NASA Conference Publication 3198, Vol. II

Goddard Conference on Mass Storage Systems and Technologies

Volume II

Proceedings of a conference held at
NASA Goddard Space Flight Center
Greenbelt, Maryland
September 22–24, 1992

NASA

NASA Conference Publication 3198, Vol. II

Goddard Conference on Mass Storage Systems and Technologies

Volume II

Edited by
Ben Kobler
Goddard Space Flight Center
Greenbelt, Maryland

P. C. Hariharan
STX Corporation
Lanham, Maryland

Proceedings of a conference held at
NASA Goddard Space Flight Center
Greenbelt, Maryland
September 22–24, 1992

National Aeronautics and
Space Administration
Office of Management
Scientific and Technical
Information Program
1993

Goddard Conference on Mass Storage Systems and Technologies

Program Committee

Ben Kobler, NASA/GSFC (Chair)
John Berbert, NASA/GSFC
William A Callicott, NOAA/NESDIS
Sam Coleman, Lawrence Livermore National Laboratories
Susan Hauser, National Library of Medicine
Sanjay Ranade, Infotech SA, Inc
Elizabeth Williams, Supercomputing Research Center
Jean-Jacques Bedet, Hughes STX
Alan Dwyer, Hughes STX
P C Hariharan, Hughes STX

Conference Coordinator

Nicki Fritz, Hughes STX

Production and Layout

Len Blasso, Hughes STX
Ann M. Lipscomb, Hughes STX

6-21-93

Donation

PREFACE

Papers presented at the Goddard Conference on Mass Storage Systems and Technologies that were submitted for publication in advance of the Conference appear in volume 1 of these Proceedings. Volume 2 contains additional papers and view graphs which were made available at the time of the Conference, as well as reports of the keynote address, the after-dinner speech, and the two panel discussions. We are grateful to all the authors for their contributions.

Dr. David Nelson, Director of the Office of Scientific Computing, Department of Energy, opened the conference with a keynote address that began by identifying projects and activities that are, or will be, generating massive volumes of data. Some of the grand challenge problems of the High Performance Computing and Communications initiative are likely to rival, or even surpass, the Earth Observing System in the amount of data they create. Managing such large archives is itself likely to prove a grand challenge. He referred to inaccessible data as the *"landfill of cyberspace."* Learning to answer unanticipated questions, revising data structures as requirements evolve, doing this in a cost-effective and practical manner in a hierarchical storage system, and dealing with distributed data bases that are networked together will tax both human ingenuity and resources.

Mass storage systems have now truly begun to be massive, with data ingestion rates approaching terabytes per day. At the same time, the identifiable unit for processing purposes (file, granule, dataset or some similar object), has also increased in size, and could begin to pose a challenge to traditional file systems that impose limits on both the size of the objects, and the number of objects in the file system. Even the casual user needs more than the object name, the size and date of the creation of the object, and the limited metadata provided with classical directory systems. Some of these issues are addressed by the IEEE Mass Storage System Reference Model (MSS RM), which is seeking to provide a framework in which hardware and software from different vendors can act cooperatively and harmoniously to store, manage and distribute data. Dr. Sam Coleman of the Lawrence Livermore National Laboratory and Mr. Bob Coyne of the IBM Federal Sector Division discussed the history and current status of the Reference Model. Version 5 of the MSS RM will appear in April 1993 as a Recommended Practice instead of as a Guide. The emphasis of the Storage Systems Standards Working Group (SSS WG) is focused on decomposing storage systems into interoperable functional modules which vendors may offer as separate products, and on defining standard interfaces through which clients may be provided direct access to storage systems services. *Bob Coyne pointed out that the data management, database, and file system development and user communities are not represented in the SSS WG, and issued a plea for their active participation in the activities and deliberations of the WG.* Those interested in the SSS WG discussions may keep abreast by sending e-mail to **ieee-mss-request@nas.nasa.gov** with the request that their name and address be included in the WG reflector. General discussions on mass storage problems are also published in the USENET newsgroup **comp.arch.storage**.

Standards are essential to ensure wide availability, multi-sourcing, and interchangeability. Mr. Al Dwyer, representing the NASA-OSSA Office of Standards and Technology, spoke about the role of this office. He was followed by Mr. Jean-Paul Emard, ANSI X3 Committee Director, Mr. Sam Cheatham of the X3B5 Committee, and Mr. Ken Hallam of the X3B11 Committee who discussed the ANSI standards-making process, the work on magnetic media standards, and the status of the optical media standards, respectively.

The sheer size of the inventories makes distributed systems attractive. Bob Coyne discussed the National Storage Asset Laboratory at the National Energy Research Supercomputer Center of the Department of Energy; this will be a testbed for network-attached storage devices. In this configuration, the devices will be nodes in a network, and will provide read/write services to authorized clients on the network without the need for the data to pass through the memory of a computer controlling the devices. Experiences from the archives at NOAA, the National

Space Science Data Center at NASA, the Eros Data Center of the USGS[1] , and the National Library of Medicine were complemented by a discussion of the information management challenge posed by the Earth Observing System. Dr. Ackerman of the National Library of Medicine pointed out that while there is much discussion of gigabit networks and petabyte-sized inventories, there are still problems today in distributing much smaller files to a user community not fortunate enough to be plugged into the latest wideband network. Browsing is a significant component of the activity at large holdings, and Dr. Ken Salem described one way to handle this.

High volume holdings require high-performance storage devices. The idea of using a Redundant Array of Inexpensive Disks to provide increased bandwidth and reliability had previously been espoused by Garth Gibson, and others, and Dr. Gibson provided a simplified explanation of it in his tutorial lecture. A natural outgrowth of the RAID idea is that of RATS (Redundant Array of Tape Systems), and Ms. Ann Drapeau of the University of California at Berkeley took up this topic in her tutorial.

Professor Mark Kryder, Director of the Engineering Research Center in Data Storage Systems at Carnegie Mellon University, Pittsburgh, PA discussed the future evolution of magnetic and magneto-optic storage systems in his talk on ultra-high density recording technologies. In cooperation with the National Storage Industry Consortium, the Center has selected the goals of achieving 10 Gbit/in^2 recording density in magnetic and magneto-optic disk recording, and 1 Tbit/in^3 in magnetic tape recording.

The National Media Laboratory (NML) has been in existence since 1989, and Dr. Gary Ashton provided an overview of its structure, scope and mission and reported on NML testing results of D-1 cassettes. A different perspective, that of the system integrator, was furnished by Mr. Richard Lee in his talk on grand challenges in mass storage.

Recent magnetic and optical recording technologies were described in a number of papers. Optical recording, traditionally available on disks, is now possible on tape. ICI Imagedata, which has pioneered the concept of the digital paper, and subjected its product to one of the largest suite of tests, now has competition from the Dow Chemical Company and from Eastman Kodak. While optical storage has generally been understood to involve ablation (pit-forming), phase change, or alloy formation (respectively the modes of the ICI, Eastman Kodak and the Dow products), Optex has a medium that uses a different technique for optical data storage. This involves excitation of electrons, and trapping the excited electrons in metastable states on a receptor ion. The method is interesting and intriguing because, unlike other technologies, it exhibits a linear response and can therefore store more than just one bit per "cell." A panel discussion on the comparative merits of magnetic and optical storage, and their future, followed these papers.

Dr. Dennis Speliotis, a veteran in the field of magnetic storage, was the after-dinner speaker at the Conference Banquet. He reminisced about his experiences over more than three decades in magnetic storage and related stories of both success and failure. His parting words were significant: the way to make progress is through evolution, not revolution; the chances of failure when one attempts a dramatic change, a drastic departure from the conventional, are very high, certainly in the short term; but small, evolutionary, step-changes are more likely to succeed.

Mr. Dale Lancaster of Convex Systems presented what the "state of the art" is in Mass Storage Technology. Drs. Elizabeth Williams and Tom Myers discussed the need for, and the nature of, the types of measurements and metrics of distributed and heterogeneous storage systems. Measurements were reported by Ms. Nancy Yeager of the National Center for Supercomputing Applications. Mr. Bill Collins of the Los Alamos National Laboratory presented an overview of the High Performance Data System being developed there and Dr. Milt Halem, from the NASA

---

1 Although John Boyd was unable to present his paper "Interim Report on Landsat National Archive Activities," it is nevertheless included in these proceedings

Goddard Space Flight Center gave a critical and comparative analysis of three application-dependent mass storage systems being built at Goddard.

Mr. James F. Berry, of the Department of Defense, chaired a panel discussion on High Performance Helical Scan Recording Systems. Representatives from Ampex, Datatape, GE, Sony and StorageTek were the participants.

The performance of the low-end helical scan tape drives was the topic of papers by Dr. Chinnaswamy, formerly of Digital Equipment Corporation, and by Mr. Gerry Schadegg of Exabyte Corporation. Exabyte now provides an on-line Technical Support Bulletin Board System (BBS). Banana Boat, as the BBS is called, can be accessed by dialing (303) 442-4323. The BBS contains information such as microcode history, technical bulletins, white papers, and articles of interest to 8 mm product users. *Mr. Schadegg advised users of 8 mm drives that those drives were not designed for 100% duty cycle, but only for 20% to 30%.* He also cautioned users that the small, handy size of the cassette should not lull them into thinking that the media does not require a controlled environment for storage, shipping and operation. Finally, tips on reducing file read latencies were discussed by Mr. R. Hugo Patterson of Carnegie Mellon University.

A number of posters were presented on the first day of the conference.

Our thanks go, in addition to the authors, to the following persons and organizations:

Dr. David Nelson, Department of Energy, the keynote speaker,
Dr. Dennis Speliotis, the after-dinner speaker,

the following session and panel discussion chairs:

Dr. Joe King, NASA/GSFC,
Dr. Mark Kryder, Carnegie Mellon University,
Dr. Milt Halem, NASA/GSFC,
Mr. James F Berry, Department of Defense,

the following members of the program committee:

Mr. Jean-Jacques Bedet, Hughes STX Corporation,
Mr. Bill Callicott, NOAA,
Dr. Sam Coleman, Lawrence Livermore National Laboratory,
Mr. Alan M. Dwyer, Hughes STX Corporation,
Dr. Susan Hauser, National Library of Medicine,
Dr. Sanjay Ranade, Infotech SA, Inc.,
Dr. Elizabeth Williams, Supercomputing Research Center,

and to:

Ms. Nicki Fritz, the conference coordinator,
Westover Consulting for conference arrangements,

and Mr. Len Blasso and Ms. Ann Lipscomb for their help with the production of this document.

We are grateful to Mr. Laurence Lueck, President of Magnetic Media Information Services, for permission to reproduce the David-and-Goliath cover art from Volume XIII, Number 1, of the *Magnetic Media International Newsletter.*

Ben Kobler, NASA/GSFC
John Berbert, NASA/GSFC
P C Hariharan, Hughes STX Corporation

TABLE OF CONTENTS

Volume I

TABLE OF CONTENTS (Continued)

Volume I (Continued)

TABLE OF CONTENTS

Volume II

TABLE OF CONTENTS (Continued)

Volume II (Continued)

Keynote Address

David Nelson
ER-7
Department of Energy
Washington, DC 20585

Practitioners of data storage are analogous to the Sisyphus of the computer world, said Dr. David Nelson, Director of Scientific Computing, Office of Energy Research, Department of Energy, in his keynote address to the Goddard Conference on Mass Storage Systems and Technologies on September 22, 1992, at NASA's Goddard Space Flight Center, Greenbelt, Maryland.

Instead of interminably rolling a boulder uphill, however, they rolled gigabytes up to the top. When they reached the summit, the data users said, "Gigabytes? We want terabytes." And they started over again. Now that they have terabytes rolled up to the top, the users say, "Terabytes? We said petabytes."

Nelson pointed to the theme of his presentation: that to do computational science and engineering, practitioners must deal with data management activities as much as anything else.

Addressing "the relationship of grand challenges and data management," he defined a grand challenge as:

> *a fundamental problem in science or engineering with broad economic or scientific impact whose solution could be advanced by applying high-performance computing techniques and resources.*

An interagency study posed the following question: What are some areas where high-performance computing really could make a difference? It yielded as examples the following grand challenges:

weather, climate, and global change
ocean sciences
atomic nature of materials
semiconductor materials and devices
superconductivity
nuclear fusion
oil and gas prospecting and recovery
efficient, clean combustion
vehicle design
vehicle signature
computer vision
speech recognition
elementary particle physics
astronomy
structural biology
human genome

Nelson displayed a recent simulation of a global ocean circulation model with a resolution of half a degree in latitude and longitude and 20 vertical levels. The strongest flows -- showing the scalar speed field -- were differentiated from the slowest by color. A number of strong flows, such as the Gulf Stream, were evident off the East coast of the United States.

Because these strong flows tend to be very narrowly defined, Nelson said, and most of the ocean is rather quiescent, a simulation to pick up these flows must use a very fine grid. Half a degree in latitude and longitude just begins to capture the mezoscale eddies.

The simulation, at the state of the art in understanding ocean circulation via computer simulation, was done on a CM-200, he pointed out. Achieving that resolution took one of the world's most powerful computers, large scale data storage and analysis, as well as some interesting mathematical improvements. "So data and simulation, mathematics and computer science all go hand in hand," he said.

Using a 3-D toroidal gyrokinetic simulation, Nelson discussed another grand challenge, this one from fusion energy, using a tokamak to simulate an ion temperature gradient instability. He described a tokamak as a doughnut-shaped confinement device with a field running around the doughnut.

In the early linear phase of the instability simulation, the perturbations are concentrated on the outside of the doughnut, while things are relatively quiescent on the inside. That is not how the real tokamak behaves, however, according to experimental data.

The scientists asked themselves whether they just did not understand the physics of the instability, or whether the problem was a limitation of their computations. More recent simulations done by massively parallel computers depict the perturbation as equally strong inside and outside during the non-linear phase, suggesting that computational power, not the understanding of physics, was the limiting factor.

Nelson stressed the significance of the ability to extract and to visualize data in the simulation as an aid to understanding. That significance cannot be understated, he said.

"If one simply looked at tables of numbers and tried to get a qualitative picture of what was happening to the transport of energy, it would be extremely tedious. Whereas by looking at a picture of this sort and seeing a filament move from the outside to the inside and back to the outside and knowing that along that filament, particles and energy are easily transported, one gets an intuitive understanding of what's going on," he emphasized. Both the simulation and the data visualization require large-scale, fast data storage systems.

Nelson then discussed some of the problems associated with another grand challenge, achieving efficient, clean combustion. The problem, he said, is that combustion is a turbulent, reacting process.

"The droplets of fuel swirl at 90 mph, change shape, evaporate, burn at their surface, form particulates. The equations that govern that process are what the mathematicians call extremely stiff," he pointed out.

By stiff one means that, in chemical kinetics, things are happening in the nanosecond or even sub-nanosecond time scale, whereas the fluid motion is more in the scale of microseconds or even milliseconds, he said. As a result, sophisticated computational techniques have to be employed that require large temporary data storage during the simulation.

"To capture even the hints of turbulence, the spatial resolution has to be well down in the sub-millimeter range," he said, and the engine involved has a cross-section of a number of centimeters.

Keeping track of all the processes involved in this problem in multispecies chemical kinetics in a turbulent flow is a very difficult computational process, he stressed, and generates huge amounts of data.

Displaying a picture of the collider detector at Fermilab, Nelson then described some of the challenges being confronted in high- energy physics. The Tevatron accelerator facility pictured is currently being used to look for the top quark, the last of the quark family. The collider detector, he explained, captures the results of collisions of particles, in this case, protons and antiprotons. The detector almost completely surrounds the collision volume to capture all resulting particles.

The volume of data being generated by these detectors is so great and is being produced so quickly, however, that the physicists have decided they simply cannot keep it all, Nelson said. Thus they are using a two- or three-stage gating process.

A signature of interesting data is looked for in real time on the fly. If the data looks uninteresting, it's thrown away. If it looks interesting, it's subjected to a second level of scrutiny and, if it passes certain tests, it is then archived and kept.for later analysis The physicists realize that they may be throwing away important data if the criteria used in the gating process are insufficiently "smart", but they believe that the cost savings justify the risk.

Nelson next turned his attention to structural biology and the catalysis of dihydrofolate to tetrahydrofolate, a process that is essential to the formation of proteins and nucleic acids. Detailed knowledge of this key to human metabolism and existence could help lead to the scientific ability to control the reaction and thus fight cancer.

An enzyme called dihydrofolate reductase catalyzes dihydrofolate to tetrahydrofolate, Nelson said. He pointed out the simulation of the dihydrofolate reductase molecule, shown as a complicated folded ribbon. The structure of this molecule is known primarily from X-ray crystallography.

"It is too large a molecule for us to be able to currently simulate it at the atomic level, to capture the electrostatic fields from each of the electrons. It is much too difficult for us to be able to fold the atoms into the molecule from first principles," said Nelson.

"For this calculation the knowledge of dihydrofolate reductase itself is extracted from the X-ray crystallography. The molecular structure is determined by undoing the diffraction pattern. Because this is a large molecule, the data and computational requirements for doing even that "simple" calculation challenge modern computers," he said. Knowing the structure of dihydrofolate reductase, scientists compute the resulting electric fields at the enzyme's "active site", and apply these fields to the much smaller dihydrofolate molecule to determine how this molecule is able to react to produce tetrahydrofolate.

This particular calculation, he emphasized, is at the state of the art in catalysis simulation. It has elucidated the mechanism whereby the electron density in the dihydrofolate is changed to make the reaction energetically favorable.

The hope now is to be able to use this understanding of how dihydrofolate reductase causes dihydrofolate to react, either to produce a commercial catalyst that would cause similar reactions for chemical synthesis or to discover how to make drugs in which the dihydrofolate reaction is inhibited.

The human genome project was Nelson's final grand challenge example. Genetic coding, he reminded conferees, is based on DNA (deoxyribonucleic acid), which has 4 base pairs coiled into the well-known double helix. Because of the twofold redundancy involved, these 4 molecules form about 3 billion base pairs. The task then is to code for 4 bases, and 3 billion of them. If one can do 2 bits to a base pair, it sounds roughly like only about 10^9 bytes, Nelson said.

This seems like a relatively small data storage problem. However, data problems exist here for several reasons. First, current biology is mainly a "soup" sort of science, he said.

"Our ability to deal atom by atom or molecule by molecule is not really there. We can't see these things. Therefore, determining the particular DNA sequence in a given strand of DNA or in a given gene is error-prone. So one is dealing with difficult data, poorly behaved data," he explained.

In addition, because different gene sequences may overlap, a clean rendition of the data is not now available.

"Therefore one has to have multiple copies of it, looking for where it overlaps and matches and using statistical techniques to weed out the inevitable errors in order that, by about the year 2000, we can extract a nice linear sequence of 3×10^9 base pairs," he said.

Because DNA is the blueprint, the diagram that builds the body, the data are even more complex. The knowledge of DNA can be used, in effect, as a scaffold on which to hang all sorts of other data. He suggested that scientists would like to be able to hang not only the chemical structure but the geometric description of the amino acid and, ultimately, of the whole protein on that scaffold.

Different individuals differ by a small fraction of a percent in their DNA. When those deviations from the standard cause serious disease, however, they can be a source of great interest, he said.

Because those deviations can have a huge effect on behavior, on makeup, and on longevity, "one wants to have a number of copies of this," Nelson said.

"So when one thinks of what DNA actually means in terms of data analysis, it becomes a lot more than 3×10^9 base pairs, or about 10^9 bytes."

Nelson then listed some current projects with the large volumes of data they generate to illustrate some of the problems in data storage and data management.

Project	Data Generation
SSC	10^{15} bytes/year
EOSDIS	10^{12} bytes/day
Global ocean	10^{13} bytes/simulation
Retail trade	10^{12} bytes/year
Genome project	10^{10} bytes/year

These efforts all involve complex data structures, Nelson said, listing the projects and some of the structures involved.

Project	Data Structure
SSC	Event statistics
EOSDIS	Multiple scalar fields over 3 spatial dimensions and 1 temporal dimension
Global ocean	Multiple scalar fields
Retail trade	Purchase transaction date
Genome project	Nested data structures with fuzzy boundaries

Nelson next listed some problems that are common to many applications.

Efficient data access -- how to get at data
Version control (updates copies when master changes)
Detection of anomalies and interesting events in real time
Relating the data manager's view to the user's view
Ability to ask unanticipated questions
Hierarchical mass storage environment

Perhaps data that cannot be accessed may be referred to as the "landfill of cyberspace," he said.

In the area of version control, he asked, "If we have distributed data and somebody finds that an instrument was miscalibrated, how does one keep track of what subsidiary databases need to be updated, and do that efficiently?"

Nelson used the example of a very large relational database, containing fifty-some tables, used in the genome project to explain why trying to extract information from a reasonable query and doing it in a way that does not require the user to employ experts to do so can be very difficult.

Because the most important things in research are usually those things we only thought of recently, the ability to reply to unanticipated questions can be very important.

Revising data structures as requirements evolve and learning how to do this through hierarchical mass storage in a way that is cost-effective and practical are also important, Nelson pointed out.

For a data manager to answer a user with "Yes, I could get it, but it'll take six months and we don't have enough disk storage on the floor," is probably not a good answer," he said.

Nelson listed some examples of the simulation data he expects to be generated in the next few years.

Project	Description	Operations	Memory	Data Access
Global ocean levels depth, 1/4 degree resolution	Century at 40	10^{17}	4 GBytes	20 TBytes
Porous media	3-D immiscible	10^{18}	1 TByte flow	4 TBytes
Ductile materials dynamics	3-D molecular	10^{18}	1 TByte	3 TBytes
Plasma physics	Numerical tokamak	10^{18}	1 TByte	100 TBytes
QCD (quantum chromodynamics)	Quenched lattice	10^{18}	8 GBytes	8 GBytes

Another dimension of the subject lies in the area of network and distributed data. The National Research and Education Network (NREN) widens the horizons for data management but also adds complexity, Nelson said.

"In the Interim NREN we're sitting at network bandwidths that are from 1.5 megabits/second up to 45 megabits/second. In a year or so, we'll have some of the links at 155 megabits/second and we think in about two years, we'll be sitting at 622 megabits/second. By 1996, if current research pays off, we'll be looking at 1 gigabit/second and up," he said.

The network pipes are expected to be full within a year or two at each stage of deployment.

Animated video can be extremely important for understanding what's going on, for getting information and understanding out of data, he pointed out, and moving such video over a network consumes bandwidth greedily.

Large data flows can also aid in building distributed databases. "Obviously, there's a tradeoff between looking at the data remotely and looking at the data in your own backyard. Clearly, some of the data will be better off elsewhere, and the consumer of the data will use the network to get at it," Nelson said.

He mentioned what he called "database fusion," or distributed databases which are fused together in the same way that sensor fusion now leads to enhanced information.

"If one could achieve database fusion - one database could contain one aspect of the problem, and another database a different aspect," he said, then perhaps a more comprehensive view could be derived.

He cited distributed data searches as another focus of interest. In the area of access rights and control, the question is, "Can someone else get into my database if he's coming in on the network?"

In propagating data updates, especially for data that's distributed worldwide, how to track down who has the data and tell him that it's wrong?

Interfaces, gateways, and translators will also be needed, Nelson pointed out, because the world is not going to adapt a uniform data description model or data structure model.

Pointing to a map of the United States, Nelson described a network of lines superimposed there to show the Interim National Research and Education Network (Interim NREN). The lines, he said, represent the network's high-speed, long-haul lines, running between network nodes. The point is that the network is virtually ubiquitous.

A schematic illustrated what the network can offer to users: access to data banks, special research facilities, remote databases, libraries, high-performance computers, and the like.

Nelson summarized that the grand challenges of observation and simulation place grand challenges on data management. Archives will quickly grow from terabytes to petabytes, he said. Data sets will grow from gigabytes to terabytes and, in some cases, already have.

Networks will widen the data horizons but add problems for the data manager, he said.

"Soon we will be dealing with systems of distributed large data archives just as we now deal with systems of large computers. The future is sure to be interesting," he concluded.

Joseph King, of NASA Goddard Space Flight Center, asked Nelson whether he was aware of any cross-disciplinary discussions of what criteria can be used for making judgments about what bytes should be kept and what can safely be discarded.

Nelson replied that he considered efforts to try to determine what data may be more interesting than others, and what process to use to establish that priority, a very worthwhile exercise. Because we function in a resource-constrained universe, we are making choices about how we deal with data all the time, he said.

"Maybe one way to avoid unnecessarily creating the `landfills of cyberspace' is by exercising intelligence on the creation of data so that we are less forced to throw it away," he pointed out.

The decision techniques specifically used by the high-energy physicists are unlikely to translate into disciplines that are not fairly similar, he concluded. However, some of the methodology that leads t those techniques may be generalizable.

Grand Challenges and Data Management

Presented to

Goddard Conference on Mass Storage Systems and Technology

by

David B. Nelson

U.S. Department of Energy

September 22, 1992

Grand Challenge:

A fundamental problem in science or engineering, with broad economic or scientific impact, whose solution could be advanced by applying high performance computing techniques and resources *

* Grand Challenges 1993: High Performance Computing and Communications, OSTP

Examples of Grand Challenges: *

Weather, climate and global change
Ocean Sciences
Atomic nature of materials
Semiconductor materials and devices
Superconductivity
Nuclear fusion
Oil & gas prospecting and recovery
Efficient, clean combustion

Vehicle design
Vehicle signature
Computer vision
Speech recognition
Elementary particle physics
Astronomy
Structural biology
Human genome

* The Federal High Performance Computing Program, OSTP, 1989

A Grand Challenge in Data Managment *

Manage, manipulate, and analyze 50-100 TByte data sets from within Petabyte archive located in 5000 sq. ft. of floor area

* Source: Milt Halem

Cross-Disciplinary Problems in Data Storage and Management

I. Large Volumes of Data

A. SSC:	10^15 Bytes/year
B. EOSDIS	10^11 Bytes/day
C. Global ocean	10^13 Bytes/simulation
D. Retail trade	10^12 Bytes/year
E. Genome	10^10 Bytes/year

II. Complex Data Structures

A. SSC:	Event statistics
B. EOSDIS:	Multiple scalar fields over 3+1 D
C. Global ocean:	Multiple scalar fields over 3+1 D
D. Retail trade:	purchase transaction data
E. Genome:	Nested data structures with fuzzy boundaries

Cross-Disciplinary Problems in Data Storage and Management (Cont.)

III. Problems Common to Many Applications

- ✔ Efficient data access
- ✔ Version control (update copies when master changes)
- ✔ Detection of anamolies and "interesting events" (in real time)
- ✔ Relating data manager's view of data to user's view
- ✔ Relating data manager's view of data to user's view
- ✔ Ability to ask unanticipated questions
- ✔ Ability to revise data structures as requirements evolve
- ✔ Hierarchical mass storage environment

Tomorrow's Data Requirements for Simulations

Project	Description	Operations	Memory	Data Access
Global Ocean	century, 40 levels 1/4 degree	10^17	4 GBytes	20 TBytes
Porous Media	3-D immiscible flow	10^18	1 TBytes	4 TBytes
Ductile Materials	3-D molecular dynamics	10^18	1 TBytes	3 TBytes
Plasma physics	numerical tokamak	10^18	1 TBytes	100 TBytes
QCD	quenched lattice (64x64x64x128)	10^18	8 GBytes	8 GBytes

NREN widens horizens for data managment, but also adds complexity

- Network bandwidth:
  - now 1.5 Mb/s to 45 Mb/s
  - soon 155 Mb/s to 622 Mb/s
  - by 1996 1Gb/s and up
- Allows large flow of data, including multimedia
  - distributed data bases
  - data base fusion
  - data searches
- Issues
  - access rights and control
  - propagating data updates
  - interfaces, gateways, and translaters

Summary

- Grand Challenges of observation and simulation place grand challenges on data management
- Archives will grow from TBytes to PBytes
- Data Sets will grow from GBytes to TBytes
- Networks widen data horizens but add problems for the data manager
- Soon we will be dealing with systems of distributed large data archives, just as we now deal with systems of large computers
- The future is sure to be interesting

Current State of the Mass Storage System Reference Model

Mr. Robert Coyne

IBM MC5600
3700 Bay Area Boulevard
Houston, TX 77058

IEEE P1244

Storage System Standards Working Group

- IEEE SSSWG was chartered in May 1990 to abstract the hardware and software components of existing and emerging storage systems and to define the software interfaces between these components

- The immediate goal is the decomposition of a storage system into interoperable functional modules which vendors can offer as separate commercial products

- The ultimate goal is to develop interoperable standards which define the software interfaces, and in the distributed case, the associated protocols to each of the architectural modules in the model

IEEE SSSWG Organization

- Chair - Bob Coyne, IBM Federal Systems Company
- Technical Editor - Dave Isaac, Mitre
- Secretary - Charles Antonelli, University of Michigan
- Treasurer - Thomas Jefferson, Sandia National Laboratory
- Archivist - Dave Tweten, NASA Ames
- POSIX Liaison - Kurt Everson, IBM Federal Systems Company
- Architecture Steering Committee Chair - Mike Milillo, E-Systems, Inc.

IEEE SSSWG Subcommittees & Chairs

- Mapping Services - Andy Hanushevsky, Cornell University
- Bitfile Server - Dave Tweten, NASA Ames
- Storage Server - Lester Buck, IBM Federal Systems Company
- Physical Volume Library - Rich Wrenn, Digital Equipment Corporation
- Physical Volume Repository - Joseph Wishner, Storage Technology Corporation
- Mover - Bob Hyer, IBM Federal Systems Company
- Storage Systems Management - Steve Louis, National Energy Research Supercomputer Center

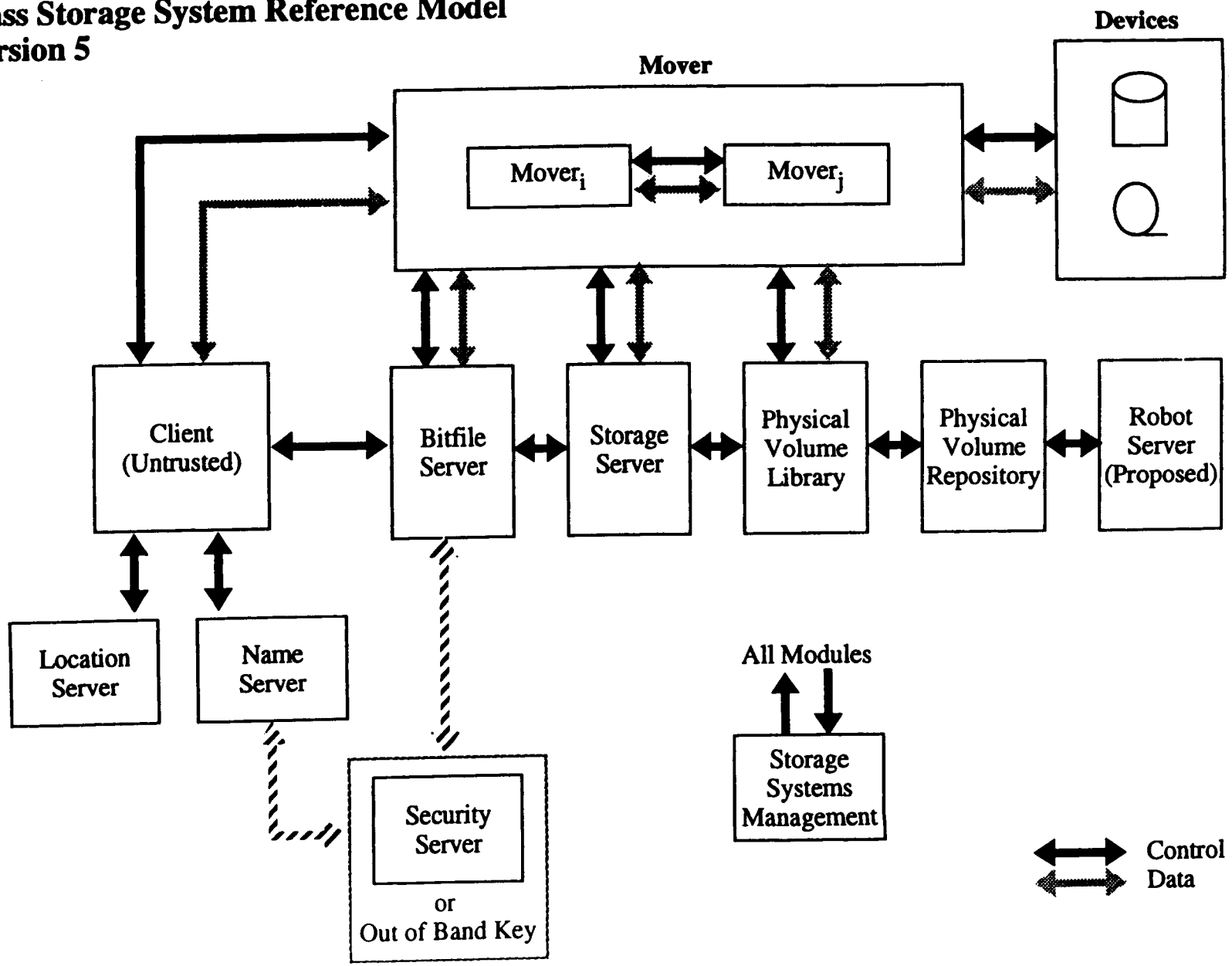
IEEE Standards Activity Board

- The IEEE SSSWG formally reported to the IEEE Mass Storage Systems and Technology Technical Committee (MSS&TC)

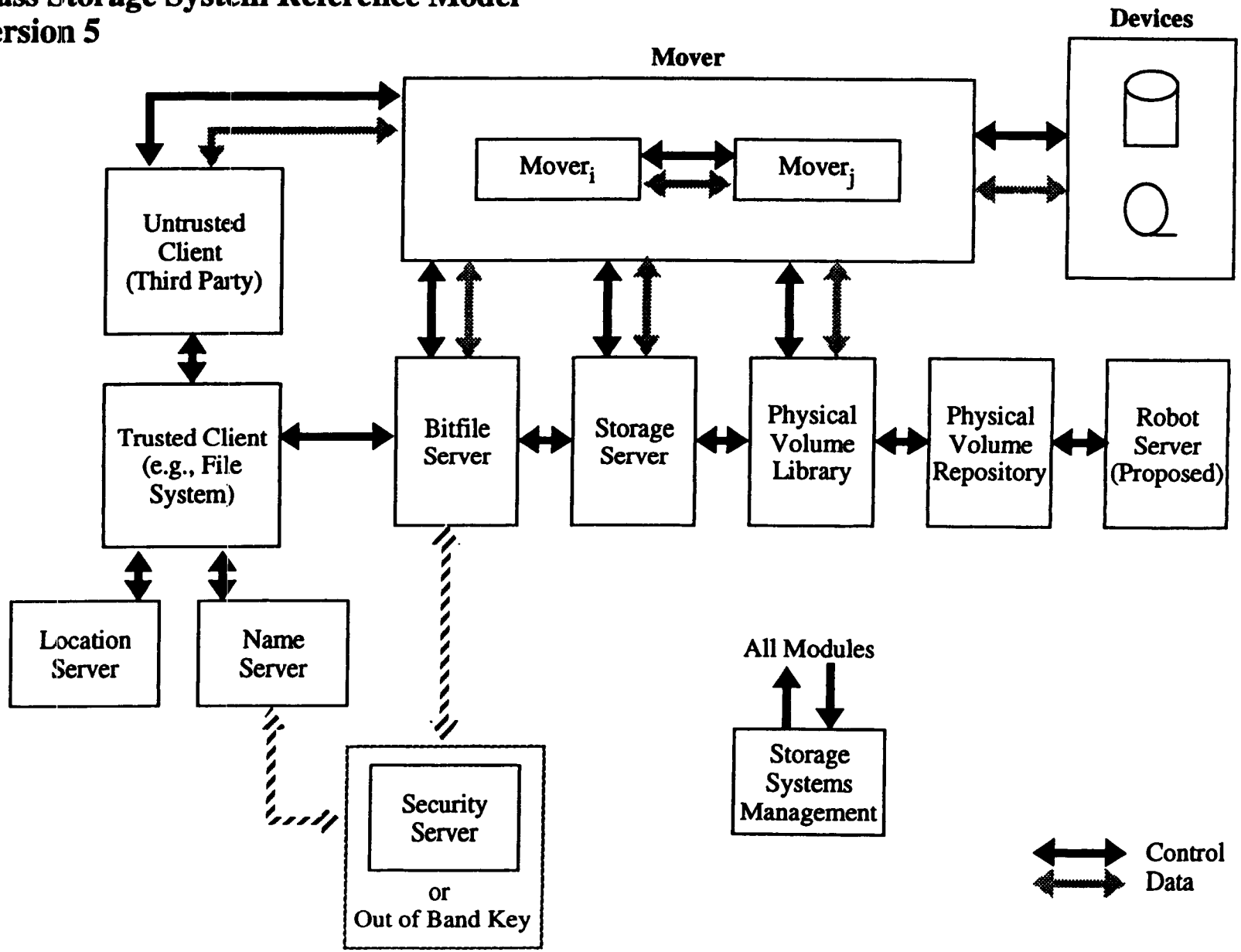
- The IEEE SSSWG now reports directly to the IEEE Standards Activity Board (SAB)

- The IEEE SAB sponsor for P1244 is Patric Savage, Shell Development Corporation

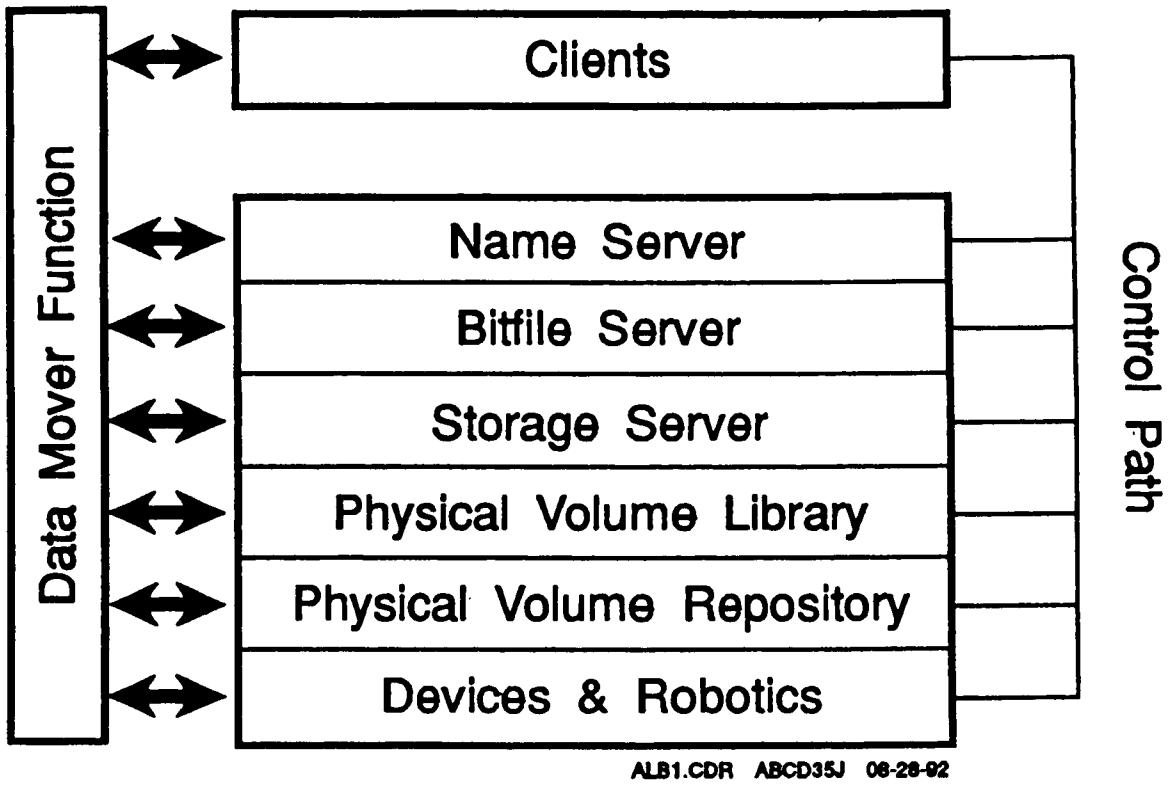
DRAFT Example
Mass Storage System Reference Model
Version 5



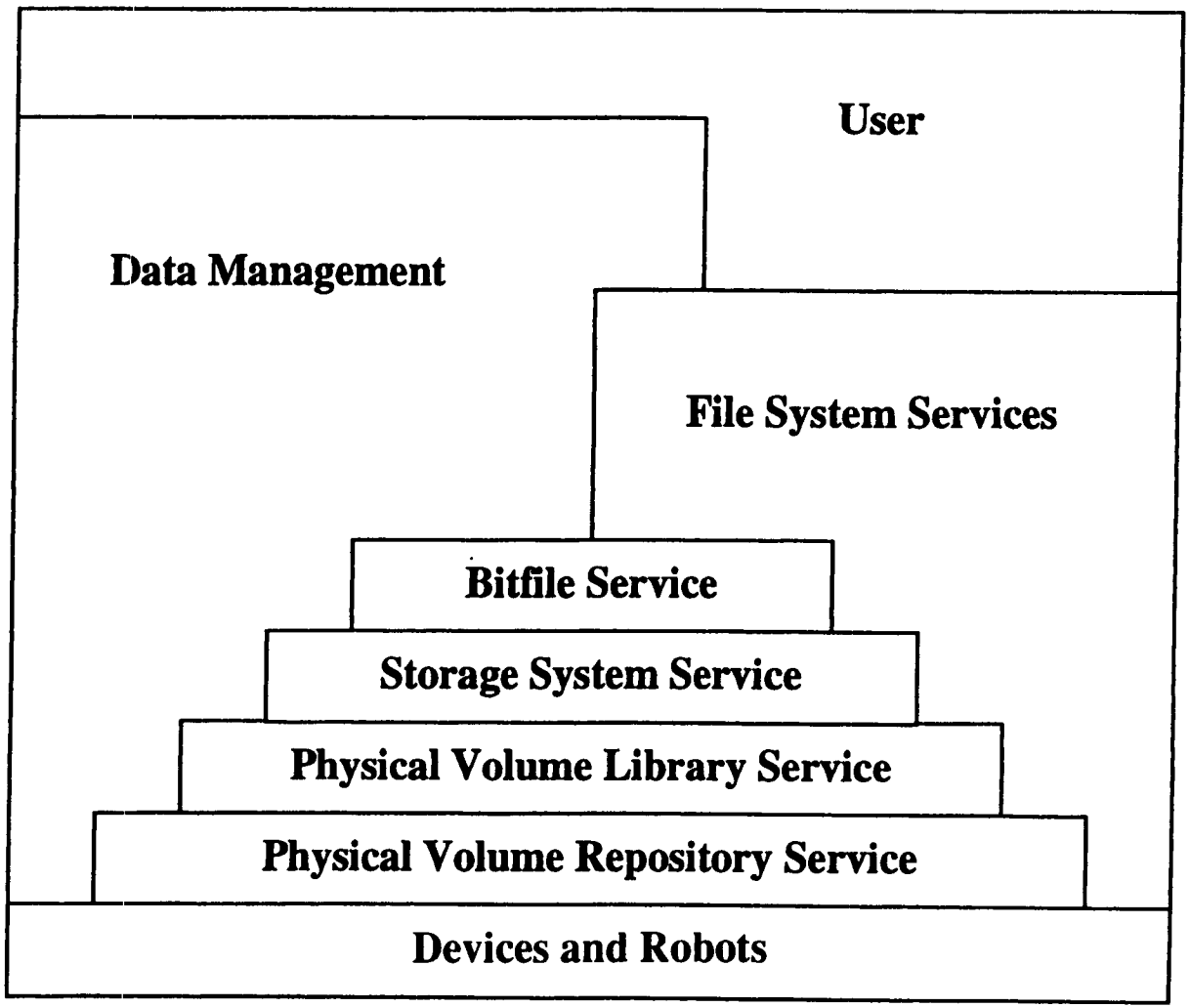
DRAFT Example
Mass Storage System Reference Model
Version 5



Layered View of the Reference Model



Layered Access to Storage Services



IEEE SSSWG Emphasis

- We are focused on decomposing storage systems into interoperable functional modules which vendors may offer as separate commercial products

- Our emphasis is on providing clients direct access to appropriate storage system services through standard interfaces

- We plan to distribute the IEEE MSSRM Version 5 at the IEEE Twelfth Symposium on Mass Storage, April 25, 1993

- We plan to assign IEEE project numbers (e.g., P1244.PVR) to each functional module interface to allow timely standards development and approval for well defined components (e.g., PVR, PVL, SSOID)

Features for MSSRM Version 5

- A generic format for unique identifier, termed Storage System Object Identifier (SSOID), has been proposed and a new PAR has been submitted to allow separate standardization of the format of these identifiers

- A division between location and name services has been established

  - protocol defined for location services
  - minimum functions defined for a compliant name server

Features for MSSRM Version 5

- A general security architecture has been established that recognizes that security is very site dependent and enables the implementation of a broad range of security options

  - each functional module interfaces with security services provided by a vendor or installation
  - general architecture provides for distributed security

- Security is divided into three separate implementation components:

  - Authentication establishes the identity of a client/server
  - Authorization calculates the access rights of a principal to a service
  - Enforcement applies the result of the authorization calculation at the point where access is actually requested

Features for MSSRM Version 5

- The Mover, formerly Bitfile Mover, transfers data between any source and sink (e.g., client, device, and/or functional module)

- Devices are no longer encapsulated by the storage server; the minimum software necessary to operate a device is the Mover

- Migration and replication of storage system objects (e.g., Bitfile Containers, Bitfiles and Virtual Volumes) are supported by the appropriate functional module with migration and replication policy administered by the Storage Systems Management

Features for MSSRM Version 5

- The Physical Volume Library has been established which is a generalization of an enterprise-wide, removable media management system (e.g., tape management system) which tracks the current locations and status of all removable volumes across multiple distributed PVRs

  - controls the actual mounting and dismounting of volumes
  - maintains mount queues
  - verifies internal volume labels
  - globally optimizes the use of drives
  - tracks the life cycle state of removable media
  - maintains scratch pools

- Robotic Services are proposed as a network interface to the basic robot operation, essentially mount/dismount physical volume in Slot X onto Drive Y and inject/eject volume into/from robot

Features for MSSRM Version 5

- The Physical Volume Repository (PVR) maintains the mapping of volume identifiers to slots and drives, optimizes apparent mount times by staging, and enforces various management and security attributes

- Each PVR manages a single repository or set of connected repositories

Features for MSSRM Version 5

- Storage system management information is structured in terms of managed objects, which encapsulate attributes, management operations, and notifications

- Several storage system management functions have been identified, including

  - migration of storage objects to a cheaper level in a storage hierarchy
  - defragmentation and repacking of Bitfiles and volumes
  - initialization, addition, and deletion of new storage resources
  - logging of relevant storage events, errors, and alarms
  - backup
  - Bitfile recovery
  - system recovery
  - capacity planning

Storage System Domain

- The Mapping Services Subcommittee has proposed that the location and name server(s) operate within a storage system domain which is the intersection of security and administrative domains

- A storage system domain contains one logical location server, a common authentication mechanism and security policy, and a common administration policy

- Name services remain an optional portion of the model because the universe of entities which can act as name servers, from hierarchical file systems through associative databases, prevent standardization of the naming semantics

IEEE SSSWG Plans

- Publish MSSRM Version 5 in 2q93

- Revise PAR 1244 to develop a Recommended Practice instead of Guide

- Identify components for Full Standards and develop them as independent standards

  - SSOID (PAR submitted to SAB 3q92)
  - PVR
  - PVL

- Hold election for IEEE SSSWG Chair at the Feb '93 meeting

Help, Help, Help !

- The IEEE SSSWG needs additional participants, especially in the areas of technical writing and editing. While our membership is growing, we need additional members to work in our subcommittees

- We need participation from the data management, database, and file system development and user communities. This group is not represented. We are not adequately addressing this important area

- If you are interested in actively participating in the IEEE SSSWG, please contact Bob Coyne (713-282-7274, coyne@houvmscc.vnet.ibm.com)

New IEEE MSS&TC Specialist Workshops

- The IEEE MSS&TC has authorized a series a Specialist Workshops on Data Management. We are looking for program committee members to plan the workshop series and coordinate with the organizations that will host the workshops.

- Michael Farrell, Director of the Center for Global Environment Studies at Oak Ridge National Laboratory, has submitted a proposal for a science data management workshop addressing interpretation, integration and interrogation of very large data bases

- Los Alamos National Laboratory, the National Security Agency, and Lawrence Livermore National Laboratory plan to submit workshop proposals.

- If you are interested in participating in an IEEE MSS&TC Specialist Workshops on Data Management, please contact Bob Coyne (713-282-7274, coyne@houvmscc.vnet.ibm.com)

The Standards Process: X3 Information Processing Systems

Jean-Paul Emard

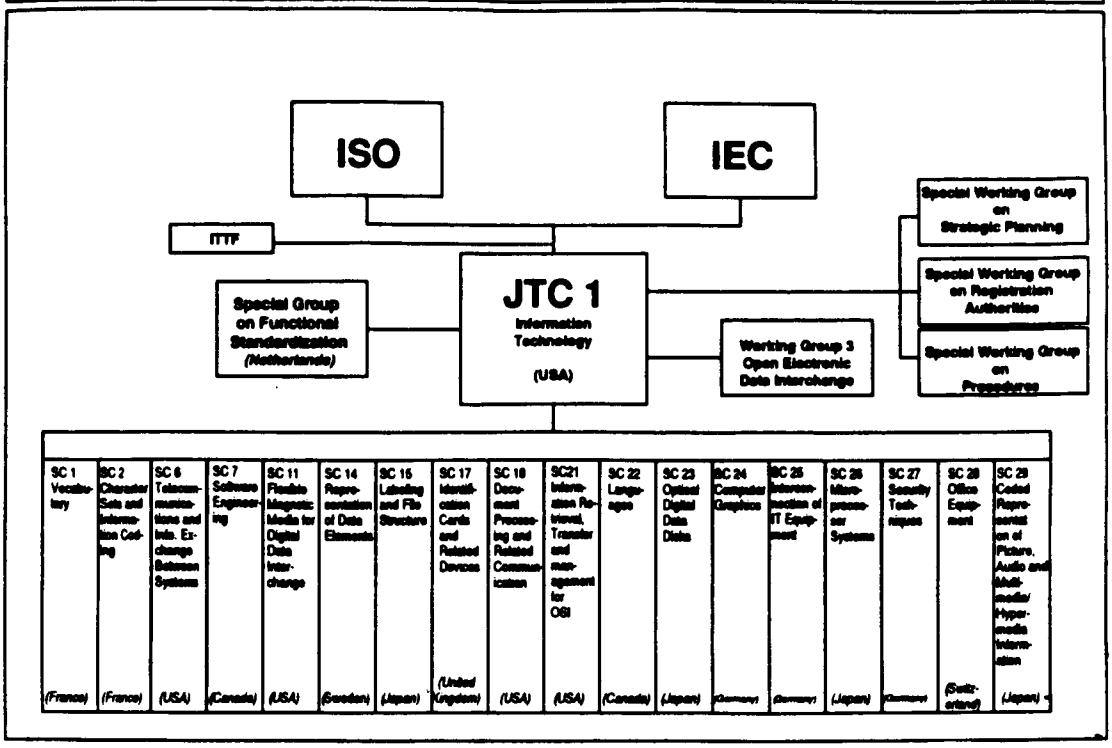
Director
Standards Secretariats

Computer and Business Equipment Manufacturers Association

1250 Eye Street, Northwest
Suite 200
Washington, DC 20005

(202) 626-5740
Fax: (202) 638-4922

JTC 1 Organizational Chart



International Organization For Standardization

ISO

- a non-treaty organization
- founded in 1946
- covers standardization in all fields (except IEC)
- 87 countries
- 73 member bodies
- 14 correspondent members

International Electrotechnical Committee

IEC

- **founded in 1906**
- **responsible for international standardization in the electrical and electronics fields**
- **41 National Committees**

ISO / IEC JTC 1

JTC 1 - Joint Technical Committee 1

- **title - Information Technology**
- **scope - "Standardization in the field of information technology"**
- **established in January 1987 (replaced ISO TC97, IEC TC83 and SC47B)**
- **18 subcommittees**
- **80 working groups**
- **chairman: Mrs. Mary Anne Lawler (USA)**
- **secretariat: ANSI**
  - **membership: (1992)** 25 - Participating members (P-members, have power to vote and defined duties)
    17 - Observer members (O-members, no power to vote; may attend meetings, and receive documents)

ISO / IEC JTC 1

Subcommittee

- established by JTC 1
- studies particular part of work assigned to JTC 1
- must comprise at least 5 'P' members
- secretariat appointed by JTC 1 from among 'P' members of SC
- Chairman is nominated by the SC Secretariat; endorsed by the nominee's National Body and by the subcommittee; and appointed by JTC 1
- members are National Bodies
- delegates represent their National Body's positions
- category A liaisons may also send delegations

ISO / IEC JTC 1

Working Group (WG)

- established by JTC 1 or SC
- undertakes a specific task
- continues in being until completion of work for which it was established
- members are individual experts designated by National Bodies
- Category A liaisons may also nominate members who must represent the liaison organizations
- WG members act as experts and do not necessarily represent their National Body's positions
- WG members shall indicate whether views expressed reflect National Body positions or personal opinions.
- convener appointed by parent committee for a three-year term
- National Body of convener must support appointment

ISO / IEC JTC 1

Other types of groups at the SC or WG level

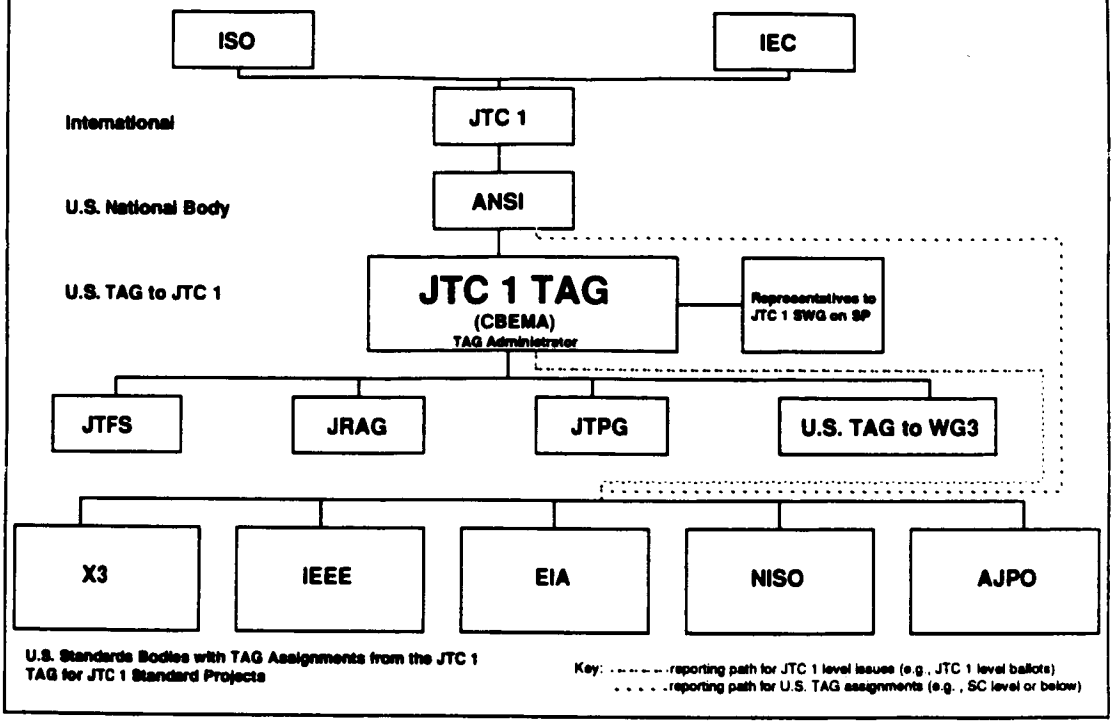
- examples are: Ad Hocs, Rapporteur, Drafting and Editing
- established by SC or WG
- membership defined by parent body
- studies precisely defined issues within the scope of its parent body
- usually reports at same or next meeting
- disbanded upon completion of assigned tasks

ISO / IEC JTC 1

ITTF - Information Technology Task Force

- joint group of ISO Central Office and IEC Central Office
- headquartered in Geneva
- responsible for the day to day planning and coordination of the activities within JTC 1

U.S. Interface to JTC 1



ANSI

U.S. National Body member of JTC 1

- **established the JTC 1 TAG to serve as the Technical Advisory Group (TAG) to ANSI**
- **coordinates development of U.S. position with responsible TAG Administrator**

Neutral Body

- **Serves as the Secretariat for JTC 1 and Subcommittees 6, 11, 18, 21**

National Organizations

American National Standards Institute (ANSI)

ANSI is the coordinator of the voluntary standards system in the United States, and represents the U.S. in the voluntary international standards developing organizations.

