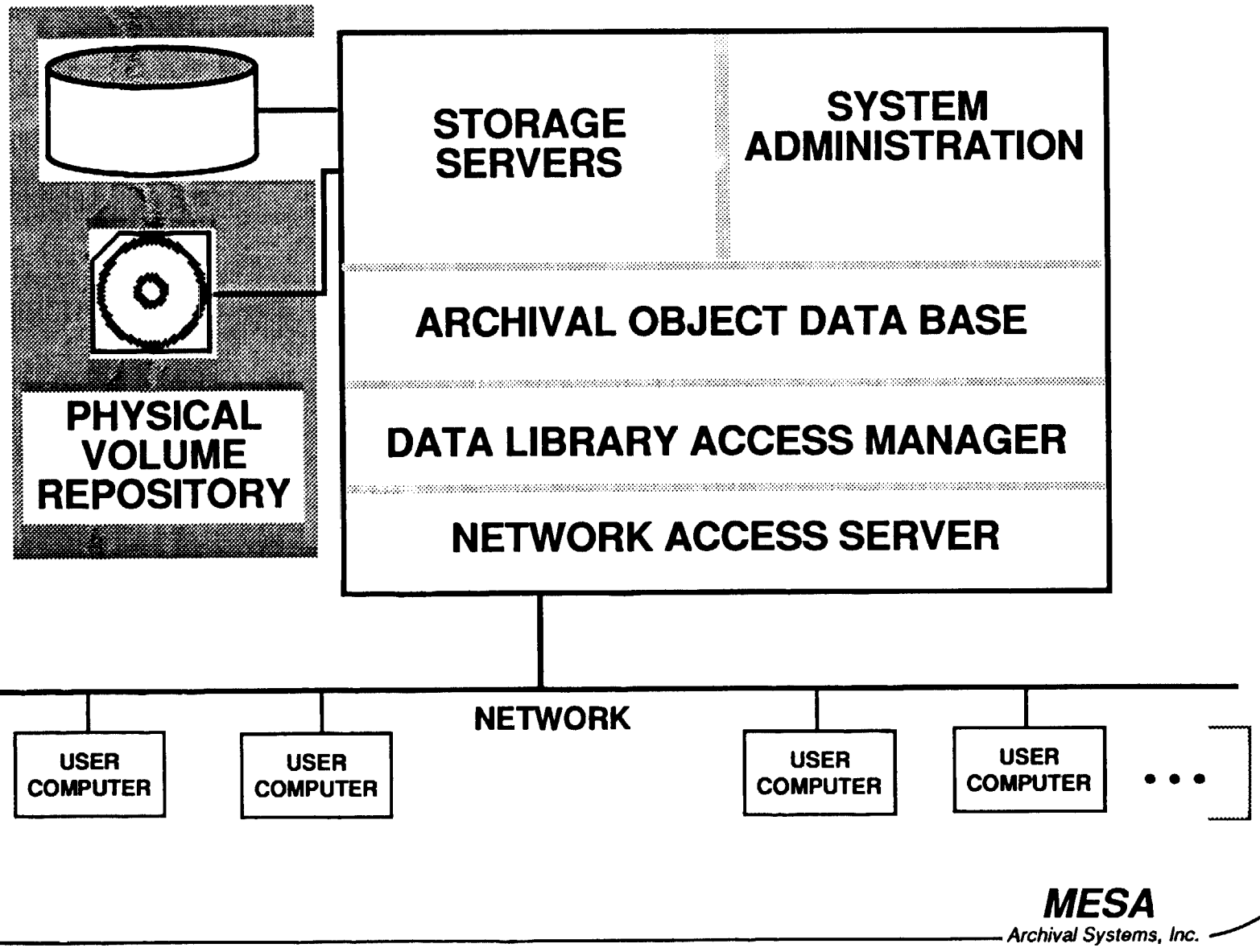
Archival Devices

- IBM 3480/90 Cartridge Tape
- STK 4400 Cartridge Tape Robot
- MTC 5400 Automated Tape Library
- DataWare Optical Disk
- Masstor Helical Scan Tape

DLS Features

- Modular implementation
- Multiple media support
- Resource accounting
- Security (Client - POSIX, System - MVS)

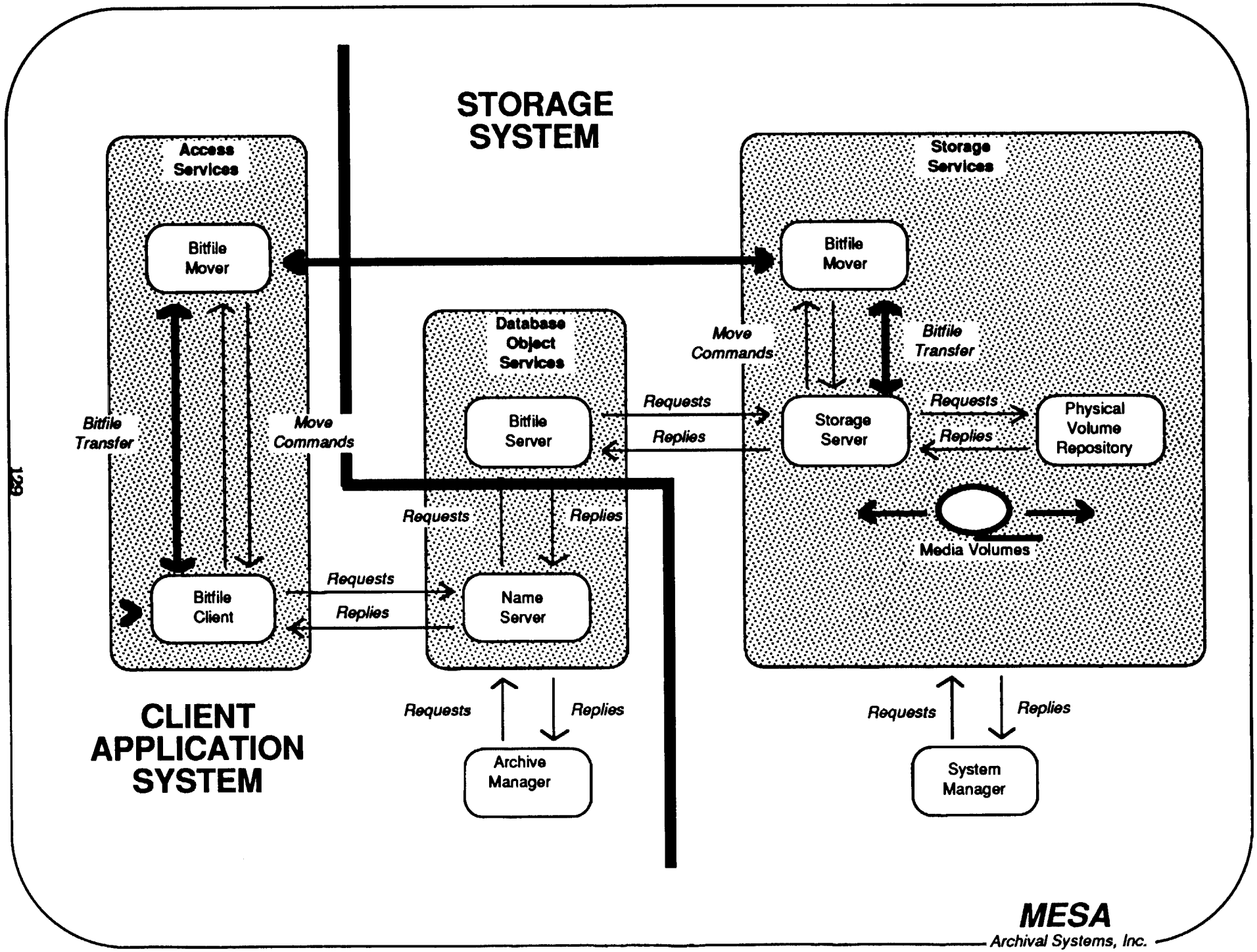
Data Library System



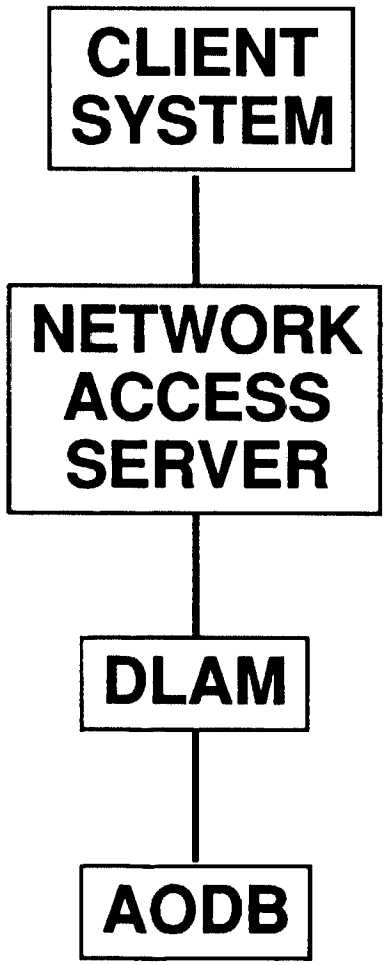
MESA
Archival Systems, Inc.

Data Library System

- Network Access Servers
  - FTP or User Access
  - Unix File System appearance
  - Gateway to DLS
- Archival Object Data Base (AODB)
  - Powerful Facilities
  - Object Orientation
- Storage Servers
  - Uniquely Mounted Media
  - Variably Mounted Media
- System Administration

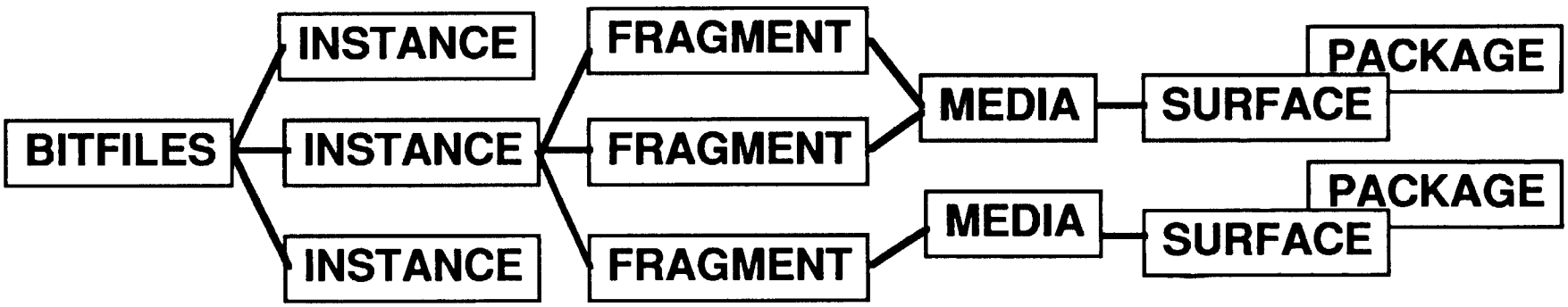

STORAGE SYSTEM
Access Services
Bitfile Mover
Bitfile Transfer
Move Commands
Bitfile Client
Database Object Services
Bitfile Server
Requests
Replies
Name Server
Archive Manager
Storage Services
Bitfile Mover
Move Commands
Bitfile Transfer
Storage Server
Physical Volume Repository
Media Volumes
System Manager
CLIENT APPLICATION SYSTEM
MESA
Archival Systems, Inc.

Client Application System



USERS
GROUPS
ACCOUNTS
DIRECTORIES
BITFILES

Storage System



Multiple Mass Storage Devices

Differing User Storage Needs

- Access Speed
- Permanent Retention
- Cost/MB
- Interchangeability

Continuing Product Evolution

- 3480 → 3490

New Mass Storage Devices

- D1/D2
- Optical tape

System Growth

- Additional Archival Devices
- Additional Connectivity Products
- Additional Operating System Support
- Transparent to Client System

NETWORK ACCESSIBLE MULTI-TERABYTE ARCHIVE

P-11

Sp-82
12/1946

NSSDC CONFERENCE
GODDARD SPACE FLIGHT CENTER
JULY 24, 1991

Fred Rybczynski
Metrum Information Storage

ROTARY STORAGE SYSTEM (RSS)

- T-120 Super-VHS media with 10.4 GB per cartridge
- Very Large Data Store (VLDS) tape drives
  - Sustained 1.0 MB/S with Read-After-Write
  - 1E-13 Bit Error Rate w/ error mgt software
  - 45 sec average file access time
- Robots with on-line capacity to 6.2 TB
- Bar code reader
- Data management computer
  - Data referenced by file name
  - Files organized by hierarchical directories
  - File and directory access protections
- Network accessible

Metrum
Information Storage
[031891.sh]

METRUM INFORMATION STORAGE SYSTEMS BASED ON STANDARDS

MEDIA
- World-wide VHS media standard
- T-120 half-inch tape cartridge
- Billions of cartridges sold

TAPE DRIVE
- SMPTE format VHS transport
- Millions of transports sold
- Commercial grade

ROBOT
- Commercial broadcast industry

Metrum
Information Storage
[082491.sh]

METRUM INFORMATION STORAGE STANDARDS - concluded

BAR CODE READER
- Standard formats

INTELLIGENT DATA MANAGER (COMPUTER)
- UNIX based
- POSIX compliance

NETWORK CONNECTIVITY
- Ethernet TCP/IP and DECNET
- HYPERCHANNEL
- FDDI
- GOSIP
- NFS

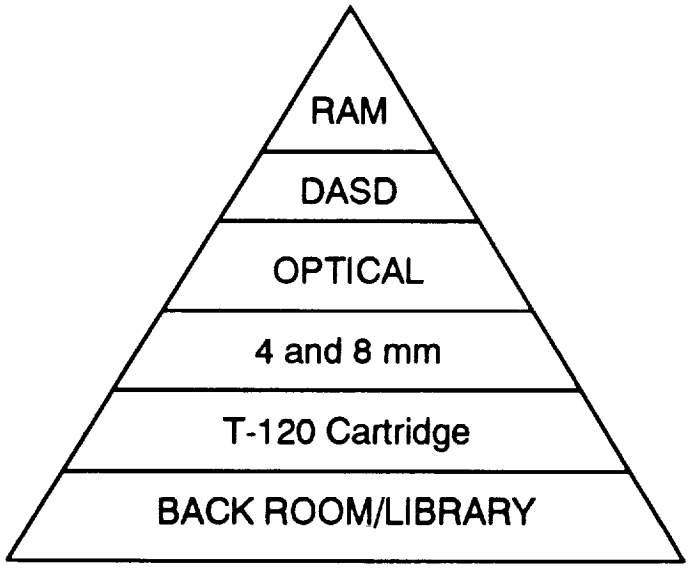
MANAGED DATA STORAGE GOALS

- Maximize data per unit volume, balanced by
  - Cost/MB of Media
  - Cost/MB of Technology
  - Required data rates
- Optimize media access
  - Unattended media access
  - Media ID verification
  - Speed of loading
- Provide ease of access to the data
  - Easily identify data to be stored/retrieved
  - Concurrent access from multiple computer systems

DATA SOURCES

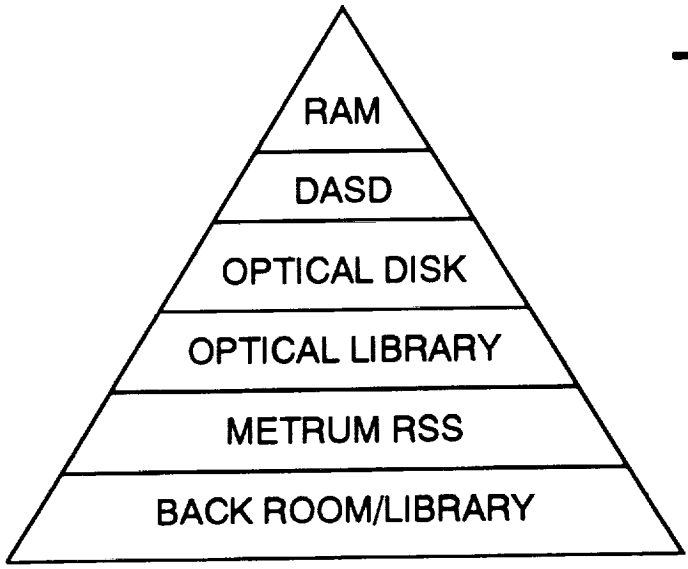
- Data acquisition systems
- Data transcription systems
- Local computer file system
- Remote computer file system

STORAGE HIERARCHY
TYPICAL MEDIA PRICE/MB



	Price/MB
RAM	$ 1000.00
DASD	$ 10.00
OPTICAL	$ 0.15
4 and 8 mm	$ 0.01
T-120 Cartridge	$ 0.003
BACK ROOM/LIBRARY	$ 0.05

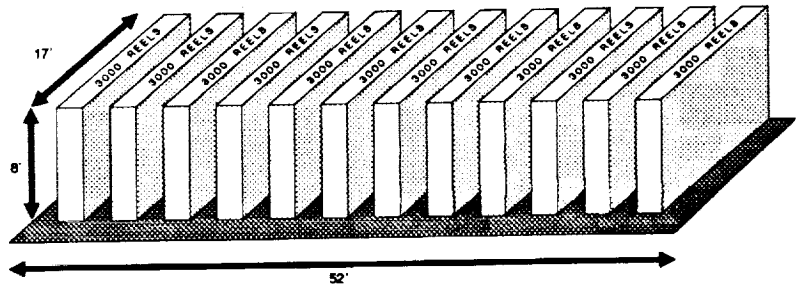
STORAGE HIERARCHY
TYPICAL ACCESS TIMES



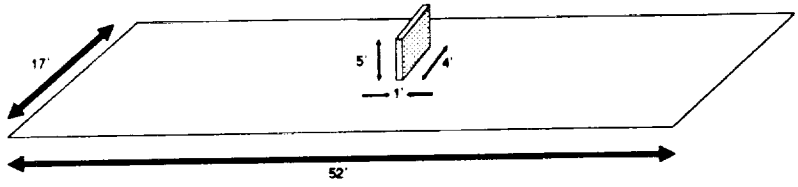
Access Time
1 micro sec
15 msec
250 msec
10+ sec
< 60 sec
min to hrs

Metrum
Information Storage
[110190 sh]

VOLUME COMPARISON
6 Terabytes



9-Track, 6250 bpi
2400 Ft, 170 MB
35,000+ Reels
850+ Sq Ft



T-120 Cartridge
10.4 GB/Cartridge
600 Cartridges
4 Sq Ft

Metrum
Information Storage
[040191 sh]

VOLUME COMPARISON

One T-120 Cartridge (10.4 GB) Equals

- 61 nine-track reels at 2400 ft/reel, 6250 bpi and 170 MB/reel
- 50 cartridges of 3480 at 200 MB/cartridge

One Cubic Foot of Storage Can Contain

- 12-15 nine-track reels
- 60 cartridges of 3480 (equals 70 nine-track reels)
- 30 T-120 cartridges (equals 1,800 nine-track reels)

RSS FEATURES

- 6.2 TB in < 20 sq ft
- System cost < 8.5 cents per MB
- Media cost < 0.3 cents per MB
- Unattended operation
- Network accessibility
- Read-After-Write reliability (1E-13 BER)

RSS DATA ACCESS

- Archive system accessed via network
- Access authorization screening (UserID & Password)
- Entire archive looks like a single huge disk
  - Data accessed by file name
  - Multi-level hierarchical directories
  - Selective directory & file protection
- File access is random
- Avg file access is 45 sec for loaded cartridge
- Cartridge load in < 8 seconds
- Intelligent queue management

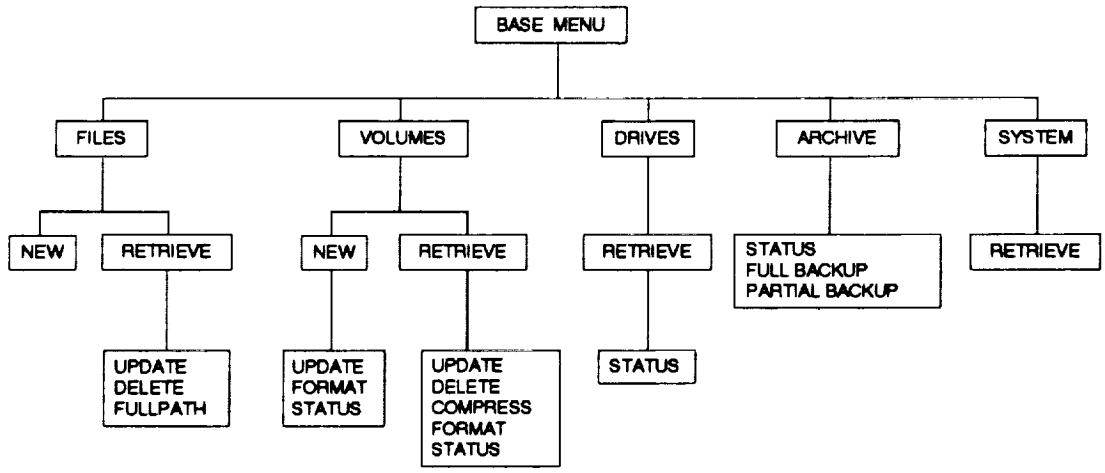
Metrum
Information Storage
[040191 slr]

RSS DATA MANAGEMENT

- Database Management System
  - 90 million files
  - 65,000 cartridges
  - Maintains system performance log
  - Comprehensive report generator
- Monitors system events
  - Errors are categorized by severity
  - Multiple notification targets, including e-mail
  - Incorporates recovery algorithms
- User data can be directed to specific cartridge(s)
- Admin functions available at console & via network

Metrum
Information Storage
[031891 slr]

RSS ADMINISTRATION
MENU MAP



Metrum
Information Storage
[052191-dr]

RSS ADMINISTRATION
BASE MENU

AMASS Administration

Version: AMASS/2.12

Files(1) Volumes(2) Drives(3) Archive(4) System(5)
Wipe(0) Help(PF2) End(PF3) Quit(PF4)>:

Metrum
Information Storage
[031601-dr]

RSS ADMINISTRATION FORMS INTERFACE

FILES / RETRIEVE MENU

AMASS Administration

Name:	1990_wp_files		
User:	100	Record:	32
Group:	101	Parent's Record	15
Mode:	rwx------	Volume:	
Type:	Directory	Volume Group:	3
Size:		Starting Block:	
Archived:	01-13-91 13:18:22	Bad Blocks:	
Accessed:	01-13-91 13:18:22	Nr of Errors:	

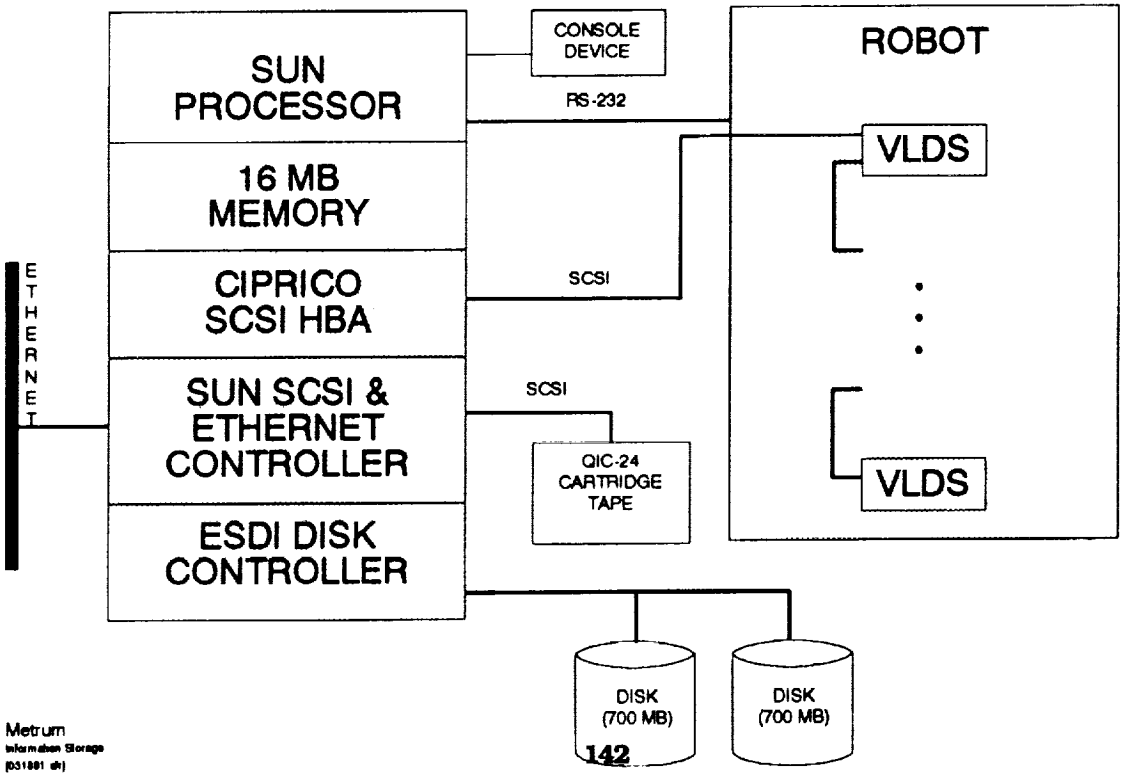
3 of 13

Files/Retrieve

Update Delete Fullpath

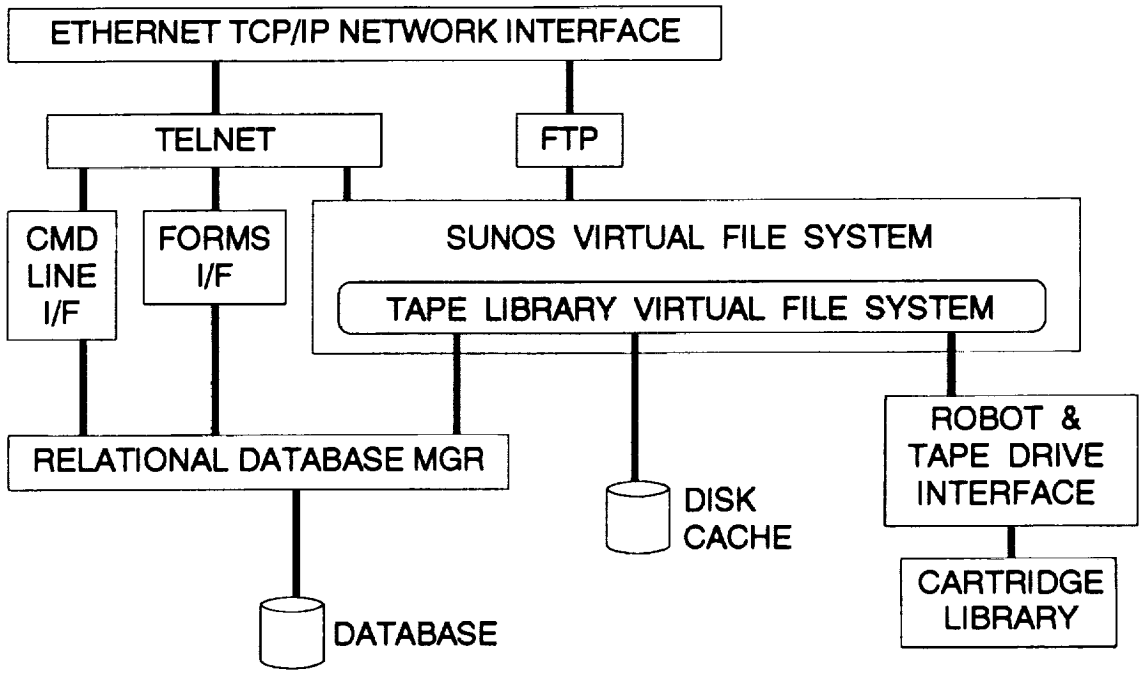
Metrum
Information Storage
[031891.sh]

RSS HARDWARE ARCHITECTURE



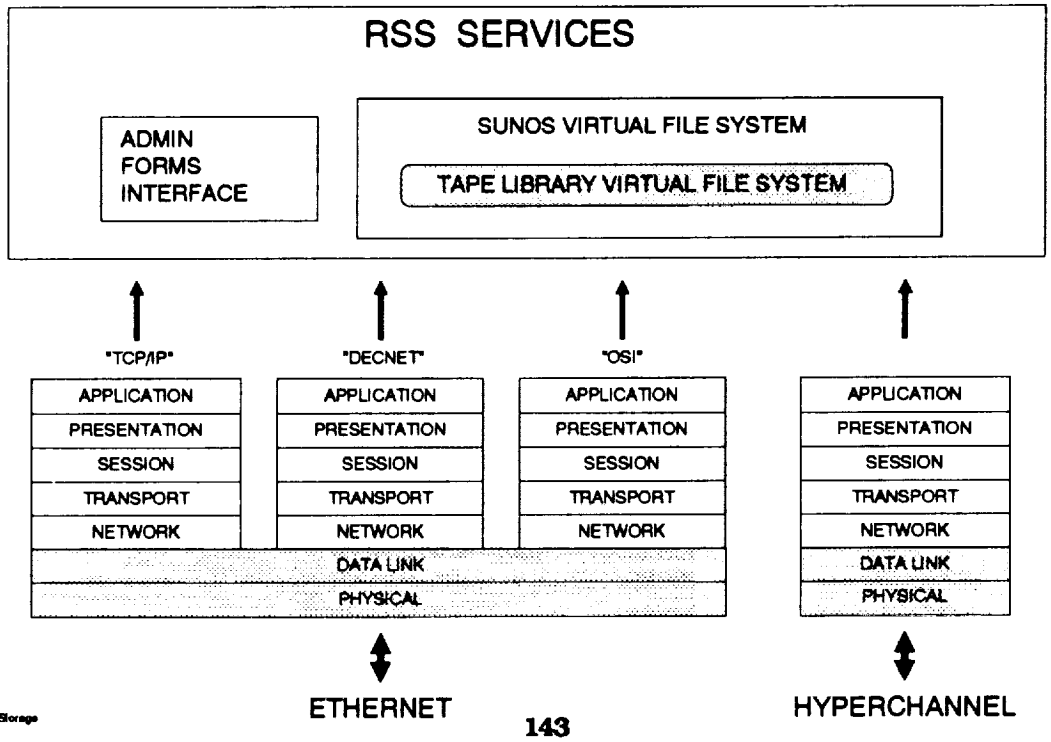
Metrum
Information Storage
[031891.sh]

RSS SOFTWARE ARCHITECTURE



Metrum
Information Storage
[062491-sh]

RSS SUPPORTS MULTIPLE CONCURRENT PROTOCOLS



Metrum
Information Storage
[031691-sh]

RSS ARCHIVE ARCHITECTURE SUMMARY

- UNIX-based operating system
- Standard network protocols
- User-directed archiving
- Supports automatic archiving
- Managed data
- Fast, easy access

Metrum
Information Storage
[031891.dh]

ICI Imagedata

ICI Optical Data Storage Tape

Robert A McLean
Joseph F Duffy

1. Introduction to flexible optical media

Optical data storage tape is now a commercial reality. The world's first successful development of a digital optical tape system is complete. This is based on the *Creo 1003* optical tape recorder with *ICI 1012* write-once optical tape media. Several other optical tape drive development programs are underway, including one using the *IBM 3480* style cartridge at LaserTape Systems.

In order to understand the significance and potential of this step change in recording technology, it is useful to review the historical progress of optical storage. This has been slow to encroach on magnetic storage, and has not made any serious dent on the world's mountains of paper and microfilm. Some of the reasons for this are:

- The long time needed for applications developers, systems integrators and end users to take advantage of the potential storage capacity.

- Access time and data transfer rate have traditionally been too slow for high-performance applications.

- Optical disk media has been expensive compared with magnetic tape.

As one of the world's major international chemical companies, ICI's strategy in response to these concerns has been to concentrate its efforts on flexible optical media; in particular optical tape.

2. Manufacturing achievements

Flexible optical media offers many benefits in terms of manufacture; for a given capital investment, continuous, web-coating techniques produce more square meters of media than batch coating. The coated layers consist of a backcoat on the non-active side; on the active side there is a subbing layer, then reflector, dye/polymer and transparent protective overcoat. All these layers have been tailored for ease of manufacture, and specific functional characteristics.

3. Media characteristics

The media can contribute to high system performance over a very wide range of capacities, depending on the drive design and media format. In addition to low cost/MegaByte, the media offers the archivability and indelibility that is vital for many storage applications. Thus, the media permits the development of systems that provide a unique set of features:

- **Low on-line cost/MB**: 10¢ to 40¢/MB depending on the format.
- **Low media cost/MB**: ½¢ to 1¢/MB at first, falling with time.
- **High performance in access time**: the low track spacing and high longitudinal speed of fixed head optical tape drives allows search rates between 2 GB/s and 20 GB/s. This compares very favorably with helical scan magnetic drives.
- **High data rate**: a single laser channel can achieve 4 MB/s with 60 mW on the media surface, and this can readily be increased with the use of multiple lasers.
- **High volumetric efficiency**: flexible optical media offers efficiencies a factor of 10 higher than advanced helical magnetic recording.

- **Indelible media**: the original information cannot be erased and altered.

- **Unlimited read cycles**: non-contact recording means that the data will survive in excess of 30,000,000 read cycles. The number of tape wind cycles is very high.

- **Long media lifetime**: the use of chemically stable materials has allowed us to predict a lifetime in excess of 15 years.

4. Media lifetime

Optical media based on organic, dye/polymer materials is well-known for its chemical stability. An unfortunate consequence of this is the difficulty in predicting a lifetime. Several accelerated aging techniques are in common usage: steady-state elevated temperature and humidity with Arrhenius extrapolation; elevated levels of temperature plus corrosive gases (the "Battelle" test); and chemical stability investigations.

We have used all of these extensively, both to attempt to predict an absolute lifetime limit, and in comparisons with magnetic tapes: iron oxide, cobalt modified iron oxide, chromium dioxide and metal particle formulations. We have also compared our media with rigid optical disks, including stability to uv light.

The Arrhenius test is performed at various constant conditions of elevated temperature and humidity (Ref. 1). The test assumes the following relationship, where failure is dominated by one chemical process with a given activation energy: $1/\text{time} = A\{e^{-E/kT}\}$. Our initial criteria for failure was a change in absolute reflectivity of 5% (Fig. 2) or a drop-off in the Carrier-to-Noise Ratio (CNR) by 3 dB (Fig. 3). This level of failure was not detected for any of the samples. A better way of detecting failure is to measure the raw BER on a drive; we are now doing this on the Creo drive.

To generate an Arrhenius estimate (Fig. 1), we looked for any deviation from the starting values that was significant compared with the measurement accuracy. There was no measurable change for the sample at 55 °C. The sample at 95 °C was well above the glass transition temperature of the PET base film, and so was badly warped, and this led to inaccurate reflectivity measurements and prevented us from doing a CNR measurement. The Arrhenius graph slope gives an activation energy of 1.12 eV. This implies that a sample kept at conditions of 20 °C and 60 %RH would survive for 393 years. A comparison with other forms of optical media is shown in Table 1.

We have used two UV stability tests. The *Blue Wool Test* was developed for dyes used in the clothing industry, and has now been formalized as British Standard BS1006. It assesses light fastness on a scale of 1 (low) to 6 (high). ICI's media measures > 4 with the test still continuing. By comparison, the average textile dye measures < 4.

Another UV test was used by Sony to test their optical disk media (Ref. 2). This exposes a sample to 120 hours of UVA light at 45 °C and 60 %RH, and is equivalent to 70 days of sunshine. We saw 1% drop in reflectivity - a result similar to that for *Sony Century Media*.

We also compared our media to metal particle and metal oxide tapes, and examined, respectively, recording characteristics and chemical stability.

The metal particle tapes were stored under accelerated aging conditions, and the recording characteristics were assessed by looking for any changes in the magnetic properties. We bought two tapes; tape "A" is consumer R-DAT and tape "B" is 8 mm video, both were different major Japanese brands. The test method consisted of storing the tapes at a constant 60 °C and 80 %RH for 3 and 6 weeks, while keeping a Control tape at room conditions. The magnetic properties were then measured in a high saturation field VSM. This work was done for ICI by the Fulmer Research Company in England, and is summarized in Table 2. The signal strength dropped by up to 22% after 6 weeks.

We performed an analogous test on the ICI optical tape out to 9 weeks (Fig. 4). The signal strength was unchanged, and the CNR curves did not change within the measurement accuracy. In addition, we used a Time Interval Analyzer (TIA) to look for an increase in the intrinsic error rate by measuring a Geometric Error Rate (GER) on unwritten media, and looking for changes. The GER is directly proportional to the sum of the defective areas. Figure 5 displays this against the detector threshold level, normalized to unwritten media at 100%. Points below 100% are dark defects, and those above are light defects. ICI media darkens upon writing and a typical detector threshold level is 60%, so the GER increased by less than a factor of 2.

Temperature cycling tests can be used to test not only the chemical stability of the media, but also the mechanical integrity of the structure. It is not possible to infer a lifetime from cycling tests. We subjected the optical tape media to 20 temperature cycles defined by Figure 6. We again looked for changes in the signal characteristics. There was 1 or 2 dB drop in the CNR (Fig. 7), which is just greater than the measurement accuracy. The time interval results in Figure 8 showed a corresponding increase in the signal jitter (standard deviation) of 3 ns. The "(+/-)" refers to the signal's rising-to-falling half cycle, and "(-/+)" is the other half cycle.

A range of oxide magnetic tapes were stored under accelerated aging conditions, and any degradation was assessed by looking for any decomposition products. This work was also done by Fulmer Research. We tested three tapes; tape "C" is iron oxide instrumentation tape, tape "D" is cobalt modified iron oxide VHS consumer video tape, and tape "E" is chromium dioxide *IBM 3480* type data cartridge tape.

The test method was to store the tapes at a constant 60 °C and 80 %RH for 3 and 6 weeks, and keep a Control tape at room conditions. Standard solvent extraction techniques were used to look for decomposition products. One meter samples were immersed for two hours in a Soxhlet extraction by Delifrene; then soluble extract was weighed. This was followed by a further two hours in acetone and weighing the extract.

The results are presented in Table 3. The initial extract is a baseline, and the percentage increase above this is a measure of chemical decomposition in the binders. These results can be put into perspective through a paper by Bertram and Cuddihy (Ref. 3). This states that after aging, an increase in extract of 1.4 % by weight corresponds with a degradation in tape performance. When our results are compared with Bertram and Cuddihy's, there was very good agreement for the Tape "C".

We performed a similar chemical stability test on the optical tape. Again, the test method was to store at a constant 60 °C and 80 %RH, and also 80 °C and 80 %RH for 1, 2 and 3 weeks, and keep a Control tape at room conditions. We were unable to detect any decomposition products at a constant 60 °C and 80 %RH, which was why the test was repeated at the more severe conditions of 80 °C and 80 %RH.

We devised an extraction technique suitable to the solubility of the dye/polymer material. This involved immersing 500 cm^2 samples for 72 hours in a Soxhlet extraction by ethanol. Samples and extract were then weighed and analyzed by sensitive FTIR techniques. Using FTIR, we could detect binder degradation only at the harsher conditions of 80 °C and 80 %RH, and this for a minor component of the protective overcoat, not the recording layer. The quantities were too small to be weighed.

Another test we used is the *Battelle Class II* accelerated aging test, performed under contract at the Battelle Institute in Ohio. This test has been correlated to aging of materials in cities and other locations where combustion byproducts form a mix of corrosive gasses. A fully assembled open reel *ICI 1012* optical tape, without the protective sealed storage and shipping box was kept at a constant 23 °C and 70 %RH in a flowing mixed gas environment consisting of 10 ppb H_2S, 10 ppb Cl_2 and 200 ppb NO_2 for 30 days. This environment has been correlated to 15 years lifetime. We detected no damage in terms of corrosion, reflectivity or modulus.

5. Conclusion

Optical tape systems can offer a unique set of attributes to potential end-users. The use of chemically stable recording materials, indelible write-once technology, plus the advantages of non-contact reading and writing, yield a very robust and archival medium. Our lifetime studies have concentrated on measurement techniques and comparisons with magnetic technologies that are independent of the drive and read/write channel. The comparisons with magnetic tape technologies are favorable:

- No deterioration in optical recording characteristics after 9 weeks, 60 °C and 80 %RH.
- Metal particle magnetic formats show up to 22 % deterioration in magnetic capability after only 6 weeks.
- No binder hydrolysis in the optical tape recording layer after severe environmental exposure and extraction process.
- Minor hydrolysis present in the overcoat after 3 weeks at 80 °C and 80 %RH.
- Clear evidence of deterioration in magnetic oxide tapes under an industry standard hydrolysis test.

We plan to extend these studies to raise the lifetime prediction for *ICI 1012* optical tape from 15 to 30 years, and to characterize any changes in the raw BER through the read/write channel on the *Creo 1003* optical tape recorder.

References

1. A B Marchant, *Optical recording: a technical overview*, Addison-Wesley Publishing Company, pp 380-382, 1990.
2. Sony Publication *Recording Principle and Reliability Study of Sony Century Media*, Sony Corporation of America, Park Ridge, NJ, April 1988.
3. H N Bertram and E F Cuddihy, "Kinetics of the Humid Aging of Magnetic Recording Tape", *IEEE Trans Magn*, vol MAG-18, no 5, pp 993-999, Sept 1982.

Table 1

Company	Media Type	Test Method	Activation Energy in eV
Sony	Metal Alloy	BER	1.5
ICI	Dye/polymer	%R and CNR	1.12
OITDA*	Multi WORM	%R	1.0
Hitachi	12" WORM	BER	1.0
NEC	Magneto-optic	DER	0.97

*OITDA = Japanese Standard Committee for the Optical Data Disk

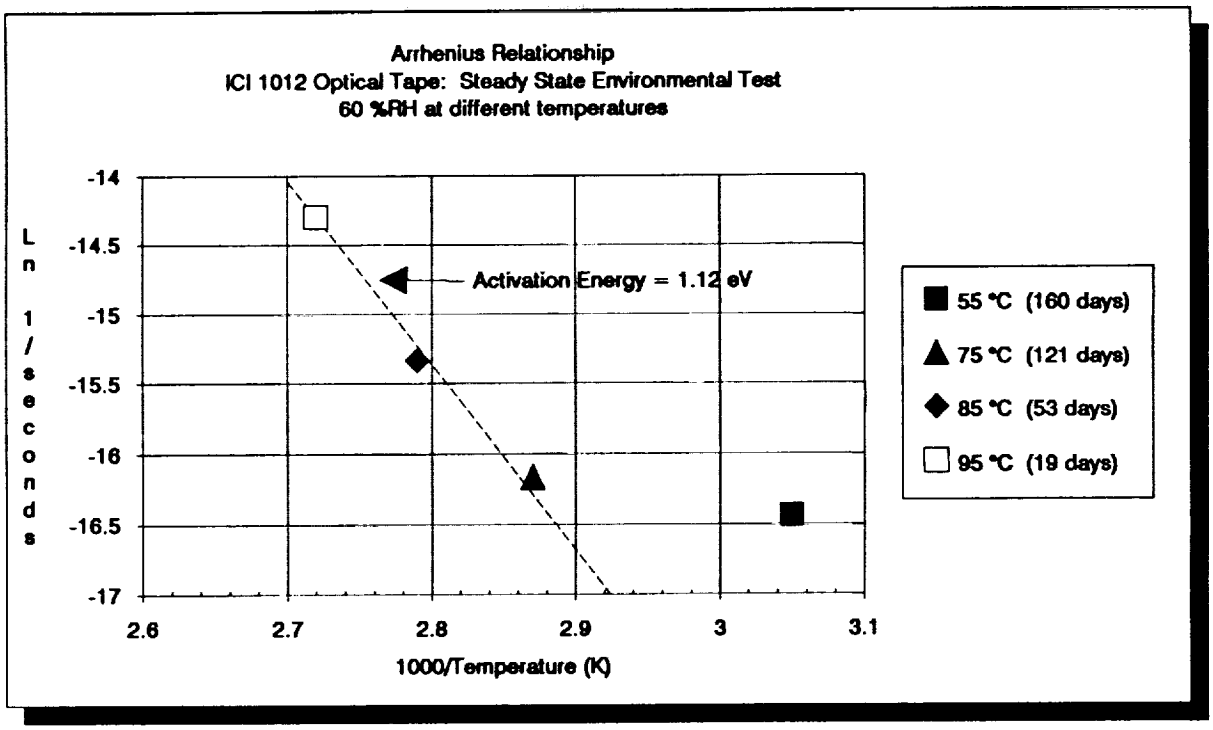
Table 2

Tape	Time	M_s % fall	M_r % fall	Sq % fall
"A"	Control	0	0	0
	3 weeks	10.64	11.30	0.74
	6 weeks	18.84	20.34	1.84
"B"	Control	0	0	0
	3 weeks	4.20	4.38	0.19
	6 weeks	22.35	22.31	-0.015

Table 3

Tape	Time	Total Extract in %	Increase in %
"C"	Control	1.20	0
	3 weeks	2.35	1.15
	6 weeks	3.49	2.29
"D"	Control	1.24	0
	3 weeks	1.45	0.21
	6 weeks	1.86	0.62
"E"	Control	1.53	0
	3 weeks	2.04	0.51
	6 weeks	2.55	1.02

Fig. 1


Arrhenius Relationship
ICI 1012 Optical Tape: Steady State Environmental Test
60 %RH at different temperatures
Ln 1/seconds
-14
-14.5
-15
-15.5
-16
-16.5
-17
Activation Energy = 1.12 eV
2.6
2.7
2.8
2.9
3
3.1
1000/Temperature (K)
55 °C (160 days)
75 °C (121 days)
85 °C (53 days)
95 °C (19 days)

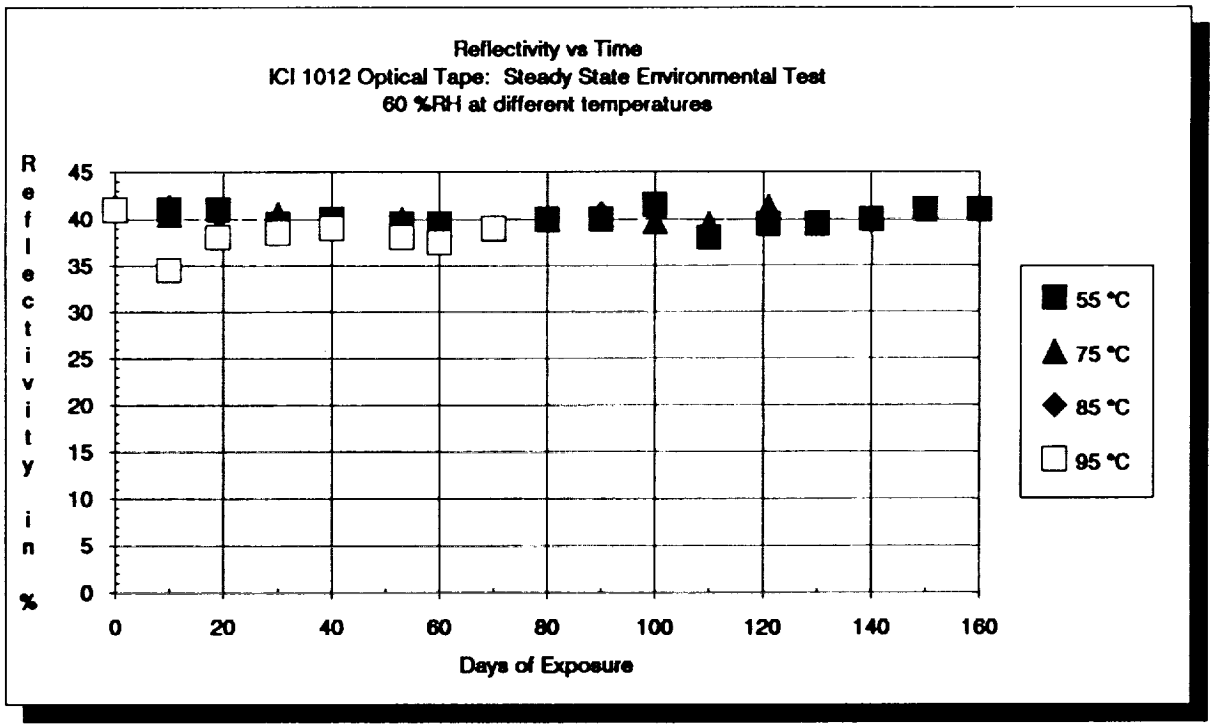
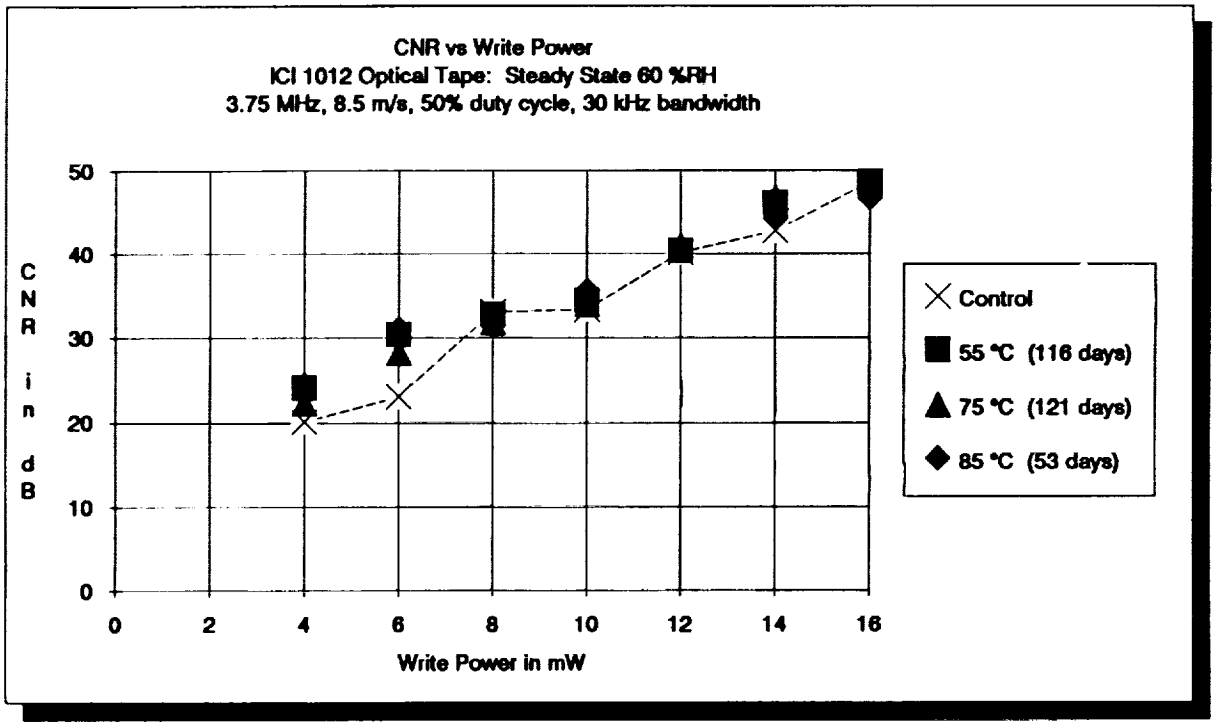

Reflectivity vs Time
ICI 1012 Optical Tape: Steady State Environmental Test
60 %RH at different temperatures
Reflectivity in %
45
40
35
30
25
20
15
10
5
0
0
20
40
60
80
100
120
140
160
Days of Exposure
55 °C
75 °C
85 °C
95 °C

Fig. 2

Fig. 3


CNR vs Write Power
ICI 1012 Optical Tape: Steady State 60 %RH
3.75 MHz, 8.5 m/s, 50% duty cycle, 30 kHz bandwidth
CNR in dB
0
10
20
30
40
50
0
2
4
6
8
10
12
14
16
Write Power in mW
Control
55 °C (116 days)
75 °C (121 days)
85 °C (53 days)

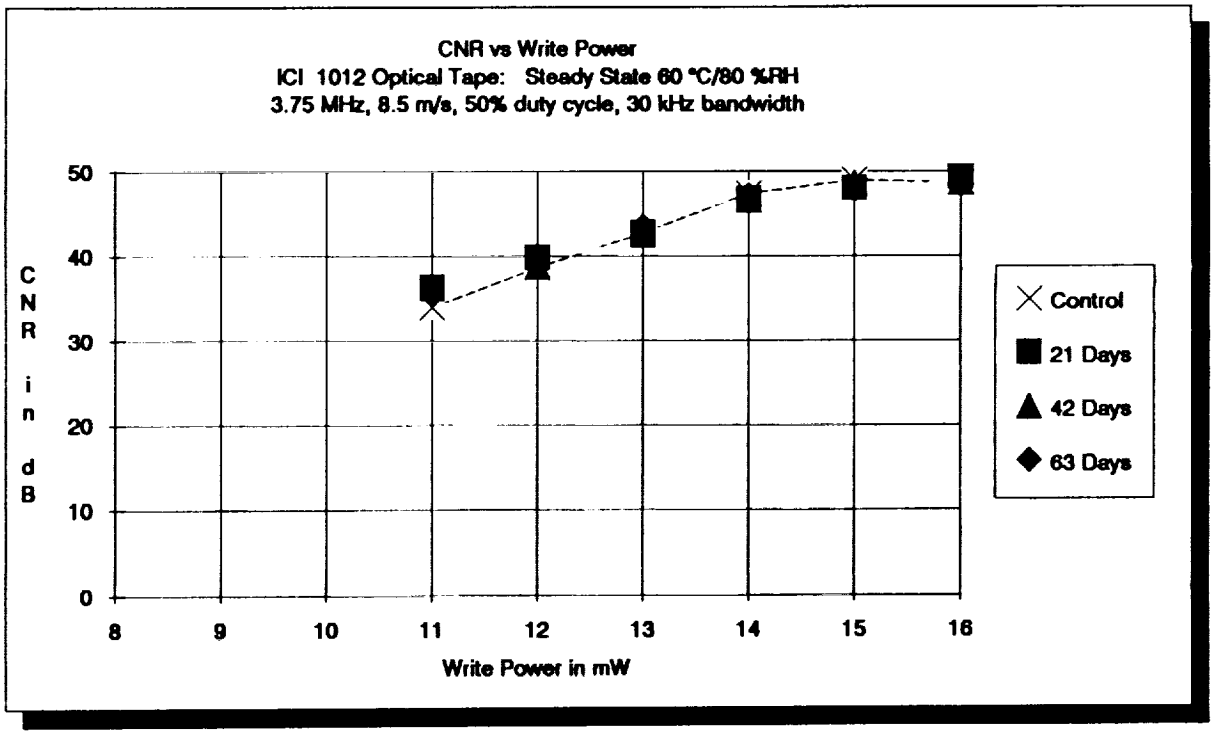
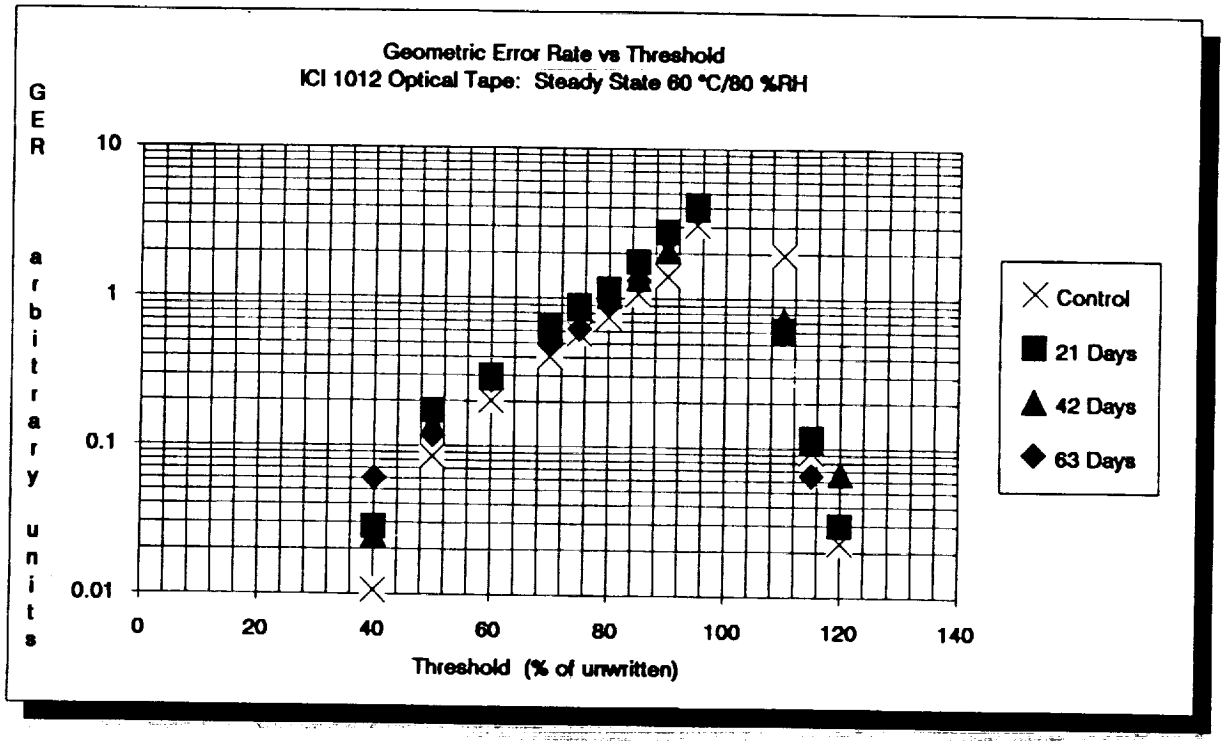

CNR vs Write Power
ICI 1012 Optical Tape: Steady State 60 °C/80 %RH
3.75 MHz, 8.5 m/s, 50% duty cycle, 30 kHz bandwidth
CNR in dB
0
10
20
30
40
50
8
9
10
11
12
13
14
15
16
Write Power in mW
Control
21 Days
42 Days
63 Days

Fig. 4


Geometric Error Rate vs Threshold
ICI 1012 Optical Tape: Steady State 60 °C/80 %RH
GER arbitrary units
10
1
0.1
0.01
0
20
40
60
80
100
120
140
Threshold (% of unwritten)
Control
21 Days
42 Days
63 Days

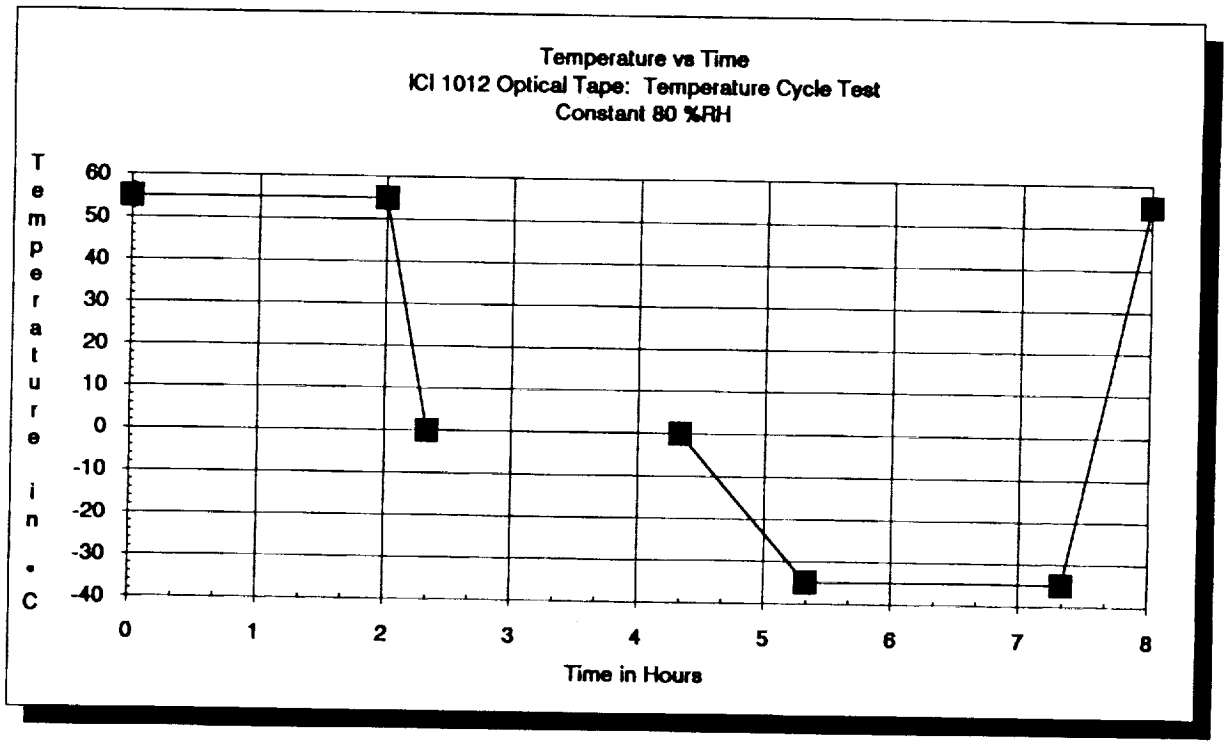
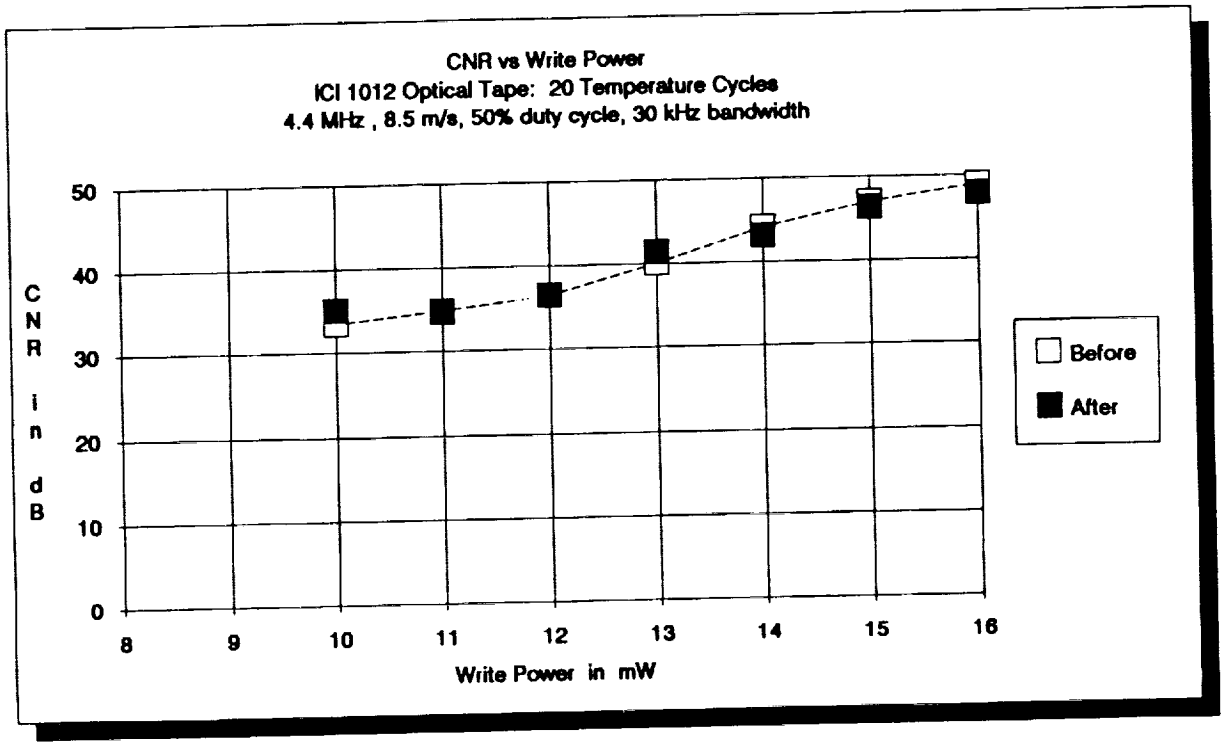

Temperature vs Time
ICI 1012 Optical Tape: Temperature Cycle Test
Constant 80 %RH
Temperature in °C
60
50
40
30
20
10
0
-10
-20
-30
-40
0
1
2
3
4
5
6
7
8
Time in Hours

Fig. 7


CNR vs Write Power
ICI 1012 Optical Tape: 20 Temperature Cycles
4.4 MHz , 8.5 m/s, 50% duty cycle, 30 kHz bandwidth
CNR in dB
0
10
20
30
40
50
8
9
10
11
12
13
14
15
16
Write Power in mW
Before
After

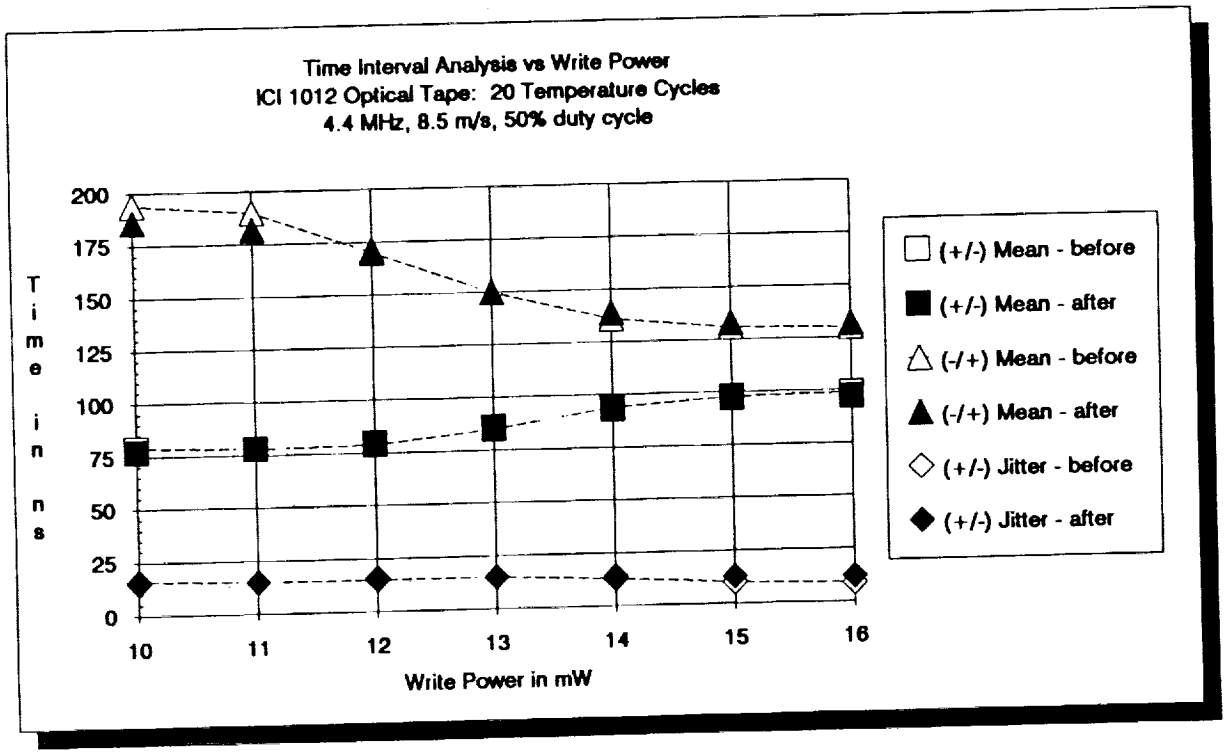

Time Interval Analysis vs Write Power
ICI 1012 Optical Tape: 20 Temperature Cycles
4.4 MHz, 8.5 m/s, 50% duty cycle
Time in ns
0
25
50
75
100
125
150
175
200
10
11
12
13
14
15
16
Write Power in mW
(+/-) Mean - before
(+/-) Mean - after
(-/+) Mean - before
(-/+) Mean - after
(+/-) Jitter - before
(+/-) Jitter - after

Fig. 8

Presentation for NSSDC Conference

Mass Storage Systems and Technologies for Space and Earth Science Applications

ATL Products Division's Entries Into the Computer Mass Storage Marketplace

July 24, 1991

Fred Zeiler
Odetics, Inc.
ATL Products Division
240 E. Palais Road
Anaheim, California 92805
714-774-6900

157

Odetics Background

- **High Tech Company Founded in 1969, Publicly Traded**
  - Serving Well-Defined Niche Markets
  - Through Variety of Product Groups
- **Roots are in Space Borne Recorders**
  - Own 80% of Marketplace
- **Evolved Into Robotics in Mid-70s**
  - AIM, Broadcast, DMS, ATL Products Divisions
  - 35% of Revenues and Growing
  - Technology and People Move From One Division to Another

Product Evolution

- **Robotics Genesis - AIM**
- **Company's High Technology Group**
  - 1979 Committed to a Six Legged Robotic System
  - 18 Months Later Demonstrated ODEX I
    - Symbol of the Corporate Commitment to Robotics
  - Demonstrates High Strength to Weight Ratios
  - All Electric, Compact, Extremely High Performance
  - Six Units Built - Three Generations of Technology
    - Predominantly for Nuclear Plant Maintenance
- **Evolution to Other Robotic Subsystems**
  - Arms, Hands, and Effectors

Product Technologies and Markets Served

- **Innovators in "Small Package" Handling**
- **Do Not Serve General Purpose Robotics Handling Market**
- **Design Intent of Our Products**
  - Move "Small Light Weight Objects" Very Quickly
  - Accent On Longevity of "Object" Being Moved
  - High Degree of Reliability
- **Necessitates**
  - Expertise in Low Mass, Light Weight, High Speed Systems
  - Requires Unique Robotic Handlers, Arms, End Effectors
  - Products Designed for Niche Markets
    - Aperture Card Storage Module Systems
    - Tape Cassettes and Cartridges
    - Optical Disks

Product Evolution
Infodetics' Aperture Storage Module Library

- **First Linear Servo Based Expandable System**
- **Modules: 10 Ft. Long By 3 Ft. Deep By 7 Ft. High**
  - 2000 Cartridges Per Two Rows or Module
  - 100 Aperture Cards Per Cartridge
  - Robotic Handler in Aisle Between Rows Within Module
- **Large System With Multiple Modules and Pass-Through**
- **5 Seconds Average "Pick and Place"**
  - Access Cartridge and Load Into Aperture Card Reader
- **Document Management System "Storage Server"**
  - Cache Microfilm Images to Disk
  - Transmitted to Work Stations for Viewing

DOCUMENT NO. 9011002
JULY, 1991

Product Evolution
Broadcast Division's TCS2000 Video Cart

- **First "Tower" Based Expandable System Introduced '86**
  - Designed as a TV Station or Network Automation System
- **Built as Part of a Joint Venture With RCA in 18 Months**
  - RCA Dropped Out, Odetics Entered End User Market
- **System Consists of:**
  - Robotics and Up to 6 Tape Recorders Per Tower
  - 225 to 300 Tapes Per Tower Depending On Formats
  - Switchers, Sequencers, Monitor and PC Based Work Station
  - Hierarchical Software
    - Real-Time Controller/Operating System, Relational DB, Playlist
- **Supports VHS, Beta, D-2 Formats**
- **Robust, Redundant and Extremely Reliable**

DOCUMENT NO. 9011002
JULY, 1991

The Cart Machine™ with Library Expansion Module

Odetics Broadcast

Broadcast Division's Newest Product
TCS90 Videocart System

- **Bookshelf Design**
  - X-Y High Speed Linear Servo System
  - One Armed Gripper With Holding Tray
- **Accommodates Combination of Cassette Sizes and Format**
  - Beta SP, VHS, or D-2
  - Small and Medium Sizes
- **Tape Recorders are Standard: Non-Modified**
- **Autoloader Accommodates Up to 8 Cassettes**
- **Fixed Size**
  - No Expansion Capabilities
- **Software From TCS2000 Migrates Directly to This Product**
  - 50 Man Years of Development

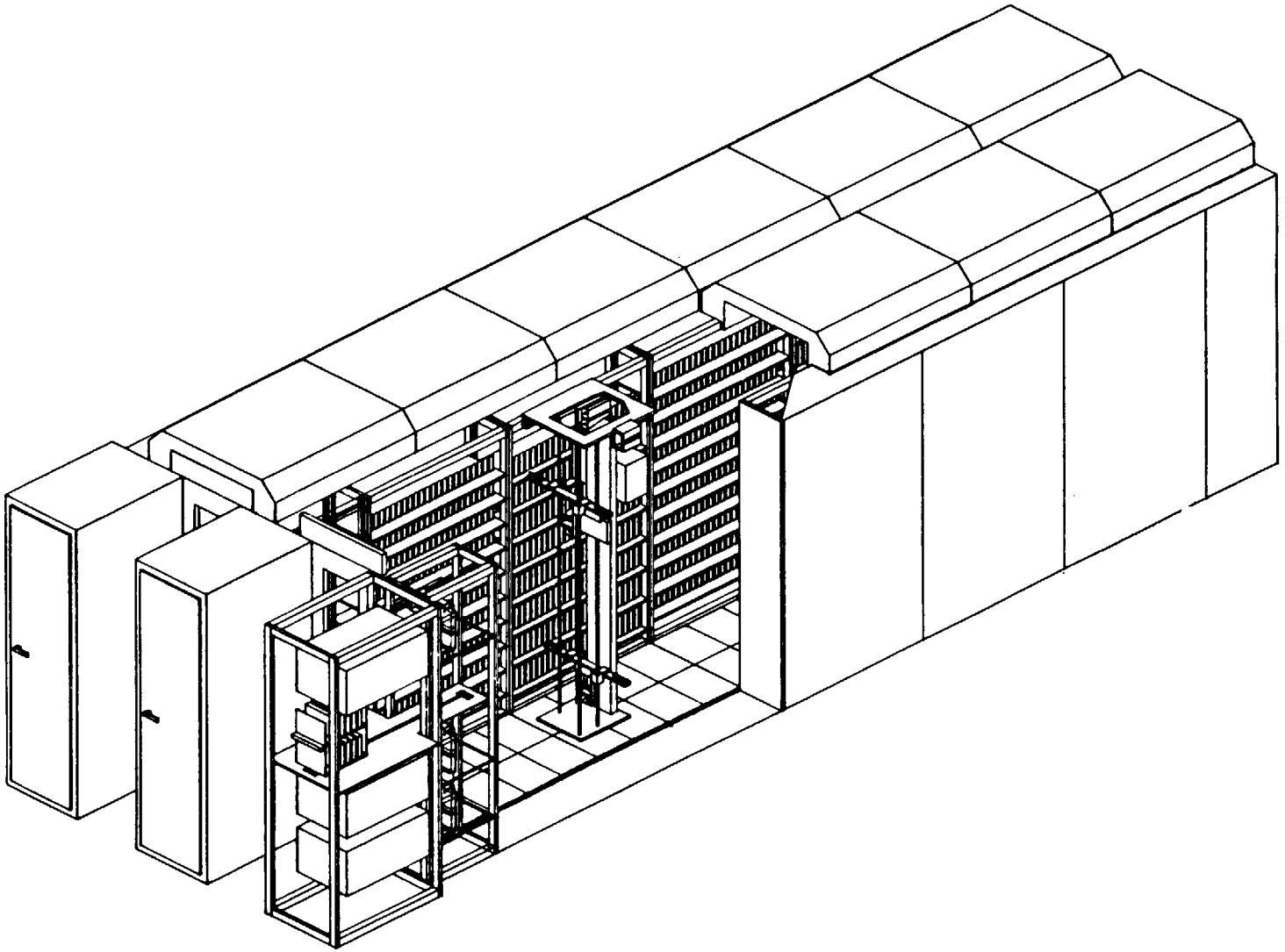
DOCUMENT NO. 9011002
JULY, 1991

ATL Products Division
Marketing Strategy

- **Serve the Evolving Computer Based Mass Storage Market**
- **Develop Tape Storage Library Subsystems**
  - Robotics, Control, Storage and Computer Interfaces
  - Support a Broad Range of Tape Sizes and Formats
  - Interface to a Variety of Tape Drives in Each Size
  - Provide Low Level Library Control and Management Software
- **Sell Through Distribution Channels That:**
  - Integrate Tape Drives
  - Add High Level Mass Storage Management Software
  - Service a Broad Range of User Markets
  - Provide "Private Labelled" and "General Purpose" Products
- **Pursue Major Market Shares**
  - High Density/High Capacity Storage Market
  - General Commercial Market By Supplying a Range of Solutions

High Density Systems Business
Product Lines

- **Developing Two Basic 19mm ATL Storage Subsystems**
  - To Serve High Capacity Markets: Terabytes and Petabytes
  - Support Small and Medium Size D-2 Cassettes
- **Expandable "Tower" Based System**
  - Broadcast System as Platform
  - Auxiliary Towers for Expansion
  - Delivered March, 1991
- **New Linear Aisle Based Expandable System**
  - 30 Months in Development: Delivery August, 1991
  - Most Advanced Robotic System On the Market



Odetics aisle-based automated tape library subsystem. Cut-away shows robot that moves up and down aisles of tapes. Tapes are stored and retrieved by the robot and placed into tape drives.

Odetics ATL Products Division

High Density Systems Business Product Lines

- **10 Minute Video On Two Technologies**
  - Copies of Video Available Upon Request

E-Systems Business Relationship

- **Exclusive Supplier of 19mm ATL Subsystems**
  - Computer Mass Storage Marketplace Only
- **Can Market Other Odetics ATL Subsystems**
- **E-Systems is the Integrator**
  - Providing Systems Expertise - ATLs, Tape Drives, Computer Integration
  - Library Management Software for the ATLs
  - Supplying Storage Server Software

3480 "Medium Size" Library
First Commercial Product Offering

- **"300" Cartridge Baseline System**
  - Expandable in Increments of Approximately 300 Cartridges
  - "1500" Cartridge Maximum
- **Up to 2 Tape Drives Available in Baseline**
  - Up to 4 Additional Drives as System Expands
  - Supports All Low Cost 3480 Tape Drives
- **Small Footprint**
  - Fits Standard 28 Inches Wide By 45 Inches Deep By 78 Inches High Cabinet
  - Very High Density Storage
- **Cartridge Autoloader and Bulk Loading**
- **RS-232C or SCSI-II Interface**
- **Serve the Distributed Computing and File Server Markets**

Storage and Library Management

- **ATLs are Driven With "Low Level" Commands**
  - Pick From Bin and Move to Tape Drive
  - Status Provided Back Through Sensors
  - Electrical Interface: RS-232C, Ethernet, SCSI-II
  - New Software Interface: SCSI-II, Chapter 16 Jukebox Commands
- **Library Management: Physical Volume Repository**
  - Input PVS and Provide Level of Intelligence
  - Management Resource and Allocation of Bins and Drives
  - Automatic Error Recovery
- **Servers and Applications Provide Next Level**
  - Storage Servers and Bit File/Client Servers
  - Backup

Conclusions

- **Odetics is and Will Be a Major Supplier of Robotic Libraries**
  - Advancing Technologies
- **By Year End, From Broadcast and ATL Products Divisions**
  - Four Different ATL Technologies and Five Products
  - Cross "Breeding" of Technologies Across Divisions
- **In the Future**
  - Broader Reach of Products and Markets Using Robotics
  - Further Transfer of Technologies at Component Levels

CORROSION OF METAL PARTICLE AND METAL EVAPORATED TAPES

Dennis E. Speliotis
Advanced Development Corp., 8 Ray Avenue, Burlington, MA 01803

Very high coercivity metal particle (MP) and metal evaporated (ME) tapes are being used in 8mm video and digital audio tape applications, and more recently in digital data recording applications. In view of the inherent susceptibility of such media to environmental corrosion, a number of recent studies have addressed their long term stability and archivability. These studies[1-4] have used an accelerated corrosion test based either on elevated temperature-humidity or polluting gas atmospheres known as Battelle tests. A comparison of the Battelle test results performed at different Laboratories reveals a large variation from one location to another[4], presumably due to incorrect replication of the Battelle condition. Furthermore, when the Battelle tests are performed on enclosed cartridges, it is quite possible that diffusion limits the penetration of the extremely low concentration polluting gaseous species to the inner layers of the tapes during the short time of the accelerated test(typically 7 to 10 days), whereas in real life these diffusion limitations may not apply. To avoid this uncertainty ,in this study we investigated the corrosion behavior of commercial 8mm MP and ME tapes when cassettes without their external plastic cases were exposed to 50°C and 80% RH for 7.5 weeks.

The effects of the corrosion were studied by measuring the error statistics at a density of 53.3kfci (2100 fc/mm) using an 8mm helical scan recorder with 0.25 micron gap MIG heads controlled by a Media Logic 4500 Digital Tape Evaluator System. This system is programmed to measure the dropouts at different threshold levels and to provide error maps for large numbers of tracks. The error statistic were measured before and after the corrosion cycle, and were compared on the basis of the change in the average number of errors per track and the error size distribution at a specific threshold level, as well as complete error maps and counts for a large number of tracks.

Our results show a large increase in errors due to corrosion for all the MP and Me tapes studies (typically by two orders of magnitude). There is also a large variation in corrosion stability among the tapes from different manufacturers. Typical results for MP and ME tapes are shown in the figures below. This large increase in errors may be due to a change in the magnetization of the tapes (particularly in the critical 0.3 micron region near the surface of the tapes which represents the area responsible for most of the signal at the 53.3 kfci recording density), or to a change in the surface morphology of the tapes, or a combination thereof.

1. E. F. Wollack et al, paper JA-3 in the Abstracts of the Intermag Conference, Washington, D.C., March 1989.
2. D. E. Speliotis, IEEE Trans. Magn., MAG-26, 124 (1990).
3. Y. Yamamoto et al, IEEE Trans. Magn., MAG-26, 2098 (1990).
4. A. Djalali et al, Proc. 1st Intl. Symposium on Corrosion of Electronic Materials and Devices, Electrochem. Soc. p.430 (1991)

Corrosion Test: 7.5 Weeks 50C, 80% RH
Dropout vs. Threshold (20/28/80)

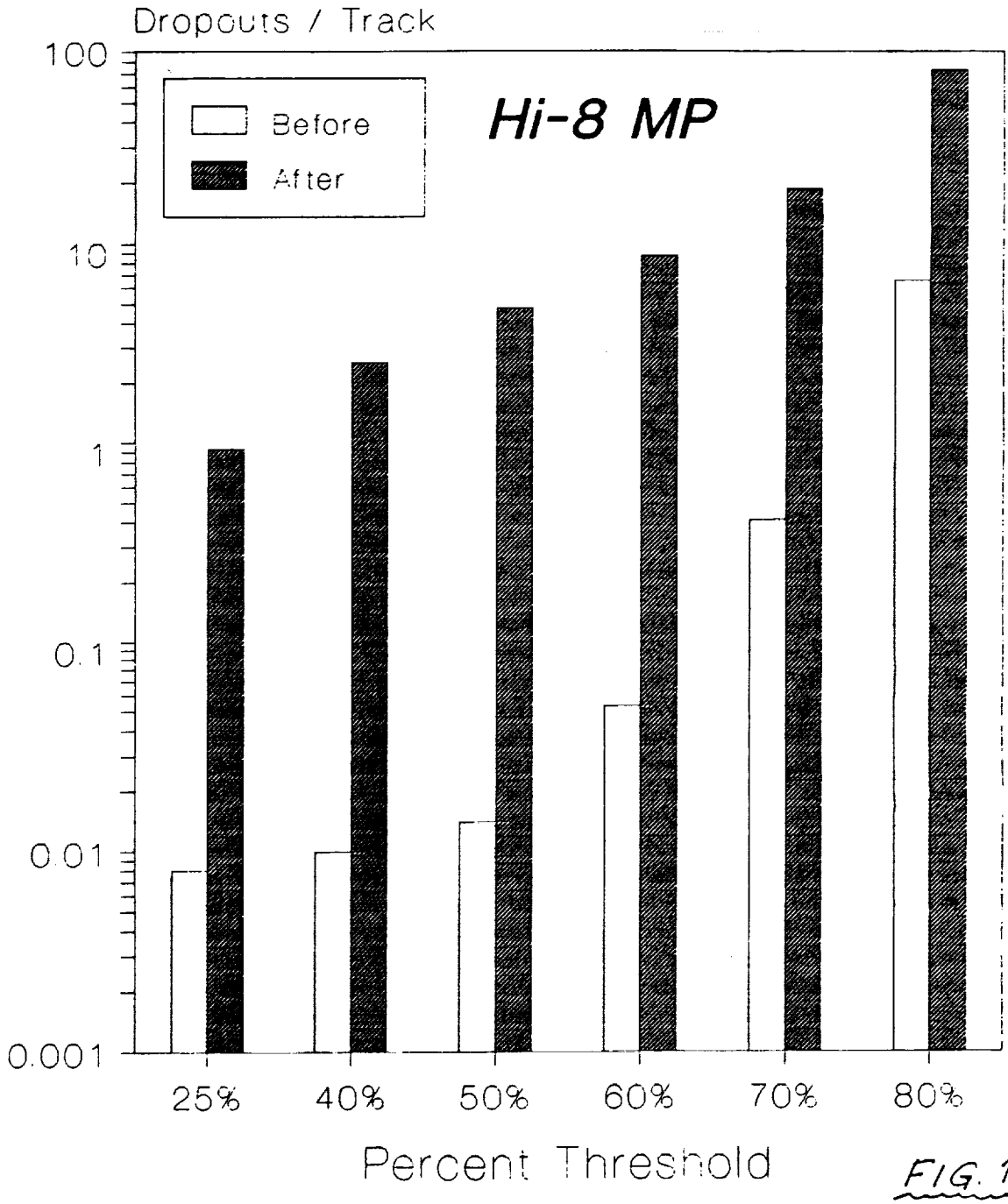


FIG. 1
Hi 8 MP Tape before and after corrosion

Corrosion Test: 7.5 Weeks 50C, 80% RH
Dropout Size Distibution (75% TH., 28G)

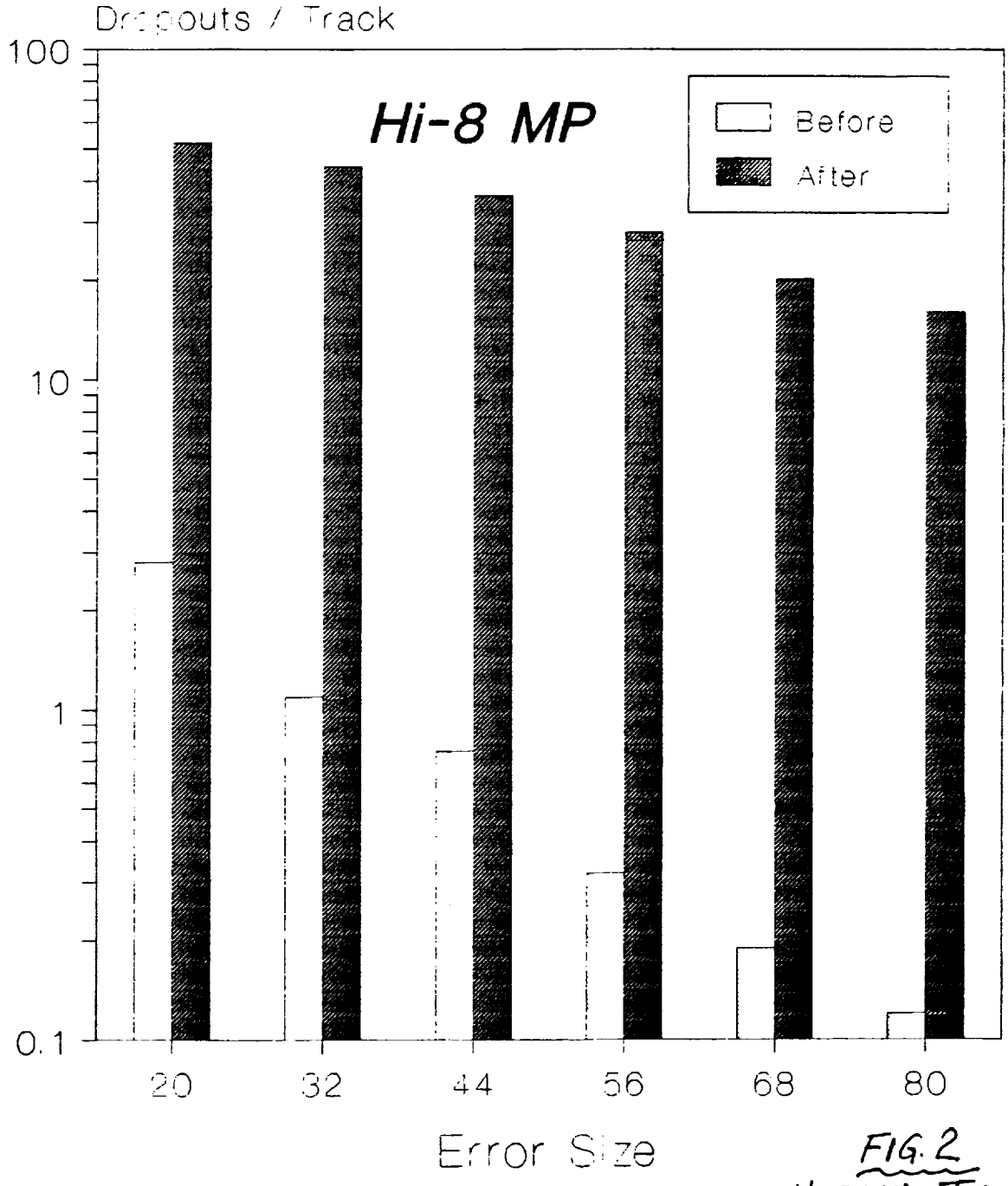


FIG. 2
Hi8MP Tape
before and after
Corrosion

Corrosion Test: 7.5 Weeks 50C, 80% RH
Dropout vs. Threshold (20/28/80)

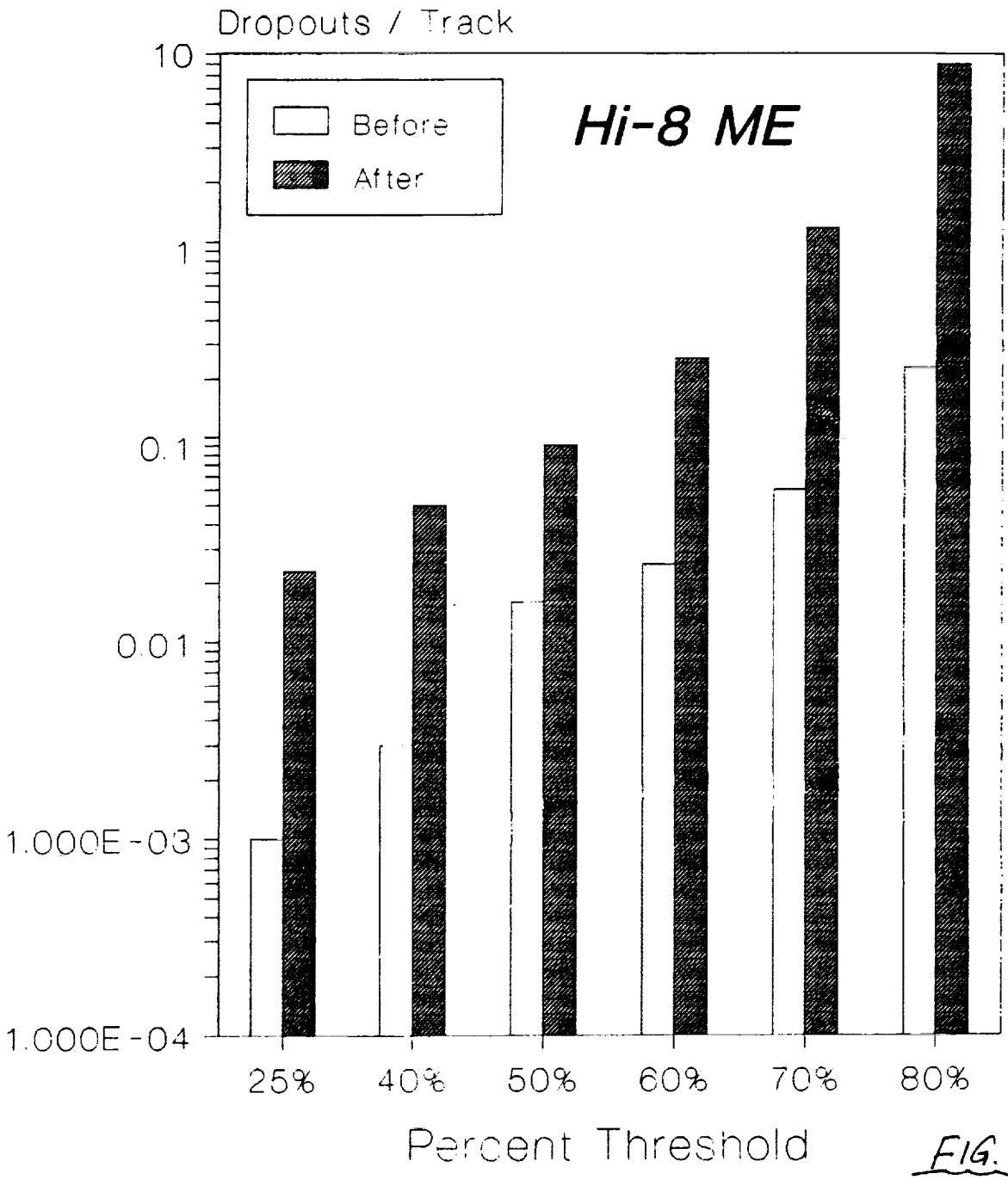


FIG. 3
Hi8ME Tape before and after corrosion

Corrosion Test: 7.5 Weeks 50C, 80% RH
Dropout Size Distibution (75% TH., 28G)

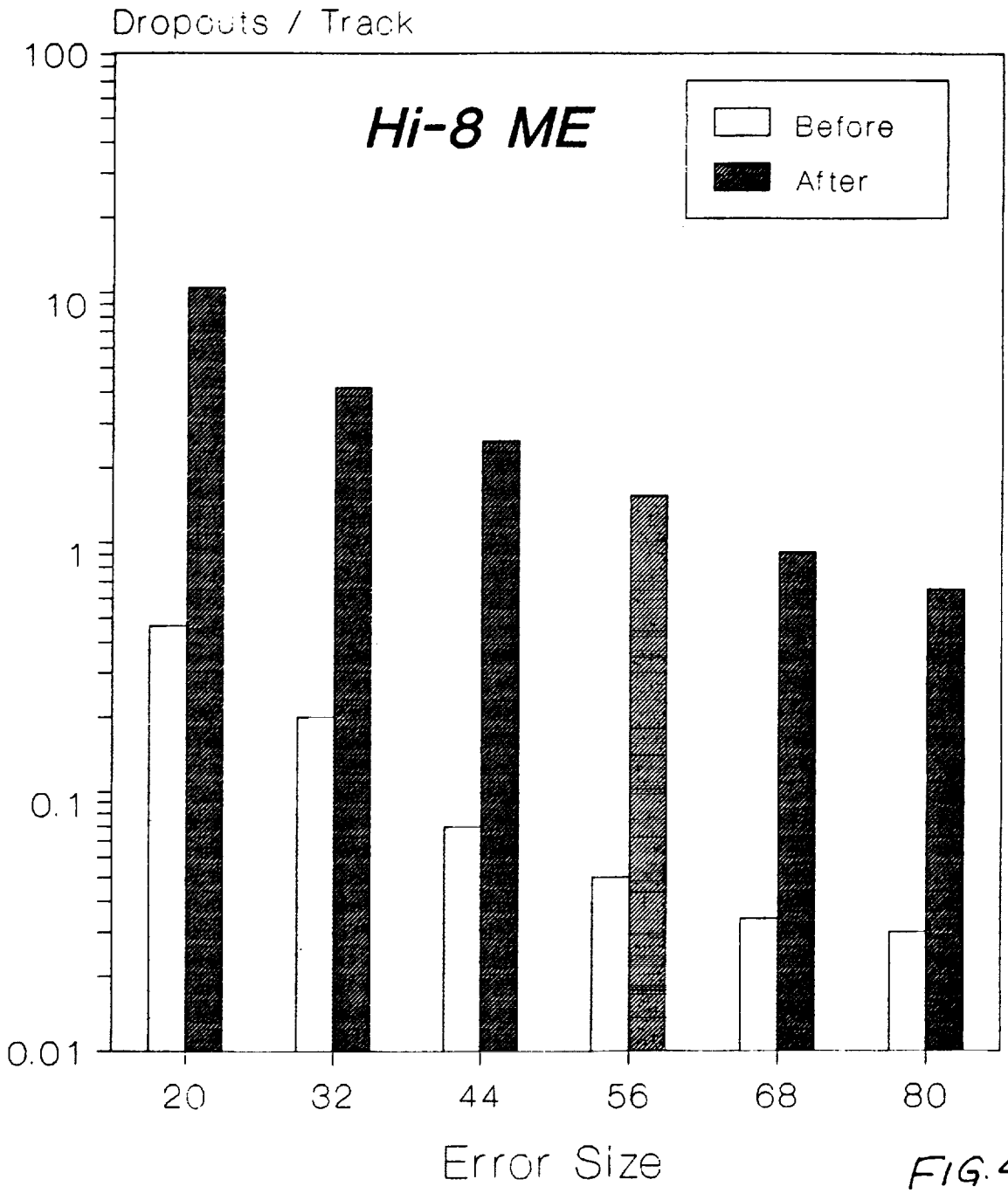


FIG. 4
Hi8 ME Tape
before and after
Corrosion

Corrosion Test: 7.5 Weeks 50C, 80% RH
Dropout vs. Threshold (20/28/80)