U.S. TAG for ISO/IEC JTC1 (JTC1 TAG)

JTC1 TAG is the group that develops U.S. positions for ANSI (as the U.S. member body) on the proposed ISO/IEC JTC1 program of work and proposed standards.

JTC 1 TAG

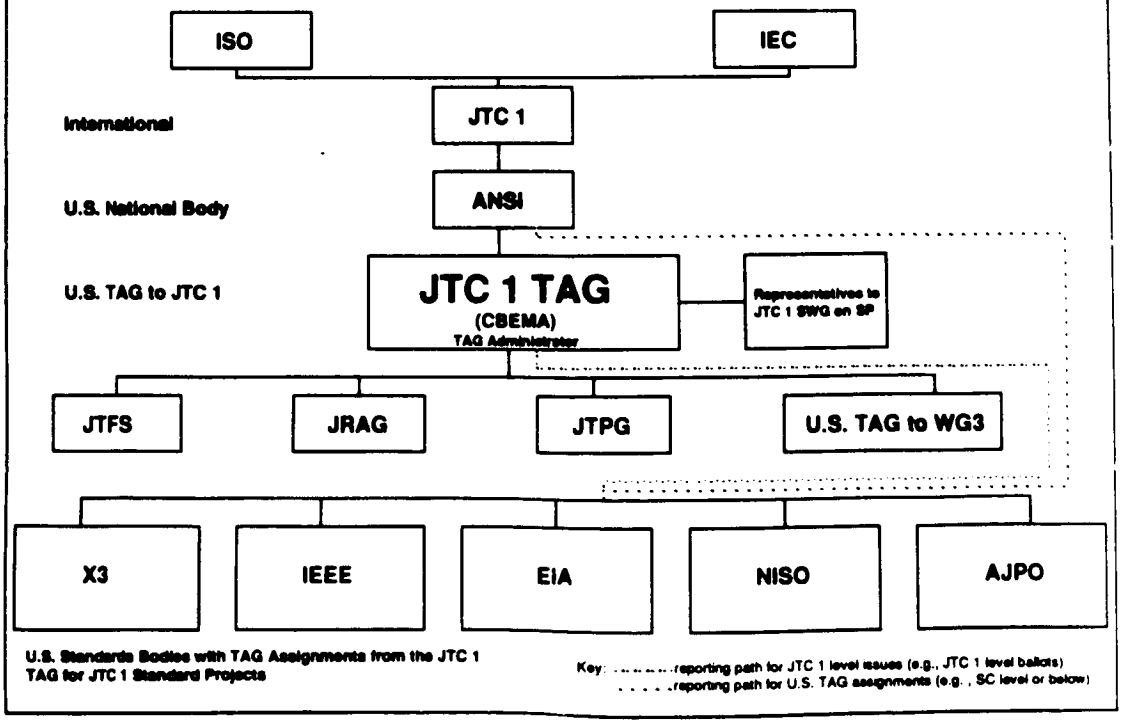
Serves as the U.S. TAG for JTC 1

CBEMA serves as the JTC 1 TAG Administrator

Has three primary responsibilities relative to ANSI's participation in JTC 1

- **Responsible for JTC 1 level U.S. positions**
- **Responsible for coordinating U.S. interests and develop consensus**
- **acts as SC, WG, or Project TAG in cases where a specific TAG assignment does not exist or is directly under the JTC 1 TAG**

U.S. Interface to JTC 1



In the United States, standards are developed three ways:

1. Canvass Method

(Department of Defense for ADA)

2. Accredited Standards Committee

(X3 - X9 - X12)

3. Accredited Standards Developing Organizations

(IEEE)

Related National Standards Developing Organizations

X9 ***Financial Services***

Secretariat: American Bankers Association (ABA)

X12 ***Electronic Business Data Interchange***

Secretariat: Data Interchange Standard Association (DISA)

T1 ***Telecommunications***

Secretariat: Exchange Carrier Standards Association (ECSA)

IEEE ***Institute for Electrical and Electronic Engineers***

The Computer Society, the Communications Society

MUMPS* Users Group

****Massachusetts General Utility Multi-Programming System***

Ada* - Department of Defense (DoD)

****Ada is a registered trademark of the U.S. government, ADA Joint Program Office***

EIA ***Electronics Industry Association***

AIIM ***Association for Information & Image Management***

NISO ***National Information Standards Organization***

ISO/IEC JTC1, Information Technology

This is a joint committee between the International Standards Organization (ISO) and the International Electrotechnical Commission (IEC) that serves as the international voluntary standards organization which develops standards in information technology.

CCITT

The International Consultative Committee for Telephony and Telegraphy (CCITT) is a treaty organization whose countries are represented by their governments. In the case of the U.S., this is the State Department.

The purpose of CCITT is to develop recommendations on questions related to technical, operational and tariff matters on facsimile, telegraph and telecommunications.

Regional Organizations

Standard Developing

CEN/CENELEC
ETSI
ECMA

Workshops

NIST
EWOS
etc.

Other Organizations Involved in IT Standards

Consortia, e.g.,

Network Management Forum
Object Management Group
Open Software Foundation
UNIX International
X Consortium
X/Open

Companies

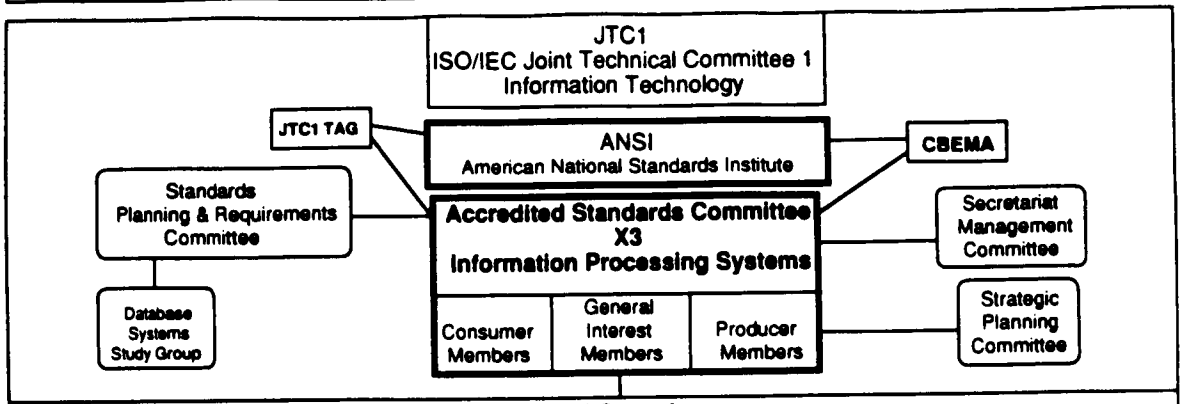
User Groups

Government

Academia

Professional Societies

X3 Organization



X3 Technical Committees

X3A1 Optical Character Recognition
X3B5 Digital Magnetic Tape
X3B6 Instrumentation Tape
X3B7 Magnetic Disks
X3B8 Flexible Disk Cartridges
X3B9 Paper/Forms Layout
X3B10 Credit/Identification Cards
X3B11 Optical Digital Data Disks
X3H2 Database
X3H3 Computer Graphics
X3H4 Information Resource Dictionary System
X3H5 Parallel Processing Constructs for High Level Programming Languages
X3H6 Case Tool Integration Models
X3H7 Object Information Mgmt
X3J1 PL/I
X3J2 Basic
X3J3 FORTRAN
X3J4 COBOL
X3J7 APT
X3J9 PASCAL
X3J10 APL
X3J11 C
X3J12 DIBOL
X3J13 LISP
X3J14 FORTH
X3J15 DATABUS
X3J16 C++
X3J17 Prolog
X3J18 REXX
X3K5 Vocabulary
X3L2 Codes & Character Sets
X3L3 Audio/Picture Coding
X3L3 Data Representation
X3S3 Data Communications
X3T2 Data Interchange
X3T3 Open Distributed Processing
X3T4 Security Techniques
X3T5 Open Systems Interconnection
X3T6 Non-Contact Info. Systems Interfaces
X3T7 Internationalization
X3T8 Fault Isolation
X3T9 I/O Interface
X3V1 Text: Office & Publishing Systems
X3W1 Office Machines
SC21 TAG - Information Retrieval, Transfer & Management for ISO
SC22 TAG - Languages

X3, Information Processing Systems

Scope:

Standardization in the areas of computers and information processing and peripheral equipment, devices, and media related thereto: standardization of the functional characteristics of office machines, plus accessories for such machines, particularly in those areas that influence the operators of such machines.

X3 Standards Development Process

Planning Phase

Milestone	Description
0	Development & submission of project proposal
1	SPARC acceptance of proposal for review
2	SPARC determination if study group required or forwards to X3
3	Study group formed if required
4	Study group recommendation re: project proposal
5	X3 ballots project proposal, response to negatives, press release

Note: For type I projects use milestones 1-5 and section 10.5 for further processing

X3 Standards Development Process

Development Phase

Milestone	Description
6	TC develops work plan
7	TC develops draft proposed standard (Project Editor is appointed)
8	TC ballots draft proposed standard
9	TC approves of draft proposed standard
10	Forward to X3 Secretariat / SPARC
11	SPARC compliance review

X3 Standards Development Process

Approval Phase

Milestone	Description
12	dpANS forwarded for public review
13	TC consideration & action on public review comments and subsequent public review
14	X3 ballot on dpANS and resolution of comments
15	X3 default ballot on unresolved negatives
16	Submission of dpANS to ANSI/BSR
17	ANSI BSR review / approval / appeal period
18	Final copy forwarded to ANSI for style review and publication
19	Potential submission of approved American National Standard for JTC1 fast track

Note: See type D and type I Flow Charts

The Standards Process: Technical Committee X3B5 Digital Magnetic Tape

Sam Cheatham

Storage Technology Corporation
Vice President
Tape and Library Systems Development
VC Technical Committee X3B5
2270 South 88th Street
MS 0275
Louisville, CO 80028

Abstract

The presentation will provide the definition of X3B5, where it fits in the national and international standards development process and how it interfaces and influences the world community of standards developers. Detail concerning the focus of the committee, how it operates and what the group sees as the future trends in the area of interchange standards utilizing the multifaceted, ubiquitous magnetic tape. Highlighted in the presentation is:

- The definition of X3B5
- Where it fits in the Information Technology Standards development arena (US).
- How it interfaces with the world community of Standards developers.
- The purview of X3B5.
- How it operates. (TC Style Guide)
- The technologies and their future directions.

The Standards Process; X3B5

In general, a technical committee such as X3B5 is defined by the projects it is authorized to develop. Specifically, Technical Committee X3B5, Digital Magnetic Tape, develops proposed standards for the interchange of data by digital magnetic tape for computer peripheral applications. These standards developments apply to three levels of digital data interchange i.e., Media, unrecorded magnetic media and its associated container for media compatibility; Physical Format, the recorded format for subsystem compatibility; Logical Format, the labels and file structure for system interoperability.

The committee is comprised of 40 voting members and 13 observers. Over the many years of its existence, the membership of X3B5 has changed. The changes that have occurred however, have been gradual, reflecting normal attrition and changes in technology. A basic cadre of committee members has always been in place and it provides the necessary continuity to ensure that the developed proposed standards are consistent and technically sound. Effective leadership in a voluntary, consensus process, in addition to technically astute contributing participants are paramount if any level of success is to be achieved. Fortunately, X3B5 has had and still has both. Evidence of the above is demonstrated by the number of projects worked concurrently (28), and the number of standards developed and maintained (28) by the committee which is under the current chairmanship of Mr. Richard Steinbrenner.

In the United States, X3B5 is one of 43 technical committees developing standards in the information technology arena. The Computer Business Equipment Manufacturers Association (CBEMA) is the Secretariat for X3 ,the Accredited Standards Committee, Information Processing Systems, which manages the standards developments within its purview ensuring that due process in developing these standards is achieved. When a developed standard meets all the due process criteria specified by the American National Standards Institute (ANSI), it is published as an ANSI Standard.

The global relationships of X3B5 are carried out via ANSI's affiliations with the various national and international standards developers. The international committee of interest to X3B5 is the International Organization for Standardization / International Electrotechnical Commission (ISO/IEC) Joint Technical Committee 1 (JTC 1). ANSI not only holds the Secretariat for ISO/IEC JTC 1, but represents the United States in those JTC 1 Sub Committees that are of interest to the U.S. . ISO/IEC JTC 1/SC11, Flexible Magnetic Media for Digital Data Interchange and ISO/IEC JTC 1/SC15, Labelling and File Structure are the committees to which X3B5 is a Co-Technical Advisory Group.

TAGs are committees accredited by ANSI's Executive Standards Council (ExSC) for participation in ISO technical activities and operate in compliance with the ANSI Criteria for the Development and Coordination of US Positions in the International Standardization Activities of the ISO and IEC. The TAG is the ANSI recognized group that has the primary responsibility for participation in the ISO Technical Committee or Subcommittee work. It is the TAGs job to recruit delegations, supervise their work, and determine ANSI positions on proposed standards.

The functions of the TAG are as follows:

- Recommend registration of ANSI as a "P" or "O" member of an ISO technical committee or subcommittee or recommend a change in ANSI membership status on an ISO technical committee or subcommittee. In this case, "P" membership to SC11 and SC15 was recommended.

- Initiate and approve US proposals for new work items for consideration by an ISO technical committee or subcommittee.

- Initiate and approve US working drafts for submittal to ISO technical committees or subcommittees (and where appropriate, working groups) for consideration as committee drafts.

- Determine the US position on an ISO draft international standard, draft technical report, committee drafts, ISO questionnaires, draft reports of meetings, etc.

- Provide adequate US representation to ISO technical committee or subcommittee meetings, designate heads of delegations and members of delegations, and ensure compliance with the *ANSI Guide for US Delegates to ISO/IEC Meetings.*

- Determine US positions on agenda items of ISO technical committee or subcommittee meetings and advise the US delegation of any flexibility it may have on these positions.

- Nominate US technical experts to serve on ISO working groups.

- Provide assistance to US secretariats of ISO technical committees or subcommittees, upon request, including resolving comments on draft international standards, draft technical reports and committee drafts.

- Identify and establish close liaison with other US technical advisory groups in related fields, or identify ISO or IEC activities that may overlap the TAG's scope.

- Recommend to ANSI the acceptance of secretariats for ISO technical committees or subcommittees. ANSI hold the Secretariat for SC11.

- Recommend that ANSI invite ISO technical committees or subcommittees to meet in the United States.

- Recommend to ANSI US candidates for chair of ISO technical committees or subcommittees and US convenors of ISO working Groups.

X3B5 also interacts with its equivalent technical committees in the European Computer Equipment Manufacturers Association (ECMA), TC17, Magnetic Tapes and Cartridges and TC19, Flexible Disk Cartridges. A number of X3B5 members are also members of the corresponding ECMA committees. This direct involvement provides the conduit to ensure that standards developed in the U.S. are technically equivalent to those developed at ECMA and subsequently at the ISO/IEC JTC 1 Subcommittee. In the U.S., liaison activities with X3 technical committees, X3B6, Instrumentation Tape, X3B8, Flexible Disk Cartridges, X3B11, Optical Digital Data Disks, and X3T9, I/O Interfaces are maintained.

The method of operation employed by the X3 Technical Committees is delineated in X3/Standing Document-2, *Organization, Rules and Procedures of X3.* This document defines the requirements for membership, officers, documentation, voting, etc. that ensure due process. In addition to the official rules and regulations that direct the standards development process within X3B5, the committee has developed a TC Style Guide for use by the various project editors.

The Guide is used to assist in the preparation of draft standards that conform to ANSI's requirements and to X3B5's unique requirements. The guide provides information on format, style, standardized text, approved definitions and conversion of units unique to X3B5's Requirements. Always viewed to be a "living document", it has been updated to:

- Encompass new common aspects brought about by helical-scan technology.
- Be compatible with the ISO/IEC Directives on the Drafting and Presentation of International Standards.
- Be compatible with the new ANSI Style Manual.
- Take into account lessons learned from experience with the ANSI Pre-Edit Process.

Another "tool of the trade' is the Model for Digital Data Interchange by means of removable computer storage media (DDI Model). The purpose of the DDI reference model is to serve as a general planning document which clarifies where specific tasks should be undertaken by standardization committees. It also serves as a conceptual tool which can be applied in building coherent sets of standards for specific digital data interchange applications. The DDI Model is comprised of four levels. Level 1 specifies the interchange requirements for the unrecorded media. Some of the requirements in this area include but are not restricted to the dimensional, mechanical, magnetic and optical properties. Level 2 specifies the interchange requirements for the recorded media. Items such as track locations, data correction techniques, modulation schemes etc., are considered. Level 3 specifies the interchange requirements for the volume identification labels, file directories and file structures of the recorded media. X3B5 concerns itself with these three levels of the model. Level 4 is required in order to accomplish general tasks, such as interchanging ASCII files on a particular medium, or specialized tasks, such as interchanging text on flexible disk cartridges or interchanging images on optical disks.

An example of an implementation of the DDI Model is as follows:

LEVEL 3		
LOGICAL FORMAT	**X3.27 -1987**	*Magnetic Tape Labels and File Structure*
LEVEL 2		
PHYSICAL	**X3.14 -1983**	*Recorded Magnetic Tape, 200 cpi, NRZI*
	X3.22 -1990	*Recorded Magnetic Tape, 800 cpi, NRZI*
	X3.39 -1992	*Recorded Magnetic Tape , 1600 cpi, PE*
	X3.157-1987	*Recorded Magnetic Tape, 3200 cpi, PE*
	X3.54 -1992	*Recorded Magnetic Tape, 6260 cpi, GCR*
LEVEL 1		
MEDIA	**X3.40 -199X**	*Unrecorded Magnetic Tape , 800 cpi, 1600 cpi, 6250 cpi*

The technologies encompassed by the activities of X3B5 include the following:

LONGITUDINAL RECORDING		HELICAL SCAN RECORDING	
1/2"	Open Reel Tape	4mm	Cartridge
1/2"	Tape Cartridge	8mm	Cartridge
1/4"	Tape Cartridge	12.65mm	Cartridge
.15"	Tape Cassette	19mm	Cartridge

The continuing evolution of the above technologies in the area of volumetric efficiency, elemental capacities and transfer rates as required by the market and the maintenance of present standards constitute a workload that extends into the next century.

The Standards Process

TECHNICAL COMMITTEE X3B5
DIGITAL MAGNETIC TAPE

Sam Cheatham
Storage Technology Corporation
VP Tape & Library Systems Dev.
VC Technical Committee X3B5

Who / What is X3B5

- DEFINED BY ASSIGNED PROJECTS
- PROJECTS ADDRESS DATA INTERCHANGE
- INTERCHANGE MEDIA IS MAGNETIC TAPE

Sam Cheatham Storage Technology Corporation

Who / What is X3B5

TECHNICAL COMMITTEE X3B5 DEVELOPS PROPOSED STANDARDS FOR THE INTERCHANGE OF DATA BY DIGITAL MAGNETIC TAPE FOR COMPUTER PERIPHERAL APPLICATIONS

Sam Cheatham Storage Technology Corporation

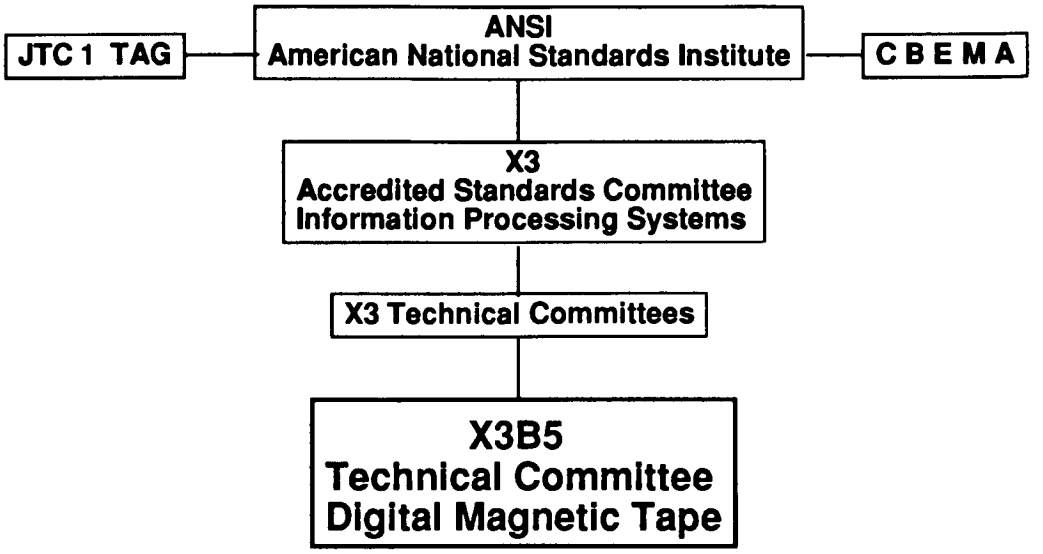
Who / What is X3B5

- OPERATING MORE THAN 25 YEARS
- RESPONSIBLE FOR 56 PROJECTS
  - 28 UNDER DEVELOPMENT
  - 18 IN MAINTENANCE MODE
  - 5 UNDER REAFFIRMATION
  - 3 UNDER REVISION
  - 1 LIAISON PROJECT
  - 1 TECHNICAL REPORT

Who / What is X3B5

- 40 VOTING MEMBERS
- 13 OBSERVERS
- 150 PERSON MAILING LIST

U.S. ORGANIZATION



Sam Cheatham — Storage Technology Corporation

U.S. ORGANIZATION

X3 Technical Committees

X3A1 Optical Character Recognition

X3B5 Digital Magnetic Tape

X3B6 Instrumentation Tape
X3B7 Magnetic Disks
X3B8 Flexible Disk Cartridges
X3B9 Paper / Forms Layout
X3B10 Credit / Identification Cards
X3B11 Optical Digital Data Disks

X3H2 Database
X3H3 Computer Graphics
X3H4 Information Resource & Dictionary
X3H5 Parallel Processing Constructs for High Level Programming Languages
X3H6 Case Tool Integration Models

SC21 TAG Information Retrieval & Management for ISO

X3J1 PL / 1
X3J2 Basic
X3J3 Fortran

X3J43 COBOL
X3J7 APT
X3J9 PASCAL
X3J10 APL
X3J11 C
X3J12 DIBOL
X3J13 LISP
X3J14 FORTH
X3J15 DATABUS
X3J16 C++
X3J17 PROLOG
X3J18 REXX

X3K5 Vocabulary

X3L2 Codes & Character Sets
X3L3 Audio / Picture Coding
X3L8 Data Representation

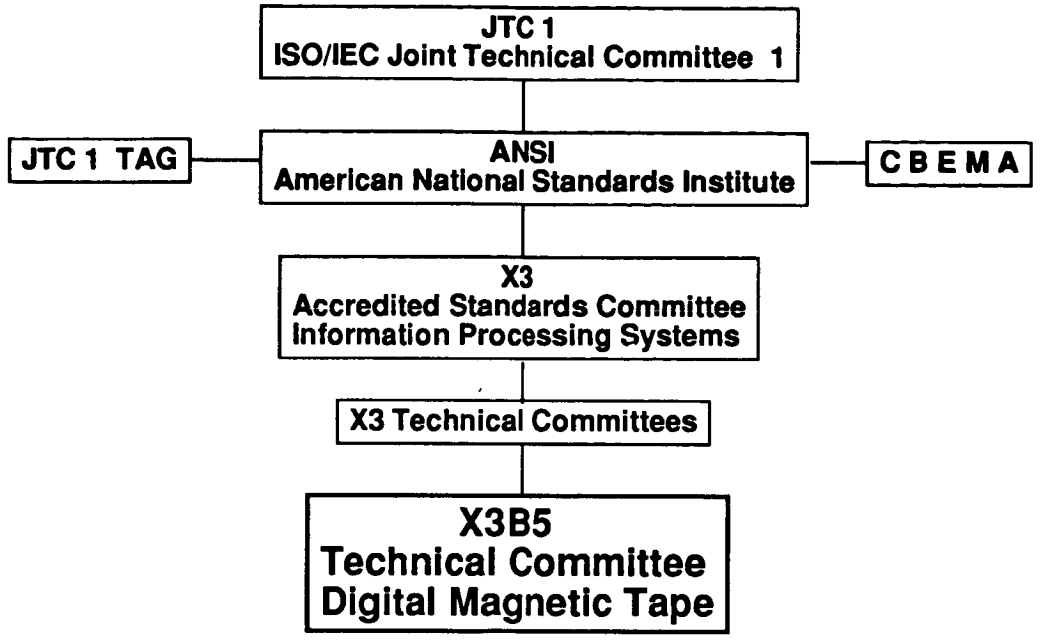
X3S3 Data Communications
X3D2 Data Interchange
X3T3 Open Distributed Processing
X3T4 Security Techniques
X3T5 Open System Interconnection
X3T6 Non-Contact Info System Interface
X3T7 Internationalization
X3T8 Fault Isolation
X3T9 I / O Interface

X3V1 Text: Office & Publishing Systems
X3W1 Office Machines

SC22 TAG Languages

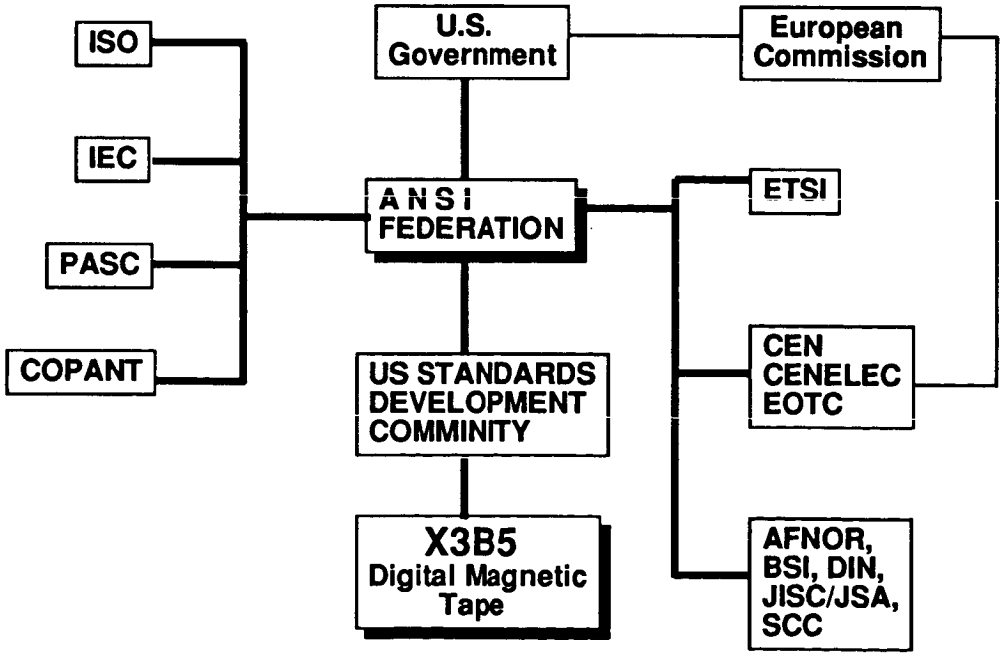
Sam Cheatham — Storage Technology Corporation

U.S. ORGANIZATION



Sam Cheatham — Storage Technology Corporation

GLOBAL RELATIONSHIPS



Sam Cheatham — Storage Technology Corporation

X3B5
Liaison Activities

ISO/IEC JTC-1

Co U.S. TAG to

SC11, *Flexible Magnetic Media for Digital Data Interchange*
SC15, *Labelling and File Structure*

ECMA

TC17, *Magnetic Tapes and Tape Cartridges*
TC19, *Flexible Disk Cartridges*

X3

X3B6, *Instrumentation Tape*
X3B8, *Flexible Disk Cartridges*
X3B11, *Optical Digital Data Disks*
X3L2, *Codes & Character Sets*
X3T9, *I/O Interfaces*

Sam Cheatham — Storage Technology Corporation

X3B5
How it Operates

Tools of the Trade

- Standing Document 2
- Membership
- Agendas
- Document Registers
- Minutes & Action Items
- DDI Reference Model
- Development Process
- Officers
- Document Distribution
- Meeting Schedules
- Voting
- TC Style Guide

Sam Cheatham — Storage Technology Corporation

X3B5
How it Operates

Procedures: SD-2

"The Object of these procedures is to achieve a concensus of the participants rather than some minimum ratio of approvals versus objections to produce technically sound standards which will be used because of their technical and economic merit and to ensure that due process in developing these standards is achieved."

Sam Cheatham — Storage Technology Corporation

X3B5
How it Operates

The Digital Data Interchange Reference Model

Purpose :

- It serves as a Conceptual Tool in Building a Coherent Set of Standards
- It Serves as a General Planning Document for the Standards Activities

Sam Cheatham — Storage Technology Corporation

X3B5
How it Operates

Level 4:
Applications Requirement

Level 3:
Volume, File & Directory Identification

Level 2:
Interchange Requirements

Level 1:
The Unrecorded Media

Sam Cheatham Storage Technology Corporation

A DDI Standards Set

LOGICAL FORMAT *X3.27-1987*

(LEVEL 3)

PHYSICAL

(LEVEL 2)

X3.14-1983	*200 CPI*
X3.22-1990	*800 CPI*
X3.39-1992	*1600 CPI*
X3.157-1987	*3200 CPI*
X3.54-1992	*6250 CPI*

MEDIA *X3.40-199X*

(LEVEL 1)

Sam Cheatham Storage Technology Corporation

X3B5
How it Operates

TC Style Guide

- **Assists in the preparation of draft standards**

  Conform to ANSI requirements
  Conform to X3B5 requirements

- **Provides information on**

  Format ***Style*** ***Units Conversion***
  Standardized text ***Approved definitions***

- **A "Living Document" Updated to**

  Encompass new common technology aspects
  Be compatible with the ISO/IEC Directives
  Be compatible with the ANSI Style Manual
  Incorporate lessons learned from recent experience

Sam Cheatham **Storage Technology Corporation**

Technologies

Longitudinal Recording	Helical Scan Recording
1/2" Open Reel Tape	4mm Cartridge
1/2" Tape Cartridge	8mm Cartridge
1/4" Tape Cartridge	12.65 mm Cartridge
.15" Tape Cassette	19 mm Cartridge

Sam Cheatham **Storage Technology Corporation**

Future Trends
(Capacity in M bytes)

Media	Technology	Present	Near Term	Future
1/2"	Tape Open Reel	180		
1/2"	Tape Cartridge	100-320	425-640	1280-2560
1/4"	Tape Cartridge	50-1350	2100	10000
.15"	Tape Cassette	20-160	410-600	1200
4mm	HS Cartridge	2000	4000	8000
8mm	HS Cartridge	300-2300	5000	
12.65 mm	HS Cartridge		20000	35000 +
19 mm	HS Cartridge		10000	

Sam Cheatham — Storage Technology Corporation

Who / What is
X3B5

TECHNICAL COMMITTEE X3B5 DEVELOPS

PROPOSED STANDARDS FOR THE

INTERCHANGE OF DATA BY DIGITAL

MAGNETIC TAPE FOR COMPUTER

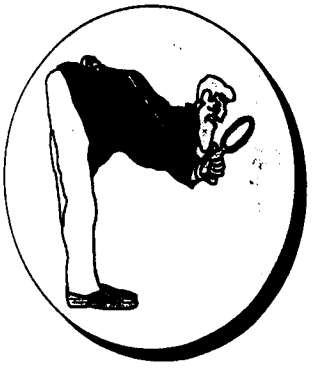
PERIPHERAL APPLICATIONS

Sam Cheatham — Storage Technology Corporation

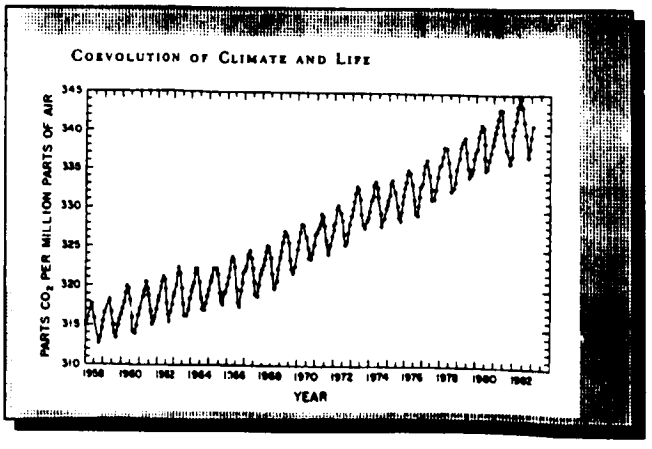
Data Management in NOAA

William M. Callicott

NOAA/NESDIS
FB4 Room 3316 OSD/5
Suitland, MD 20233



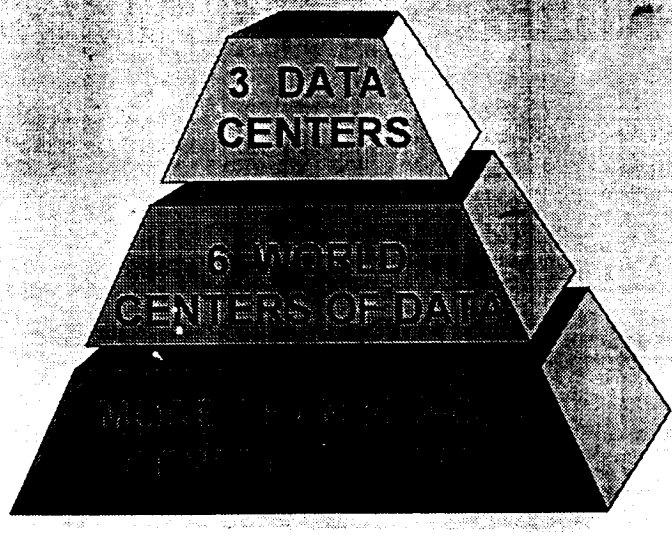
"DATA IS THE PRODUCT OF INVESTIGATION"



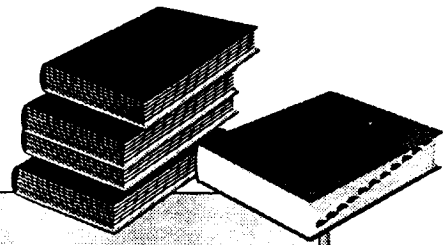
SOMETIMES, DATA CAN HAVE SUBSTANTIALLY MORE IMPACT THAN ITS ORIGINAL USE

(C.D.KEELING, SCRIPPS INSTITUTION OF OCEANOGRAPHY)

NOAA DATA MANAGEMENT

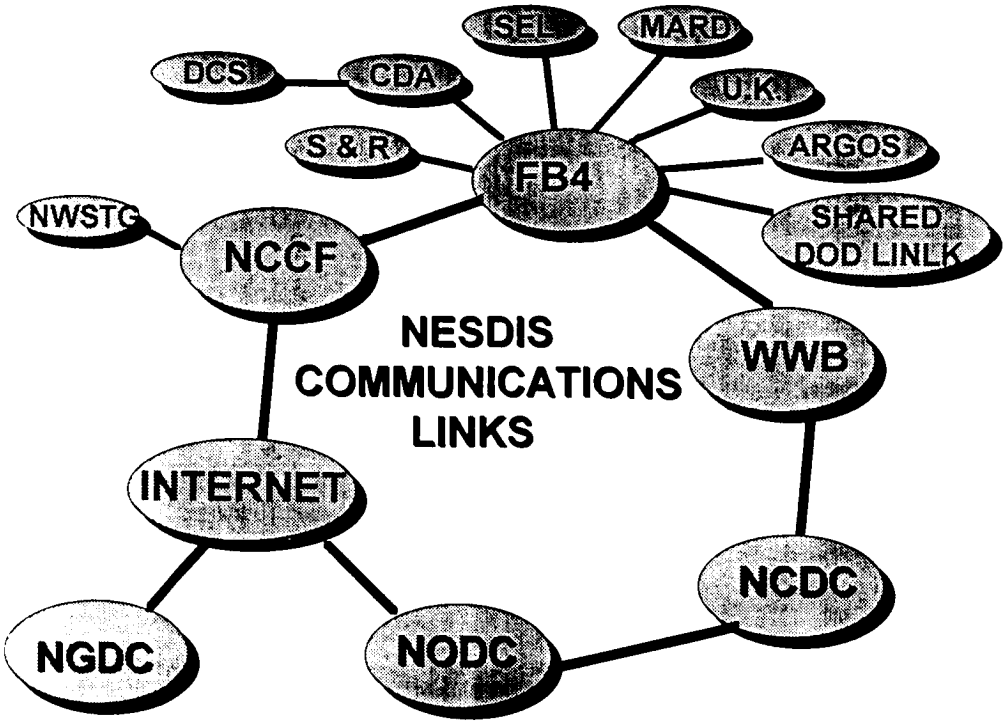


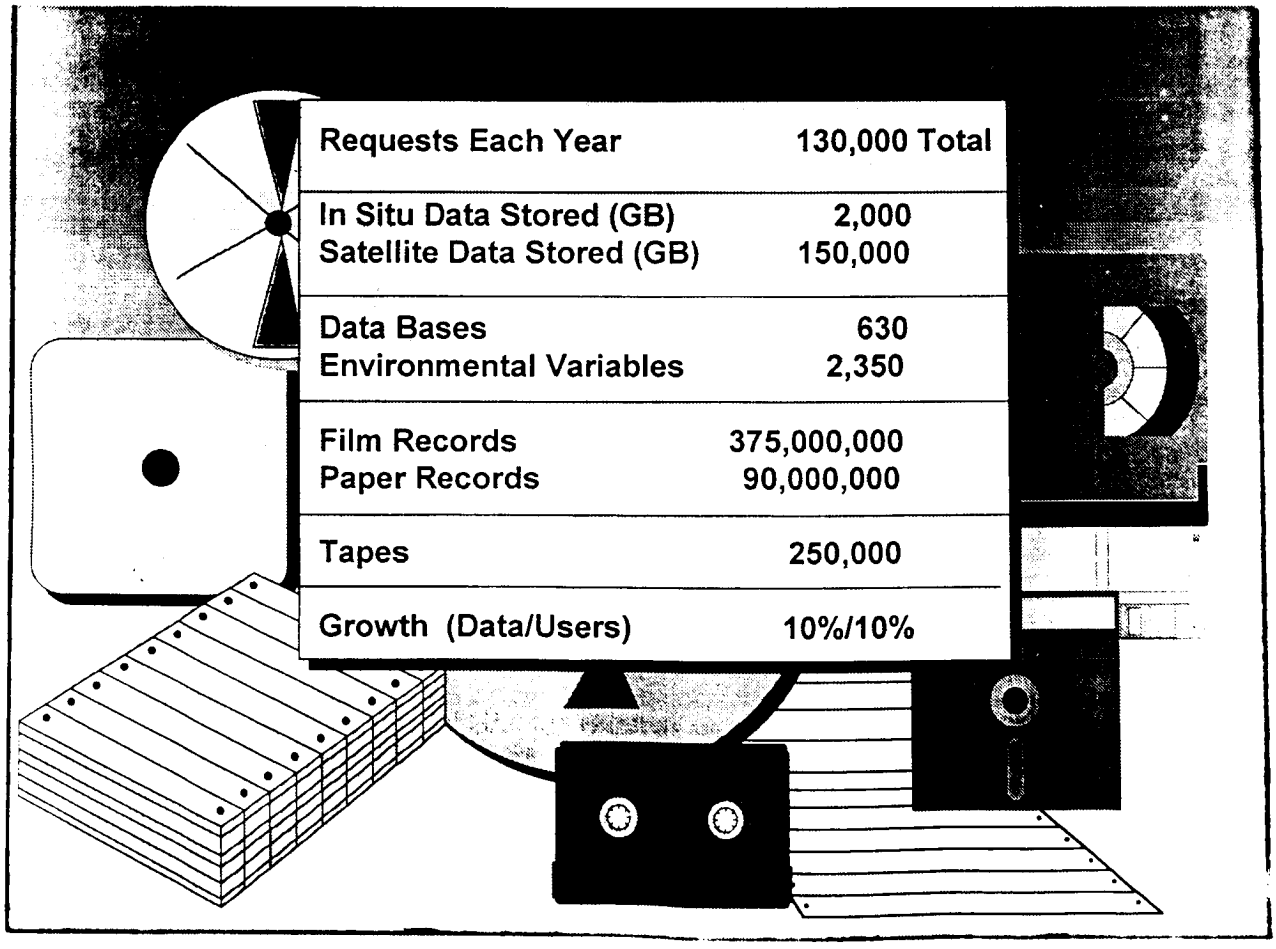
NOAA LIBRARY



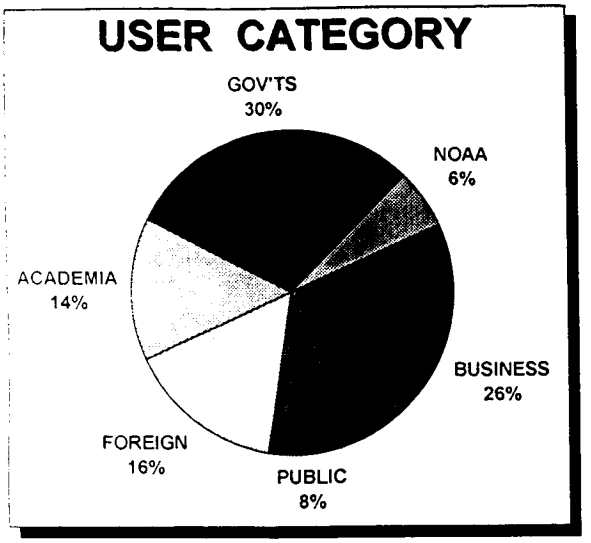
1,000,000 BOOKS
35,000 REPORTS
9,000 SERIAL TITLES
1,500 JOURNAL SUBSCRIPTIONS
1,000 RARE BOOKS
25,000 USER CONTACTS

CENTRAL LIBRARY IN ROCKVILLE, MARYLAND
REGIONAL LIBRARIES IN MIAMI AND SEATTLE

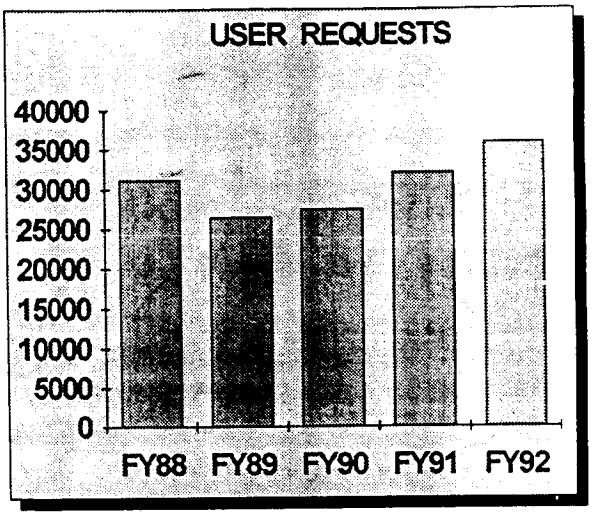




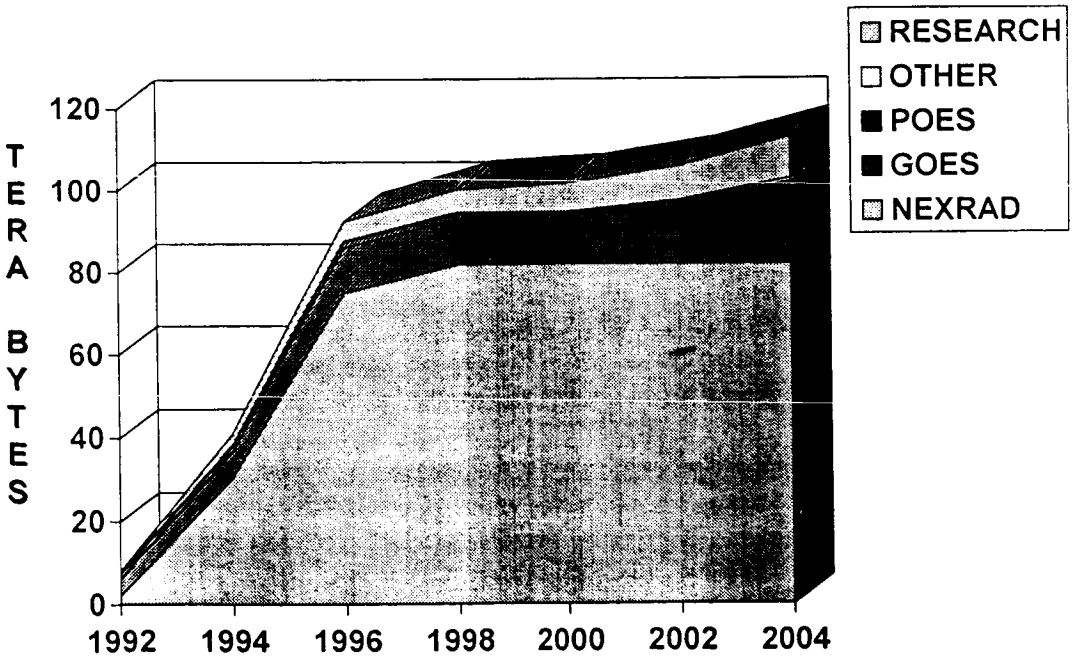
Requests Each Year	130,000 Total
In Situ Data Stored (GB)	2,000
Satellite Data Stored (GB)	150,000
Data Bases	630
Environmental Variables	2,350
Film Records	375,000,000
Paper Records	90,000,000
Tapes	250,000
Growth (Data/Users)	10%/10%



PROFILES OF CENTER ACTIVITY



ANNUAL DATA VOLUMES



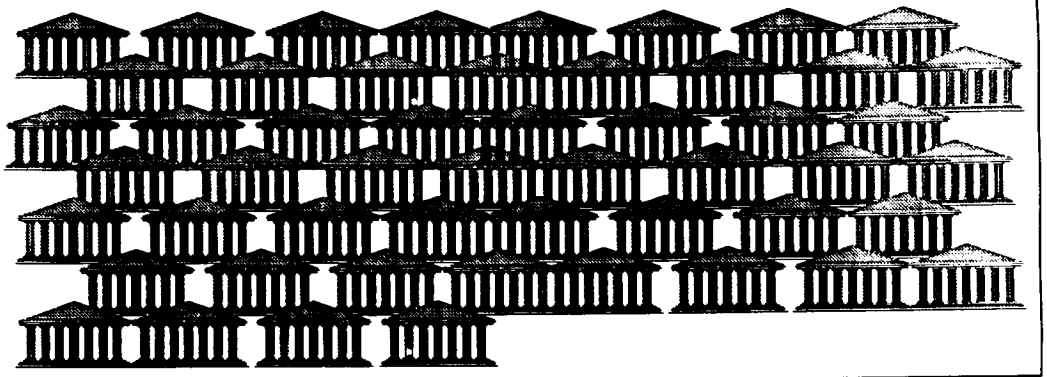
ANNUAL LIBRARY OF CONGRESS UNITS

1980 (1)

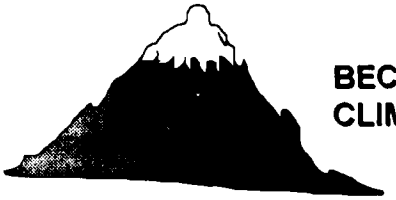
1990 (2)

1996 (8)

2002 (52)

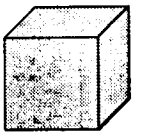


WHY SAVE THE DATA???