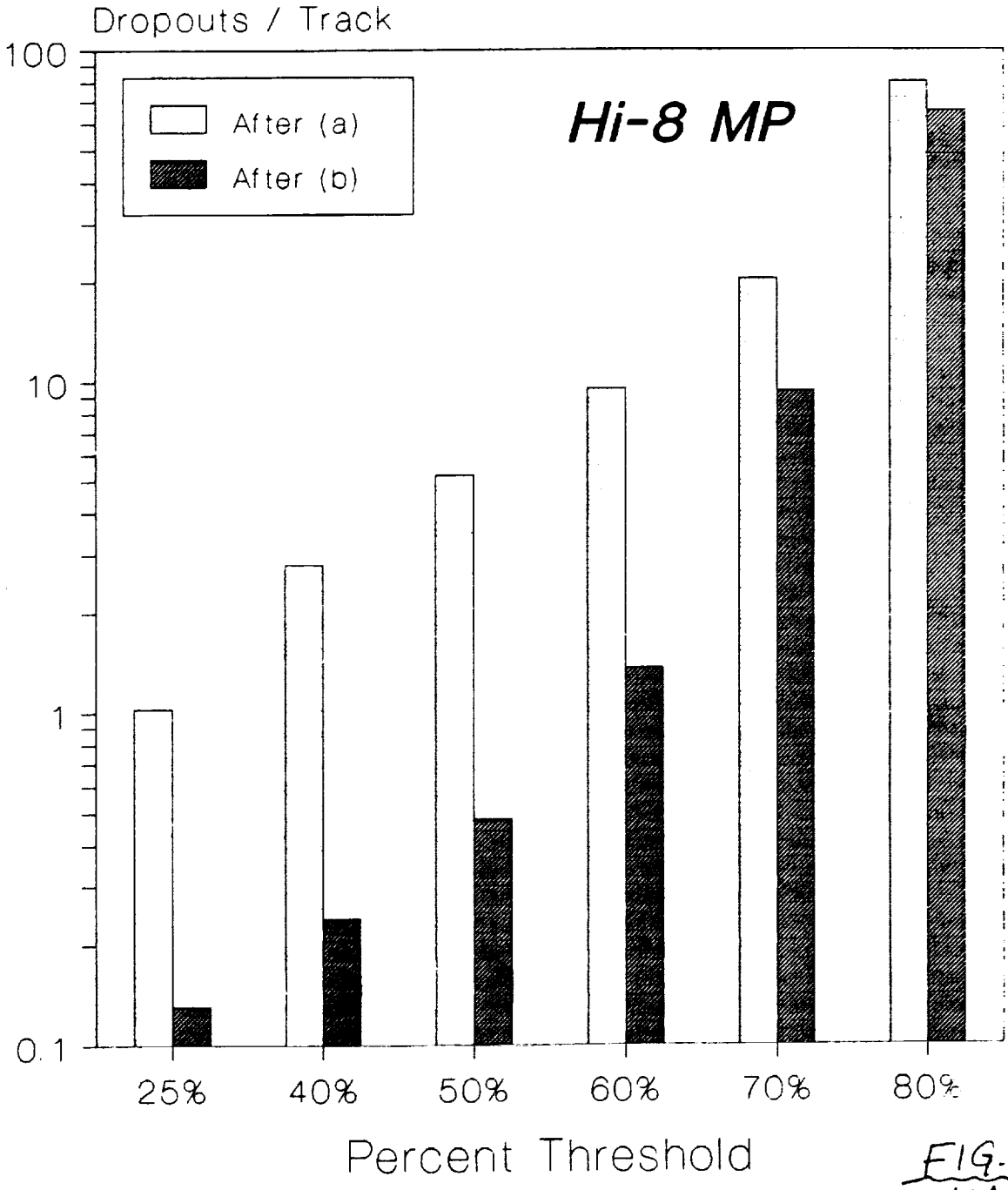


FIG. 5
Two different Hi 8 MP Tapes after corrosion

Corrosion Test: 7.5 Weeks 50C, 80% RH
Dropout Size Distibution (75% TH., 28G)

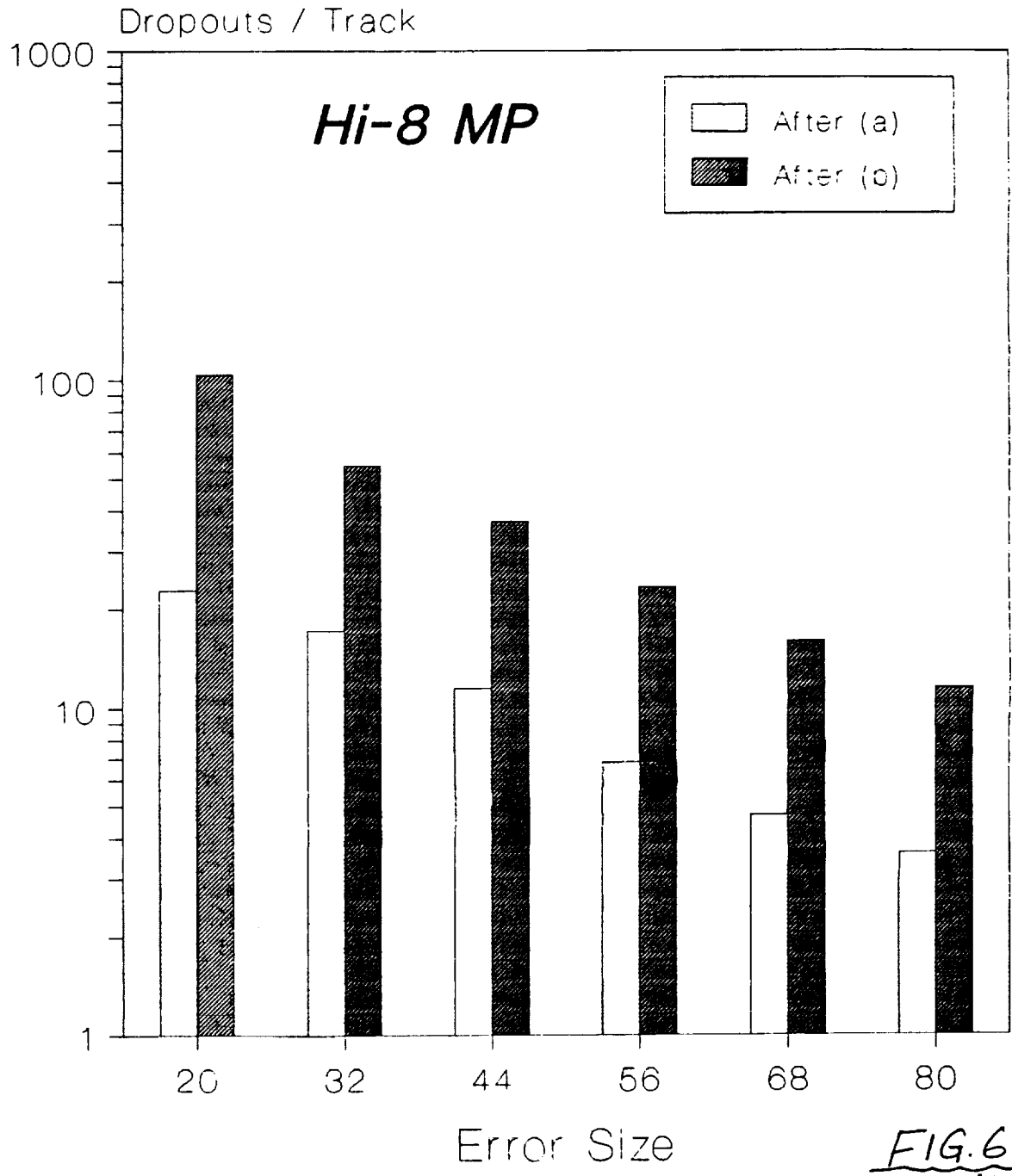


FIG. 6
Two different Hi8MP Tapes after corrosion

C-3

Corrosion Test: 7.5 Weeks 50C, 80% RH
Dropout vs. Threshold (20/28/80)

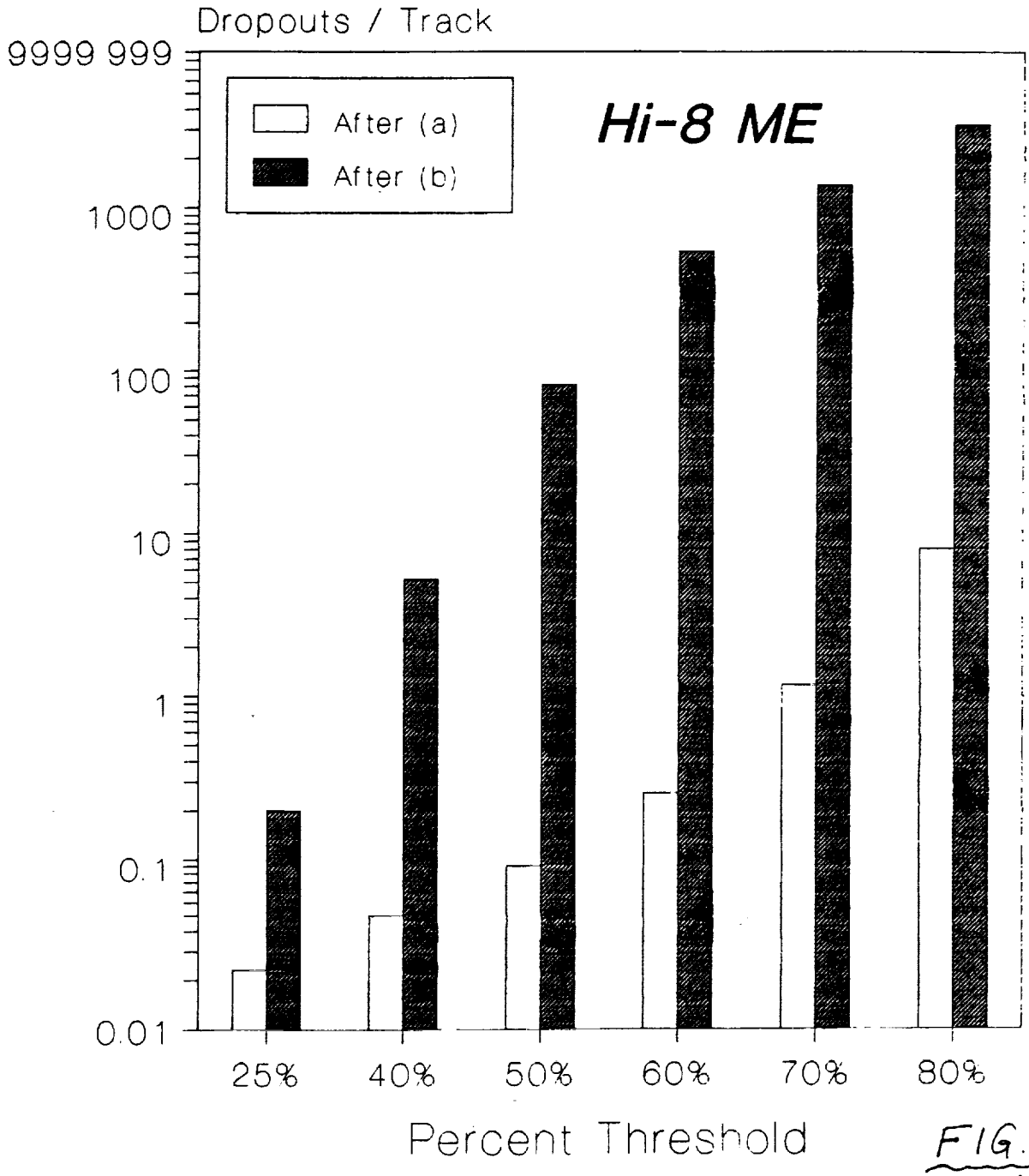


FIG. 7
Two different Hi8ME Tapes after Corrosion

Corrosion Test: 7.5 Weeks 50C, 80% RH
Dropout Size Distibution (75% TH., 28G)

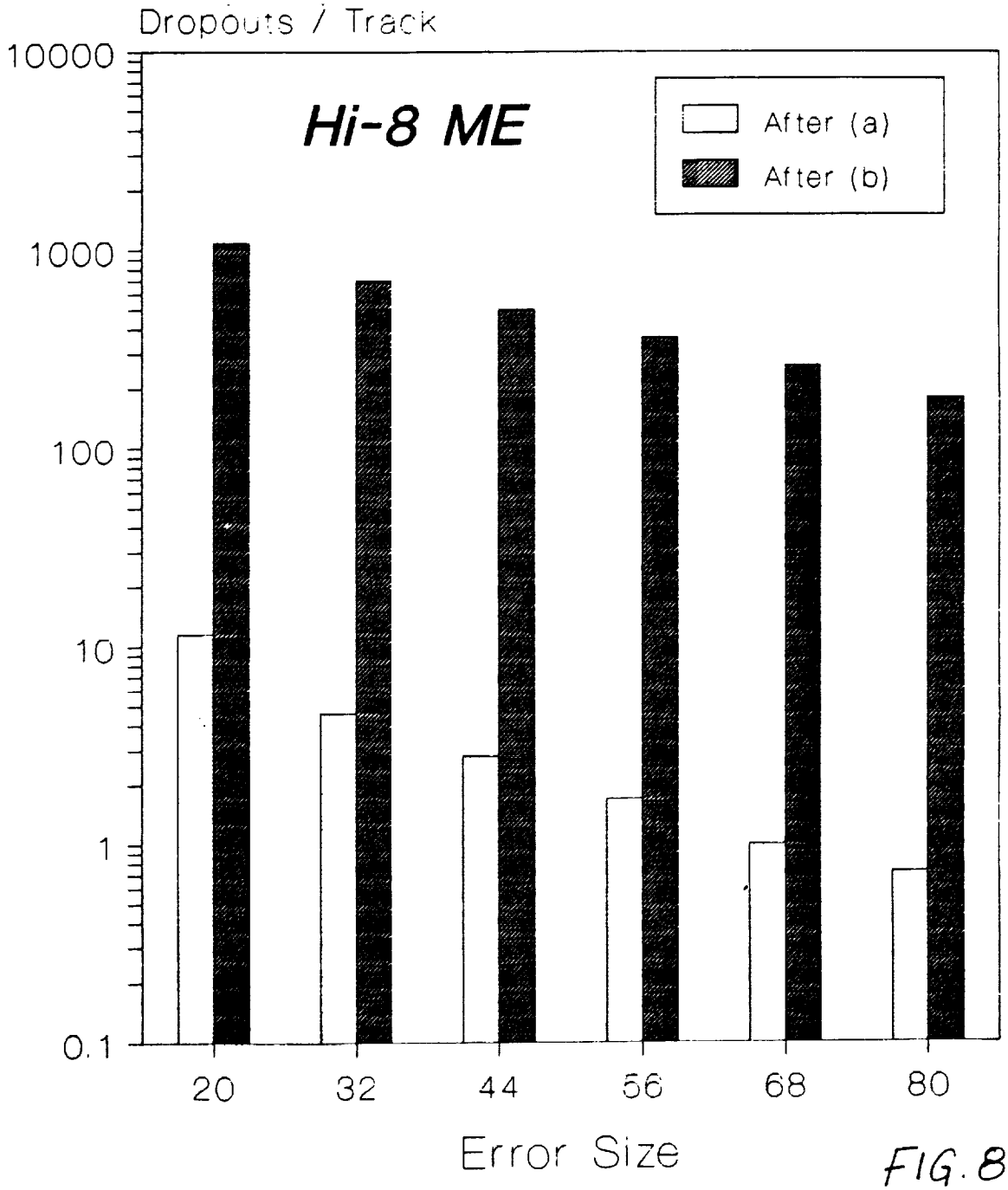
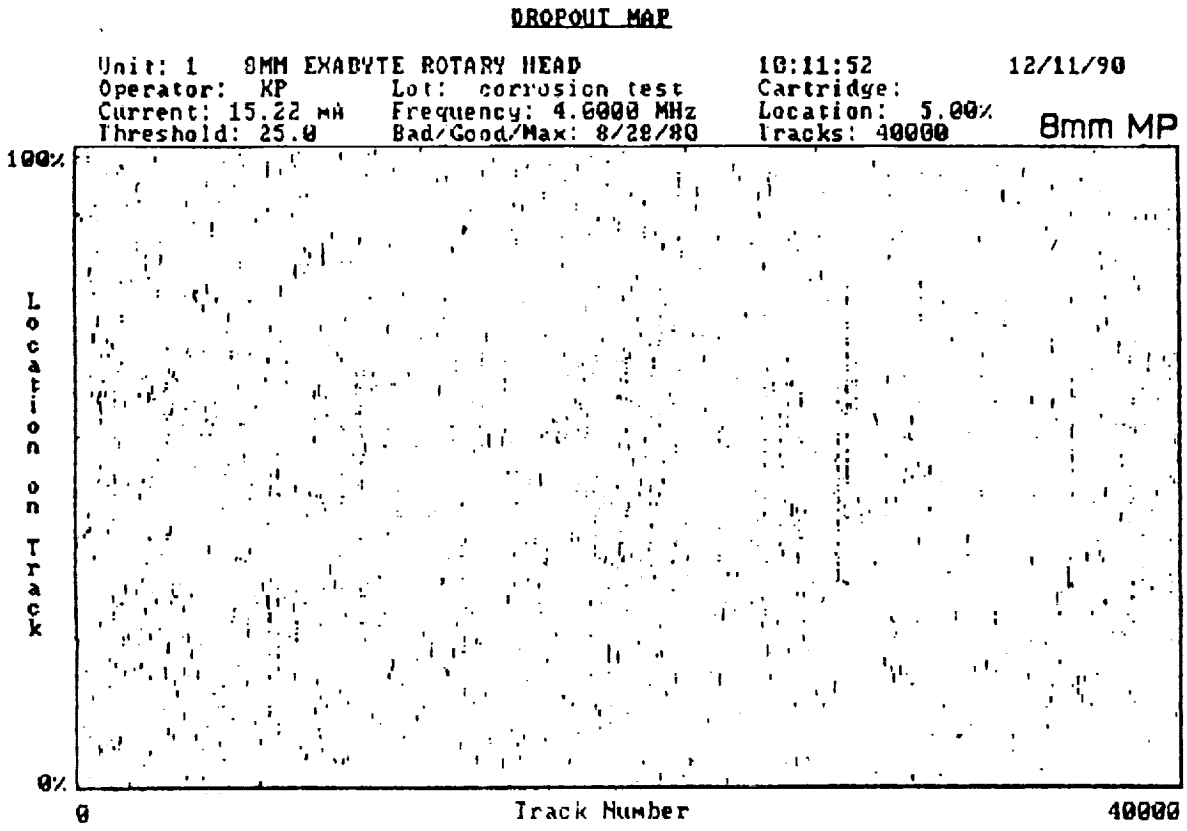
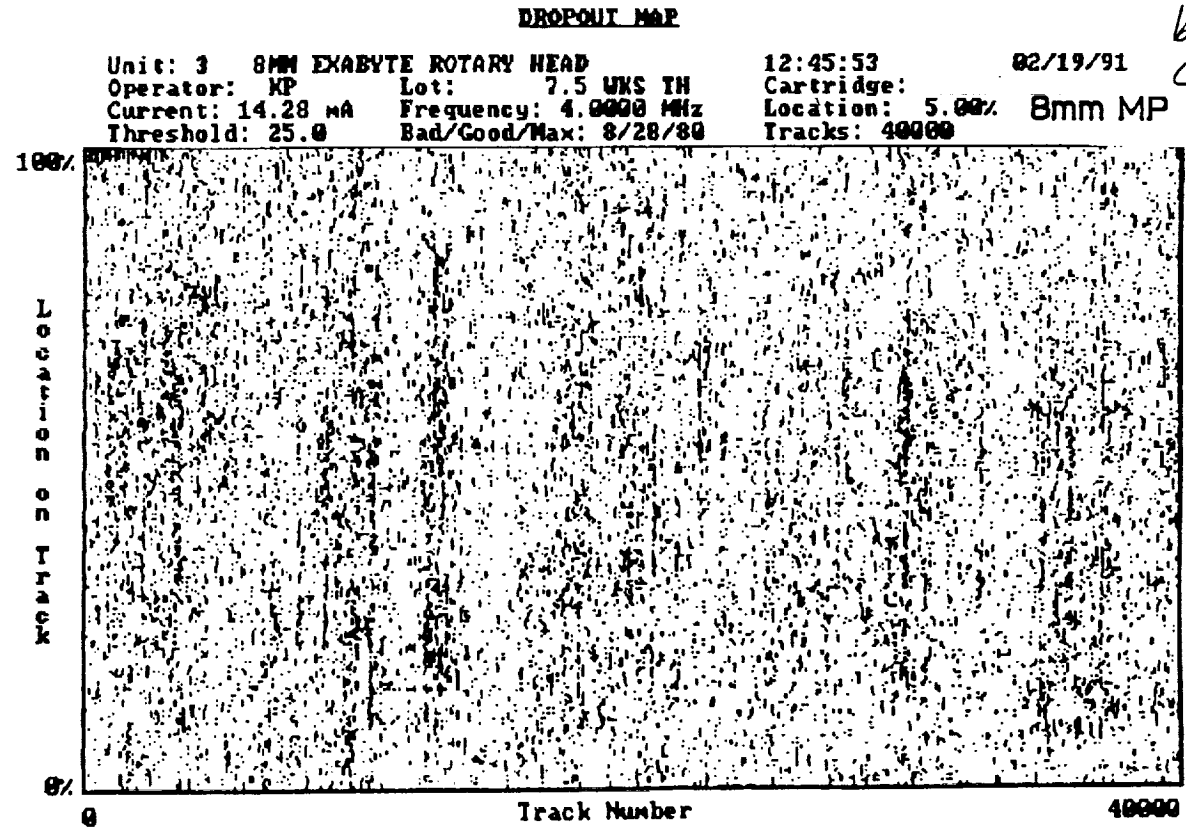


FIG. 8
Two different
Hi8ME Tapes
after Corrosion



Before

FIG. 9
Error maps of 8mm MP Tape before and after corrosion



After

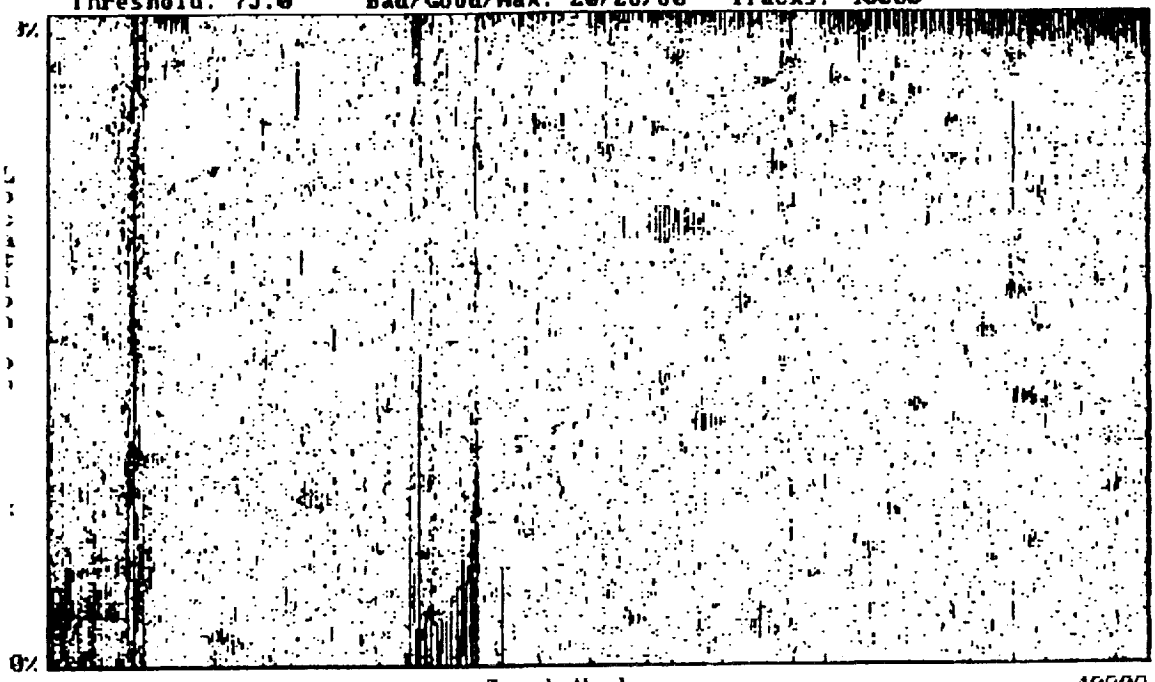
FIG.10

Error maps of two different Hi8ME tapes after corrosion

DROPOUT MAP

Unit: 1 8MM EXABYTE ROTARY HEAD 08:07:45
Operator: DES Lot: 7.5 UKS TH Cartridge:
Current: 13.40 mA Frequency: 4.0000 MHz Location: 0.50%
Threshold: 75.0 Bad/Good/Max: 20/28/80 Tracks: 40000

HI-8 ME



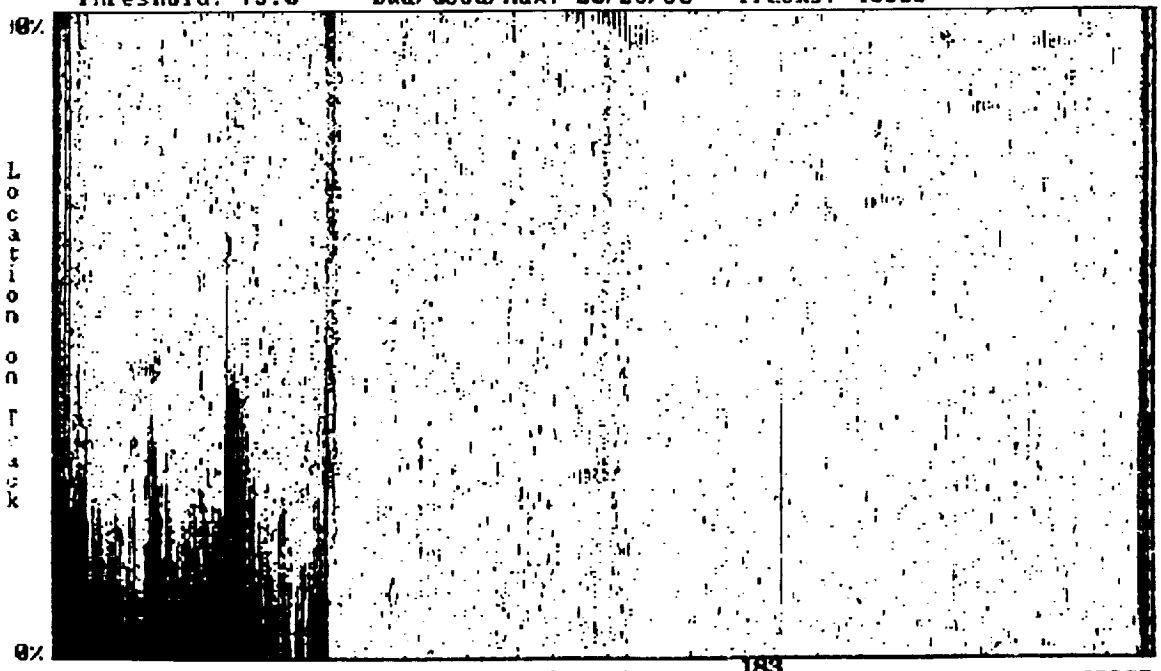
Total Errors: 58540

After (a)

DROPOUT MAP

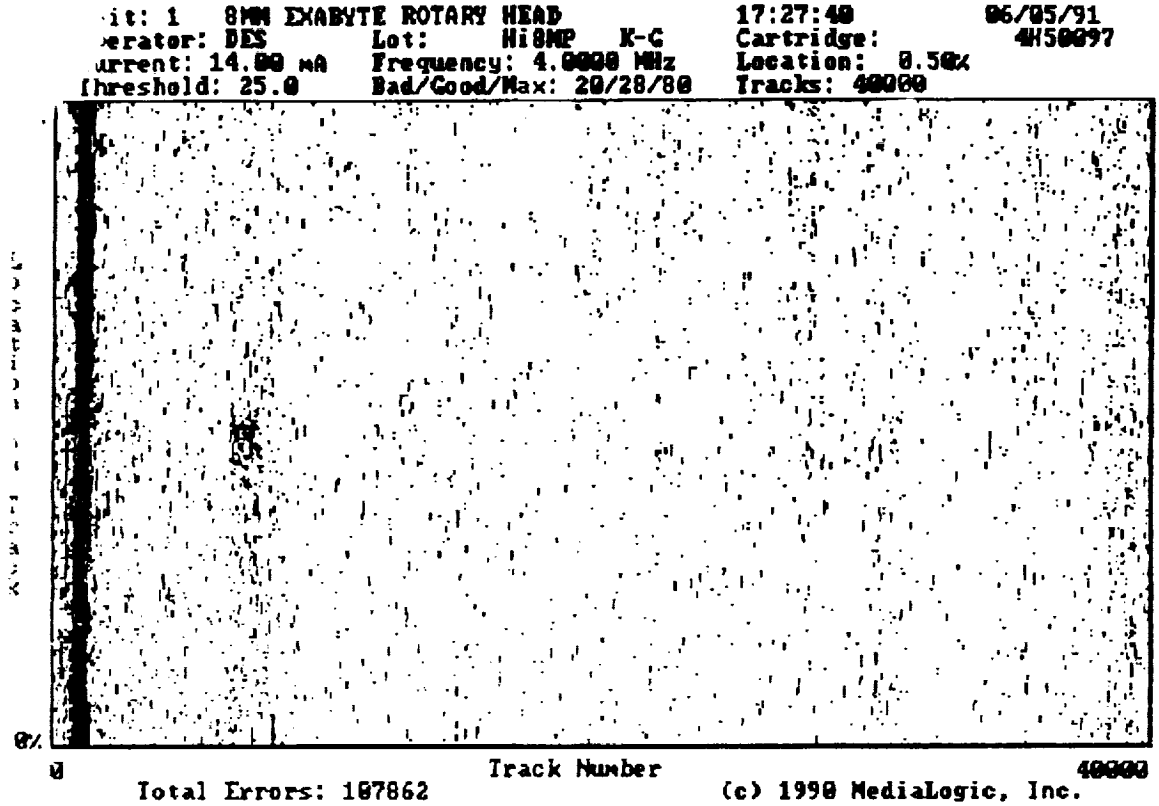
Unit: 1 8MM EXABYTE ROTARY HEAD 11:05:25
Operator: DES Lot: 7.5 UKS TH Cartridge:
Current: 11.93 mA Frequency: 4.0000 MHz Location: 0.50%
Threshold: 75.0 Bad/Good/Max: 20/28/80 Tracks: 40000

HI-8 ME



Total Errors: 388968

After (b)



DROPOUT MAP

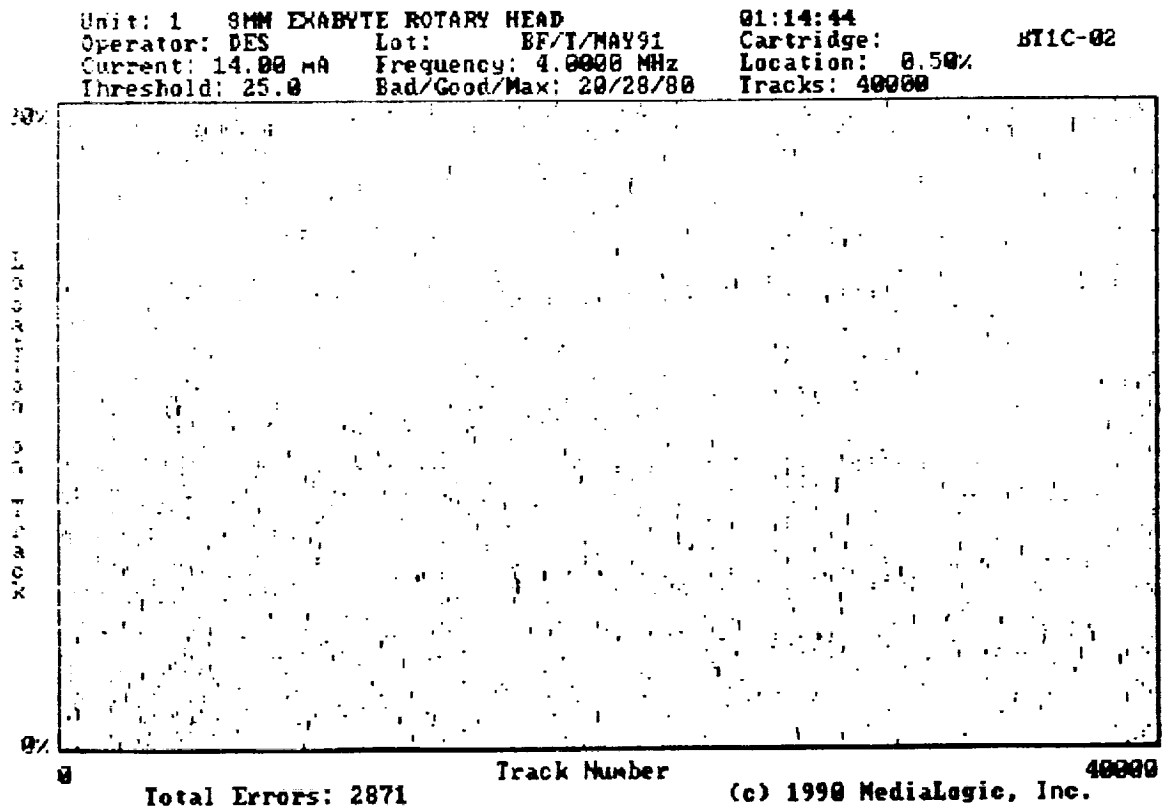
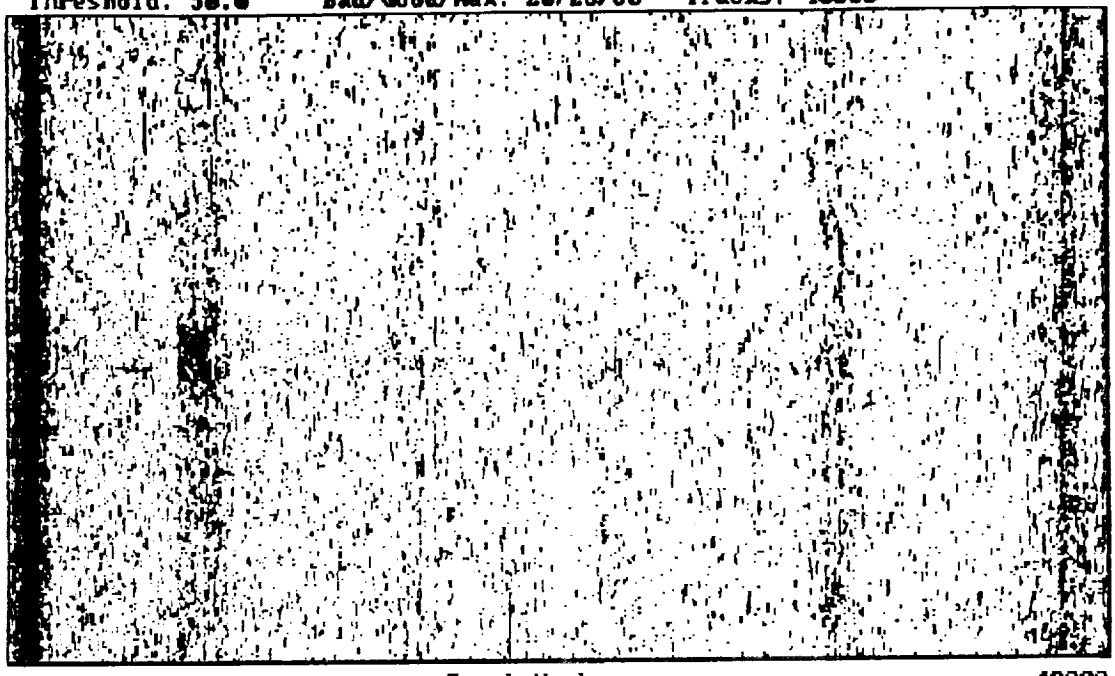


FIG. 11

Error maps of Hi8MP and Ba Tapes after 1 me testing.

DROPOUT MAP

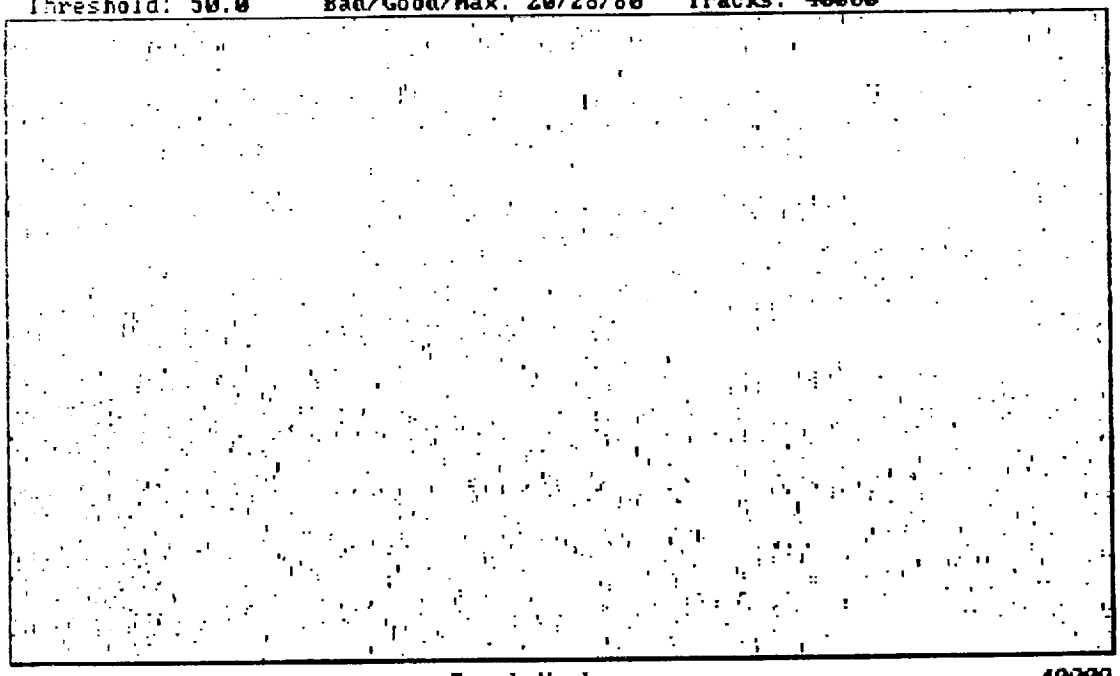
Unit: 1 8MM EXABYTE ROTARY HEAD 16:58:41 06/05/91
Operator: DES Lot: Hi8MP X-G Cartridge: 4H50097
Current: 14.00 mA Frequency: 4.0000 MHz Location: 0.50%
Threshold: 50.0 Bad/Good/Max: 20/28/80 Tracks: 40000



Total Errors: 236298 (c) 1990 MediaLogic, Inc.

DROPOUT MAP

Unit: 1 8MM EXABYTE ROTARY HEAD 16:15:47 06/05/91
Operator: DES Lot: BF/I/MAY91 Cartridge: BT1C-02
Current: 14.00 mA Frequency: 4.0000 MHz Location: 0.50%
Threshold: 50.0 Bad/Good/Max: 20/28/80 Tracks: 40000



Total Errors: 3154 (c) 1990 MediaLogic, Inc.

FIG. 12
Error maps of Hi8MP and BaFe Tapes after 1 mon testing.

A SYSTEM APPROACH

TO

ARCHIVAL STORAGE

March 20, 1991

John W. Corcoran

Introduction

When I found Bill Doyle had scheduled me to talk between Bill Mulari and Dennis Speliotis, I wondered if I were the thorn between two roses, or the rose between two thorns.

The topic was provoked by the frequent repetition of two questions. The first, raised by Bill and Dennis, is:

Can D-2 iron particle tape be used for archival storage?

The other related question is:

How can acceleration factors relating short-term tests to archival life times by justified?

The reports of possible corrosion of metal particle tape which have been presented over the last 18 months have raised serious doubts about its use in the computer data community. Probably those doubts are the reason for this symposium.

Now Ampex is primarily concerned with iron particle media used in the D-2 video standard. The standard was developed starting in 1985 by a SMPTE group -- including broadcast studio users and manufacturers of recorders and tape -- to fill a need for a high density, high data rate, digital recorder. Six years later, the best evidence of the video industries' viewpoint on D-2 stability is that a major studio is re-recording its master tapes -- 2-inch Quad and C format -- onto D-2. In video, format obsolescence -- not archival stability -- is the life limitation. The originals are 10 to 30 years old.

Ampex Recording Systems is now transferring D-2 video technology to data storage applications, and encountering concerns about corrosion that the title of this symposium suggests. To protect the D-2 standard, RSD, with the cooperation of most of the tape manufacturers, had Battelle test all four in the Class II environment. Error rates were measured before and after the test on both exposed and control groups. Correlating the before and after data on the groups shows no degradation on a 28-day test -- 14-year equivalent life.

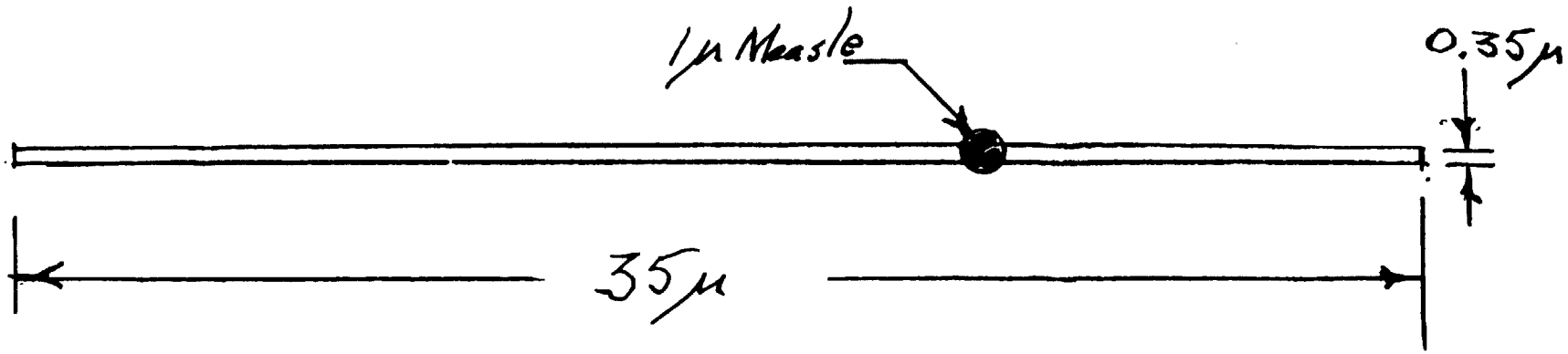
Similar results have been reported in Trans Mag and other journals by Sony, Maxell, Fuji and DEC. What is the explanation for

the contradiction between these reports and those of Bill Mulari and Dennis Spellotis? (Dennis is now inn a straddle position; he is co-author of the DEC paper that concludes metal particle tapes are stable when stored in cassettes.)

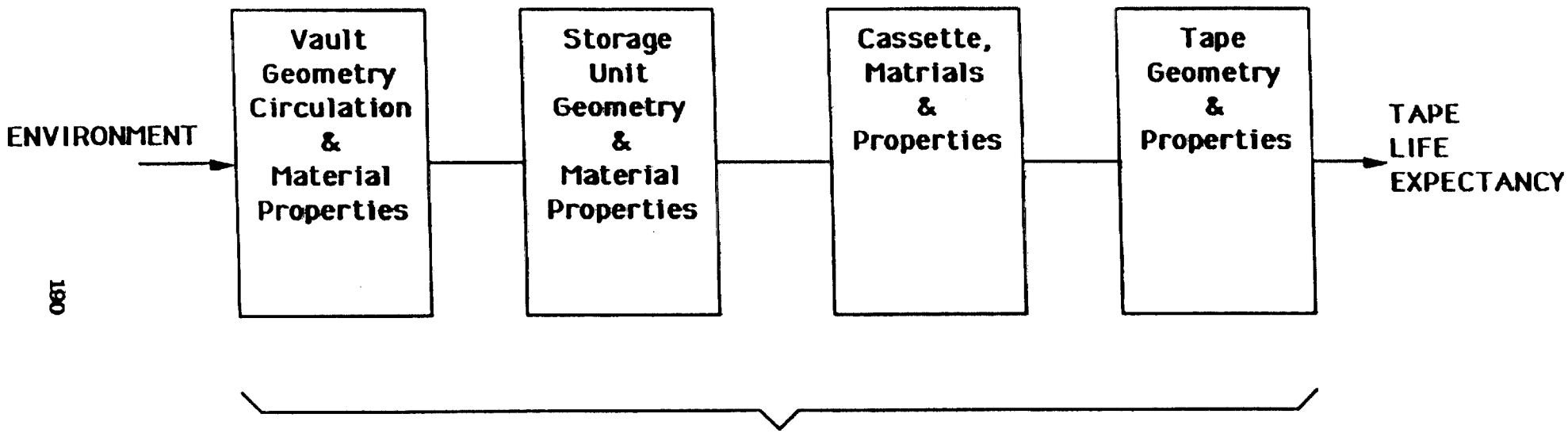
Well, this question takes us back to the first two and the topic of this talk. As is commonly the case in such differences, the results were obtained under different conditions. As will be shown, tapes stored in cassettes are invulnerable to corrosion. That is not surprising; cassettes were developed to protect tapes from the hazards of reel-to-reel recorders.

Bill Abbott commented that the measurements of the shielding effect of cassettes showed a "qualitatively different type of behavior than 'naked' tapes." To explain this requires a system approach.