BECAUSE IT'S THERE...WHY PEOPLE CLIMB MOUNTAINS

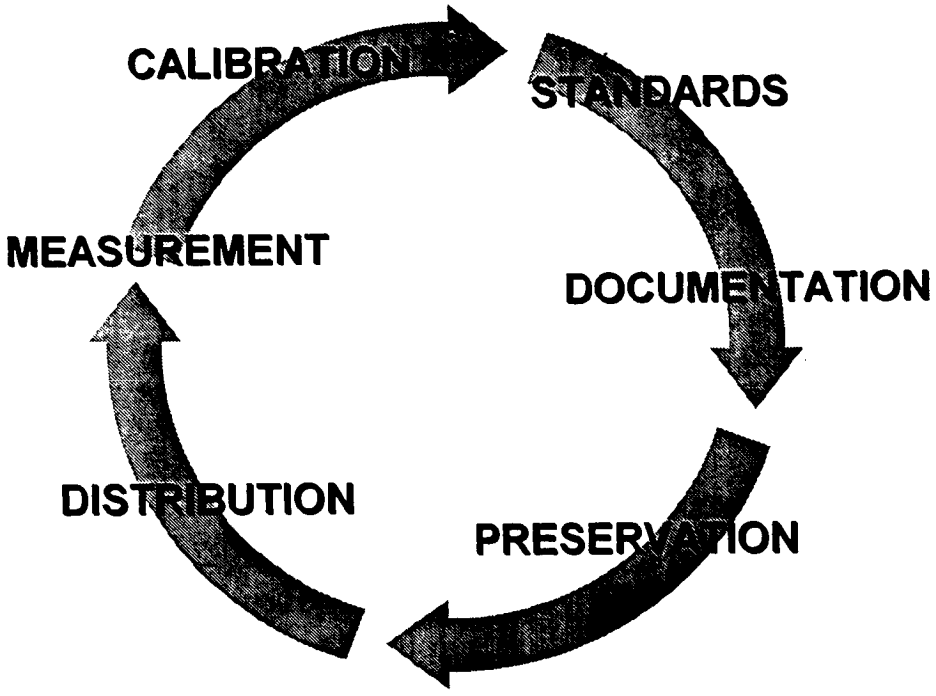
BECAUSE THE PROGRAM COST SO MUCH IN THE FIRST PLACE...WHY PEOPLE HAVE SO MUCH JUNK STORED AWAY



A NEW TECHNOLOGY MAKES IT FEASIBLE TO COST EFFECTIVELY SAVE THE DATA

REMEMBER, DATA IS THE PRODUCT OF INVESTIGATION....WE DON'T WANT TO THROW AWAY THE EVIDENCE

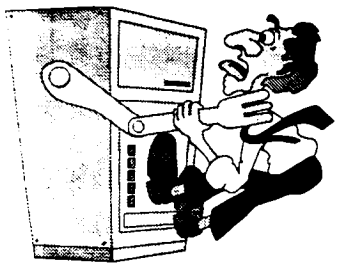
CRITICAL FACTORS FOR MANAGING DATA



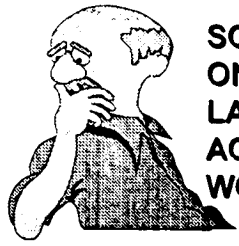
TECHNOLOGY LEVERAGING

- LEVERAGED BY THE CONSUMER MARKET
  - 4mm DAT.... Audio
  - 8mm EXABYTE... Video Camcorder
  - 1/4 inch QIC... Audio
  - 1/2 inch video... VHS video
  - 5 1/4 inch CD ROM... Audio CD
  - 5 1/4 inch read/write optical... Audio CD
  - 9 to 14 inch optical... Laserdisk video
- LEVERAGED BY INDUSTRIAL APPLICATIONS
  - 19mm Helical scan... Broadcast video
  - DCRSi 1 inch ... Instrument recording
  - ICI Optical Tape... Woolen dye industry
- MADE FOR COMPUTERS
  - 7 and 9 Track Computer Tape
  - 3480 and 3490 Cartridge Tape
  - Hard and Floppy Disks

CONFLICTING VIEWS ON MEDIA REQUIREMENTS

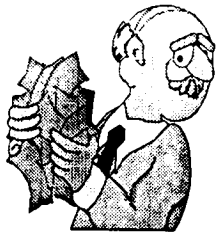


OPERATIONS - LEAST COST, CAN KEEP UP WITH THE FLOW



SCIENTIST - MINIMAL COST, ON-LINE RANDOM ACCESS, LARGE CAPACITY, FAST ACCESS, WORKS ON MY WORKSTATION

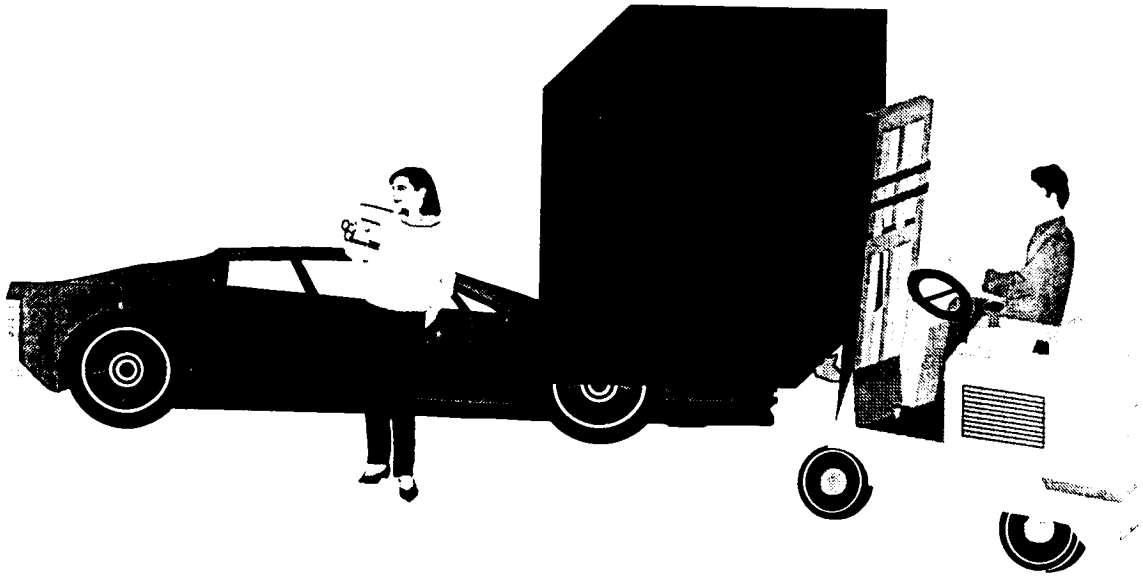
DATA MANAGER - LEAST LOGISTICS, LONG SHELF LIFE

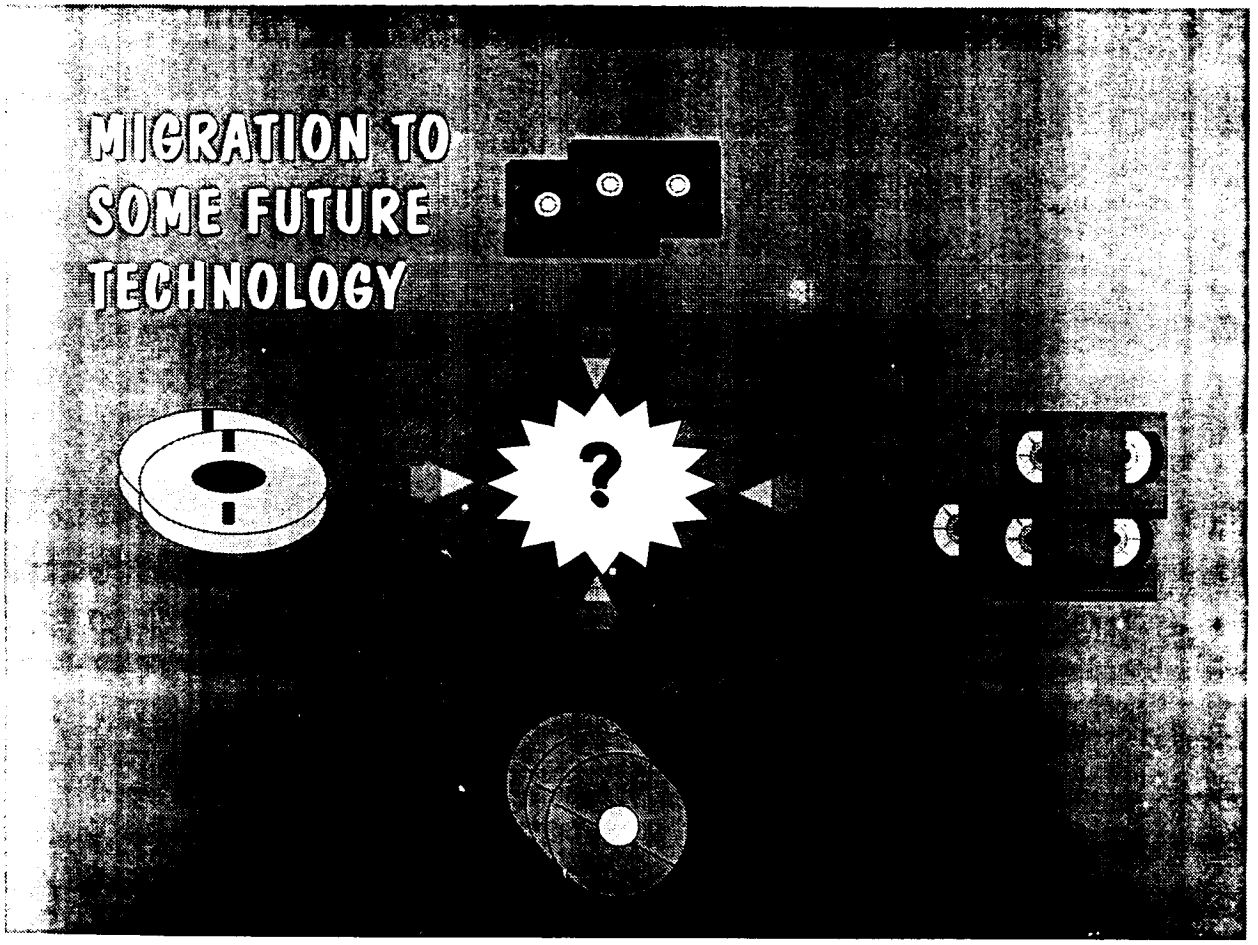


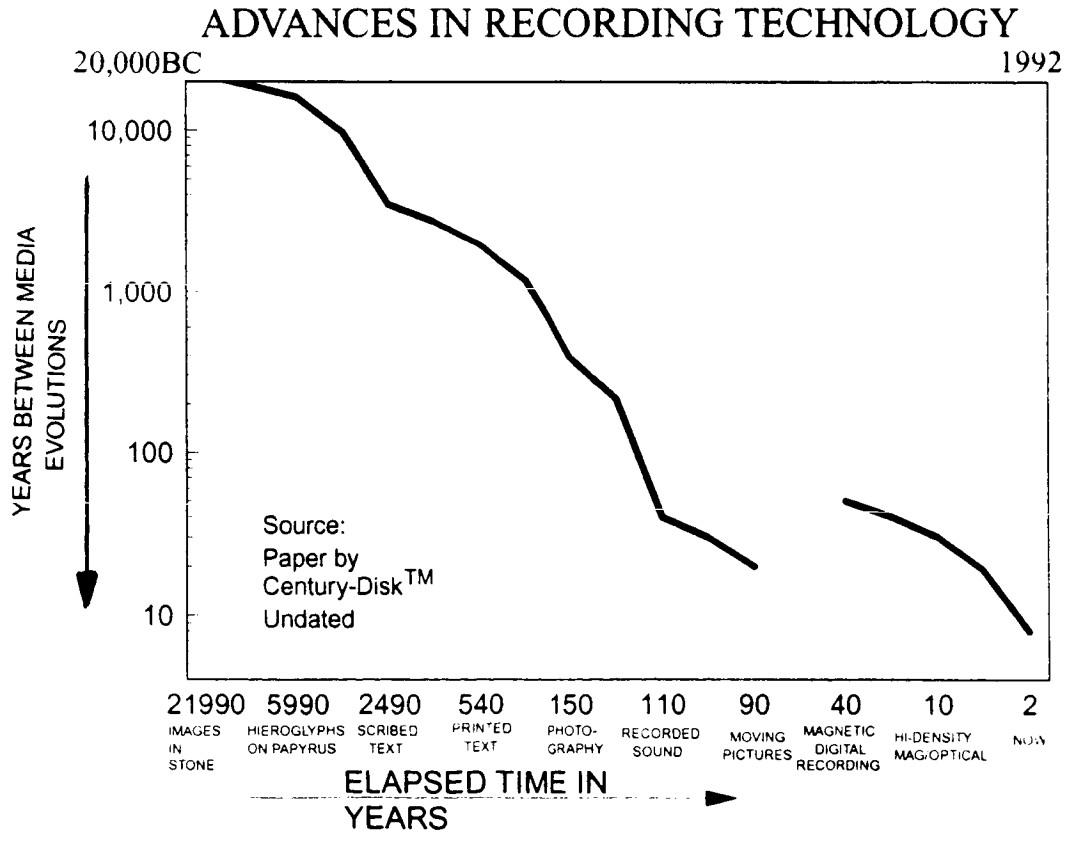
FINANCIAL MANAGER - NO COST

BUREAUCRAT - WHAT, ME WORRY?, NOT ON MY WATCH

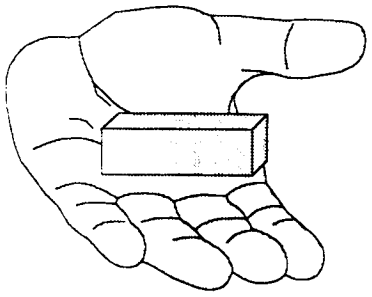
IT'S GOTTA FIT!




MIGRATION TO
SOME FUTURE
TECHNOLOGY
?

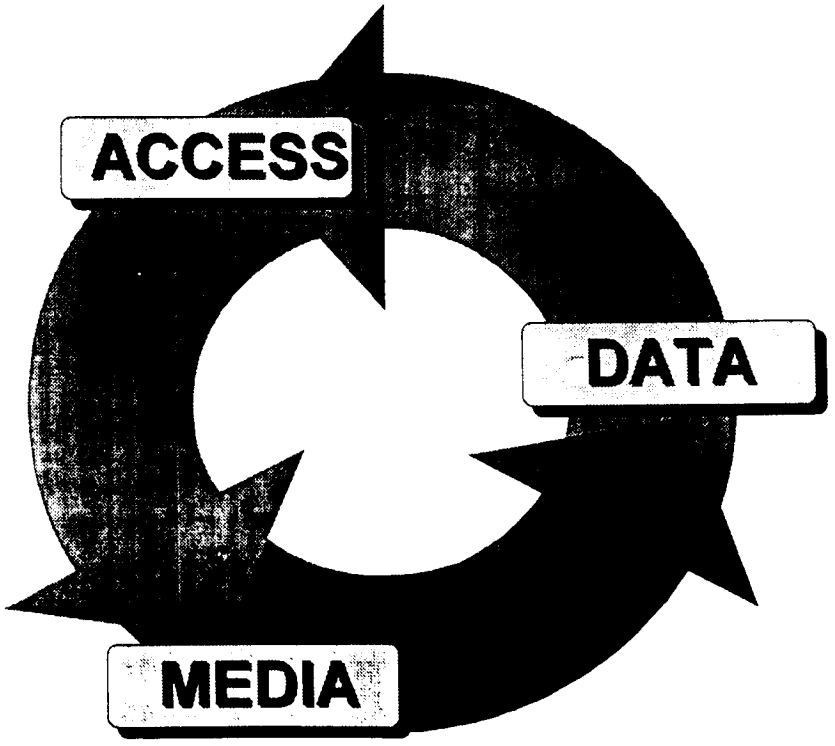

ADVANCES IN RECORDING TECHNOLOGY
20,000BC
1992
10,000
1,000
100
10
YEARS BETWEEN MEDIA EVOLUTIONS
Source:
Paper by
Century-Disk™
Undated
21990
IMAGES IN STONE
5990
HIEROGLYPHS ON PAPYRUS
2490
SCRIBED TEXT
540
PRINTED TEXT
150
PHOTO-GRAPHY
110
RECORDED SOUND
90
MOVING PICTURES
40
MAGNETIC DIGITAL RECORDING
10
HI-DENSITY MAG/OPTICAL
2
NOW
ELAPSED TIME IN YEARS

LIKE TO HAVE



SMALL AND PORTABLE
25 GB OR GREATER CAPACITY
XFER AT 50 Mbps OR FASTER
SYSTEM INDEPENDENT STANDARD
WORKSTATION COMPATIBLE
> 20 YEAR SYSTEM LIFE
DYNAMIC ACCESS
REPROCUCIBLE
"ROBOTICABLE"
< $0.50 PER GIGA BYTE

NOW!



Analysis of the Data and Media Management Requirements at the NASA National Space Science Data Center

Ron Blitstein

Hughes STX Corporation
4400 Forbes Boulevard
Lanham, MD 20706

TEXT WAS NOT MADE AVAILABLE FOR PUBLICATION

ACCESSING EARTH SCIENCE DATA FROM THE EOS DATA AND INFORMATION SYSTEM

Kenneth R. McDonald
Sherri Calvo

NASA Goddard Space Flight Center
Code 902.1
Greenbelt, MD 20771

INTRODUCTION

NASA's Earth Observing System (EOS) is designed to support the interagency Global Change Research Program through its Scientific Research Program (EOSSRP), Space Measurement System (EOSSMS), and Data and Information System (EOSDIS). The EOSDIS is responsible for the mission and instrument planning, scheduling, and commanding associated with EOS data acquisitions, the routine production, archive, and distribution of EOS data products, and the access to correlative data that may be archived by external data systems and organizations. The concepts related to the functions, architecture, and services of the EOSDIS that have emerged and are evolving are a direct consequence of the characteristics of both the scientific investigations associated with global change research and the community that is conducting those investigations.

A number of factors distinguish global change research from other scientific programs and endeavors. It is interdisciplinary, including studies in all of the Earth sciences and investigations of the interrelationships of different Earth processes. It is not limited to environmental science but also encompasses the analyses of the socioeconomic impacts of global change and the environment's response to human activities. Global change research requires massive sets of geophysical observations from numerous sources over the longest time periods available. The data sources include remotely sensed and in situ observations and predictions from numerical models and analyses. Because of the data demands of global change research, the program must integrate the existing collections of observations that are held by a variety of agencies and organizations with the future acquisitions of the EOSSMS and other programs. The numerous Federal agencies and research institutions involved in the Global Change Research Program are further evidence of the scope of the effort.

The purpose of this paper is to present an overview of the EOS Data and Information System, concentrating on the users' interactions with the system and highlighting those features that are driven by the unique requirements of the Global Change Research Program and the supported science community. However, a basic premise of EOSDIS is that the system must evolve to meet changes in user needs and to incorporate advances in data system technology. Therefore, the development process which is being used to accommodate these changes and some of the potential areas of change will also be addressed.

EOSDIS Program Requirements

Archive Contents

The EOSDIS will hold most of the Earth science data and data products from NASA activities and other data required for the production and effective use of these data. It will hold all of the data and data products from the EOS Space Measurement System and precursor missions.

The majority of NASA's heritage Earth science data will be migrated to EOSDIS and where it is not, EOSDIS will provide pointers and an access path to it. The current projection for the daily EOSDIS data rate by the end of this century is on the order of a terabyte per day, with a total archive of about two petabytes. EOSDIS will also contain metadata and browse products of its holdings, a software library of data production and analysis algorithms and tools, and a documentation library.

User Community

The EOSDIS user community is as extensive and diverse as the disciplines from which it is drawn. The number of users involved directly in Earth science research are estimated to be as many as 10,000 and when the education, applications, and government users are included the number grows to 100,000 or more. The characteristics and therefore the requirements of the users will also vary greatly. They will access the system from workstations, personal computers, and terminals. Some will use the system several times a week while others will only log on occasionally. Their experience level with on-line data systems and with the data itself will range from novice to expert.

Services

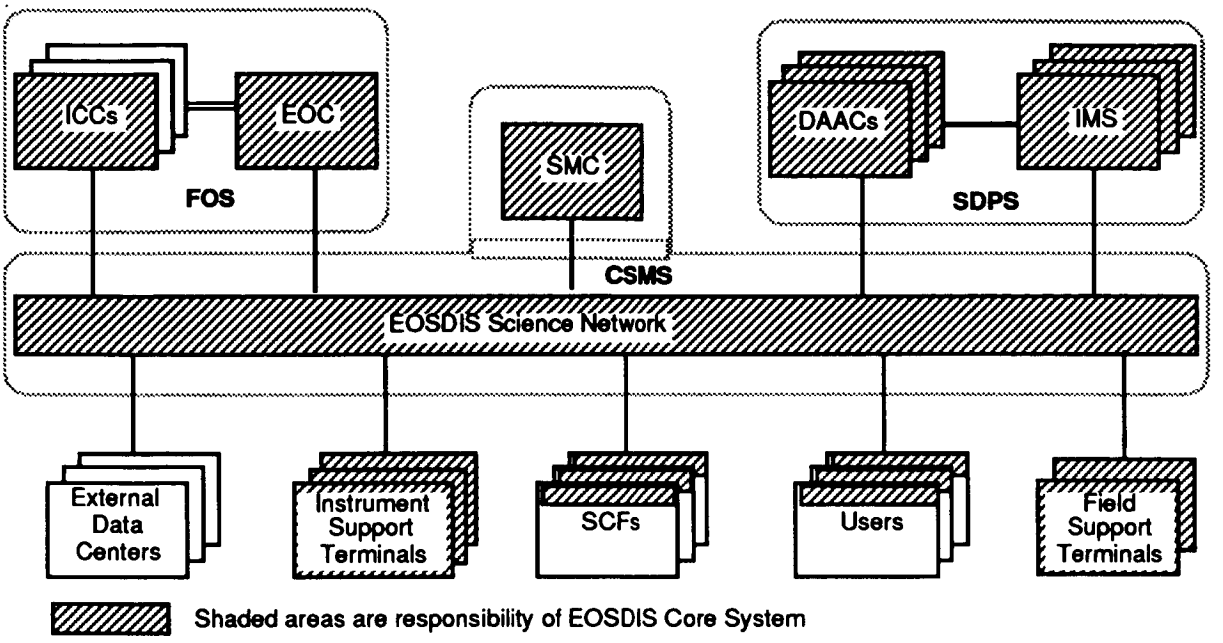
The EOSDIS will provide an end-to-end set of data production, management, and distribution services for the supported science community. It will supply the systems to perform the command and control of the EOS satellites and instruments. It will provide or augment the facilities that the science community will use to develop the algorithms that process the EOS data. The systems that will perform the routine generation of EOS products and the creation of special products are part of EOSDIS. The system will perform the archive and distribution functions for the EOS products, including those auxiliary data that are used in their generation. The information management services of the EOSDIS will provide the users with the capability to locate and order any of the data in the archives and will provide access to Earth science data held at external data systems through its interoperability with the interagency Global Change Data and Information System. Finally, the EOSDIS will provide the connectivity between its different elements and the overall management of its resources.

EOSDIS Conceptual Architecture

The process of defining the requirements of the EOSDIS has included the performance of a Phase A Conceptual Design Study and two Phase B Preliminary Design and Resource Estimates. The results of these analyses formed the basis of the conceptual architecture of the EOSDIS (Figure 1) which was included in the Phase C/D Functional and Performance Specification for the EOSDIS Core System (ECS). The architecture divides the EOSDIS into three segments and shows the interfaces between the segments and the external elements, shown as the Science Computing Facilities (SCFs) at the users' home facilities, the external data systems, and the users. The Instrument Support Terminals (ISTs) and the Field Support Terminals (FSTs) represent specialized interfaces that will be developed to support instrument operations and field campaigns, respectively.

The Flight Operations Segment (FOS) manages and controls the EOS platforms and instruments through the EOS Operations Center (EOC) and one or more Instrument Control Centers (ICCs). The ICCs will be used to schedule and command the more complex observatory instruments. The EOC will perform these functions for the survey instruments and perform the overall coordination of the platform and instrument operations. The Communications and System Management Segment consists of the EOSDIS Science Network (ESN) and the System Management Center (SMC). The ESN is responsible for the internal communications between the EOSDIS elements and the SMC monitors the overall resource usage of the system.

Figure 1 - EOSDIS Conceptual Architecture



The third segment of the core system is the Science Data Processing Segment (SDPS) which is composed of three functional elements. The Product Generation System (PGS) provides the systems and software to generate the higher level data products from the EOS observations. The Data Archive and Distribution System (DADS) stores the EOSDIS data products and auxiliary data and fills data requests. The third element of the SDPS is the Information Management System (IMS) which is the users' interface to all data and services of the EOSDIS. The IMS manages and provides the users with access to all of the information required to search, select, and order any of the EOSDIS data products and to construct and submit requests for data acquisitions and standard data processing options.

The SDPS is shown as a distributed system with the PGS and DADS elements coupled to form a Distributed Active Archive Center (DAAC) with a corresponding IMS element at each DAAC (the IMS element is somewhat arbitrarily shown outside of the DAAC to depict its system-wide interface to the EOSDIS in addition to its local information management function). This distribution was motivated by several of the Program requirements. First, to centralize a data collection of the size and scope of EOSDIS would be to create a huge, monolithic archive of unprecedented dimensions. The fear that such a center would be unresponsive to the evolving requirements of the user community and the specific needs of the individual user, was quickly pointed out by the EOSDIS science advisory panel. Secondly, building a centralized EOSDIS would diverge from the existing distribution of Earth science data where the archives are held at centers of discipline expertise. Instead, the distributed EOSDIS elements will augment these centers as shown in Table 1. Another advantage that distribution offers is the ability to easily expand the system by the addition of new active archive centers. This option has already been exercised by the addition of the DAAC at Oak Ridge National Laboratory. Finally, the scope of global change research dictates that much of the data required by the science investigators will be held outside of EOSDIS by other agencies and institutions. The technical challenges of providing access to distributed data will have to be addressed.

Table 1 - EOSDIS Distributed Active Archive Centers

Center	Heritage Systems	Areas of Interest
Goddard Space Flight Center	NASA Climate Data System, Pilot Land Data System, Coastal Zone Color Scanner Data System	Upper atmosphere, atmospheric dynamics, global biosphere, and geophysics;
Langley Research Center	ERBE processing	Radiation budget, aerosols, and tropospheric chemistry
EROS Data Center	Global Land Information System, Landsat processing	Land processes imagery
University of Alaska - Fairbanks	Alaska SAR Facility System	Sea ice, polar processes imagery (SAR)
University of Colorado	Cryospheric Data Management System	Cryosphere (non-SAR)
Jet Propulsion Laboratory	NASA Ocean Data System	Ocean circulation and air sea ice interaction
Marshall Space Flight Center	WetNet	Hydrologic cycle
Oak Ridge National Laboratory	Based on CDIAC , ARM , & Surface Water Survey Data Center data management procedures	Biogeochemical dynamics

Access Characteristics

The EOSDIS Program goals and the science user community input have been the basis for the definition of the access requirements and the EOSDIS element designs. The two elements that are directly involved in providing this access are the IMS and the DADS. The DADS data distribution must deliver the data with a response that supports the scientific research process and in the desired format. All of the IMS functions are directed at providing the science users with a functionally complete and robust interface to the data and services of the system.

The IMS design process has identified those functions that the users will require to gain access to the information, data, and services of the EOSDIS. The first function is a user interface that supports a dialog between the user and the system and conveniently supports the interchange of information. The interface to all functions should follow a consistent style and must provide an appropriate level of assistance to guide the user through the session.

The information search functions of the IMS must be capable of answering a wide variety of user queries. At the highest level, the users will want to identify data sets that can be used in their particular area of research. The users will enter information that describes their research such as discipline, parameter, and area of interest into the system, and a directory function will identify those data sets that meet the criteria. The directory will provide overview descriptions of the data sets which may be within the EOSDIS or held by external systems. Users will typically need more specific information to actually order data from the system. An inventory function will identify instances of a data set (granules) that specifically meet the users' data needs. To adequately narrow the search, this function may require users to enter additional criteria such as the specific time and location of interest and data quality requirements. In addition to identifying data sets and data granules, the IMS will perform additional functions that assist in the selection and use of the data. Visual aids will include the display of browse data, which are subset or subsampled versions of the data that have been

produced to allow the user to preview a data granule, and coverage maps showing the geographical location and areal extent of one or more data sets or data granules. Another information access function, referred to as the guide, will provide the capability to access a variety of text and other supplementary information to assist in the selection and use of the data.

The IMS functions also include the generation and delivery of data and service requests to the other elements of the system. Through the user interface of the IMS, the user will be able to construct three types of data requests. For those instruments that have a variable duty cycle, the user can construct a request for a future acquisition. The user interface will also present distribution and delivery options which together with the results of an information search allow the user to construct a request for archived data products. The third type of request is for data processing and will allow the user to select from a set of standard options. In addition to data requests, the IMS will provide the interconnectivity with cooperating external Earth science archives where agreements and standard procedures have been established. Finally, throughout the IMS, an automated "Help" service will be available to provide as much on-line assistance as possible, tailored to the users' needs, and complementary to the user services functions.

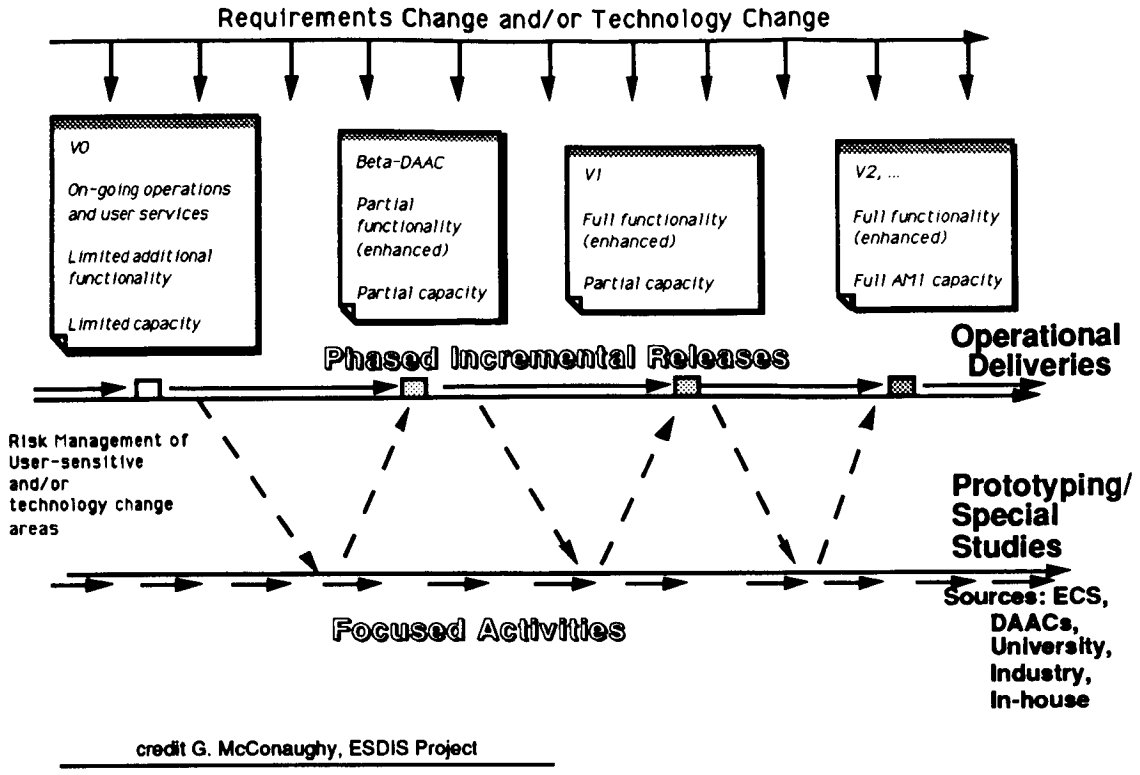
The concepts and conceptual designs of the EOSDIS access functions reflect the nature and characteristics of the Global Change Research Program and the Earth science research community. Since the program is interdisciplinary and the community ranges from students in Earth science to the experienced investigator, the system must provide sufficient background information in addition to the minimal search and order functions. Because the users will have different levels of experience with this system and with data systems in general, the help facility must be tailorable to the particular user. The users' local environment will also vary, with some relying on terminals to access the EOSDIS and others using powerful workstations allowing enhanced visualization and other capabilities.

The different characteristics of the scientific investigations will also place widely divergent demands on the system. Some will be global in scale with long time ranges of interest, while others will be site or phenomenon studies with much smaller data requirements. The nature of the study and the preferences of the user will determine the mode in which the user interacts with the system and will be reflected in the degree of coupling between the EOSDIS and the user's local analysis systems. At one extreme, the two are decoupled and the investigator uses the EOSDIS to select and deliver large bulk orders of data that will be managed and processed locally. This corresponds to a "personal library" model and typically relaxes the response time required of the EOSDIS. At the other extreme, the two systems will be completely integrated and the local analysis system will rely on the EOSDIS to manage and supply its data resources. The integrated model may require that smaller data volumes be delivered with each request but also implies significantly more frequent and faster deliveries. This impacts the data storage methods at the DADS and the required network performance.

Development Approach

The EOSDIS Program requirements and goals have had a significant impact on the approach that has been taken to define, design, implement, and operate the system. An overview of the approach is shown in Figure 2. In recognition that the definition and development cycle of such a large, complex system would be a lengthy process and to take advantage of the experience base that exists from the development and operation of the heritage systems at the identified DAAC sites, the concept of EOSDIS Version 0 emerged. The general goal of Version 0 is to provide "lessons learned" for EOSDIS through the integration and augmentation of the capabilities at the DAACs and the development of selected prototypes.

Figure 2 - EOSDIS Development Cycles



The Version 0 effort consists of DAAC development activities and a number of system-level tasks. Each DAAC is responsible for the design and implementation of the product generation, data archive and distribution, and local information management functions that are required to support their existing data and the data from precursor missions. The development is governed by a set of requirements and an architecture and operations concept that have been developed by the team of EOSDIS Project and DAAC system engineers. In addition to satisfying the local data and information system requirements, this effort provides the mechanism to jointly address the program goals and objectives and a forum to exchange information on technology and approaches.

This system engineering team has also defined the objectives and scope of the information management, networks, and data format system-level tasks. The Version 0 IMS system-level task is working with the DAAC teams to implement an operational prototype that provides a cross-DAAC data search and order capability. This prototype will provide a preliminary version of each of the information access functions and the data access function which allows users to place requests for archived data products. The Version 0 formats task has evaluated many of the data formats that are currently being used by the science community and is working with the DAACs to reach consensus on the standard formats that will be used in the Version 0 timeframe. The networks task is analyzing the bandwidth requirements to perform the Version 0 on-line access and distribution functions and is responsible for the enhancement of user and inter-DAAC connectivity.

The Version 0 activity is only the first step in the EOSDIS development process. The EOSDIS Core System Project will develop the Beta-DAAC and subsequent versions of the system through a series of phased, incremental releases. At the same time, the project will be supporting a series of prototypes and special studies that will focus on particular topics that address user-sensitive areas and emerging technologies. These prototypes will be performed by the ECS contractors, other industry participants, the DAACs, academic institutions, or

members of the in-house staff. The results of these activities will continuously feed into the implementation of the operational systems. The science community will be constantly involved in the specification and evaluation of the prototypes and releases.

Conclusions

The breadth and complexity of the Program goals and the size and diversity of the Earth science research community place unique data and information access requirements on the EOSDIS which impact all aspects of its design and development. The distributed architecture is a reflection of the data volumes to be supported and the broad scope of participants, in terms of both institutions and investigators. The flexibility of the access functions and the wide range of capabilities that they must provide are necessary because of the differing needs of the individual science users. However, the greatest impact is on the overall approach that is being used in the development of EOSDIS.

The EOSDIS conceptual architecture and functional requirements describe the current understanding of global change research and the role to be played by the data and information system. However, there are many aspects of this process that are not yet well understood and it is believed that the process will evolve over the course of the program. The EOSDIS development plan recognizes and reflects these factors. More direct interaction with the users is required to more accurately define the current requirements and is a primary justification of the Version 0 effort. Technology is rapidly changing in many ways that will enable the science community to better pursue its research and that will potentially alter the requirements and characteristics of its access to EOSDIS. The intention of the prototyping plan is to give the users the opportunity to use and evaluate the technological advances. Finally, the delivery of the EOSDIS through a series of releases will allow the systems to evolve in response to changes in user requirements and data systems technology.

RECORDING AND WEAR CHARACTERISTICS OF 4 AND 8 MM HELICAL SCAN TAPES

Klaus J. Peter
Media Logic Inc.
310 South Street
P.O. Box 2258
Plainville, MA 02762

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

INTRODUCTION

Performance data of media on helical scan tape systems (4mm and 8mm) is presented and various types of media are compared. All measurements were performed on a standard MediaLogic model ML4500 Tape Evaluator System with a Flash Converter option for time based measurements. 8mm tapes are tested on an Exabyte 8200 drive and 4mm tapes on an Archive Python drive; in both cases the head transformer is directly connected to a Media Logic Read/Write circuit and test electronics. The drive functions only as a tape transport and its data recovery circuits are not used.

Signal to Noise, PW 50, Peak Shift and Wear Test data is used to compare the performance of MP (metal particle), BaFe (Barium Ferrite), ME (metal evaporated). ME tape is the clear winner in magnetic performance but its susceptibility to wear and corrosion, make it less than ideal for data storage. (See also : Corrosion of MP and ME Tapes, D. Speliotis and K. Peter, Journal of the Magnetics Society of Japan Vol. 15 Supp. S2 1991)

EXPERIMENTAL DATA

Fig. 1 shows PW50 performance comparison between MP, BaFe and ME tapes. PW50 is the pulse width in nano seconds measured at 50% amplitude of an isolated flux reversal; a narrow pulse width is required for high recording densities when as many pulses as possible are squeezed closely together. In addition, the write current was varied over a 6 mA range, roughly covering the broad peak of a saturation curve, to show how sensitive performance is to write current variations (see Fig. 2).

ME tape not only easily outperformed all other tapes but also showed the least sensitivity to write current variations, making it ideal from the drive designers' point of view. The reason for this is most likely the extremely thin magnetic layer and corresponding low SFD (Switching Field Distribution). PW50 typically was 115 nsec at 16 mA write current; at a head/media velocity of 3.8 m/sec, this translates to a PW50 of 43.7 um. Barium Ferrite came in second at 148 nsec (56.2 um), but showed the characteristic slope which all thicker coated media exhibited. This means that PW50 becomes worse as write current increases, making drive and head design more critical. HI 8 MP (fine grain MP) had a PW50 of 154 nsec (58.5 um), while regular MP was approximately 160 nsec (60.8 um).

Fig. 3 to 6 compares peak shift between types of tape; again, ME is the clear winner with 26.3 nsec peak shift while MP is second with 40 nsec. BaFe had the highest (worst) peak shift at 51.2 nsec which is counter to what one would expect based on PW50 and frequency response data. Peak shift was measured using a 00110011 repeating pattern at 4 MHz which means that the unshifted spacing between flux reversal pairs is 125 nsec. The reason for more than expected peak shift on BaFe tape is not clear and more work needs to be done. One possible explanation may be that the samples tested had more surface roughness (see fig. 7 HF modulation) and in effect increased the head/media spacing or that some as yet unidentified negative interactions are occurring between particles.

Signal to Noise ratio was measured up to 10 MHz to compare the potential suitability of various tapes for higher recording densities. The measurement is made at any spot frequency over a 10 KHz bandwidth; the bandpass filter used has a 30 dB per octave rolloff. The S/N contribution of the low noise read preamp is 0.5 nV/rt Hz; this represents 50 nV over 10KHz bandwidth, allowing an 80 dB dynamic range below a 500 uV signal which is well below media noise.

Fig. 8 is the S/N data for 8 mm tapes; ME again is the star performer, with HI-8 MP 6dB below ME at 9.5 MHz. BaFe is only 2 dB below HI-8 MP at high frequencies but performs worse at low frequencies due to lower output amplitude. Regular MP is last with about 12 dB below ME.

4 mm S/N data (Fig. 9) gives much the same picture except that no ME tape was available. S/N curves drop off faster at high frequencies partly due to a lower head/media velocity (16%) and possibly due to head performance. It is significant that despite lack luster low frequency S/N performance, BaFe appears ready to overtake all other tapes when the graph is extended beyond 10 MHz. No double layer BaFe tapes with better surface finish were available at the time the data was taken, but in the near future we expect to fill in this gap. The coated double layer MP tape used a low coercivity underlayer (gamma ferric oxide) and MP as upper layer; the performance appears to come close to ME when interpolated to 8mm data.

Signal to noise measurements are difficult to perform accurately on helical scan systems since only 16.7 mS for 8mm and 7.5 mS for 4mm is available to allow write/read recovery, set VCO frequency, allow RMS detector to settle and take reading. The repeatability of S/N measurements on our tester was within 1/4 dB. With a spectrum analyzer, the sweep must be triggered by the start of the read track and the bandwidth and sweep rate settings carefully chosen to obtain accurate S/N data.

The wear test data was obtained by making repeated passes over the same 1000 tracks of tape in a continuous write/read process, i.e. the pattern was rewritten each time and dropout errors counted. For dropout error measurement, an all 1's pattern is recorded; on read, the analog head signal is amplified and the peak amplitude of each positive and negative flux transition is compared to the TAA (track average amplitude), and if it is below 50% of TAA, a single bit error is generated. The dropout error block definition Bad/Good/Max allows the user to set the error block size or duration; for example, a setting of 1/5/180 means that an error block count is initiated when a single flux reversal is below the set threshold (50%) but a second error block is only counted after 180 bad flux reversals; after 5 good flux reversals, the bit error counter is reset. A setting of 125/16/250 means that an error block is counted only after 125 bad bits, and a second error block is counted after 251 bad bits; if 16 good bits occur, the bad bit counter is reset.

Fig. 10 shows a typical wear test result on 8 mm tape of 1000 passes over 1000 tracks. Due to burnishing effects of head and drum, the dropout error rate of a new tape will typically drop by as much as a factor of 10 after a few hundred passes. The 4 mm wear test shows a similar reduction of errors due to the burnishing effect of head/drum rotation (see fig. 11). Except for a few mechanical cartridge failures, most brands of MP and BaFe tapes endured the entire 20,000 passes they were subjected to without catastrophic failure, however some tape brands exhibited substantial error rate increases after a few thousand passes. We attributed this sudden rise in

error rates to a form of stiction caused by particles adhering to the head due to extreme smoothness of head and media interface (see fig. 12).

One type of tape which failed catastrophically after about 700 passes, is ME (see fig. 13). In this case, the thin evaporated metal coating just wears through much sooner than the thicker coated substrates and makes this type of tape not really suitable for applications requiring many passes. For video applications it may be good enough.

Another interesting phenomenon noticed on some brands of 4 mm tapes was termed "end of test problem". As fig. 14 through 18 shows, there is a continuous increase in large dropout errors from 81 to 454 over 40,000 tracks; the increase is concentrated at the end (right hand edge) of the error map. In fig. 17, this "end of test problem" is clearly visible since the error map was extended to 45,000 tracks which is beyond the original test. The reason for the bunching of errors at the end of a test must be explained by additional wear caused by stopping/reversing and rewinding of the tape.

It is interesting to note that no inherent advantage in wear characteristics were evident due to a smaller wrap angle (90 deg) of 4 mm versus 8 mm (180 deg). Could it be that the larger wrap angle provides a more stable air bearing? The raw uncorrected dropout error rate of 4mm is an order of magnitude higher than on 8mm when measured under the same conditions which is the reason an extra level of error correction is used in 4mm systems. This however does not imply that 4mm is in any way inferior, it only means that recoding density and error correction were optimized differently.

To summarize, in order to evaluate tape performance and compare apples and apples, a common reference point and test method needs to be established. Despite the complexities of helical scan recording and the availability of many types and brands of tape, it is possible to collect valid data which can be used to make logical choices for one's own application.

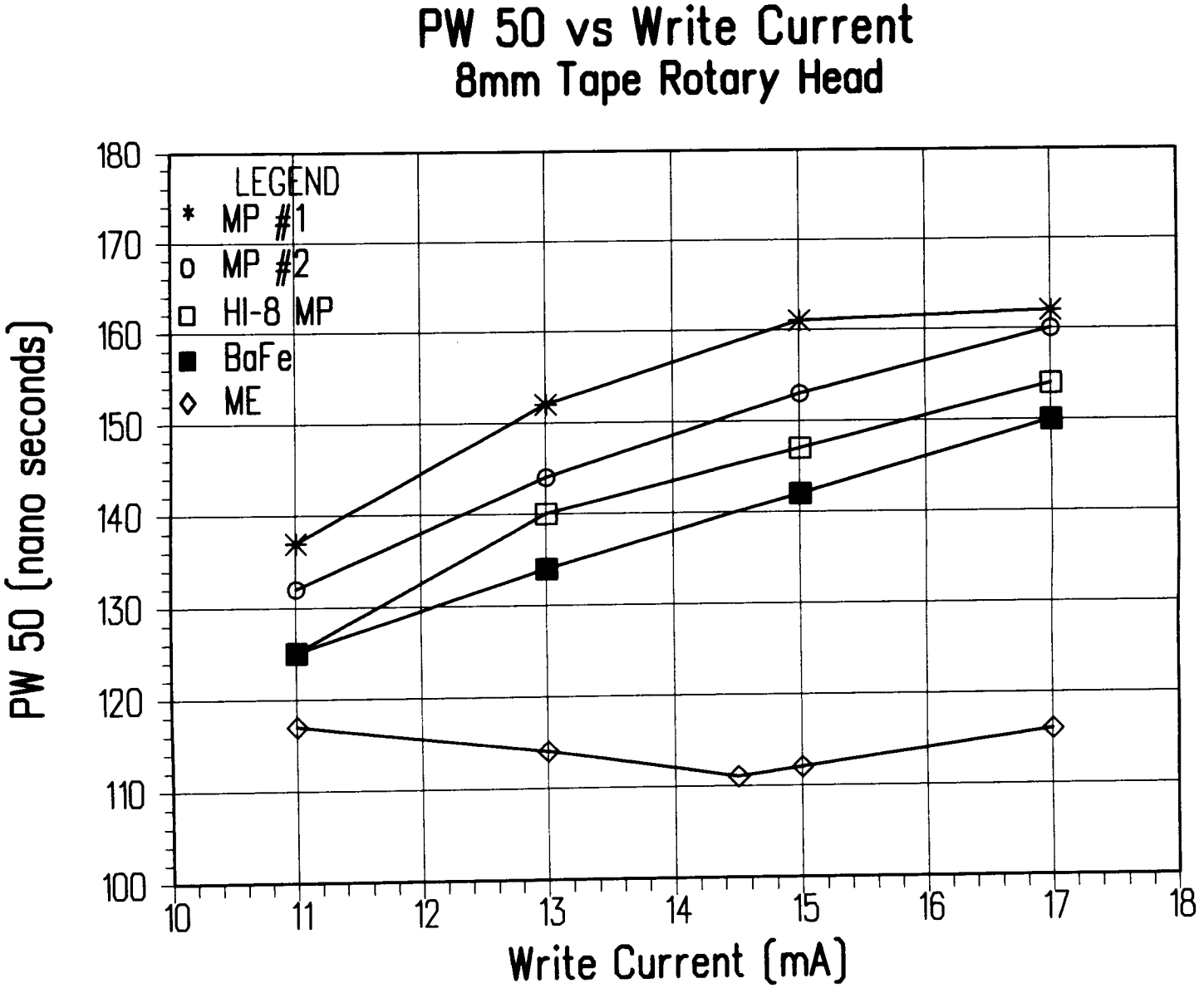

PW 50 vs Write Current
8mm Tape Rotary Head
LEGEND
* MP #1
o MP #2
□ HI-8 MP
■ BaFe
◇ ME
PW 50 (nano seconds)
180
170
160
150
140
130
120
110
100
10
11
12
13
14
15
16
17
18
Write Current (mA)

FIGURE 1

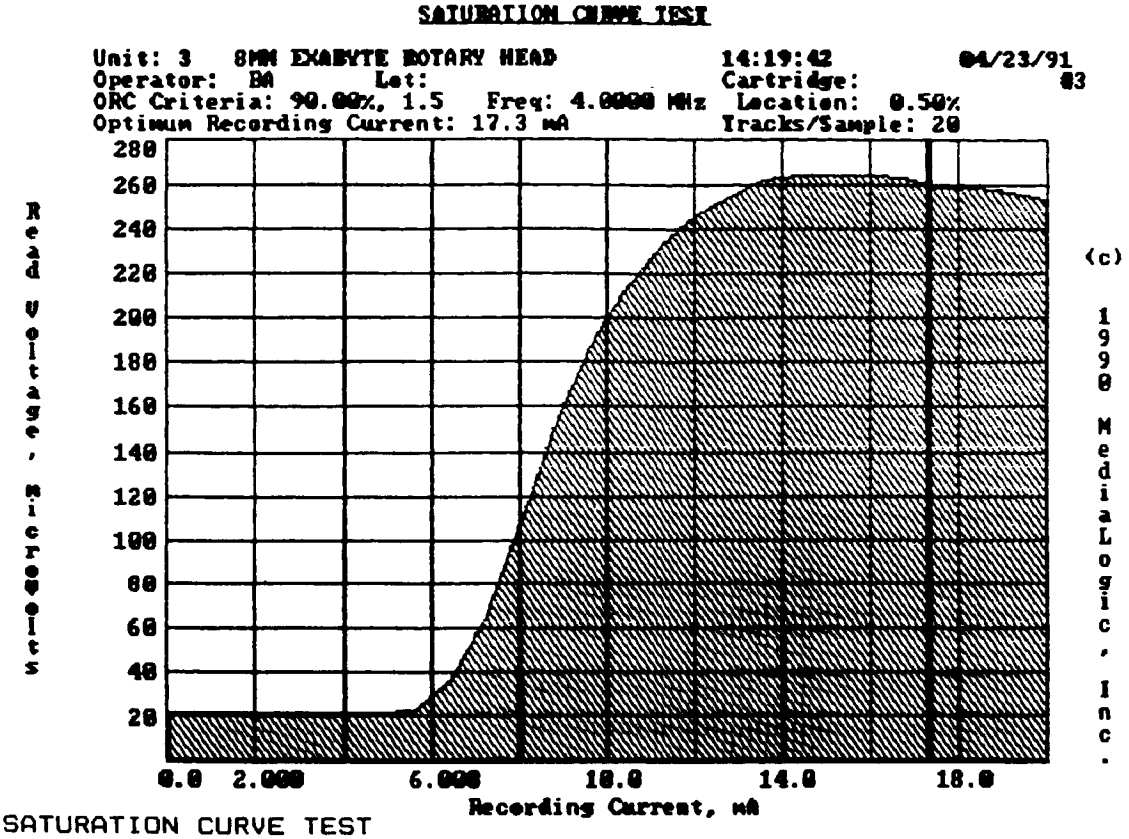


FIGURE 2

SATURATION CURVE TEST

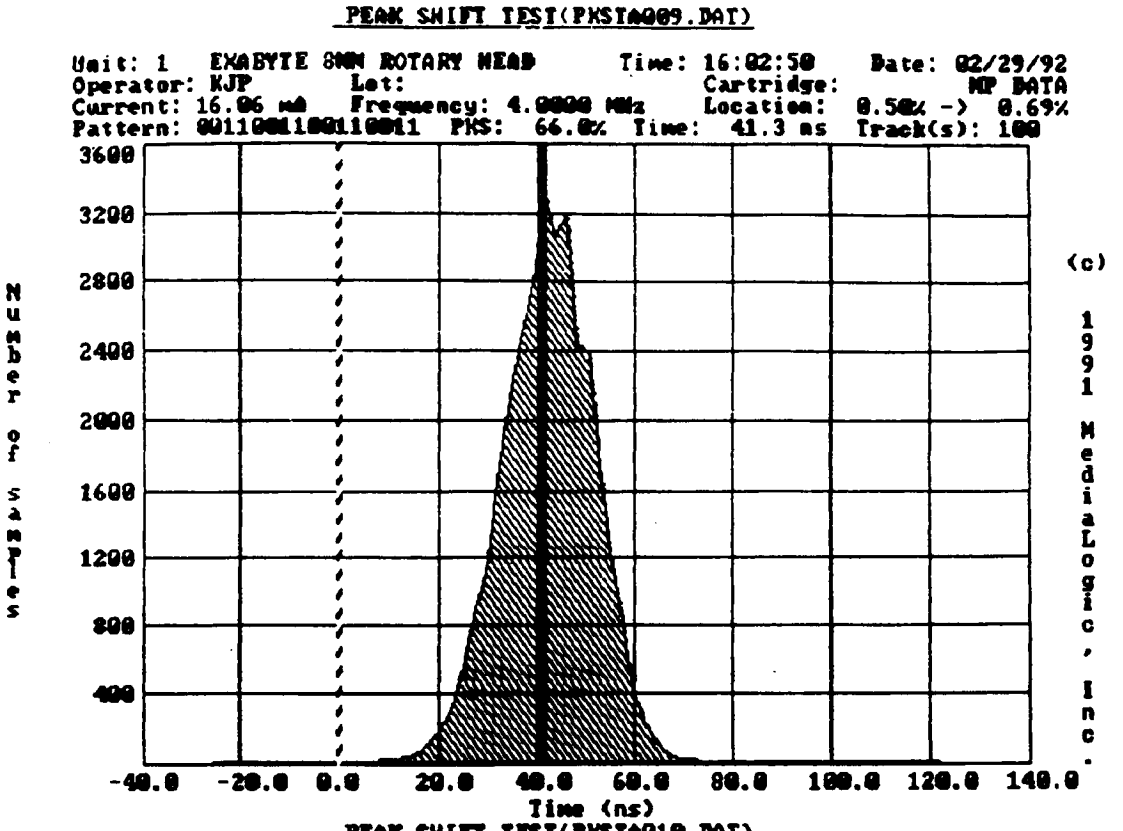
Unit: 3 8MM EXABYTE ROTARY HEAD 14:19:42 04/23/91

Operator: BA Lot: Cartridge: #3

ORC Criteria: 90.00%, 1.5 Freq: 4.0000 MHz Location: 0.50%

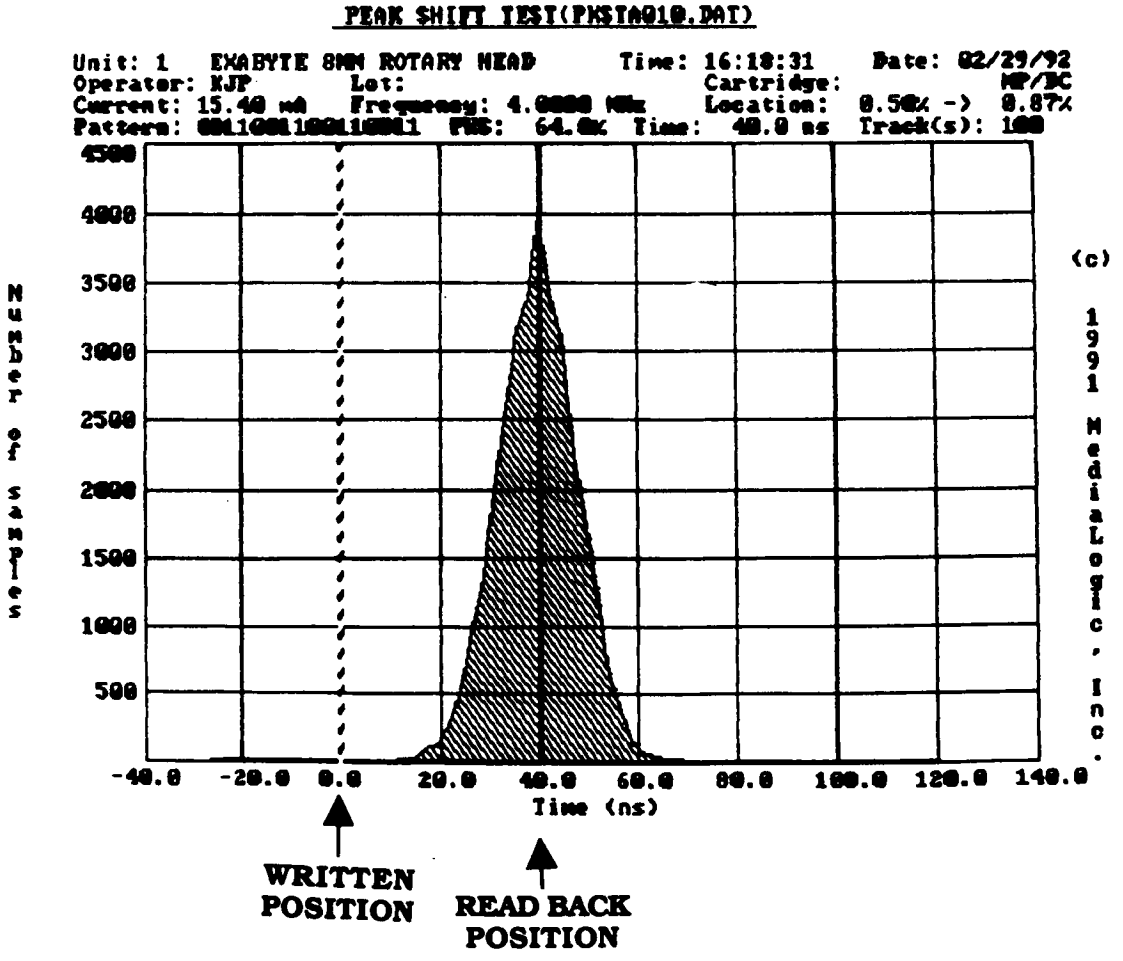
Optimum Recording Current: 17.3 mA Tracks/Sample: 20

0)	current =	0.00 mA,	read voltage =	21.1	microVolts
1)	current =	0.80 mA,	read voltage =	21.1	microVolts
2)	current =	1.60 mA,	read voltage =	21.1	microVolts
3)	current =	2.40 mA,	read voltage =	20.7	microVolts
4)	current =	3.20 mA,	read voltage =	21.1	microVolts
5)	current =	4.00 mA,	read voltage =	21.1	microVolts
6)	current =	4.80 mA,	read voltage =	21.1	microVolts
7)	current =	5.60 mA,	read voltage =	22.1	microVolts
8)	current =	6.40 mA,	read voltage =	34.6	microVolts
9)	current =	7.20 mA,	read voltage =	62.9	microVolts
10)	current =	8.00 mA,	read voltage =	106.4	microVolts
11)	current =	8.80 mA,	read voltage =	150.4	microVolts
12)	current =	9.60 mA,	read voltage =	187.1	microVolts
13)	current =	10.40 mA,	read voltage =	212.5	microVolts
14)	current =	11.20 mA,	read voltage =	232.5	microVolts
15)	current =	12.00 mA,	read voltage =	246.4	microVolts
16)	current =	12.80 mA,	read voltage =	255.0	microVolts
17)	current =	13.60 mA,	read voltage =	262.1	microVolts
18)	current =	14.40 mA,	read voltage =	264.3	microVolts
19)	current =	15.20 mA,	read voltage =	265.4	microVolts
20)	current =	16.00 mA,	read voltage =	265.4	microVolts
21)	current =	16.80 mA,	read voltage =	263.6	microVolts
22)	current =	17.60 mA,	read voltage =	259.6	microVolts
23)	current =	18.40 mA,	read voltage =	258.9	microVolts
24)	current =	19.20 mA,	read voltage =	256.8	microVolts
25)	current =	20.00 mA,	read voltage =	253.9	microVolts