Comparison of a Bif Cell & a Measle



BLOCK DIGARAM OF AN ARCHIVAL STORAGE SYSTEM

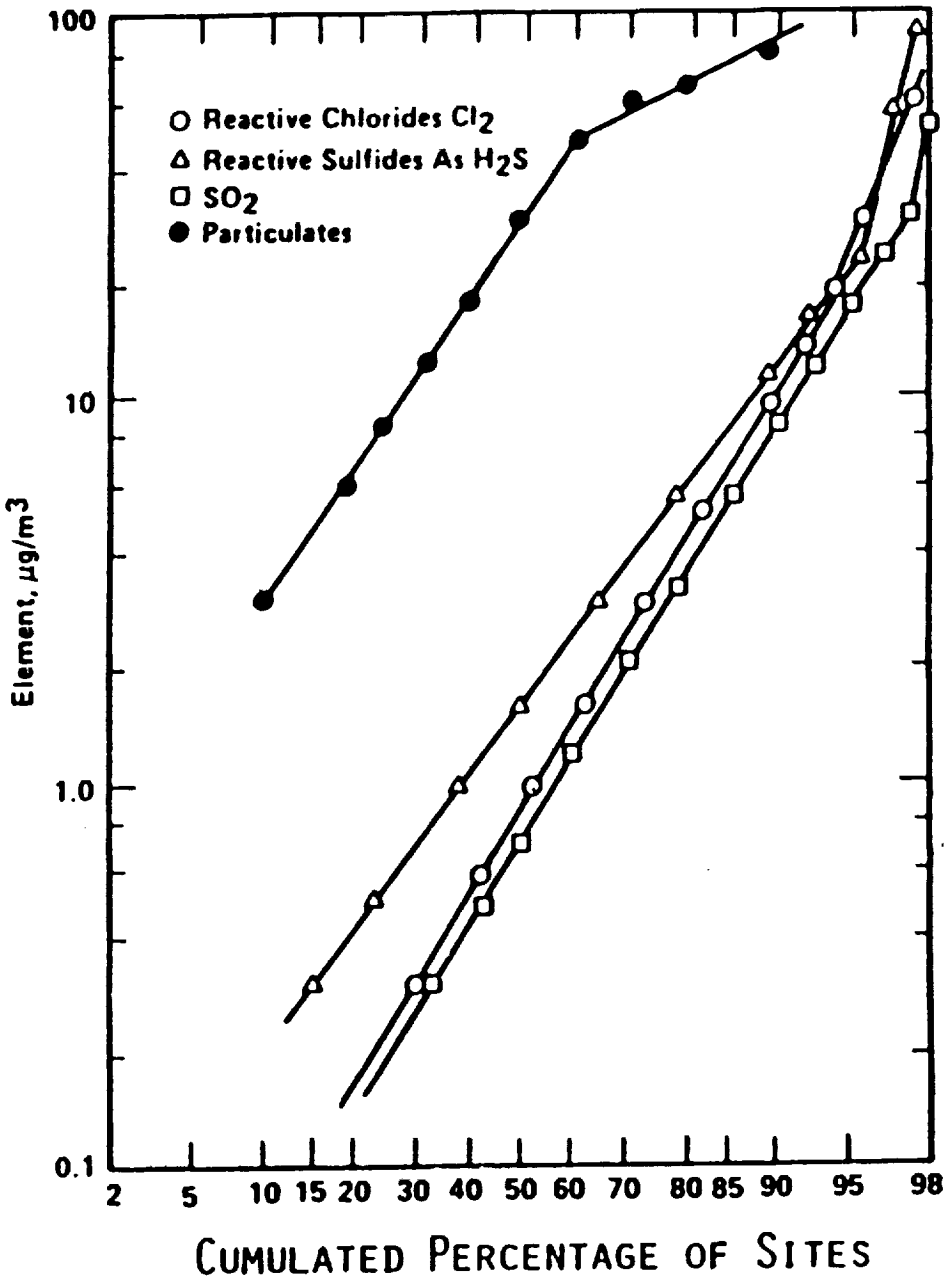


160

Concentration of Agents
Flow Rates
Reaction Rates
Time Constants
Attenuation Factors

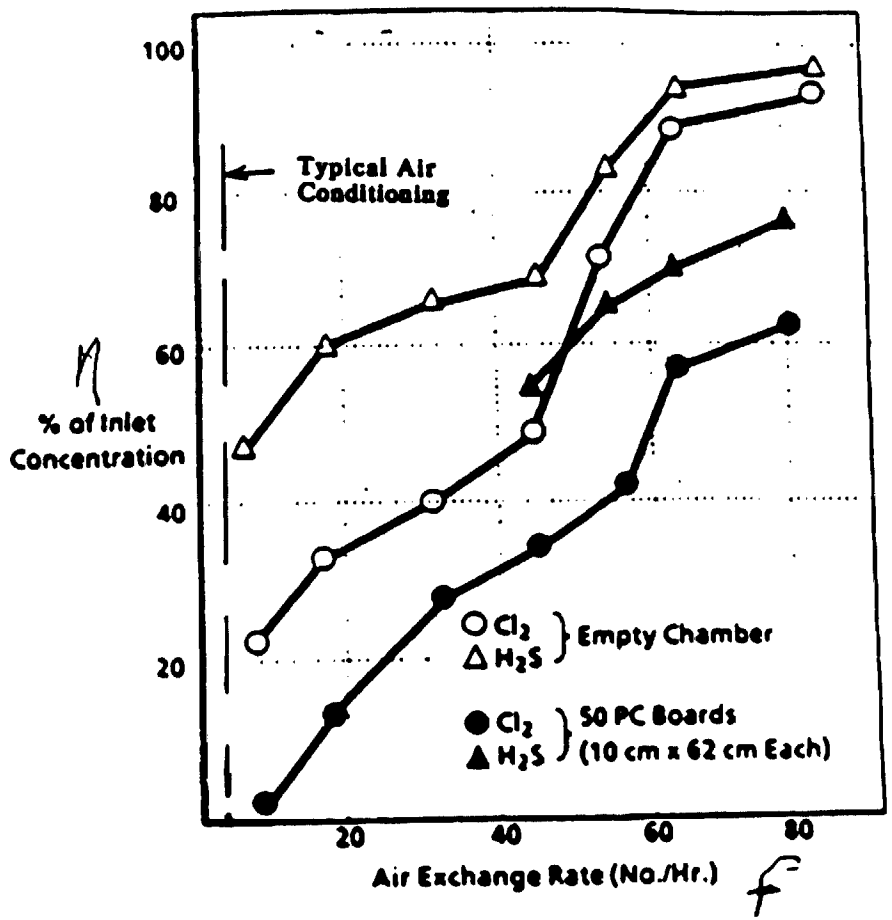
} ≈

Potential
Current



WORLDWIDE INDOOR POLLUTANT DISTRIBUTIONS FOR ELECTRICAL AND ELECTRONIC EQUIPMENT

ADSORPTION IN BATTELLE TEST CHAMBERS



Gas Adsorption Rate

$$\frac{dQ}{dt} = \{[G]_{in}-[G]_{out}\}Vf$$

$$= k\,A\,[G]_{Avg}$$

$$\frac{2}{V}\Sigma\, k_i\, A_i = \frac{1-\eta}{\eta} f$$

Adsorption Cofefficents

k_i cm/sec

Gas	Chamber Walls	Circuit Boards
Cl_2	0.20	0.093
H_2S	0.086	0.035

GENERAL CORROSION REACTION

$$M + G \longrightarrow MG$$

M Metal
G Gas

RATE FOR SECOND ORDER REACTION (Brackets indicate concentration)

$$\frac{d\,[M]}{dt} = \frac{d\,[G]}{dt} = -k\,[M]\,[G]$$

For [G] constant (as in Battelle chamber)

FOR FIRST ORDER REACTION

$$M = M_o \exp\, -t/\tau$$

$$\tau = \frac{1}{k\,[G]}$$

For low gas concentrations, the time constant is very long. Also, for $t \leq \tau$ the decay rate is constant.

Time Constants for Vaults Without Air Circulation

For Diffusion
in a tunnel of length
d :

$$\tau = \frac{d^2}{3\Delta}$$

For chlorine in air, $\Delta \approx 0.1\ cm^2/sec$ and for $d = 10\ m$, $\tau = 10^3\ hr$ or 40 days.

For Thermal Pumping

A daily temperature variation of 3°/k will cause air to flow in and out of any unsealed container or space, thereby carring corrosion agents into the container. Using perfect gas law, the daily mass flow is

$$\frac{1}{M_c}\ \frac{\Delta M}{\Delta t} = \frac{\Delta T}{\overline{T}\ \Delta T}$$

where M_c is the mass of air in the container and $\overline{T}$ is the average absolute temperature. The time constant is the reciprocal of these terms, about 1200 hours, 50 days.

For Barometric Pumping

Similarly

$$\tau = \frac{\overline{P}\ \Delta t}{\Delta P}$$

For $\overline{P}$ = 30 in. Hg, $\Delta P = 1in$ $\Delta t = 3$ days, $\tau = 90$ days.

AMPEX

ERROR RATE MEASUREMENT

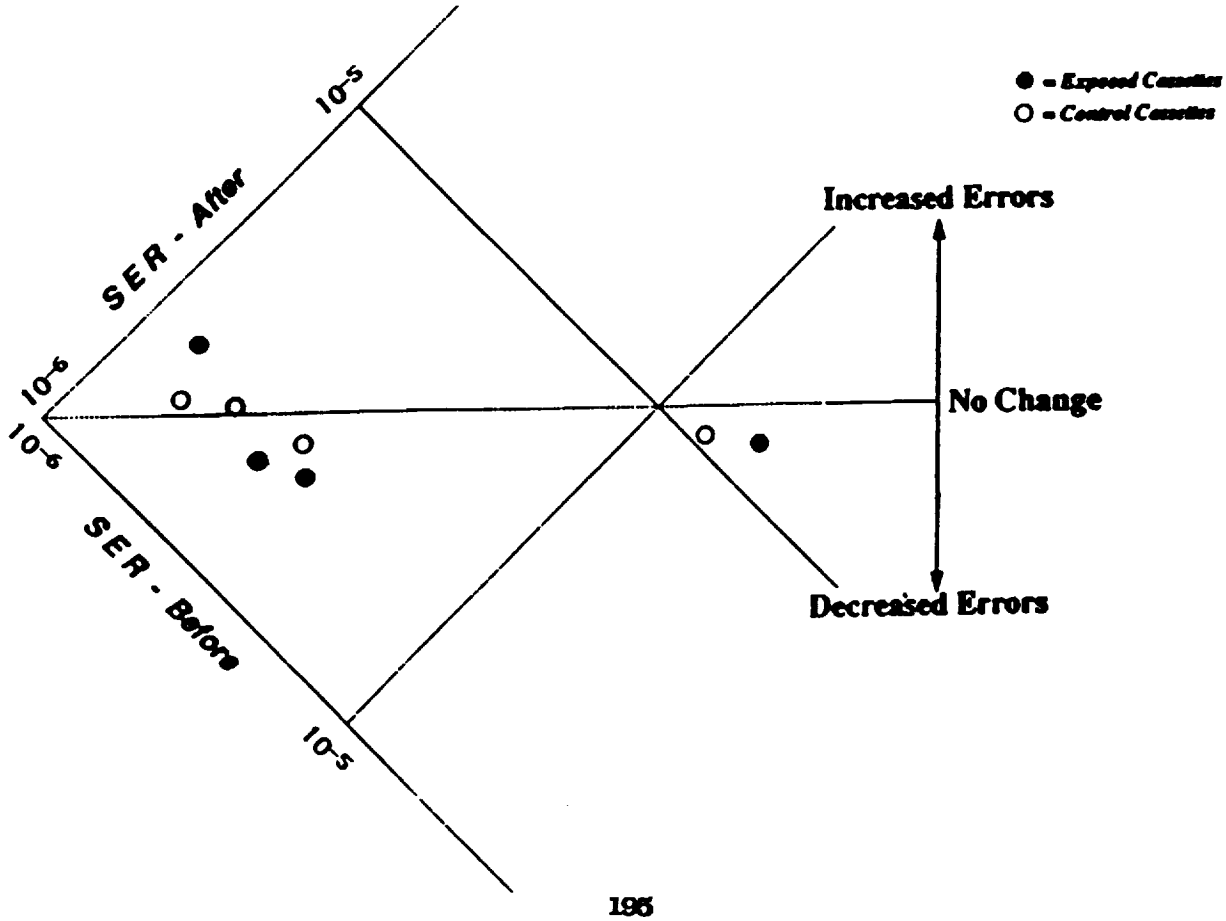
ALL ERROR DATA PRESENTED IS "RAW" TO PERMIT DETECTION OF ANY CHANGE IN THE MEDIA. (WITH ERROR CORRECTION APPLIED, MOST CASSETTES WOULD RUN ERROR-FREE, AND CORROSION EFFECTS WOULD BE UNDETECTABLE).

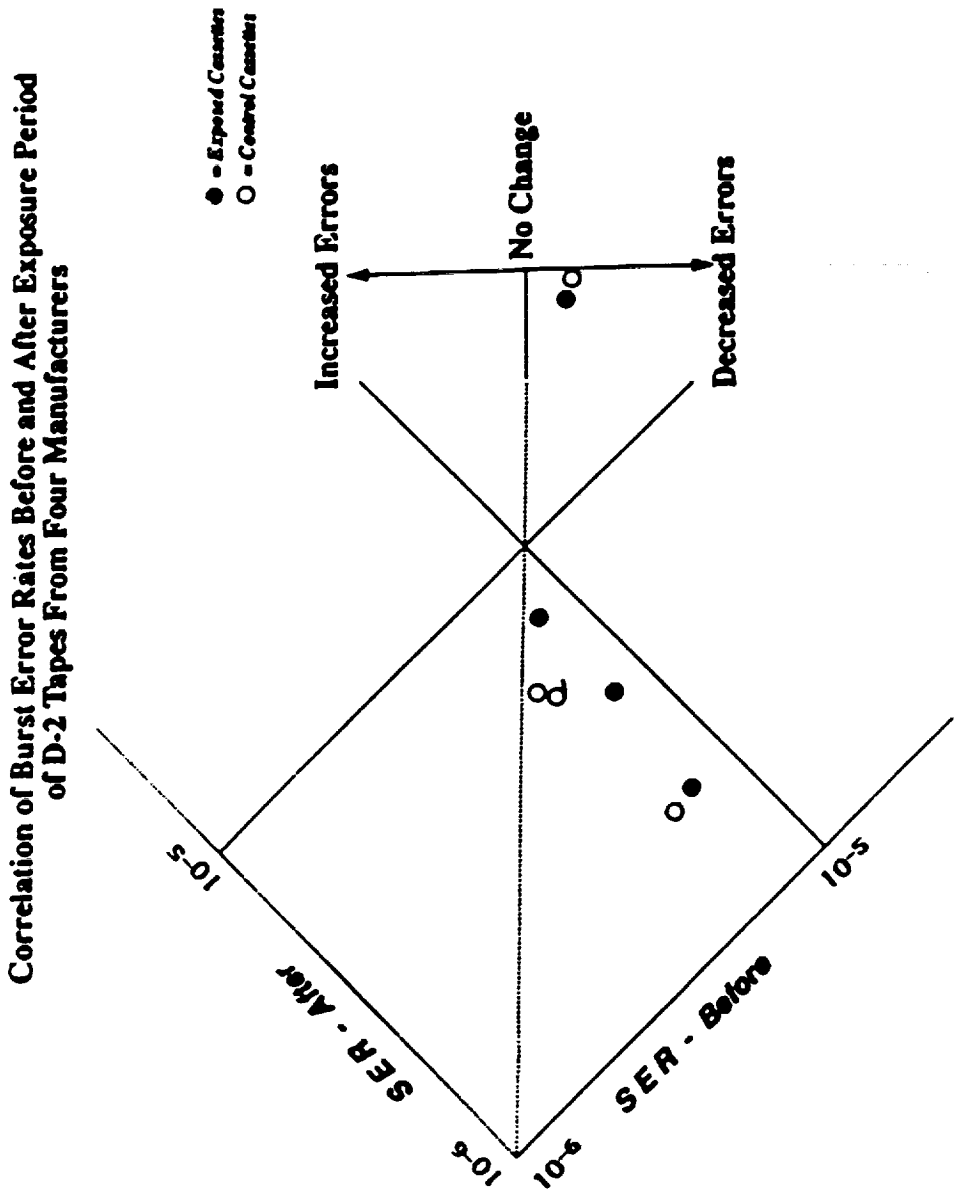
SYMBOL (OR BYTE) ERROR RATE (SER) IS USED FOR D-2 DATA BECAUSE THE REED-SOLOMON ERROR CODING IS SYMBOL ORIENTED.

BURST AND NON-BURST ERROR DATA ARE PRESENTED. NON-BURST DATA INDICATE ERRORS IN INDIVIDUAL BITS DUE TO LOSS OF SNR, MEASLES, ETC. BURST ERRORS ARE PRIMARILY DUE TO SURFACE DEFECTS.

POST TEST DATA ARE PLOTTED AS A FUNCTION OF PRE-TEST FOR THE SAME CASSETTE. POINTS BELOW THE "NO CHANGE" LINE INDICATE IMPROVEMENT.

Correlation of Non-Burst Error Rates Before and After Exposure Period of D-2 Tapes From Four Manufacturers




Correlation of Burst Error Rates Before and After Exposure Period
of D-2 Tapes From Four Manufacturers
● - Exposed Cassettes
○ - Control Cassettes
Increased Errors
No Change
Decreased Errors
10-5
SER - After
10-6
10-6
SER - Before
10-5

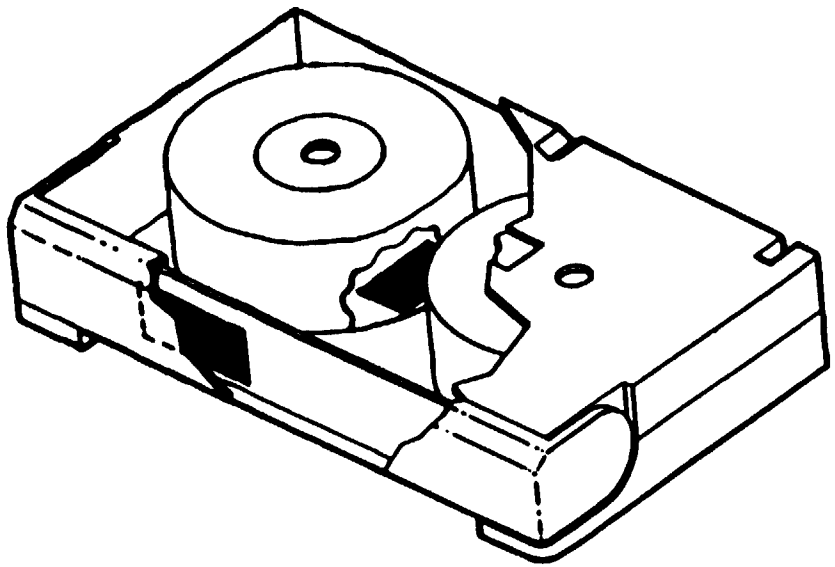
SHIELDING TEST

AMPEX

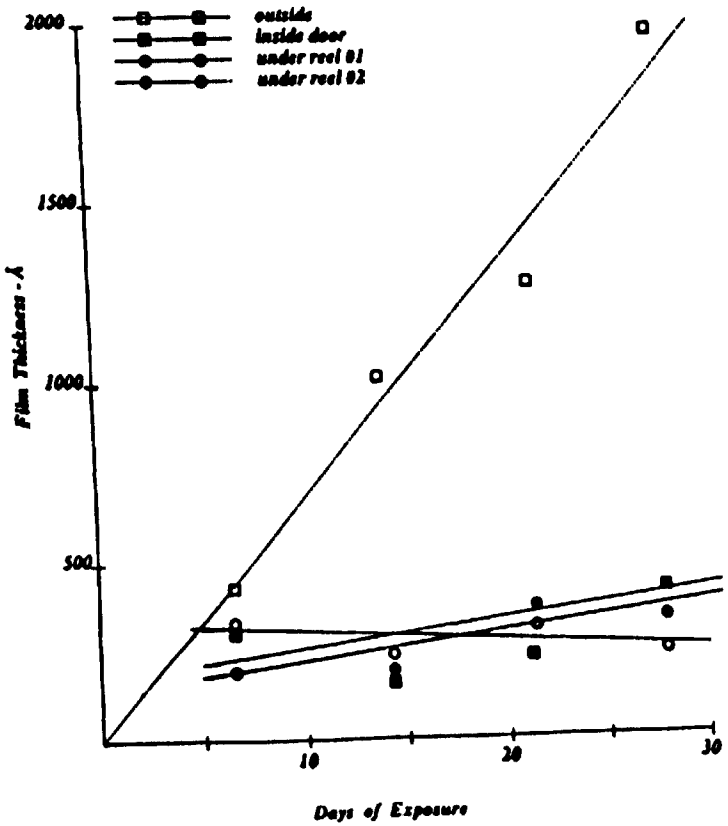
STANDARD D-2 CASSETTES (SMALL SIZE) WERE FITTED WITH BATTELLE METAL COUPON SAMPLERS AND EXPOSED FOR 28 DAYS. A COMPARISON OF RESULTS OF THREE INTERNAL SAMPLERS WITH AN EXPOSED ONE INDICATES THE CASSETTE HAS ATTENUATED THE CORROSIVE ENVIRONMENT TO NEGLIGIBLE PROPORTIONS.

THE DIFFERENCE BETWEEN THE PRESENT RESULTS AND THOSE REPORTED BY SPELIOTIS FOR "NAKED" TAPE ARE ATTRIBUTABLE TO THE CASSETTE SHIELDING FACTOR.

Sampler Locations on D-2 Cassette

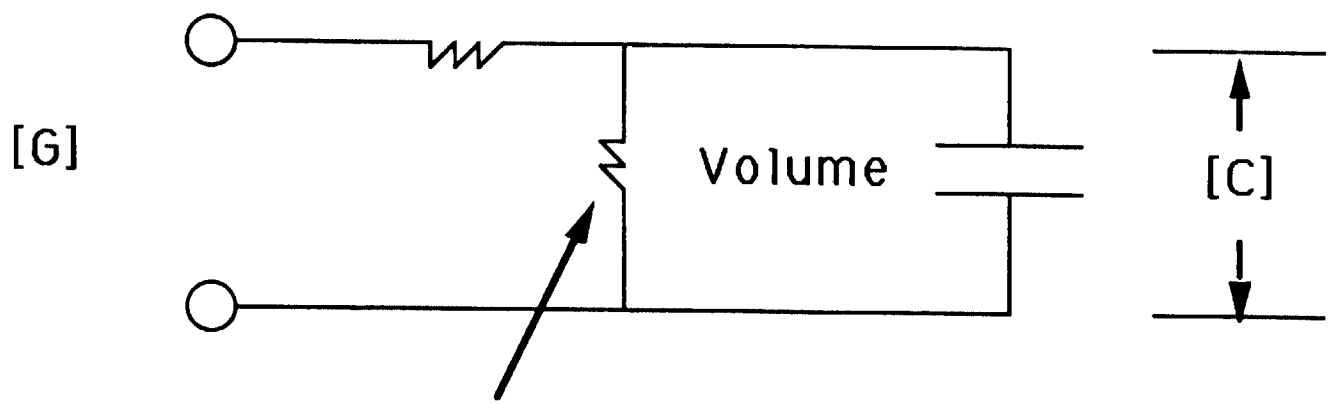


Silver Sampler Film Measurements for Various Locations on D-2 Cassette



CASSETTE IMPEDANCE MODEL

Diffusion Resistance $R_D = \frac{\ell}{\Delta A_x}$



$$\frac{[C]}{[G]} = \frac{1}{\frac{K_w A_w \ell}{\Delta A_x} + 1}$$

Attenuation Factor

DIFFUSION INTO REELS

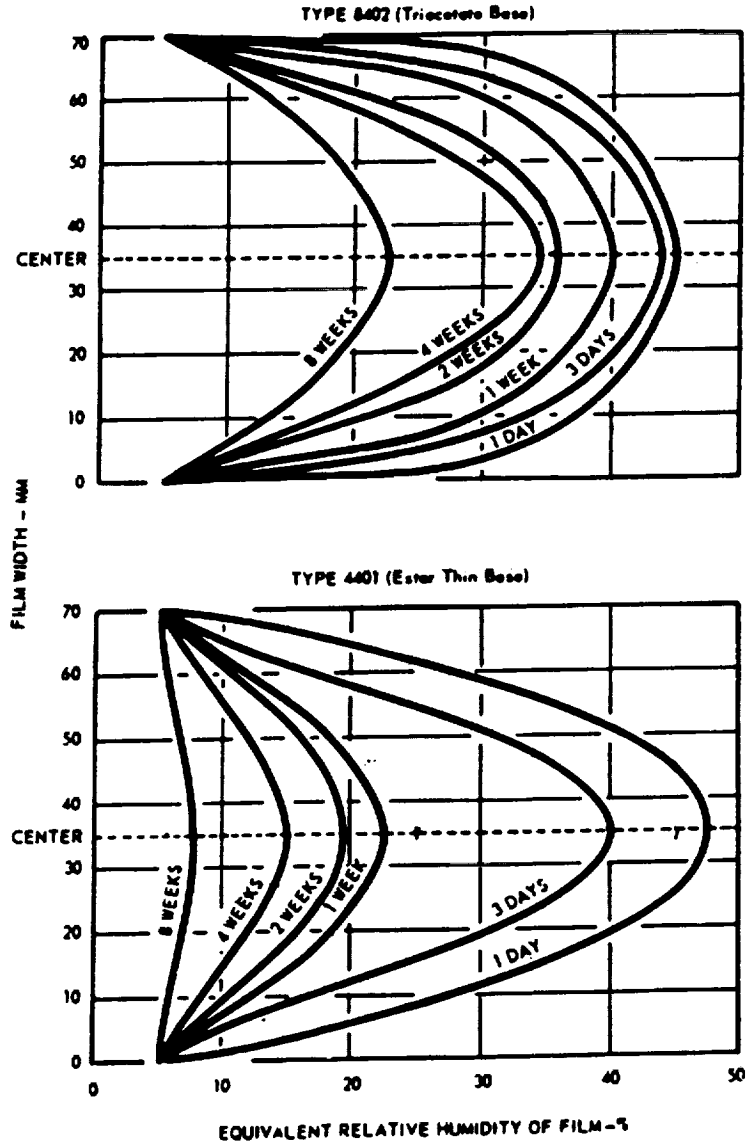


Fig. 6.10 Moisture Gradients During Conditioning of Aerial Film Rolls at 70°F -5% RH

Data from "Manual of Physical Properties" - Eastman Kodak Co.

DETERMINATION OF TIME CONSTANTS FOR MOISTURE DIFFUSION

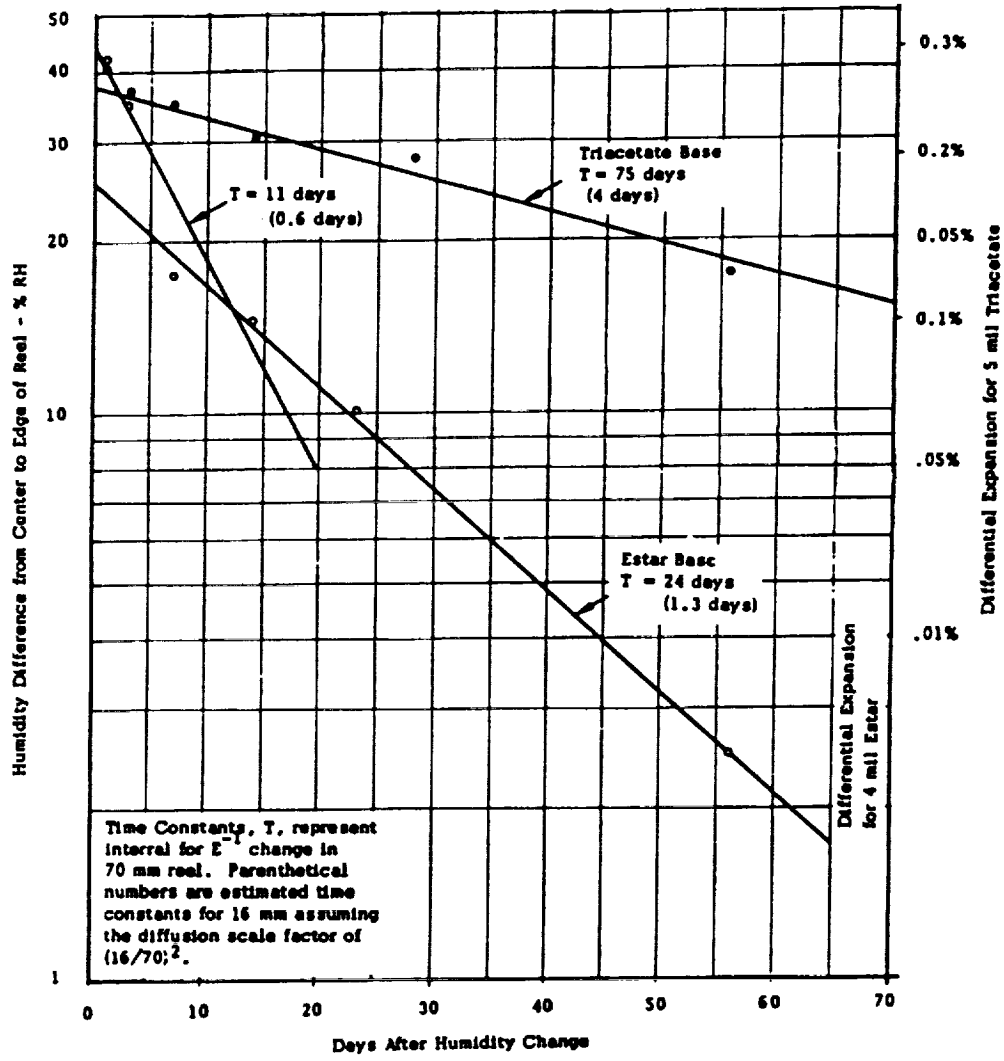


Fig. 6.11 Difference in relative humidity between center to edge of a 70 mm reel following a step change from 70°F - 50°$_\lambda$ RH

$$\tau = T \left\{ \frac{W}{70\ \text{mm}} \right\}^2$$

FROM BERTRAM & ESHEL

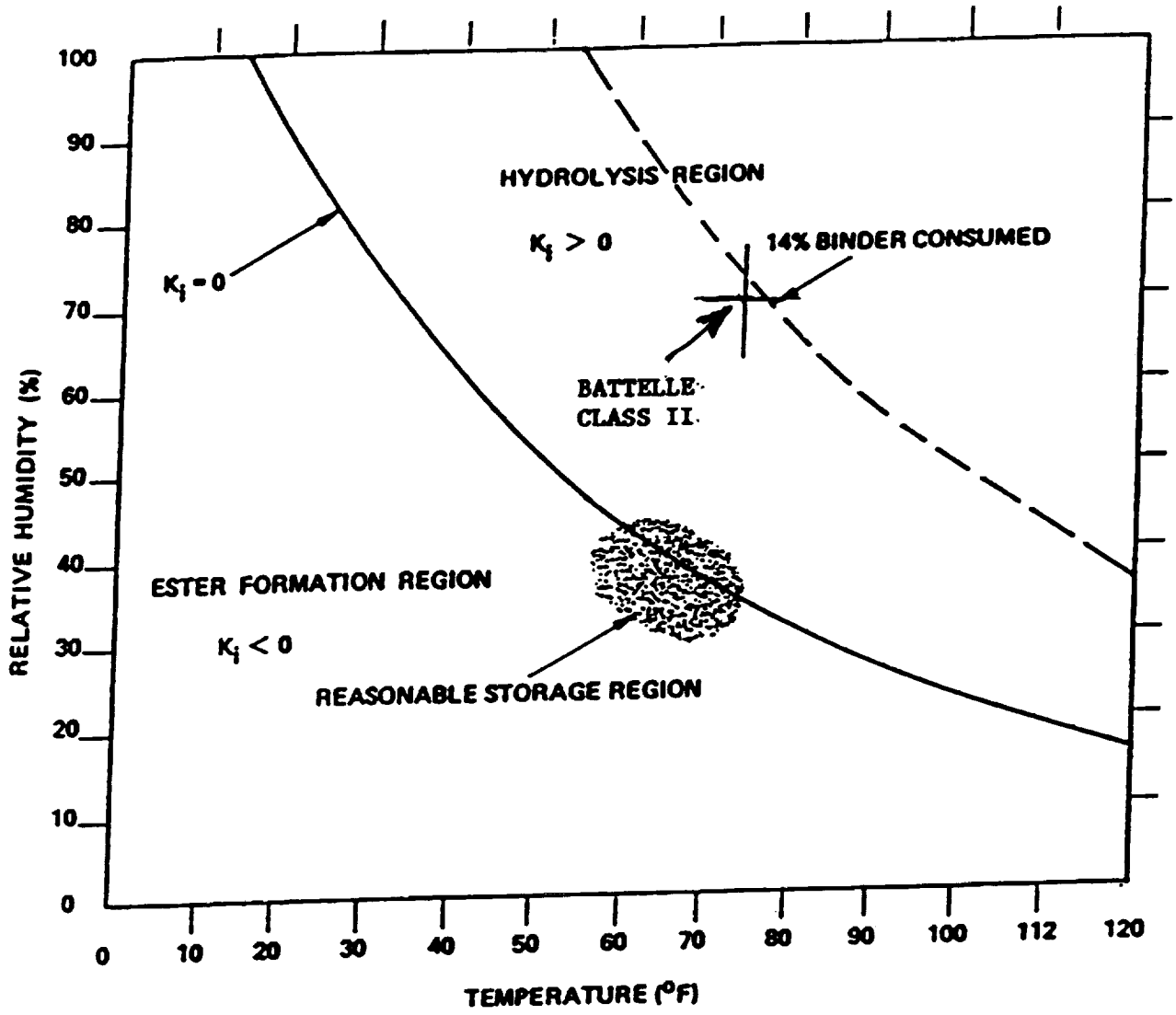


Figure 37 Hydrolysis equilibrium curves vs humidity and temperature. $K_i = 0$ indicates initial equilibrium. Dashed curve denotes 14% binder hydrolysed from an initial 6.7%.

Conclusions

1. The archival life of four brands of D-2 tape have been shown to exceed 14 years in the Battelle Class II environment. No evidence of corrosion was found.
2. The cassette is a necessary element for achieving this life. Dangling tape out of a cassette invites failure.
3. Extended exposure of any type of tape to high temperature and humidity causes binder degradation and coating failure. Archival storage is not possible in such environments.
4. An archival storage system includes: the tape, its cassette, other protective enclosures (if used), the storage vault, and its material parameters and the environment. Methods of determining time constants and attenuation factors for estimating storage life have been suggested.
5. There is a need for research -- perhaps here at the University of Alabama -- to determine adsorption coefficients of tapes, cassettes and other materials used in archival storage to use in life predictions.

REFERENCES

1. Abbott, William G. - Development and Performance Characteristics of Mixed Flowing Gas Test Environment - IEEE Trans on Components Etc., VII, p. 22 Mar. 1988.
2. Cuddihy, Edward F. - Aging Correlation (RH + T) Relative Humidity (%) + Temperature (C) - Corrosion Science, V27, No. 5, p 463-474, 1989.
3. Cuddihy, Edward F. - Aging of Magnetic Recording Tape, IEEE Trans on Mag V16, No. 4, p 558-568, July 1980.
4. Bertram, Neal and Eskel, A. - Recording Media Archival Attributes (Magnetic) Contract No. F 30602-78-C-0181, Final Technical Report to Rome Air Development Center, Ampex Report RR-80-01, 30 Nov. 1979.
5. Speliotis, Dennis E., - Corrosion of Particulate and Thin Film Media, IEEE Trans Mag V26, p 124-6, Jan. 1990.
6. Wolf, Irving - Environmental Stability of D-2 Tape, THIC, Mar. 1990.
7. Thomas, Robert G. - SMPTE Study Group on New Magnetic Media, SMPTE Journal, p 1242, Dec. 1986.
8. Hatai, A. et al. - Archival Stability of Metal Video Tape/Betacam SP, Symposium on Archival Stability, Hull, Canada, May 1990.
9. Yamamoto, Y. et al. - Study of Corrosion Stability in Metal Particle Tape, IEEE Trans Mag V26, p 098-100, 1990.
10. Mukaida, Y., Practical Stability of Metal Powder Magnetic Tape, THIC, June 2, 1990
11. Metal Tape Stability, Sony, THIC, June 19, 1990...
12. Djalali, A.; Judge, J. S.; Speliotis, D., et al. Study of the Stability of Metal Particle Data Recording Tapes - Fall Meeting of Electrochemical Society, Seattle, Oct. 14, 1990.

Stability of Co-γ-Fe_2O_3 Tape

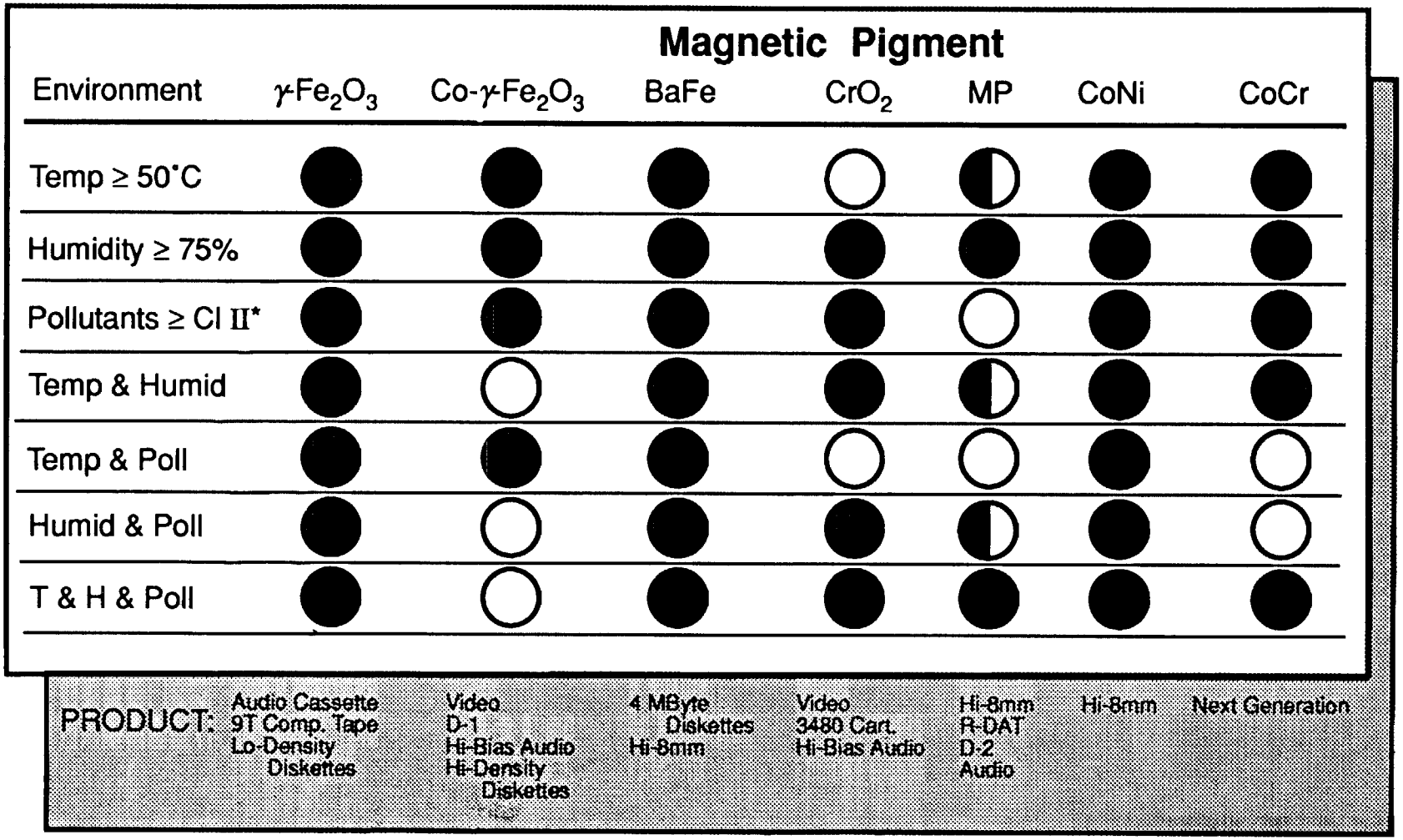
National Media Laboratory

Darlene M. Carlson

NSSDC Conference

7/24/91

Overview of Archival Stability of Recording Media



KEY: ● = GOOD — No corrosion or signal loss problems expected
○ = FAIR — May be suitable if some signal loss and/or bit errors can be tolerated
◒ = POOR — Unsuitable for storage under this condition

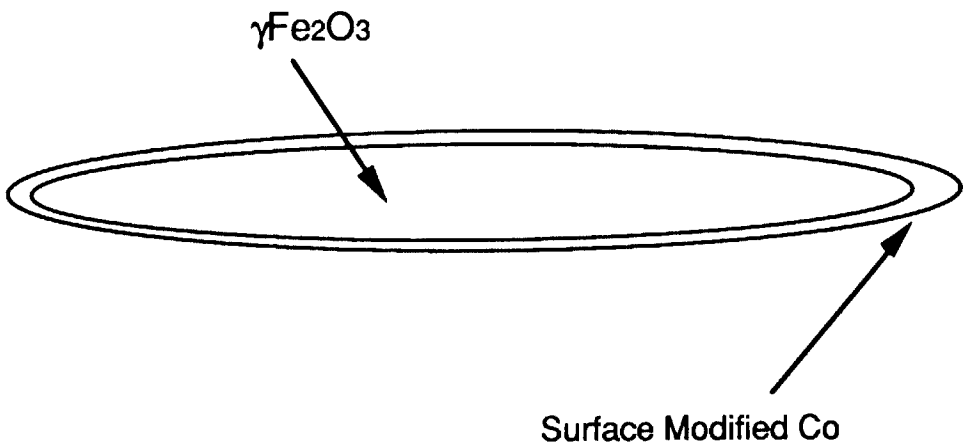
NOTES: *Interim recommendations, based on results of NML and others, as of January, 1991.*
Does not include possible binder and substrate problems not specific to media type.

**Battelle Class II Environment*

National Media Lab
1991

Co-γ-Fe_2O_3 Pigment

National Media Laboratory



Surface Adsorbed Co
Hc (oe)- 450-900
σs (emu/gm)- 84
l/d- 5:1 to 10:1

Storage Systems Using Co-γ-Fe_2O_3 Tape

National Media Laboratory

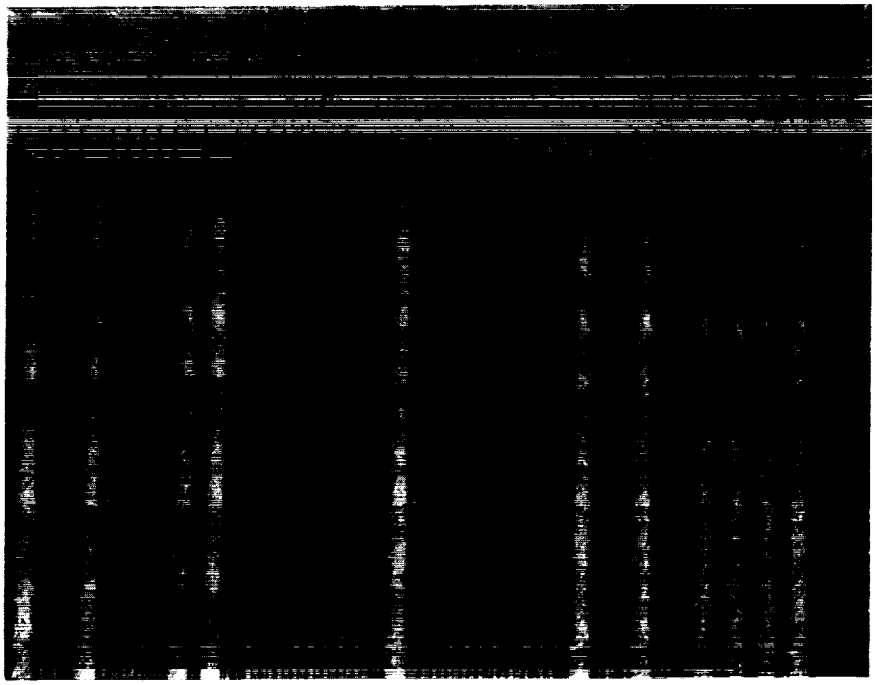
System ID	Head	Media Format	Thickness (μ)	Capacity/Pkg (Gbits)	Durability (passes)
QIC 525	Fixed	1/4" Cart.	11	4.2	10,000
QIC 1350	Fixed	1/4" Cart.	13	10.8	10,000
Interferometrics	Fixed	1"x14" reel	16	5500	1,000
DCRsi	Trans.	1" Cart.	25	380	200
VHS	Helical	1/2" Cart.	20	32	NI
Digidata	Helical	1/2" Cart.	20	43	NI
VLDS	Helical	1/2" Cart.	16	80	NI
ID1	Helical	19mm Cart.	16	300	300
DCR	Helical	1" Cart.	25	240	200

* NI - No Information

NML DMC 7/24/91

Recording Bit Density Comparison

Linear, Low Density: 2060 bits per inch, 20 mil width; 165X.

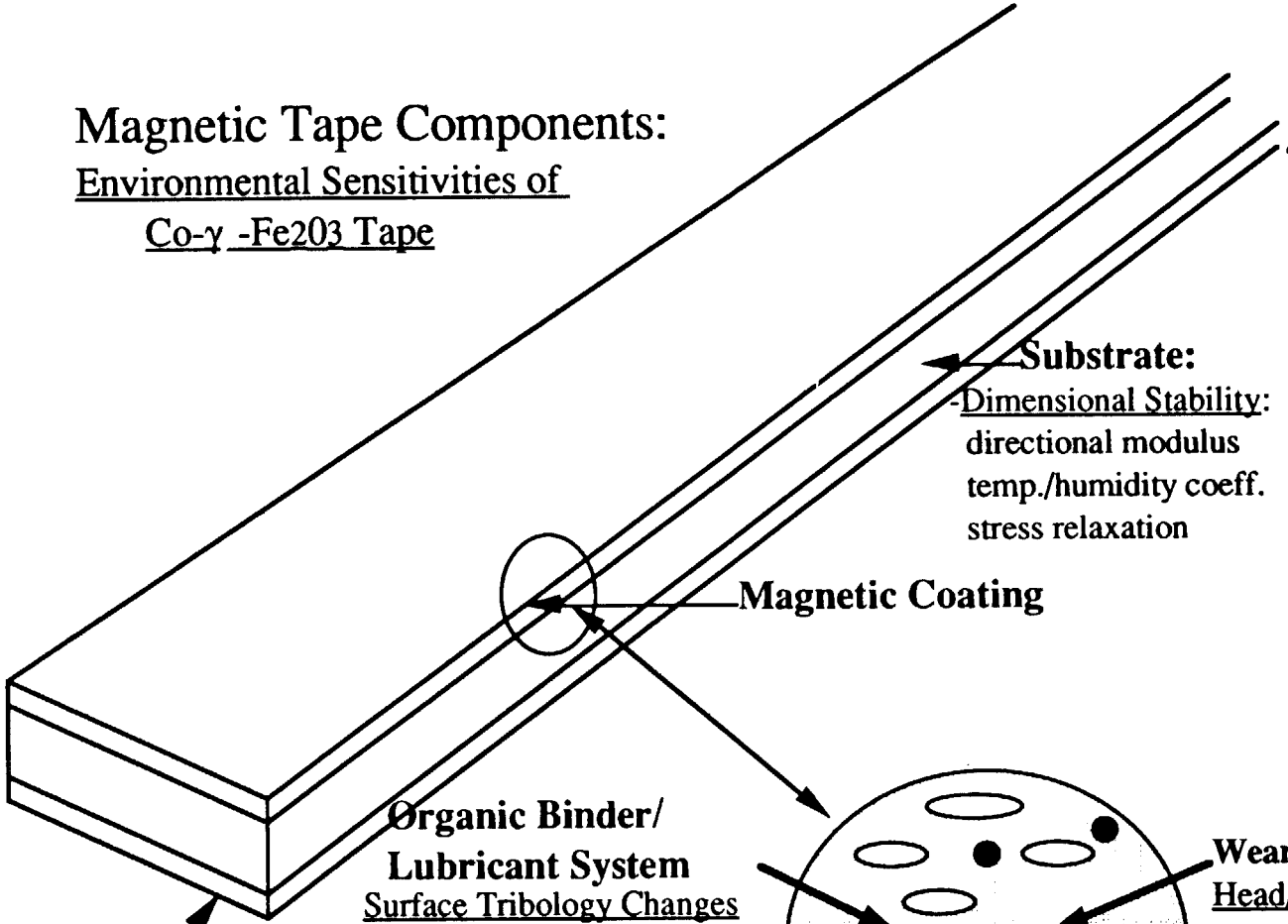


Helical, High Density: 45,000 bits per inch, 1.2 mil width; 165X.

National Media Laboratory

Magnetic Tape Components:

Environmental Sensitivities of
Co-γ -Fe2O3 Tape



Substrate:
-Dimensional Stability:
directional modulus
temp./humidity coeff.
stress relaxation

Magnetic Coating

Organic Binder/ Lubricant System
Surface Tribology Changes
stick-slip
swelling...
Composition Changes
hydrolysis
molecular wt. change
lubricant loss

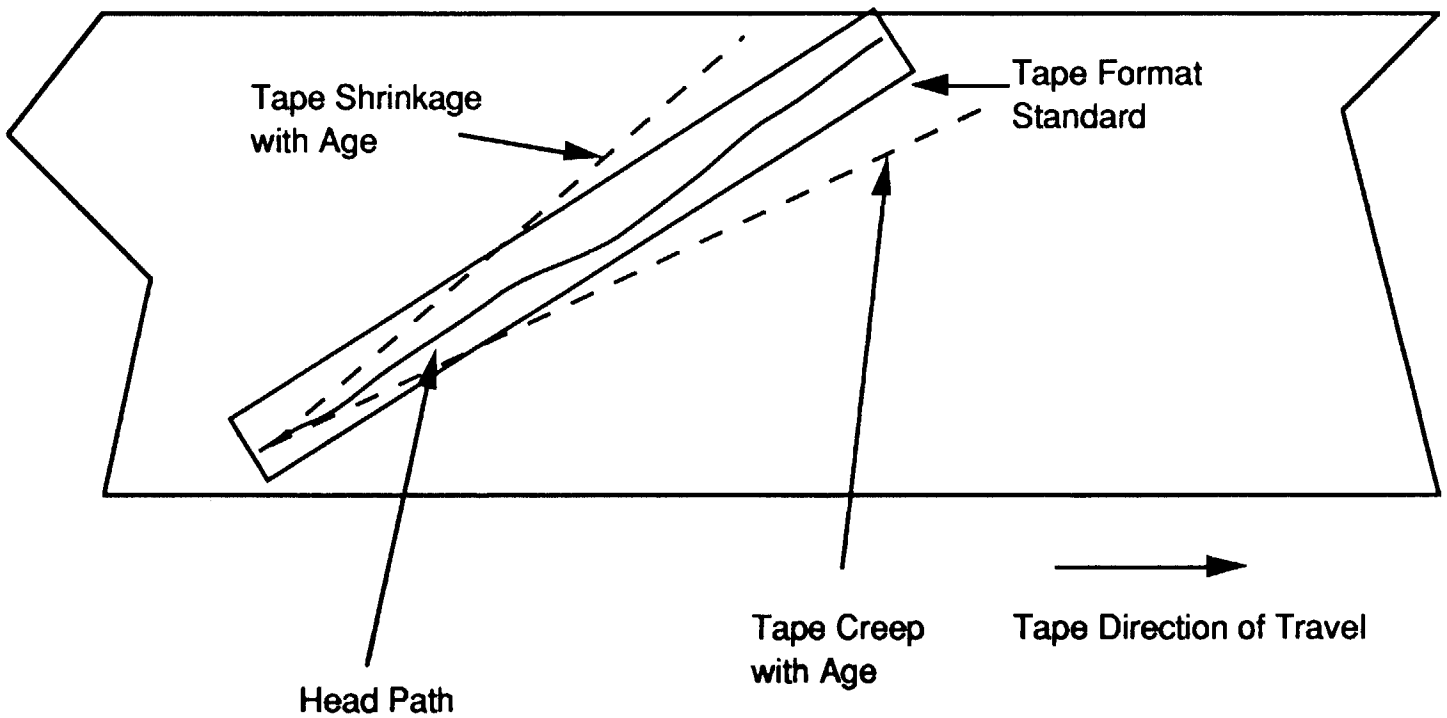
Backcoat:
Backcoat/Magcoat
Adhesion (blocking)

Wear particles
Head/tape interface
brown staining
head wear

Magnetic Pigments
Loss of Magnetic Moment
diffusion (Co- Fe2O3)
stress demagnetization

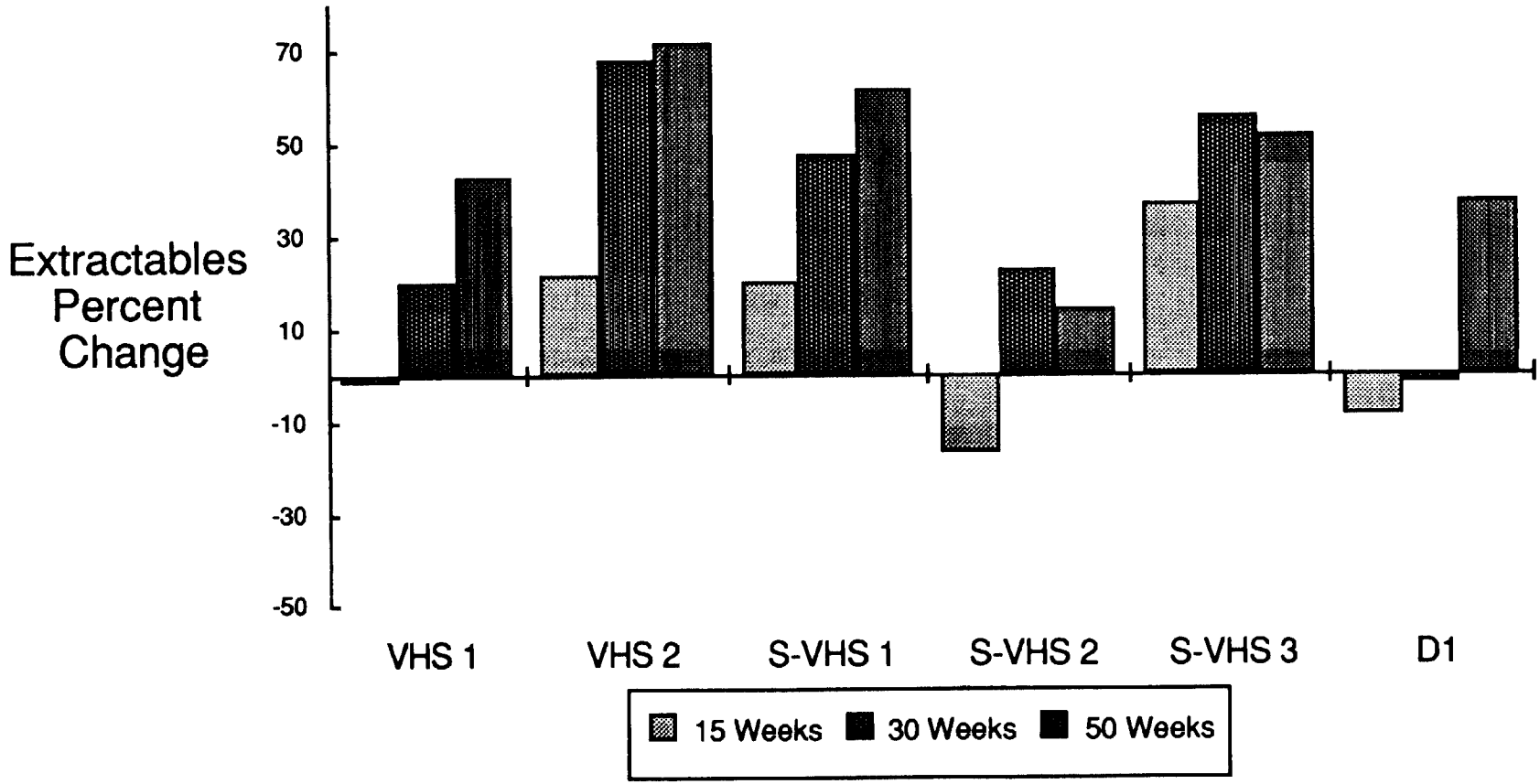
National Media Laboratory

Dimensional Stability and Crossplay



National Media Laboratory

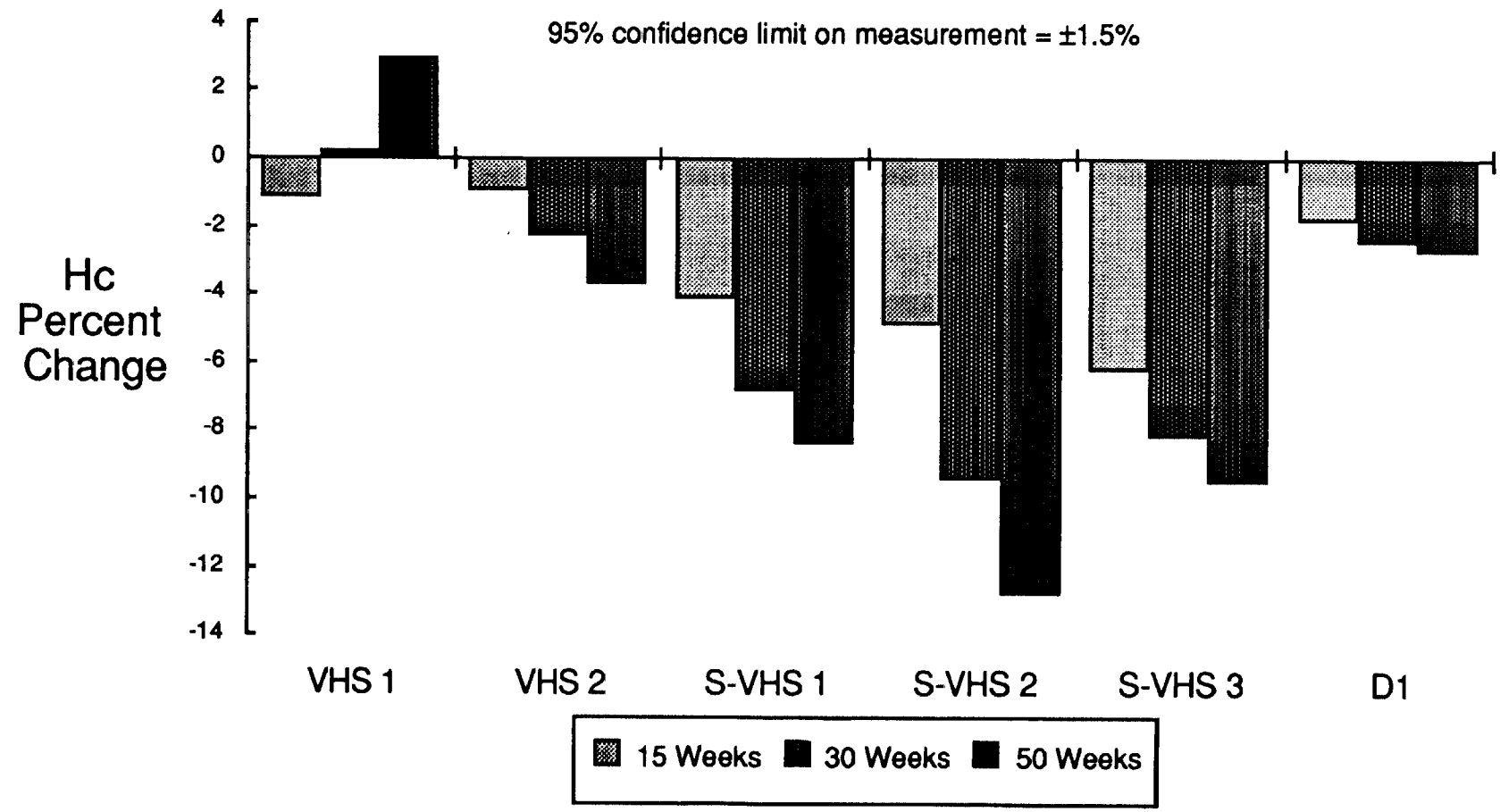
58°C - 90% RH
Acetone Extraction



NML DMC 7/24/91

National Media Laboratory

58°C - 90% RH Coercivity

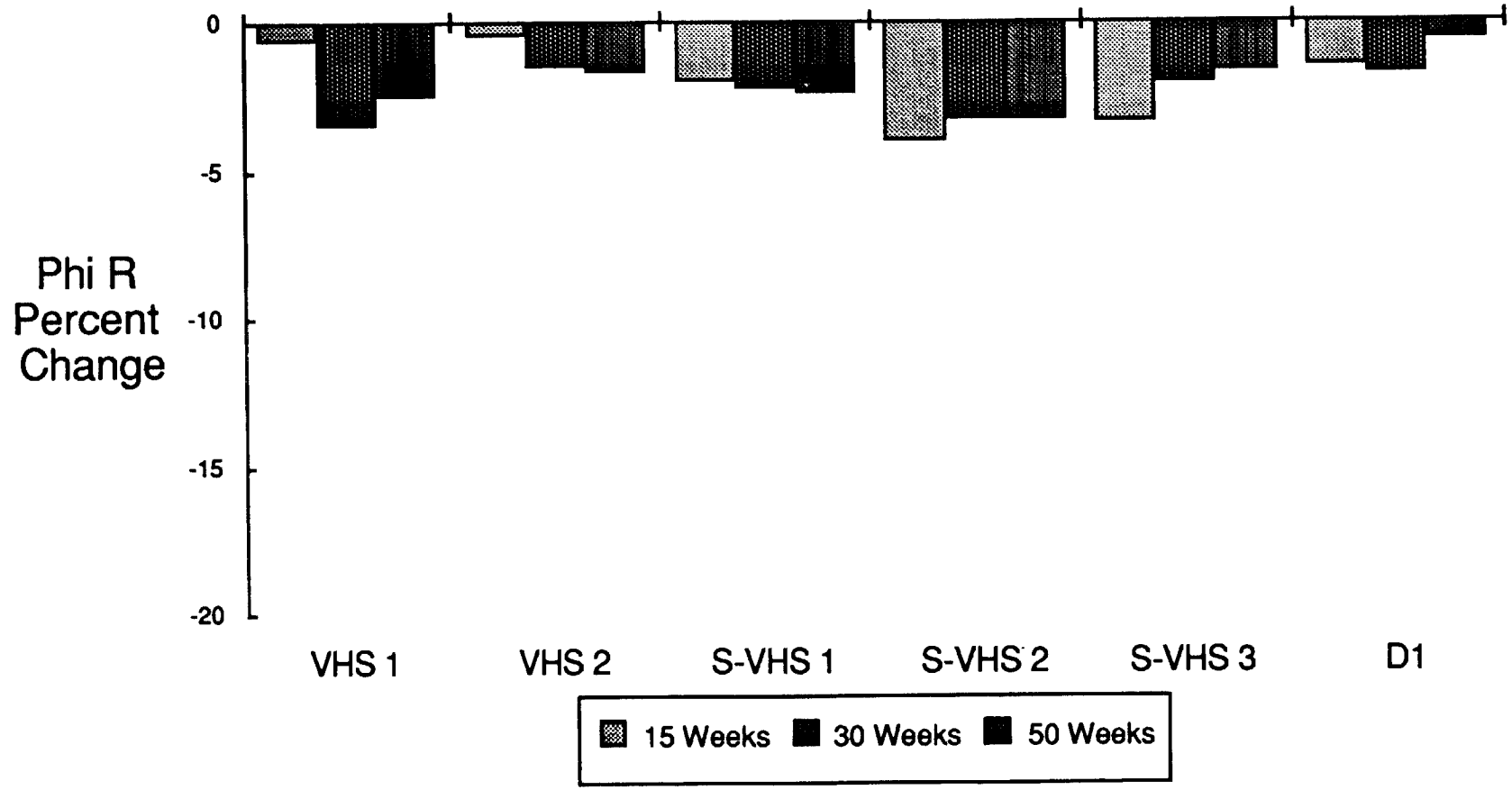


NML DMC 7/24/91

National Media Laboratory

58°C - 90% RH
Phi R

95% confidence limit on measurement = ±6%



213

NML DMC 7/24/91

Summary
Co-γ-Fe_2O_3 Tape Stability

National Media Laboratory

- Archive Stability: 10 yrs in a dry, controlled environment
  - Retensioning
  - Cleaning
  - Recertification
- Archive requirements are system specific
  - Wide range of system performance
  - User requirements outside recommended controlled environments
  - Distributed Storage Systems require Shipping
  - Anomalous events

S15-35
12/952
p. 5

19 mm DATA RECORDERS SIMILARITIES AND DIFFERENCES

A WHITE PAPER FROM AMPEX CORPORATION

by

Steve Atkinson

January 28, 1991

19 mm DATA RECORDING FORMATS

Confusion over the use of non-video 19 mm data recorders is becoming more pronounced as we enter the world of high performance computing. What is the difference between ID-1, ID-2, MIL-STD-2179 and DD-2? What is the proper machine for my application? How do I integrate it into my environment? These are all questions the user community needs answered. This paper attempts to address these issues and clear up any misconceptions there might be about 19 mm tape recorders.

HISTORY

The MIL-STD-2179 and ID-1 tape formats are modifications of the D-1 tape format standard used in the television and motion picture industry. The development of D1 based instrumentation standards was driven by the need of military and government users for high speed digital recorders to capture data. The Military Standard 2179 (MIL-STD-2179) was supported by a group of manufacturers and military representatives (primarily the United States Navy) to provide such a recorder. The first prototypes of these recorders are becoming available now for Beta testing in a number of military programs. Loral, Sony, Datatape, Honeywell, and Schlumberger are all working to deliver these machines. An ANSI standard known as ID-1 also exists which is being implemented by same manufacturers. It differs slightly from the 2179 implementation but is not interchangeable due to the use of azimuth recording, which is not used in the D1 TV format or the MIL-STD-2179 format.

Products to support the television D1 format became available in early 1987 and were followed by D2 format products in late 1988. The two TV digital formats differ in that D1 supports component recording while D2 supports composite recording. In laymen's terms, this means that D1 records video as individual data streams and D2 records it as a single stream. The Ampex ID-2 product is based upon the unmodified D2 NTSC video format, but is supplemented with user interfaces specific for instrumentation applications. As such, the Ampex ID-2 products can serve as either a video or instrumentation recorder. D1 recorders are produced for the video industry by Sony and BTS and D2 recorders are manufactured by Ampex with Sony and Hitachi. The D2 recorder is the dominant 19 mm machine in the video industry constituting over 75% of the digital market today and is expected to increase in market share to over 80% by the early 90's.

The need to support high speed data storage and retrieval for the supercomputer industry was identified by Ampex as a major market opportunity as data creation and storage grew exponentially in the 1980's. Hence, the Data D2 (DD-2) development was undertaken as a joint development by Ampex and E-Systems. The basic difference between the DD-2 recorder and instrumentation recorders is that DD-2 is designed as a computer peripheral with computer interfaces and file structures as opposed to an instrumentation recorder which records input as a single uninterrupted stream of data. Each class of machine has its strengths and weaknesses depending on the mission the user wants to accomplish.

COMMON FEATURES

DD-2, MIL-STD-2179, ID-2, and ID-1 physically record data in very similar manners. All implementations use a recording technique know as helical scan. Helical scan recorders lay down data at an angle on tape and provide improved storage density over longitudinal formats and also provide powerful error correction capabilities. Ampex was a pioneer in the helical scan technology used in today's home VCR's.

Each of these scans of recorded data are called tracks. They are grouped together to form track sets, the lowest addressable unit of data from a physical standpoint. The

helical scan data tracks are accompanied by three longitudinal tracks. In the video version, these tracks contain control, time code and audio information. The instrumentation and data recorders use these tracks for control, timecode, and for file labels in the case of DD-2.

All four 19 mm formats are implemented in high performance machines with several versions tailored for specific missions. Features commonly supported include high rate data storage and retrieval, robotic compatibility, and positioning to data at speeds of 30 to 60 times playback speed. For DD-2 at 60 times playback speed, this translates to over 800 megabytes per second (MB/s) or the equivalent of four 3480 cartridges every second.

With these similarities in mind, it is easier to discuss how the DD-2 computer storage peripheral and 19 mm instrumentation recorders use the 19 mm platform to accomplish different tasks with different derivations of the same basic technology.

INTERFACES

The applications for instrumentation recorders such as ID-1, and computer peripherals such as DD-2, drive the design of their interfaces. Instrumentation machines are commonly used for data capture which means that data will come in a continuous stream with the recorder turned on when the data starts and off when the data stops. This alleviates the need for any buffering on the recorder since the data should arrive in a continuous mode at a steady rate. The standard implementation is a 16-bit wide data front-end which operates in a synchronous mode. By implication, this means that the recorder operates as the master with the input or output source acting as the slave. This is because the recorder expects to record and retrieve data at a steady, uninterrupted stream. Typically, it handles different data rates by the ability to record/retrieve data at selectable, binary rates of up to approximately 30 MB/s in some implementations. Therefore, it is up to the source to buffer the data for smoothness within a small percentage of the recorder's data rate needs. While this is a good implementation for the field and downlink data recording it was designed for, it poses severe problems in a computer environment.

Most computer operating systems simultaneously execute multiple programs in an interactive, time sharing environment. Input and output to a peripheral is not supported as a steady, uninterrupted stream, but instead is dependent upon the dynamics of system loading, computational rates, and application mix. These operating environments do not include the dedicated data buffering and strict command scheduling required to match rates with a data peripheral. The logic of these operating systems is designed to use compute cycles to serve applications -- not peripherals.

Instrumentation recording on a computing platform requires a dedicated computer, executing only from a predetermined set of controlled applications. Development of instrumentation interfaces, controllers, and special device drivers are required to connect a given peripheral. Memory must be dedicated to support the strict input/output rates of the peripheral.

The DD-2 recorder was designed with these interface shortcomings in mind. The DD-2 recorder from Ampex uses the ANSI standard Intelligent Peripheral Interface Level 3 (IPI-3) interface to send data to and retrieve data from the recorder. This standard computer interface defines command, control, and status for both disk and tape peripherals. The subset of the IPI-3 command set dealing with tape devices is used for the DD-2. A large buffer is included as part of the interface electronics. This buffer performs the rate smoothing to provide a sustained data rate of 15 MB/s with a burst rate of 20 MB/s. Using the IPI-3 command set and the buffering capability allows the host to request data units as small as a single byte so that the host computer does not have to retrieve an entire physical block. The IPI-3 implementation also allows up to 8 recorders to be "daisy-chained" on a single control interface. While one recorder is transferring data, the others can be given positioning commands to minimize the

time spent waiting for the subsequent file transfers. Other standard interfaces will be supported in the future. In summary, the DD-2 looks like a standard tape peripheral to the host computer.

DATA FORMAT

As stated earlier, both DD-2 and 19 mm instrumentation recorders physically record data on tape in very much the same way, helical scans grouped as track sets. Once again, the differences lie in the fact that the DD-2 is a computer peripheral recorder. Both machines can retrieve data by track sets if that is the search parameter they are given. The DD-2 also makes use of the longitudinal tracks to implement ANSI 9-track file labeling. Files on the DD-2 can be retrieved by using file labels rather than track set ID's. This allows software already developed on the host computer to manipulate 9-track tape and 3480 cartridges to also make use of DD-2 cassettes. Minimizing the impact of incorporation of DD-2 into existing computer is one of the highest priorities of the DD-2 development.

The traditional concept of tape volumes is incorporated into the DD-2 design while not in 19 mm instrumentation formats. Each DD-2 physical cassette looks like one or more logical volumes to the host computer. These logical volumes can be edited and appended to look like traditional 9-track volumes. Another somewhat related innovation is that the cassettes can be loaded and unloaded without rewinding to beginning of tape or end of tape. This greatly decreases the amount of search time spent positioning to the beginning of a file.

Inter-record gaps are another format difference between 19 mm instrumentation recorders and DD-2. Because most 19 mm instrumentation machines record data at binary rates, the gaps between recorded data are of variable size. This leads to less efficient use of tape as it is indeterminate how much tape is left unrecorded in between data records. The DD-2 uses a consistent gap to further improve upon the substantial data density advantage DD-2 holds over the ID-1/MIL-STD-2179 class of recorder.

ERROR RATE

Image capture and retrieval is one of the most common uses of ID-1 and MIL-STD-2179 recorders. When manipulating images on a host computer, errors on tape can be recovered from by reconstructing the lost part of an image from the remaining parts which surround it. For that reason, the Bit Error Rate (BER) of 10^{-10} specified for MIL-STD-2179 and ANSI ID-1 recorders is sufficient for most of the applications in which it is used but may require tape certification to achieve. This BER is not always adequate for computer applications. DD-2 will deliver an error rate approximately three orders of magnitude better by guaranteeing less than one error event in every 10^{-12} bytes read.

All commercial 19 mm recorders have the capability to read what has been written on tape immediately after it is written. However, the implementation of a data buffer in the DD-2 interface makes verification of the data written on tape possible. When an uncorrectable error is detected by the read-after-write verification function, the data will be rewritten to tape a selectable number of times by the DD-2 controller. When retrieved, invalid data is ignored with only the corrected data returned to the host. This capability is coupled with three levels of powerful Reed-Solomon coding to achieve the improved error rate over 19 mm instrumentation formats. For example, the extended burst correcting capability of DD-2 provides a 50-times improvement over ID-1, minimizing errors due to tape defects.

In addition, the DD-2 will monitor the stress level of the error detection and correction (EDAC) equipment and store the statistics in the recorder. When requested from the host computer, it will be returned as status. In this manner, the host can monitor the "health" of an individual cassette and record the data contained on it to a new cassette when appropriate.

MEDIA

The media used by data storage peripherals and instrumentation recorders is the same as used in the video industry. The cassettes used by all 19 mm recorders contain tape which is 19 millimeters wide (about 3/4 inch). The main difference is that DD-2 and ID-2 use a metal oxide based tape with a total thickness of 13 micrometers, while the ID-1 and MIL-STD-2179 use the older iron oxide formulation with a total thickness of 16 micrometers. 19 mm cassettes are very similar in appearance to home VHS cassettes except that VHS uses 1/2 inch tape. D2 tape is manufactured by Ampex, Sony, Fuji, and Maxell with 3M expected soon to enter the market. D1 tape is available from Ampex, Fuji, and Sony. The important point is that both are supported by more than one source which should result in both a reliable future supply and low user costs.

The storage density of D1 and D2 based recorders is a function of the recorder and tape used. A storage density comparison is contained in this section for simplicity's sake. Both DD-2 and ID-1/MIL-STD-2179 have three sizes of cassettes: small, medium and large. The small D2 cassette contains 25 gigabytes (GB) of user data with a medium containing 75 GB and a large holding 165 GB. This compares to 14 GB, 44 GB, 92 GB for the small, medium, and large D1 cassettes respectively.

A concern has been raised on the shelf life of metal oxide tape due to the high content of iron particles in the tape. Metal particle tape has been in existence over 15 years and has been used in commercial products for over 10 years. All 8 mm Camcorders, as well as other video and data storage products, are based on the metal oxide tape. Ampex has recently completed accelerated life tests which simulate a 14-year archive. No degradation in BER or magnetics was detected with the tape in the cassette. Details were presented by Ampex at the Tape-Head Interface Conference (THIC) proceedings on 9 January 1991. Copies of the presentation are available from Ampex. Tests by other manufacturers also support these results.

SUMMARY

DD-2 and 19 mm instrumentation recorders have missions for which each is well designed. While the differences may appear subtle, understanding the difference between the two is the key to picking the right recorder for your particular application.

AN EMPIRICAL APPROACH TO PREDICTING LONG TERM BEHAVIOR OF METAL PARTICLE BASED RECORDING MEDIA

By

Allan S. Hadad
Ampex Recording Media Corporation
Redwood City, California

Narrative
Submitted for the

National Space Science Data Center's
Conference on
Mass Storage Systems and Technologies for Space and Earth
Science Applications

Goddard Space Flight Center
Greenbelt, Maryland
July 23-25, 1991

AN EMPIRICAL APPROACH TO PREDICTING LONG TERM BEHAVIOR OF METAL PARTICLE BASED RECORDING MEDIA

Alpha iron particles used for magnetic recording are prepared through a series of dehydration and reduction steps of α-Fe_2O_3-H_2O resulting in acicular, polycrystalline, body centered cubic (bcc) α-Fe particles that are single magnetic domains. Since fine iron particles are pyrophoric by nature, stabilization processes had to be developed in order for iron particles to be considered as a viable recording medium for long term archival (i.e. 25+ years) information storage. The primary means of establishing stability is through passivation or controlled oxidation of the iron particle's surface.

The usual technique of producing the protective layer is through re-oxidation of the iron particle's surface after synthesis starting with a mixture of 0.1% O_2 and 99.9% N_2. The oxygen content is slowly increased to 20% (the composition of air) so as to maintain the reaction at room temperature. This results in a particle that is stable in air provided it is not subjected to any form of 1) mechanical abuse that could disturb the outer layer, or 2) source of heat that would initiate combustion.

The nature of the passive layer on iron particles has been found to consist of either Fe_3O_4, γ-Fe_2O_3 or a mixture of the two. The thickness of the passivation (or total) oxide layer is typically about 3.0 nm. The condition of passivity slows down the oxidation of iron, but, the formation of the iron-oxide layer around metallic iron particles does not preclude further oxidation. The continued stability of iron particles is controlled by the integrity of the oxide layer and the kinetics of diffusion of iron ions through it.

Since iron particles used for magnetic recording are small, additional oxidation has a direct impact on performance especially where archival storage of recorded information for long periods of time is important. Further stabilization chemistry/processes had to be developed to guarantee that iron particles could be considered as a viable long term recording medium.

In an effort to retard the diffusion of iron ions through the oxide layer, other elements such as silicon, aluminum and chromium have been added to the base iron to promote more dense scale formation or to alleviate some of the non-stoichiometric behavior of the oxide or both.

The presence of water vapor has been shown to disrupt the passive layer, subsequently increasing the oxidation rate of the iron.

A study was undertaken to examine the degradation in magnetic properties as a function of both temperature and humidity on silicon-containing iron particles between 50-120°C and 3-89% relative humidity. The methodology to which experimental data was collected and analyzed leading to predictive capability is discussed.

EXPERIMENT

To study the effect of temperature and humidity on the stability of passivated iron particles as a function of silicon content, three particle batches were prepared from the same precursor containing 0.81, 0.95 and 1.12% silicon by weight.

As stated earlier, the oxidation process of iron particles is diffusion controlled. The increase in oxide thickness is determined by how fast iron ions can migrate to the surface (passivation kinetics). It is assumed that the reduction in σs is proportional to the oxide thickness; therefore, Fick's second law of diffusion for concentration dependent systems can be applied:

$$Y^2 \simeq D\ t$$

where:

Y = diffusion distance in cm

D = diffusivity in cm^2/sec

t = diffusion time in sec.

For this analysis, the data are plotted as specific magnetic moment versus the square root of time. The resultant slope is therefore the rate of degradation expressed as EMU/gram-$sec^{0.5}$ for a given temperature and humidity condition.

Analysis of the interactive effects of temperature, humidity and percent silicon on the degradation of magnetic properties of iron particles were determined using a statistical experimental design package run on an IBM AT compatible computer. The characteristic measured was the $\log_{10}$ of σ_s loss per $second^{0.5}$ (rate of degradation).

The procedure of data analysis consists of selecting an initial mathematical model (usually a quadratic equation to analyze interaction and curvature effects of the form $Y = C_o + C_1X_1 + C_2X_2 + C_{13}X_1X_3 \ldots. + C_{11}X_{12} + C_{22}X_{22} \ldots.)$ where C_n is a constant and X_n is the level setting of the respective factor. The model is then used to "fit" the data.

RESULTS AND DISCUSSION

Dry Conditions. The first series of oxidation rate experiments were run at the "dry" conditions for the selected temperatures. Oxidation at various temperatures allowed generation of σs vs. time curves for iron particles containing 0.81, 0.95 and 1.21% silicon sample at various temperatures.

Figure 1 shows the change in specific magnetic moment with time for the 1.12% silicon sample at various temperatures. As temperature increased, the rate of degradation of σs also increased. Differences in the rate of degradation at a given temperature between the three samples studied containing different amounts of silicon are slight, indicating that the iron particles all behave the same at low humidity, independent of Si content or initial σ_s value.

Activation energies for degradation of the 0.81, 0.95 and 1.12% silicon-in-iron powder samples were determined to be 0.383, 0.263, 0.333 eV, respectively when the rates of degradation were plotted against the absolute temperature (Arrhenius plot). Thus, the energy requirement to initiate the oxidation reaction for the various Si-contents is essentially the same and all degrade at essentially the same rate as a function of temperature. However, these activation energies are only ~10% of the value for diffusion of iron through any of the possible iron oxides. The implication is that the primary mechanism for the loss in magnetic properties is not due to lattice diffusion but rather diffusion along some short circuit path such as linear or planer defects present in the oxide shell.

Moisture Effects. Magnetic moment loss is shown as a function of humidity at constant temperature (70°C) for the 0.81, 0.95 and 1.12% silicon samples in Figures 2, 3 and 4 respectively. As the humidity increases, the rate at which the magnetic moment degrades increases markedly with decreasing Si content.

Degradation is most rapid for the 0.81% Si-containing iron particles. The sample containing 0.95% silicon is more resistant to degradation until the relative humidity is above 59% at 70°C. Finally, the sample containing 1.12% silicon is the most resistant to degradation exhibiting almost the same rate of loss at 7% relative humidity as it does at 74% relative humidity.

There appears to be a threshold value of relative humidity that is observed at the 0.95% silicon level and possible the 1.12% level, above which the degradation rate increases very rapidly.

As silicon content increases, the critical humidity (the humidity where rapid degradation occurs, at a given temperature) also increases. However, the degradation ceases at σ_s value of 50-60 EMU/gr for the 0.81% silicon sample (Figure 2). The trend appears to be the same (a saturation limit) for the 0.95 and 1.12% Si particles also.

Interaction of Temperature, Humidity and Silicon. The three parameters of this study (temperature, percent relative humidity and percent silicon content) interactively influence the degradation of the iron particles. Thus, statistical analysis was performed to quantify the contribution of each variable and present it in a manner that could be easily visualized. The measured characteristic was the $\log_{10}$ of the degradation rate in specific magnetic moment (EMU/gram-second$^{0.5}$) as derived from the results of the experimental matrix.

The design predictor equation determined using XSTAT is as follows:

$$\log_{10}(\text{Rate}) = -14.76 + 25.26(X_{Si}) + 0.04176(X_{RH}) + 0.0015(X_T * X_{RH}) - 0.05963(X_{Si} * X_{RH}) + 0.000068(X_T)^2 - 12.49(X_{Si})^2 + 0.000219(X_{RH})^2$$

where:

X_i = concentration of silicon in weight percent

X_T = temperature in degrees Celsius

X_{RH} = percent relative humidity

The percent variance explained (how well the regression equation predicts the data) was 96.03% so the model can be considered as quite good. Solving the predictor equation via the computer produces two dimensional contour plots of percent relative humidity versus temperature at constant silicon content. The values of constant degradation rate were transferred to a psychrometric chart (Figures 5, 6 and 7) so the relationships between absolute humidity (expressed as pounds of water per pound of dry air), percent relative humidity and temperature as a function of silicon content could be observed. The non-linear degradation behavior becomes very obvious when presented in this manner. A rate value can be taken from the contour plots for a given temperature, humidity and percent silicon content and be used to determine how long it would take for the magnetic moment to degrade to some predetermined value.

For a constant absolute humidity (above the critical value, depending on silicon content) increasing temperature will cause the degradation rate to decrease. It appears that relative humidity is the controlling factor. The corrosion rate appears proportional to the thickness of the adsorbed water layer on the surface of the test specimens. The higher the relative humidity, the thicker the adsorbed layer, hence the faster the corrosion rate.

The behavior observed in the contour plots follows an expression of the form Rate = A_e $(b\%RH)_e$-(Q/RT), where at low humidity temperature is the controlling factor, then a transition occurs where humidity effects prevail. It is obvious that special precautions should be taken when attempting to predict long term, low temperature behavior based on high temperature data with a system that is humidity sensitive. Actual archival stability could be less than expected unless humidity control is considered. The value of statistical analysis to the interpretation of oxidation behavior is clearly evident.

CONCLUSIONS

This study has shown the effects of percent silicon content, temperature and percent relative humidity on pre-passivated fine iron particles used for magnetic recording.

When the iron particles were exposed to various temperatures between 50 and 120°C at very low humidity it was observed that the degradation rates were not affected by either the silicon content or the initial value of the magnetic moment. TEM micrographs revealed that the oxide layer grew thicker leading to a condition of passivity.

It was shown that for a given temperature there exists a critical relative humidity value above which the degradation rate of magnetic moment increases markedly. The presence of silicon appears to increase the critical humidity value at which rapid degradation occurs. When the magnetic moment degrades to 50-60 EMU/gram it remains constant for additional exposure time.

A parametric expression has been proposed to relate silicon content, temperature and humidity to the initial rate of specific magnetic moment degradation for the iron particles used in this study.

Figure 1.

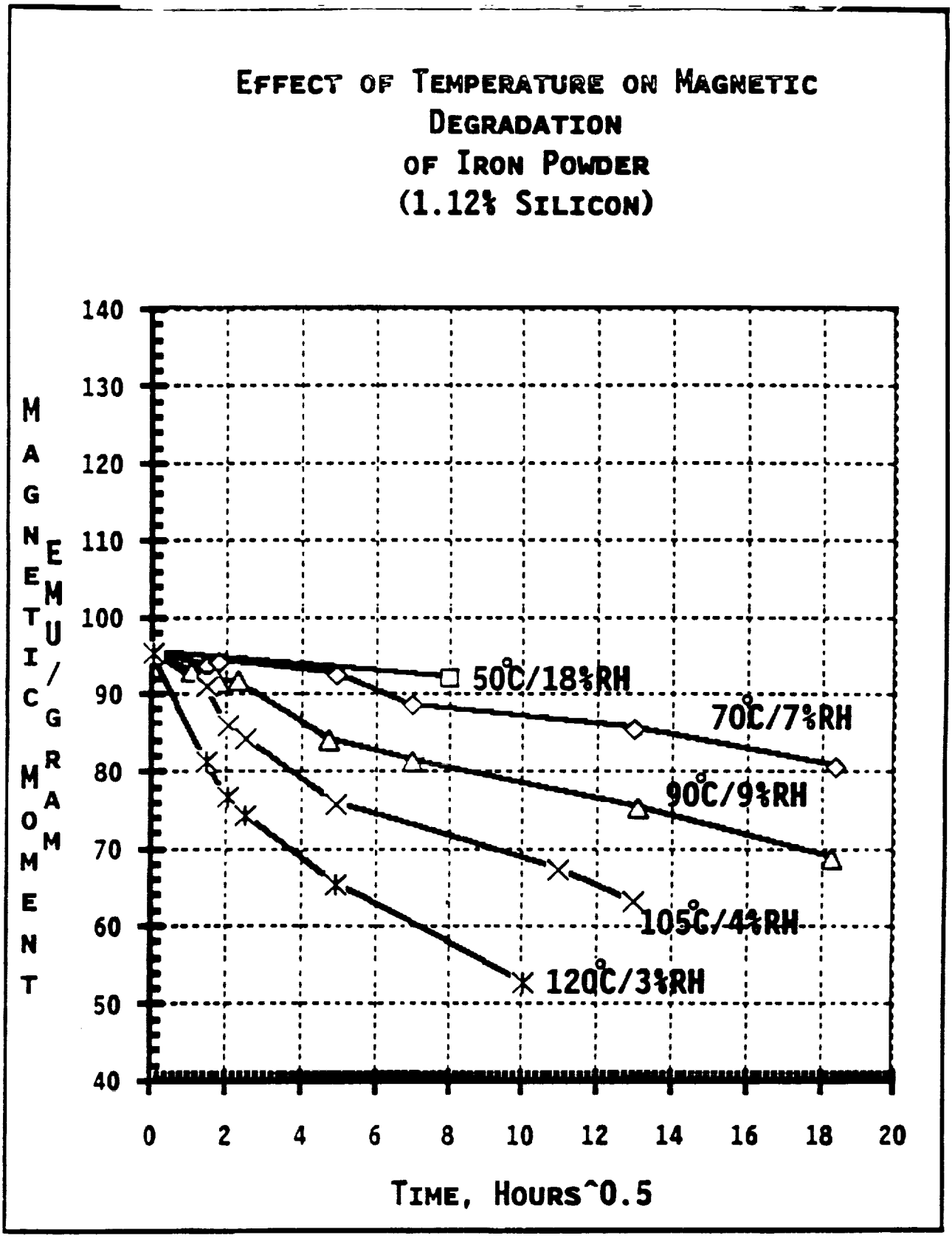

EFFECT OF TEMPERATURE ON MAGNETIC DEGRADATION OF IRON POWDER (1.12% SILICON)
MAGNETIC MOMENT EMU/GRAM
140
130
120
110
100
90
80
70
60
50
40
0
2
4
6
8
10
12
14
16
18
20
TIME, HOURS^0.5
50°C/18%RH
70°C/7%RH
90°C/9%RH
105°C/4%RH
120°C/3%RH

Figure 2.

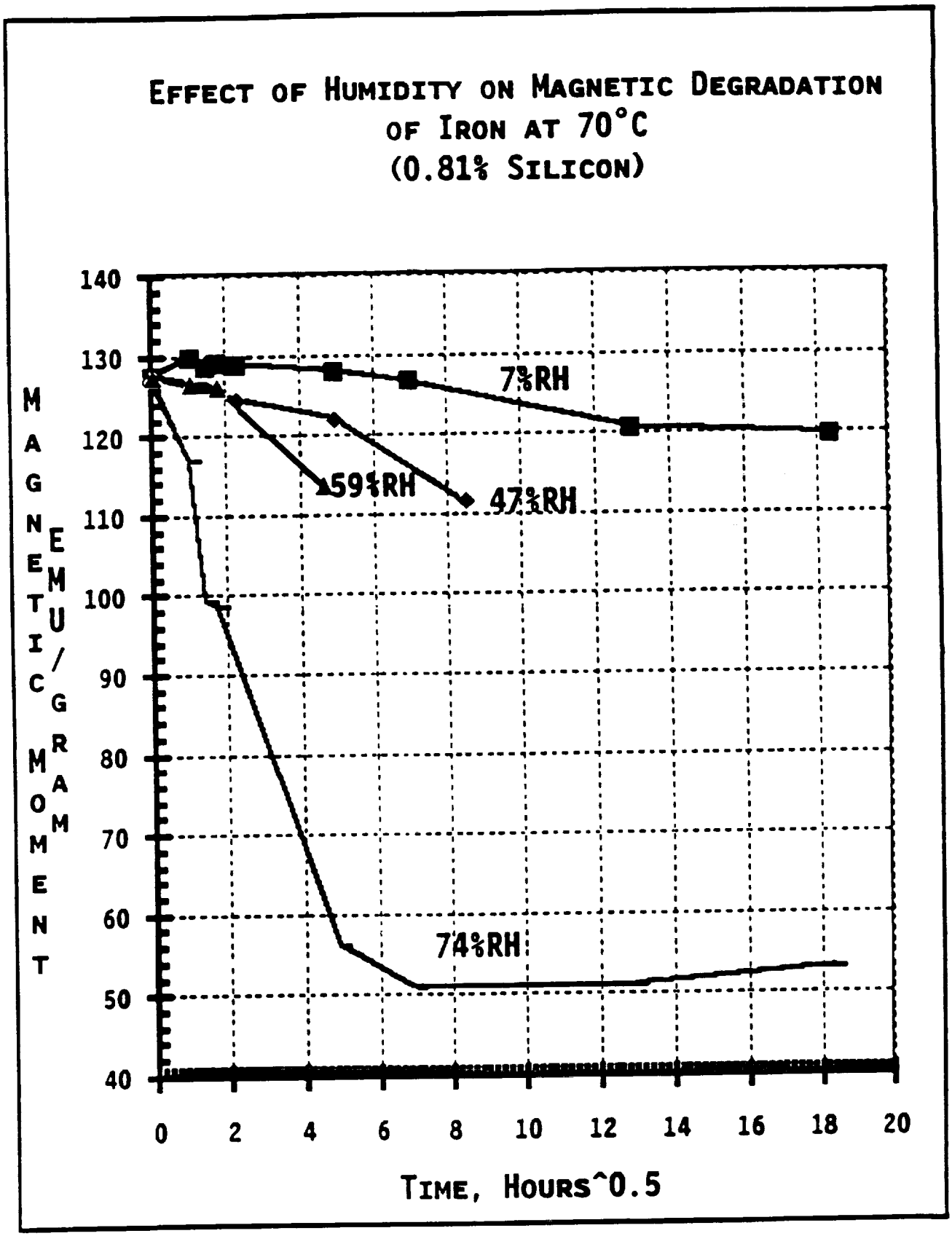

EFFECT OF HUMIDITY ON MAGNETIC DEGRADATION
OF IRON AT 70°C
(0.81% SILICON)
MAGNETIC MOMENT EMU/GRAM
140
130
120
110
100
90
80
70
60
50
40
7%RH
59%RH
47%RH
74%RH
0
2
4
6
8
10
12
14
16
18
20
TIME, HOURS^0.5

Figure 3.

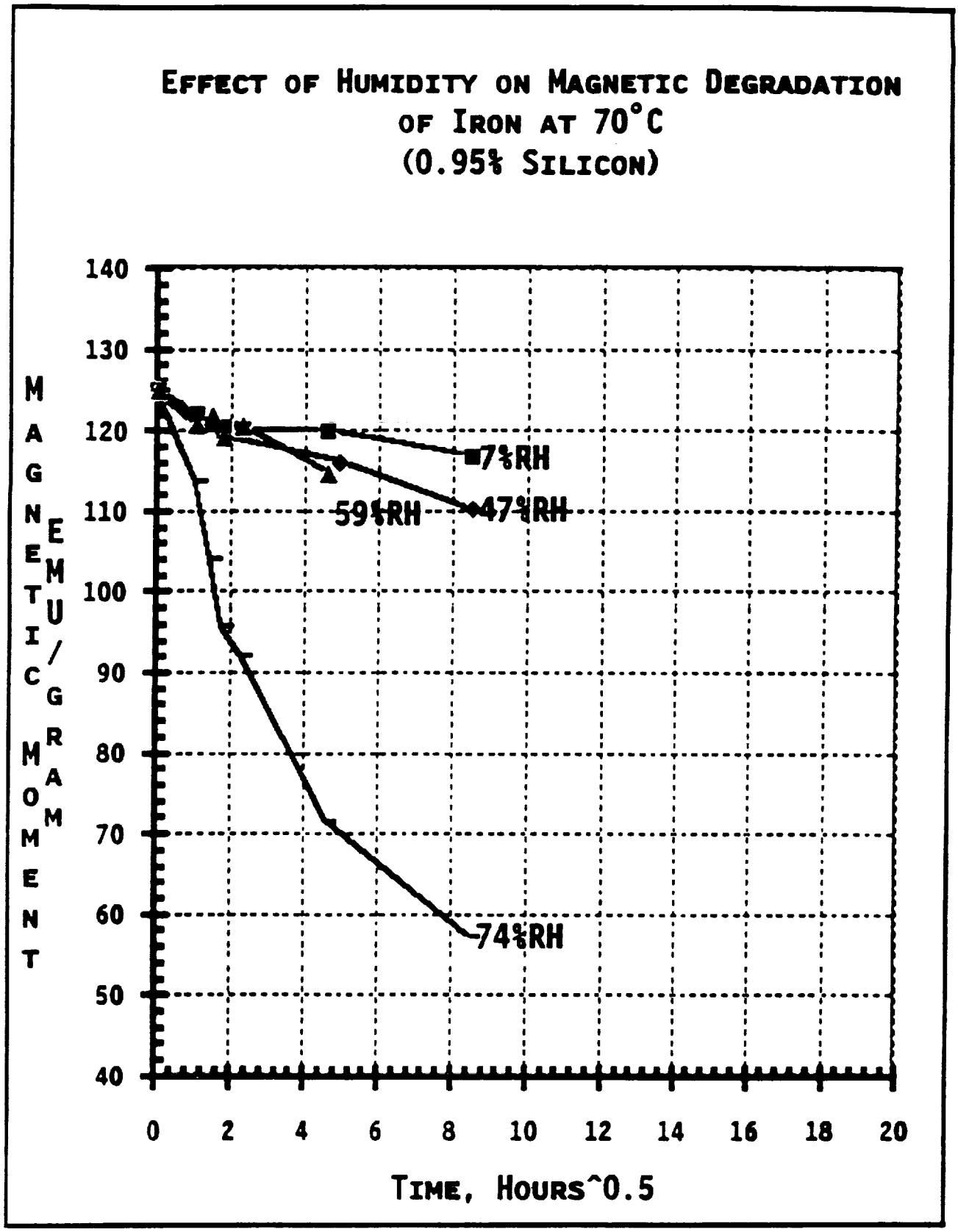

EFFECT OF HUMIDITY ON MAGNETIC DEGRADATION
OF IRON AT 70°C
(0.95% SILICON)
MAGNETIC MOMENT EMU/GRAM
140
130
120
110
100
90
80
70
60
50
40
0
2
4
6
8
10
12
14
16
18
20
TIME, HOURS^0.5
7%RH
59%RH
47%RH
74%RH

Figure 4.