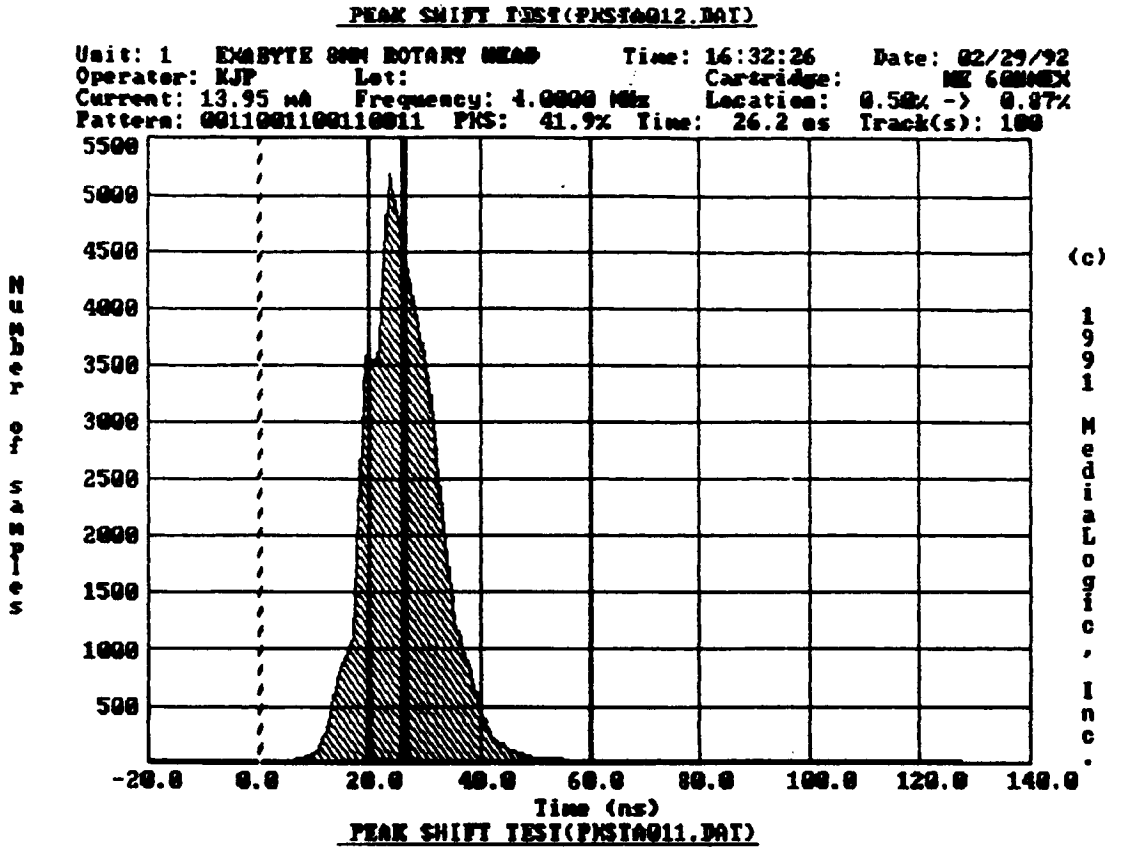
8 MM MP

FIGURE 3



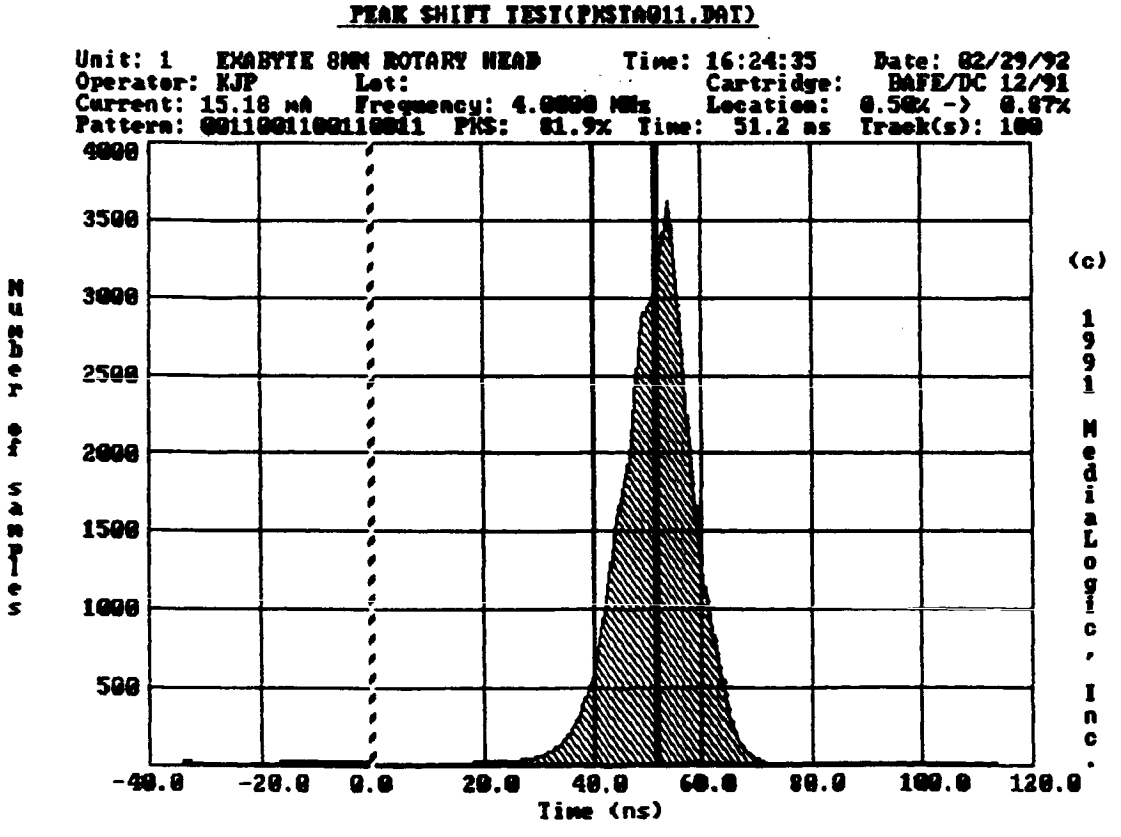
8 MM MP

FIGURE 4



8 MM ME

FIGURE 5



8 MM BAFE

FIGURE 6

```
 TAA AND MODULATION TEST (TAAMA019.DAT)
Unit: 1   EXABYTE 8MM ROTARY HEAD        Time: 14:16:21     Date: 03/12/92
Operator: KJP        Lot:                          Cartridge:
Current: 16.00 mA    Frequency: 4.0000 MHz         Location:  1.00% ->  1.28%
                                                   Tracks: 100

Track Average Amplitude : 178.17 uV  <----
ANSI Modulation         :   3.24 %                          SAFE
ECMA Modulation         :   4.82 %
HF   Modulation         :  15.28 %  <----
================================================================================

 TAA AND MODULATION TEST (TAAMA020.DAT)
Unit: 1   EXABYTE 8MM ROTARY HEAD        Time: 14:29:01     Date: 03/12/92
Operator: KJP        Lot:                          Cartridge:
Current: 16.00 mA    Frequency: 4.0000 MHz         Location:  1.00% ->  1.41%
                                                   Tracks: 100

Track Average Amplitude : 174.12 uV  <----
ANSI Modulation         :   2.40 %                          SAFE
ECMA Modulation         :   3.10 %
HF   Modulation         :   9.96 %  <----
================================================================================

 TAA AND MODULATION TEST (TAAMA022.DAT)
Unit: 1   EXABYTE 8MM ROTARY HEAD        Time: 14:36:23     Date: 03/12/92
Operator: KJP        Lot:                          Cartridge:
Current: 16.00 mA    Frequency: 4.0000 MHz         Location:  5.00% ->  5.42%
                                                   Tracks: 100

Track Average Amplitude : 183.16 uV
ANSI Modulation         :   3.79 %
ECMA Modulation         :   4.92 %                          SAFE
HF   Modulation         :  10.81 %  <----
================================================================================
```

FIGURE 7

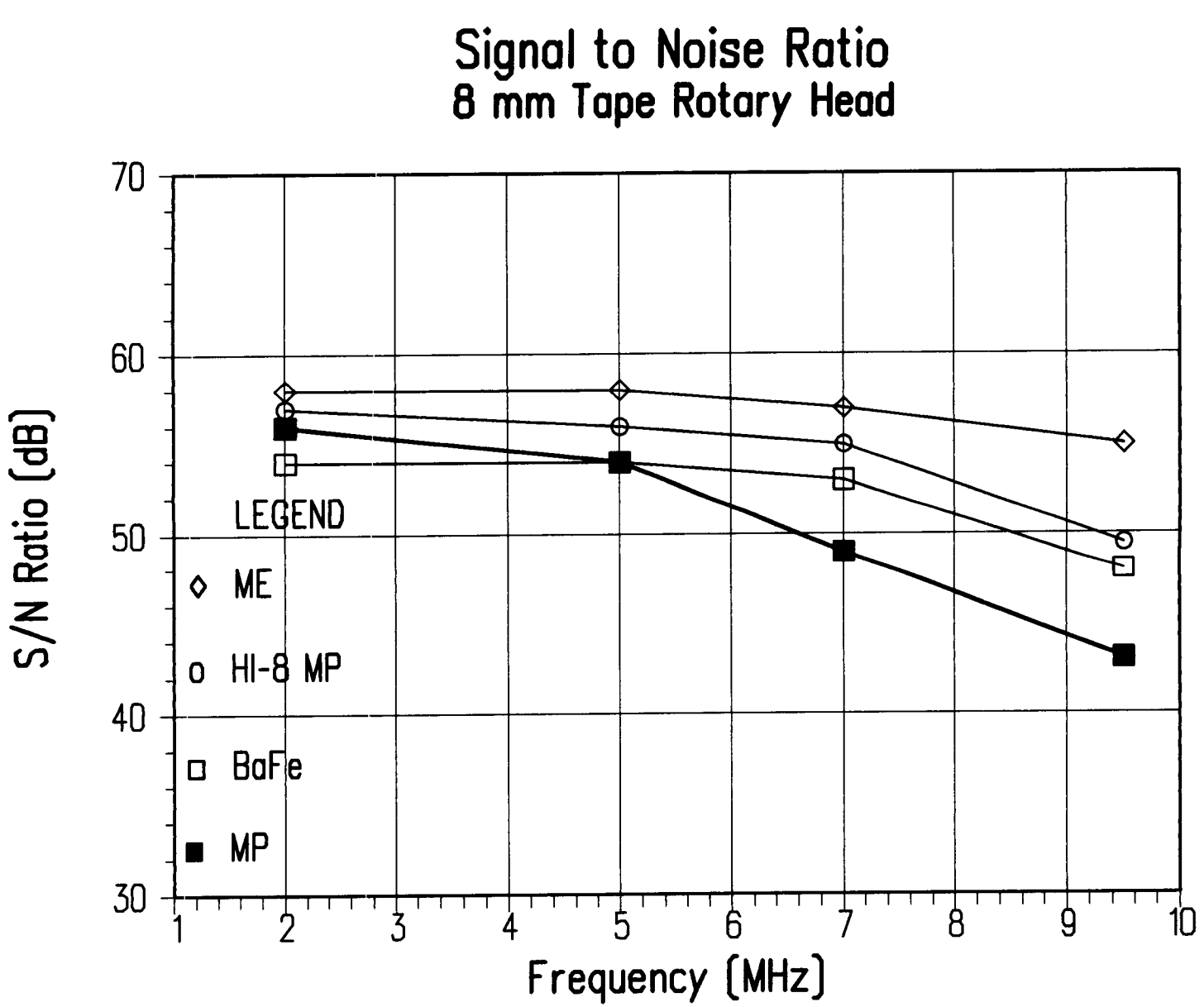

Signal to Noise Ratio
8 mm Tape Rotary Head
S/N Ratio (dB)
70
60
50
40
30
LEGEND
ME
HI-8 MP
BaFe
MP
1
2
3
4
5
6
7
8
9
10
Frequency (MHz)

FIGURE 8

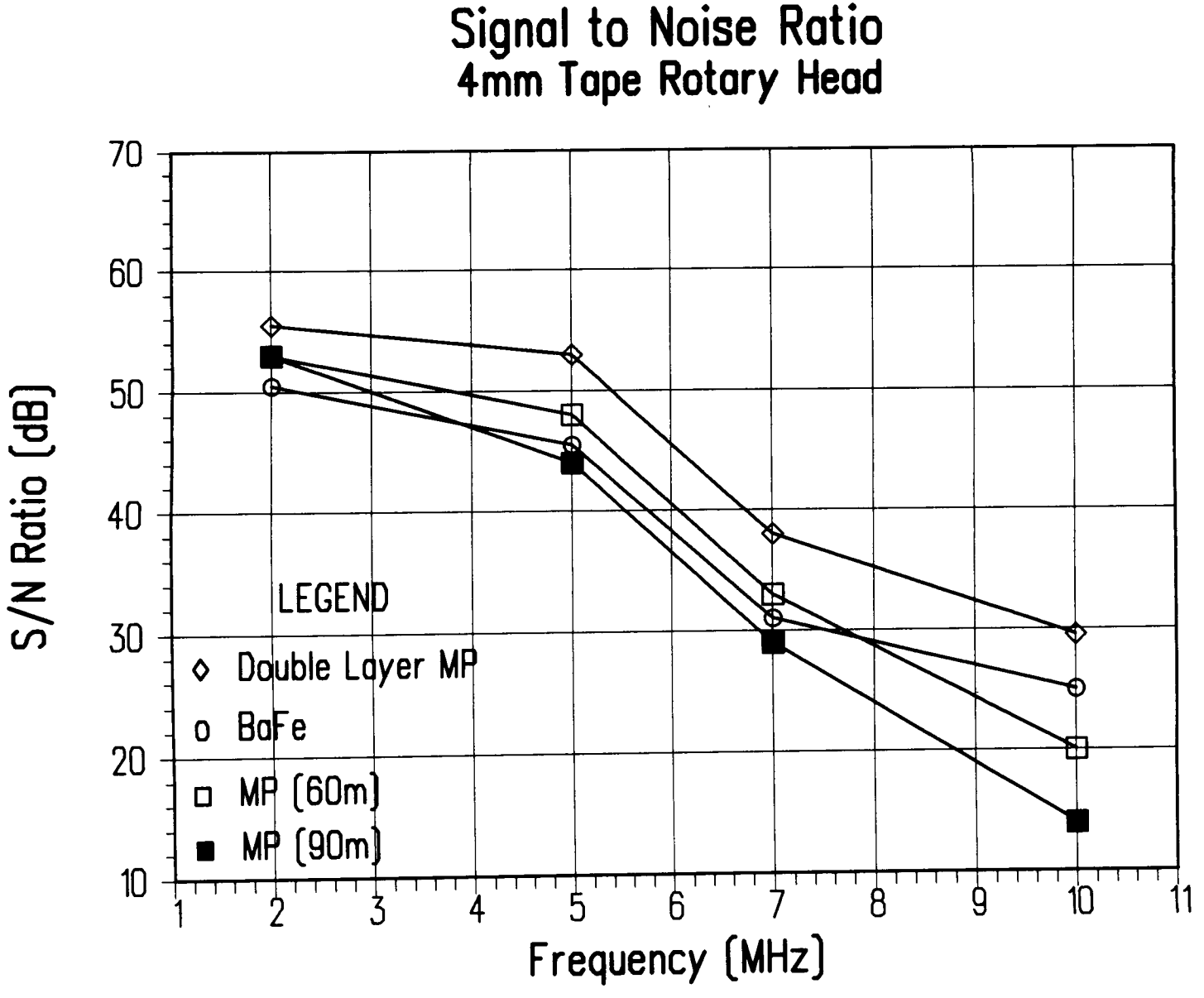
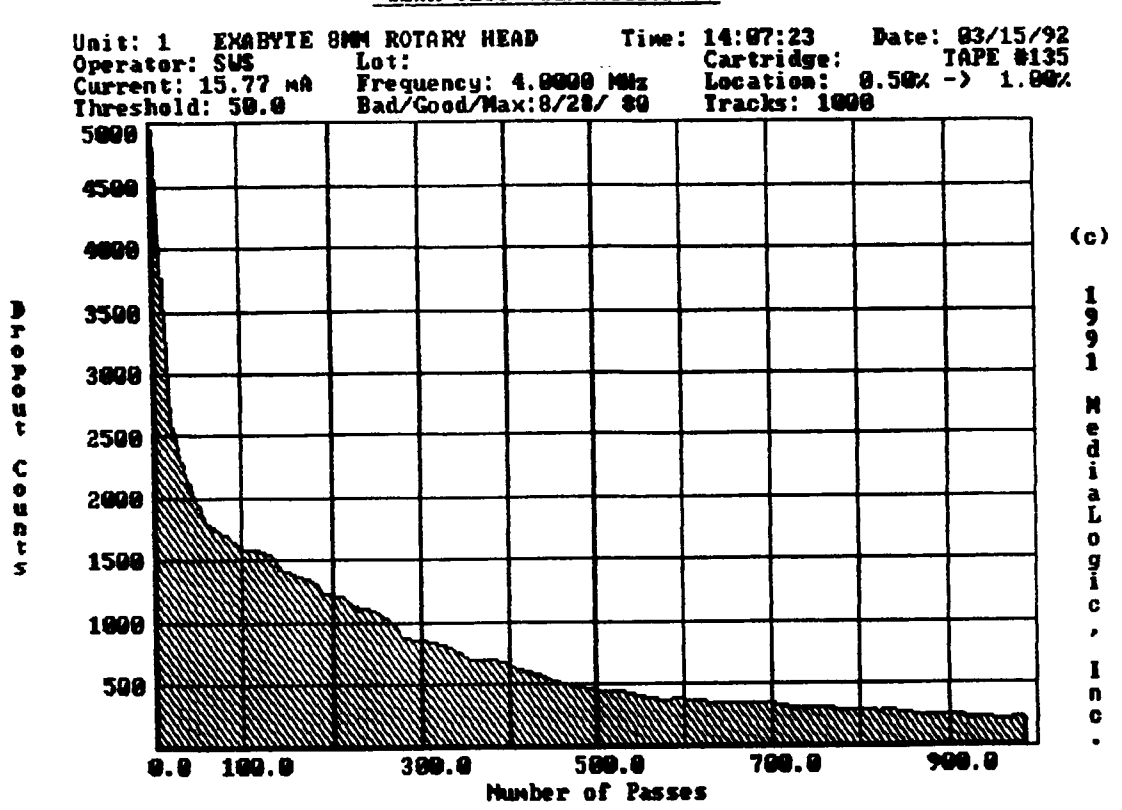

Signal to Noise Ratio
4mm Tape Rotary Head
S/N Ratio (dB)
70
60
50
40
30
20
10
LEGEND
◇ Double Layer MP
o BaFe
□ MP (60m)
■ MP (90m)
1
2
3
4
5
6
7
8
9
10
11
Frequency (MHz)

FIGURE 9



8 MM MP

FIGURE 10

WEAR TEST (WEARA002.DAT)
Unit: 1 EXABYTE 8MM ROTARY HEAD Time: 14:07:23 Date: 03/15/92
Operator: SWS Lot: SONY Cartridge: TAPE #135
Current: 15.77 mA Frequency: 4.0000 MHz Location: 0.50% -> 1.00%
Threshold: 50.0 Bad/Good/Max:8/28/ 80 Tracks: 1000

Pass Num.	Counts	Pass Num.	Counts	Pass Num.	Counts
1	4976	393	673	785	263
18	2607	404	617	796	264
34	2164	417	579	816	263
59	1797	432	567	821	254
72	1712	452	512	833	263
85	1680	464	486	851	241
97	1576	480	477	869	235
113	1578	505	432	883	234
129	1531	521	429	897	233
140	1423	527	430	913	227
169	1343	548	390	925	206
181	1302	562	372	940	208
189	1229	578	354	962	196
210	1208	599	362	985	202
229	1091	601	349	988	191
237	1094	621	347		
251	1060	633	327		
270	961	651	325		
278	864	670	327		
300	840	679	329		
312	825	701	325		
326	803	709	301		
341	755	729	289		

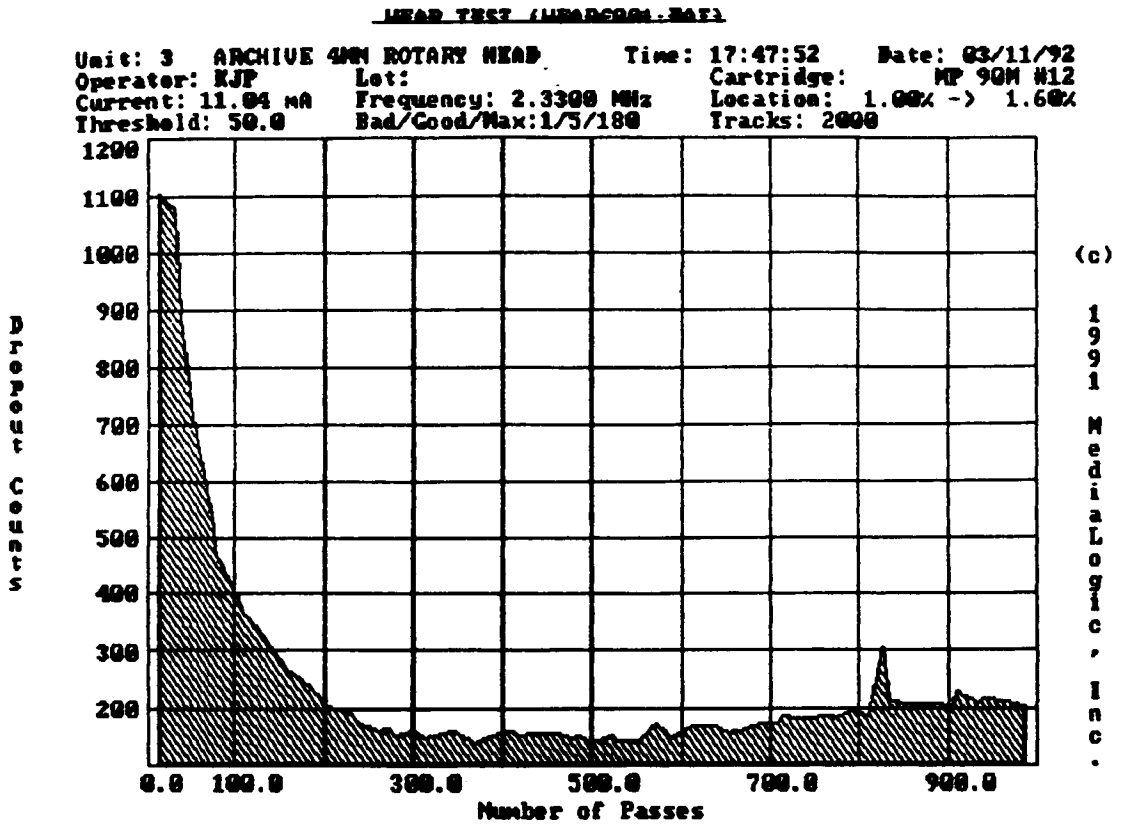
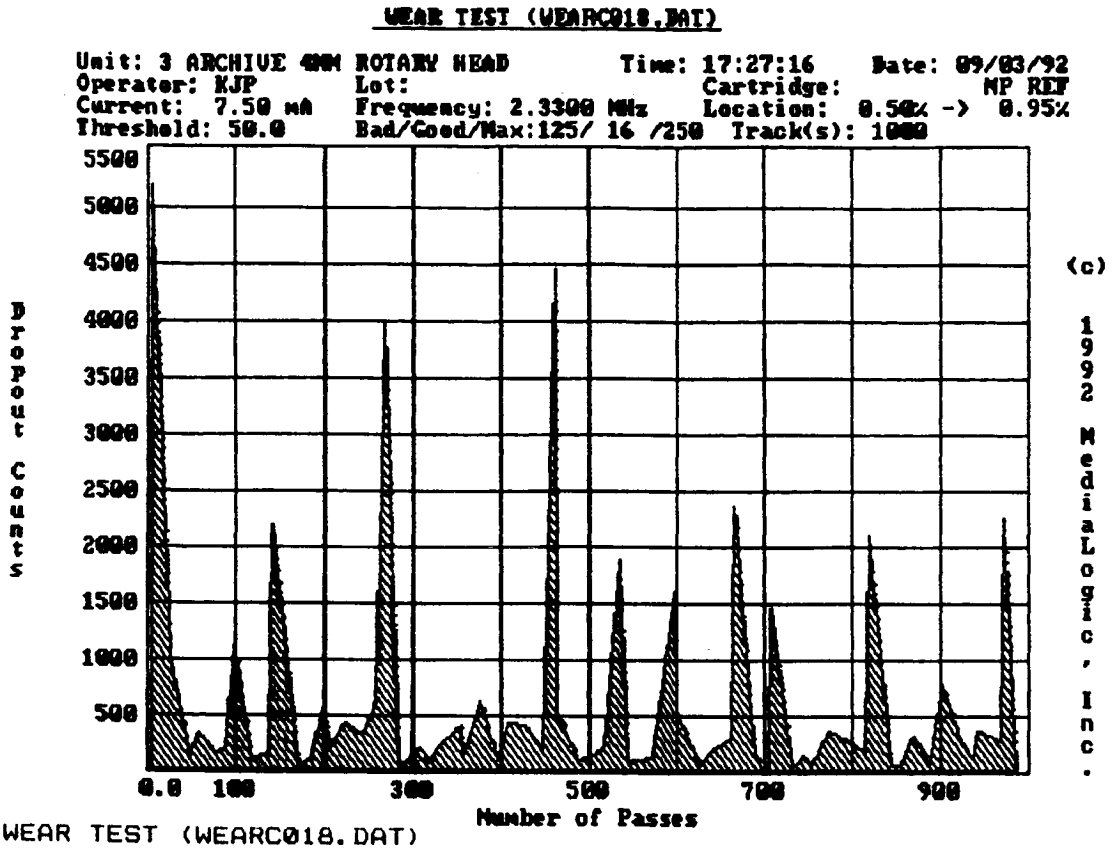

HEAD TEST (HEADCOM1.DAT)
Unit: 3 ARCHIVE 4MM ROTARY HEAD Time: 17:47:52 Date: 03/11/92
Operator: KJP Lot: Cartridge: MP 90M #12
Current: 11.04 mA Frequency: 2.3300 MHz Location: 1.00% -> 1.60%
Threshold: 50.0 Bad/Good/Max:1/5/180 Tracks: 2000
1200
1100
1000
900
800
700
600
500
400
300
200
Dropout Counts
0.0
100.0
300.0
500.0
700.0
900.0
Number of Passes
(c) 1991 MediaLogic, Inc.

FIGURE 11



20 K PASSES

FIGURE 12

WEAR TEST (WEARC018.DAT)
Unit: 3 ARCHIVE 4MM ROTARY HEAD Time: 17:27:16 Date: 09/03/92
Operator: KJP Lot: Cartridge: MP REF
Current: 7.50 mA Frequency: 2.3300 MHz Location: 0.50% -> 0.95%
Threshold: 50.0 Bad/Good/Max:125/ 16 /250 Track(s): 1000

Pass Num.	Counts	Pass Num.	Counts	Pass Num.	Counts
2	5173	395	75	775	340
24	995	407	401	797	253
47	95	427	375	816	134
57	312	447	174	821	2077
75	140	461	4443	847	43
86	170	463	487	857	43
97	1100	488	69	871	304
116	88	505	131	890	65
136	153	518	212	905	765
140	2170	535	1856	923	323
170	21	542	48	937	111
183	101	570	103	944	342
198	540	571	155	970	251
206	187	598	1574	974	2246
221	416	601	507	987	19
242	293	629	31		
258	579	640	177		
267	3976	662	268		
283	19	666	2343		
307	181	690	114		
317	63	706	38		
329	204	709	1426		
353	375	733	23		
355	93	746	120		
375	600	755	43		

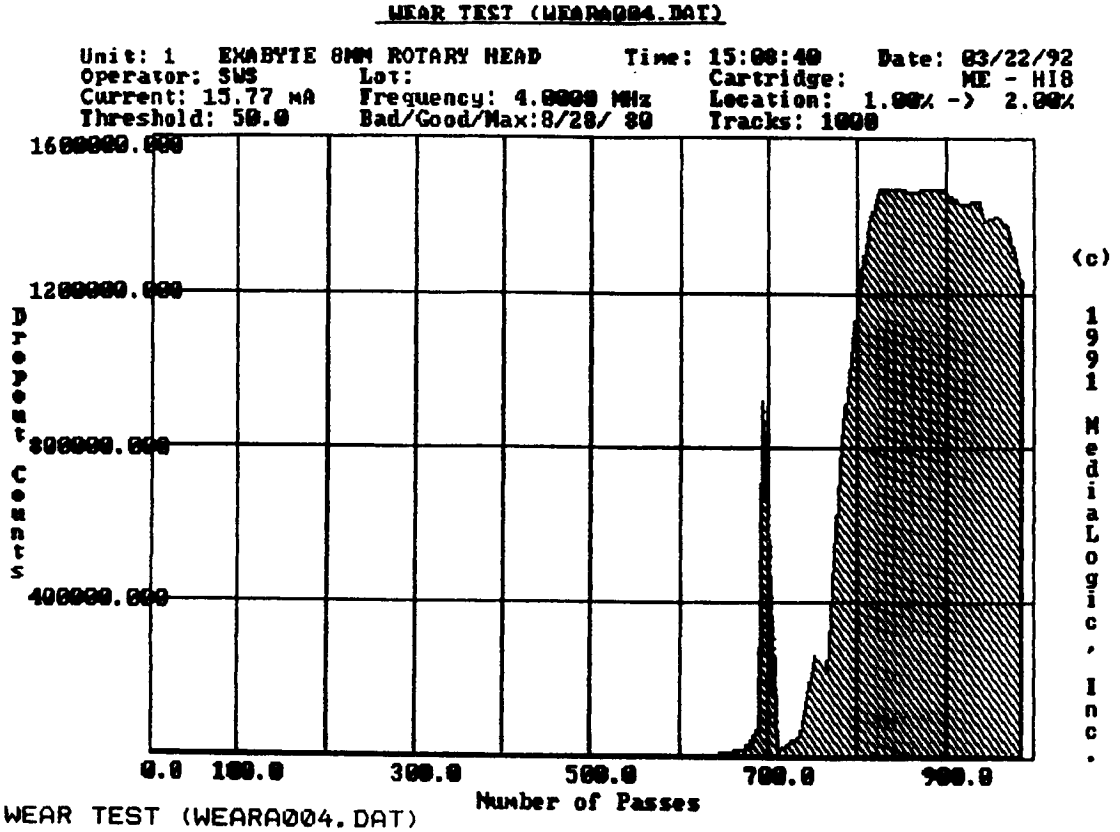


FIGURE 13

WEAR TEST (WEARA004.DAT)

Unit: 1 EXABYTE 8MM ROTARY HEAD Time: 15:08:40 Date: 03/22/92
Operator: SWS Lot: Cartridge: ME - HI8
Current: 15.77 mA Frequency: 4.0000 MHz Location: 1.00% -> 2.00%
Threshold: 50.0 Bad/Good/Max:8/28/ 80 Tracks: 1000

Pass Num.	Counts	Pass Num.	Counts	Pass Num.	Counts
1	37	400	41	785	814773
23	36	406	27	800	1158717
42	35	418	29	814	1384420
57	35	434	30	826	1473750
63	34	448	23	844	1474072
86	34	463	26	851	1473706
98	32	492	26	863	1471346
110	33	497	23	880	1473884
126	31	513	22	894	1472254
149	31	537	26	921	1437701
167	30	545	28	938	1440926
173	31	562	27	946	1390048
197	33	576	26	960	1403681
202	33	588	23	971	1379884
226	30	602	44	987	1232467
238	36	622	30		
258	30	642	406		
271	32	651	1058		
286	26	674	8604		
307	26	691	59384		
316	27	695	925003		
327	26	712	5831		
347	26	739	47945		
355	25	754	256125		
383	29	768	210260		

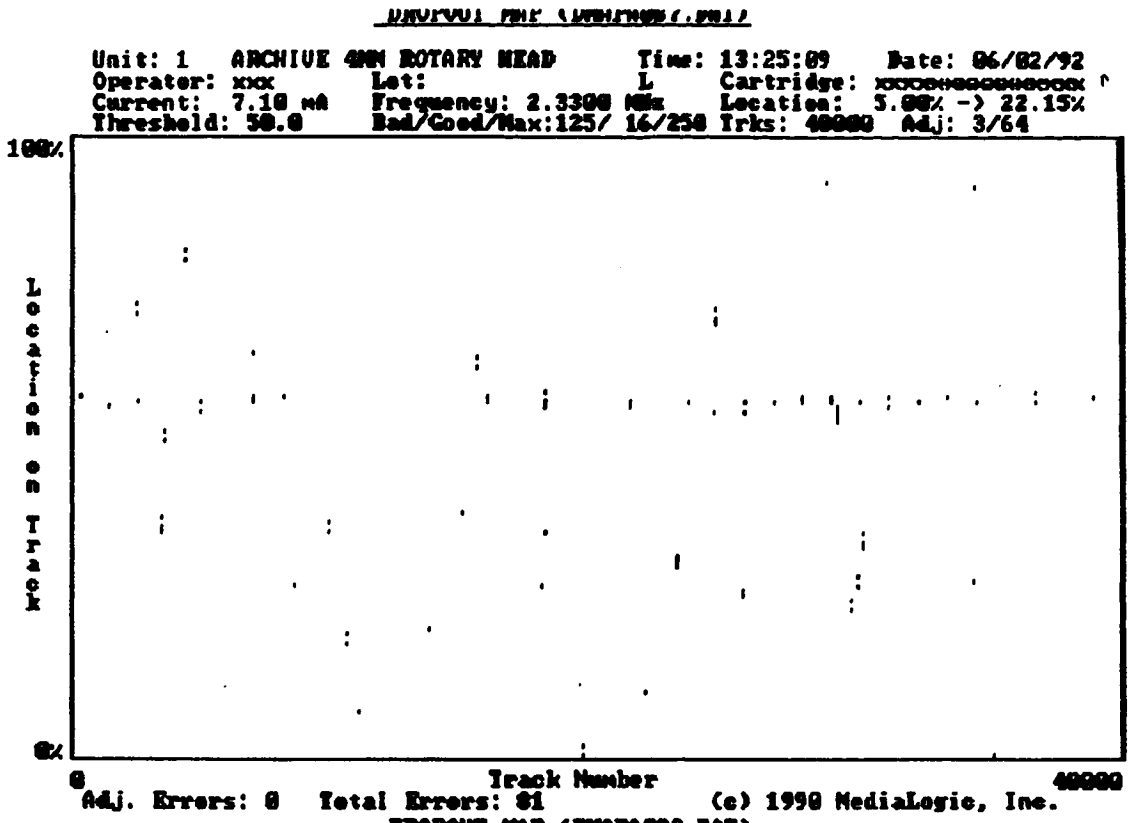


FIGURE 14

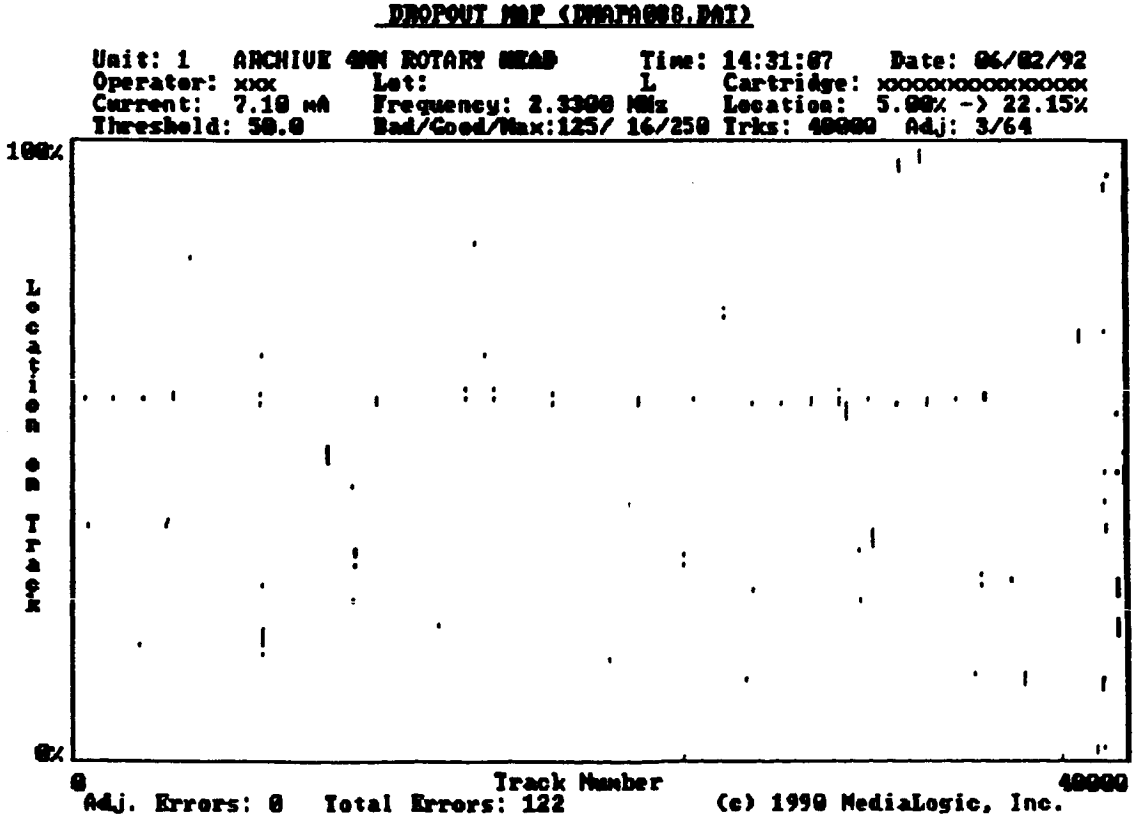


FIGURE 15

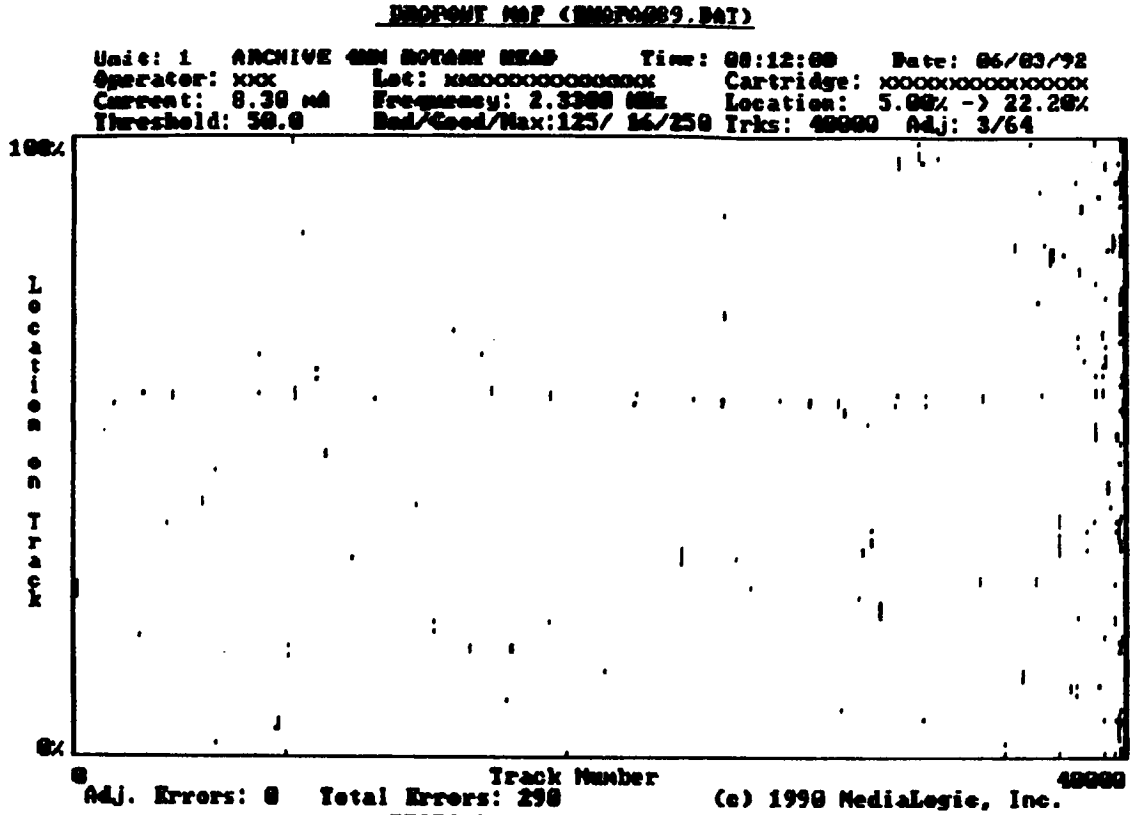


FIGURE 16

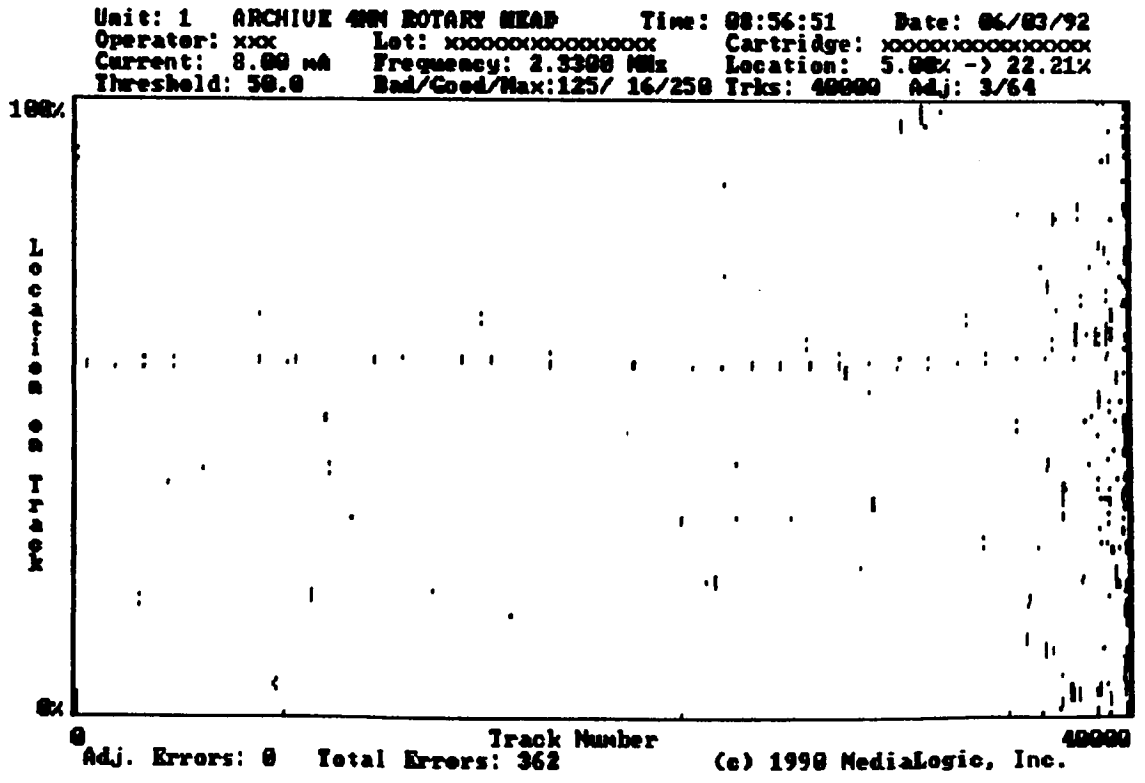


FIGURE 17

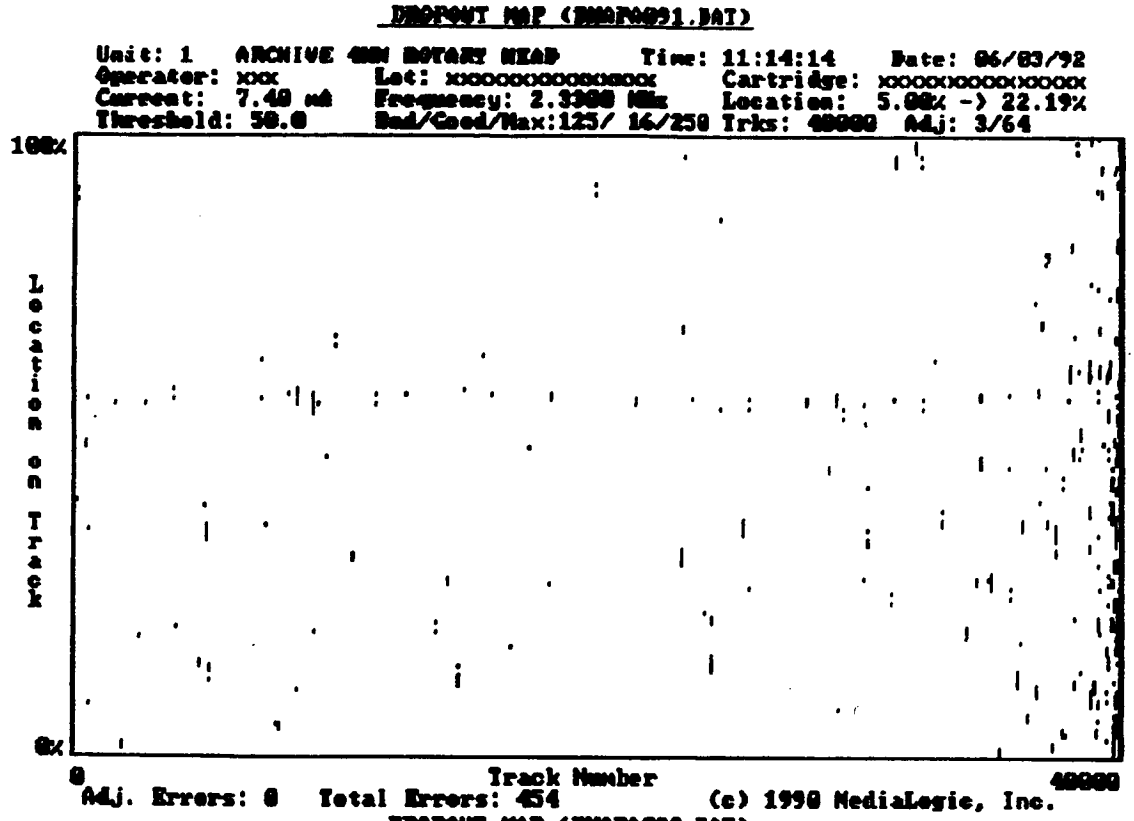


FIGURE 18

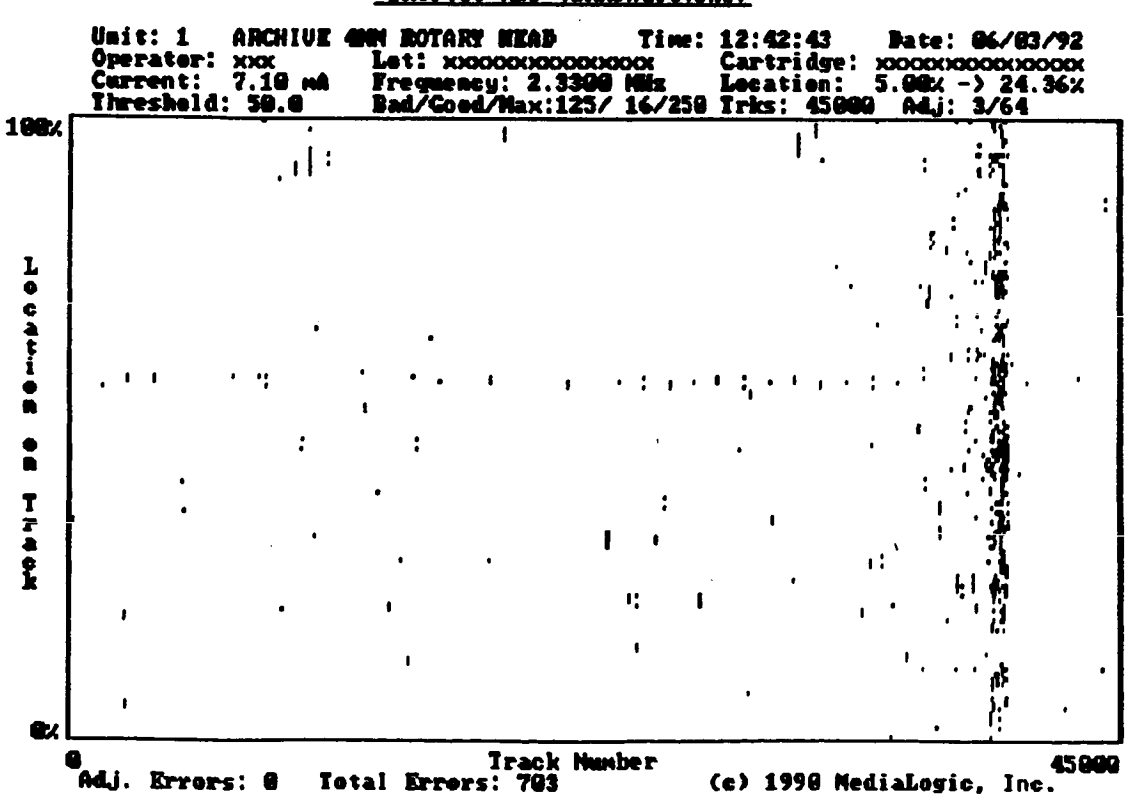


FIGURE 19

Striped Tape Arrays

Ann L. Drapeau

Computer Science Department
University of California
571 Evans Hall
Berkeley, CA 94720

alc@cs.berkeley.edu

Striped Tape Arrays

Ann L. Drapeau
alc@cs.berkeley.edu

Motivation

- Applications require **high throughput** (100 MB/sec), **massive storage** (Terabytes, Petabytes)
- Technology Trends
  - Magnetic tape: high capacity, low bandwidth
  - Robots: automatic loading of tape cartridges
- **Striping: a technique for increasing throughput**
- Issues in striping effectively
- Tape array reliability

Outline

- Introduction to Striping
- Applications
- Tape Technologies
- Robots
- Access Times
  - Drive and Robot Measurements
- Striping Options and Issues
- Reliability Issues
- Summary

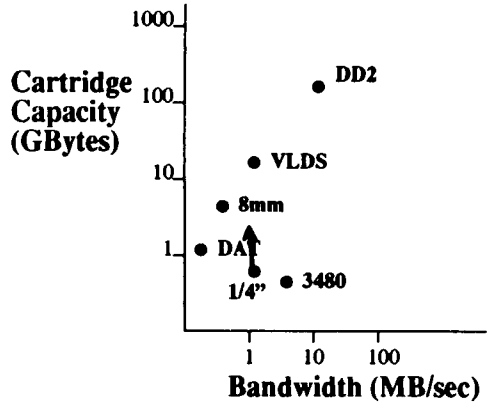
Data Striping

- **Spread data from individual files across several devices**
- Advantages:
  - **Increase bandwidth to a single file**
  - **Reduce latency of large accesses**
  - Allows independent "smaller" accesses
  - Easy to incorporate error correction
- Problems:
  - Increase latency of some accesses
  - Synchronization

Do Applications Need Striped Tape?