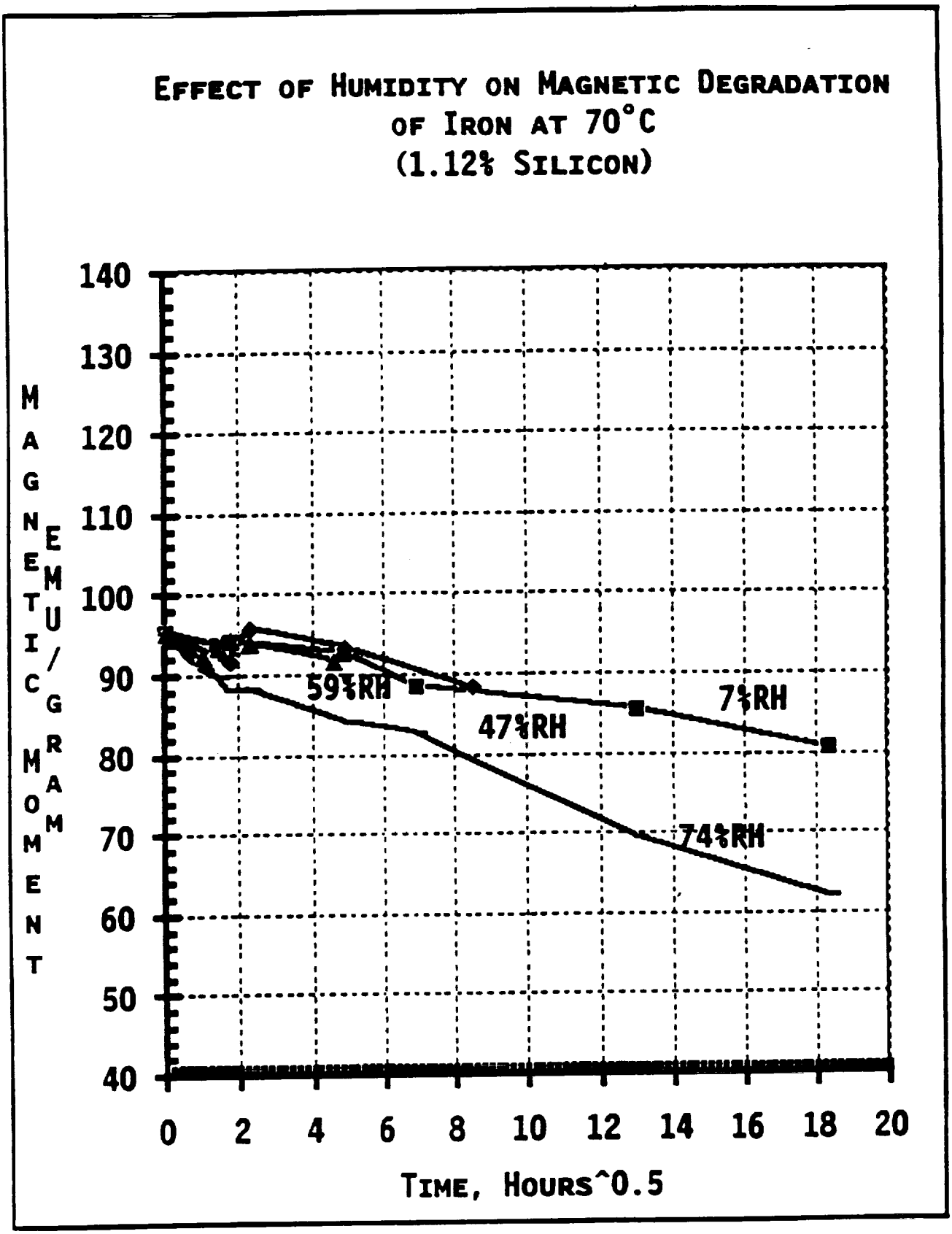

EFFECT OF HUMIDITY ON MAGNETIC DEGRADATION OF IRON AT 70°C (1.12% SILICON)
MAGNETIC MOMENT EMU/GRAM
140
130
120
110
100
90
80
70
60
50
40
0
2
4
6
8
10
12
14
16
18
20
TIME, HOURS^0.5
59%RH
47%RH
7%RH
74%RH

Figure 5.

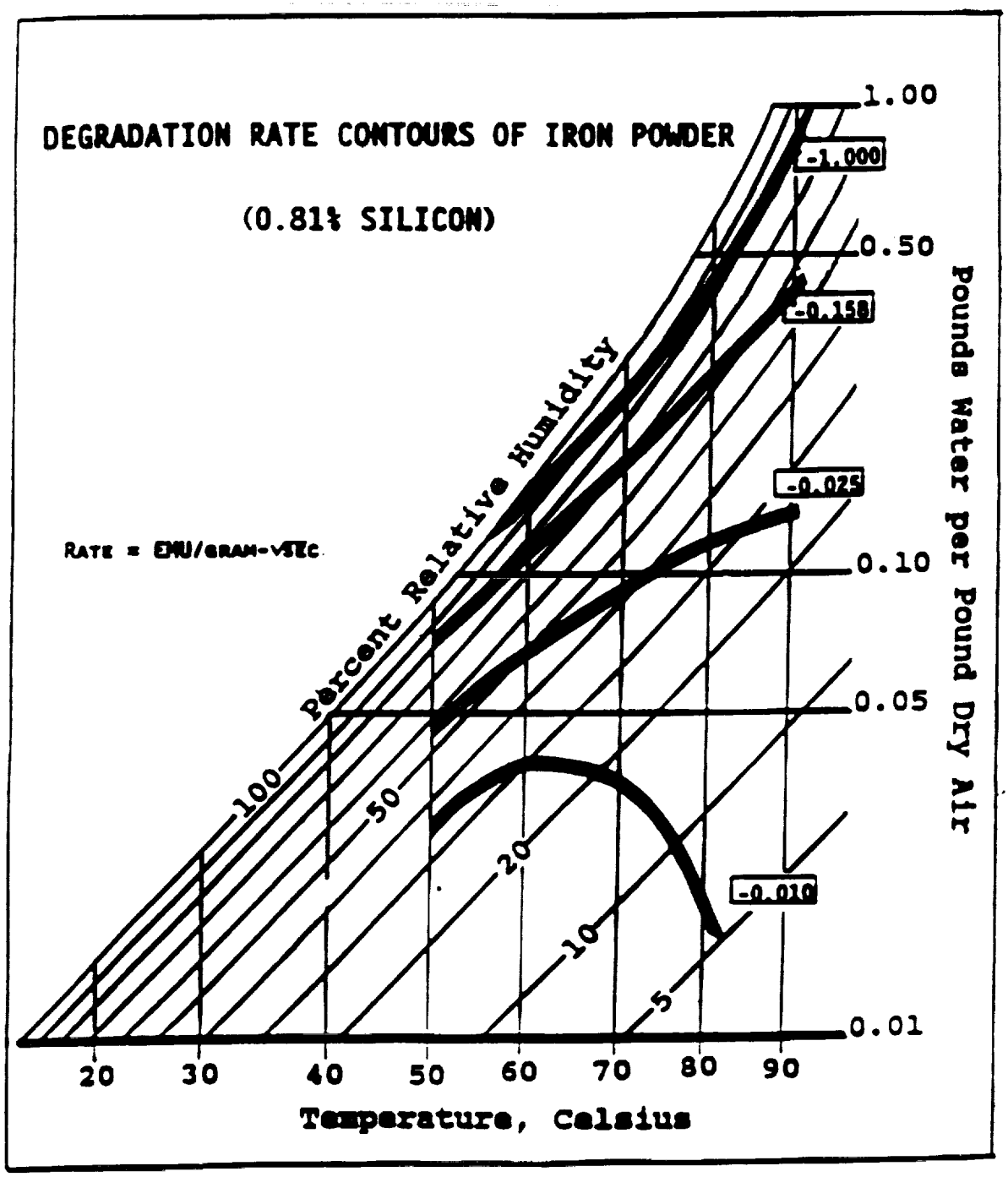

DEGRADATION RATE CONTOURS OF IRON POWDER
(0.81% SILICON)
RATE = EMU/GRAM-√SEC
Percent Relative Humidity
100
50
20
10
5
1.00
0.50
0.10
0.05
0.01
-1.000
-0.158
-0.025
-0.010
20
30
40
50
60
70
80
90
Temperature, Celsius
Pounds Water per Pound Dry Air

Figure 6.

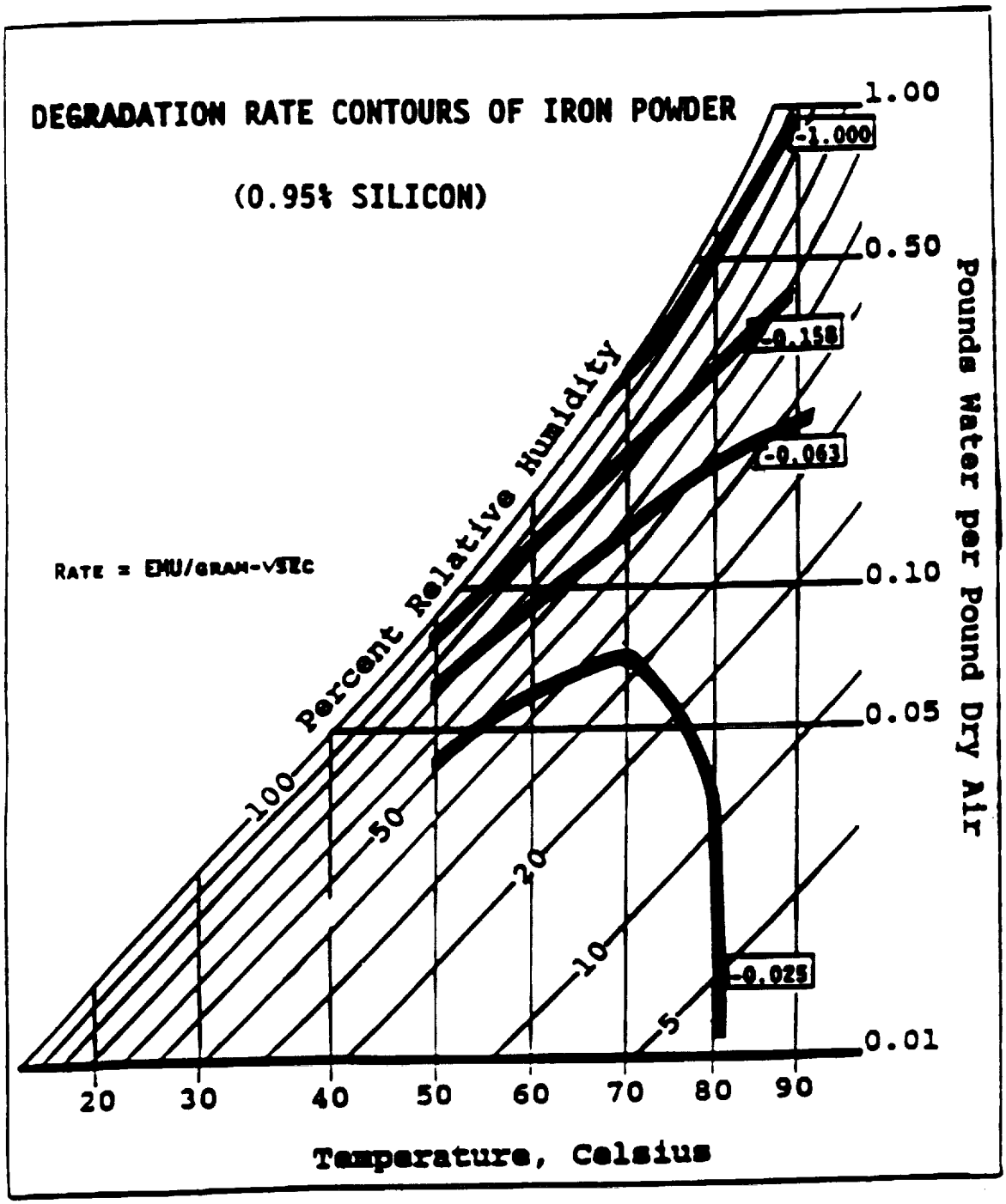
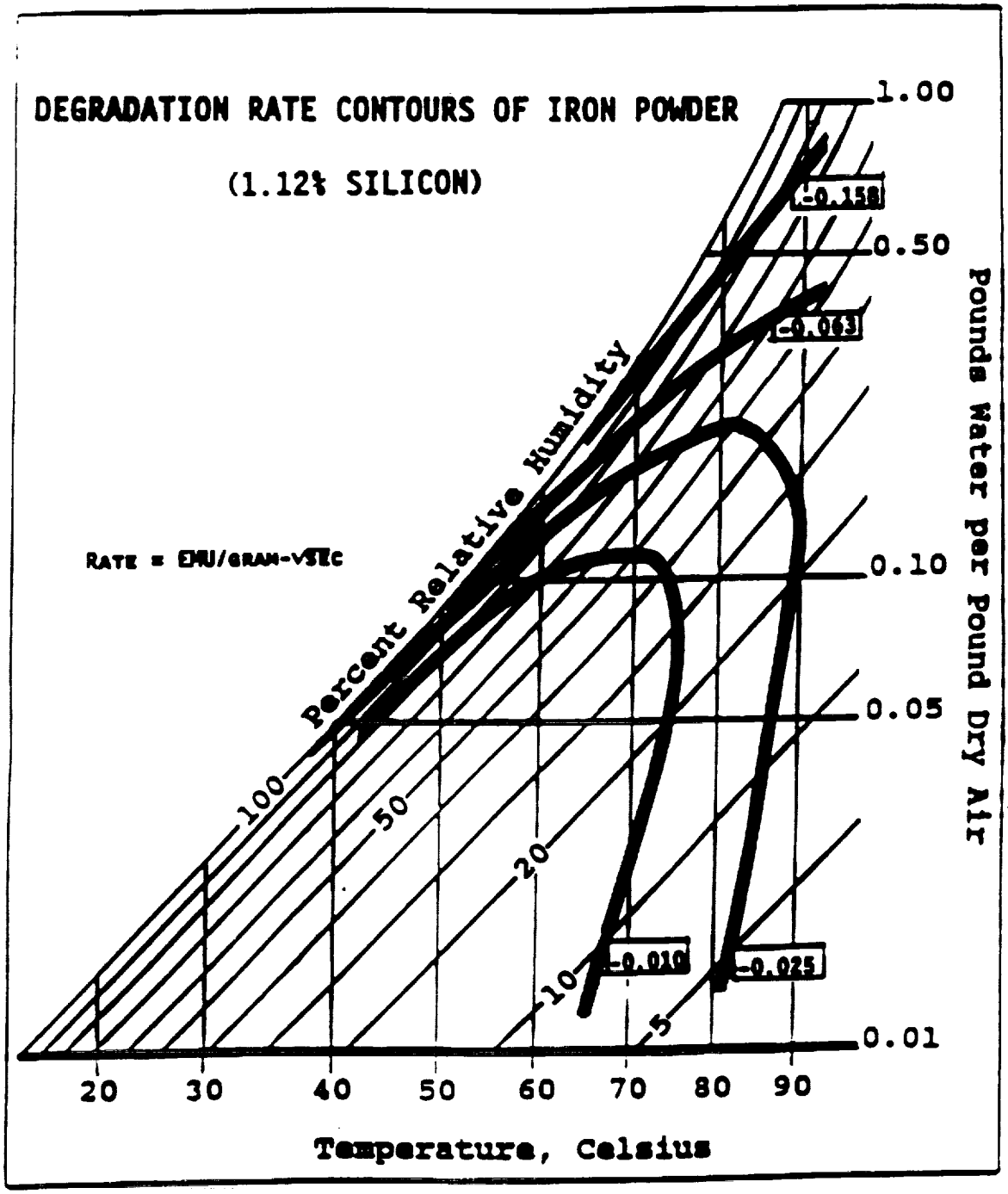

DEGRADATION RATE CONTOURS OF IRON POWDER
(0.95% SILICON)
RATE = EMU/GRAM-SEC
Percent Relative Humidity
100
50
20
10
5
1.000
0.158
0.063
0.025
1.00
0.50
0.10
0.05
0.01
Pounds Water per Pound Dry Air
20
30
40
50
60
70
80
90
Temperature, Celsius

Figure 7.


DEGRADATION RATE CONTOURS OF IRON POWDER
(1.12% SILICON)
RATE = EMU/GRAM-WEEK
Percent Relative Humidity
100
50
20
10
5
1.00
0.50
0.10
0.05
0.01
0.158
0.063
0.010
0.025
Pounds Water per Pound Dry Air
20
30
40
50
60
70
80
90
Temperature, Celsius

The Role of HiPPI Switches in Mass Storage Systems: A Five Year Prospective

T. A. Gilbert

Network Systems Corporation
Vienna, Virginia

Introduction

New standards are evolving which provide the foundation for novel multi-gigabit per second data communication structures. The lowest layer protocols are so generalized that they encourage a wide range of application. Specifically, the ANSI High Performance Parallel Interface (HiPPI) is being applied to computer peripheral attachment as well as general data communication networks.

This paper introduces the HiPPI standards suite and technology products which incorporate the standards. The use of simple HiPPI crosspoint switches to build potentially complex extended "fabrics" is discussed in detail. Several near term applications of the HiPPI technology are briefly described with additional attention to storage systems. Finally, some related standards are mentioned which may further expand the concepts above.

The High Performance Parallel Interface

History

The HiPPI standard evolved from efforts begun and still lead by individuals at The Los Alamos National Laboratory. Originally known as HSC or "High Speed Channel", HiPPI was derived from the Cray Research HSX supercomputer channel.

The original framers of what has become the HiPPI standard had several objectives in mind which in retrospect have been crucial to the rapid acceptance of this standard by many users and vendors:

- An interface capable of data transfer in the gigabit per second range. HiPPI is defined for 800 Mbps and 1.6 Gbps rates.

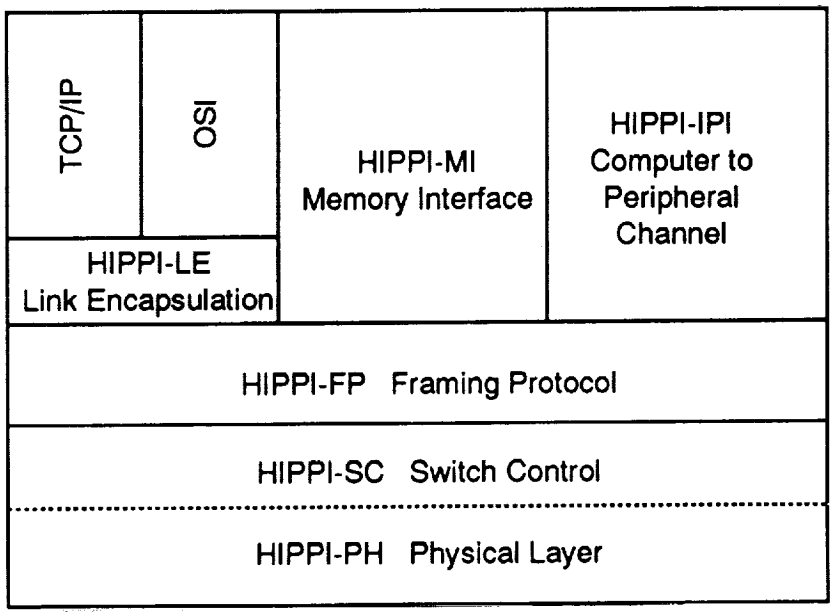
- standard interface which could be implemented by a broad range of vendors without the need for exotic or expensive technology. HiPPI physical layer interfaces can be built from off the shelf components which have been available for two decades.

- standard which is stratified such that the most fundamental common layers impose the least possible restriction on the nature of the digital datastream. HiPPI is being proposed for use in traditional networks, for the attachment of peripherals to host channels, for digital HDTV, and for connecting isochronous streams of imagery and digitized voice.

HiPPI standards efforts are under the auspices of ANSI X3T9.3 which this year will finalize most if not all of the constituent standards relevant to the directions discused in this paper. Related or follow on efforts are discused below.

The ANSI HiPPI Standards Suite

The X3T9.3 committee has defined six HiPPI component standards. Three are common to all others and comprise what may be likened to the media access layer in the ISO Open Systems Interconnection protocol model. However, this analogy implies that HiPPI is but another data link component in the traditional data communications hierarchy. It can serve that role but this understates its generality of application as discussed below.



The six standards are:

HIPPI-PH - The physical layer definition which includes mechanical and electrical interface definitions. It also specifies the signaling rates of 800 and 1600 Mbps. Important HIPPI-PH characteristics are:

- 800 or 1600 Mbps isochronous interface
- parallel 32 or 64 bit wide data line interface
- 25 meter maximum cable length
- simplex interface
- parity and LRC data protection
- ready resume flow control

HIPPI-SC - An optional extension of the physical layer standard which defines a switch control interface. HiPPI connections may be switched to achieve multi-point connectivity. Multiple addressing modes are defined.

HIPPI-FP - Defines a common framing protocol for all other standards.

HIPPI-LE - The link encapsulation definition designed to support traditional data communication protocols such as TCP/IP and OSI. LE essentially creates an IEEE 802.2 LLC compatibility layer on top of HIPPI-FP.

HIPPI-IPI - This is really more of a place holder to designate the use of ANSI IPI2 or IPI3 channel protocols over a HiPPI connection.

HIPPI-MI - Is a memory interface definition which provides for a communication controller to mediate memory to memory data transfers. MI attempts to avoid the overhead in traditional protocols and create mechanisms useful for cooperative processing.

HiPPI Technology

A handful of equipment vendors have actually shipped HiPPI compliant products to date. However, many more have announced intentions to do so over the next year. Products are available as of the first half of 1991 to begin implementing several of the advanced applications mentioned later. Examples of existing products are described in this section.

Computer Channels

IBM was the first computer manufacturer to announce and ship a HiPPI channel for their mainframe products. Subsequently, other vendors in the technical computing market have begun to deliver HiPPI channels. Most notable has been Cray Research who have also aggressively pursued software support in their standard operating system UNICOS.

Peripherals

One of the earliest effects of the HiPPI standards effort was to stimulate peripheral manufacturers efforts. Broad support of a high performance channel by the computer vendors immediately created a "plug compatible" peripheral market. Disk arrays, tape cartridge drives and frame buffers are early examples of announced product which also require the high data transfer rates achievable with HiPPI.

Switches

Network Systems has been an active member of the X3T9.3 committee since its inception and was perhaps the first vendor to ship a HiPPI compliant product in the form of a switch. HiPPI switches provide for the very rapid connection of input channels to output channels. Currently, products support up to eight input and eight output ports per chassis. Switches may be cascaded to form larger fabrics as described below. Thirty-two port switches have been announced for availability later this year.

Extenders

HiPPI's twenty-five meter cable length imposes a severe restriction on most applications. Several companies have delivered fiber extenders for full rate HiPPI channel extension. Using either multi-mode or single-mode fiber pairs, distances of several kilometers can be reached. Extenders may attach switches to one another enabling switched high speed connections over campus distances.

Work has recently begun at Network Systems to couple HiPPI fabrics using SONET (Synchronous Optical NETwork) facilities at the OC-12 signaling rate which is about 622 Mbps. This will initially be targeted to metropolitan area distance requirements. As the technology matures, it is intended that this interface would incorporate an ATM cellifier and data rates up to OC-24 which is 1.244 Gbps. ATM will support variable data rates and the creation of virtual circuits to multiple remote destinations.

Gateways

Traditional internetworks have firmly entrenched the role of bridges and routers in any but the simplest of networks. As the potential applications for HiPPI grow to demand extended "fabrics", perhaps over geographic distances, there will be a need for gateways engineered to operate at HiPPI rates.

Network Systems is currently developing a family of HiPPI gateways as part of its work in the Carnegie Mellon NECTAR project. In NECTAR, the gateways are known as CABs for Communication Accelerator Boards. Indeed, one of the projected uses for HiPPI gateways involves the interfacing of existing bus based systems to the fabric; this was the original intent of the CAB in the NECTAR architecture.

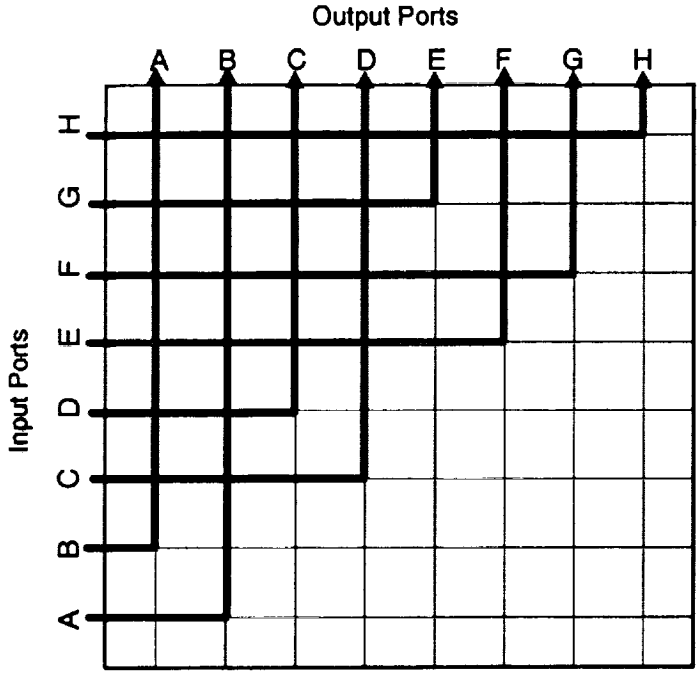
CABs will also exist within HiPPI networks to provides various types of bridging functions. For instance, where long haul extenders are inserted into a network it may be prudent to interface each end via a CAB. The CABs keep a permanent HiPPI connection up between them. Each CAB is prepared to accept HiPPI connections from the user side for forwarding over the extender. This design avoids the latency necessary to establish an end to end HiPPI circuit before the first word can be transmitted. The existence of the CAB will generally be transparent to the user nodes.

Other functions proposed for CABs include security functions to enforce network level access control. Current research is focused on ways CABs may be used to perform outboard protocol assist functions for host computers.

Building Crosspoint Switch "Fabrics"

HiPPI is fundamentally a connection oriented interface standard. One must actually create a HiPPI connection (via control circuits in the physical layer) before data can flow. This is true even for point to point HiPPI cable connections. The basic idea of a crosspoint switch is familiar to anyone with even the barest understanding of telephony. Any input port may be switched in some fashion to any not-busy output port. Once connected, data may flow at the nominal port rate without regard to other connections though the switch.

So far, Network Systems HiPPI switches are true crosspoint switches in that there are no shared data paths. All ports may simultaneously move data at the nominal rate without any contention effects providing for impressive aggregate throughput. Also, the switches are "non-blocking" internally which means that as long as the output port is not busy any input port may connect regardless of other connections in the switch.



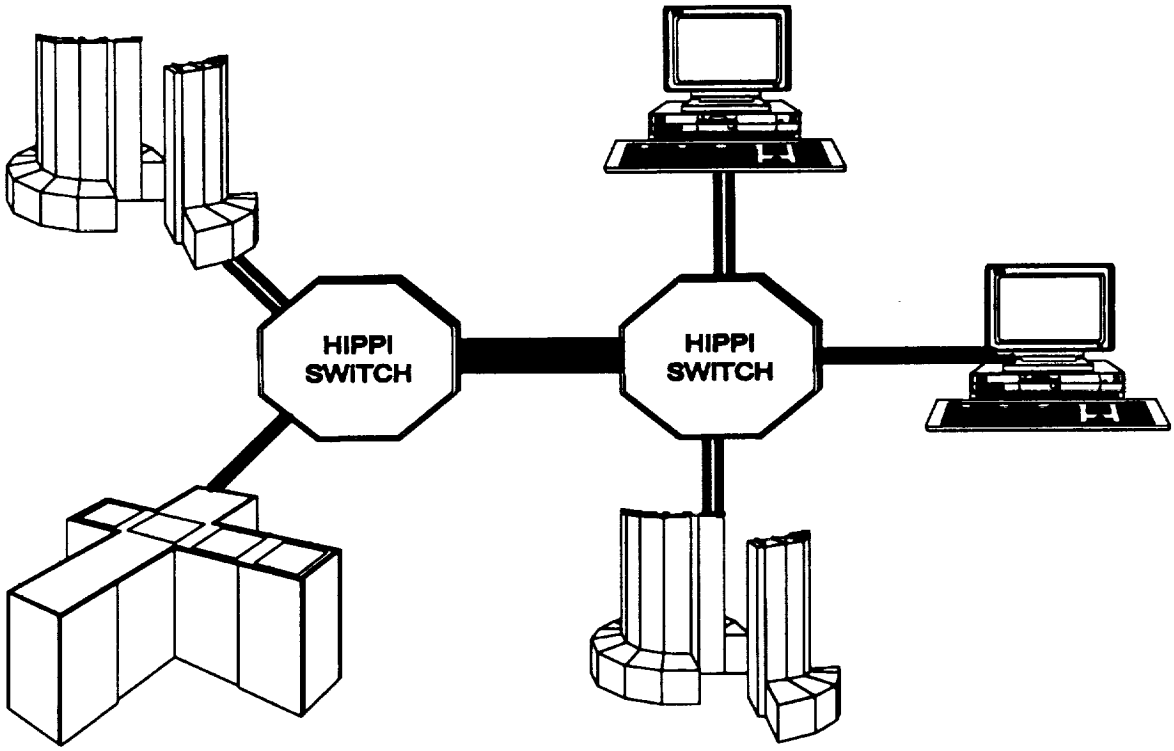
For near term applications in backend networking for supercomputers or attachment of peripherals, single stage switches with four to thirty-two port pairs are probably adequate. However, the limits of board to board connector technology means that we are rapidly approaching the limits of current switch architecture. Therefor, requirements which dictate greater HiPPI connectivity will probably use multi-stage switches constructed by cascading existing switches.

Cascading HIPPI Switches

The output port of a HiPPI switch may be connected to the input port of another (or the same) switch. At each stage, the input port may be switched to any not busy output port of that switch. Switches are designed to propagate the necessary switching signals from input to output such that the existence of the multiple switching stages is essentially transparent to the end points. Once a HiPPI circuit is established through a multi-stage switch fabric, the only noticeable difference from direct cable connections would be a negligible amount of additional data latency.[1] The process of creating a HiPPI switch connection is dependent upon the switch interpreting an in-band address designated by the originator. Note that in a multi-stage switch configuration, each prior stage becomes the originator for each subsequent switch stage until the end-point is reached.

---

[1]Current HiPPI switch products add approximately 160 nsecs of latency to data. This is roughly comparable to the latency due to 25 meters of cable.



Addressing

The basis for HiPPI connection switching is something called the "I-Field" in the HIPPI-PH standard. The I-Field is the contents of the 32 bit wide address circuits of the HiPPI channel at the time the connection request control circuit is raised. The high order octet carries control flags and the low order twenty-four bits are used for the actual addressing.

The HIPPI-SC standard defines two modes of addressing. Either may be used to create multi-stage HiPPI switch connections.

Source Routed Addressing

In source routed addressing, each switch stage examines the several low order bits in the I-Field necessary to address an output port. For instance, an eight port box requires three bits to address ports 0 through 7.

To support multi-stage switching, the switch can optionally rotate the field to bring the next *N* bits into position for the next switch stage. Preservation of the path information is important for the last stage switch. It may be set to automatically create a reverse HiPPI circuit for dual simplex connections.[2]

Source routed address interpretation in switches will typically be performed in hardware providing for very high performance switching.[3] The disadvantage of source routing is that the end point systems must keep a record of the switch fabric topology. The route to each resource will be different for each from-point complicating address table administration. For small networks, this has not been judged to be a problem. But recently, requirements have started to surface for multi-thousand port HiPPI fabrics.

Isomorphic Addressing

Most people are more familiar with isomorphic addressing than the source routing approach. This is the same concept as in Ethernet networks. Each attachment to the network has a unique address which is unrelated to the network topology and need not change when the node is moved to a new point on the network.

Second generation switches support the use of isomorphic addressing which is selected with one of the flag bits in the I-Field. The address portion of the I-Field is split into two twelve bit fields; a to address and a from address. The "to" address is interpreted by each switch stage to determine the next outbound port. Obviously, isomorphic addressing limits HiPPI switch fabrics to a maximum of 4096 addressable nodes.[4]

With isomorphic addressing, boundary nodes are relieved of the need to know about the network topology. Instead they rely upon the collective knowledge contained in the switch forwarding tables. The HiPPI standards do not specify how these tables are created or inserted into the fabric. This is the subject of a current project at Network Systems concerned with switch management.

---

[2]Many of the planned HiPPI applications do not require duplex connections. For instance, frame buffers are essentially simplex, write only devices. Connectionless protocols which use IEEE 802.2 procedures also do not require immediate reverse connections.

[3]The original Network Systems P8 first generation switch is capable of establishing source routed connections in 240 nsecs.

[4]Notice that for multi-stage switch arrays only boundary ports need to consume isomorphic address space. Inter-switch ports may be addressed if necessary using source routed addressing modes.

Switch Management

The switch management project is focused on the practical details of constructing and using arbitrarily large switch fabrics. It is also directing switch features which contribute to the resiliency of the fabric when inevitable failures occur.

Auto-configuration

The foregoing discussion of isomorphic addressing makes clear the necessity of some automatic means for a large multi-switch network to configure itself. By this, we mean the creation of the forwarding tables using connectivity data received from neighboring switches. This is analogous to the techniques used by spanning tree bridge networks to automatically discover the "best" path to a destination.

This process is also intended to support alternate pathing since most practical HiPPI fabrics will contain many possible ways to route a connection from the originating port to a destination. Frequent updating of the tables through the automatic process also provides for routing around failed components. Lastly, the switch management features will provide a means for address resolution similar to that done in internetworks.

None of the switch management features will preclude the use of the HiPPI network for attachment of simple peripherals. Participation in advanced services by boundary nodes is optional.

Additional Services

Closely related to switch address management are the provision of two additional services under consideration. Multi-cast delivery of data is an outgrowth of the address resolution function. It will be possible for boundary nodes to be joined to a multicast group. A sending node may address a HiPPI connection to a multi-cast group address. The switches will provide a best efforts delivery to each node in the multi-cast group.

Network access control services will be provided through forwarding table management. This will allow an administrator to restrict the possible connections from any boundary node.

HiPPI Applications

The HiPPI standards are still being finalized and related products have only been available for a short time. There are many applications for which HiPPI has been proposed. Few of these have been proven for

practical application as of mid 1991. However, the following should be considered representative of the potential breadth of use for this new technology.

Device Connections

Since HiPPI is directly descended from the Cray HSX channel, it seems obvious that it will be used as an open standard computer to peripheral channel. Currently available disk array controllers capable of 500 to 800 Mbps transfer rates clearly demand HiPPI rates. High density tape cartridge systems can read and write in the hundreds of megabits per second range. Some types of telemetry recording devices are being adapted to HiPPI which are capable of Gbps rates.

Another special type of peripheral is the frame buffer used to image animated high resolution displays of complex scientific data. At 24 frames per second, this application requires over 700 Mbps data rates.

The availability of HiPPI switches leverages the advantage of a multi-vendor standard peripheral channel. Any peripheral on a HiPPI switch fabric is potentially shareable by any other nodes on the fabric. Although this sounds like the old Block Mux Channel switch often seen in IBM shops, the rapid switching rates and high transfer rates make this a feasible application even in supercomputer environments.

Backend Networks

The earliest "production" uses of HiPPI are expected to be computer to computer file blasting applications. Standard protocols such as TCP/IP will be supported by most computer vendors who have HiPPI channels on their hosts. This, in turn, will allow higher speed FTP and NFS based data access from host based file servers.

There is a general misconception that TCP/IP is not capable of achieving gigabit per second network speeds. However, multiple researchers have found that there is no intrinsic reason that TCP/IP should not perform in the super gigabit range.[5] In most instances, poor implementation or operating system interference have delivered disappointing network performance.

---

[5]See "How Slow Is One Gigabit Per Second?" by Craig Partridge; BBN Systems and Technology Corporation, Report No. 7080, June 5, 1989.

The availability of high performance networks based upon HiPPI is expected to stimulate vendor efforts in improving protocol performance. Cray Research has, so far, been the leader in this effort.

Backbone Networks

Interestingly, the rapid connect processing of the HiPPI switches makes them suitable for the delivery of short message traffic. It is entirely feasible to "dial-up" a HiPPI connection for each datagram. Each port on current switch products can potentially deliver several million short packets per second.

Todays bridges and routers are not capable of forwarding millions of packets per second. However, HiPPI switches are relatively inexpensive and provide a high performance "media" for the interconnection of high performance bridge routers. Network Systems will deliver HiPPI interfaces for its bridge routers towards the end of this year.

Isochronous Data Routing

A fascinating application of HiPPI involves the transfer of arbitrary digital information. As long as the peak transfer rate requirement does not exceed the HiPPI burst rate of 800 or 1600 Mbps, virtually any type of data can be carried. Continuous or bursty, chunked or non-protocoled, HiPPI imposes minimal constraints on the datastream.

Examples of digital data types considered for HiPPI channels are:

- Digital High Definition TV
- Digitized voice
- Imagery
- Telemetry data

Potential Storage Subsystem Application

The client server model has been applied to files servers from PCs to supercomputers. Despite this success it has serious flaws in the current implementations. One or more computers manage a catalog of files on behalf of one or more client systems so as to facilitate sharing. However, the management computer is also used to retrieve (read) the data from storage peripherals and send a copy (write) to the client.

The server computer is clearly a bottleneck to performance. This design does not scale well and in the supercomputer range literally requires a supercomputer to provide effective file service.

Since the advent of HiPPI switch attachable storage media such as RAIDS and cartridge tape systems, a new file server model has begun to evolve. The obvious but essential idea being that the management computer and the client computers can share direct access to storage peripherals. Access to catalog information by clients need not be across the HiPPI fabric since it is a low bandwidth application.

Most who first consider this concept are aghast that the storage peripheral is left so exposed to unmediated access. The fear of unauthorized access or worse, erasure of valuable data immediately arises.

However, let's consider the following:

- The catalog information will probably exist on private media for optimized access by the server system.
- HiPPI is inherently a simplex media (with flow control). A "read-only" connection can be established from the peripheral to the client system to prevent unauthorized erasure.
- HiPPI switch fabrics will support access control mechanisms such that connection to specific ports may be restricted to specified clients.
- Adequately intelligent peripherals may be instructed by the server computer to stage data, create a simplex connection to the client and then transfer the data as flow controlled via the HiPPI connection.

Many objections can be raised about this concept but equally many solutions have been discussed. No one has yet demonstrated such a system but the author has reason to believe that a commercial implementation will be available in less than a year. The advantages, both technical and economic are so compelling that it must be taken seriously.

Related Emerging Standards

Although this paper has focused on HiPPI because it is here now, there are other standards that will augment or in some cases replace HiPPI for similar needs.

Fiber Channel

Fiber channel is also an emerging computer/peripheral interface spanning a wide performance spectrum up to roughly a Gbps. Like HiPPI, it is also fundamentally a point to point, connection oriented interface.

Network Systems expects to see a demand for fiber channel to HiPPI bridges. Fiber Channel is also well suited for multi-pointing via switches.

SONET

The Synchronous Optical NETwork standards have been adopted by most telecommunications companies on a world wide basis. Signaling rates and multiplex framing standards have been defined from 51.84 Mbps (OC-1) to 2.488 Gbps (OC-48). The large telephony market is expected to create a supply of inexpensive SONET standard components which may be used for data oriented applications.

SONET is also seen as the basis for a national communications infra-structure capable of supporting gigabit per second data applications. As previously stated, a HiPPI over SONET bridge is under development at Network Systems.

ATM

Asynchronous Transfer Mode is associated with SONET and is also promulgated by the telephony industry. Based upon cell relay concepts, ATM will eventually support the economic carriage of bursty data over wide area or metropolitan virtual circuits.

Currently envisioned data applications hide the existence of the cell fabric from the user. The effect, however, will be to allow the cost effective extension of gigabit scale networks over geographic distances.

Conclusion

The HiPPI standards and HiPPI switches are expected to have a significant near term impact on the design and use of mass storage systems. The least optimistic projections recognize the availability of a widely supported standard which offers an order of magnitude improvement over currently available data rates for access to data. Additionally, the creation of an open computer peripheral channel standard is stimulating the development of high performance, cost competitive peripherals accessible from many computer platforms.

More far reaching is the possibility of new client server implementations for mass storage access. The first step implementations are expected this year, with multiple vendor support for direct device access by

1993. Related standards promise geographic access to mass storage libraries at gigabit per second data rates by the mid 1990s.

THE NATIONAL SPACE SCIENCE DATA CENTER
- AN OPERATIONAL PERSPECTIVE -

Ronald Blitstein, ST Systems Corporation/NSSDC
Dr. James L. Green, NSSDC

ABSTRACT

The National Space Science Data Center (NSSDC) manages over 110,000 data tapes with over 4,000 data sets. The size of the digital archive is approximately 6,000 GBytes and is expected to grow to more than 28,000 GBytes by 1995. The NSSDC is involved in several initiatives to better serve the scientific community and improve the management of current and future data holdings. These initiatives address the need to manage data to ensure ready access by the user and manage the media to ensure continuing accessibility and integrity of the data.

This paper will present an operational view of the NSSDC, outlining current policies and procedures that have been implemented to ensure the effective use of available resources to support service and mission goals, and maintain compliance with prescribed data management directives.

INTRODUCTION

The NSSDC is a heterogeneous data archive and distribution center operating in a dramatically changing scientific and technological environment. For most of its thirty year history, it has operated as a batch-oriented library providing custom support for the ingest and distribution of data. Its rate of growth, as measured in volume of data held and request activity have been steady but modest when compared with expected future activity. The NSSDC responds to approximately 3000 requests for data per year. Some requests are supported through on-line or near-line capabilities, but many are filled through the replication and distribution of data tapes or images. During the past five years, over 8500 individual requestors have been provided data, with over thirty percent of them repeat customers. The average volume of data distributed with each request has increased dramatically from 900 MB to 1500 MB during this period. As a data archive, the NSSDC has established policies for media and data management that strive to ensure the continued integrity and availability of its data holdings. These policies cover the ingest, archive, maintenance, and migration of data, as well as the management of the supporting documentation, software, and metadata necessary to meaningfully access and use the data.

INGEST AND ARCHIVE ENVIRONMENT

All data currently received at or generated by the NSSDC enter the archive through a data ingest process. This process requires that two copies of each data volume are made to the current "technology pair" (described below) of media identified by the data center. A copy is retained locally and the other sent off site to the backup archive currently maintained at the Washington National Records Center (WNRC). After the copies are made and validated, the original data volume is not retained. This procedure ensures that the data are written to new, archive-quality media and enables the NSSDC to accurately track each creation date.

As an integral part of the ingest process, entries are made in catalog and inventory data bases. These data bases track spacecraft, experiment, and data set attributes of value to researchers and browsers of the NSSDC's data holdings. Additionally, media-specific information is entered which enables the data center to locate and retrieve desired data files and manage media characteristics, usage, and maintenance actions necessary to ensure the continued integrity and technological currency of the media.

The concept of "technology pair" is one developed by the NSSDC in response to the accelerated obsolescence of recording media resulting from the rapid development and introduction of new storage technologies. The frequency with which new technologies are being introduced makes it difficult to identify and evaluate the archivability of any one media/format before another enters the marketplace. The concept defines the archivability of new media in terms of several factors. These criteria evaluate the appropriateness of any media through the extended lifecycle expected of an archive facility.

- Degree of standardization
- Availability of hardware/software
- Error detection/correction
- Integrity as a function of age
- Capacity
- Transfer rate
- Compatibility with robotic load devices

The NSSDC has identified two technologies, 9-track/6250 bpi and IBM 3480 cartridges, as its current media of choice for institutional archival purposes. Additionally, the data center has installed capability to support near-line archival of data on 12-inch WORM platters, each with a capacity of 2 GB or 6.2 GB. As a new medium is selected, acquired, and installed to support the full spectrum of operational requirements, it will replace the oldest technology then in place (eg. 9-track). Together with the other currently supported medium (eg.

IBM 3480 cartridge), will then comprise the new technology pair. This conservative approach ensures that unforseen problems with new media do not jeopardize the total archive holdings, and that orderly migration of data from one media to another is supported.

Historically, the NSSDC was often resource constrained and the ability to generate and maintain a backup copy of all data was often beyond its reach. This occurred during periods of relatively low rate of change in the technological environment, and the "push" from the commercial sector to adopt new media technologies did not exist. Most of the data ingested at the NSSDC during this period came from missions in progress, and the project scientists provided back up capability with their copies of the data. Today's policies reflect a new philosophy in stewardship, where the total responsibility for data management lies with the primary archive data center. To implement these policies, the NSSDC has taken actions to maintain the integrity of its current holdings and prepare for the massive amount of data from future missions. These actions include a comprehensive data restoration effort, migration of data to near-line accessible media, improvement in research tools, and proactive involvement in the data management planning of future missions.

DATA RESTORATION

Through its data restoration effort, the NSSDC is currently migrating its older data holdings to new technology pairs. Success in this effort has been outstanding. To date, data recorded on approximately 25 percent of the media volumes in the archive have been migrated with greater than 98 percent of the integrity preserved. These media volumes were in 7-track and low density 9-track formats, many 20 years old or more. The success of this effort was unexpected, and many feared that the data would be "lost on earth". But as a result of basically sound storage procedures and the development of an appropriate set of procedures, a more optimistic view is emerging.

The development of data and media management guidelines is very important. A great deal of the success in the data restoration effort can be attributed to the environmental conditions in which the data were archived. NSSDC is continually reviewing its policies in these areas to gain increasing benefit from advances in technology and collective knowledge. Current areas of interest include:

- Pre-certification of archival tapes
- Use of specialized off-site archive facilities
- Increased use of robotic near-line storage for media maintenance
- Error detection and correction
- Data compression

NEAR-LINE DATA ACCESS

The NSSDC is responding to an ever increasing number of scientific inquiries by placing requested data in an public-access retrieval account on the NSI wide area network. Two strategies are being employed to provide this high level of data retrieval, near-line and on-line mass data storage. An example of the success obtainable from effectively managed near-line storage can be found in the data management for the International Ultraviolet Explorer (IUE) mission. The NSSDC has loaded all the IUE data, consisting of over 70,000 unique star images and spectra, in the IBM 3850 Mass Store device operated by the NASA Space and Earth Sciences Computer Center. An interactive system on the NSSDC VAX cluster allows remote users to order data from the electronic Merged Observer Log. This order is processed off-line, where the data of interest are located on the mass store, transferred over the network from the IBM to a public VAX account and a message sent to the requestor that the requested data are ready for retrieval. This process typically takes less than one work day to complete. The use of the mass store is being phased out, and the IUE data will soon be available through a automatic near-line retrieval capability on the NASA Data Archive and Delivery Service (NDADS) optical disk juke box. In its final configuration, NDADS will manage data, meta data, and documentation, all stored within the same system.

On-line access of NSSDC-held data is currently possible for smaller, often requested data sets. The NSSDC On-line Data and Information Services (NODIS) provides public access to data sets that can be researched, viewed, and retrieved by a requestor during a single interactive NSI-net session. Both Earth and space science data are currently available in this manner, including Nimbus Ozone and merged OMNI data, as well as access to the NASA Master Directory.

RESEARCH TOOLS

Proper data management is only part of the picture. To facilitate the research of data, the meta data needs to be afforded an equal level of support. The researcher needs to know of the existence of possible data of interest, and how to use it once located. Through tools developed by the NSSDC, the researcher is able to spend less time and effort on these actions, and more time doing scientific research on the data. NSSDC is involved in the development and dissemination of NASA-wide directory and catalog information, and has installed versions of its Master Directory in numerous data centers throughout the world.

Once located, the data and their formats must be understood if useful research is to be conducted. NSSDC has promoted the correlative use of data across missions through sponsorship of Coordinated Data Analysis Workshops

(CDAWs). In support of these workshops, Common Data Format (CDF) tools have been developed and implemented to allow the researcher to focus on the content of the data and develop meaningful relationships among data having different resolutions and areas of coverage. With CDAW 9.4 held recently, the latest in CDF and graphical display tools were demonstrated.

Another important research tool is data browse. Browse capabilities have been built into the data organization strategies used extensively for data available on CD-ROM. The NSSDC currently maintains approximately two dozen titles on this medium, supporting research in the Earth, space, and planetary sciences.

PROACTIVE DATA MANAGEMENT FOR FUTURE MISSIONS

Data archives have a responsibility to manage future as well as current data holdings. Through its experiences in data restoration, on-line access, and tool development, the NSSDC is sensitized to problem-avoidance strategies for future missions. The data center has developed a cost model to estimate the resource requirements of data ingest, archival, management, and distribution. It is using data from this model to identify future missions requirements for inclusion in the appropriate Project Data Management Plan (PDMP). The PDMP is a multi-lateral agreement that is executed for all future NASA missions. Data management issues addressed by this plan includes the level of service to be provided by the archive, the nature, volume, and frequency of the data to be ingested, the type of media, expected request activity, etc. This process is enabling the NSSDC to reliably estimate future costs for these missions; a critical element when one considers the very large volumes of data that missions such as EOS and the Space Station will generate. PDMPs have been developed for several of the newer missions, including Magellan and Gamma Ray Observatory.

SUMMARY

The course of future scientific research can not be predicted, nor can the data needs of this research. As a national data archive, the NSSDC must not only ensure the continued integrity of the data intrusted to it, but must also ensure the continuing evolution in its ability to provide the correct data to the user in the correct way. As the volume of its data holdings increases, the shift from specialized service to a uniform spectrum of generic services must continue. The NSSDC is pursuing this goal through various initiatives in mass storage, networks, media management, tool development, and standards advocacy.

As is often true, the hardware capabilities and the technological sophistication necessary for very large mass storage systems is rapidly being

developed. In short duration project environments, the selection, installation, and implementation of viable systems is relatively easy. But in the view of an archival data center, such as the NSSDC, the massive volume of non-homogeneous data from hundreds of missions for which it is responsible make this a very difficult procedure. The selection of high capacity storage media must be accompanied with corresponding strategies to ensure the integrity of the data for many years. The current media transfer rates and the requirement for the generation of backup copies effectively doubles the volume of data to be managed. Higher density storage without accompanying capabilities in robust error detection and correction that provide lossless recovery of data may be inappropriate for permanent archives. Frequent migrations of data from one media to another, especially if accomplished in a true automated fashion, are attractive alternatives to the manual processes widely used today, but pose enormous requirements of inventory and catalog data bases that have visibility across all the various archive systems in use in any facility (or even across facilities in a distributed archive environment).