- **Large scientific archives (NASA EOS)**
  - High sustained bandwidth (100 MB/s)
  - Total storage very large (Petabytes)
  - Would benefit from striping **throughput**
- **Interactive access to large data sets (Sequoia)**
  - Researchers across California
  - Want reasonable response time over network
  - Total storage large (Terabytes)
  - Striping would **reduce large access latency**

Tape Technologies



Tape Technologies

Technology	Capacity per Cartridge	Cost	Bandwidth
1/4"	2 GB	$ 1000	3 MB/sec
1/2" 3480	480 MB	$ 20000	6 MB/sec
4mm DAT	1.3 GB	$ 1000	183 KB/sec
8mm Exabyte	5 GB	$ 3000	500 KB/sec
1/2" Metrum VLDS	14.5 GB	$ 40000	2 MB/sec
Ampex DD2	165 GB	$150000	15 MB/sec

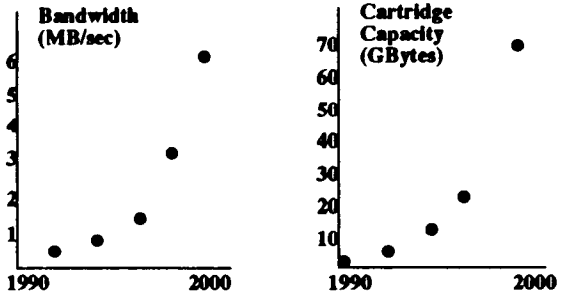
Linear Recording: 1/4" cartridge, 1/2" 3480

Helical Scan: DAT 4mm, 8mm, 1/2" VLDS, 19mm D2

Tape Tradeoffs: No "Perfect" Drive

- **Inexpensive** helical scan drives have **low bandwidth** (DAT, 8mm)
- **Inexpensive** serpentine drives have **moderate bandwidth** (1/4")
- **High capacity** drives have **long access times** (helical scan, 1/4")
- Drives with **short access times** are **low capacity** (1/2" 3480)
  - Moderate price and bandwidth
- **High bandwidth** drives **very expensive** (DD2)
  - **Bandwidth not high enough**
  - **Very high capacity**

Future Tape Drives (8mm)



- Source: Harry C. Hinz, Exabyte Corp.
- **Changes:** increase track density, decrease track width & pitch, reduce tape thickness, increase rotor speed

Robots

- **Large Libraries:**
  - many cartridges, several drives
  - expensive
  - one or more robot arms
- **Carousels**
  - around 50 cartridges, one or two drives
  - moderate cost
- **Stackers**
  - around 10 cartridges, one drive
  - inexpensive

Robots

	Metrum RSS-600 (1/2" VLDS)	Spectra Logic STL-8000H Carousel (8mm)	Exabyte EXB-10 Stacker (8mm)
# Drives	up to 5	1 or 2	1
# Cartridges	600	45	10
Total Capacity (GBytes)	>6000	225	50
Cost	$540,000 (2 drives)	$27,500 (1 drive)	$7000
Avg. Robot Access Time (sec)	8	10	<20

Tape Access Time (Cartridge Switch)

- Access time =

  rewind time +
  eject time +
  robot unload +
  robot load +
  device load +
  fast search +
  transfer time

- Measured three tape drives, one robot:
  Accurate access time models for simulation

Drive Measurements

Drive Load and Eject Times

	4mm DAT	8mm Exabyte	Metrum VLDS
Mean Load Time (sec)	16	35.4	28.3
Mean Eject Time (sec)	17.3	16.5	3.8

Data Transfer Rates

	4mm DAT	8mm Exabyte	Metrum VLDS
Read Rate (MB/sec)	0.17	0.47	1.2
Write Rate (MB/sec)	0.17	0.48	1.2

Rewind and Search Behavior

	4mm DAT	8mm Exabyte	Metrum VLDS
Rewind Startup (sec)	15.5	23	15
Rewind Rate (MB/sec)	23.1	42.0	350
Search Startup (sec)	8	12.5	28
Search Rate (MB/sec)	23.7	36.2	115

- Constant startup
- Approximately linear search/rewind

Tape Access Time Example
(Exabyte EXB8500 Drive, EXB-120 Robot)

- Average Access time =

rewind time	(1/2 tape)	(75 sec)	+
eject time		(17 sec)	+
robot unload		(21 sec)	+
robot load		(22 sec)	+
device load		(35 sec)	+
fast search	(1/2 tape)	(84 sec)	+
transfer time			

- **Not including data transfer: 4 minutes!**

Options for Striped Tape

- Within a robot
  - + cartridges in stripe kept together
  - - few readers, robot arms
  - - single point of failure
- Between robots
  - + several robot arms used in access
  - - harder to keep cartridges together
- Between small robots (stackers)
  - + highest proportion arms to readers and cartridges

Striping Issues

- Configuration depends on workload
- **Interleave factor crucial:**
  - Too small: cartridge switches increase latency (Long access times -- big penalty)
  - Too big: lose potential parallelism
- **Workloads that will benefit from striping**
  - Large archives
  - Interactive systems with large avg. request size
- **Striping will hurt** performance of some accesses
  - Interleave much smaller than average request
  - High load/scarce readers

More Striping Issues

- Striping with improved devices/robots
  - **Higher bandwidth drives**
    - Bandwidth, aerial density may increase 30X by end of decade
    - Less need for striping?
    - **Still get throughput benefits**
  - **Faster access times (drives and robots)**
    - faster load, eject, search, rewind, robot arms
    - no rewind before eject
    - cartridge switch penalties reduced
    - **striping more effective**

Synchronization Issues

- Drives retry after failed writes
  - Bad tape would retry indefinitely
  - Pat Savage (Shell Oil): after write error, retry on all tapes in stripe
- If "RAID-5" (large interleaving)
  - Single cassettes may satisfy smaller requests independently
  - Large requests spanning several tapes may be out of synchronization by minutes
  - Buffer space required to hold stripe units while request completes

Reliability Issues: Tape Media

- **High rates of raw bit errors**
  - before internal ECC
  - one in 10^5 bits
- **Dropouts**
  - Debris
    - Slicing of tape
    - Particles in atmosphere
    - Start/stop wear
  - Nonhomogeneous Tape Coating

Uncorrectable Bit Error Rates

Drive	Bit Error Rate
1/4"	10^{-14}
4mm DAT	10^{-15}
Exabyte 8mm	10^{-13}
Metrum VLDS	10^{-13}
Ampex DD2	10^{-12}

- **Error rates after ECC**
- Terabyte approximately 10^{13} bits
- **MSS will contain uncorrectable errors!**

- **Need Error Correction**
  - Easy to implement in striped systems
  - **How much?**
  - How reliable are error rates?
  - How will ECC affect performance?
- Error Rates Increase with Wear
- **Tapes last around 2000 passes**
- Severe wear: tape unreadable
- **If tapes are rewritten often, need to copy tapes periodically**

Reliability: Tape Heads

- Drive design includes tape/head wear
- Accumulate debris
  - tape debris
  - atmosphere
  - tape coating (friction, humidity)
- Wear with tape medium helps clean heads
- **Heads last around 2000 hours of tape contact**
- Algorithms for
  - Periodic head cleaning
  - Fast replacement on failure

More Reliability Issues

- Other drive problems

Megatape 1991 Repair Statistics (8mm)

Repair type	%
Replace heads	44
Tape mechanism (reel motors, tape tension, etc.)	21
Card failure	17
Other (firmware, power supply, etc.)	14
No defect found	4

- Robot reliability
- Support hardware

Summary

- **Applications want high sustained throughput**
- **Technology Trends:**
  - Tape drives increasing in capacity, bandwidth (currently inadequate)
  - Robots allow automatic handling of cartridges
- **Striping:**
  - Increased throughput
  - Reduced latency of large requests
- **Striping configurations:**
  - Within or between robots
  - Tradeoffs: ratio of readers, robot arms, tapes

- **Striping issues:**
  - Interleave factor for best performance
  - Effect of improved drives, robots
  - Synchronization problems
- **Reliability Issues:**
  - Media Wear
  - Head Wear
  - Other drive failures
  - Robot failures
  - Error correction needed: how much?

Ultra-High Density Recording Technologies

Mark H. Kryder
Data Storage Systems Center
Engineering Research Department
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Introduction

The Engineering Research Center in Data Storage Systems at Carnegie Mellon University in cooperation with the National Storage Industry Consortium has selected goals of achieving 10 Gbit/in^2 recording density in magnetic and magneto-optic disk recording and 1 terabyte/in^3 in magnetic tape recording technologies. This talk will describe the approaches being taken and the status of research leading to these goals.

Future Recording Technologies

The capacities and performance which could be achieved from magnetic tape, magnetic disk and magneto-optic disk drives, assuming storage densities of 1 TByte/in^3 on magnetic tape and 1 Gbit/in^2 on magnetic disk and magneto-optic disk are illustrated in Fig. 1.

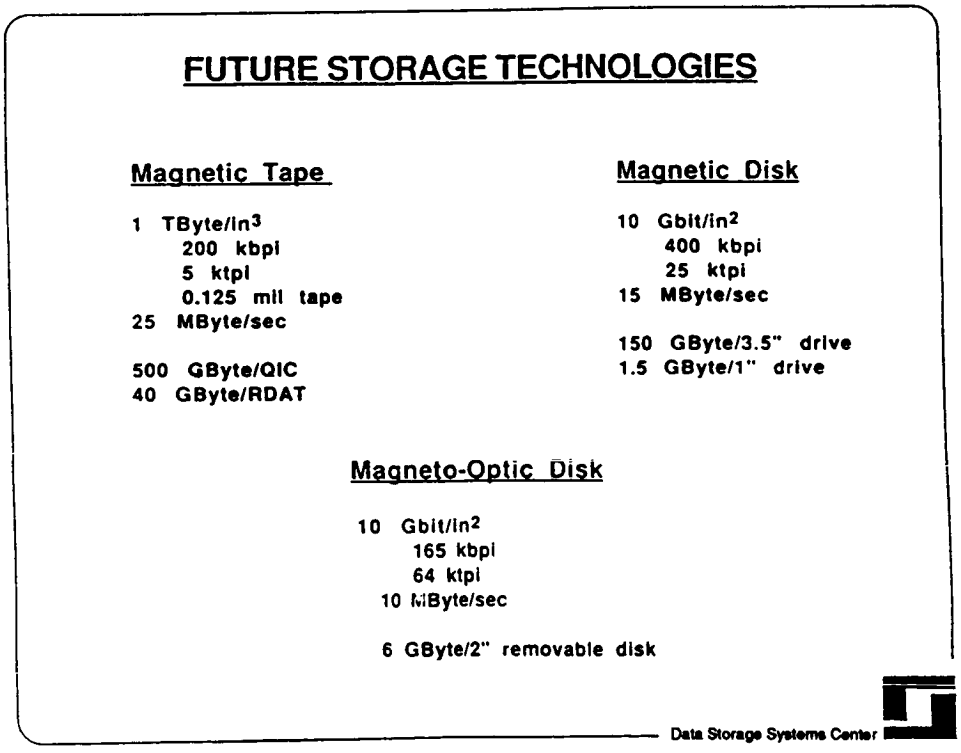


Fig. 1. Possible configurations for future ultra-high density magnetic tape, magnetic disk and magneto-optic disk drives.

It is proposed that 1 TByte/in^3 can be achieved on magnetic tape by using 0.125 mil tape and 1 Gbit/in^2 areal recording density. It is noted that 1 Gbit/in^2 areal recording density has already been achieved on a magnetic disk [C. Tsang et al., *IEEE Trans. Magnet.*, MAG-26, 1689 (1990)]. It is suggested that a linear bit density of 200 kbpi and a track density of 5 ktpi will be approximately what are used. Building a transport capable of handling tape only 0.125 mils thick will be a challenge, but is believed to be possible. Data rates of 25 MByte/sec could be achieved from a single helical scan head or from a longitudinal recorder with multiple track heads.

Although cartridge sizes may well change significantly by the time such technology is available, it is interesting to note that a storage density of 1 TByte/in^3 would make it possible to store 500 GBytes in a Quarter Inch Cartridge and 40 GByte in a RDAT cartridge.

A storage density of 10 Gbit/in^2 on a magnetic disk drive would make it possible to store 150 GBytes in a 3.5-inch disk drive having 12 disks in it, or 1.5 GBytes on a single 1-inch disk. A data rate of about 15 MByte/sec could be achieved from the 3.5-inch disk by spinning it at 3600 rpm or from a 1-inch drive by spinning it at 10,800 rpm.

With a storage density of 10 Gbit/in^2, a 3.5-inch drive will probably be as large a disk drive as one would desire to build. A capacity of 150 GBytes is a lot for one spindle and larger disks would make the data rate too high for the semiconductor channel electronics. One inch or smaller drives are likely to be the high volume products when such storage density is available.

Although magneto-optic disk drives could in principle be made equally as small as hard drives, it is doubtful that they will use media smaller than 2 inches, because the media is removable. Smaller removable disks would be too easily lost. At 10 GBit/in^2 a 2-inch disk could store about 6 GBytes. Since magneto-optic recording uses a lower linear bit density than magnetic recording, a disk rotation rate of 10,800 rpm would be necessary to achieve a data rate of 10 MBytes/sec, still slower than a 1-inch magnetic hard disk spinning at the same speed.

Scaling Magnetic Recording Technology

The width of a recorded transition in magnetic recording media is affected by the media properties, the head field gradient and demagnetizing effects, as illustrated in Fig. 2. The finite switching field distribution, SFD, in the media convolved with the recording head field gradient cause a finite width transition. Demagnetizing fields in the media tend to broaden the transition. In a medium of thickness d exhibiting perfect squareness, the transition width may be written as

$$a = \sqrt{\frac{\delta^2}{16} + \frac{M_r \delta (d + \delta/2)}{\pi H_c}} - \frac{\delta}{4}$$

where M_r is the remanent magnetization of the medium, H_c is the coercive force, and d is the head-to-medium spacing. That the transition width broadens with increasing $M_r d/H_c$ is due to demagnetizing effects, while the dependence on $d+\delta/2$ is due to the reduction in head field gradient with increased spacing of the head from the medium.

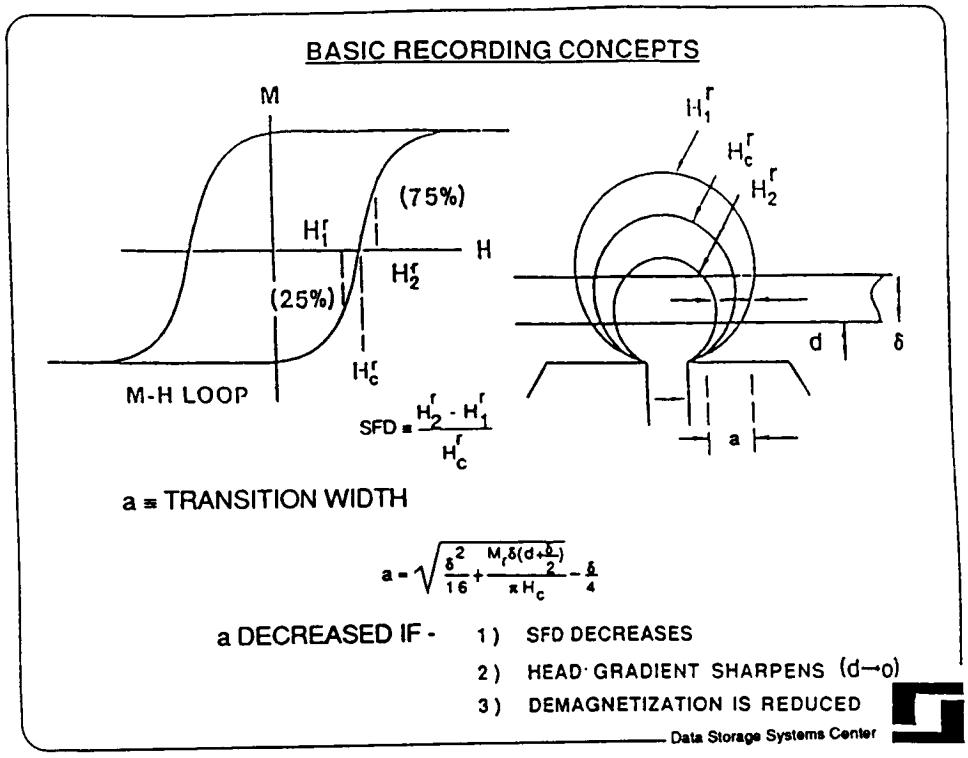


Fig. 2. Factors which determine the recorded transition width.

In addition to the widening of the recorded transition produced by increased head-to-medium spacing, increased spacing leads to a very rapid decrease in playback signal, as illustrated in Fig. 3. The readback voltage from the head varies exponentially with the ratio of $d+\delta/2$ to b, the spacing between flux changes. This causes a reduction in signal of about 27 dB for every $(d + \delta/2)/b$. This extremely rapid fall-off in signal with head-to-medium spacing means that, when the recording density is increased, it is extremely important that the head-to-medium spacing is simultaneously reduced.

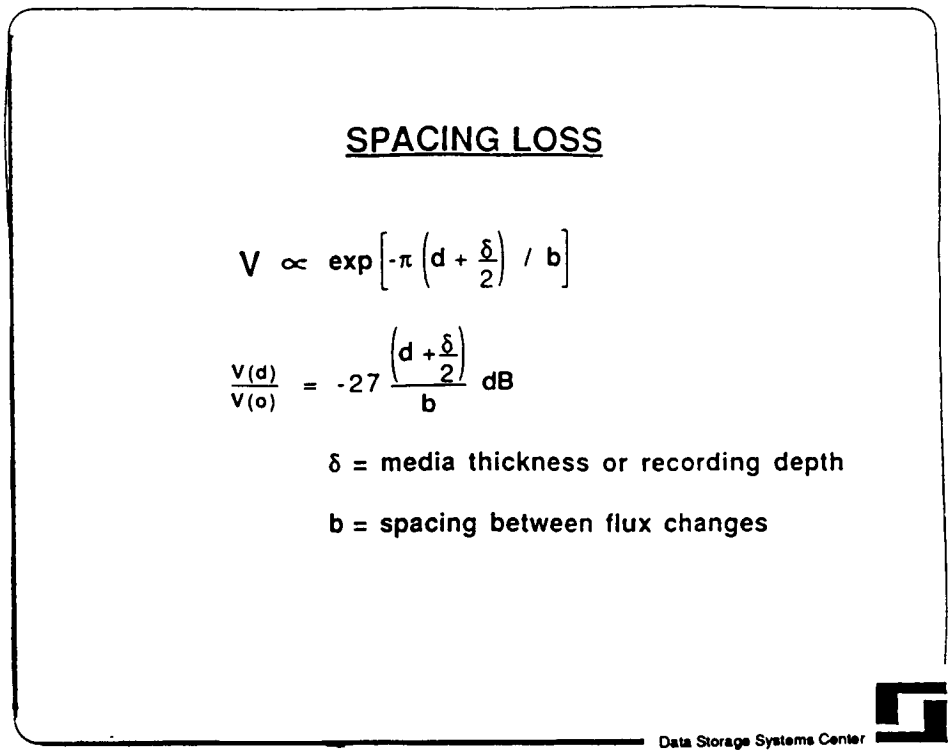


Fig. 3. Spacing loss in magnetic recording.

In magnetic tape recording systems, the head runs in contact with the tape. Even so a finite head-to-medium spacing results as the head moves from asperity to asperity on the tape surface. To achieve a very small head-to-medium spacing, the tape must be very smooth.

Magnetic hard disk systems built today use an air bearing to fly the head slightly above the surface of the media. Head-to-medium spacings of the order of 4 microinches (100 nm) are being used today. To achieve 10 Gbit/in^2 recording density it is expected to be necessary to also run disk heads in contact with the medium. One approach to achieving this is illustrated in Fig. 4. A low mass secondary slider which runs in contact with the media is built into a larger primary slider which flies above the disk surface. The low mass of the secondary slider and its weak coupling with the more massive primary slider enable it to run in contact with the media without causing significant wear. This entire assembly can be micro-machined from single crystal silicon. An alternative approach is to use a whisker-like flexible probe head such as that being pursued by Censtor Corporation [H. Hamilton, et al., *IEEE Trans. Magnet.*, MAG-27, 4921 (1991)].

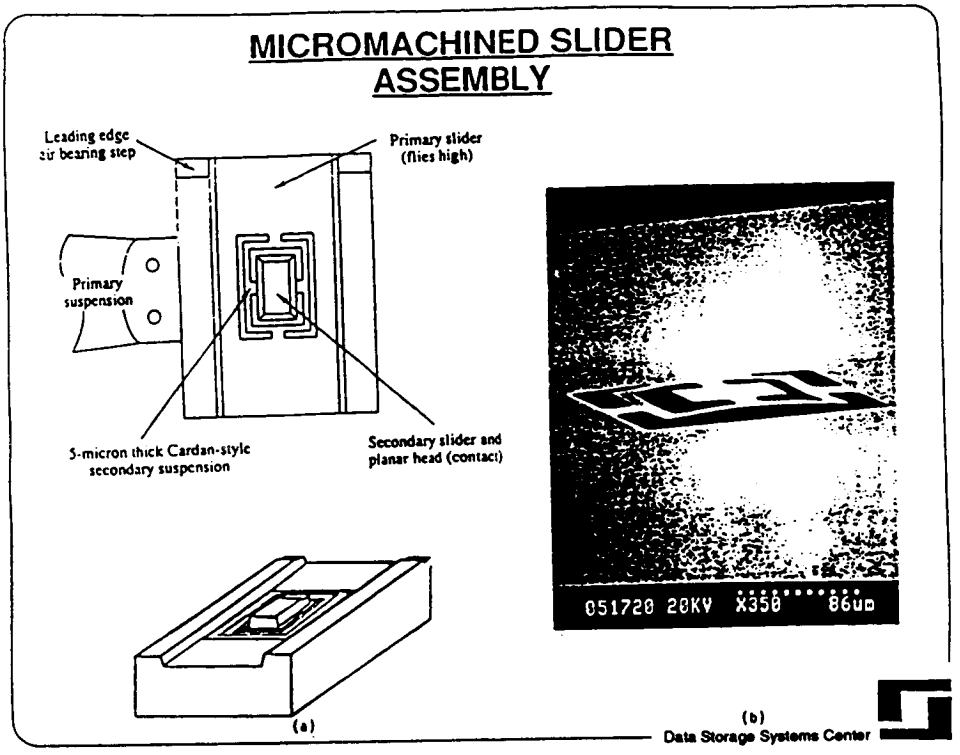


Fig. 4. (a) A diagram of a micromachined slider for contact recording in a disk system. The low-mass secondary slider runs in contact with the medium.

(b) A SEM micrograph of a micro-machined secondary slider.

The ultimate noise level in a magnetic recording system is set by the noise properties of the medium. The media power signal-to-noise ratio was shown by Mallinson [*IEEE Trans. Magnet.*, MAG-5, 182 (1969)] to be approximately equal to the number of particles sensed by the recording head at any instant in time. Thus the power signal to noise ratio is given by

$$PSNR = nWb\delta,$$

where n is the number of magnetic particles per unit volume, W is the recorded trackwidth, b is the spacing between flux changes (one half the wavelength of recording) and δ is the medium thickness. The particle or grain sizes necessary to achieve 20 dB or 30 dB of signal-to-noise ratio in magnetic tape and disk media with densities of 1 Gbit/in^2 and 10 Gbit/in^2 recording densities, respectively are shown in Fig. 5. These particle and grain sizes are believed to be achievable.

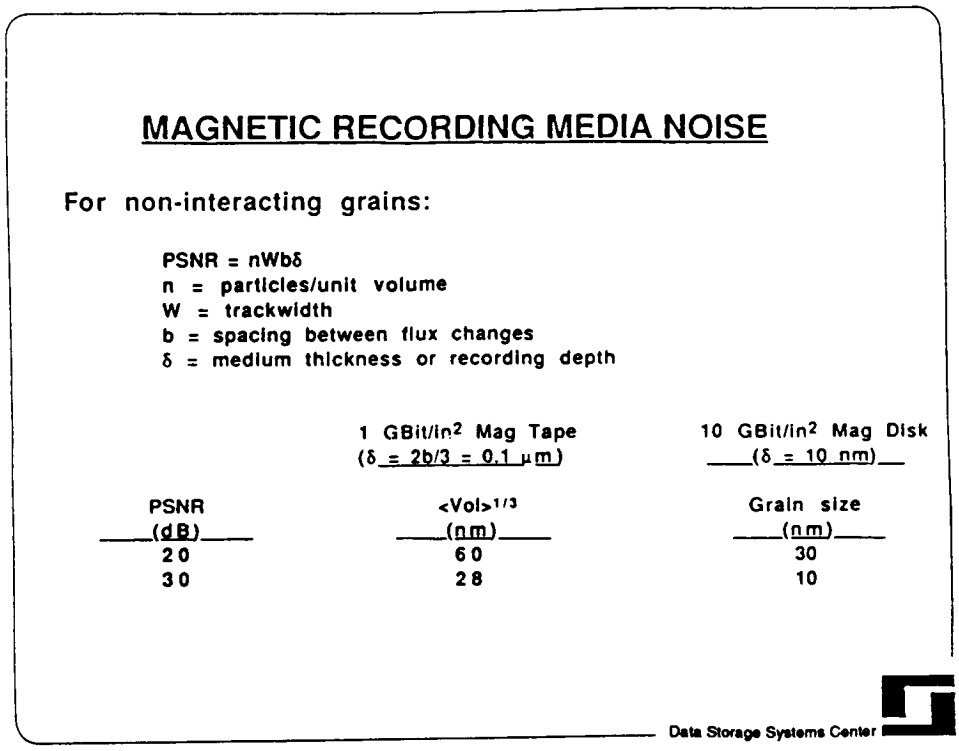


Fig. 5. The particle sizes necessary to achieve given power signal to noise ratios in 1 Gbit/in^2 and 10 Gbit/in^2 recording densities in magnetic tape and disk systems, respectively.

The recorded transition width which might be expected in future barium ferrite and metal evaporated tape media for 1 Gbit/in^2 recording density are shown in Fig. 6. The depth of recording into a medium is given approximately by 2b/3 [J. C. Mallinson, *IEEE Trans. Magnet.*, MAG-5, 182 (1969)]. Hence the effective medium thickness for the barium ferrite media is taken to be about 0.1 micrometer, which is approximately 2b/3 for 200 kbpi recording. If a head-to-medium spacing of 25 nm or 1 microinch is used, then a transition width parameter of 8 nm results for barium ferrite media with 250 kA/m (3125 Oe) coercivity.

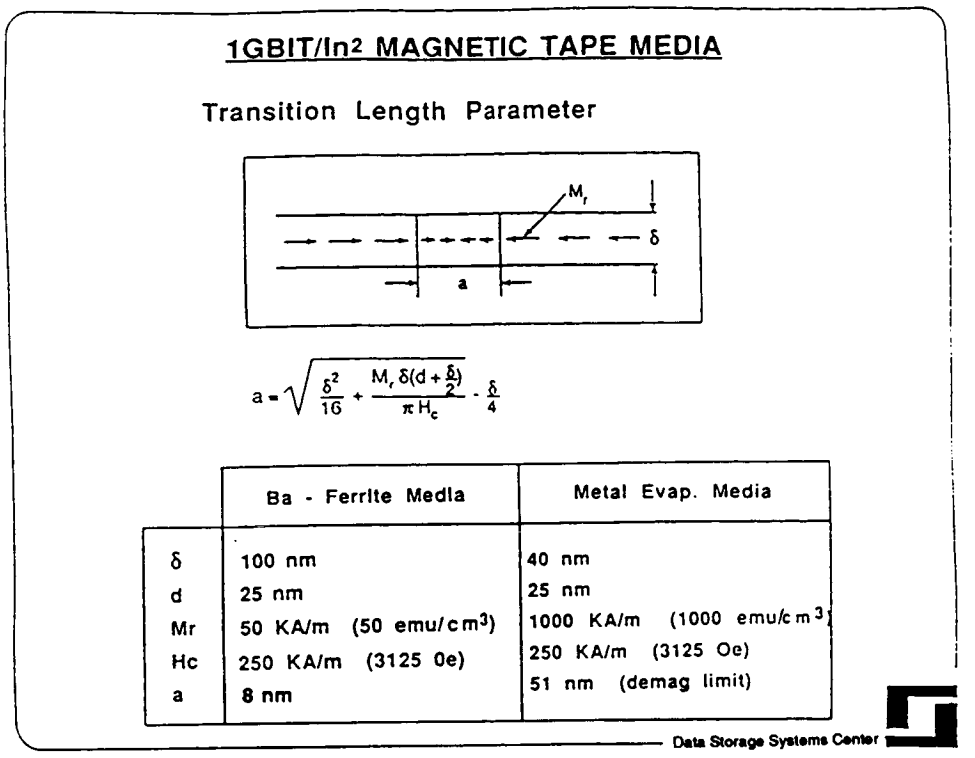


Fig. 6. The transition length parameter for barium ferrite and metal evaporated recording media designed for 200 kbpi recording density.

The metal evaporated media was assumed to be less than 2b/3 in thickness in order that the transition length parameter a be less than b. Unless the media thickness is reduced to 40 nm, demagnetizing effects will limit the recording density to less than 200 kbpi in metal evaporated media with a remanent magnetization of 1000 kA/m (1000 emu/cm^3).

The transition length parameter for a possible 10 Gbit/in^2 thin film disk medium is calculated in Fig. 7 and compared to parameters of media used in the 1 Gbit/in^2 recording demonstration by IBM [C. Tsang et al., *IEEE Trans. Magnet.*, MAG-26, 1689 (1990)]. With a (1,7) code, a recording density of 300 kfci yields a linear bit density of 400 kbpi and a bit spacing of 83 nm. A head-to-medium spacing of 20 nm (0.8 microinch) and a medium thickness of 10 nm (0.5 microinch) ensure that the spacing loss does not become significantly worse than in the 1 Gbit/in^2 demonstration. Both the media remanent magnetization and coercivity are increased, yielding a transition length parameter of 14 nm for the 10 Gbit/in^2 media. This is well below the bit spacing of b = 83 nm.

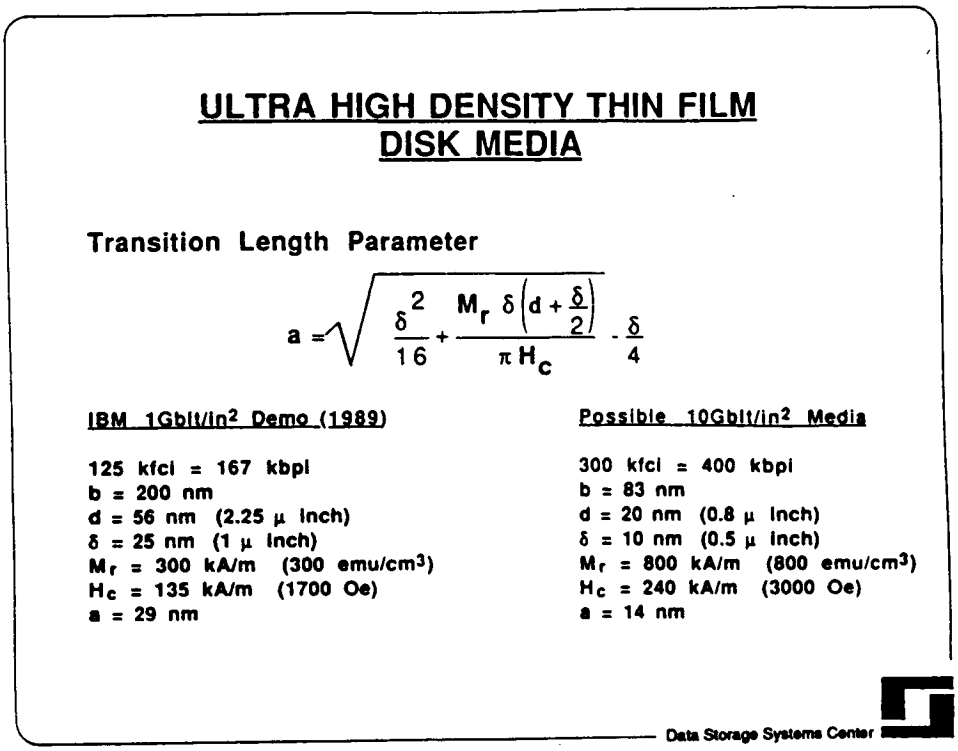


Fig. 7. The recording media parameters for 1 Gbit/in^2 disk recording as demonstrated by Tsang et al. [*IEEE Trans. Magnet.*, MAG-26, 1689 (1990)] and possible future 10 Gbit/in^2 disk recording.

Thin film disk media with parameters similar to those required for 10 Gbit/in^2 recording are currently under development. One such medium is SmCo/Cr, which was described by Velu and Lambeth [E. M. T. Velu and D. N. Lambeth, *IEEE Trans. Magnet.*, Mag-27, 2706 (1992)] at the 1992 Intermag Conference. Velu and Lambeth noted that the $(11\bar{2}0)$ face of SmCo provided an excellent lattice match to the (110) face of Cr. Hence they deposited Cr at low temperatures where the (110) orientation is typically obtained and then deposited SmCo on top of it. The result was a high coercivity medium with reasonable squareness and extremely low noise. The carrier to integrated noise ratio, CINR, and the overwrite performance of representative disks of this media are shown in Fig. 8. The CINR remains above 25 dB out to 50 kfci. The slight roll off above 35 kfci is believed to be due to limitations of the recording head and spacing loss rather than media noise problems, as measurements of the media noise showed no increase with recording density. In spite of the high coercivity of the medium (2320 Oe in Fig. 8), the overwrite performance is very good. Approximately 50 dB of overwrite is obtained when a 7.3 kfci signal is overwritten by a 14.6 kfci signal. Although the media for which data are plotted in Fig. 8 only have coercivity of up to 2320 Oe, Velu and Lambeth have made media with coercivities in excess of 3000 Oe. To date the recording performance of these ultra high coercivity disks have not been tested, as no recording heads with sufficiently high magnetization to prevent saturation have been available to test them.

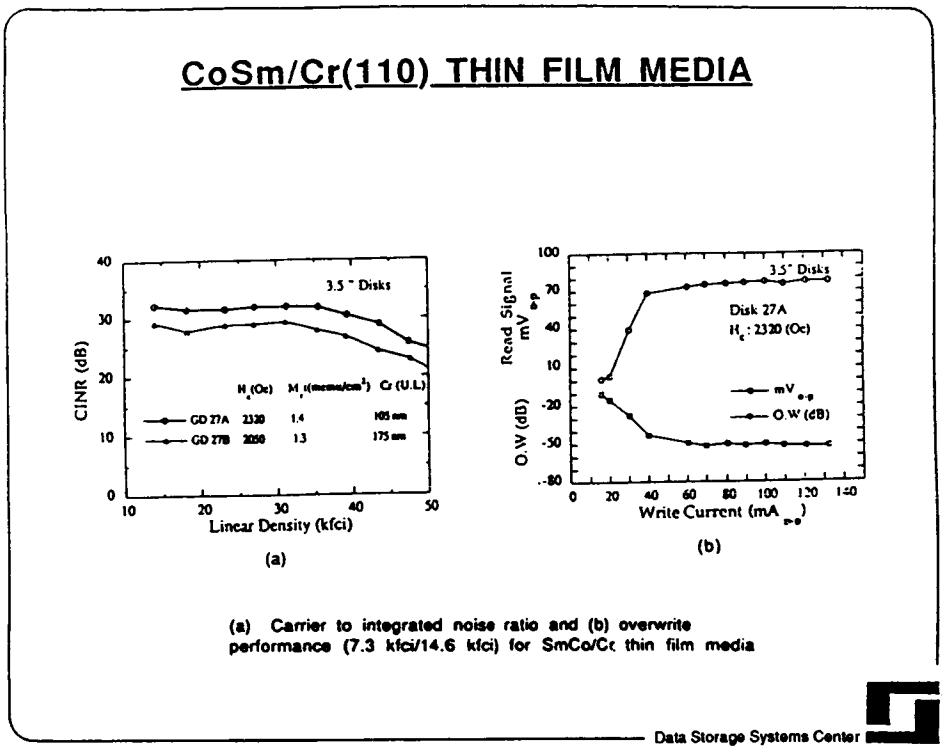


Fig. 8. (a) Carrier to integrated noise ratio and

(b) overwrite performance (7.3 kfci/14.6 kfci) for SmCo/Cr thin film media.

Recording head saturation is a problem which can be solved with higher magnetization soft magnetic alloys. Jeffers [*Proc. of IEEE*, 74, 1540 (1986)] pointed out that metal recording heads typically showed saturation effects when the deep gap field of the head H_g reached about 80% of M_S. By using the Karlqvist equations to describe the longitudinal head fields, setting $H_g = 0.8\ M_S$ and requiring that the longitudinal field be equal to the coercivity at the back of the media, the maximum coercivity media which may be recorded may be calculated and shown to be

$$H_{c,max} = 0.25\ M_S \tan^{-1} [g/2(d+\delta)]$$

This relationship may be used to calculate the maximum coercivity which may be recorded with a given recording head as shown in Fig. 9. It is seen there that, in a 100 nm thick tape medium and using a head-to-medium spacing of 25 nm (1 microinch), the maximum coercivity recording tape media which may be written with a recording head having a gapwidth of 200 nm and a magnetization of 800 kA/m is 138 kA/m (1720 Oe); whereas, if the magnetization were increased to 1600 kA/m, media with coercivity as high as 275 kA/m (3440 Oe) could be used.

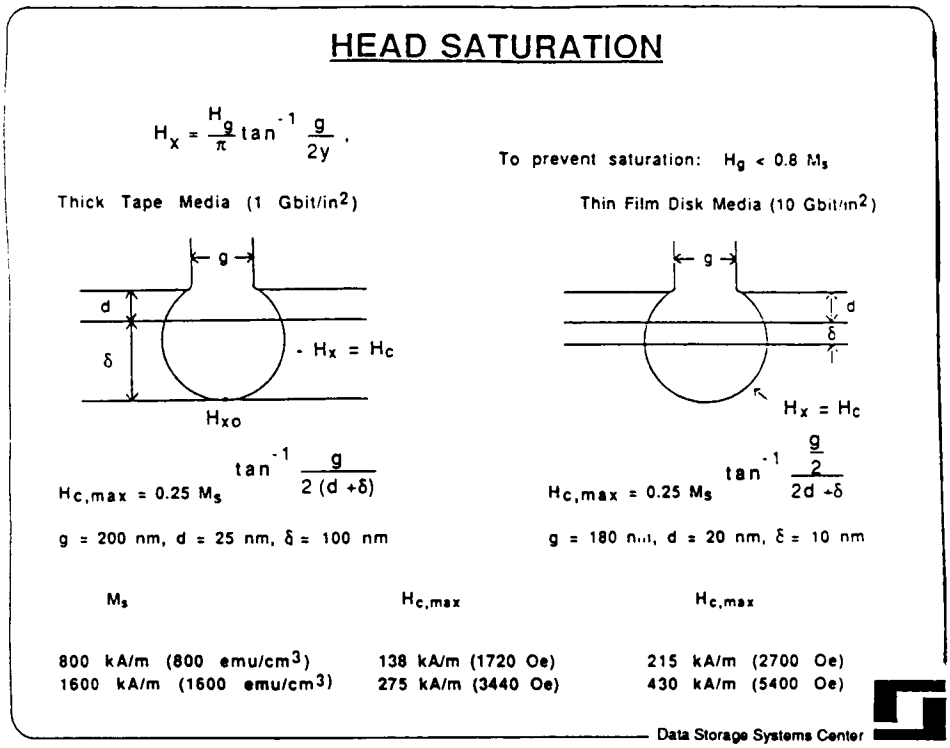


Fig. 9. The maximum coercivity tape and disk media which may be recorded with a given magnetization metal recording head.

For thin film disk media the maximum coercivity is calculated differently. To ensure complete saturation, the longitudinal field contour corresponding to the $H_x = H_c$ contour is made to fully penetrate the thin film media. In this case the maximum coercivity which may be used is

$$H_{c,max} = 0.25\ M_S \tan^{-1} [g/2(2d+\delta)].$$

In this case a head with a gapwidth of 180 nm and made of magnetic material having magnetization of 800 kA/m could be used with media having coercivity as high as 215 kA/m (2700 Oe) if the head-to-medium spacing and medium thickness were chosen to be be 20 nm (0.8 microinch) and 10 nm (0.4 microinch), respectively, as in Fig. 7. Increasing the magnetization to 1600 kA/m would make it possible to use media with coercivity up to 430 kA/m (5400 Oe). Thin media are thus seen to significantly reduce the problem of head saturation.

High magnetization alloys with good soft magnetic properties and high frequency performance appear to be available for ultrahigh density recording requiring a saturation magnetization of 1600 kA/m (B_S = 20,000 Gauss). Multilayer FeAlN/SiO_2 thin films have been made with coercivities of the order of 20 A/m (0.25 Oe) and flat high frequency response to beyond 200 MHz, as shown in Figs. 10 and 11. These materials also exhibit very low magnetostriction and have been shown to be similar to Permalloy in their corrosion characteristics.

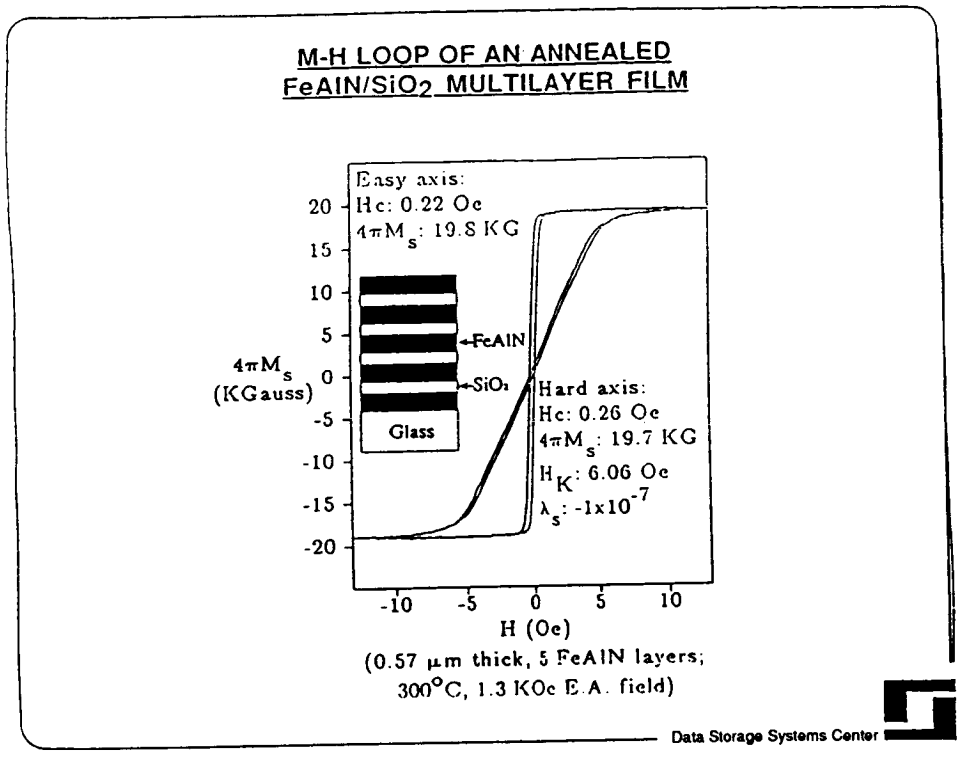


Fig. 10. The M-H loop of a FeAlN/SiO_2 multilayer film.

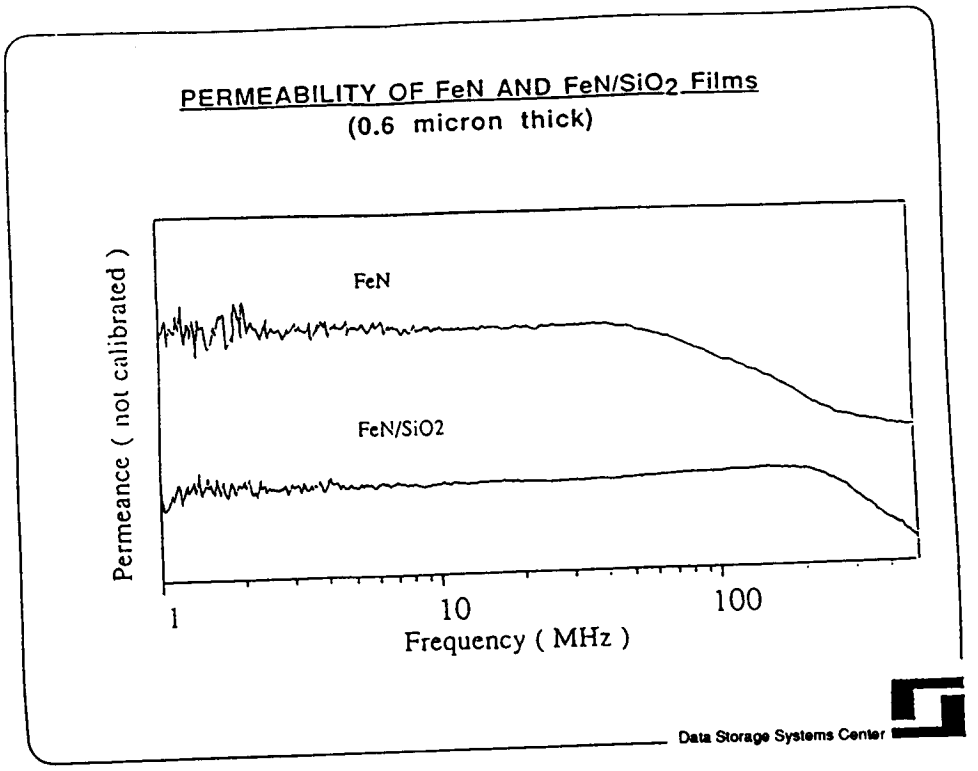


Fig. 11. The relative permeabilities of a 600 nm thick FeN film and a FeAlN/SiO_2 multilayer material as a function of frequency. The vertical scale for the two materials is displaced. Both materials have similar low frequency permeability.

Scaling of Magneto-Optic Recording Technology

A schematic diagram of a magneto-optic recording system is shown in Fig. 12. The recording medium consists of a magnetic thin film with preferred axis of uniaxial anisotropy perpendicular to the film plane. This medium has very high coercivity near room temperature, but low coercivity at temperatures near 200°C. To record on the medium, the beam of a diode laser is focussed onto the medium in the presence of an externally applied magnetic field directed opposite to the initial magnetization direction of the medium and pulsed for a short duration. The energy absorbed by the medium heats it to above 200°C, where the coercivity of the medium is low, and the magnetization in the heated region reverses in response to the applied magnetic field. Thus by controllably pulsing the diode laser, reverse domains, corresponding to bits of information may be recorded into the medium. To read previously recorded information out of the medium, the polar Kerr magneto-optic effect is used. The same diode laser is used, but at lower power output so the energy does not disturb the recorded information. The light from the diode is plane polarized and, upon reflection from the medium, suffers a change in polarization orientation which is dependent upon the direction of magnetization in the medium. This change in polarization state is converted to a change in light intensity by the analyzer and then detected with a photodetector.

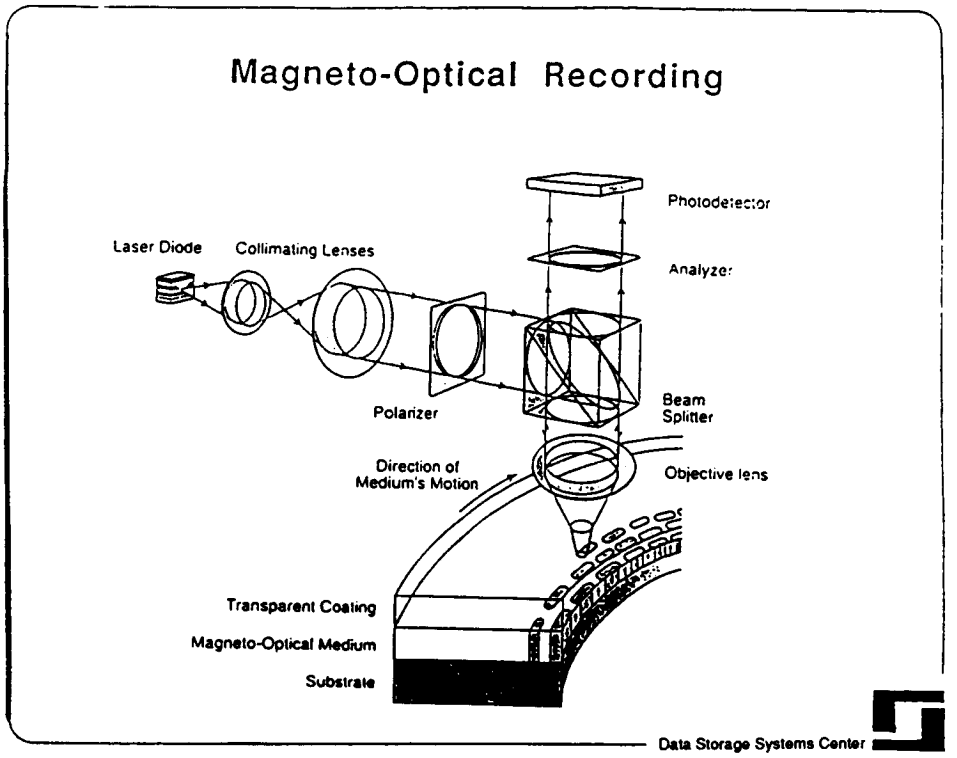


Fig. 12. A schematic diagram of a magneto-optic recording system.

Future magneto-optic recording devices are expected to have considerably higher densities than presently manufactured devices. Increased areal bit density is expected from a number of factors, as shown in Fig. 13. The spot size in magneto-optic drives manufactured today is about 0.8 micrometer; however, in the future spot sizes of 0.2 micrometer are expected. Smaller spot sizes are expected to result from the use of blue, rather than infra-red, light sources and higher numerical aperture objective lenses. In addition, the use of run-length limited pulse-width modulation codes, such as the (2,7) block code, instead of simple pulse-position codes, are expected to increase the linear bit density by a factor of 3. Track density will also be increased by a factor of 2 through the use of shorter wavelength light and by a factor of 1.3 through higher numerical aperture objectives. Moreover, alternative track following servo techniques could reduce the guard band between tracks to the order of the rms deviation of the tracking servo, increasing the track density by another factor of 1.6. Finally, zone bit recording, which is commonly employed in magnetic disk drives today, and packs the bits at equal density on the inner and outer radii of the disk, could increase the total storage capacity by another 50%. Multiplying all these factors together, approximately a factor of 50 improvement in storage density on a magneto-optic disk results, making a storage density of 10 Gbit/in^2. As will be shown later, magnetic super-resolution or optical super-resolution could potentially enable even higher densities.

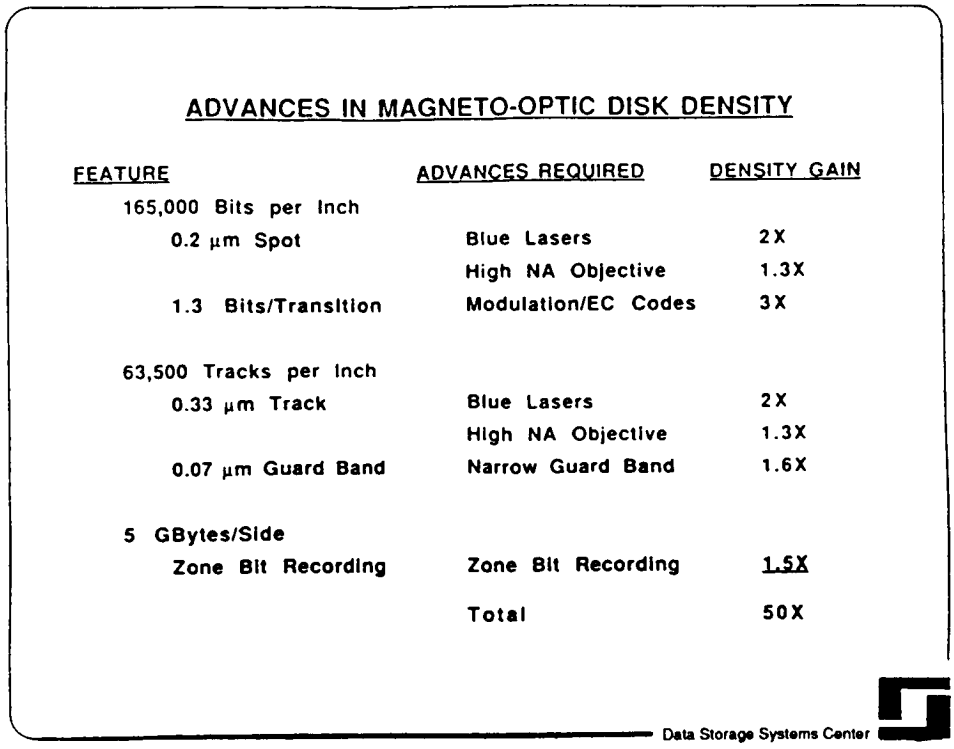


Fig. 13. Factors which are expected to contribute to future increases in the storage density of magneto-optic recording devices.

The increases in bit and track densities which arise from shortening the wavelength of light and increasing the numerical aperture of the objective lens are easily understood by considering the minimum resolvable spot size as determined by the laws of diffraction of light and illustrated in Fig. 14. The minimum linewidth which may be resolved with an optical system is given by

$$D = C\lambda/\sin\theta$$

where λ is the wavelength of the light used and $\sin\theta$ is the numerical aperture of the objective lens. The constant C can have values ranging from about 0.31 to 0.61, depending upon the modulation transfer function which is desired. It may be seen that halving the wavelength reduces the minimum resolvable spot size by a factor of two while also narrowing the minimum track width by a factor of two. Similarly, increasing the numerical aperture of the objective by a factor of 1.3 from the present value of 0.5 to 0.65, causes a similar increase in bit and track densities.

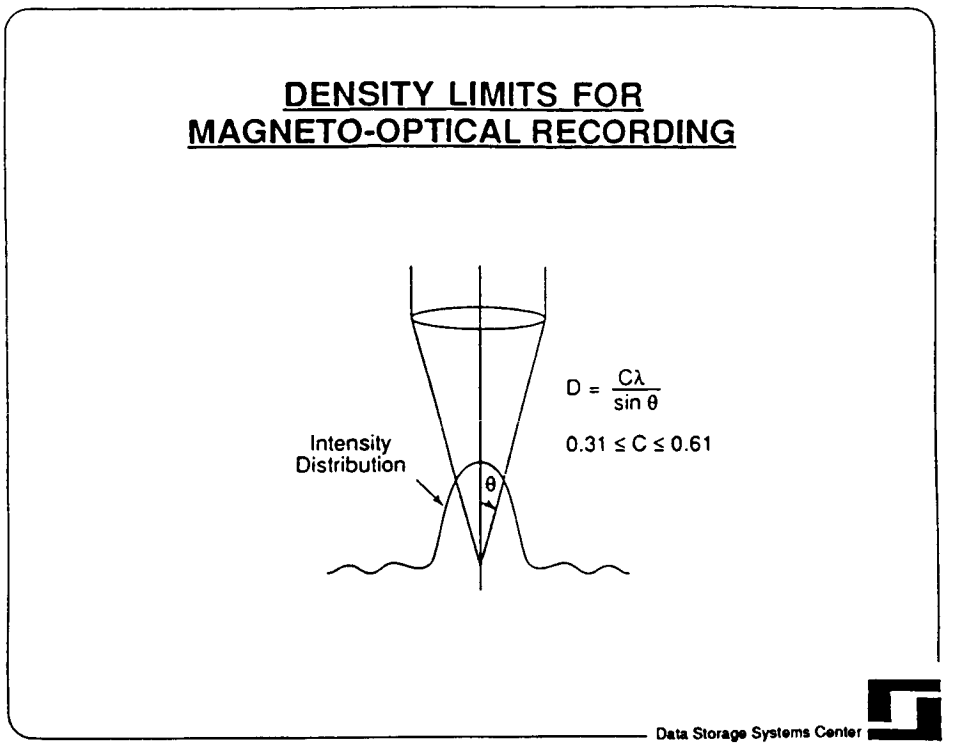


Fig. 14. The minimum resolvable linewidth in an optical system.

Narrowing the guard band between tracks can potentially be done by a number of techniques. Presently manufactured media use grooves between tracks to define the tracks and provide a position error signal, which is used for track following. Current disks use a track pitch of 1.6 micrometers, but a laser beam size of only 0.8 micrometers. Thus there is effectively a guard band 0.8 micrometers wide between tracks. However, the rms deviation of the track following servo on optical drives has been measured to be less than 0.05 micrometers. Thus the groove between tracks is extremely wasteful of the surface area of the disk.

Alternative servo techniques can be expected to be developed which enable one to reduce the guard band. One possibility is to use a sector servo such as that pictured in Fig. 15. With this servo system, no grooves are used. Rather, isolated pits are located at the beginning of sectors and used to provide the position error signal for track following. When the laser is properly positioned, it follows the solid line through the pits and the output signals from the pits are equal in magnitude. However, if the laser is off track and follows the dotted line, the output signal from pit A is larger than the output signal from pit B. On the other hand, if the laser follows the dashed line the output signal from pit B is larger than from pit A. The difference in the output signals from pits A and B can thus be used as a position error signal. Since these pits are not continuous, tracks can be spaced closer together. In principle, the upper edge of one pit could serve as pit A for the upper track while the lower edge serves as pit B for a track immediately below it, effectively doubling the track pitch.