MASS STORAGE SYSTEM EXPERIENCES AND FUTURE NEEDS AT THE NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Summary of the Presentation to the Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications July 23-25, 1991

by

Bernard T. O'Lear
Manager, Systems Programming
Scientific Computing Division
National Center for Atmospheric Research

3/5/91

This is a summary of the presentation given at the Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications. The presentation was compiled at the National Center for Atmospheric Research (NCAR), Boulder, Colorado. NCAR is operated by the University Corporation for Atmospheric Research and is sponsored by the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author and do not necessarily reflect the views of the National Science Foundation.

This presentation is designed to relate some of the experiences of the Scientific Computing Division at NCAR dealing with the "data problem." A brief history and a development of some basic Mass Storage System (MSS) principles are given. An attempt is made to show how these principles apply to the integration of various components into NCAR's MSS. There is discussion of future MSS needs for future computing environments.

NCAR provides supercomputing and data processing for atmospheric, oceanic and related sciences. This service is provided for university scientists and for scientists located at NCAR. There is a total of about 1200 users.

The data problem for this community can briefly be summarized as follows; Historical atmospheric data is archived, programs are saved and the data which model the atmosphere, oceans and sun are saved. The NCAR storage experience is based upon current supercomputing megaflop rates which produce a number of terabytes archived on a yearly basis. There is a history of data growth and file growth. The NCAR data storage experience has been as follows; There are about 500 bytes of information archived for each megaflop of computing. When NCAR had an X-MP/48, the archive rate for the utilized megaflop compute rate was 3 terabytes per year. The installation of a Y-MP8/864 increased the archival rate to 6 terabytes per year. Forecasting future computing configurations and atmospheric models being planned we are now approximating a 30-50 terabyte archive per year rate by the year 1993 or 1994.

Data has been saved in many forms over NCAR's existence and then migrated to machine-readable media. Some of the data has come from handwritten logs, from punch cards, half-inch tape. All of this has been collected and is now archived on IBM 3480 cartridge tape. One of the basic principles for archiving this data is to identify certain classes of data. Archive data is kept forever. Long-term data is kept for 10 to 15 years. Near-term data is kept for 1 month to 1 year and a category called scratch data is killed after 1 month and cannot be recovered automatically by the system.

One of the other basic principles that has been identified is that dataset sizes continue to grow as a function of supercomputing sizing. The amount of data that can be saved is bound in storage by media capacities. That is, these criteria are established for determining which data will be saved and for how long because there is not an infinite media capacity at this time. Our experience has shown that every 10 to 15 years the data in the MSS will need to be migrated to a new media base because of changing systems and obsolescence of existing media. Usually the media or the drives cannot be purchased anymore. This migration takes place not because the data is bad on the media, but because the drives will not be available.

Another problem is that a number of companies have provided the capability for this massive storage, but the small companies tend to disappear within five years. The drive components that have been furnished for mass data storage disappear in five to eight years no matter what company they come from.

The next basic principle is that the migration of the mass storage system data to a new media base, which is now several ten's of terabytes, is not a trivial operation. The migration does not take place in a short amount of time. For instance, one-time migrations can run for long periods of time, necessarily years to move terabytes data. It is very difficult to guarantee that the data is migrated absolutely without reading it back, which is time consuming. These migrations are very costly and in my opinion shouldn't be done. We have developed the concept of "DATA OOZE," and we prefer this technique over migration right now. The way DATA OOZE works is that it is a continuous movement of data within the system. The data is moving across the

storage hierarchy and across the changing media types under the control of the MSS. The migration path for this data in the hierarchy can be from memory to solid state disk to high speed disk to disk arrays or farms, and from there out to some kind of tape. Later on as new data storage media become available, the data is migrated onto these media in real time, since every day some amount of the data is migrated as it is being used.

Our conclusions from these experiences have been that new components and media types are integrated according to the following rules; Use standard components. The standards may be real or de facto and apply in the areas of channels, interfaces, operating systems, media, etc. We look for media that is easy to obtain and is cost effective. We look for the long-term viability of the vendor and multiple sources for the many system components. In the area of mass storage system integration we look at access speeds, ease of expandability, heterogeneous host access, maintenance costs, media costs and systems costs.

There are a number of future growth issues for the NCAR MSS. The Scientific Computing Division (SCD) continues to develop future configuration scenarios. These scenarios try to anticipate the functional requirements we anticipate providing for our scientific community. There are three key components we need to address: network services and access, the large scale computing (Big Iron), and the data archives. Of course, these all play within the context of distributed computing.

The near-term issues for the NCAR MSS focus on some immediate upgrades which will deal with the MSS growth for a couple of years. The entire archive will be migrated onto double density 3490 and 3490-compatible media. The mid-90s to late 90s became more interesting because of the expanding interest in archiving vast data collections.

The issues of future growth will be centered in three areas of ongoing development: the various MSS software packages, the data storage components and the networks.

The questions then become how all of these components get assembled and which ones do we plan to use. Will SCD be able to construct on effective peta-byte MSS by the end of the decade? Which of our basic principles can we apply to insure that such a system can be built?

# # #

1

Mass Storage System Experiences and Future Needs at The National Center for Atmospheric Research

Conference on Mass Storage Systems
and Technologies for
Space and Earth Science Applications

July 23-25, 1991

Bernard T. O'Lear
Manager, Systems Programming
Scientific Computing Division
National Center for Atmospheric Research

2

The following presentation was compiled at the National Center for Atmospheric Research Boulder, Colorado

- The National Center for Atmospheric Research is operated by the University Corporation for Atmospheric Research and is sponsored by the National Science Foundation.
- Any opinions, findings, conclusions, or recommendations expressed in this talk are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Introduction

The experiences of a scientific center dealing with

"The Data Problem"

- **Brief history**
- **The current computing environment**
- **Development of some basic principles**
- **How the principles apply**
- **Future needs for future computing environments**

History

NCAR provides supercomputing and data processing for atmospheric, oceanic, and related sciences:

- **At universities**
- **At NCAR**
  - **Totals about 1200 users**

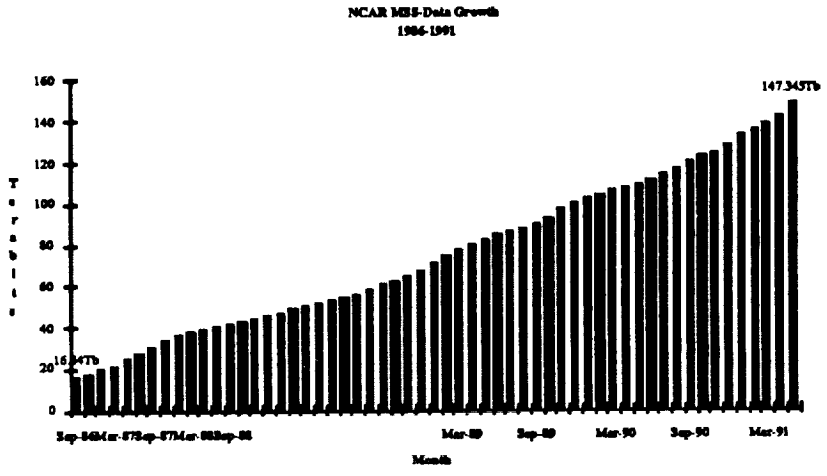
The data problem for this community

- **Save and archive historical atmospheric data**
- **Save programs and data which model the atmosphere, oceans, and sun**

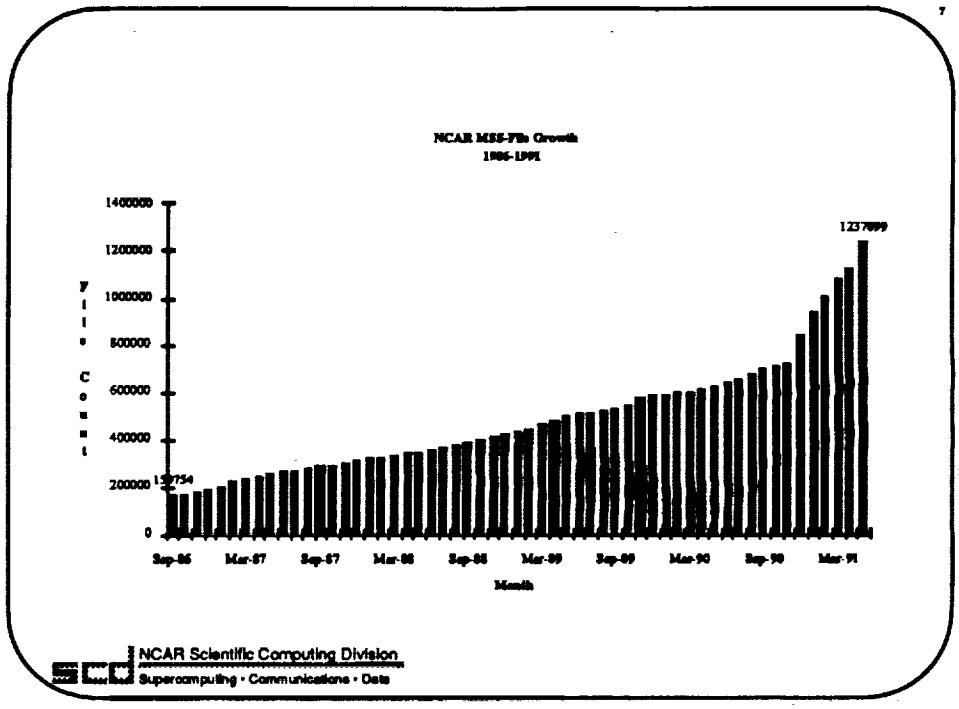
The NCAR Storage Experience

- 500 Bytes per million flop
- Archival rate for model output
  - 4 TBytes/year with X-MP/48
  - 8 TBytes/year with Y-MP8/864
  - 40 TBytes for climate simulation

NCAR Scientific Computing Division
Supercomputing • Communications • Data



NCAR Scientific Computing Division
Supercomputing • Communications • Data



NCAR Scientific Computing Division
Supercomputing • Communications • Data

NCAR Mass Storage Systems (MSS)

Usage Data

101,700 tape cartridges in use
Over 18.5 Tbytes of data stored
Over 710,000 files
Average file length 26.2 MB

Fast Path

≤ 2 minute delivery

NCAR Scientific Computing Division
Supercomputing • Communications • Data

History (Continued)

Data saved in many forms -- then migrated to machine readable media:

- **Handwritten logs ------> Punched cards**
- **Punched cards ------> One-half inch tape**
- **One-half inch tape ------> AMPEX TBM tape**
- **AMPEX TBM tape ------> IBM 3480 tape**
- **IBM 3480 tape ------> IBM 3490-E tape**
- **IBM 3490-E tape ------> ? ? ?**

Basic Principles

Identification of data classes:

- **Archive data = keep forever**
- **Long-term data = keep 10-15 years**
- **Near-term data = keep 1 month to 1 year**
- **Scratch data = kill after 1 month**

11

Basic Principles (Continued)

- Dataset sizes continue to grow as a function of supercomputer sizing
- Dataset sizes are constrained in Storage by media capacities
- Every ten to fifteen years, the data in the MSS will need to be migrated to a new media base

13

Basic Principles (Continued)

Migration:

- Not because the data is bad on the media...
- But because the drives will not be there
  - Half life of a start-up company = 5 years
  - Half life of drive electronics = 5-8 years

13

Basic Principles (Continued)

The migration of MSS contents ("n" tera-bytes) to new media is not a trivial operation

One-time migrations:

- **Run for long periods of time (years)**
- **Are difficult to guarantee**
- **Are costly**
- **Shouldn't be done**

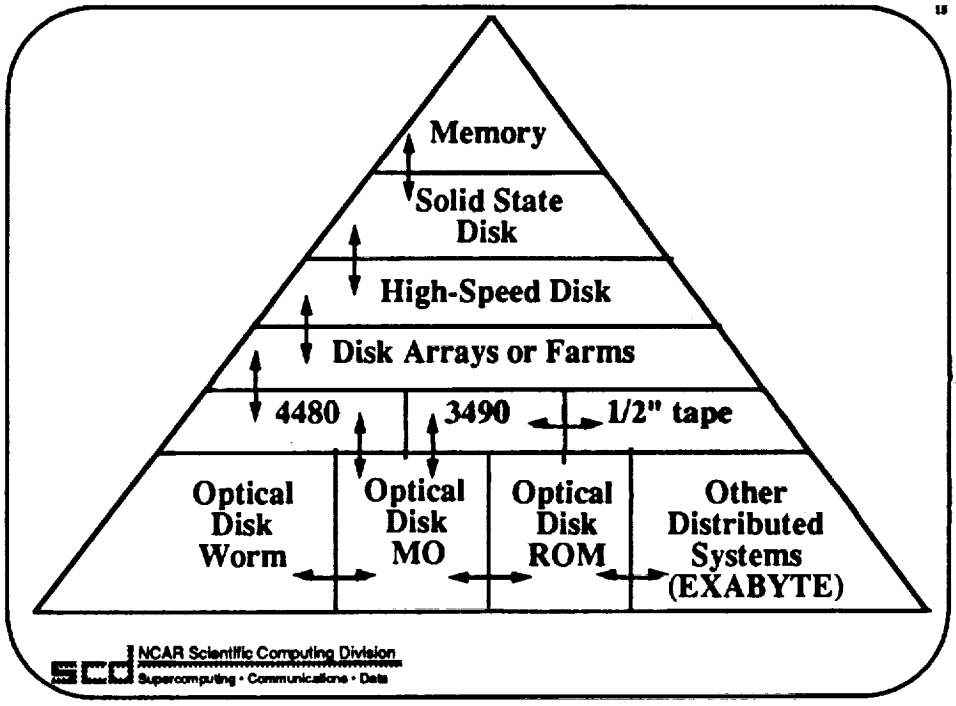
14

Basic Principles (Continued)

Data OOZE preferred over migration

Data OOZE is a continuous movement of data within the system:

- **Data movement across:**
  - **The storage hierarchy**
  - **The changing media types**



CONCLUSIONS

New components and media types are integrated according to these rules:

- Standards (real or de facto)
  - For channels and interfaces (IBM, IPI, HIPPI, SCSI)
  - For media
- Long-term viability of vendor
- Multiple source availability for media (drives?)

NCAR Scientific Computing Division
Supercomputing • Communications • Data

17

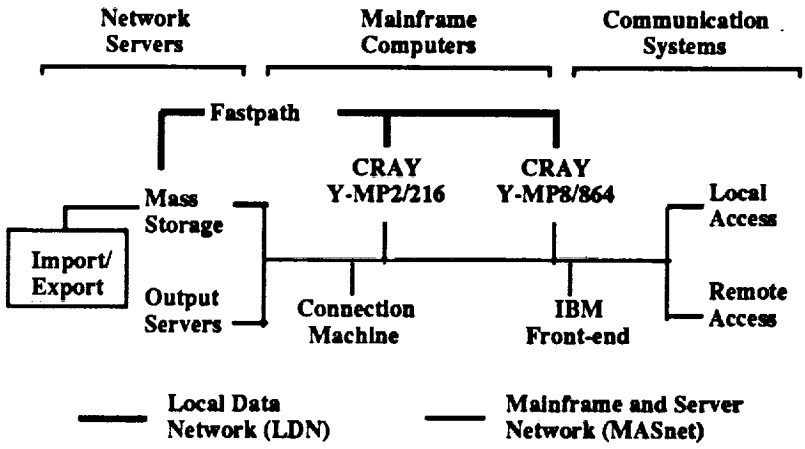
MSS Integration

- Access speeds (sometimes)
- Ease of expandability
- Multiple heterogeneous host access
- Maintenance costs
- Media costs
- System cost

18

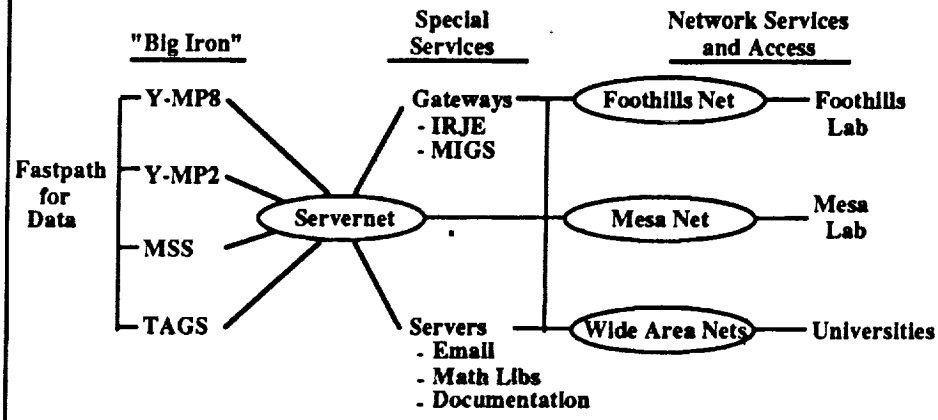
FUTURE GROWTH
ISSUES
IN THE
NCAR MASS STORAGE SYSTEM

Functional Diagram

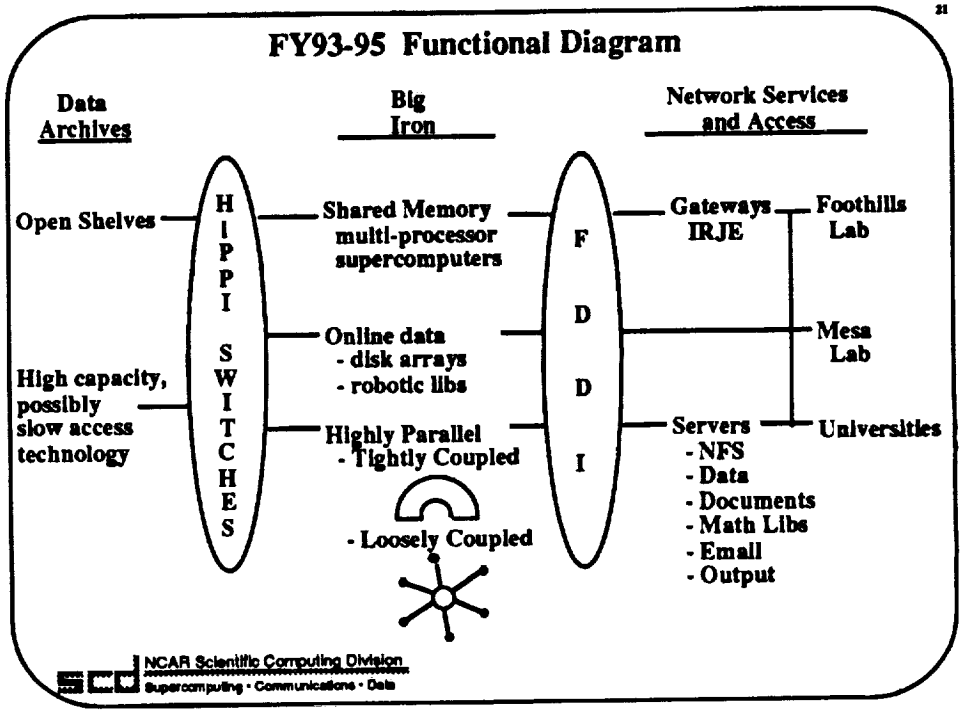


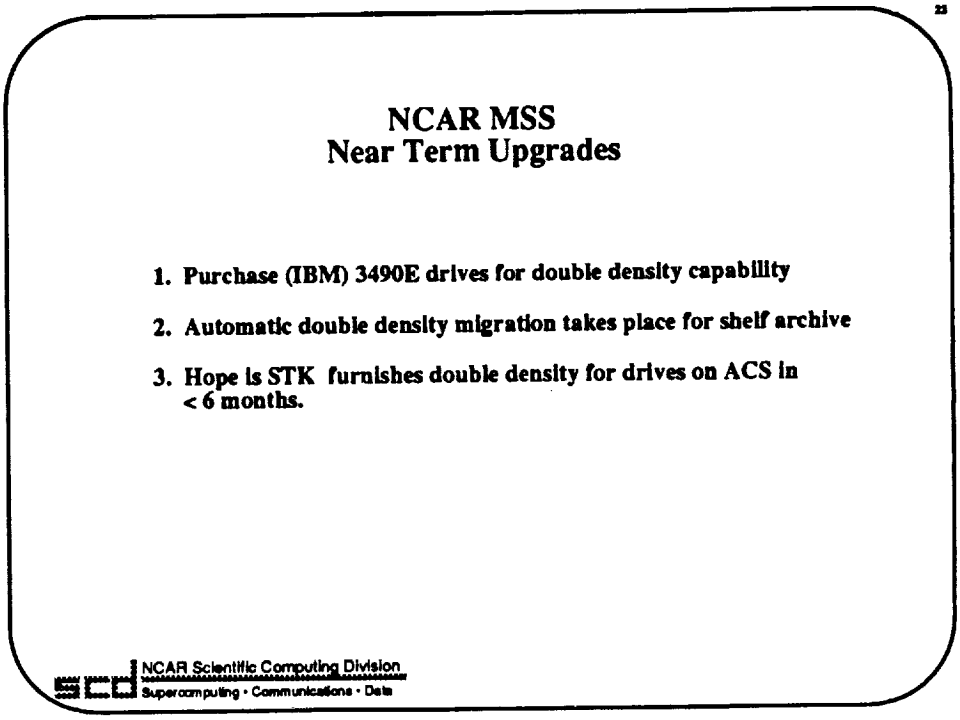
NCAR Scientific Computing Division
Supercomputing • Communications • Data

Functional Diagram of the NCAR Computing Complex "____.UCAR.EDU"



NCAR Scientific Computing Division
Supercomputing • Communications • Data


FY93-95 Functional Diagram
Data Archives
Big Iron
Network Services and Access
Open Shelves
High capacity, possibly slow access technology
HIPPI SWITCHES
Shared Memory multi-processor supercomputers
Online data
- disk arrays
- robotic libs
Highly Parallel
- Tightly Coupled
- Loosely Coupled
FDDI
Gateways IRJE
Servers
- NFS
- Data
- Documents
- Math Libs
- Email
- Output
Foothills Lab
Mesa Lab
Universities
NCAR Scientific Computing Division
Supercomputing • Communications • Data


NCAR MSS
Near Term Upgrades
1. Purchase (IBM) 3490E drives for double density capability
2. Automatic double density migration takes place for shelf archive
3. Hope is STK furnishes double density for drives on ACS in < 6 months.
NCAR Scientific Computing Division
Supercomputing • Communications • Data

NCAR MSS

The Issues of Future Growth are dependent upon:

1. Future MSS Software
   a. Distributed MSS
   b. Large archives (Peta-Byte)
2. Future Data Storage
   a. The media
   b. The drives
   c. The robotics
3. The Network and Channels
   a. HiPPI
   b. Fibre channel standards
   c. fabric (switch)

1. Future MSS Software

a. Distributed MSS

- UNITREE (DISCOS)
- Infinite Storage Architecture (EPOCH)
- Distributed Physical Volume Repository (EPOCH & STK)
- EMASS (E-SYSTEMS)
- NAStore (NASA, Ames)
- NETARC and AWBUS (CDC)
- SWIFT (IBM)
- DataMesh (Hewlett Packard)
- M (DS) [2]NASA Goddard

b. Peta-Byte Archives

- How do we build them?

2. Future Data Storage

a. THE MEDIA

- The 3M National Media Laboratory (Media Database)
- Government funds and goals
- Private sector participation
- Standards being developed

b. THE DRIVES

- Being developed for the media
- 10-year life span
- Attachable to various robotics

c. THE ROBOTICS

- StorageTek is the leader
- ODETICS
- EXABYTE and others

3. The Network and Channels

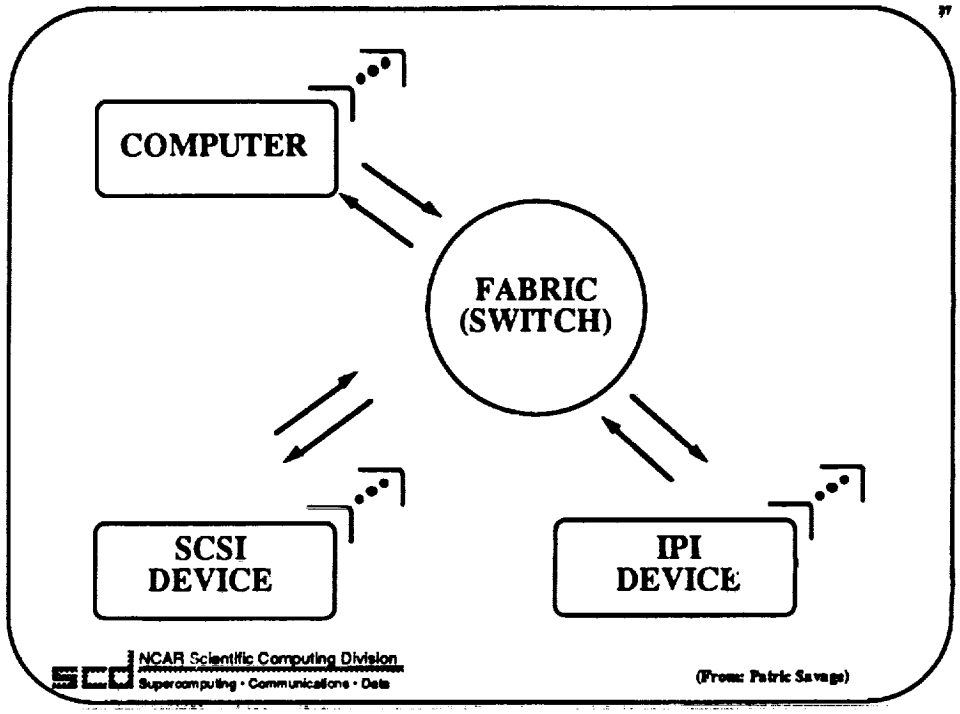
a. Standards moving fast for HIPPI

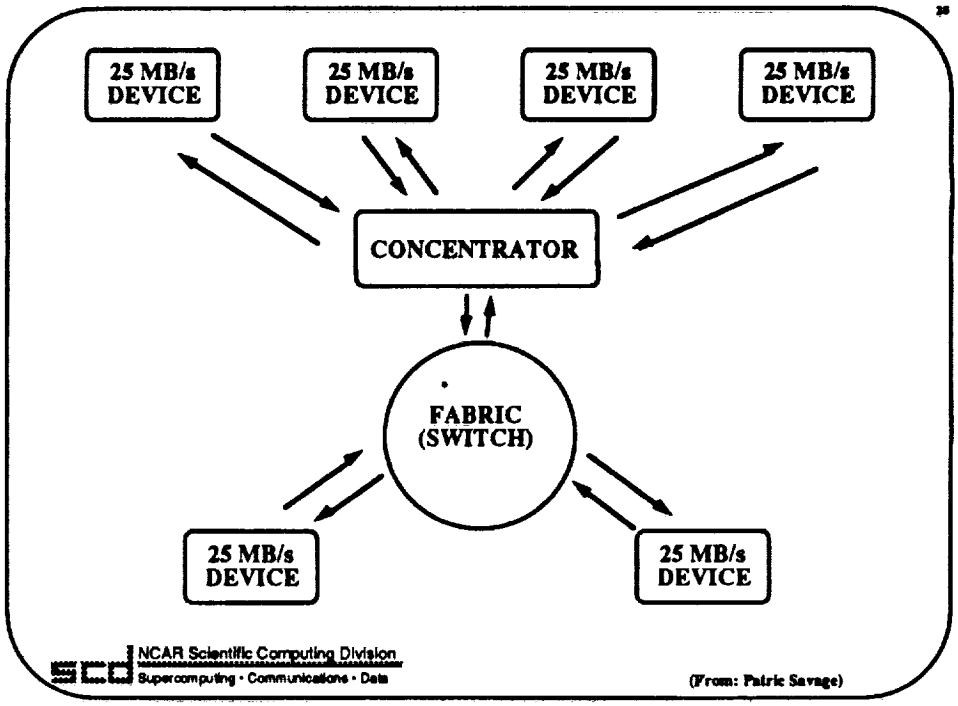
- The HIPPI switch

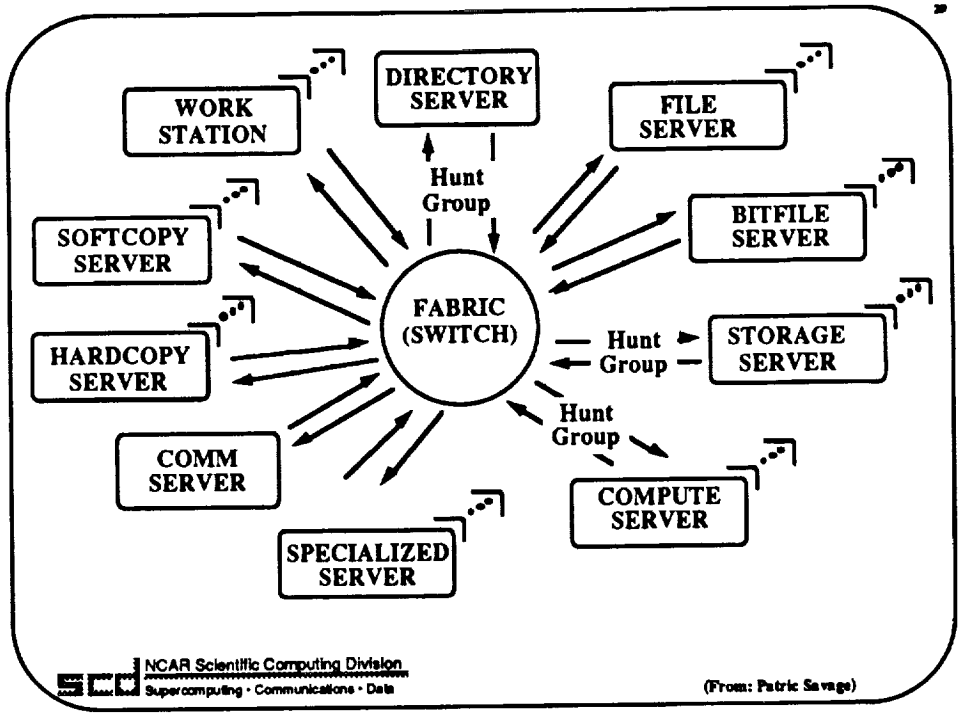
b. Fibre Channel Advantages

- Length to 10 kilometers
- General Protocol
  - HIPPI
  - SCSI
  - IPI
  - Others
- Security
- Immune to Electrical Disturbance

c. Fabric Switch


COMPUTER
FABRIC
(SWITCH)
SCSI
DEVICE
IPI
DEVICE
NCAR Scientific Computing Division
Supercomputing • Communications • Data
(From: Patric Savage)


25 MB/s
DEVICE
25 MB/s
DEVICE
25 MB/s
DEVICE
25 MB/s
DEVICE
CONCENTRATOR
FABRIC
(SWITCH)
25 MB/s
DEVICE
25 MB/s
DEVICE
NCAR Scientific Computing Division
Supercomputing • Communications • Data
(From: Patric Savage)


WORK STATION
DIRECTORY SERVER
FILE SERVER
Hunt Group
SOFTCOPY SERVER
BITFILE SERVER
FABRIC (SWITCH)
HARDCOPY SERVER
Hunt Group
STORAGE SERVER
Hunt Group
COMM SERVER
COMPUTE SERVER
SPECIALIZED SERVER
NCAR Scientific Computing Division
Supercomputing · Communications · Data
(From: Patric Savage)

S20-82
121957
P, 20

N93-15047

Storage Needs in Future Supercomputer Environments

Notes for the presentation by:

Sam Coleman
Lawrence Livermore National Laboratory

July 25, 1991 at the
NASA Goddard "Mass Storage Workshop"

Introduction

The Lawrence Livermore National Laboratory (LLNL) is a Department of Energy contractor, managed by the University of California since 1952. Major projects at the Laboratory include the Strategic Defense Initiative, nuclear weapon design, magnetic and laser fusion, laser isotope separation and weather modeling. The Laboratory employs about 8,000 people. There are two major computer centers: The Livermore Computer Center and the National Energy Research Supercomputer Center.

As we increase the computing capacity of LLNL systems and develop new applications, the need for archival capacity will increase. Rather than quantify that increase, I will discuss the hardware and software architectures that we will need to support advanced applications.

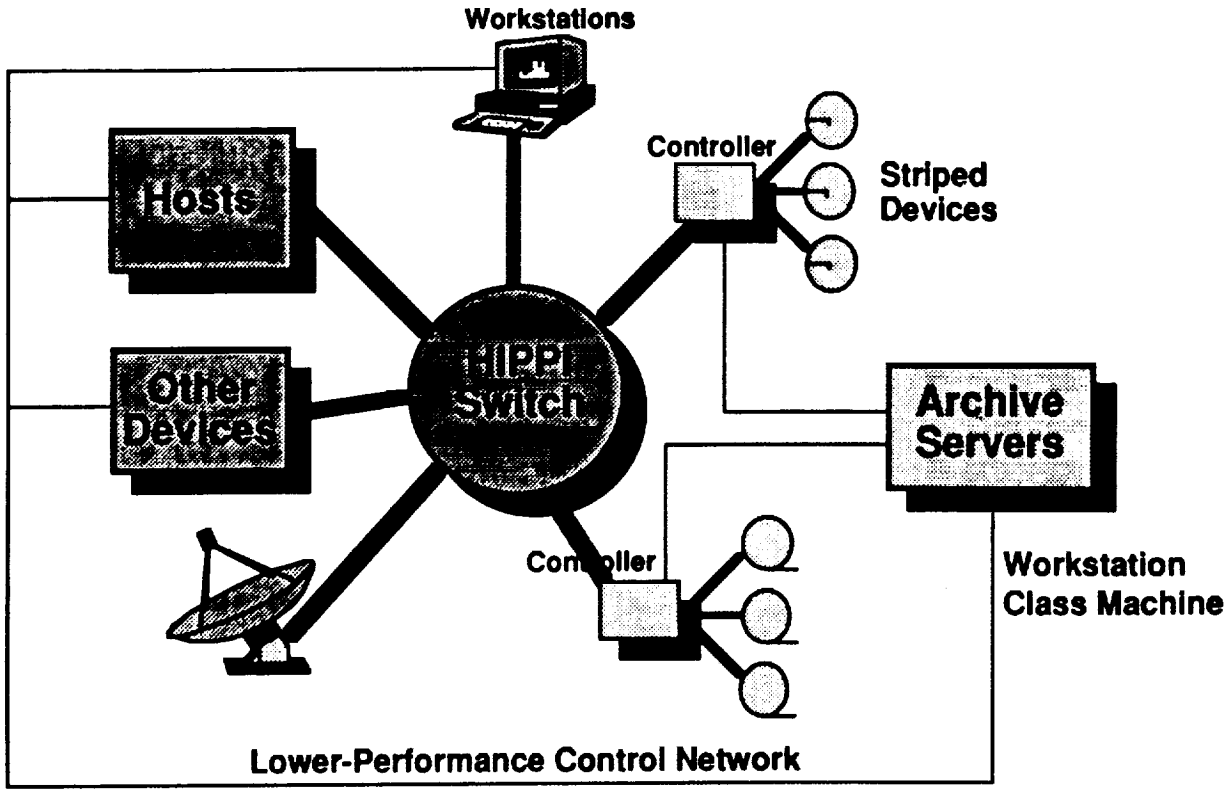
Storage Architectures

The architecture of traditional supercomputer centers, like those at Livermore, include host machines and storage systems linked by a network. Storage nodes consist of storage devices connected to computers that manage those devices. These computers, usually large Amdahl or IBM mainframes, are expensive because they include many I/O channels for high aggregate performance. However, these channels and the devices currently attached to them are individually slow; storage systems based on this architecture will become bottlenecks on HIPPI and other high-performance networks. Computers with the I/O-channel performance to match these networks will be even more expensive than the current machines.

The need for higher-performance storage systems is being driven by the remarkable advances in processor and memory technology available on relatively inexpensive workstations; the same technology is making high-performance networks possible. These advances will encourage scientific-visualization projects and other applications capable of generating and absorbing quantities of data that can only be imagined today.

To provide cost-effective, high-performance storage, we need an architecture like that shown in Figure 1. In this example, striped storage devices, connected to a HIPPI network through device controllers, transmit large blocks of data at high speed. Storage system clients send requests over a lower-performance network, like an Ethernet, to a workstation-class machine controlling the storage system. This machine directs the device controllers, also over a lower-performance path, to send data to or from the HIPPI network. Control messages could also be directed over the HIPPI network, but these small messages

would decrease the efficiency of moving large data blocks; since control messages are small, sending them over a slower network will not degrade the overall performance of the system when large data blocks are accessed (this architecture will not be efficient for applications, like NFS, that transmit small data blocks).



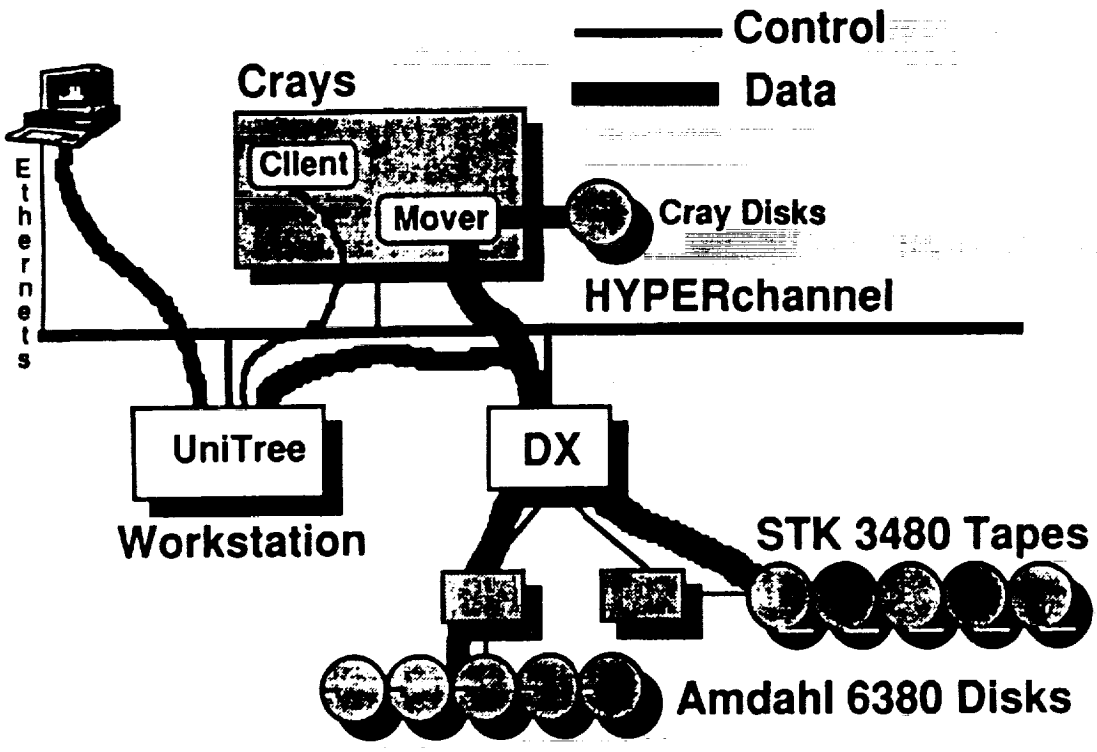
A High-Performance Storage Architecture
Figure 1

To make the architecture in Figure 1 efficient, we will need the following components:

- Programmable device controllers imbedding relatively high-level data-transfer protocols;
- High-performance, possibly striped, archival storage devices to match the performance of the HIPPI network. These devices should be faster than the D1 and D2 magnetic tapes being developed today;
- High-capacity media, with at least the capacity of the largest D2 tape cartridges;
- Robotics to mount volumes quickly;
- Devices and systems that are more reliable than the 1-in-10^{12} error rates quoted today; and
- Devices that are less expensive than the current high-performance devices.

In short, we need reliable, automated archival devices with the capacity of Creo optical tapes (one terabyte per reel), the performance of Maximum Strategy disks (tens of megabytes per second), and the cost of 8mm tape cartridge systems (less than $100,000).

As a step toward the Figure 1 architecture, we are investigating the architecture shown in Figure 2; we will connect existing storage devices to our Network Systems Corp. HYPERchannel, controlled by a workstation-based UniTree system. Even though the hardware connections are



An Interim Storage Architecture at LLNL
Figure 2

available today, the necessary software is not. In particular, there is no high-level file-transport software in the NSC DX HYPERchannel adapter. As an interim solution, we will put IEEE movers' on our host machines, allowing direct file-transport to and from the storage devices over the HYPERchannel. The UniTree workstation will provide service to client workstations and other network machines. This is acceptable, in the near term, because most of the archival load comes from the larger host machines. This architecture will replace the Amdahl mainframes that we use to control the current archive.

Software Needs

To implement high-performance storage architectures, we need file-transport software that supports the network-attached devices in Figure 1. Whether or not the TCP/IP and OSI protocols can transmit data at high speeds is subject to debate; if not, we will have to develop new protocols.

From the human client's point of view, we need software systems that provide transparent access to storage. Several transparencies are described in the IEEE Mass Storage System Reference Model document:[1]

Access
Clients do not know if objects or services are local or remote.

Concurrency
Clients are not aware that other clients are using services concurrently.

Data representation
Clients are not aware that different data representations are used in different parts of the system.

Execution
Programs can execute in any location without being changed.

Fault
Clients are not aware that certain faults have occurred.

Identity
Services do not make use of the identity of their clients.

Location
Clients do not know where objects or services are located.

Migration
Clients are not aware that services have moved.

Naming
Objects have globally unique names which are independent of resource and accessor location.

Performance
Clients see the same performance regardless of the location of objects and services (this is not always achievable unless the user is willing to slow down local performance).

Replication
Clients do not know if objects or services are replicated, and services do not know if clients are replicated.

Semantic
The behavior of operations is independent of the location of operands and the type of failures that occur.

Syntactic
Clients use the same operations and parameters to access local and remote objects and services.

The IEEE Reference Model

One way to achieve transparency is to develop distributed storage systems that span clients environments. In homogeneous environments, like clusters of Digital Equipment Corp. machines, transparency can be achieved using proprietary software. In more heterogeneous supercomputer centers, standard software, running on a variety of machines, is needed. The IEEE Storage System Standards Working Group is developing standards (project 1244) on which transparent software can be built. These standards will be based on the reference model shown in Figure 3. The modules in the model are:

Application
Normal client applications codes.

Bitfile Client
This module represents the library routines or the system calls that interface the application to the Bitfile Server, the Name Server, and the Mover.

Bitfile Server
The Bitfile Server manages abstract objects called bitfiles that represent uninterpreted strings of bits.

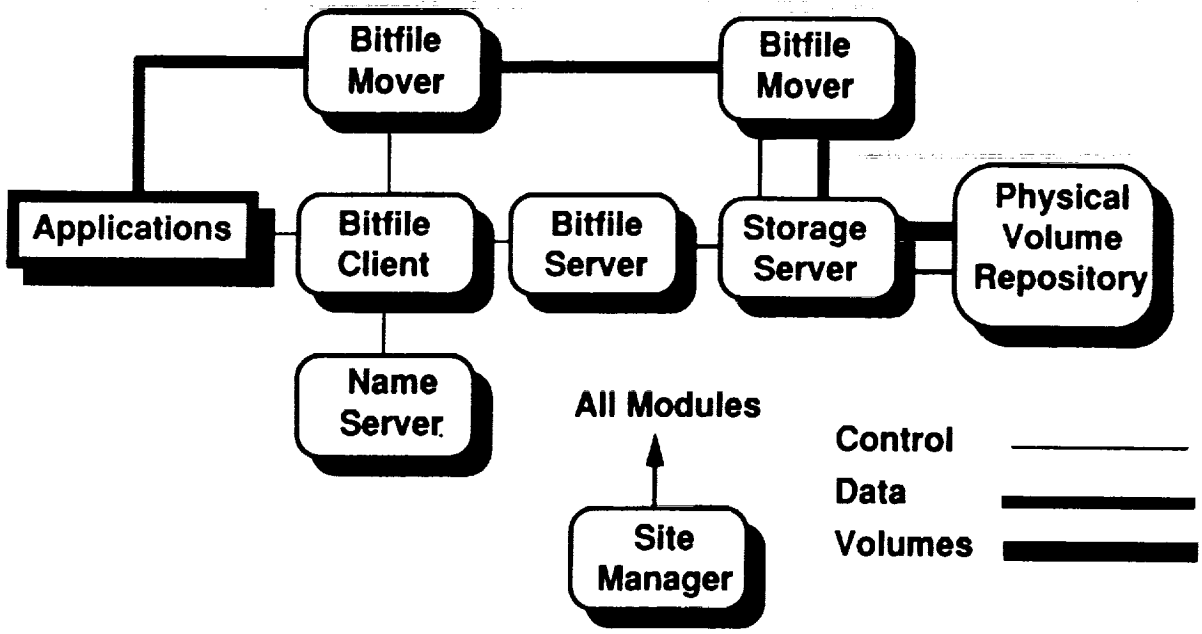
Storage Server
The module that manages the actual storage of bitfiles, allocating media extents, scheduling drives, requesting

volume mounts, and initiating data transfers.

Physical Volume Repository
The PVR manages physical volumes (removable disks, magnetic tapes, etc.) and mounts them on drives, robotically or manually, upon request.

Mover
The Mover transmits data between two channels. The channels can be connected to storage devices, host memories, or networks.

Site Manager
This modules provides the administration interface to all of the other modules of the model.



The IEEE Mass Storage System Reference Model
Figure 3

The key ideas that will allow standards based on the reference model to support transparency are:

- The Mover separates the data path from the control path, allowing the controller-to-network path shown in Figure 1.
- The Name Server isolates the mapping of human-oriented names to machine-oriented bitfile identifiers, allowing the other modules in the model to support a variety of different naming environments.
- The modularity of the Bitfile Client, Bitfile Server, Storage Server, and Physical Volume Repository allows support for different devices and client semantics with a minimum of device- or environment-specific software.

I would like to encourage people attending the Goddard conference to support the IEEE standards effort by participating in the Storage System Standards Working Group. For more information, contact me at:

Sam Coleman
Lawrence Livermore National Laboratory
Mail Stop L-60
P. O. Box 808
Livermore, Ca. 94550
(415) 422-4323
scoleman@llnl.gov

Until standard software systems are available, there are steps that the storage industry can take toward more transparent products. The Sun Microsystems Network File System and the CMU Andrew File System provide a degree of transparency. Work on these systems to improve their security and performance, and to provide links to hierarchical, archival systems, will improve their transparency. I would suggest that software vendors strive to provide operating-system access to archival storage systems, possibly through mechanisms like the AT&T File System Switch.

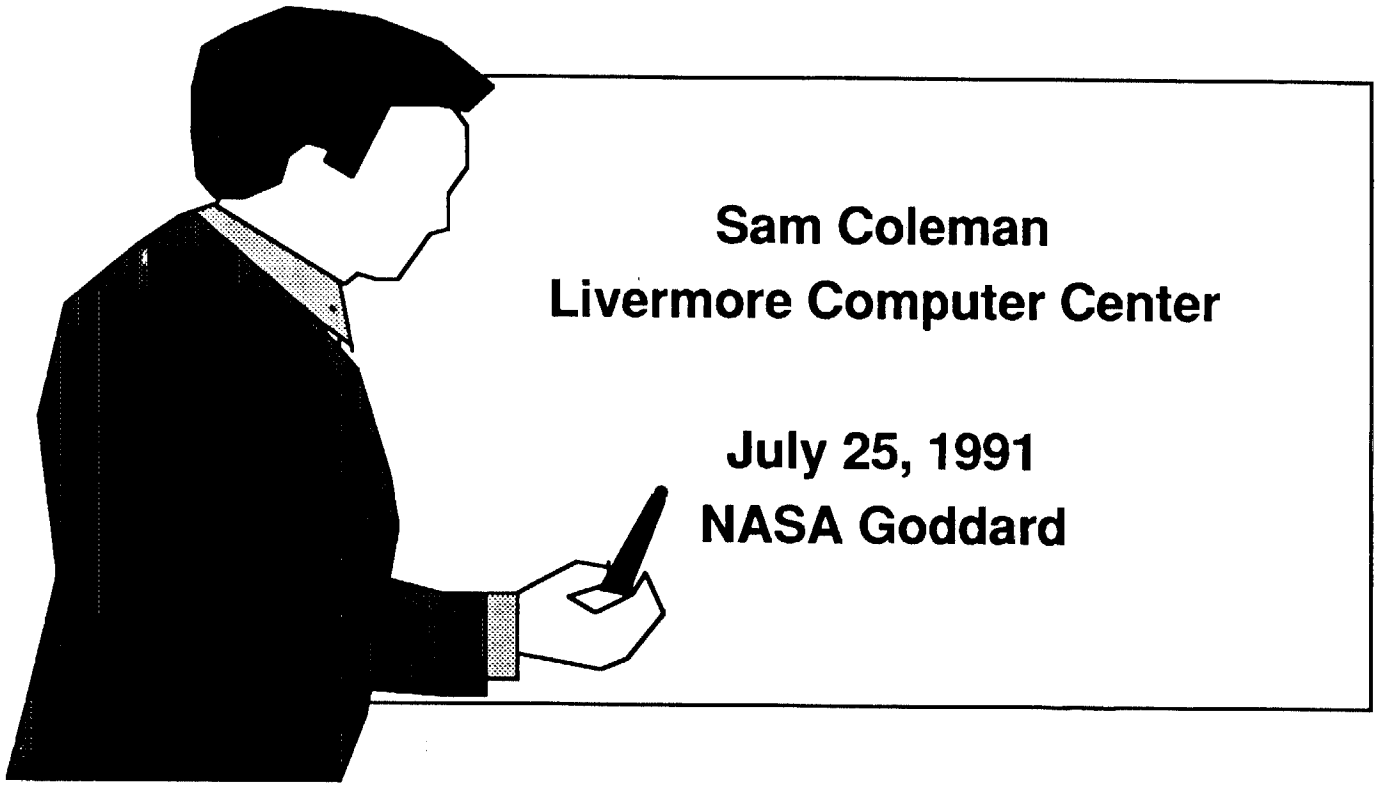
To learn more about all of the storage issues that I have mentioned, I would encourage you to attend the 11th IEEE Mass Storage Symposium in Monterey, California October 7-10, 1991. For details, contact:

Bernie O'Lear
National Center for Atmospheric Research
P. O. Box 3000
Boulder, Colorado 80307

Reference

1. Coleman, S. and Miller, S., editors, *A Reference Model for Mass Storage Systems*, IEEE Technical Committee on Mass Storage Systems and Technology, May, 1990.

Storage Needs in Future Supercomputer Environments



Sam Coleman

Livermore Computer Center

July 25, 1991

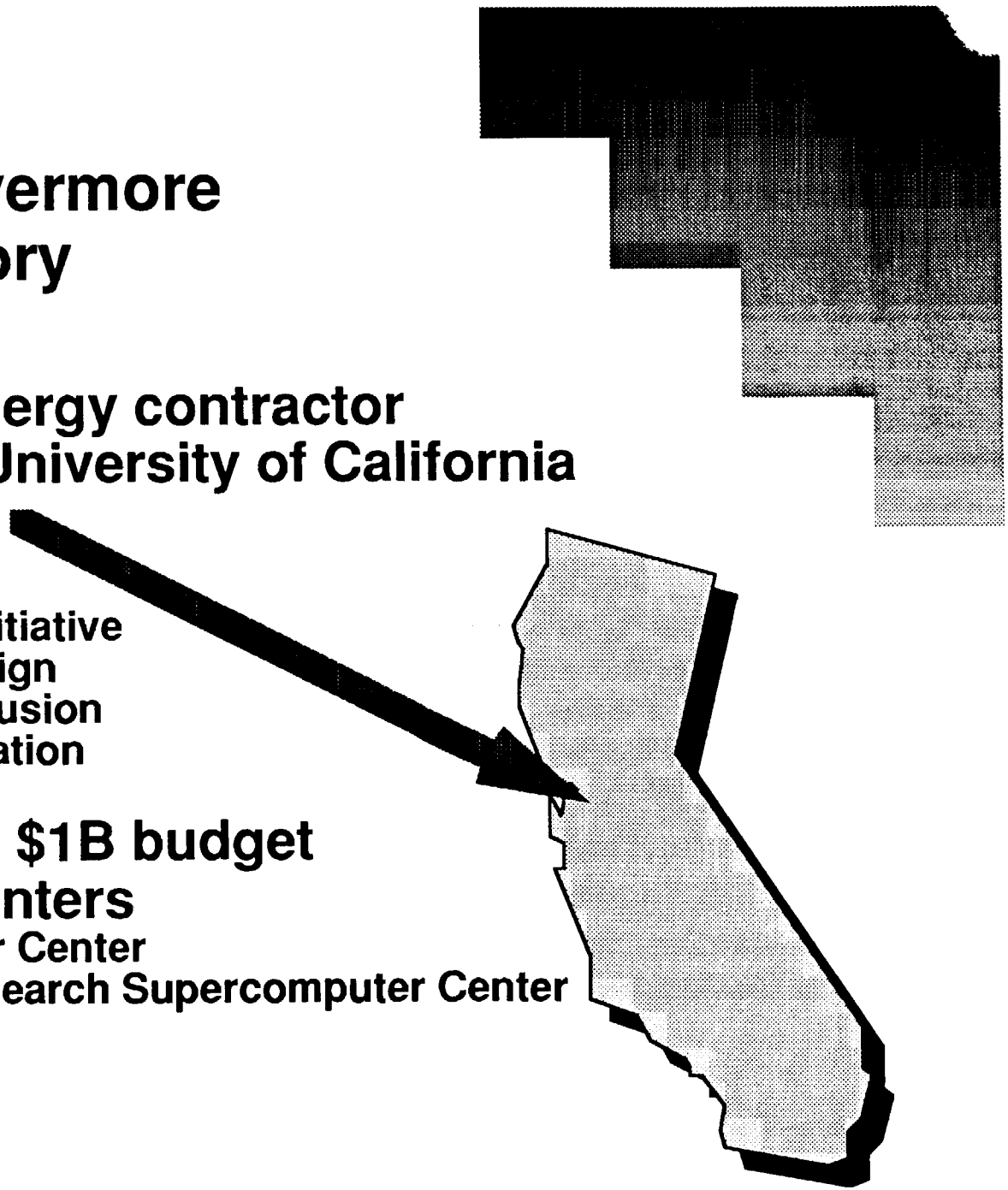
NASA Goddard

University of California

Lawrence Livermore National Laboratory

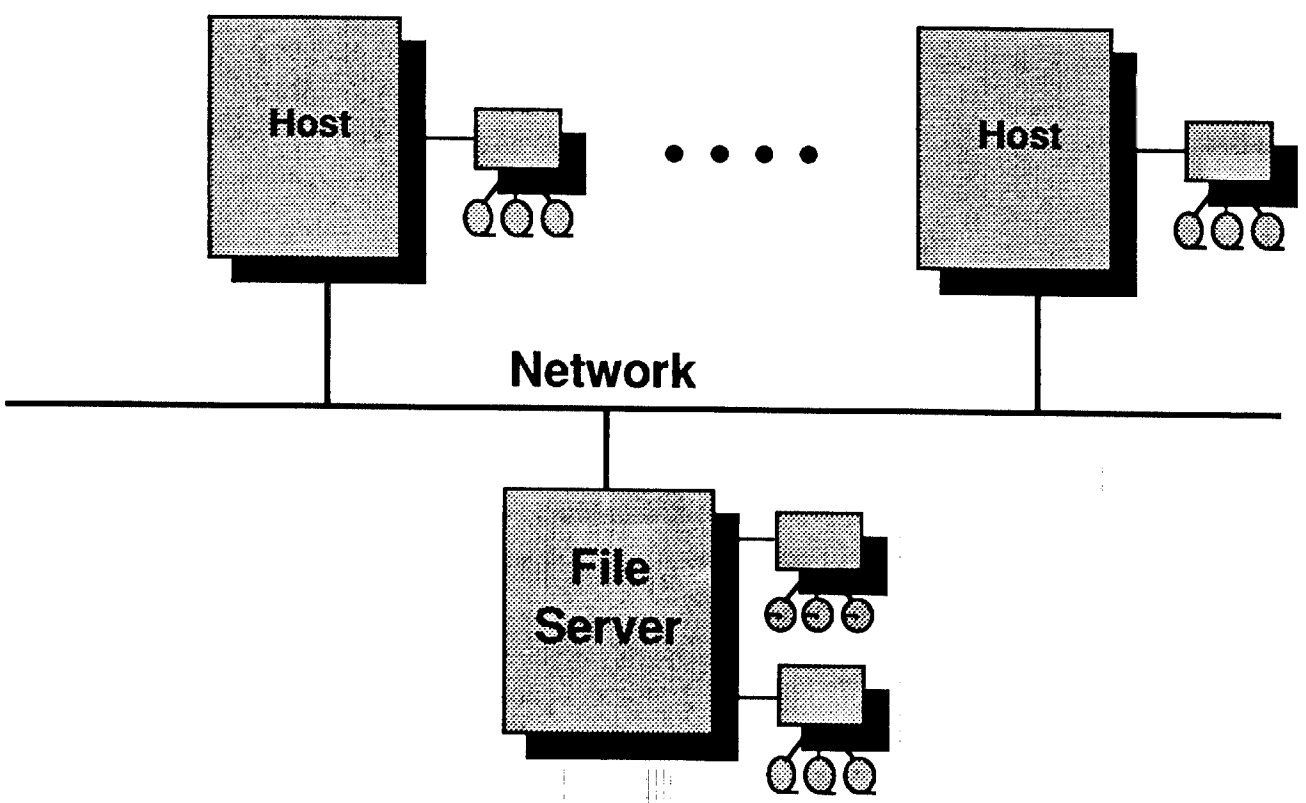
The Lawrence Livermore National Laboratory

- Department of Energy contractor
- Managed by the University of California
- Founded in 1952
- Major projects
  - Strategic Defense Initiative
  - Nuclear weapon design
  - Magnetic and laser fusion
  - Laser isotope separation
  - Weather modeling
- 8,000 employees, $1B budget
- Two computer centers
  - Livermore Computer Center
  - National Energy Research Supercomputer Center

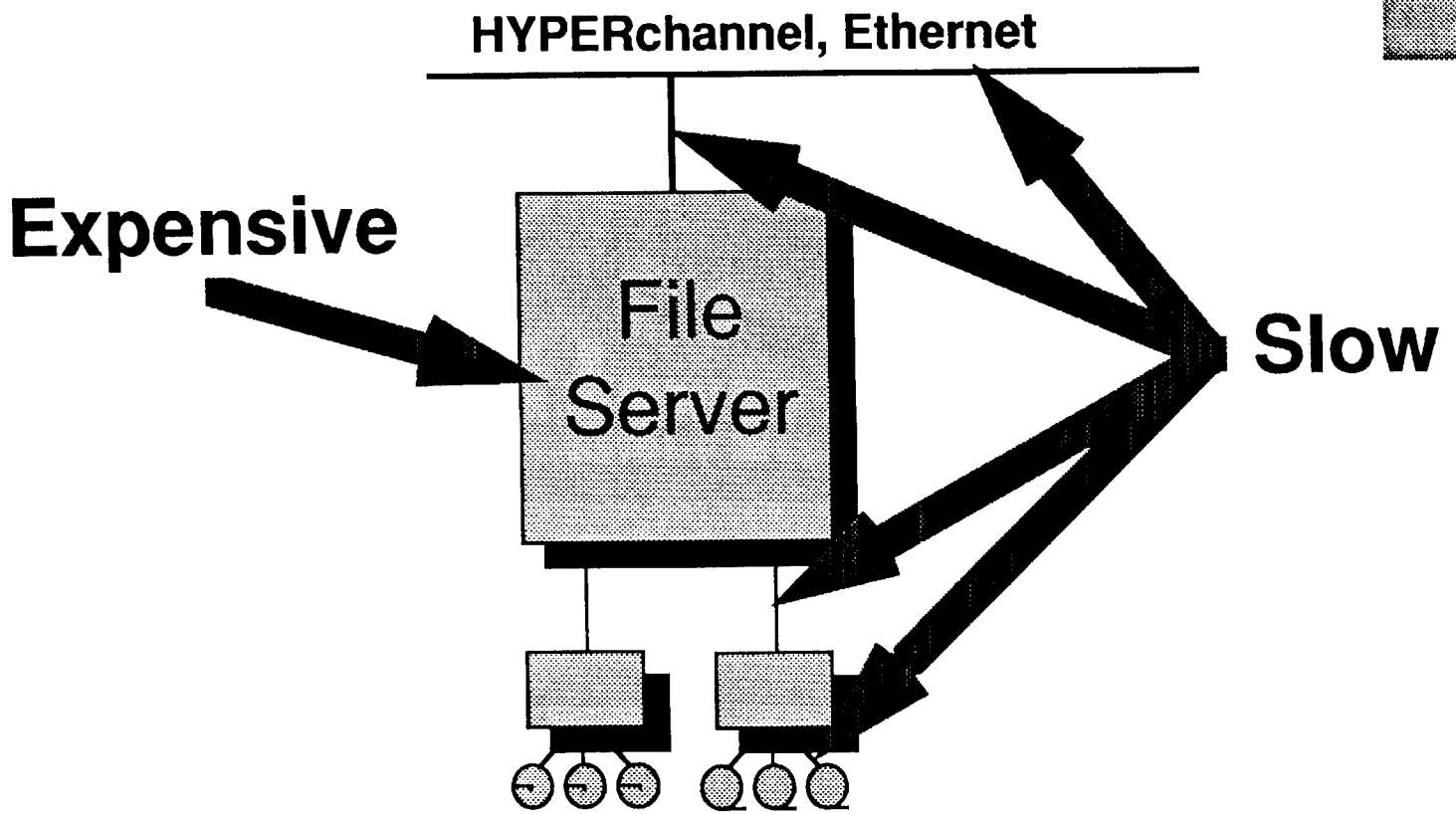


University of California
Lawrence Livermore
National Laboratory

Traditional Supercomputer Storage Architecture



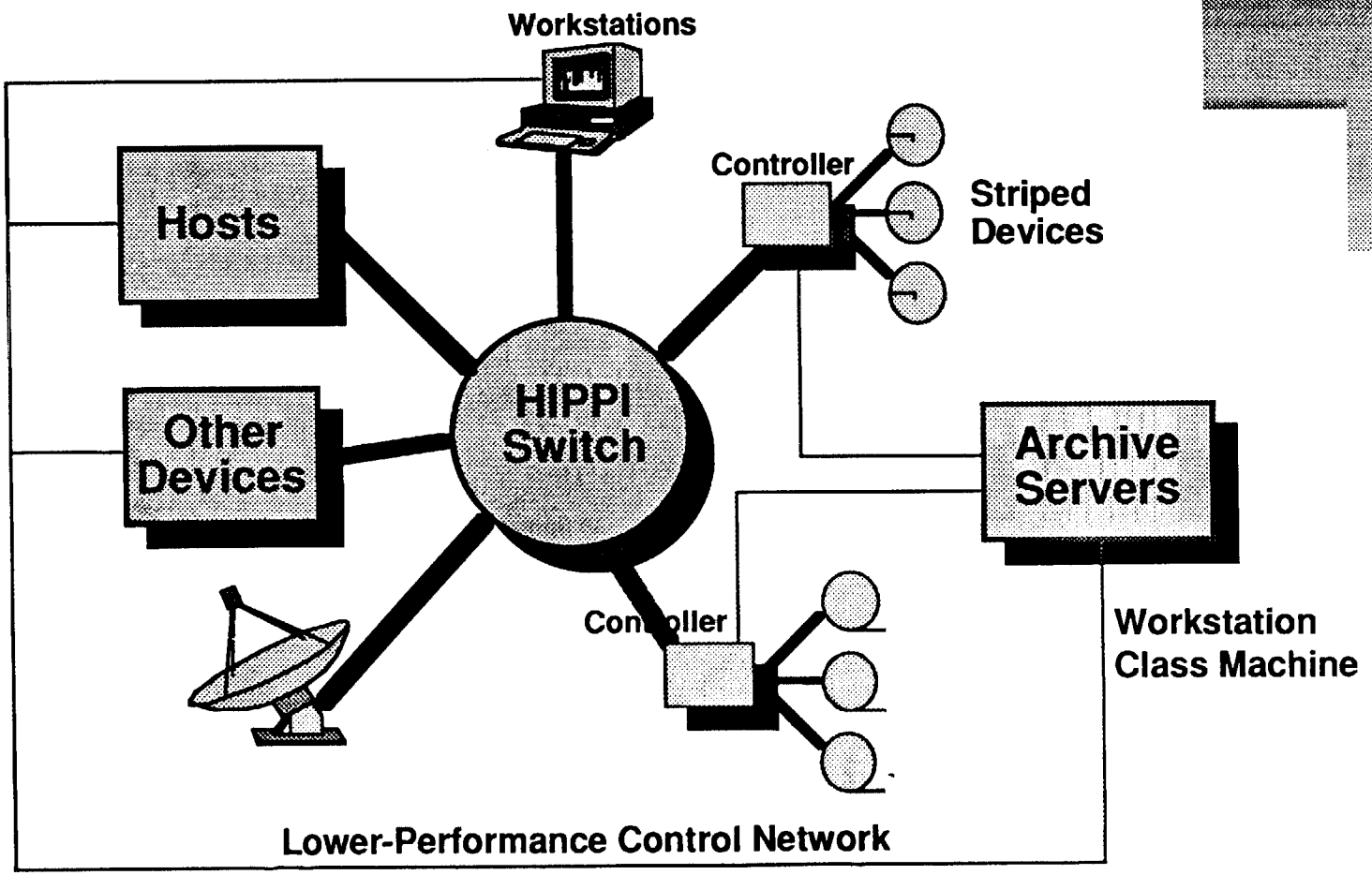
Problems with the Traditional Architecture



The Need for Higher-Performance Storage

- Rapidly increasing CPU performance
- Exploding main memory sizes
- High-performance networks
- Scientific visualization
- New applications (e.g. Mission to Planet Earth)

A High-Performance Storage Architecture

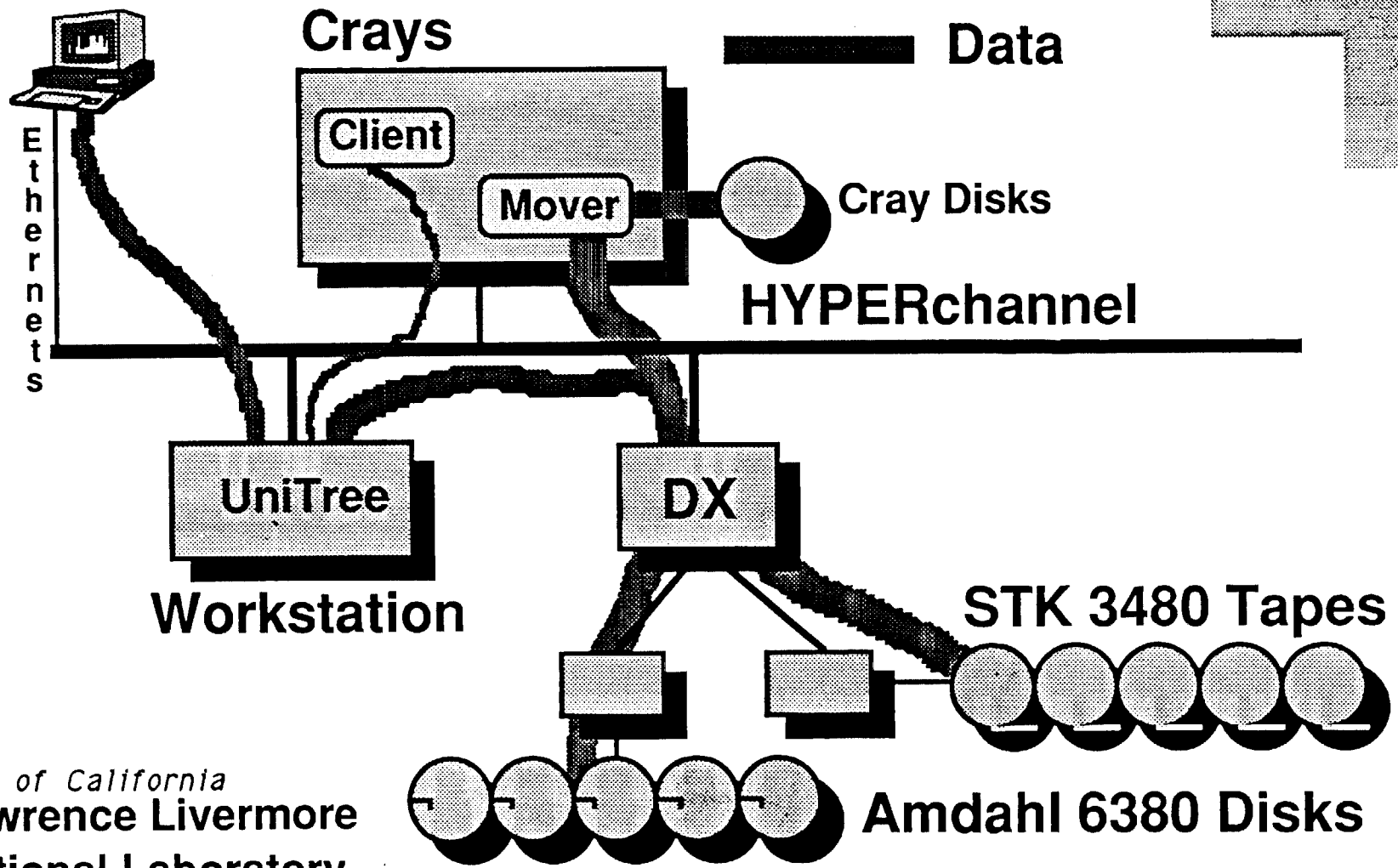


University of California
Lawrence Livermore National Laboratory

What is Needed

- Programmable device controllers
  - For protocols above IPI-3
- Striped devices (RAID)
- HIPPI-speed archival devices
  - Faster than D1, D2 tapes
  - Striped tapes?
- Higher-capacity media
- Increased reliability
- Cheaper devices, maintenance

In the Meantime.....



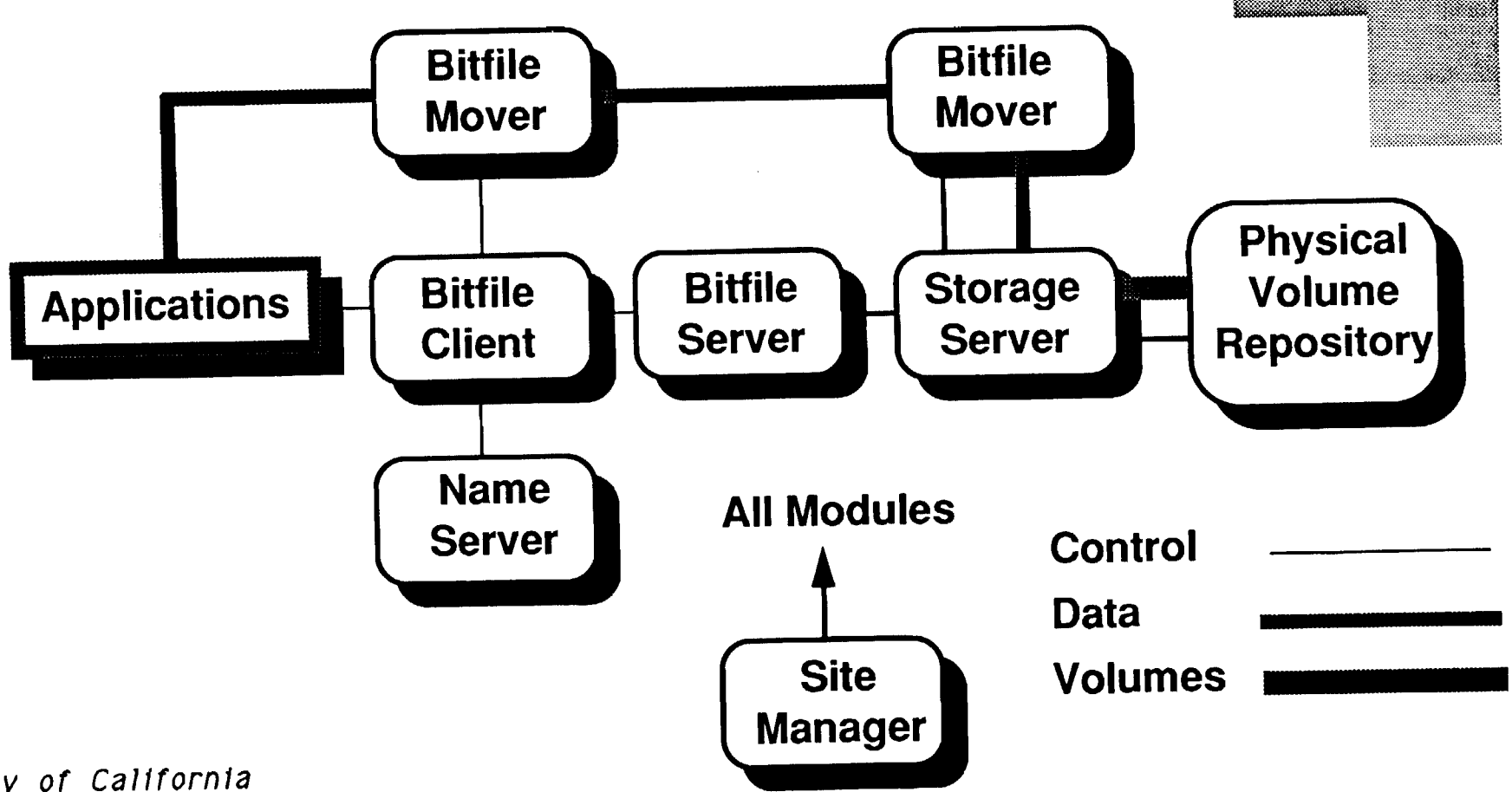
University of California
Lawrence Livermore
National Laboratory

Software Needs

- **Support for network-based devices**
  - Direct data paths
  - High performance protocols
- **Transparent, distributed systems**
  - Network-wide naming environments
  - Performance transparency
  - Device-, location-, operating system-, network-independence
- **Portable, Standard Software!**

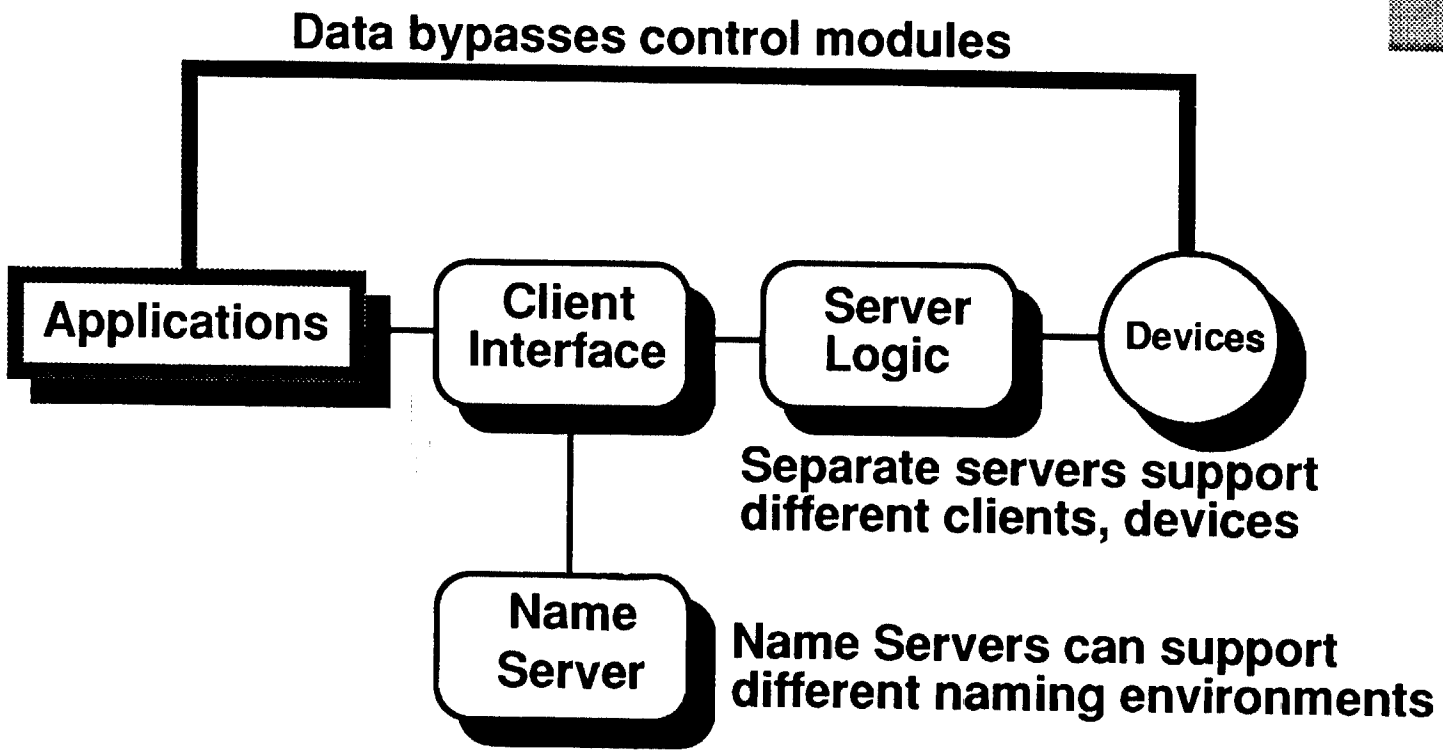
University of California
Lawrence Livermore
National Laboratory

The IEEE Mass Storage System Reference Model



University of California
Lawrence Livermore
National Laboratory

The Significant Modularity



University of California
Lawrence Livermore
National Laboratory

In the Meantime.....

- We need to go beyond FTP
  - Sun Network File System
    - Need to improve security, performance
  - Andrew File System (AFS, IFS)
    - Need to integrate with archival systems
  - File system switch (virtual file system)
    - Need to provide hierarchical, archival storage

Summary of Important Issues for Future Storage Systems

- High-performance architectures
  - Network-attached devices
- Device striping technology
- Transparent, distributed software architectures
- Software standards
- Open Systems

University of California
Lawrence Livermore
National Laboratory

To Learn More

- **Attend the 11th IEEE Mass Storage Symposium**
  - **October 7-10, 1991**
  - **Monterey Sheraton Hotel, Monterey, CA.**
  - **Arranged by**
    - **Bernie O'Lear**
    - **National Center for Atmospheric Research**
    - **P. O. Box 3000**
    - **Boulder, Colorado 80307**

Requirements for a Network Storage Service

Suzanne M. Kelly and Rena A. Haynes

Sandia National Laboratories
Albuquerque, NM

INTRODUCTION

Sandia National Laboratories provides a high performance classified computer network as a core capability in support of its mission of nuclear weapons design and engineering, physical sciences research, and energy research and development.

The network, locally known as the Internal Secure Network (ISN), was designed in 1989 and comprises multiple distributed local area networks (LANs) residing in Albuquerque, New Mexico and Livermore, California. The TCP/IP protocol suite is used for inter-node communications. Scientific workstations and mid-range computers, running UNIX-based operating systems, compose most LANs. One LAN, operated by the Sandia Corporate Computing Directorate, is a general purpose resource providing a supercomputer and a file server to the entire ISN.

The current file server on the supercomputer LAN is an implementation of the Common File System (CFS) developed by Los Alamos National Laboratory. Subsequent to the design of the ISN, Sandia reviewed its mass storage requirements and chose to enter into a competitive procurement to replace the existing file server with one more adaptable to a UNIX/TCP/IP environment.

The requirements study for the network was the starting point for the requirements study for the new file server. The file server is called the Network Storage Service (NSS) and its requirements are described in this paper. The next section gives an application or functional description of the NSS. The final section adds performance, capacity, and access constraints to the requirements.

APPLICATION DESCRIPTION

This application description section defines the functions and capabilities of the NSS. After describing the NSS perspective, NSS functions are developed from both the end-user and operations/maintenance viewpoints. NSS characteristics are also described.

NSS Perspective

The NSS shall support a hierarchy of data storage. The storage levels shall include an on-line facility and an archival facility. A back-up capability for both on-line and archival

files is also required. The on-line facility will be the primary storage system for the NSS. As files age or space limit thresholds are crossed, files will be migrated to the slower access, but denser archival facility. Access, capacity, and performance requirements for the on-line and archival facilities are given in the next section. Both on-line and archival data access must be functionally transparent to end-users; for example, if files are migrated from one facility to another, the user should not need to know the facility name to access the files. The NSS shall have the capability of ensuring that the most active user data resides on the storage level with the fastest access time appropriate for the file size.

NSS Functions

The NSS will provide data storage, retrieval, and access services to two major classes of customers. The first class represents the end-users of computing systems at Sandia. This set of customers is primarily concerned with the functionality and flexibility of services provided. End-users are also interested in the ease of use and accessibility of the services as well as the integrity of their data. Besides actual computer users, this group includes processes on other network service nodes that utilize NSS facilities.

The second class of customers represents the operations and maintenance personnel. This set of customers is primarily concerned with the reliability and performance of the NSS as well as maintainability issues. Other areas of concern for this group include accounting, security, and space management functionality. The following sections describe the functional requirements of the NSS from these two points of view.

End-User Functional Requirements

The NSS shall support a standard set of functions for user file access within a UNIX environment. User files shall be maintained in logically hierarchical directory structures, and users will be able to create and delete their own tree structures.

Many UNIX-based systems have a signed 32-bit field for calculating file offsets. Based on this constraint, the required maximum file size is at least 2.1475 gigabytes (2^{31} bytes). The supercomputer may generate files an order of magnitude or more larger than 2^{31} bytes, but current industry standards do not support directly accessing such large files. A size limit of 2^{31} x 10 bytes is desired for the NSS so that as industry standards change, the NSS will be able to support larger files.

NSS data files shall have the capability of being opened, closed, read from, or written to, from within a user program via the Network File System (NFS). If any extensions to NFS are needed, they should be limited to the NSS software, but security

constraints may require changes to NFS software on other nodes. File access from a network node will permit record-level I/O and will not require that a file be staged entirely to/from local data storage. All physical storage levels on the NSS shall be transparent to the end-user except for perhaps an initial access time, e.g., mount time associated with accessing an archived file.

File-transport-level access to user data will be provided within the context of the File Transport Protocol (FTP) supported from TCP/IP. This level of access will guarantee delivery of the entire file to the destination node's local storage area or return an error. FTP must meet security requirements as specified in the next section. While the user interface to the FTP shall be consistent on all nodes, the FTP may utilize additional protocols besides TCP/IP.

The basic philosophy of the NSS is to treat files as bit streams. Only the standard FTP and NFS data formatting features will be supported. Automatic encryption and compression routines will not be available on the NSS.

Independent of the hierarchical storage levels of on-line and archival is a back-up capability. The back-up capability will be used in two modes:

1. An operational back-up of the on-line disks will be taken to permit recovery in case of media failure. These back-up copies will remain in the Central Computing Facility.

2. Users may request specific files or subtrees of files to be backed up. The file(s) may be in the on-line or the archival facility. In a periodic (perhaps nightly) run, these files will be copied to the back-up media and subsequently sent to off-site storage. An option should allow the back-up media to be segregated by user id.

In addition to file access and back-up capabilities, the NSS will provide end-users with file management functions. These will include the capabilities of retrieving user file/directory information; setting, changing, and retrieving user file/directory access permissions, including ownership; establishing and changing some accounting information at the file or subtree level; creating, deleting, and renaming user files/directories; and copying or moving files or subdirectories to another directory in the user's hierarchy or to a directory in another user's area if permissions allow. (Note that rename and move are two distinct logical functions although both are accomplished with one command in the UNIX environment.) The capabilities of automatically maintaining file revisions (versions), marking files as undeletable or deletable, and comparing files are also desired. The file information that will be maintained and can be reported on includes file name, user id of owner, accounting information, date and time created, date and

time last accessed, date and time last modified, file type, number of links to the file, and file length in bytes. Date and time last accessed should only be updated when accessed by a non-system process. If a file is a link file, then the resolved path name will also be available. Information shall be available on a single file basis or on multiple files, for example, by using wildcard notation in the query. Wildcarding at the directory level is required. Options will be available to sort the file information output and to obtain user subtree information.

Access permissions on files should allow the owner of a file to specify read, write, or execute permissions for specific users or groups of users. Access permissions may be specified on a single file or multiple file (for example, on a subtree of files) basis. A universal, or public-type, read access mechanism is desired.

The user interface to NSS file services shall be readily available on any UNIX system that has access to the ISN. Knowledge of the network topology shall not be required for file access from the user level. Shell level commands will permit wildcard or template specification of files. Redirection capabilities are desired so that a file transfer can be initiated from one node for delivery to a process on another node, but security constraints may restrict this redirection capability.

The preceding paragraphs have dealt primarily with the NSS capability, flexibility, and ease of use requirements for end-users. The NSS shall also maintain the integrity of users' data and provide protection mechanisms to detect and prevent the corruption of data. File-level locking capabilities during write operations are required, and record-level locking mechanisms are desired. NSS availability shall be on a twenty-four hour basis 365 days a year except for periods required for system maintenance.

Operations Functional Requirements

There will be no dedicated operator attending the NSS, although there will be a centralized operations center where the NSS can be monitored. Due to this unattended nature of the NSS, any normal movement between on-line and archival storage levels shall not require operator intervention. The centralized operations center will handle exception conditions. Except for importing and exporting back-up media, the back-up process should also be automatic.

A single master file directory shall be maintained for all files known to the storage system, including files in archival storage. It is desirable for this directory to be maintained as a tree structure. The master file directory will be journaled and archived, and utilities will be available to recover the master file directory in case of data corruption or catastrophic failure. Operations personnel will be able, in a privileged

mode, to access, list, and modify the contents of the master file directory.

Capabilities for backing up and restoring the on-line storage, as well as specific user files or subdirectories, will be provided. Utilities to print or dump individual data files will also be available. The NSS must allow other network nodes to be able to mount at the NSS directory level as well as at the file system level.

The NSS file management software shall maintain a hierarchy of storage levels where files are stored based on file size and frequency of use. Space management and file migration utilities shall be available for discretionary use by operations personnel. Automatic aging and migration capabilities with configurable parameters that operators can modify shall be provided as well as tools for monitoring storage and access performance.

Since reliability is a key concern for operations, the NSS shall include external and internal redundancy features. The NSS must continue to operate in a degraded mode upon a single subsystem failure such as an operator console or a disk controller.

Performance tools shall also be provided to allow operations staff to monitor and obtain statistics on file accesses, effective transfer rates, and media access.

To support integrity as well as reliability, the NSS shall maintain an audit trail of file operations, permit access to files only after access rights are verified, and maintain recovery files that permit recovery of the master file directory without loss of data.

Any recovered or unrecovered hardware or software errors shall be recorded in an error log along with the date and time of occurrence and additional diagnostic information. Utilities shall be available to analyze error logs and produce reports for operational staff.

Since the NSS will operate within the secured SNL environment, security features described in the next section will be supported. Operations personnel will be able to obtain logs of security events, and alarm mechanisms will be activated if security violations are detected.

Support for accounting functions is required to enable operations staff to charge customers for NSS usage. This support will include utilities to determine space/time utilization for files at each storage level as well as networking or data channel usage. Accounting algorithms will be modifiable and different storage levels will be charged according to the cost for storage /retrieval services. The ability to easily obtain and modify accounting information for files, subdirectories, and files within subdirectories is required.

The evolutionary nature of the ISN requires software maintenance staff to have the ability to maintain and extend the software initially provided in the NSS. The software development environment shall allow interactive development of site specific utilities and extensions. Documentation provided shall include high level design documents, detailed internals documentation of the operating system and the file management software, and operations reference manuals. User reference manuals are also required to be available. Source code for the file management software shall be provided in machine readable form to Sandia software maintenance personnel. Source code for the NSS operating system is desired, but is required if security features or the file management software require hooks or modification to the operating system. In addition to the source code, any compilers, assemblers, or loaders needed to convert the source code into machine executable code is required. The associated "build macros" or equivalent are to be supplied. Software support tools including high-level language compilers, editors, debuggers, and computer assisted software engineering tools shall also be provided.

NSS Characteristics

This section describes the characteristics of the NSS, including the hardware and system software environments.

The hardware environment shall include sufficient processing, memory, and support peripherals/subsystems to generate and run the normal operating system in addition to the file management/control system and any performance or diagnostic programs. Hardware expansion to support at least 300 gigabytes of on-line and at least 5 terabytes of archival storage shall be field upgradable. Expansion recommendations will be defined in terms of cost per increment, time for installation, impact on installed hardware and software, and procedures for installation.

Memory shall be field upgradable to double its initial capacity. All memory components, initial and expanded, shall be protected by at least single bit error correction and double bit error detection and reporting.

Support peripherals shall include a tape unit, a printer, two operator consoles, and at least 3 system programmer terminals. The tape unit shall support 1600 and 6250 bpi, and accommodate 2400 foot reels.

The operator console will be the primary operational interface to the system. A hardcopy capability is desired for all operator input from and output to this console. Hardware to support a remote operator capability is required to provide status information and limited operator commands to operations personnel supporting the NSS. The NSS hardware environment will include all interface hardware to support communications to the ISN, as

well as any equipment that may be needed to support an optional direct link to the supercomputer.

The system software environment for the NSS will include all software necessary for operating, controlling, modifying, and maintaining NSS functionality. It is desirable that the operating system be UNIX-based. In any case, the NSS shall be compatible with UNIX network nodes and shall support POSIX (IEEE 1003.1-1988) file operations from these nodes.

SYSTEM DESIGN CONSTRAINTS

This section discusses significant factors that bound, or constrain, the design of the NSS. Areas that bound the design include the network interface, performance, capacity, and security.

Network Interface

The NSS is to be located on the supercomputer LAN portion of the ISN. A Network Systems Corporation (NSC) DX-technology HYPERchannel is the network backbone for the supercomputer LAN. IP routers from NSC and cisco Systems, Inc., provide connectivity to other distributed LANs. The NSS must support connections to at least two NSC "N" series adapters. These adapters will be used for communicating with other nodes on the ISN. One exception is that a dedicated link may be required between the NSS and the supercomputer. The following subsection specifies performance requirements for file transfers between the NSS and the supercomputer which may not be achievable using the HYPERchannel backbone and its TCP/IP protocols.

Performance

Performance values will differ depending on the characteristics of the network node accessing the NSS. In the case of the supercomputer, it is assumed that the NSS is the limiting partner, and, therefore, the performance values are required goals for the NSS. For other nodes, the particular node architecture or the network topology are assumed to be the limiting factors. Therefore, the NSS is to support the stated performance requirements for the non-supercomputer nodes but the actual performance may vary.

File transfer performance figures are specified assuming the file resides in or is destined for the on-line storage facility. The particular implementation of the archival system will drive its file transfer performance figures. For example, a possible implementation of the archival system may require that a file be staged onto on-line storage before it can be transferred. If this is the case, the performance figures from archival storage are the sum of the on-line performance figures plus the transfer rate between the on-line and archival storage.

Record-level access is to be provided by NFS. Performance figures are assumed to be more driven by the design of NFS rather than NSS I/O channel speeds, for example. Thus, it is believed to be impossible to dictate a meaningful record-level performance figure, as it may be incompatible with NFS design constraints.

File-level access can be grouped in two classes: 1) transfers to/from the supercomputer, and 2) transfers to/from non-supercomputer nodes. Performance requirements for each of these groups are specified.

Supercomputer file-level transfers will include transfers of large files (greater than 100 megabits) and other files (less than 100 megabits). For files containing less than 100 megabits, the transfer must complete in at most 3 seconds. For files greater than 100 megabits in size, the user-perceived, disk-to-disk, transfer rate of ONE file must be 50 megabits per second, which may require the dedicated path mentioned previously.

File transfers between the NSS and non-supercomputer nodes will use FTP/TCP/IP. The NSS must support individual file transfer rates (user-perceived) of at least 10 megabits per second, which is consistent with the communication bandwidth of the distributed LANs.

As mentioned previously, the user-perceived file transfer rates are not specified for archived files. However, it is mandatory that the effective transfer rate between one on-line device and one archival device be at least 10 megabits per second. To provide for future archival technologies, higher I/O bandwidth, up to 200 megabits per second is desired.

Other transactions to the NSS, such as ls (list), must be responded to within two seconds. Like record-level access discussed previously, this performance will be probably be constrained by software design issues rather than channel speeds, etc. However, this requirement will probably require that file administrative information be stored on on-line disk rather than archival media.

Capacity

The determination of on-line capacity requirements is based on a performance objective. For the current file server, this objective has been to maintain a 95% on-line media "hit" rate on file retrievals. This objective has been consistently met by storing files on on-line disks for 30-120 days. Thirty days is used for the smallest files and 120 days is used for the largest files. After 30-120 days of inactivity, a file is moved to archival media.

The same objective of a 95% on-line media "hit" rate for retrievals will be used for the NSS. The current system manages this performance with 90 gigabytes of formatted on-line capacity.

The NSS will be configured with 100 gigabytes of on-line storage initially and additional storage will be added as needs dictate and funding is available.

Archival capacity requirements were also estimated based on the characteristics of the current file server. The ratio of archived data to on-line data has been 10 to 1. Therefore, the NSS will initially be configured with 1 terabyte of archival storage.

Back-up media must be removable. Therefore, the total back-up capacity is theoretically unlimited. However, each storage unit, e.g., tape, must hold a minimum of 150 megabytes. In order to back up the largest files, multiple volumes must be supported.

Security

The NSS must enforce the following security rules.

1. Users must be authenticated before accessing the system.

2. All processes will have a classification level associated with them.

3. Processes may not access data for which they are not authorized.

4. Processes may not read data that is at a higher classification.

5. Processes may not write data that is at a lower or higher classification.

6. Access to classified or sensitive data must be audited. Audit information must include user id, type of access, date and time of access, and file name. Both authorized and unauthorized accesses to classified or sensitive data must be recorded.

7. Access to classified data requires two independent levels of controls. One of these controls can be the user logon password. The second type of control can be a file access key, access control list, or equivalent mechanism. File access keys must be protected at the classification level to which they permit access.

8. Hardware protection mechanisms must prevent processes from accessing physical memory locations outside of their program images.

9. Software mechanisms must assure that left-over data on magnetic media cannot be retrieved by an ordinary process.

10. All devices and media, e.g., tapes or printed material, that contain classified data must be labeled. Approved procedures for removing data must be followed before declassifying or removing any storage device/media that contains classified data.

The NSS must provide an access control capability for every file in the system. This includes files residing on the on-line or archival facilities as well as files placed on back-up media.

The NSS must distinguish between multiple (at least four) classification levels of data. The capability of maintaining a category of information for all files on the NSS is also desired. This capability should allow several categories (from 4 to 64) for each classification level.

The NSS must be able to identify the processing level of a user making a file system request. Process classification levels are identical to those defined for data classification levels. The processing level will be used as well as the access control capability to decide whether to grant or deny access to data stored on on-line, archival, or back-up media. The policy to be used for granting access to data is as follows:

1. Users may only access data to which they have been permitted.

2. Even if a user has permission to access a file, read or execute access will be granted only if the user's classification level is greater than or equal to the classification level of the data.

3. Even if a user has permission to access a file, write, update, or delete access will be granted only if the user's classification level is equal to the classification level of the data.

These policy rules will apply to file attribute accesses as well as to file data accesses.

An audit trail of all NSS activity, except for successful ls commands, is required. This activity includes file management requests, privileged mode accesses or logons, and logons from system programmer terminals. Log entries must contain at a minimum:

date and time of request,
type of request,
user id,
requesting node,
requesting process classification level,
file or directory name, and
file or directory classification level.

Protection mechanisms must be available to protect privileged access to the NSS. In general, the NSS will not run user codes. Privileged operator commands should require a special access mechanism, e.g., logon or password, before executing. Privileged access to the NSS will be limited to a minimum number of Sandia personnel required for NSS operations.

Since the NSS will be a node on a DOE accredited packet-switched network, security features will be available in the networking software. The security features will include the capability of determining a packet security classification level. This information must be checked with NSS request classification level, and any detected violation of access security policy will be logged and a security alarm will be activated. Additionally, some application level utilities, e.g., telnet, may not be available on the NSS.

A security alarm feature is required to identify and highlight actions or requests that could be penetration attempts, i.e., security events. An example of a security alarm is to send a highlighted, unscrollable message to an operator console when a security event is detected. Security events must be logged in the audit trail as well as activating the security alarm. Security events that will activate the security alarm will include violations of Sandia data security policy, violations of privileged access policy, and violations of network security policy.

ACKNOWLEDGEMENT

This work was supported by the United States Department of Energy under contract DE-AC04-76DP00789.

TRADEMARKS

UNIX is a registered trademark of AT&T.

NFS is a trademark of Sun Microsystems, Inc.

HYPERchannel is a trademark of Network Systems Corporation (NSC).