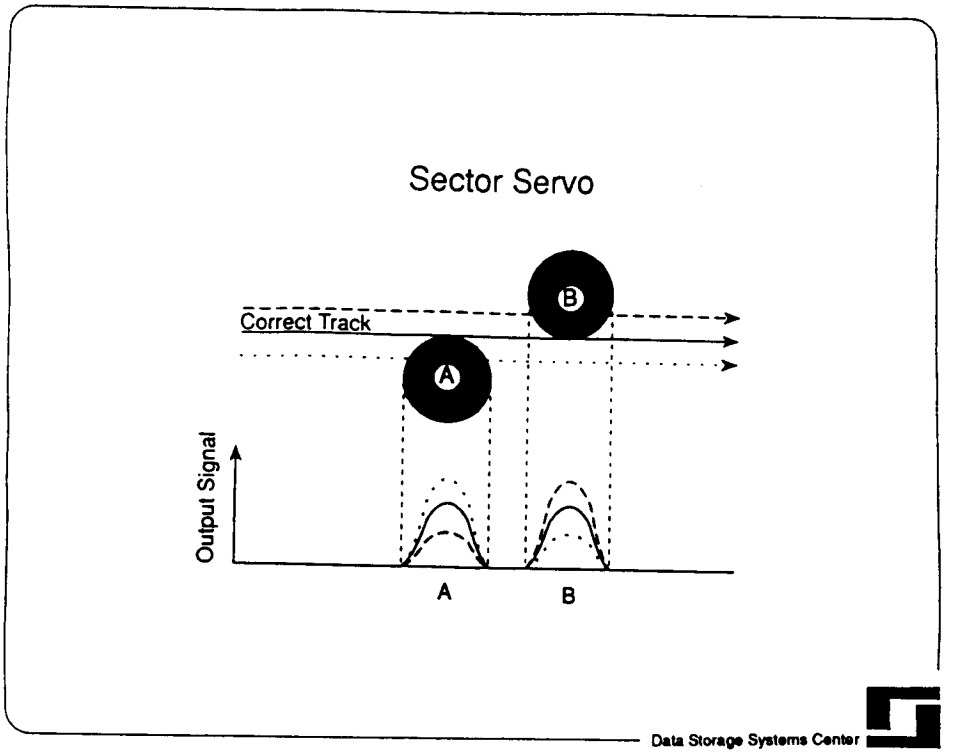


Fig. 15. A possible approach to sector servo which would allow narrower track pitch than continuously grooved media.

Techniques to increase the areal bit density of magneto-optical recording by more than the factor of 50 illustrated in Fig. 13 include magnetic and optical super-resolution. With these techniques it is possible to exceed the limit of resolution determined by focussing optics alone. Magnetic super-resolution makes it possible to write and read marks smaller than the optically resolvable spot size even while using a large head-to-media spacing. Optical super-resolution uses an aperture, smaller than the diffraction limit, in very close proximity to the magneto-optic medium to define the spot size. This latter technique requires a similar head-to-medium spacing as magnetic recording to achieve an equivalent bit spacing.

Magnetic super resolution is achieved by using multilayer exchange coupled media like that shown in Fig. 15 [M. Ohta, et al., *J. Magn. Soc. Jpn.*, 15, Suppl. No. S1, 319 (1991)]. The properties of the three layers are summarized in Table 1.

Table 1 Magnetic Properties of the Films used to Make Up the Magnetic Super Resolution Disk.

Layer	Material	Thickness (Å)	T_C (°C)	H_C (KOe)
readout	GdFeCo	300	>300	0.1
switching	TbFeCoAl	100	~140	
recording	TbFeCo	400	~250	>10

In this technique, recording is performed with a magnetic head flying close to the media to provide a magnetic field which is modulated at the recording frequency of the laser. Since the magnetic field, and not the laser spot size, determines whether the magnetization is directed upward or downward, mark size can be smaller than the laser spot. The information to be stored is recorded into all three exchange coupled layers shown in Fig. 16.

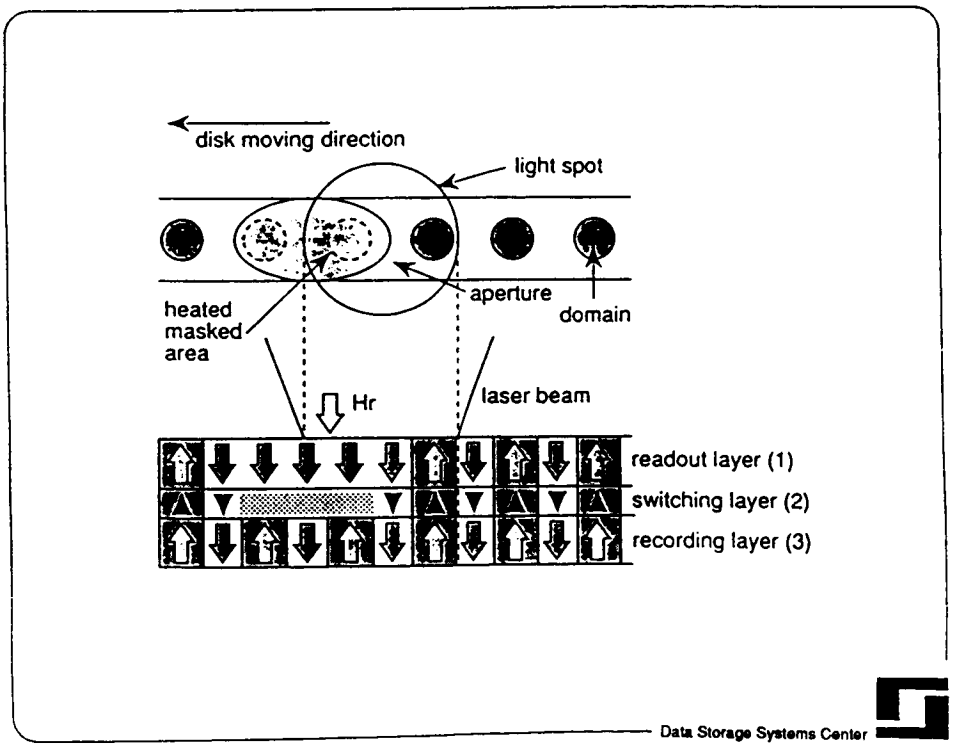


Fig. 16. Readout by Magnetic Super Resolution [M. Ohta, et al., *J. Magn. Soc. Jpn*, 15, Suppl. No. S1, 319 (1991)].

To readout the information, a magnetic field H_r is applied downward as shown in Fig. 16. Sufficient energy is applied from the laser to reach the Curie temperature T_C of the switching layer. This breaks the exchange coupling between the readout and recording layers and allows any downward directed domains recorded in the readout layer to switch. Since the light spot encompasses the switching domain, when the domain switches, there is a change in the net magneto-optic rotation of the light, which may be detected. Hence, domains smaller than the beam diameter can be detected.

The readout mechanism is nondestructive, because when the switching layer cools below its Curie temperature, it again exchange couples the recording and readout layers. Since the recording layer has a very large coercivity and the readout layer has a small coercivity, the exchange coupling forces the readout layer to replicate the pattern in the recording layer.

Optical super-resolution is achieved by defining a very small aperture and bringing the magneto-optic medium into the near-field region of the aperture. In such a case the resolvable spot size is determined primarily by the aperture size, which can be much smaller than can be obtained with focussing optics. Recently, this technique was used by Betzig et al. [1992] to demonstrate recording and readback of an array of magnetic domains having a density of 45 Gbit/in^2. Even higher densities are believed possible.

Betzig et al. [1992] used an optical fiber which had been drawn down so it had approximately a 20 nm diameter at the end to define the aperture as illustrated in Fig. 17. The sides of the fiber were coated with an aluminium reflector in the region near the aperture to prevent the effanescent light energy from escaping where the fiber was too small to support an optically guided wave. By using polarizing optics and scanning the fiber above the surface of a Co/Pt multilayer film, magnetic domains as small as 60 nm were written and readback as shown in Fig. 18. By using a flying head this technique could, in principle, be used to achieve storage densities above 100 Gbit/in^2.

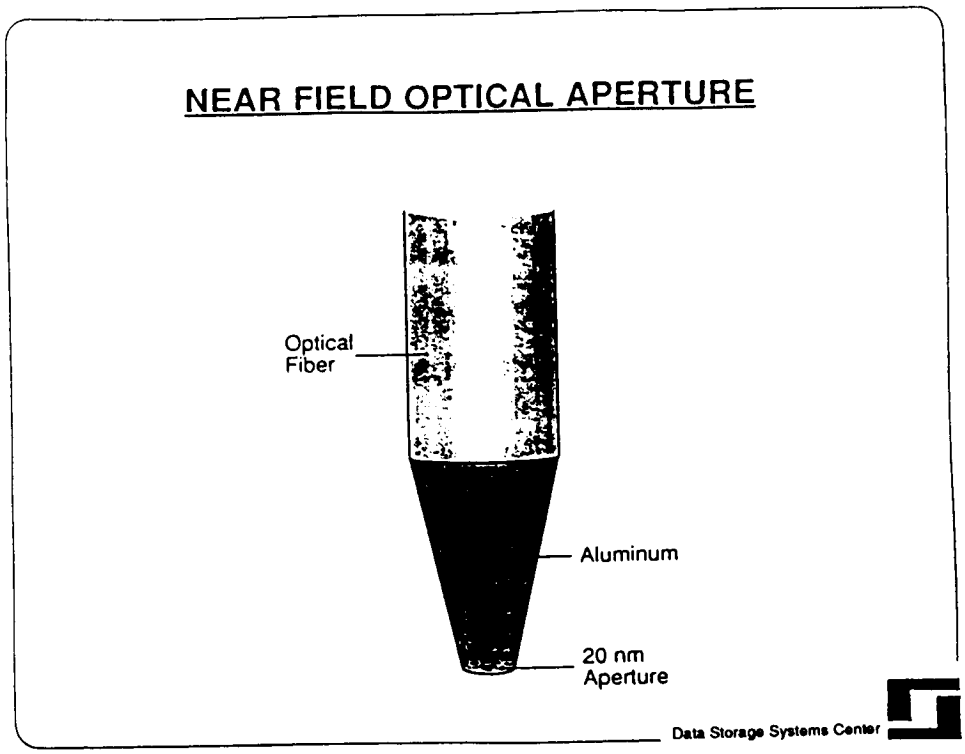


Fig. 17. An optical fiber is drawn down to produce a 20 nm aperture.

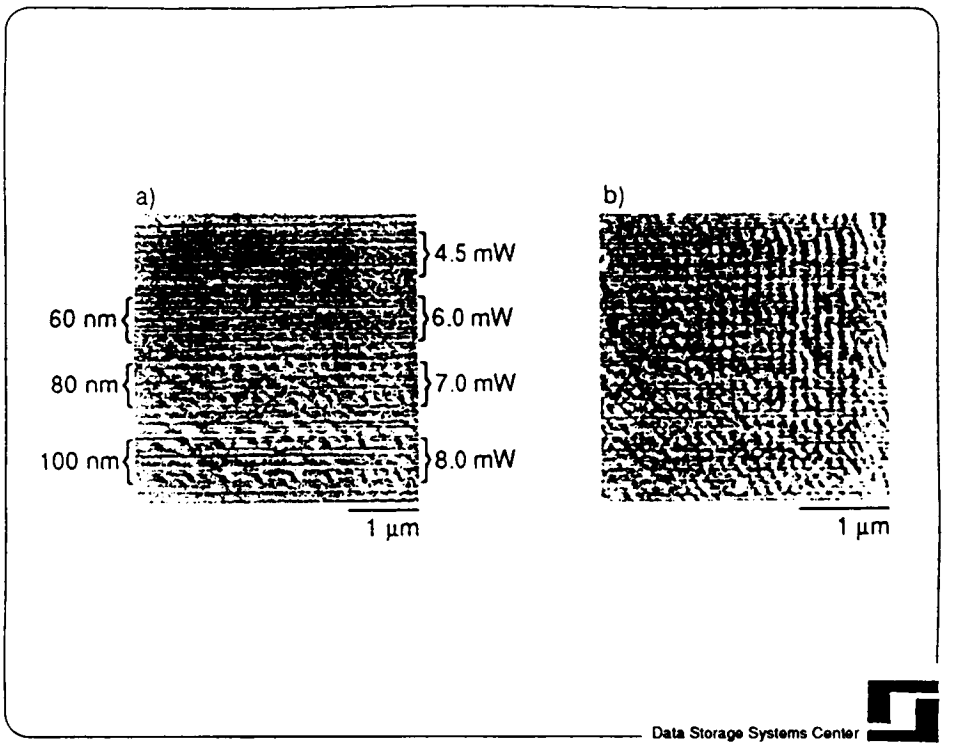


Fig. 18. (a) Submicrometer domains recorded using optical super resolution at different write powers and head-to-media spacings.

(b) A 20 X 20 array of domains with 120 nm periodicity in both directions, correpsonding to a storage density of ~45 Gbits/in^2 [Betzig, et al., *Appl. Phys. Lett.*, 61, 142 (1992)].

Conclusions

Magnetic recording technology has increased storage density by a factor of 50,000 over the past 35 years since it was first used in a disk format for computer data storage; however, there is no sign this rapid pace is slowing. Indeed fundamental limits, set by superparamagnetism are estimated to be several orders of magnitude from where we are today. Recent product announcements and developments in research labs suggest that the rate of progress is likely to accelerate. Storage densities of over 1 GBit/in^2 are likely by the end of this decade and densities of 10 GBit/in^2 appear likely in the early 21st Century. Multilayer thin film media, thin film write heads utilizing high magnetization thin film multilayer materials and magnetoresistive read heads are expected to be among the new technologies which make this occur.

Magneto-optic recording has only recently been introduced to the marketplace. It offers high areal density (about 2 X 10^8/in^2) on removable disk media. Presently the media costs about $0.50 per megabyte. Although areal storage densities are high, the volumetric storage densities are less than in rigid magnetic disk drives containing multiple disks. Data rates and access times are currently about a factor of two or three times slower than on rigid disk drives, and costs at the drive level are higher than for rigid disks. The technology, however, is young and both cost reductions and improvements in storage density and performance are expected. Higher numerical aperture objectives, improved signal processing, improved track following servos and shorter wavelength laser sources combined with improved media for short wavelength recording are expected to increase the storage density. Over an order of magnitude improvement is likely. Both higher bit density and higher spindle rotation rates will lead to higher data rates. The higher spindle rotation speed will also shorten the latency time. A removable 2 inch magneto-optic disk which will store a few GBytes of data with data rates and access times only slightly poorer than those of rigid magnetic drives appears likely by the year 2000. The use of magnetic and/or magneto-optic super-resolution could eventually lead to magneto-optical drives with areal storage densities well beyond 10 Gbit/in^2.

National Media Laboratory
Media Testing Results

Bill Mularie

Gary Ashton

National Media Laboratory
3M Center
Building 235-3B-30
St. Paul, MN 55144-1000

Presentation Topics:

1. Overview of National Media Laboratory
2. Results of D-1 Testing

National Media Laboratory

U. S. Government Users:

Industry:

- 3M Storage Media Laboratory
- 3M Hardware and Electronic Resources
- Ampex Recording Systems
- DataTape

University:

- Center for Magnetic Recording Research
  University of California, San Diego
- Center for Materials for Information Technology
  (MINT)
  University of Alabama

NML
NATIONAL MEDIA LAB

P.O. Box 3[illegible] St. Paul, MN 55[illegible]
(612) 78[illegible], Fax (612) 78[illegible]

Purposes of NML Activities

1. Help set reasonable performance/reliability expectations for advanced recording hardware and media
2. Give the PO's data to assist in making program choices
3. Help translate government program needs to recorder/media industry
4. Irritate the industry into doing better for data recording

NML
NATIONAL MEDIA LAB

P.O. Box 3[illegible] St. Paul, MN 55[illegible]
(612) 78[illegible], Fax (612) 78[illegible]

Critical Issues in Government Mass Storage

Skyrocketing requirements:

- **Platform data rates > 1 Gigabit/sec.**
- **Storage of terabytes/day.**
- **Archive of > 10 years.**
- **Government leads industry by 3-5 years.**

NML
NATIONAL MEDIA LAB
P.O. Box 33015 St. Paul, MN 55133-3015
(612) 779-0440, fax (612) 779-0440

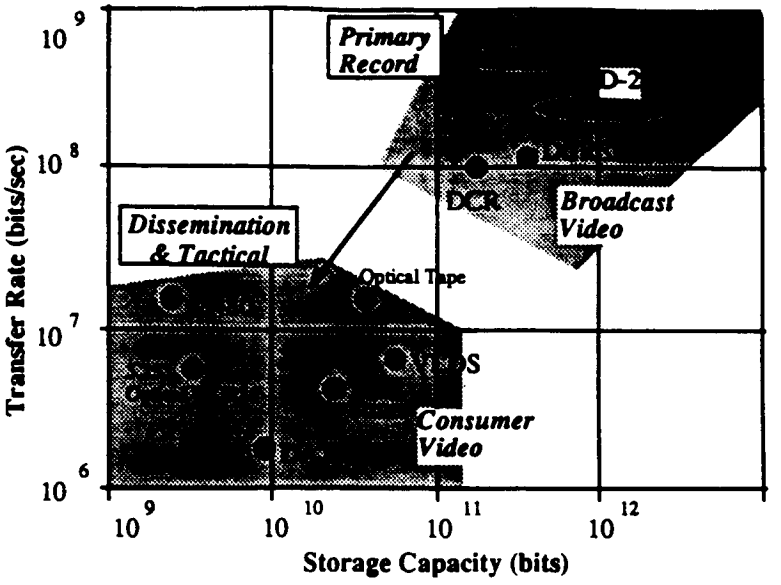
Data Storage Comparisons

Requirement: Data Rate 800 MBits/Sec

Performance of Media

	Data Rate (MBits/Sec)	Capacity (MBits)	1 Day Storage (Units)	Parallel Hardware Requirement
Floppy Disk 	0.5	16	4,200,000 Diskettes	1600 Drives
12" Optical Disk 	2.4	6,000	11,500 Disks	334 Drives
Magnetic Cassette (DTC) 	100	232,000	298 Cassettes	8 Recorders

NML
NATIONAL MEDIA LAB
P.O. Box 33015 St. Paul, MN 55133-3015
(612) 779-0440, fax (612) 779-0440



NML
NATIONAL MEDIA LAB

Principal "Drivers" for Advanced Recording Systems

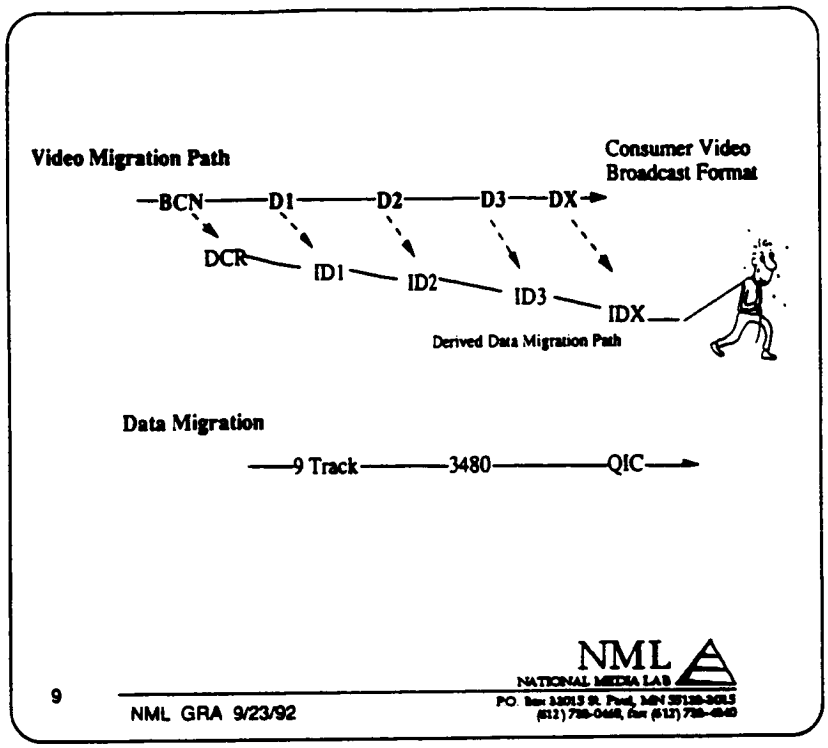
Video

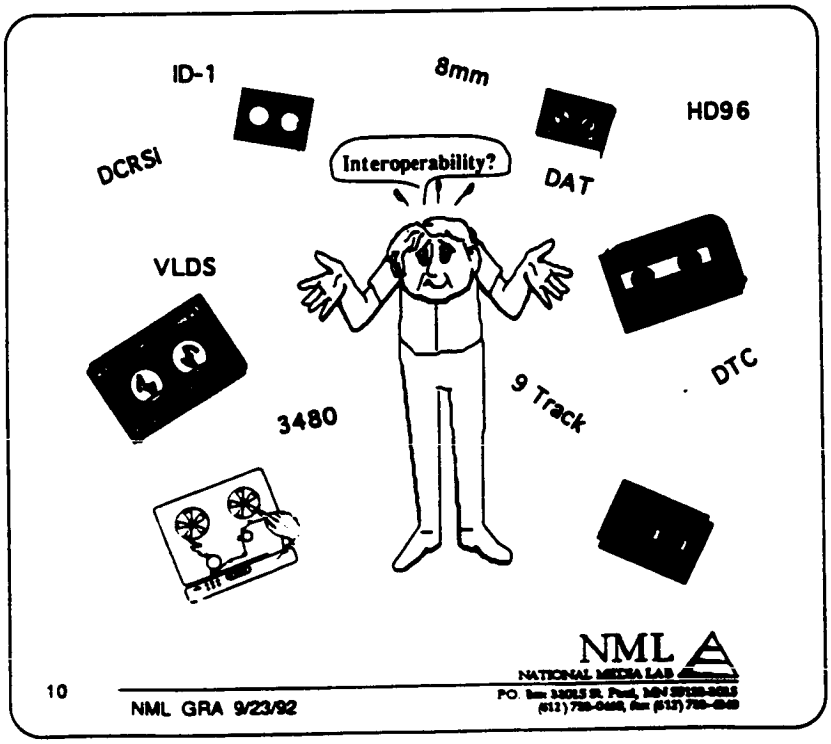
- Increasing Capacity/Cassette
- Decreasing System Form Factor

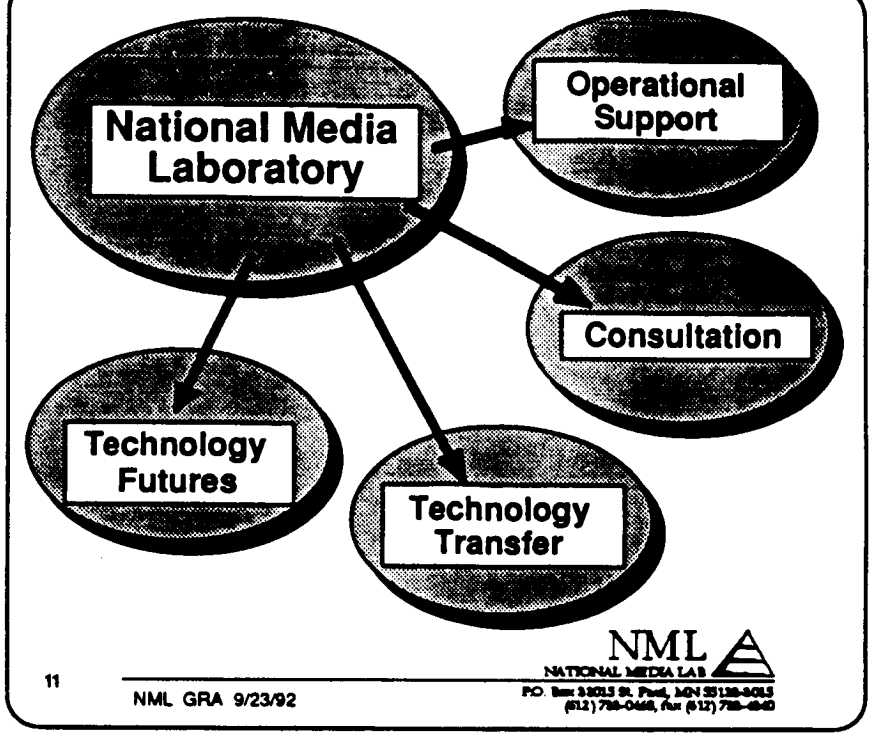
Data

- Reliability
- Environmental Stability
- Archivability

NML
NATIONAL MEDIA LAB


Video Migration Path
Consumer Video Broadcast Format
BCN
D1
D2
D3
DX
DCR
ID1
ID2
ID3
IDX
Derived Data Migration Path
Data Migration
9 Track
3480
QIC
NML
NATIONAL MEDIA LAB
9
NML GRA 9/23/92


ID-1
8mm
HD96
DCRSI
Interoperability?
DAT
VLDS
DTC
9 Track
3480
NML
NATIONAL MEDIA LAB
10
NML GRA 9/23/92



NML
NATIONAL MEDIA LAB
P.O. Box 30015 St. Paul, MN 55130-0015
(612) 758-0468, Fax (612) 758-4640

National Media Laboratory Testing Results:

Moisture Content of D-1 Tapes When Delivered

NML
NATIONAL MEDIA LAB
P.O. Box 30015 St. Paul, MN 55130-0015
(612) 758-0468, Fax (612) 758-4640

Overview

1. Why Look at D-1 ?
2. Goals
3. Information Available
4. Moisture Content
5. Conclusion

NML
NATIONAL MEDIA LAB

P.O. [illegible]
(612) 758-[illegible], Fax (612) 758-[illegible]

1. Why Look at D-1 ?

New Format With Little Experience And Data

- MIL-STD-2179
- ANSI X3.175-1990 (Tape Format)
- ANSI X3B.5/90-133 (Tape Cartridge)

Professional Video Use ≠ Data Storage Use

Military Use ≠ Commercial Use

ATARS And JSIPS Use Of D-1 Cassettes

NML
NATIONAL MEDIA LAB

P.O. [illegible]
(612) 758-[illegible], Fax (612) 758-[illegible]

2. Goals

Evaluate D-1 Media And Cassettes

ATARS & JSIPS

Determine Environmental Window

Measure Media Properties

Improve Packaging

NML
NATIONAL MEDIA LAB
P.O. Box 33015 St. Paul, MN 55133-3015
(612) 736-6168, fax (612) 736-6140

3. Information Available

Initial Evaluation of D-1 Tape and Cassette Characteristics

Packaging Plan for D-1 Cassettes

Packaging Tests of Commercial D-1 Cassettes and Cases

Relative Humidity of Sony and Ampex D-1 Tapes when Delivered

Resistivity Characteristics of Ampex and Sony D-1 Tape

Modulus (Stress-Strain Curves) of Sony and Ampex D-1 Tape

Width and Weave Characteristics of Sony and Ampex D-1 Tape

Shrinkage of Sony and Ampex D-1 Tapes

Friction Characteristics of Ampex and Sony D-1 Tapes

Vibrating Sample Magnetometer (VSM) Tests on Sony and Ampex D-1 Tape

M-H Meter Tests on Sony and Ampex D-1 Tape

Surface Roughness of Sony and Ampex D-I Tape

Coating and Substrate Thickness of Sony and Ampex D-1 Tape

Stiffness of Sony and Ampex D-1 Tapes

Magnetic Print-Through Effects in Sony and Ampex D-1 Tapes

Thermal and Hygroscopic Time Constants of Sony and Ampex D-1 Tape Cassettes

Data Diskette of: Commercial D-1 Cassettes & Media Test Data: 1990 - 1991 Data

Data Diskette of: Commercial D-1 Cassettes, Media, & Packaging Test Data: 1991 - 1992 Data

To Request Reports Contact:

National Media Laboratory

P.O. Box 33015

Saint Paul, MN 55133-3015

Phone: (612) 736-6183

NML
NATIONAL MEDIA LAB
P.O. Box 33015 St. Paul, MN 55133-3015
(612) 736-6168, fax (612) 736-6140

4. Interest in Moisture Content

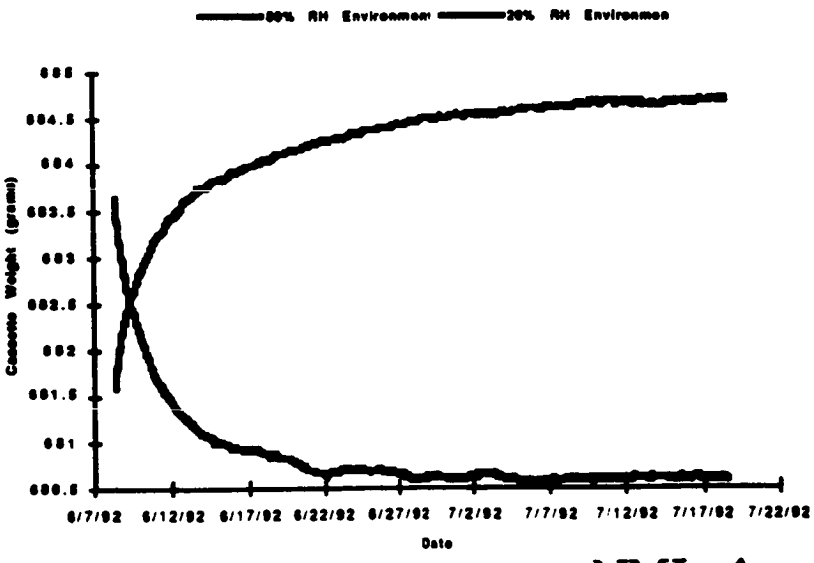
Conditioning Needed ?

Time To Condition ?

Archive Evaluation

NML
NATIONAL MEDIA LAB
P.O. Box 63015 St. Paul, MN 55163-0015
(612) 758-0468, fax (612) 758-4649

4.1 Behavior of D-1 Cassettes



NML
NATIONAL MEDIA LAB
P.O. Box 63015 St. Paul, MN 55163-0015
(612) 758-0468, fax (612) 758-4649

4.2 Solution to Diffusion Equation

$W(r,t) = Wa + (Wi - Wa)\, F(r,t)$

Wa = Equilibrium Weight in Ambient Humidity

Wi = Initial Sample Weight

$F(r,t=0) = 1$

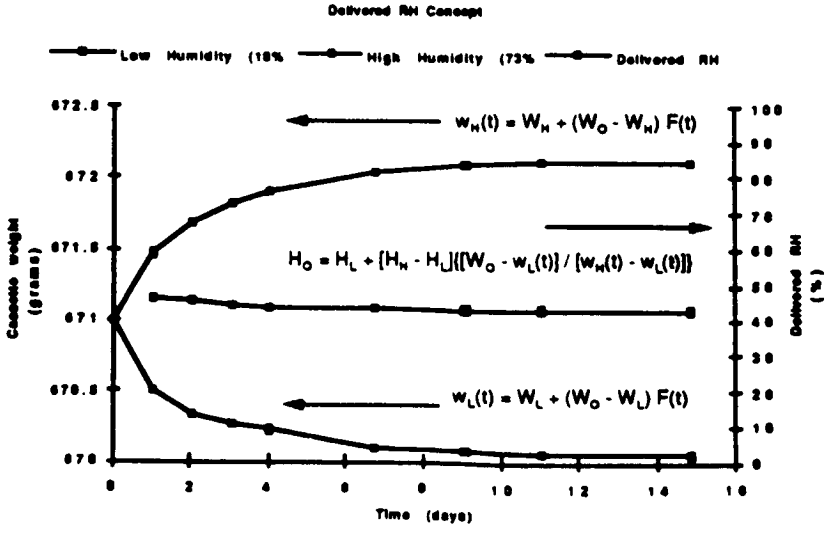
$F(r,t=\infty) = 0$

$\partial F(r,t=\infty)/\partial t = 0$ (Steady State at $t=\infty$)

Sample Geometry Is In F(r,t)

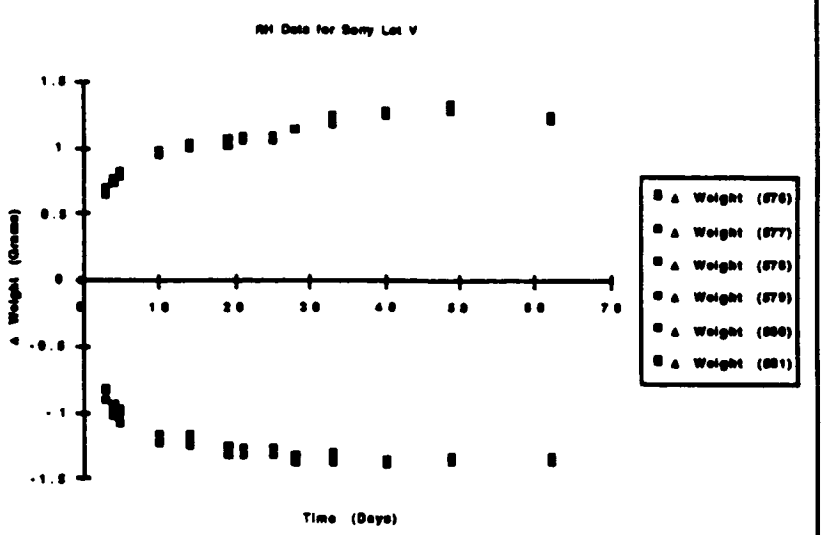
NML
NATIONAL MEDIA LAB
P.O. Box [illegible] St. Paul, MN [illegible]
(612) [illegible], Fax (612) [illegible]

4.3 Moisture Content Theory



NML
NATIONAL MEDIA LAB
P.O. Box [illegible] St. Paul, MN [illegible]
(612) [illegible], Fax (612) [illegible]

4.4 Moisture Content Data



NML
NATIONAL MEDIA LAB

P.O. Box [illegible], St. Paul, MN [illegible]
(612) 786-[illegible], fax (612) 786-[illegible]

4.5 Moisture Content Results

		Relative Humidity (%)	
Size	Lot	Average	Std. Dev.
Ampex Large	J	52.60	1.8
Ampex Medium	X	52.63	0.6
Ampex Medium	Y	55.14	1.4
Ampex Medium	Z	51.10	1.3
Sony Large	L	54.05	1.8
Sony Medium	T	45.90	2.7
Sony Medium	U	45.48	4.0
Sony Medium	V	52.6	1.0

NML
NATIONAL MEDIA LAB

P.O. Box [illegible], St. Paul, MN [illegible]
(612) 786-[illegible], fax (612) 786-[illegible]

5.0 Conclusion

NML's Charter

D-1 Media Information is Available

Moisture Content Determination

Basic Understanding of Tape Moisture Experiment

NML
NATIONAL MEDIA LAB
P.O. Box 63015 St. Paul, MN 55163-0315
(612) 758-0468, fax (612) 758-4640

1992 NASA GSFC Conference on Mass Storage Systems and Technologies

Grand Challenges in Mass Storage "A System Integrator's Perspective"

Dan Mintz
8500 Leesburg Pike, Suite 402
Vienna, Virginia 22182
(703) 893-1030
Fax: (703) 893-1069
70322.1065@CompuServe.com

Richard Lee
Post Office Box 1293
Ridgewood, New Jersey 07451
(201) 670-6620
Fax: (301) 670-7814
73130.2011@CompuServe.com

September 23, 1992

What are these Grand Challenges?

- Develop more Innovation in Approach
- Expand the I/O Barrier
- Achieve Increased Volumetric Efficiency & Incremental Cost Improvements
- Reinforce the "Weakest Link" -Software
- Implement Improved Architectures
- Minimize the Impact of "Self-Destructing" Technologies

WJ CULVER CONSULTING
A SYSTEMS INTEGRATION COMPANY

Data Storage Technologies, Inc.

Grand Challenges in Mass Storage
"A System Integrator's Perspective"

Our Definition of Mass Storage

- We Define Mass Storage as any Type of Storage System Exceeding 100 GBytes (0.1 TB) in Total Size (not off-line), Under the Control of a Centralized File Management Scheme

The Growing Importance of Systems Integrators

- Potential Systems Solutions are Becoming Increasingly Complex
- Open Systems Architectures Allow Multi-Vendor Solutions
- In-House Technical Staff are Tied Up Making the Current Technology Work on a Daily Basis
- In-House Technical Staff Members Have Difficulty Keeping Up with New and Alternative Technologies

WJ CULVER CONSULTING
A SYSTEMS INTEGRATION COMPANY

Data Storage Technologies, Inc.

Grand Challenges in Mass Storage
"A System Integrator's Perspective"

Todays High Performance Computing Environment

- A Hodge Podge of Many Different Types of CPU;
- Vector
- Scalar
- Parallel & Massively Parallel
- CISC
- RISC
- Visualization Engines

Todays High Performance Computing Environment (cont'd)

- Interconnection by Elaborate Networking Schemes;
- HyperChannel
- FDDI/CDDI & ATM
- HiPPI
- Ethernet & Token Ring
- Kluge

WJ CULVER CONSULTING
A SYSTEMS INTEGRATION COMPANY

Data Storage Technologies, Inc.

Grand Challenges in Mass Storage
"A System Integrator's Perspective"

Todays High Performance Computing Environment (cont'd)

- Volumes of Bitfile Data are Being Produced at Rates Beyond our Wildest Hallucinations
- Local and Network Disk, and Tape Systems are Overwhelmed
- Dedicated and Intricate Software Schemes Have Been Developed to Manage Data
- The Growing Impact of Scientific Visualization
- New Fiscal Realities

How do we Develop More Innovation in Approach?

- To "Innovate" Requires Abandonment of Many Practices of the Past
  "New Challenges Require New Thinking"
- CPU Price/Performance Capabilities have become a "Double edged sword"
  "Desktop Supercomputers with PC I/O Ports"

WJ CULVER CONSULTING
A SYSTEMS INTEGRATION COMPANY

Data Storage Technologies, Inc.

Grand Challenges in Mass Storage
"A System Integrator's Perspective"

How do we Develop More Innovation in Approach? (cont'd)

- Mass Storage Solutions Require a Coordinated Effort by all Facets of the Data Center
  "Plugging in the latest-greatest box and software buys only short term results"

- Cost Factors Drive the Most Effective Solutions Today
  "Doing more with less has spawned innovative thinking across the board"

Expanding the Input/Output Barrier

- I/O Capabilities Must Begin to Keep Pace With Processor Speed
  Processor Power has Increased 25% Per Year (CAGR), While I/O Rates Have Remained at .250 - 7 MB/s (with Few Exceptions) for Many Years.
- RAID, DASD-like devices, 19mm & 1/2" Helical Scan Tape, and Other New Storage Systems Cannot Achieve Their Potential w/o Solving the I/O Bottleneck.
  Operating System (O.S.) Software and Low-Bandwidth Peripheral Channels Must Undergo Significant Improvements to Meet the Challenges of the '90's

WJ CULVER CONSULTING
A SYSTEMS INTEGRATION COMPANY

Data Storage Technologies, Inc.

Grand Challenges in Mass Storage
"A System Integrator's Perspective"

Expanding the Input/Output Barrier (cont'd)

- HiPPI, FDDI, CDDI and ATM Offer Hope In Increasing the Bandwidth of Interconnecting Peripherals, But Do Not Solve the O.S. Software or I/O Channel Limitation Problems.
  Direct Connection to High Bandwidth Internal Buses, and More Simplistic O.S. I/O Calls are Required.

Achieving Volumetric Efficiency & Incremental Cost Improvements

- Increasing Volumetric Efficiency of Storage Systems Reduces Operations and Transportation Costs Dramatically.
  RAID, DASD-like Devices and Helical Scan Tape Provide Orders-of-Magnitude of More Storage Capacity per Square Foot With Increased Bandwidth Thrown in For Virtually No Cost!
- DASD-like Devices and Helical Scan Tape Incrementally Reduce the $/MB of Capacity to a Fraction of More Traditional Devices, while also Providing More Capacity Per Unit Volume and Higher I/O Bandwidths.
  Almost Like Having a "Free Lunch"

WJ CULVER CONSULTING
A SYSTEMS INTEGRATION COMPANY

Data Storage Technologies, Inc.

Grand Challenges in Mass Storage
"A System Integrator's Perspective"

Reinforcing the "Weakest Link" - Software

- Hierarchical File Management Software Packages are Available From Many Manufacturers. All are Unique; Many are Proprietary; and Some Comply With Emerging Standards (IEEE MSRM 4.0, OSI ,etc.)
  Those Developed by CPU Manufacturers Are The Most Mature (Cray, IBM, etc.)
  Those Developed by Independents and Small Companies Offer the Most Features and Benefits (UniTree, EPOCH, E-Mass, etc.), but are also the Most Immature and Risky from a Business Perspective.

Reinforcing the "Weakest Link" - Software (cont'd)

- The Newest IEEE Mass Storage Reference Model (Version 5.0) Potentially Provides the Means to Tame the Hierarchical File Management Software (FMS) Beast.
  Logical Layers and Task Partitioning Unbundles the Entire Package from one Provider. Key Segments Can be Provided from Developers with Specialized Expertise i.e. Security, PVR's, Bitfile Movers. etc.
- Government Mandates (POSIX, GOSIP and OSI) Must be Tempered Against Established Practices and Protocols i.e. TCP/IP

WJ CULVER CONSULTING
A SYSTEMS INTEGRATION COMPANY

Data Storage Technologies, Inc.

Grand Challenges in Mass Storage
"A System Integrator's Perspective"

Implementing Improved/Advanced Hardware Architectures

- Dedicated File Server CPU's (Mainframe, Mini-Super, and Supercomputers) are Too Expensive and Inefficient to Solve Current and Future Mass Storage Requirements.
  HiPPI, FDDI and ATM Fabrics Provide for Direct Interconnection of Source-to-Sink in Data Intensive Environments
- The "Redundant Array of Inexpensive/Independent Whatevers" Concept Can be Applied to Many Facets of the Data Center.
  e.g. Disk (RAID) and Tape (RAIT) Drives, Independent Computers (RAIC/Clustering), and Data Centers themselves (RAIR).

Minimizing the Impact of "Self-Destructing Technologies"

- Revolutionary Advances in Computer Technology have Produced a Nasty By-Product known as; "Self-Destructing Technology".
  Pursuit of the latest, greatest technological solution for each new program has blinded many to the fact that a significant portion of the technology used in the last program has "self-destructed", while no one was paying attention.
- The Balance of Maturity in an Approach vs. Maintaining One's Technological Edge Produces Serious Conflicts in How to Proceed.
  i.e. Running the COTS Juggernaut

WJ CULVER CONSULTING
A SYSTEMS INTEGRATION COMPANY

Data Storage Technologies, Inc.

Grand Challenges in Mass Storage
"A System Integrator's Perspective"

Conclusions

- In Order to Survive into the Future, a New Order Must Emerge in the way we Develop and Manage Technology for Computing and Data Storage.

- Systems Integrators Face Grand Challenges in Adapting Technology to Meet Their Customers' Wide Variety of Needs. ***Business as Usual Will Not Work.***

- Programs like EOSDIS will Force a New Paradigm on the Computer Marketplace. ***Manufacturers and Systems Integrators Should Not Ignore the Fate of the Dinosaur in Respect to Change***

WJ CULVER CONSULTING
A SYSTEMS INTEGRATION COMPANY

Data Storage Technologies, Inc.

Kodak Phase-Change Media For Optical Tape Applications

Yuan-sheng Tyan, Donald R. Preuss, George R. Olin, Fridrich Vazan, Kee-chuan Pan, and Pranab. K. Raychaudhuri

Electronic Imaging Storage Research Laboratories, Eastman Kodak Company, Rochester, New York 14650-2017

INTRODUCTION

The amount of data generated and handled in our lives is increasing at an astonishing rate. The need to store, manage, and in some cases archive this ever-increasing amount of data has become a challenge. Advances in conventional methods of data storage such as paper, microfilm, magnetic tape, and magnetic disk will solve some of these problems, but it has become increasingly apparent that other solutions are desirable. One such possible solution is optical tape. It offers the potential of being a high-capacity, low-cost, and low-maintenance technology for not only storing but at the same time archiving data. The introduction of CREO's optical tape recorder[1] demonstrated the technical feasibility of optical tape systems. Advances in hardware and media in costs and performance are needed to make this technology widely accepted in the future.

The SbInSn phase-change write-once optical medium developed by Eastman Kodak Company is particularly suitable for development into the next generation optical tape media. Its performance for optical recording has already been demonstrated in some of the highest performance optical disk systems (the 10.2 GB per disk 14" Kodak Optical Disk System 6800 and the 20 GB per side Ar laser based GE Juke Box system at NASA Huntsville). Some of the key performance features are:

- Good writing sensitivity. At 0.3 nJ/mm^2, it is among the most sensitive optical media. Since high-capacity storage systems need high data rate, and the rate of optical recording is in most cases limited by available laser power, high recording sensitivity is desirable.

- Good signal-to-noise ratio. Since the rate of read back in well-designed systems is limited by signal-to-noise ratio, this medium will allow high rate readback.

- Wide wavelength response. Being a metallic alloy, the SbInSn alloy has optical properties that are only weakly dependent on wavelength. The same alloy used for 830 nm recording in the Kodak 6800 system is used for the GE Juke Box using 488 nm Ar laser recording. The wavelength dependence will increase somewhat if a scratch-rsistant overcoat is used. It is still possible, however, that the same formulation designed for the IR lasers today will be compatible with the shorter wavelength lasers of the future to further increase the recording capacity.

- Good resolution. The phase-change recording process is based on transformation of atomic structure and not on macroscopic materials flow. The recording principle is simple and is based purely on the thermal crystallization effect. Any part of the thin-film that is heated above the crystallization temperature is crystallized, and the other parts stay amorphous. The sharpness of the transition is determined by the temperature profile produced by the heating laser beam as well as the dependence of crystallization rate on temperature. The SbInSn material has low thermal conductivity, which results in sharp temperature profiles. It also has an unusually large temperature dependence of its crystallization rate. These two effects lead to

extremely sharp transitions. Experimentally we have demonstrated that the transition width is only on the order of nanometers. The mark size and hence the recording density depends only on how small a spot we can heat with the write beam. It will not be limited by media properties.

In addition, this storage medium also appears to be exceptionally stable. This paper presents some of the studies on the fundamental stability of the medium to demonstrate its suitability as an archival medium.

THE SbInSn MEDIUM

The SbInSn medium is an alloy thin-film prepared by DC-magnetron sputtering using an alloy target of the same composition. As deposited, the film shows no ordering under x-ray or E-beam diffraction and is considered amorphous. Since the amorphous structure is not thermodynamically stable, the film will convert eventually into a crystalline phase. When this happens, the optical properties will change. The indices of the amorphous phase at 830 nm are about 5.07 + 2.62i and those of the crystalline phase 3.64 + 5.52i. The reflectivity of the alloy will, therefore, increase as a result of crystallization (Fig. 1). The crystallization process can be accelerated by heating. The optical recording process uses a focused laser beam to heat selected areas on the thin-film and record information in the form of higher reflectivity crystalline marks. Figure 2 shows an optical micrograph of some laser recorded marks. Figure 3 shows laser- recorded marks under TEM where the featureless amorphous background and the crystalline marks are clearly visible.

Since the recording process involves only a solid-state crystallization process, the construction of the recording element is flexible. In the Kodak 6800 system, the 14" disk has an aluminum substrate with both surfaces coated with a surface-smoothing polymer layer. The phase-change layer is coated on top of the surface-smoothing layer. A thin polycarbonate membrane is then suspended about 0.35 mm over the phase-change layer to provide dust defocusing. The disk thus has an air-sandwich structure with first surface recording. The 14" disks supplied to NASA, on the other hand, utilize glass substrates. The phase-change layers are coated directly onto the substrates. Two pieces of coated substrate are then laminated together, with phase-change layers facing each other, using adhesives. These disks are thus in a solid-sandwich structure for through-substrate recording. To construct an optical tape, it is envisioned that the phase-change layer will be coated on a suitable thin-support and that a scratch-resistant layer will be applied over the surface of the phase-change layer. The scratch-resistant layer is used mostly to prevent defects being generated because of the tape-to-tape contact during the wind-unwind operations. In a well-designed optical tape recorder, there should not be the kind of head-media contacts as experienced in magnetic recorders.

STABILITY STUDIES

To evaluate the stability of the medium our approach is to first make a list of conceivable degradation mechanisms. We then designed and performed experiments to try to quantify their rates. For a phase-change medium, the following are the most probable mechanisms:

- spontaneous crystallization of the amorphous phase
- extended mark growth
- growth of the subthreshold heated region
- restructuring of the amorphous phase
- restructuring of the crystalline phase
- corrosion in high temperature and humidity
- oxidation

We will examine these mechanisms in the following sections.

SPONTANEOUS CRYSTALLIZATION OF THE AMORPHOUS PHASE

The principle of phase-change recording is based on the strong temperature dependence of the crystallization rate. For use as an optical recording medium, a phase-change material needs to have a low rate of crystallization at around room temperature so that the medium can be kept for years without spontaneous crystallization. It is also necessary that the rate will increase significantly at elevated temperatures so that during the recording process, the material can be heated by the laser beam to complete the crystallization process in a few nanoseconds. The temperature dependence of the kinetics of the crystallization process can be studied by studying the thermal crystallization process. A sample is placed on a hot plate and heated at a constant temperature ramp rate while its optical reflectance is continuously monitored. At a certain temperature the sample will crystallize and an abrupt increase in reflectance will be observed. This temperature is called the crystallization temperature and is the temperature at which the rate of crystallization matches the rate of heating. When the temperature ramp rate is increased, the crystallization temperature will also increase. By studying the dependence of crystallization temperature on heating rate, the dependence of crystallization rate on temperature can be deduced.

Figure 4 shows the results of such a study. Here $\log(\beta/T^2)$, where β is the heating rate, is plotted against $1/T$. A straight line as indicated in the figure suggests a thermally activated behavior where the slope of the line is proportional to the activation energy. The actual data can only be collected over a very limited temperature range (at about 170°C) where the experiments lasted from a few minutes to a few days. It is seen that the temperature dependence is exceedingly strong. A three-orders-of-magnitude increase in the heating rate only results in an increase of crystalline temperature by a mere 14°C. The calculated activation energy is almost 200 kcal/mole, or over 8 eV. The origin of such an exceptionally high activation energy is not understood at the present time. From these data, it is possible to estimate the amount of time needed to complete 50% crystallization as a function of temperature (Fig. 5). If the observed temperature dependence can be extrapolated to lower temperatures, then the time for 50% crystallization at 50°C is predicted to be more than 10^{30} years, which is longer than the predicted lifetime of the universe. Obviously something else will happen before this, but the data do indicate that the stability of the amorphous phase by itself should not be of a practical concern.

If the temperature dependence can be extrapolated to higher temperatures, then it is predicted that the material needs only to be heated to about 230°C to have the crystallization completed in 10^{-10} seconds, a rate required for high-speed recording. This is to be compared with ablative type optical media that have to be heated to much higher temperatures in order to achieve melting or vaporization of the materials. The small heating requirement is the main reason for the superior writing sensitivity of the SbInSn medium.

EXTENDED CRYSTALLINE MARK GROWTH

Crystallization process in SbInSn proceeds via the normal nucleation and growth process. It is possible, therefore, that even though the amorphous phase is stable by itself, growth of the recorded crystalline marks will destroy the data before the spontaneous crystallization of the amorphous phase. Figure 6 shows the TEM micrographs of a laser-recorded mark before and after heating for an extended period of time at high temperatures. Indeed irregular growth of the crystalline boundary is observed even though the amorphous background remains unchanged. We call this the extended mark growth to distinguish it from the uniform mark growth to be discussed later.

To study the kinetics of the extended mark growth, a series of samples with prerecorded marks were heated at various temperatures for extended periods. At selected time intervals, microscopic examinations of the prerecorded marks were performed to determine the extent of mark growth. The results indicated that the growth rate is essentially constant with time and it increases exponentially with temperature. Figure 7 shows a plot of the time needed for a 3

mm increase of mark diameter as a function of temperature. It also shows the estimated time needed for 100 nm growth. We labeled the latter the end of life because an irregular growth of such an extent would increase jitter significantly and an appreciable degradation in phase-margin in a pulse-length-modulated recording system would have resulted at this stage. For pulse-position-modulation system or fixed-bit-cell recording system this estimate could be overly conservative.

The data in Fig. 7 indicated that the predicted lifetime based on this conservative criterion is over a thousand years at 50°C. Extended mark growth will not, therefore, limit the lifetime of this medium.

GROWTH OF THE SUB-THRESHOLD HEATED REGION

In addition to the extended mark growth that is observed when the samples are heated to higher temperatures, we also observed another kind of mark growth at relatively lower temperatures. In this case there seems to be a uniform growth front. The marks will become larger with time but no increase in jitter is observed as a result of this growth. The growth has a logarithmic dependence on time. It is fast initially, but slows down significantly afterwards and hence is self-limiting. It also has a temperature dependence. Figure 8 shows the observed rate of change at various temperatures. For example, at 80°C the mark will grow in diameter by a total of 13 nm after a day, but it is only predicted to grow by 26 nm after 300 years.

Evidence suggests that this growth is caused by a subthreshold heating effect. The laser spot used for optical recording has a Gaussian-like intensity profile that generates a Gaussian-like temperature profile in the material. Within a radius near the center, $r < r_c$, the material is heated to a high enough temperature that it becomes crystallized to form a written spot. Just outside the transformed region the material has not crystallized, but it has received substantial heating from the writing laser beam that apparently has caused changes in its properties. It is the growth into this region that is observed when the marks are aged.

Since the growth is predictable and limited, it is not a lifetime-limiting effect. Rather, proper budget in the phase-marging has to be allowed to account for this effect if PWM recording scheme is used. For PPM or fixed-bit-cell recording the impact would be even less.

RESTRUCTURING OF THE AMORPHOUS PHASE

The as-deposited films are in a vapor quenched high free-energy state. The free energy of the films can be reduced by local rearrangement of the atoms without actually crystallizing the films. This effect has been documented, for example, for magneto-optic thin-films.[2] In the MO films, the restructuring causes a change in the magnetic properties. In the SbInSn thin-films, the restructuring reduces the writing sensitivity. The observed reduction in sensitivity again has a logarithmic time dependence and a slight temperature dependence. Figure 9 shows the observed rate of recording power increase at various temperatures. At 60°C an about 10% drop in writing sensitivity is expected over the first year of media life. Since most recorders utilize write calibration strategies, this effect will not affect the lifetime of the media.

RESTRUCTURING OF THE CRYSTALLINE PHASE

In high-rate optical recording processes the total laser exposure time at a given spot on the media is on the order of nanoseconds. Such a small amount of time is usually insufficient for the materials to reach thermal equilibrium. As a result, metastable crystalline phases are obtained. In the case of SbInSn alloy, the laser recording process results in a NaCl-like cubic crystalline structure. The equilibrium state, however, is expected to be a mixture of several crystalline phases. Upon aging a gradual transformation of the laser-written marks into more stable crystalline phases has been observed. Associated with the phase transformation is a slight decrease in the reflectance of the crystalline marks and hence the recording contrast or

the readback signal (the carrier). Figure 10 shows that the change of reflectance depends logarithmically on time and is faster at higher temperatures. Figure 11 shows the resulting degradation in carrier signal at various temperatures. At 60°C, for example, we observed a 1.2 dB degradation after one year. From the data in the figure, we predict a total of 1.5 dB degradation after three hundred years.

CORROSION AT HIGH TEMPERATURE AND HUMIDITY

The SbInSn thin films are exceptional in their corrosion resistance in high temperature and high humidity environments. Incubation of bare films without any protection layer at 70°C, 70% RH for over two years has resulted in no observable corrosion. The rate of corrosion cannot, therefore, be easily quantified. In practical products where an overcoat protection is likely to be used the corrosion rate should be further slowed. It is judged, therefore, that corrosion degradation of these films should not be a concern.

Figure 12 is a TEM micrograph of an SbInSn thin-film that had been incubated for over 4000 hours at 70°C, 70% RH. The film was coated on an injection-molded polycarbonate substrate and recorded with the laser-written marks shown in the picture. The incubation was carried out with the SbInSn film directly exposed to the environment. After the incubation, the polycarbonate substrate was dissolved in a solvent so that the SbInSn thin-film could be mounted on a copper grid for the TEM observation. This abusive sample preparation procedure caused the cracks and stains in the picture. There is, however, no sign of corrosion, nor the crystallization of the amorphous region, nor the extended growth of the prerecorded marks. These results are consistent with the predictions as noted in previous sections.

OXIDATION

The corrosion resistance of the film is believed to be a result of self-passivation: the oxidation process resulted in the growth of a native oxide film that protects the films from further oxidation. This mechanism is commonly observed among most corrosion-resistant metals or alloys. For thin-films, however, a concern is whether the passivating oxide layer will reach a thickness that will affect the performance of the films.

Since the passivating layer is most likely a transparent oxide, one easy way to monitor the thickness of the oxide layer is to monitor the optical density of the thin-film, which gives a good representation of the unoxidized alloy film. Figure 13 shows a plot of the optical density of a 48-nm-thick thin-film heated at different temperatures in air as a function of time. The optical density again has a logarithmic dependence on time, and the temperature dependence appears to be small. The projected oxide thickness after 300 years is less than 5 nm even at high temperatures. In a disk or tape construction, although the maximum change in reflectance between the amorphous and the crystalline phases is observed at about 80 nm film thickness, the actual recording performance is rather constant within a 15% thickness range. Thus a 5 nm decrease in phase-change layer thickness due to oxidation is not expected to have much impact on the recording performance. The long-term incubation studies have confirmed this prediction.

14" DISK DATA

Although all the fundamental studies mentioned above suggest that the SbInSn alloy is a very stable recording medium, real-life confirmation is still necessary. We have subjected many 14" disks to various testing environments, and so far have not had any indications that the projections based on these studies are incorrect. Figure 14 shows the results of extended incubation of several disks at 70°C, 90% RH. It shows virtually no change in uncorrected error rate over the 84-day incubation period.

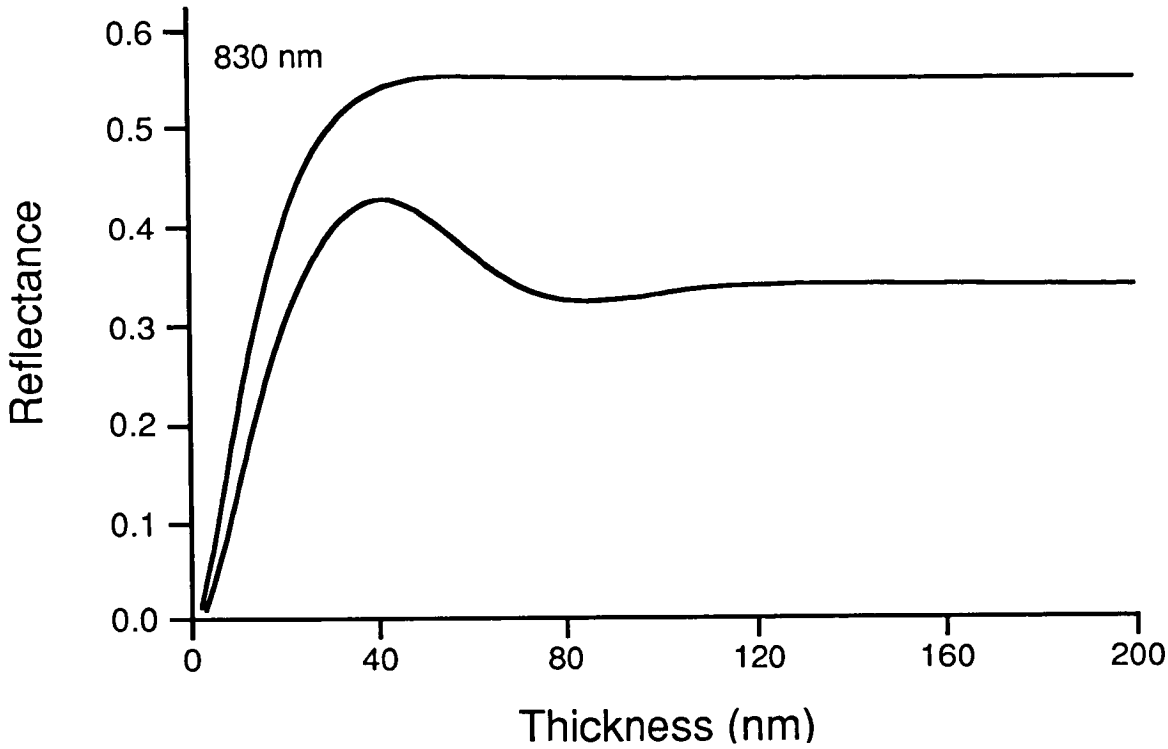
SUMMARY

Extensive degradation mechanism studies suggest that the SbInSn phase-change alloy thin-film is very stable and it will not be the lifetime-limiting factor in any disk or tape medium construction. This superior stability as well as the many virtues listed in the Introduction section makes this alloy an ideal candidate for optical tape applications. Extensive efforts are underway at Eastman Kodak Company to promote optical tape technology based on this material.

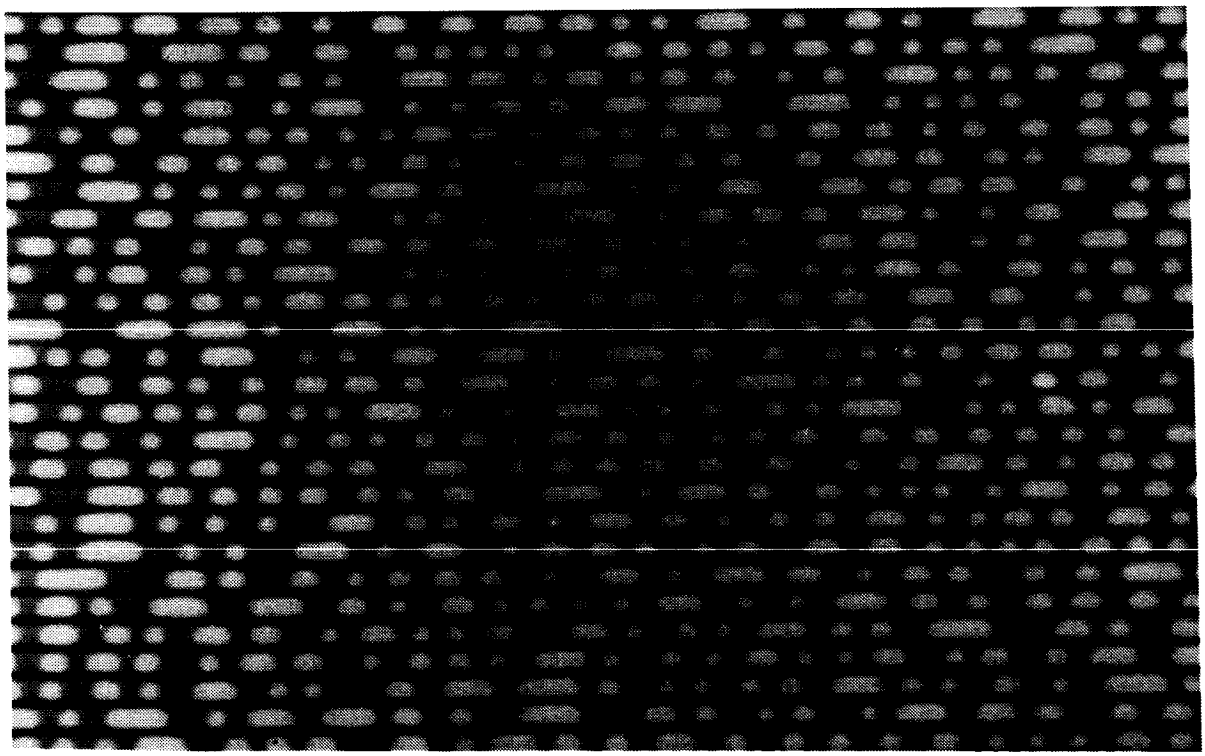
REFERENCES

1. D. Gelbart, Optical Data Storage, SPIE Vol. 1316, p.65 (1990).

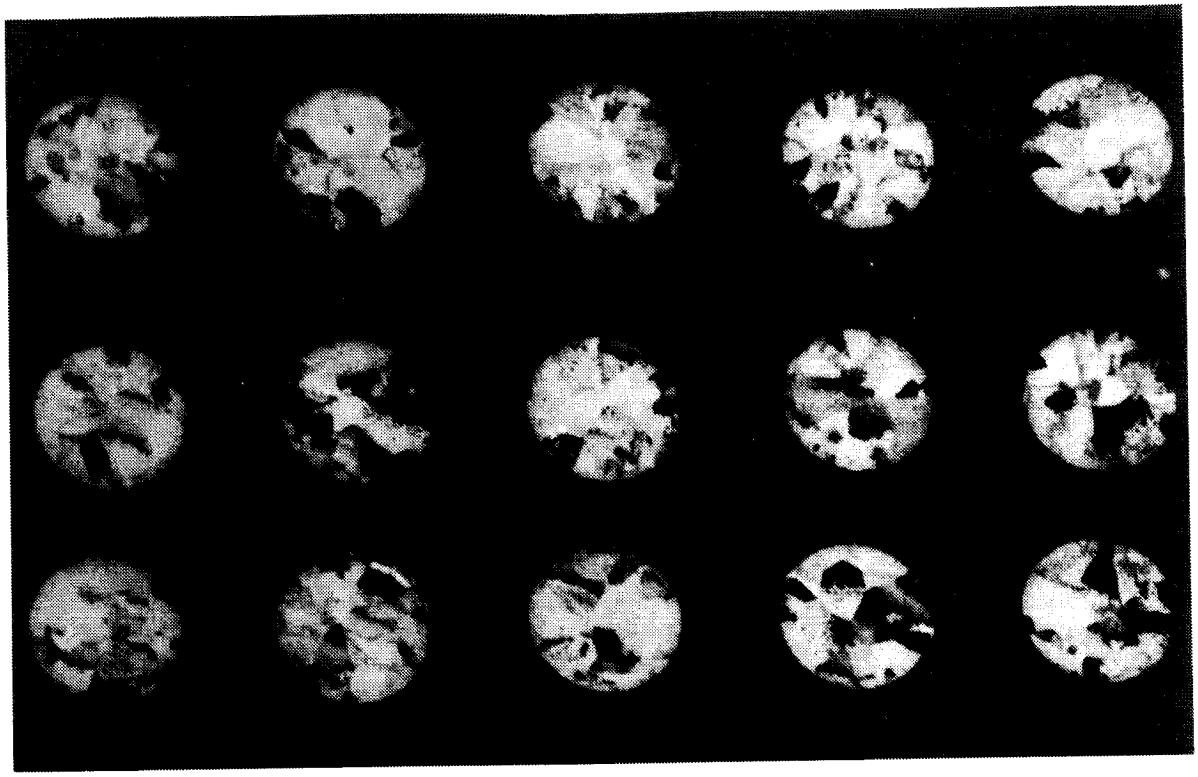
2. N. Ogihara, K. Shimazaki, Y. Yamada, M. Yoshihiro, A. Gotoh, H. Fujiwara, F. Kirino, and N. Ohta, Jpn. J. Appl. Phys., V. 28 (1989), Supplement 28-3, pp 61-66.



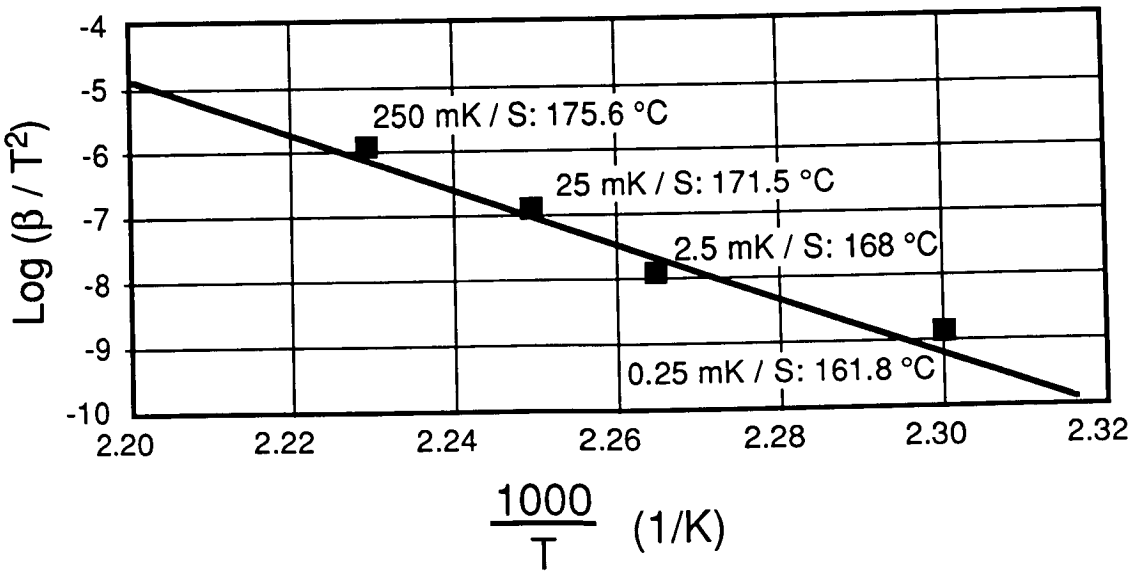
1. **Calculated reflectivity of an SbInSn thin-film for both amorphous and crystalline states as a function of film thickness.**



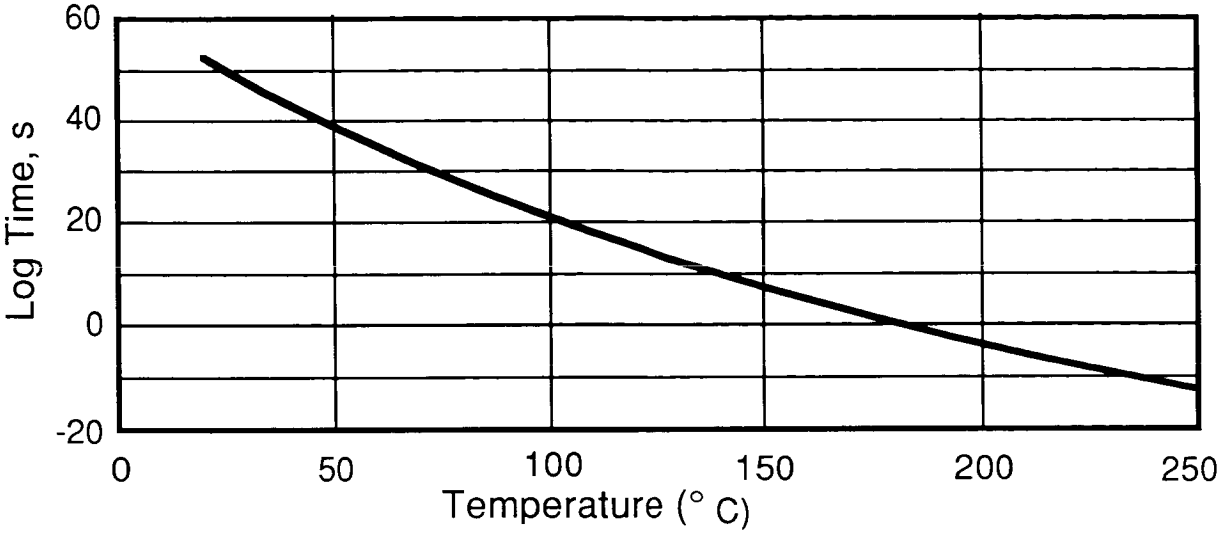
2. **Photomicrograph of some laser-recorded marks in an SbInSn thin-film. (Bright field reflection; the white features are recorded marks. The groove width is 1.6 mm)**



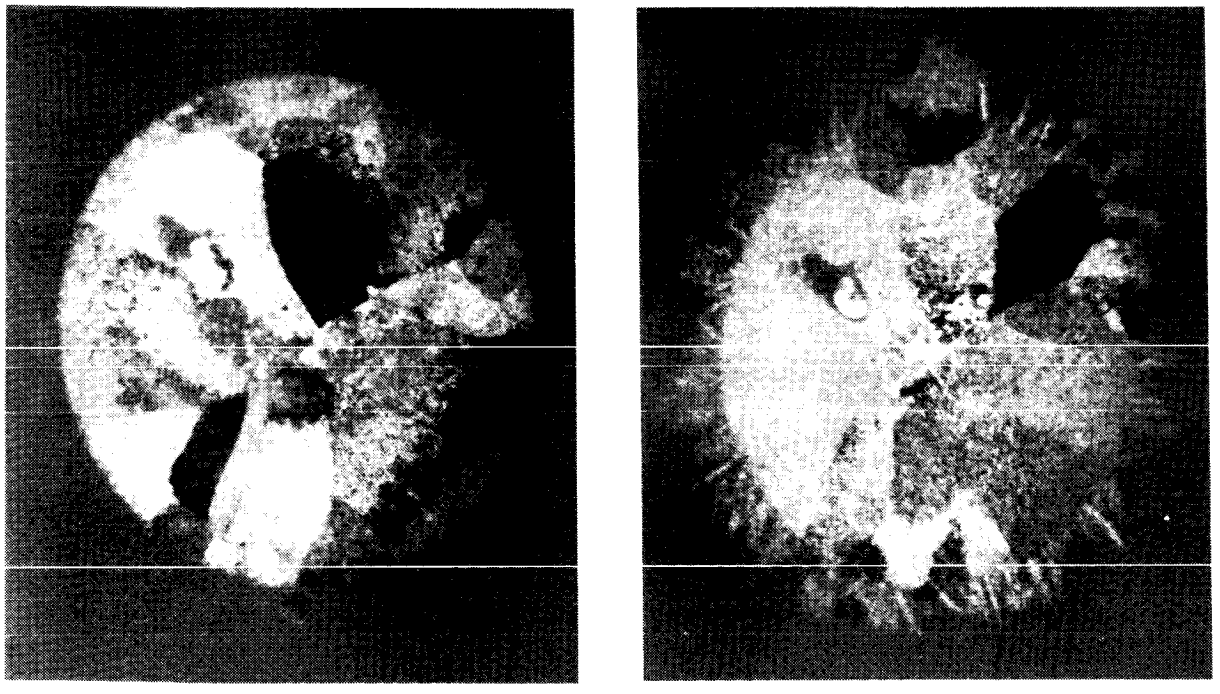
3. **TEM micrograph of some laser-recorded marks. Structure within the marks is due to diffraction of electron beam by crystallites of different orientations.**



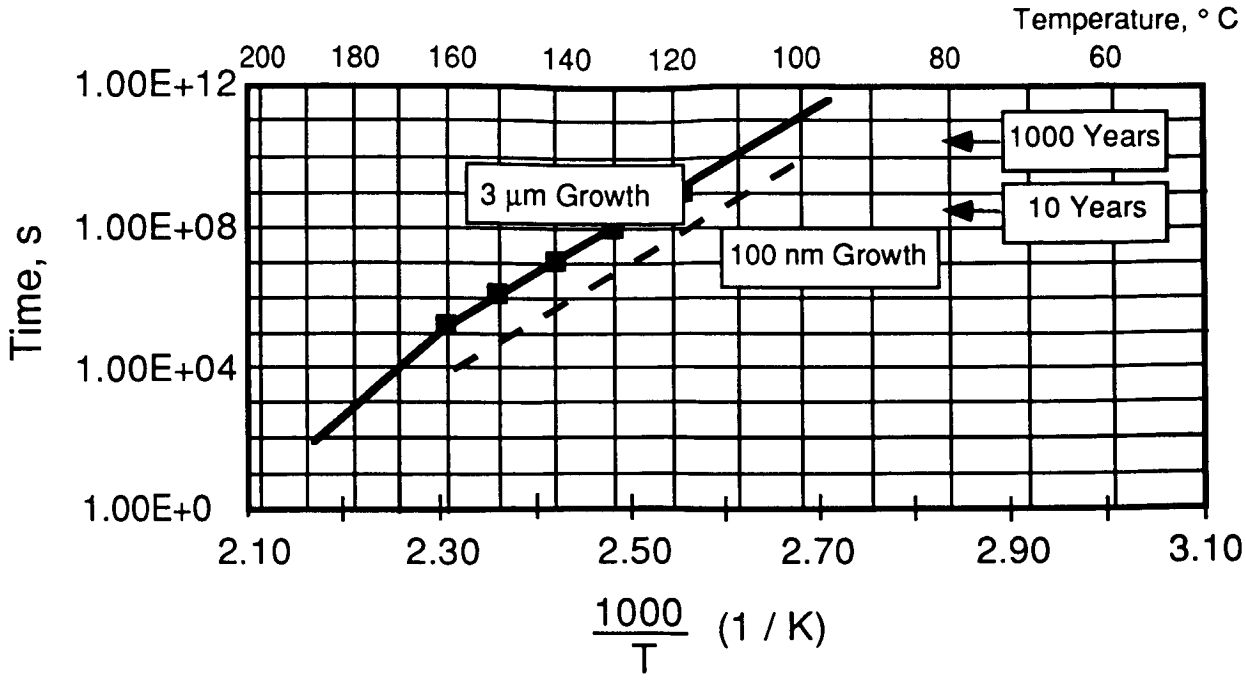
4. **The dependence of measured crystallization temperature on heating rate. The linear dependence indicates a thermally activated behavior.**



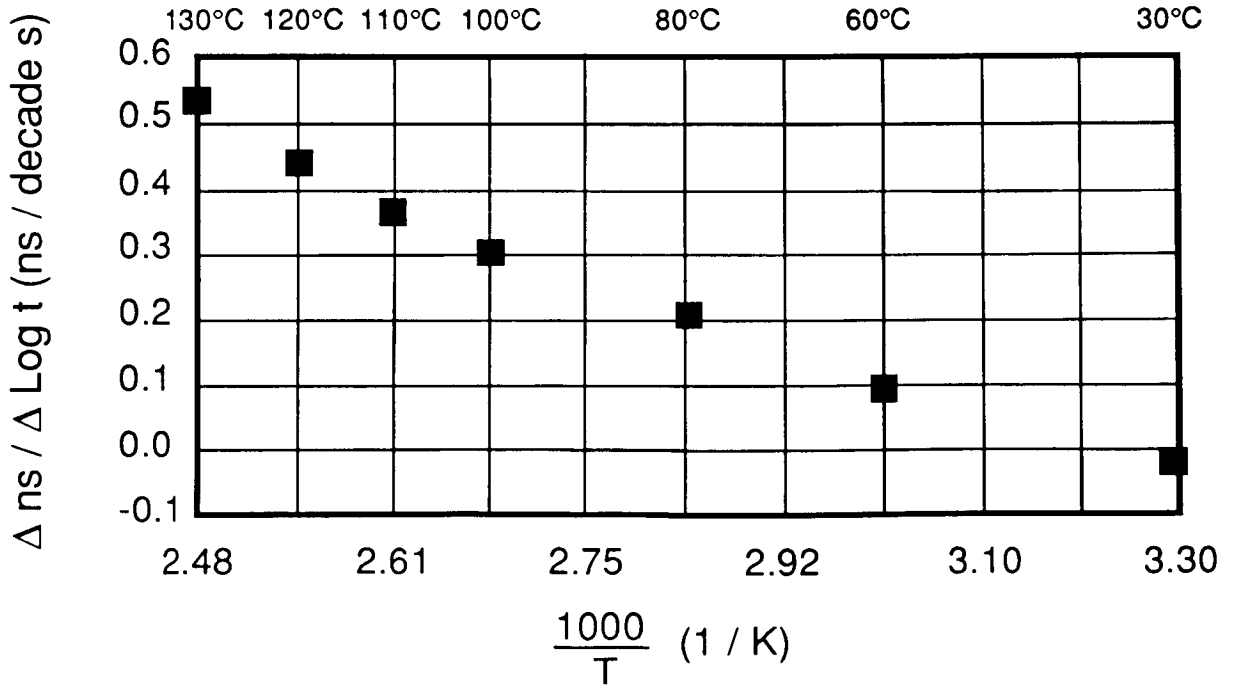
5. **Estimated time required for crystallization to proceed 50% as a function of temperature.**



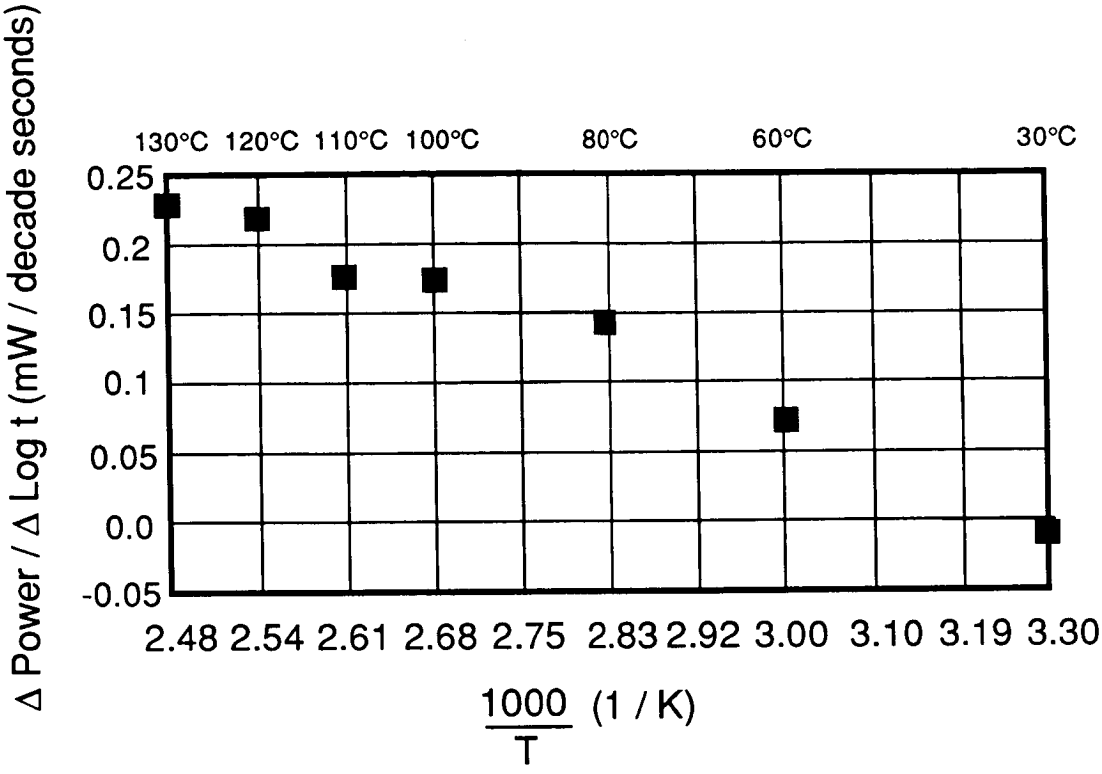
6. **TEM micrograph of (a) a laser-recorded mark in its unheated state, and (b) a mark that has been heated to 140°C for an extended period of time and that experienced extended growth of the crystalline phase beyond the original mark boundary.**



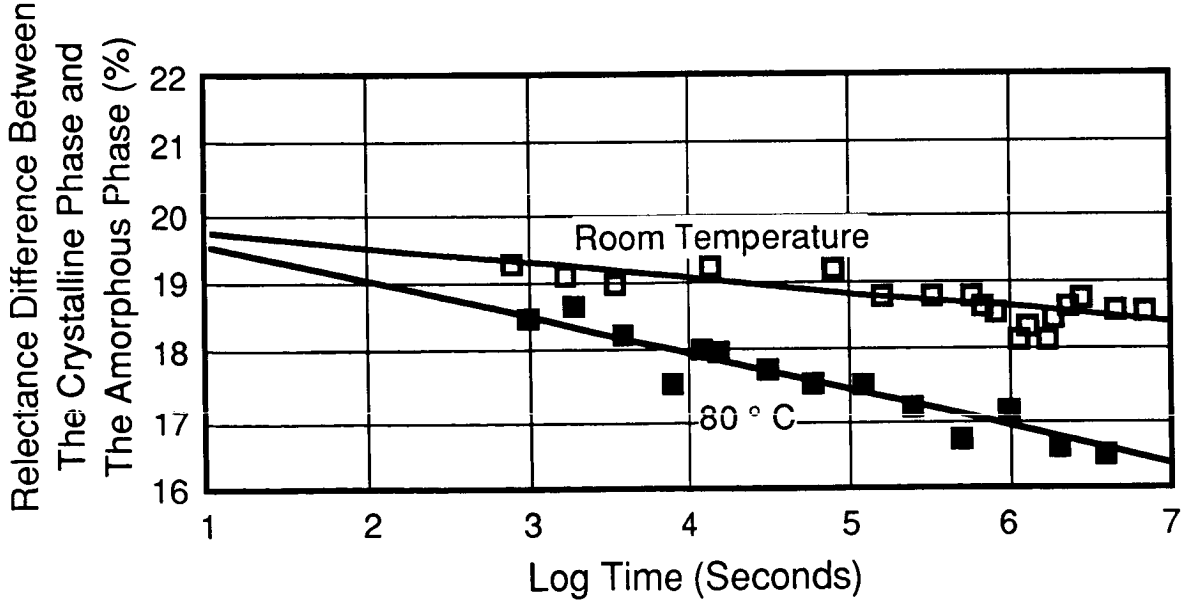
7. **Time needed for extended mark growth as a function of temperature. All points for 3 mm growth are actual data except the one at 120°C, which is extrapolated from the amount of growth observed so far. The line for 10 nm growth is estimated from the 3 mm data based on constant growth rate with time. The break in slope at 160°C is believed to be real from the many repeated experiments.**



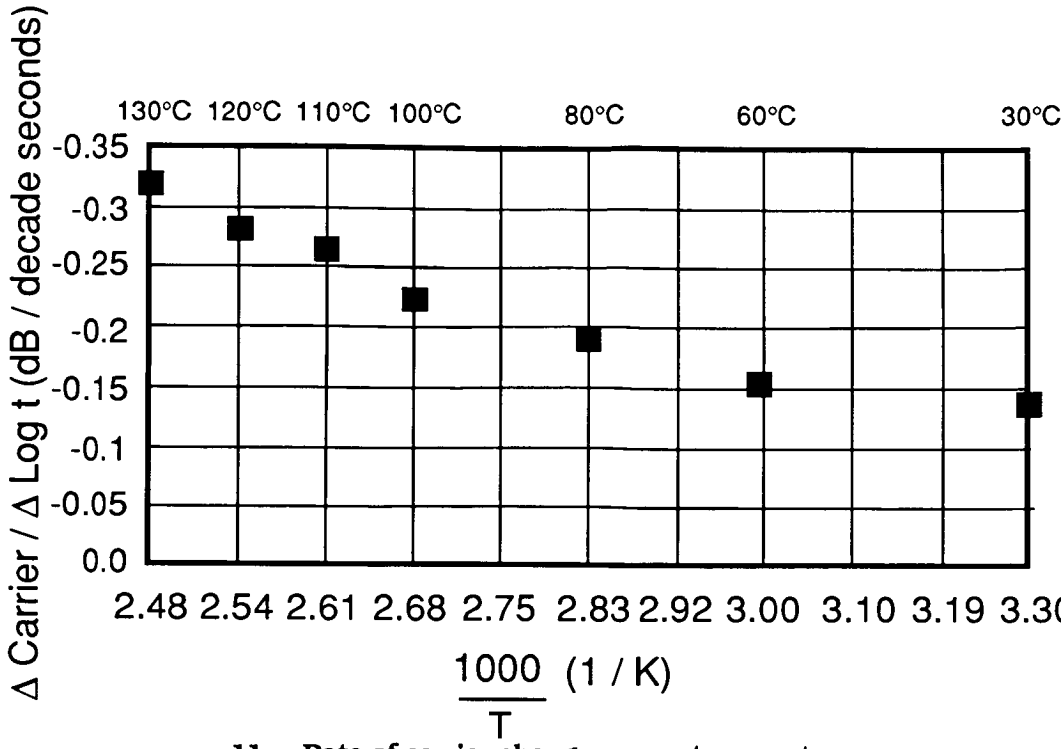
8. **Predicted growth rate due to subthreshold heating during the writing process. The rate is expressed in ns/decade, which can be converted to nm/decade of mark diameter growth by multiplying the values by the linear speed of the media, 13 m/s. The growth depends logarithmically on time which means equal amount of growth is expected between any 10^n to 10^{n+1} s.**



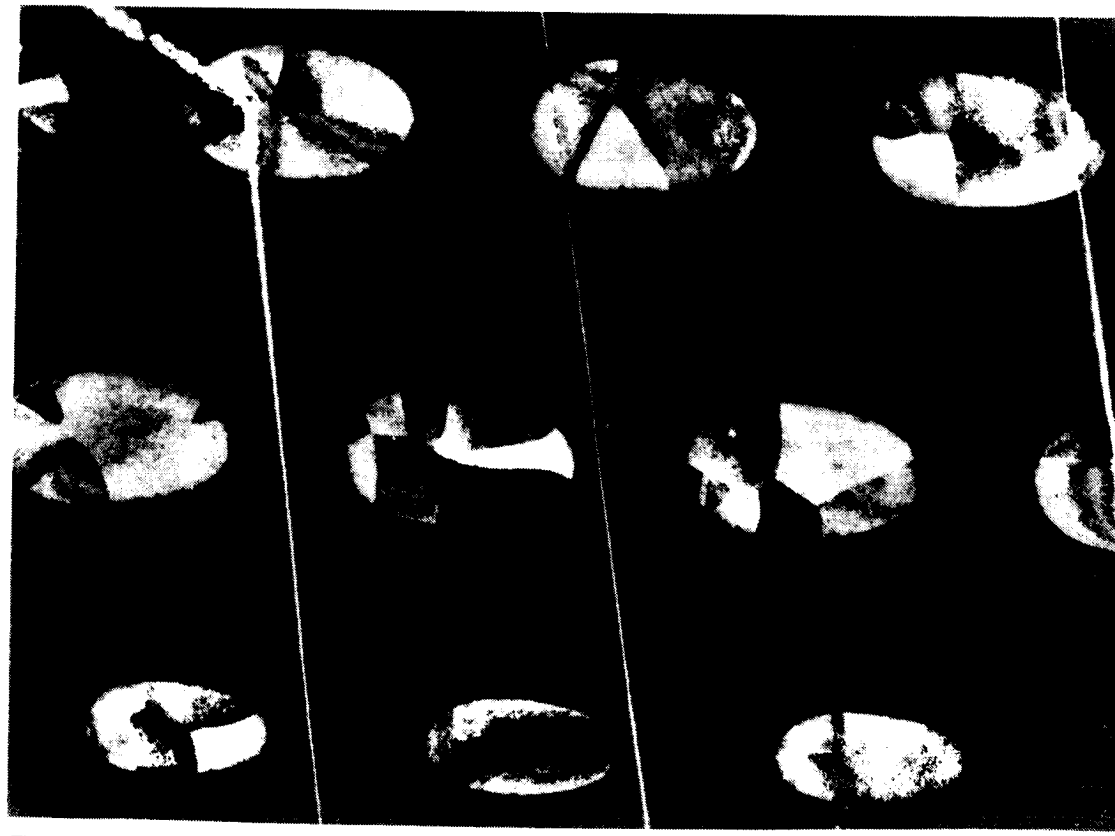
9. Change of optimum recording power as a function of time at various temperatures. The optimum recording power is defined as the power to achieve 50% duty cycle readback using 50% duty cycle input signal.



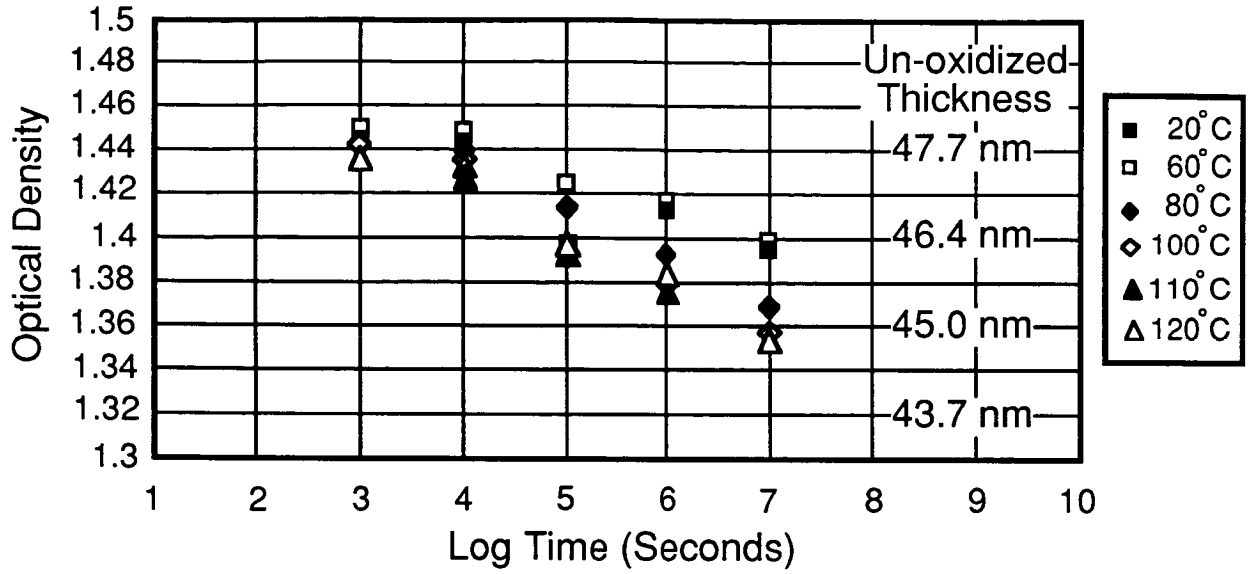
10. Change of the reflectivity of the cubic crystalline phase with time at room temperature and 80°C.



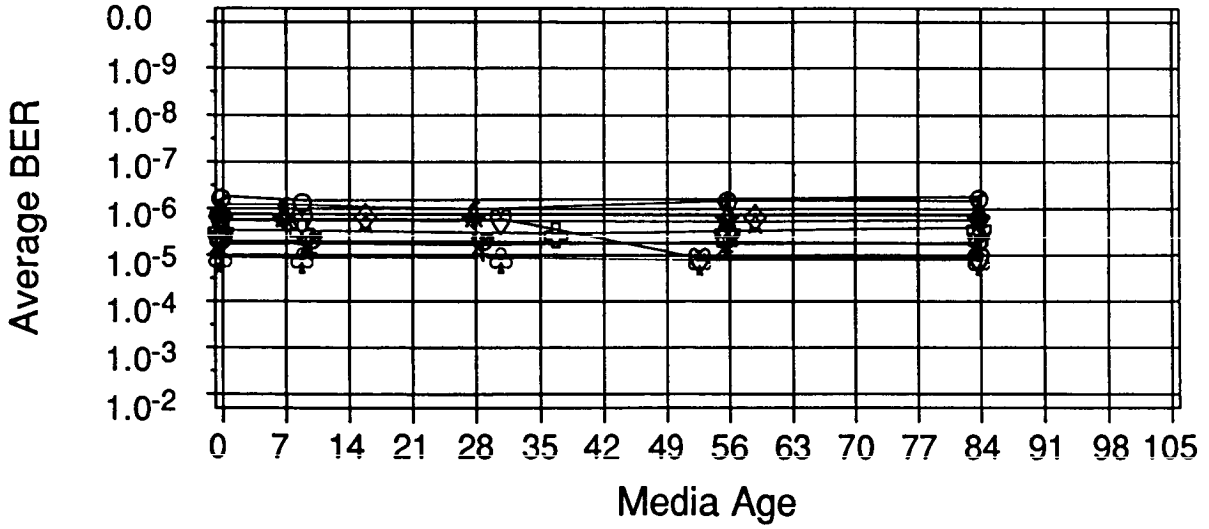
11. Rate of carrier change versus temperature.



12. TEM micrograph of laser-recorded marks after 4000 hours incubation at 70°C, 70% RH. The cracks and dark stains are results of preparing samples for the TEM studies.



13. Change of optical density for a 48-nm-thick SbInSn thin-film. The decrease in optical density is interpreted as a result of oxide formation. The remaining optical density is taken as proportional to the unoxidized film thickness.



14. Change of bit-error rate (BER) as a function of incubation time at 70 °C 90% RH for several Kodak 14" disks .

Electron Trapping Optical Data Storage System and Applications

Daniel Brower, Allen Earman and M.H. Chaffin

Optex Corporation
2 Research Ct. Rockville, MD 20850

Abstract

A new technology developed at Optex Corporation out-performs all other existing data storage technologies. The Electron Trapping Optical Memory (ETOM™) media stores 14 gigabytes of uncompressed data on a single, double-sided 130mm disk with a data transfer rate of up to 120 megabits per second. The disk is removable, compact, lightweight, environmentally stable, and robust. Since the Write/Read/Erase (W/R/E) processes are carried out photonically; no heating of the recording media is required. Therefore, the storage media suffers no deleterious effects from repeated Write/Read/Erase cycling.

This rewritable data storage technology has been developed for use as a basis for numerous data storage products. Industries that can benefit from the ETOM data storage technologies include: satellite data and information systems, broadcasting, video distribution, image processing and enhancement, and telecommunications. Products developed for these industries are well suited for the demanding store-and-forward buffer systems, data storage, and digital video systems needed for these applications.

Electron Trapping Overview

The advent of digital information storage and retrieval has led to explosive growth in transmission, compression, and high capacity random access storage of data. A key limitation for growth of digitally based systems is the slow advance of erasable data storage technologies. New storage technologies are required that can provide higher data capacity and faster transfer rates in a more compact format. Magnetic disk/tape and current optical data storage technologies fail to provide all of the higher performance requirements of digital data applications.

The Electron Trapping Optical Memory (ETOM) Media developed at Optex Corporation out-performs all current data storage technologies. ETOM is a novel erasable data storage media which utilizes the phenomenon of electron trapping[1]. Electron trapping is common in a class of luminescent materials known as IR stimulable phosphors. They are composed of a host lattice, typically an alkaline-earth sulfide, and two rare earth dopants (the luminescent and trapping centers). Data storage is a fully photonic process which involves the interaction of light with the dopant ions and their electrons within the media.

The Electron Trapping Optical Data Storage (ETODS) System uses two wavelengths of light to accomplish the Write/Read/Erase processes. The fundamental process responsible for the storage of data is the transfer of electrons between the two types of

dopant ions. The write/read/erase processes are fully reversible and occur at the atomic level within the crystal lattice structure.

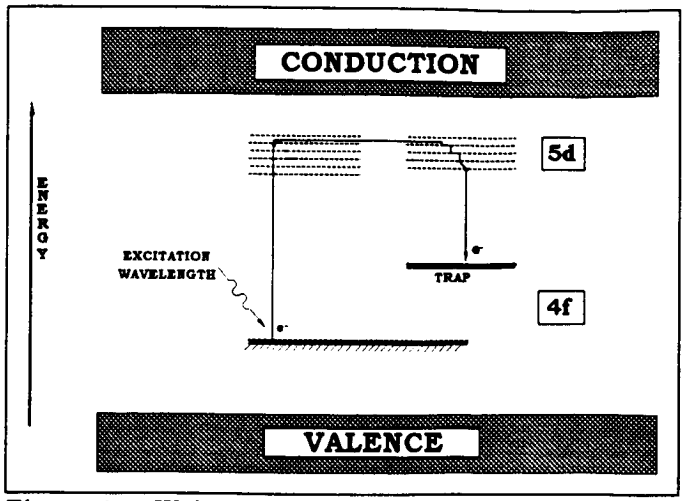


Figure 1a. Write operation

The **write** process involves raising an electron from the ground state of the luminescent ion to its excited state. This electron then migrates to a nearby trapping ion and falls to the ground state of that ion. Figure 1a illustrates the write process. It is important to note that the dopant ground and first excited states are within the bandgap of the host lattice. The ground states of both ions have a 4**f** configuration which is highly localized; therefore they are very stable. So, a trapped electron, barring excitation with a stimulation source, will remain trapped indefinitely. The excited states of both ions have a 5**d** configuration and are much more extended. These extended orbitals overlap orbitals of nearby atoms in the lattice allowing for electron transfer through the lattice. The **read** process is the reverse of the write process; the only difference is the wavelength needed to detrap the stored electrons. The trapped electron will remain trapped until a photon of the read light source excites it from the ground state to the excited state. From here it can migrate back to a luminescent ion and relaxes to the ground state. The transfer back to the ground state is accompanied by the emission of a photon which is detected by the electro-optical drive system and indicates stored data. Figure 1b illustrates this process. All traps need not be emptied in this process, thus allowing for multiple read passes prior to a refresh step. **Erasure** is carried out by simply increasing the Read laser power to completely detrap all stored electrons.

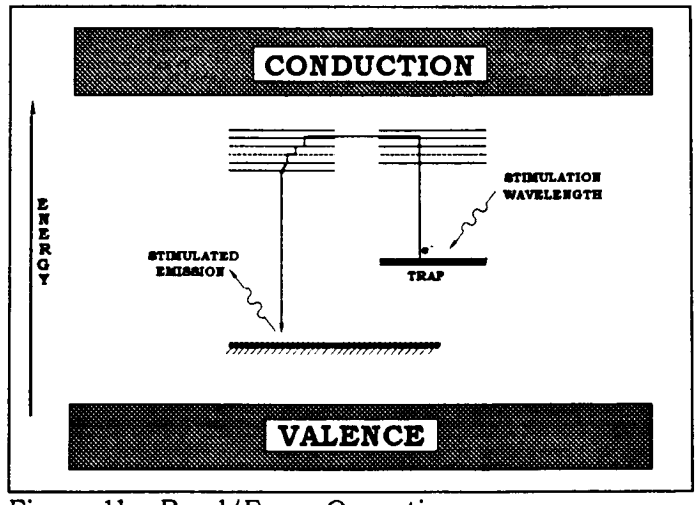


Figure 1b. Read/Erase Operation

Performance Features

The ETOM technology has many advantages over current erasable data storage technologies. Since the write/read/erase processes are carried out 100% photonically, no heating of the recording media is required. This provides a distinct improvement in media sensitivity and transition rates over current technologies. As a

result the media requires less energy to write a mark. A decrease in dwell time and/or laser power translates to higher data rates and decreased laser cost. The time to create a mark and read it back is also decreased since the write and read processes are photonic and not thermal.

ETOM media is also durable. It must be shielded from external light sources, so it is stored in a compact, light-tight cartridge. This forms a rugged package that easily can be removed from the drive. The media itself is environmentally stable as long as the proper stoichiometry is maintained during processing. Normal thermal flucations do not have an effect on the performance or stability of the media. In fact, the substrate and protective coatings would be destroyed prior to the loss of data. Thermal detrapping will occur only above 370 °C[2].

Analog performance

In a typical optical recording head, the optical stylus beam is formed using an objective lens with a Numerical Aperture (N.A.) of about 0.5. This lens forms a spot on the disk surface that is approximately 1.0 μm in diameter using conventional laser wavelengths. Within the media volume illuminated by the focused optical stylus beam, many dopant pairs will participate in the data storage process. The typical dopant pair density in an ETOM film is 10^6 pairs/μm^3 for a dopant concentration of 300 ppm. Therefore, there are approximately 10^6 potential "traps" per cubic micron—the volume illuminated by the optical stylus beam. Essentially this corresponds to an effective domain of approximately 10^{-6} μm^3. The number of traps filled during the write process depends on the field strength of the illumination and the absorption of the material. In this manner, the number of traps filled may be controlled by varying the intensity of the illuminating light.

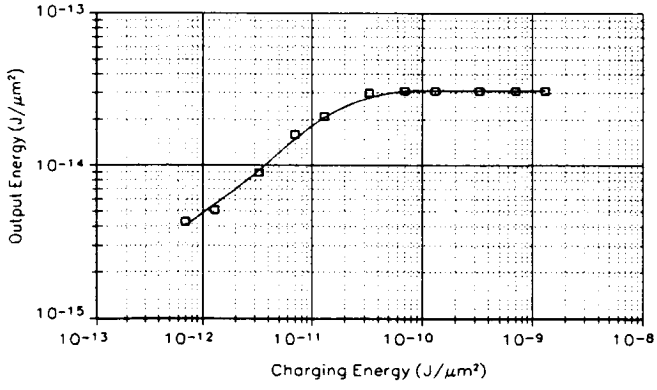


Figure 2. ETOM write saturation.

The linear response of ETOM material is illustrated in Figure 2. The graph shows emission energy ($J/\mu m^2$) as a function of excitation (charging) energy ($J/\mu m^2$) for a constant stimulation energy[1]. The stimulation occurs at some time period after excitation is terminated. ETOM material illuminated by optical radiation at the excitation wavelength responds linearly over a broad region until a saturation level is reached. Saturation in this case is defined as maximum filling of available "traps".

For conventional two-state binary digital recording, the write energy is alternated between the region of no response and saturated response. It is the linear response

---

[1] The shape of the curve resembles the Hurter-Driffield, or DLog(E), curve for photographic film. In the case of the H-D curve, post-development film optical density is plotted as a function of exposure. Photographic film typically is characterized by a long linear portion of the curve where increased exposure results in increased density. At a point, called saturation, maximum density is achieved and further increases in exposure produce no further increase in film density.

region, however, that provides ETOM with its unique ability for analog signal recording. Although full analog recording is possible—we routinely demonstrate analog video recording on an ETOM disk—discrete multi-level recording provides many salient advantages. Discrete information coding provides greater noise immunity for a given band-limited data channel, and provides the opportunity to scale the information transfer rate while maintaining the integrity of source information.

Multi-level digital recording

Multi-level or non-binary digital recording[3,4] takes advantage of the exceptionally wide dynamic range of ETOM media. This method of coding source information can provide up to four times the data capacity and transfer rate using four discrete amplitude levels. Multi-level recording is implemented by controlling the write laser illumination energy so that the number of filled traps in a data feature can be none, partial, or complete. Figure 3 shows the result of recording a monotonically increasing and decreasing "gray" scale on an ETOM disk. The highest amplitude pulse represents the saturated recording level. Here thirteen discrete levels are shown. Several of the levels are not distinguishable due to non-uniformities in the coating of this particular sample. However, note that the pulse width remains uniform for all pulse amplitudes, illustrating the independence of pulse width and amplitude.

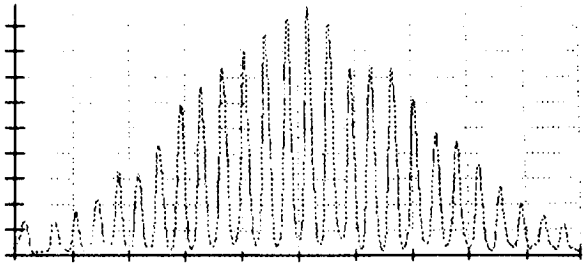


Figure 3. Recorded gray-scale.

Pulse Amplitude Modulation (PAM) combined with Pulse Duration Modulation (PDM), or Pulse Position Modulation[5] (PPM), provides a larger data channel symbol set for coding source information. In typical digital recording channels, information is carried in the phase of the recorded signal—either width or position of the pulses. Combining PAM and PDM—or PPM—creates a matrix arrangement of data channel symbols where information is carried in *both* the phase and the amplitude of the recorded signal. A simple MFM-type code provides a good example. In Modified-Frequency-Mark (MFM) coding there are three possible phase symbols—1.0,1.5, and 2.0 bit cells long[6]. The equivalent 4-level code provides four amplitude symbols and three phase symbols for a total of twelve (12) code symbols. (One of these symbols—the "0"—is redundant so that there are only eleven usable symbols.) This yields more than three times as many usable symbols for the same physical mark dimension. Although this is a very simple explanation of a simple code system, it is evident that there is a significant capacity and data rate enhancement through the use of such coding.

Using a 4-level coding system, it is possible to enhance the capacity—the average number of source bits per recorded feature—by up to four times. Since the average number of bits per feature is increased, the average number of bits per second also is increased by the same factor if the number of features/second—the "mark" rate—is unchanged. This leads to a 4× increase in user data transfer rate.

When data is read out from the ETOM disk, the filled-trap density decreases proportionally with the intensity of the read-laser illumination. Therefore, it possible to fully erase, or to partially erase, the data while reading. In the case of partial erasure, the various amplitude symbols decrease proportionally. After several read steps—depending on the intensity of the read illumination—the number of stored electrons will fall below a critical detection threshold. At this point data is lost. In ETOM technology the data must be refreshed to retain information integrity. This refresh step is simply a read step followed with a restoring write step which can be performed automatically. Since the ETOM-based optical system contains the two laser beams—one for the write wavelength and one for the read wavelength—no additional optics are required for the refresh operation. Figure 4 illustrates the optical stylus arrangement for a data refresh operation.

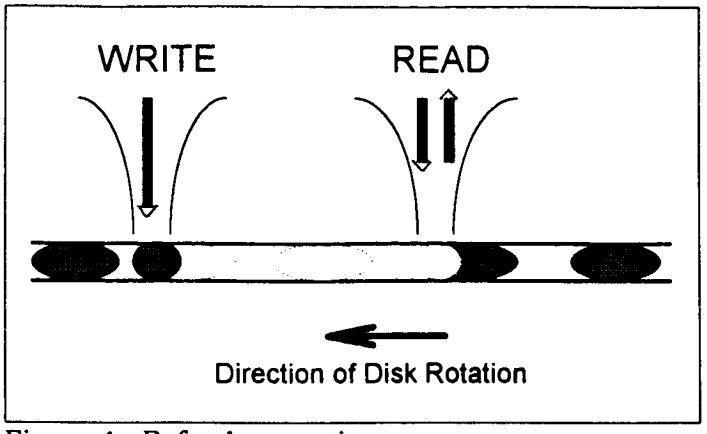


Figure 4. Refresh operation

Using quaternary coding it is possible to store 14 gigabytes of data on a double-sided 130mm substrate with a transfer rate of 120 megabits per second. This capacity and transfer rate coupled with the features of erasability and random access forms a very versatile data storage system.

Applications

The ETOM media is capable of both analog and digital recording. An ETODS system with its massive data capacity, high transfer rate, and its ability to directly access random data fits a number of video and data storage applications. Industries that can benefit from the ETOM data storage technologies include: satellite data and information systems, broadcasting and video distribution, telecommunications, and computer data storage.

Satellite Data and Information Systems

ETOM-based products are ideally suited to the huge data requirements of imaging applications. An ETODS store-and-forward downlink system can be used to acquire and re-transmit data from satellites. There is a continuing pressure to increase the number of on-board information generating devices and to increase the resolution of the images collected. A higher capacity system with a greater transfer rate is needed.

The data collected during a single orbit of a polar orbiting satellite, for example, must be dumped in 10 minutes as the satellite travels from horizon to horizon. Faster transfer to the ground station would make it possible to collect more data per pass. Currently, the satellite dumps approximately 9 gigabytes of data for a single pass at a rate of 2.66 megabytes/sec[7]. A single 130 mm ETOM substrate could store this

amount of data with 5 gigabytes to accommodate future growth. The ETOM-based system could accommodate this data rate and volume with the added benefit of random access to any bit of data.

While the initial application of an ETODS system would most likely be a store-and-forward system at a ground-based downlink site, it might also find use in on-board satellite operations. The system would weigh less than current tape systems and contribute no additional EMI. Since the detrapping process is not temperature sensitive, operation at the temperatures encountered in space flight should be possible.

Broadcasting and Video Distribution

A digital video recorder based on ETOM materials holds particular promise for this industry. An ETODS system could store many hours of video on a single substrate with random access to any segment. This capability is unmatched in the industry and makes possible a variety of video applications that have simply not been practical without ETOM.

Today the broadcasting and video distribution industries are faced with technological advances that require high speed and high capacity storage devices. The switch from analog to digital signals and new products such as HDTV will virtually obsolete existing video tape storage products. An ETOM-based digital video recording system would complement many of these new technologies and will create many new applications in the video/broadcasting marketplace.

Video Buffering
An ETOM-based digital video recording system developed for TV cable system "head ends" would both simplify downloading of satellite signals and insertion of local commercials prior to re-transmission to subscribers. Cable operators could use this system to manipulate the hundreds of channels of digitally compressed programming that will be available in the near future.

An ETODS system is quite attractive in this application. A digital video recording system based on 4x subcarrier sampling of standard NTSC composite color video (i.e., the D-2 standard) requires approximately 1 GB per minute of digitized video frames, and a transfer rate of 120 Mb per second. A 130 mm ETOM disk could store 14 minutes without compression and offers 50 ms access time to any frame. With a data compression technique such as the proposed MPEG-2 standard, the same disk could store 18 hours of compressed digital video programming.

Video Post-Production
A digital video recording system for post-production editing would offer random access editing of digital video at much higher speeds than existing serial access tape machines and at far less cost than the usual tape-based products (e.g. D-1 and D-2 machines).

Video Distribution
A digital video recording system integrated into a conventional in-home cable TV converter box would enable the delivery of true "Video-On-Demand," because it would allow the consumer complete control over the time of viewing. The lack of true Video-On-Demand is widely believed by many industry experts to be the primary reason for the slow acceptance of Pay-Per-View programming. Hours of digitally compressed programming could be downloaded in minutes via cable, fiber optic, or Direct Broad-

cast Satellite (DBS). Consumers would then have the in-home convenience of Pay-Per-View combined with the breadth of selection and control of viewing time offered by video tape rental. Moreover, with adequate auditing and reporting safeguards built into the ETOM-based system, motion picture producers could offer first-run movies directly to the home viewing audience.

Consumer Products
A merger of the television and the computer is expected during the next decade. Consumer products based on the ETOM technology could also store video games and interactive video and multimedia applications.

Telecommunications

The merger of television and the computer and the linking of homes and offices using high performance optical fibers will create opportunities for ETOM-based systems in telecommunications. An example is the potential early use of ETOM-based buffer storage units in fiber optic systems. Increased capacity buffer storage will be important when data is downloaded from very high capacity "trunk" lines for regional and local distribution of data.

Computer Data Storage

While advanced digital video recording offers an excellent near-term opportunity for the commercial introduction of ETOM-based products, the computer data storage industry is perhaps the largest market in which ETOM technology could have an impact. ETOM's data density, data transfer rate, and potentially low cost, would make ETOM-based products attractive to a broad range of current computer users. Today's computer systems are becoming more and more sophisticated, yet their performance is often peripheral-limited. Although there are many different data storage devices (tape, optical, magnetic 'hard'" drives) that *partially* meet a computer user's needs (random access, removability, cost, performance, etc.) not a single one of them meets all his needs. ETOM technology allows for the development of products which meet all of the user's critical needs.

Summary

Perhaps the largest impact that ETOM-based products would have on technology stems from the fact that it is a truly photonic technology. Light is simply a better messenger than electricity. Optical fiber can carry far more information than traditional cable. Computer logic devices based on photonic processes are under investigation to make computers operate faster. Optical video and data storage as well can now benefit from the application of photonics.

Because of its high performance, an ETODS system would provide the data storage break-through needed for satellite data and information systems, broadcasting and video distribution, telecommunications, and computer data storage.

References

1. D. T. Brower and R. E. Revay, "Tuning of read/write/erase processes in Electron Trapping Optical Memory media," *Optical Data Storage '92, Proceedings of the SPIE*, ed. D. B. Carlin and D. B. Kay, vol. **1663**, pp. 86-91, Feb. 1992.

2. Urbach, F., Preparation and Characterization of solid luminescent materials, ed. Fonda and Seitz, John Wiley, Chap. 6, pp. 115-140, 1948.

3. A. M. Earman, "Optical data storage with electron trapping materials using M-ary data channel coding," *Optical Data Storage '92, Proceedings of the SPIE*, ed. D. B. Carlin and D. B. Kay, vol. **1663**, pp. 92-103, 1992.

4. T. Kasami, T. Takata, T. Fujiwara, and S. Lin, "On Multilevel Block Modulation Codes," *IEEE Transactions On Information Theory*, vol. **37**, no. 4, pp. 965-75, July 1991.

5. M. Ozaki, T. Furukawa, K. Tanaka, and T. Kubo, "An Effective Reproducing Method on Digital Optical Disk," *6th Conference on Video, Audio, and Data Recording*, pp. 105-11, 1986.

6. H. Kobayashi, "A Survey of Coding Schemes for Transmission or Recording of Digital Data," *IEEE Transactions On Communications Technology*, vol. COM-19, no. 6, pp. 1087-1100, Dec. 1971.

7. Personal conversation with R.L. Brower, July 2, 1992.

Panel Discussion on Magnetic/Optical Recording Technologies

Dr. P C Hariharan of Hughes STX was Moderator of a Panel Discussion on Magnetic/Optical Recording Technologies held September 23, 1992, at the Goddard Space Flight Center in Greenbelt, Maryland.

Members of Dr. Hariharan's panel were

Mr. Martin Clemow, Penny & Giles Data Systems, Inc.
Dr. Jean-Marc Coutellier, Thomson CSF / Laboratoire Central de Recherches
Mr. Bruce Peters, Datatape Inc.
Dr. Dennis Speliotis, Advanced Development Corporation
Mr. John W. Corcoran, Corcoran Associates
Mr. Allen Earman, OPTEX Corporation
Mr. William Oakley, Lasertape Inc.
Mr. Andrew Ruddick, ICI Imagedata
Dr. Yuan-Sheng Tyan, Eastman Kodak Co.

DR. HARIHARAN: Allen Earman is Manager of Systems Development Engineering at the Optex Corporation and got his MS in optics from the University of Rochester and BS in electrical engineering from the Virginia Polytechnic Institute and State University. He has worked in the field of optical recording for over 16 years and has presented four technical papers on the subject, most recently one titled, "Optical Data Storage with Electron Trapping Materials Using M-ary data channel coding." He has one use patent and another pending and has cochaired two optical data storage conferences. Mr. Earman is a member of SPIE and the IEEE LEOS, Magnetics and Information Theory Societies.

John Corcoran received the bachelor of electrical engineering degree from the Manhattan College and the master of electrical engineering from the Polytechnic Institute of Brooklyn. He migrated to California to work with Beckman and Whitley on high-speed photography. Subsequently, he worked on optical recording in the Advanced Technology Division of Ampex and was then converted to magnetic [inaudible]. He retired last year but keeps occupied on questions of archival storage, error characteristics, etc.

The remaining participants have already been introduced when they presented their papers.

I had hoped that I would be able to put up a chart from a paper that Mark Kryder was invited to write in 1989 in which he was asked to make some prognostications. But those of you who heard his paper earlier this morning will know what has been achieved by this year; and, if you had read his 1989 paper, you will also know that most of those achievements were expected by the year 2000, and not by the year 1992. So, one of the things that I would like to do in the discussions today is for the panel assembled here to state what they think the technology is capable of doing in the next 5 years, or may be the next 10 years. Mark has already given us his views. And in a couple of years, we'll have most of these people back, including Mark, and we'll ask them to sit back and tell us whether their predictions are on course, ahead or behind, and why.

capacity, data rate, generally getting much better performance from the available media in the immediate future. I don't see any reason why that should change as things go on. *I think Dennis is quite correct: any predictions we make are going to be short of what will actually really happen.*

MR. RUDDICK: I think there's another dimension to this issue and one that sometimes — I won't say it's forgotten about -- but we've heard a lot earlier on about some of the fundamental limitations of the technology, and I could prepare slides showing the fundamental limitations of optical technology as it applies to laser wavelengths, et cetera, et cetera.

The other important dimension that I think is important, however, is some of the engineering issues that need to be also considered to turn the fundamental limitations into real products. For example, some of the issues around substrate performance are going to be critical if any of these fundamental magnetic limitations are going to be achieved. I think that broadens the debate away from, if you like, the physics and the chemistry to some of the real practical engineering tasks associated with turning bright ideas into real products that are going to solve the data storage problems that we've heard about. And that makes predictions even more difficult because, it's easy to predict in some ways about the physics and the chemistry. It's not quite so easy to predict, for example, what laser powers are going to be available in five years' time, prices that are going to be appropriate for optical data storage. It's not easy to predict, for example, what the quality of substrates is going to be like in five years' time for the demands that these high-density storage systems are going to place. So, I think that's one comment I'd like to make.

DR. COUTELLIER: In my case, I have a double difficulty in answering that question, since we are dealing with both magnetics and optics in our system. In fact, it is a major advantage, for the following reasons. The density of information which can be recorded on tape is currently limited to 1 square micron per bit. The longitudinal resolution of the magnetic head is much better than that of conventional optics. In the latter case, it is limited by light diffraction phenomena. In the new Kerr readout component which I presented this afternoon, the longitudinal resolution is given by the width of the non-magnetic gap located between the two magnetic poles, as it is for a conventional inductive readout head. An optical head provides much higher transverse resolution. As a matter of fact, the magneto-optical layer constituting our Kerr transducer is continuous all along the tape width, the transverse resolution is then limited by the laser light diffraction phenomena. So, for both longitudinal and transverse resolutions, we use the best of magnetics and optics. This is one of the major advantages of the new recording system. The recording density is limited only by the tape pigment and coating.

MR. OAKLEY: I have three comments. The first one is a shootout between optical and magnetic. I don't believe laser power to be a limiting factor. The reason for that is that just within the infrared laser diode domain, it has already been demonstrated that by putting a small crystal amplifier in the laser cavity, the single-mode laser powers can be raised to approximately 10 watts. And that takes the data rates to terabytes per second. The cost is minimal. If you're talking about shorter wavelengths, then the doubling efficiency is approaching 50 percent, so several watts in the ultraviolet is probably possible within a few years.

To address the issue of ultimate capacities, I'd like to point out that all we've discussed today, except for Optex, is a spatial limitation to storage. Within the optical domain, you have another dimension you can use, which is the actual color of the laser itself. In the research labs, people have demonstrated spectral hole-burning techniques whereby shifting the laser wavelength very slightly allows a second bit to be recorded in the same location. Theoretically, this will allow something like 10,000 bits to be stored in each square micron of media, optically. And that raises the capacity of optical storage to whatever the number is. It's some horrendous number. So, the spectral modulation is a dimension that no one's even considered yet, in terms of limitations. Thirdly, in terms of technology change, the rapid technology change, I think we can solve that by just invoking the present financial, fiscal environment, which will limit technology change by rationing investment. So the technology can be very stable there for a long time.

MR. PETERS: I might offer a slightly different perspective, that is, I think that regardless of what technology or product implementation you're talking about, we're going to be hard pressed to keep up with the demand. If you look at what's happening in our ability to collect data, as well as our appetite to process it, we're going to really have to hustle on all fronts to keep up with it. And I dare say that in five years, users will still be complaining about data overload. With respect to magnetics, I would say that we're not limited by technology, but only in implementation, and there too only in the short term. Implementations will be driven by economics and our ability to deliver real, useful products. But there always seems to be that engineer who has a different approach that is very viable right around the corner to solve a particular implementation problem that we have. So, with respect to magnetics, I would say that we're still seeing very real potential for orders of magnitude areal density improvements and other factors that you would normally associate with them.

MR. EARMAN: Every couple of years or so, we try to take a measure of what we think the predicted improvements are going to be over the next couple of years. Maybe we do this every year now. But every time we do that, we overlook things that change very rapidly or new discoveries that happen from time to time. For example, just two years ago people were wishing short-wavelength lasers were available that could increase the capacity of a disk by four times or maybe six times or whatever. The recent work in blue and green lasers has been extremely exciting and extremely fast-changing. Also, in the spatial domain, we mentioned earlier, we can determine fundamental imaging limits based on wavelengths and the numerical aperture of the lens. However, AT&T's recent announcement of the stretched fiber end introduces a whole new field of investigation in that area. With the end of the fiber and using near-field rather than a far-field imaging, the capacity again goes up by a factor of perhaps three orders of magnitude. As my colleague also mentioned, about the multispectral recording, and as we talked about earlier, non-binary recording, all contribute to greater capacities and this is in leaps and bounds, not just gradual changes. So, trying to put an expectation on where we're going in the next five to eight years, or by the turn of the century, gets harder and harder, and makes more fools of the predictors.

DR. TYAN: I guess I'm chicken. I'm supposed to be on the optical side. But I would say that magnetics is probably going to be here forever. Optical is never going to replace magnetics totally. But I think we have learned in the past two days that the storage industry has become more and more diversified. And people are not going to judge the technology in the future just based on capacity. Maybe people will not be satisfied with 2000 hours of head life, or 2000 passes of tape life. And other kinds of considerations will become very important. So I think in terms of the future, there will be a need for both optical and magnetic. Just a matter of which technology will be the best for a particular application.

MR. CORCORAN: It's kind of awkward to be "tail-end Charlie" in such a distinguished group. I'll make a couple of comments, though. I've watched over the years the projections of resolution limits in optics and magnetics move slowly upward, and it never seems to stop. There's increasing sizes of memory required and there's the question increasing costs. The cost of memory itself has declined enough so we can keep expanding memory. I think the harder thing I see is how can we control the costs of the equipment that we're going to build so that America can produce things that will sell in a world market. In some ways I think that that's perhaps a worse challenge than anything we can do to increase the size of memories.

DR. HARIHARAN: Are there any comments or questions from the audience? Yes?

PARTICIPANT: [inaudible]

DR. HARIHARAN: Atomic force microscopy?

PARTICIPANT: Where does that technology go, and who will take advantage of it -- the magnetic guys or the optical guys?

DR. HARIHARAN: Well, Mark has a comment on that.

DR MARK KRYDER *(Carnegie Mellon University, and Session Chair):* I can make a comment on that. The problem with that technology is basically the same one that is faced by the AT&T/Bell Labs experiment that I was talking about earlier. All of them rely on gates of electric transducers and basically IBM's present, published best values is 100 KHz in a case where they were actually moving the field emission tip relative to the recordig medium. They project that they might be able to get it up into the megabit or maybe, even a few megabits, per second. But that's about as far as they can go. So there is a real limitation with regard to moving the atoms around on the surface as to what sort of data rate they can get out of a system that does that. I think that's the obstacle there, and that's why I suggested this morning, with the near-field optical scheme, that the preferred approach is to use a flying head to make it look like a disk again, and so forth.

DR. HARIHARAN: Dennis, do you have a comment?

DR. SPELIOTIS: Not specifically for that, but if you consider electron beams, which is a technology that was not mentioned here, diffraction limits go down by four or five orders of magnitude. You have interactions that are unimaginable in the optical field; electron beams interact with anything. Deflection is trivial -- maybe too easy. You have everything, except that you need a vacuum and nobody's pursuing it. There's another dimension that we did not even discuss at all.

DR. HARIHARAN: You did work in electron beam technology, didn't you?

DR. SPELIOTIS: Right.

DR. HARIHARAN: We don't have the holographic memory people here either. There's an outfit in Texas that has been working on this for quite a while. Does any other member of the panel have a comment on this?

MR. CORCORAN: I could comment that between 1960 and 1973, Ampex built about six different electron beam recorders for various analog situations. And they achieved up to 100 MHz analog bandwidth, but there doesn't seem to be any real market. I think we're going to go more digital. We won't go holographic.

DR. HARIHARAN: I saw another hand raised over the back. No? If not, we'll go to a point that Dennis raised while the optical people were giving their talk. It is remarkable that there seems to be no activity from the Japanese on optical tape. Do you wish to comment on that, Dennis?

DR. SPELIOTIS: To me, it's remarkable, because a couple of years ago (I don't know exactly what the situation is now in Japan) there were over 100 companies or laboratories in Japan pursuing magneto-optics, mostly on rigid disk formats, for the development of materials for magneto-optics. Over 100 companies! And in the U.S., hardly anybody--only two or three companies. With that tremendous Japanese involvement in magneto-optics, there is hardly anything on optical tape. It is strange and I don't know if anybody has any insight into that.

MR. RUDDICK: I don't know whether it's great insight or not. We spent some time looking at magneto-optic tape as a technology in terms of feasibility and, I think, have concluded that the lifetime issues are such that it probably isn't a viable way to go. The corrosion characteristics of those materials tend to be very rapid and in the context of a rigid disk, which you conceal and enclose, it's a manageable issue. In the context of an open-reel tape or cartridge tape with a very thin protective coating perhaps, it seems much more of an issue, and my judgment is that wouldn't be the way you would go if you were developing an erasable optical tape product. On the more general comment about the fact that there is no optical tape activity in Japan, our information suggest that isn't strictly true and that there are some companies who are active. But they are promoting their activities toward consumer applications for high-definition digital TV. I'm sure we'll hear more about that in the next few years.

PARTICIPANT: Anybody out there putting optical tape in a cassette format?

DR. HARIHARAN: Bill Oakley has tried using it in a 3480 cartridge and you want to know whether it'll be in a two-reel cassette?

PARTICIPANT: Yes.

DR. HARIHARAN: Fine.

MR. RUDDICK: Yes, there are people working on that problem, and we are working with a number of companies who are looking at that issue. We can't talk about it, because it's not public, but there are activities.

MR. OAKLEY: I'd like to comment from the systems standpoint, which is that when you're talking about putting 100, perhaps 500 gigabytes of data in a single cartridge, you have all your eggs in one basket. In a multiuser environment, that is completely ludicrous. It may well be that we'll end up with both single-user, very large systems like CREO and also multi-user systems with a wall of small drives with only 10 gigabytes per cartridge, you know, in a microcartridge. So it'll go in both directions.

DR. HARIHARAN: Do you have a question?

DR. SPELIOTIS: One of the issues that Hari raised is that it seems like the pace of technology is accelerating, and the product cycles become shorter and shorter. A question: Why is that happening? What is driving it? Is the demand in the data storage or other applications requiring this kind of fast pace? Is it the threat of other technologies that pushes some of the technologies to advance faster? What are the reasons behind this, because it seems to be getting to a pace where the economics will not be there to sustain this growth. I mean, if the product cycles are so short, people will not be deriving the revenue out of these product developments to sustain further growth, so we're going to sort of commit suicide eventually. So, why are we doing this? What is driving us? Can there be some order and logic in this kind of pace? I'm puzzled by it, and I don't have an answer. Maybe some people have opinions about this.

DR. HARIHARAN: Linda?

MS. LINDA KEMPSTER *(Strategic Management Resources):* I think that's what's happening. Speaking as President of the local chapter of the Association for Information and Image Management, and I have 800 people in my chapter, the world I address is a paper management world. I have a person who is in my chapter that measures the number of documents by Washington Monuments. OK? It's sort of like the Library of Congress. She said, "I have 2 1/2 Washington Monuments full of paper." This person will never come out with an RFP to go on optical disk. The cost— I mean, it would be the national treasury here. But when she finds out, and I've been educating them — on what kind of economics magnetics can bring, because the people with paper problems have never thought about putting paper solutions on magnetics — then they'll come out with that kind of requirement for systems that can be responded to by the high-density magnetic tape solutions. So what's happening is that the tape people, who have always addressed or found their home in instrumentation data problems, and solving those problems — are now being forced — not forced, but encouraged -- to look at other markets where we have tremendous paper repositories that are trying to get on something electronic. Those folks have bumped into a very expensive optical disk ceiling, and they're looking for other solutions, and I think that's what's going to drive your market. That's where your solutions with tape are going to go, and that solution is going to go from magnetic tape to optical tape, because they still want to have the archive ability and what they think of as nonerasability and so forth. So it's going to be the paper industry, the microfilm, the people that used to put all their stuff on microfilm. Those people are going to be driving your industry, and that's where your customer base is coming from. Any other questions?

DR. HARIHARAN: Yes, Bill.

MR CALLICOTT *(NOAA):* I don't think [inaudible]

DR. HARIHARAN: You're saying that data which is not online and which is so voluminous is useless?

MR CALLICOTT: That's correct.

DR. HARIHARAN: OK. Mark?

DR. KRYDER: I think that's exactly the right answer. Let me comment that every new PC has come out at that rate, in 18 months. And that's what's driving [the pace of development]. Every PC has to have a new disk drive in it. It has to have a higher capacity and so forth. So, very clearly, that' what's driving it. I don't think that's negative, though. If you want my opinion, as far as the U.S. is concerned, the truth is, the disk drive industry resides in the U.S., and that's the one that has the short cycle time. The one which hasn't had the short cycle time has been tape -- VHS and so forth. All those are offshore. If we allow these industries to have such a long product lifetime that it comes down to only cost of manufacturing, somehow the community in this room, the U.S. people, we don't do well in that. In innovating and bringing out new products quickly we are good -- in fact we've done well in the disk drive business.

DR. HARIHARAN: Any other comments? Yes, Dave?

DR DAVID ISAAC *(MITRE)*: [inaudible] IBM disk drives [inaudible] to the capacities of individual disks. The driver there seems to be access times, rather than capacity. Will we reach a natural limit in terms of cartridge size?

DR. HARIHARAN: Excuse me, did everyone hear the question, or should I ask Dave Isaac to come to the front and repeat it? Dave?

DR DAVID ISAAC: Hari, the question I had for the panel. In my exposure to IBM disk drives over the past 5 or 10 years, I've seen the capacities of the individual disk drives remain around a couple of gigabytes rather than growing larger and larger because the platter sizes have dropped from 12 inches to 10 inches to 5 1/4 to 3 to 2. It keeps going down to offset the increase in areal density. And the prime mover there seems to be access times rather than capacity. So I was wondering if we see these capacities going up in tapes with the helical scans and the opticals and the potential there. Really, are we going to reach an actual limit in terms of cartridge size? Will the cartridges start getting smaller? Is there a natural limit in terms of how much data people want to handle in one chunk in tape, or is the application of tape so different from that of disk that it really doesn't apply?

DR. HARIHARAN: Well, we saw the Sony nontracking technology tape yesterday, which is less than the size of a credit card. Let me pass the microphone along to the other members here.

DR. SPELIOTIS: The Sony cartridge — the so-called SCOOPMAN is the size of a stamp. So it is getting very much smaller than previous cartridges, but the question is: does that help throughput? Probably not. There is a real problem, I think, in the tapes. As I see it, how do we combine the best features of helical scan, which is typically large-capacity and slow throughput rate and low cost for the data cartridges, with the 3490 technology, which is very high cost, very high throughput, but relatively low capacity. We need an imaginative solution that would combine these two. If we can get that, I think it will answer a critical need for the next few years for several applications.

MR. CLEMOW: I'd like to make two points here. One is, I think that the capacity per unit item — whatever it may be, cartridge, cassette — from the tape point of view, is going to be driven by the application. I think you alluded to the fact that there is a limit to the maximum size per unit that people require for their application. I think that will be the natural breakpoint, if you like, that we'll get to that limit and then people will want the size reduced. And that's where the second point I'd like to make comes in. Particularly with tape drives that involve heads, reel motors, all the rest of it. You've got a problem with the mechanics. And you'll be getting into

micromechanics, and that is yet another field to consider. It's an unknown one at the moment, certainly from my point of view. It's a whole new technology, and I think the limit on the size with tape cassettes, tape cartridges, will be determined by the mechanics of the situation. Again, you come down to a cost. You get into a whole new field. It's an unknown at the moment.

MR. RUDDICK: I guess my response is that's a very complicated equation, and I'm not sure I understand all the issues around it. But certainly, the granularity of the data is something that I think does have a definitive size, depending on the application and the system design. The 1-terabyte capacity tapes certainly have their niche and their application, and the 100-gigabyte, 50-gigabyte, 3480 cartridge similarly. And, as I said earlier, we also have development programs looking at smaller formats. And again I think those sizes with those capacities and that granularity, also have their applications in different systems. So I don't think I can give a simple answer to that question. It's very application-system-dependent.

DR. COUTELLIER: The amount of data that can be stored is proportional to the magnetic surface you have in your cassette. And one way of decreasing the access time would be to change the form factor of the cassette itself by using smaller tape lengths and higher widths. This raises the question of track-following on wider tapes. You will have to deal with shrinkage, temperature effects, things like that. What kind of track-following servo will be implemented on the system? It could be a mechanical adjustment of the head on the tape. It could also be, as we can do in our recorder, an electronic track-following system. We have indeed a very nice feature available: if the optics is made such that more than one CCD pixel is associated with each track, we get many samples per track. If we are able to recognize on which pixel a servo track is located, it is then possible to shift the samples electronically in registers, and dispense with mechanical adjustments.

DR. HARIHARAN: The Sony people have shown that you don't really need to track anymore.

DR. COUTELLIER: Yes, but then the helical track length has to be made short to reduce the number of bits which have to be stored before data recovery can be attempted. The tape width is reduced, so to compensate, it has to be made longer to store the large number of bits needed for computer backup, HDTV, etc. This raises the problem of access time once again.

DR. HARIHARAN: But the NT cassette is the size of a postage stamp; even though it is not being used for digital data storage now, it does hold 693 Mbytes of digital music, and thus exceeds in capacity the popular CD-ROM.

MR. OAKLEY: I'd like to answer a question that really wasn't asked but is fundamental to this. The question to ask was posed in terms of latency for disks. I'd like to point out that the intent of the MCC program in Austin, Texas, is to replace disks. That's a crystal holographic system, and their intent there is to have a 10-gigabyte secondary memory with latency of less than 1 microsecond. So the hierarchy of a future system, consists of CPU, and a holographic crystal memory. And that crystal memory is supported by a tape system. So the real question is, How big is your holographic memory, and what is the cycle time on that? It seems to me if you have a 10-gigabyte holographic memory, what you need to do is load data sets, 10-gigabyte sized, to that every few seconds. So tape array's the answer to that, with maybe 10 seconds access time on the tape array. That's a prediction, by the way, about the demise of disk drives for large systems.

MR. PETERS: My comment is really a follow-on to a comment that was just made. I think we're going to see new things in the future, just like we've done in the past and maybe at an increasing rate. One of the things we will see is well-packaged, hybrid storage subsystems that actually are combinations of the various technologies and media for specific applications.

MR. EARMAN: The tape-based systems, unlike the disk-based systems - well, actually even the disk-based systems to a certain point - are really application driven. And depending on which application you're using determines the size of the acceptable amount of data that can be stored. The example that Bill mentioned earlier was that of 500 gigabytes on a single cartridge

on a multiuser system. It would be potentially catastrophic because of the loss of information if something happened to that one drive or that one cartridge. We also can draw an analogy to the consumer video system, such as the VHS and Beta systems. The original concept behind the VHS cartridge was to get 2 hours of video on a single cartridge because that would cover most movies that were available. So, you could store a single movie on a single tape. And of course with the 3X increase in capacity by going to the slower speed, you were able to get up to 6 hours, or 3 movies, on the tape. There hasn't been a lot of drive to push the capacity much beyond that, because if you could possibly put 18 movies or something like that, on a single tape, the loss of that tape would be catastrophic to your movie library, whereas a loss of 1 to 3 movies is not so significant. And that same idea, of a minimum volume and a maximum volume, also came along with the D2 systems. The medium cassette for the D2-based digital video contains about 94 minutes of capacity in digitized, full-rate (*i.e.*, not compressed) video. Again, that was driven by having a minimum of 90 minutes on a single cartridge, even though that D2 cartridge is physically larger than the VHS cartridge.

DR. TYAN: Maybe just a follow-on to your comment here. When the technology becomes available that you can do more, you can store more. You can generate new applications, too. For example, in video cassette, people are working on tapes which can store HDTV; it requires the storage of much more information. Another example is for document storage. I think it goes both ways.

MR. CORCORAN: In my mind, the size of the file and the access times are inherently linked. You don't want to make a file too big or you just make your ability to find an item very difficult. It's all right in the TV system -- we have a D2 cassette, which I believe holds 3 hours. And it's quite large. But they are putting them in some storage systems. I kind of think it's monstrous. It's much better to handle a block of perhaps 26 gigabytes, which is the small cassette in the same class. It's a more convenient package for the storage of data.

DR. HARIHARAN: Let me ask another question here. The issue of VHS was brought up. VHS was a competitor to the Beta format introduced by Sony. Is there some such competition going on which has caused the displacement of D1 by D2 and of D2 by D3? Will the use of these technologies for data storage result in some of us being left with unsupported or unsupportable hardware and/or format?

MR. EARMAN: I'll make a comment on that. I'm not very familiar with the D3 configuration, but regarding the D2 and the D1, right now, they're used for quite different purposes. The D2 is based mainly on composite video, and the D1, on component video. For general purpose usage, the D2 composite video is sufficient. But for most production houses, where television, commercials, whatever, especially with a lot of computer-animated artwork, is being done, component video is really gaining ground over composite video. There's more of a desire to have D1 systems in those setups. So there's still quite a bit of competition for different purposes and different uses between D1 and D2.

DR. HARIHARAN: Larry Lueck has recently predicted that magnetic recording is a dying duck. Now you are saying that D1, in particular, is not a dying duck and it's not a dead duck. That right? Allen?

MR. EARMAN: No, not in the near future.

DR. SPELIOTIS: The question I think is very interesting and I would like to see a show of hands: what is the opinion of the audience? Is magnetic recording going to die sometime soon and be replaced by a completely solid state, electronic type memory, or will it just keep growing and electronic buffers will coexist and support and extend its usefulness?. What is the opinion of the audience? Because some people are predicting that we are in a dying industry and we're going to see in the very immediate future a decline in the volume and applications and revenues of this industry. Does anybody feel that the industry is threatened by solid state to the extent that it would go out of existence by the end of the decade or something like this?

MS LINDA KEMPSTER: I would think that the magnetic tape industry is going to die when every MIS data processing guy back in the lab, doing the same thing every day, dies.

MS LINDA KEMPSTER: I was introduced to optical disk in 1983-84. When I described optical disks to the MIS folks they said, "No way am I putting my data on that shiny blue disk, I've got my tapes." As long as you have a whole generation of those people, you will have a tape market. Maybe they'll change to optical disks in another generation down the line. As long as you've got some of those old tape dudes, you're going to have tapes. They're going to be around.

DR. SPELIOTIS: But there are a lot of new tape dudes coming along. So they'll never die.

MS LINDA KEMPSTER: This is true, as tape formats, capacities and applications expand, there will be *new* tape dudes and dudettes.

DR. HARIHARAN: Thanks. Pat [Mr Patric Savage of Shell] Would you like to make a comment along the lines of: We take care of our eggs by watching our baskets? Would you please come to the front so everyone can hear you?

MR SAVAGE *(Shell):* We do put all of our eggs in one basket, but we watch that basket!

MR. OAKLEY: Hari, when Pat gets done, I have a comment I'd like to make as well, about the massive upcoming proliferation of tape systems.

DR. HARIHARAN: OK.

MR SAVAGE: What was that last question? The problem that we have in our industry is exemplified by the scene that we have at Shell. Some handful of years ago, we converted to 3480. We didn't convert; we stopped usign round tape and started using 3480. We now have about 900,000 3480 cartridges of permanent data. And we still have, unconverted, about 1 million round tapes of 6250 data. We have migrated all of our 1600 bpi stuff and all previous data, so we only have 6250 round tapes and 3480 right now. The problem that we have obviously is that migrating a million reels of round tape is an extremely expensive and long-enduring project. You simply can't do it overnight. It's labor-intensive. It's the kind of thing that you would like to be able to migrate now, and not have to do that again for another 30 years. On the other hand, if we were to migrate it to 3480 right now, we know the 3480 technology is already obsolescent as the 3490 technology takes over. We would be converting to something that, yes, exists now, but the data on those round tapes is not likely to be used very much during the lifetime of 3480 technology, so it would soon again have to be migrated. I'd like to put it in an icebox and sock it away for a long time. On the other hand, whatever I put it in, I want to be very, very sure that at the end of that archiving period or epoch, there are still drives around that I can read my data back with. Probably not much concerned with the retrieval data rate at that time, because we will probably copy it from the archive to the currently in-use medium. So, migrating to a new technology that probably will not last for more than 5 or 6 years is a very costly thing for me to be looking at. I would like to migrate to something that is as permanent as possible but with definite assurance that drive technology will be clearly available at the end of that archiving period. I would like to make another comment. I believe that what I heard here from the speakers and what I know of the MCC holographic storage in doped crystals are that these kinds of technologies could actually be brought into effective marketing position. I have enough vision to see that the MCC stuff, in principle, could literally take over everything that the rest of you guys are doing. That's all of the optical, all of the magnetic, and all of the magneto-optic. In principle, it could do that. It would take some money to make that happen. And it really could happen somewhere around the end of the decade.

DR. SPELIOTIS: But the idea has been around for 30 years, so what makes you think that it's going to happen by the end of the decade?

MR SAVAGE: I know. I've been tracking it myself for 30 years, Dennis. The difference is the clever little advances that have been made in the new ideas that MCC has brought into this area. My latest contact with them was about a year ago. I left believing that if they had enough money and had enough vision, that they could bring it to pass. It's got all the qualities that we really want for deep archiving -- high bandwidth, low cost, random access, all those good things.

DR. HARIHARAN: Infinite stability?

MR SAVAGE: Yes, infinite stability.

DR. HARIHARAN: Inexpensive too?

DR. SPELIOTIS: Motherhood.

MR SAVAGE: All those good things.

DR. HARIHARAN: You said you had a comment, Bill?

MR. OAKLEY: OK. I'd like to change the direction of the discussion slightly by pointing out that in the computer environment, historically we've seen diminishing interest in very large systems. Thirty, twenty, even ten years ago, we had one CPU serving thousands of users, and that system is rapidly disappearing with the advent of workstations and distributed computing. What we're talking about here today is massive data centers. The data being distributed over networks is very much like the old multiuser system, but it turns out that with the optical tape type technology, if you can get 100 gigabytes in a 3480 cartridge, you can replicate that quite cheaply. You can mail it across country overnight for a few dollars. And 100 gigabytes, by the way, is equivalent to 24 hours' continuous transmission on an Ethernet line. So perhaps the way of the future is not to have very large databases with massive fiber-optic, high-speed data nets but just to use a whole pile of FedEx envelopes and a stack of 3480 cartridges. Every user gets his own 200 or 300 gigabytes that he's interested in.

DR. HARIHARAN: Bill, my thesis advisor was a graduate student at Cambridge University in England, and he was one of the users of the very earliest computers called EDSAC — Electronic Delay Storage Automatic Calculator, and the memory there was mercury delay lines. He used to tell me that if more than two people got into the computer room, the temperature went up sufficiently that the number of bits that could be stored in the delay line changed. Now, we already have 5 million miles of fiber in this country, I am told, and if we can put in another 50 million miles of fiber and we do get the gigabit network, in principle I can ask Sam Coleman to start pumping in the data at Lawrence Livermore and have regenerative repeaters here in Baltimore or Goddard and recirculate the thing and I can have a memory, of the grand-old type that they were using in the early machines, but with substantially higher capacities, available. Anybody connected to the network can capture the bits as they flow by. Maybe it's slow access.

MR. OAKLEY: I can't respond to that. What I had in mind was that with the rapid growth in workstations, if there would be a market developed in using optical tapes in workstations, then the cost of the 10-15-megabyte/second, 100-gigabyte, tape drive, in that kind of volume, is going to drop down to maybe $1000 per tape drive. So that means the user with an autoloader, just 10 cartridges, each at a 100 gigabytes, is going to have, for maybe $1500, a desktop terabyte system. And that will impact the use of networks.

DR. HARIHARAN: I'm sure that those things are going to occur, because workstations are getting more powerful and Andy Heller has threatened to unleash the Godbox -- giga everything on your and my desk--gabits, gigabytes, gigaflops, gigainstructions per second. That should to come to pass. And he said it's possible within the next 3 years for a price of around $30,000. There's going to be a lot of activity in, not just processing data, but in reprocessing data that we already have. I'm sure we'll find ways to put to use the new networks, the high-capacity links that we have, and we'll make pretty good use of those.

Introduction of Dr. Dennis Speliotis (After-Dinner Speaker)

DR HARIHARAN: Ladies and gentlemen, it is my pleasure to introduce Dr Dennis Speliotis as our after-dinner speaker. He was born in the Peloponnesus in Greece, but came to the US for his college education, obtaining his bachelor's and master's degrees in electrical engineering from the University of Rhode Island and the Massachusetts Institute of Technology, respectively, and his Ph.D. in solid state physics from the University of Minnesota. After graduation, he worked at IBM till 1967 before joining the University of Minnesota as Associate Professor. There he founded the Magnetics Research Laboratory. He then became a co-founder, director, vice president and general manager of Micro-Bit corporation. In 1967 he founded Advanced Development Corporation and continues as its president. He started Digital Measurement Systems in 1984, and has been able to sell the measuring equipment made there even to the Japanese.

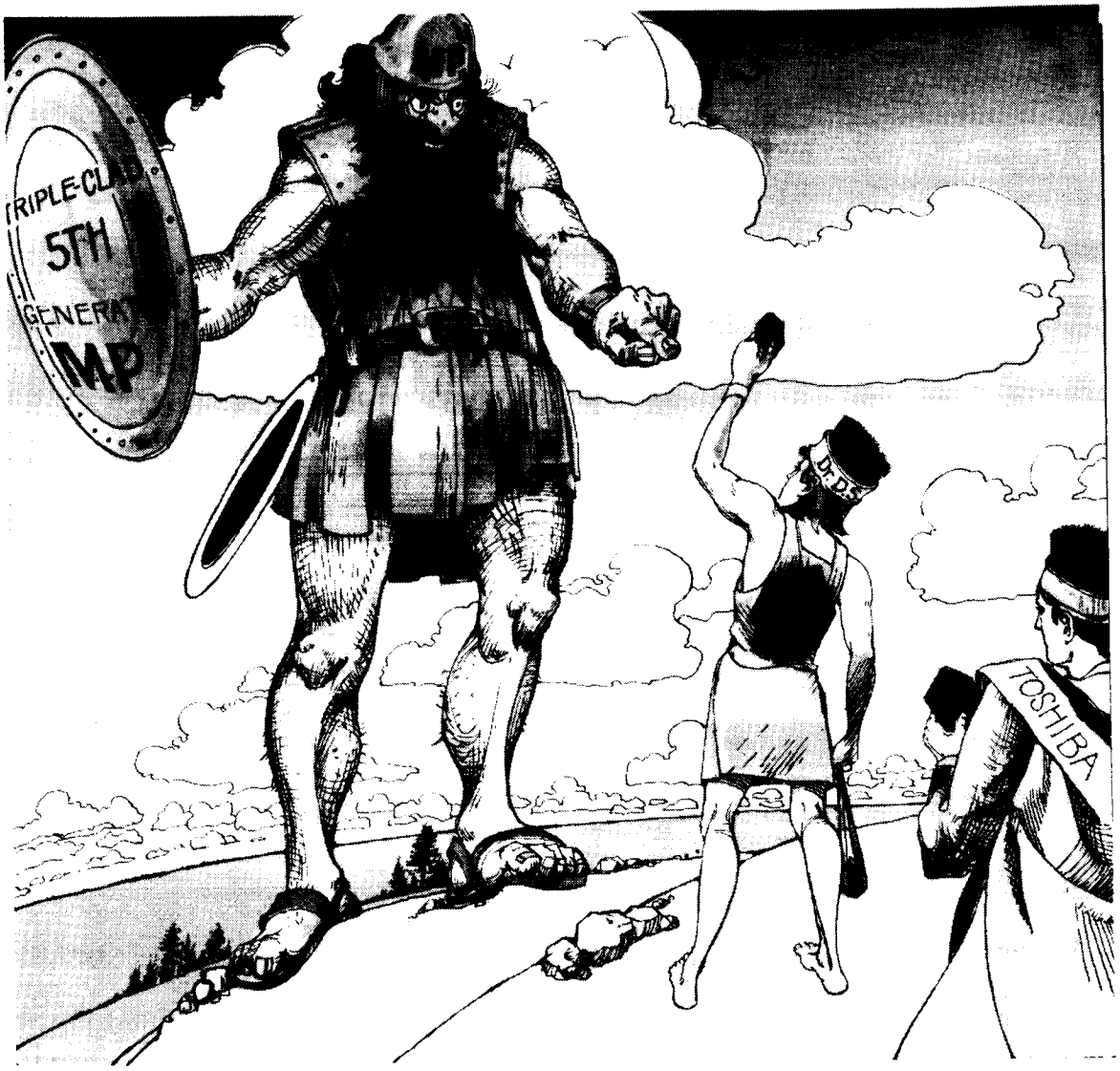
Dr Speliotis is the author of over 150 technical papers, has been an IEEE Distinguished Lecturer, and has been an invited technical speaker at over 30 international conferences. He has also organized numerous conferences and symposia, and is as energetic and productive today as he was 30 years ago. In September this year, I had the privilege, along with a select group of others, to proceed on a pilgrimage in search of the elusive hexagonal Barium Ferrite platelets. Larry Lueck has captured the essence of the battle Dr Speliotis has been waging on behalf of this elixir of magnetic recording in the cover art of his MMIS Newsletter. Let me add that the First International Symposium on Barium Ferrite in Kalamata, Dr Speliotis' birthplace, was a great success, and brought together some of the keenest intellects at work in the field of magnetic recording.

Without further ado, let me present to you Dr Speliotis.

VOLUME XIII NUMBER 1

MAGNETIC MEDIA

INTERNATIONAL NEWSLETTER



Debating the merits of MP and Barium Ferrite in advanced media configurations.

MMIS

Researched, written and published by MAGNETIC MEDIA INFORMATION SERVICES

Data Storage: Retrospective and Prospective

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

We study history to learn from its lessons so we don't repeat the mistakes. Ironically, however, as Pat Savage of Shell Development remarked, sometimes it seems that the lesson we learn from history is how to repeat the mistakes more precisely. In this brief talk, I would like to reminisce a bit about the history of magnetic recording, and use the lessons of the past to look into the future.

Magnetic recording is an extraordinary technology. It is so pervasive and so difficult to replace or supplant that it cannot even replace itself. In the beginning of the computer era, peripheral memories and mass storage were dominated by magnetic recording devices and systems. Today, some forty years later, magnetic recording is still the overwhelmingly dominant technology, and will continue to be the dominant technology well into the beginning of the next century. The pace at which this technology has changed is sometimes exasperatingly slow, and sometimes extraordinarily fast. For most implementations of the technology, however, the old concepts continue to exist almost *forever*, with small evolutionary changes along the way. New implementations usually do not replace the old ones, but simply expand the horizon and coexist with the old ones. Because the technology is so deeply and widely embedded, it has acquired enormous inertia to change, and a broadband frontal attack by a new technology is destined to fail. Evolution succeeds, bu revolution does not!

Early tapes and disks utilized primarily particulate γ-Fe_2O_3 (gamma ferric oxide) media. Forty years later, γ-Fe_2O_3 is still widely used in low and high performance applications, including the top-of-the-line IBM 3390 large disk drives, in spite of the tremendous progress in more advanced particulate media (chromium dioxide, Co-modified γ-Fe_2O_3 , metal particles, [illegible] have not been replaced by the newer magnetic recording techniques, what are the chances that a new and totally different technology is going to replace them?

After graduating from college, my first industrial assignment was to develop plated tape - a thin metallic film of Co-P deposited electrolessly onto mylar substrates. It was argued that the old γ-Fe_2O_3 technology was nearing its end, and thin films with their higher coercivity and magnetization would soon replace it. It did not happen then, and it did not happen for the next 25 years. Today we see some metal-evaporated (ME) tapes, but they have inferior wear and corrosion properties. The head-media interface is a very difficult problem for a thin metallic film rubbing against the head, and particulate media will continue to dominate tape technology for the foreseeable future. Evolution works, but revolution does not!

A perennial dream of peripheral storage architects has been to eliminate the famous memory access gap between the sub-microsecond access of main memories and the tens of milliseconds access of disk files. If a technology could be found to close the access gap, it would also eliminate the electromechanical devices inherent in disk and tape systems which tend to be bulkier, power inefficient, and less reliable than totally electronic systems. Of course, cost must always be a primary consideration. Per-bit costs of main memories are relatively high because of the discreteness of the bits, and of the wired access paths to the bits. On the other hand, magnetic recording systems offer low per-bit costs because the storage media do not require bit discreteness, and many millions or billions of bits share a common write/read sensor by moving the bits to the sensor, or the sensor to the bits, or a combination of both. The non-discreteness of the storage media and the sharing of a large number of bits by a single sensor, seem to be prerequisites in order to achieve low per-bit costs. Essentially inertialess beams of photons or electrons can be used to electronically address large numbers of bits in

otherwise homogeneous media. The interaction of the beam with the media is used to write and read the bits. Absolute positioning of the beam is not necessary, but repeatability is. By the middle 60's the newly developed laser had caught the imagination of the memory architects, who believed that beam-addressable memories would soon replace the "old" electromechanical magnetic recording storage devices. Everybody climbed on the laser bandwagon, and basic work on magnetic recording stopped. Why spend any effort and money on the old horse if it was to die soon? How wrong can one be? Laser beam memories faced a host of problems ranging from deflection to storage materials, and did not succeed - at least not at that time.

Some of the problems and limitations facing laser-beam memories derive from fundamental diffraction limits and from the limited types of interactions of photons with materials. Electron beams are much less limited by these constraints. They have no practical diffraction limits, they are very easy to deflect - in fact, they are too easy to deflect, which becomes a liability - and they interact strongly with all materials (ferromagnetic, ferroelectric, semiconductor, etc). The main drawbacks are that they require a vacuum, and they lack electron emitter sources that provide high current density, mono-energetic, collimated beams. Consequently, the depth of focus and the depth of field are limited by aberrations. To circumvent these problems, it is necessary to employ very precise and expensive focusing and deflection systems, which raise the entry price and the physical size of economically feasible memory modules. Therefore, even though the basic reasons were quite different than those facing the laser, several valiant attempts to develop electron-beam addressable memories in the late 60's and through the 70's failed, much like the laser-beam addressable memories had failed before them. The great memory-access gap was still wide open, and in fact getting wider, as the integrated semiconductor memories began to replace the ferrite core and to dominate main memory technology.

The attacks of the beam technologies on fortress ***Magnetic Recording*** were fueled to a large extent by the hope of achieving electronic access at low cost by *bringing the sensor to the bits* and sharing its relatively high cost among millions or billions of bits. But what about reversing the strategy, and bringing the bits to the sensor electronically? This approach gave rise to the *bucket brigade* technologies of magnetic bubbles and charge-coupled semiconductor devices (CCD's). These approaches would not actually eliminate the access gap, but they would shorten it significantly, while eliminating electromechanical systems. In spite of great investment and effort, bubbles and CCD's attained very limited success, and the fortress ***Magnetic Recording*** was still intact and looming more unassailable than ever before. The challenges presented by the new technologies and general market demands had, in fact, contributed to strengthen the fortifications and to raise the walls of the fortress, thus rendering it more impregnable. The lessons for new technologies trying to gain market in an area dominated by a firmly established and broadly based technology are:

(i) Frontal attacks across the entire market area are prone to failure, while selective attacks in specific sectors may have a better chance of success.

(ii) It takes a very long time to develop a new technology, particularly if it requires the synthesis of new materials.

(iii) Do not be so absorbed in developing the components of a new technology that you forget to consider their integration and how they interface in a total system.

(iv) A lonely technology has a much lower chance of success than a technology with broad industrial interest.

(v) Never underestimate the opposition by assuming it will stand still during the time you are developing your new technology and thereafter.

There is little doubt that magnetic recording is a mature technology. Consequently, we might expect that making large strides and fast progress would be more difficult. In fact, however, the strides being made today are bigger and bolder than they have ever been in the past:

Very thin film media
Extremely high coercivities
Multilayer film and particulate media
Magneto-resistive heads
1-, 2- and perhaps 10-Gbit/in^2 areal densities
Contact recording on rigid disks
Superfine metal and oxide particulate media
Perpendicular and quasi-perpendicular recording systems

Yes, the technology is mature on account of its longevity, and the breadth and depth of its accomplishments. But it is not getting old, slowing down, or about to disappear or be replaced by a new technology any time soon. It is more vibrant and more vigorous today than ever before in the last forty years that I remember. We have much greater rate of advancement, more new products, new developments, new expectations, and more things happening more quickly today than ever before. And, all along, the old products coexist with the new and hardly anything gets replaced. But there seems to be a distinct evolutionary trend in magnetic recording media

- from low coercivity to high coercivity
- from particulate to thin film
- from oxide to metal
- from thick to thin

and a corresponding rapid change in heads from ferrite and MIG to thin film and MR. This evolution, in the case of the media, has brought about a whole set of problems relating to noise and corrosion which are inherent to the metals. It would seem that, ultimately, the oxides offer more advantages compared with the metals (easier coercivity and anisotropy control, immunity to corrosion, and low noise), and I would predict that the oxides will dominate the future media. On the other hand, longitudinal recording, which requires ever-increasing coercivity and decreasing media thickness in order to achieve optimization, will gradually be replaced by perpendicular recording, which does not require extreme optimization of the magnetic parameters and of the thickness of the magnetic media.

The longevity, the dominance, and the vigor of magnetic recording, simply demonstrate the extraordinary power of the technology, which has sufficient base and momentum to carry it well into the next century. This is not to say that other technologies, such as magneto-optics and semiconductors, will not have an impact. But that impact will be felt primarily in certain areas and will not be an across-the-board displacement. Floppy disks, for example, may be impacted by magneto-optics, and very small rigid disk systems may be replaced by flash memories, but regular rigid disk and tape will probably not be affected in the foreseeable future. In the future (one to two decades), it is probable that some new technology will emerge which will challenge the main stream of magnetic recording. My opinion is that such a technology will have to be electronic (not electromechanical), and the storage media will be three-dimensional as compared to the two-dimensional technologies currently in existence.

Measurements Over Distributed High Performance Computing And Storage Systems

Elizabeth Williams
Supercomputing Research Center
17100 Science Drive
Bowie, Maryland 20715-4300

Tom Myers
Department of Defense
9800 Savage Road
Ft. Meade, Maryland 20755-6000

1.0 Introduction

The rapid pace of technological change and the move toward "open systems" is making the process of acquiring systems much more complex. Traditionally, functional and performance requirements have been carefully described for systems to be acquired and the systems usually have come from a single vendor. The process worked as long as the requirements remained nearly static and systems changed slowly over their life time. There generally has been no need for a requirement to provide measurements and performance monitoring to see that requirements were met over the long term. Measurements that were available were often left over from development.

In the future the requirements for many systems are expected to change more quickly, and parts of the systems, acquired from multiple vendors, will evolve to meet those changing needs. There is a desire to ask for life-time measurements of systems in request for proposals (RFPs) when systems are being acquired. Thus, there is a need for measurements and performance monitoring as an integral part of the system to ensure that requirements are met over the long term after acceptance.

This paper gives a strawman proposal for a framework for presenting a common set of metrics for supercomputers, workstations, file servers, mass storage systems, and the networks that interconnect them. Production control and database systems are also included. Though other applications and third party software systems are not addressed, it is important to measure them as well.

The capability to integrate measurements from all these components from different vendors, and from the third party software systems has been recognized and there are efforts to standardize a framework to do this. The measurement activity falls into the domain of management standards. Standards work is ongoing for Open Systems Interconnection (OSI) systems management; AT&T, Digital and Hewlett-Packard are developing management systems based on this architecture even though it is not finished. Other efforts include the Storage System Management Sub-committee of the Mass Storage System Working Group and the UNIX International Performance Management

Working Group [1]. In addition, there are the Open Systems Foundation's Distributed Management Environment and the Object Management Group. A paper comparing the OSI systems management model and the Object Management Group model has been written [2]. Though most of the standards effort has been on the mechanisms for gathering and reporting measurements, we expect to cooperate with these standards making efforts. The work reported here is ongoing.

The IBM world has had a capability for measurement for various IBM systems since the 1970's and different vendors have been able to develop tools for analyzing and viewing these measurements. Since IBM was the only vendor, the user groups were able to lobby IBM for the kinds of measurements needed. However, in the UNIX world of multiple vendors, a common set of measurements will not be as easy to get.

In this paper we distinguish between metric and measurement. A **measurement** is a quantity that is directly measured while a **metric** is a quantity that can be derived from a set of measurements. Our focus is on using low level vendor specific measurements to support a set of higher level metrics that are common across a variety of vendors. The set of measurements to support the common metrics should in general be the minimum that is provided. Most systems should also make available measurements of unique aspects of the system that are not covered by the common set. For example, measurements on vectorization and hit ratios for memory hierarchies may not be in the common set of metrics but such measurements are desired.

2.0 Uses for Measurements

Measurements of systems are, of course, useful in many other ways than just to support system acquisition. They can be used to support day-to-day operations, management decisions and planning, and performance monitoring. The following are seven types of uses we have identified:

(1) distributed computing system scheduling,

(2) fire-fighting - solve immediate problems to provide acceptable response time and resource allocation to all processes,

(3) tuning systems for current workloads,

(4) capacity planning,

(5) allocating resources,

(6) looking for trends and characterizing workloads,

(7) verifying system strategies are working or assumptions about workloads are valid, e.g. locality of reference,

(8) validating accounting reports.

In analyzing how measurements are used, the following three points are very important. (1) For fire-fighting and tuning, a systems administrator must be able to **link** a particular "event" to a set of user commands.The systems administrator should be able to know when a resource is responding slowly and which process is causing the problem. We stress that it is important to be able to link particular events of interest back to user commands though we know that it is sometimes difficult. (2) Process as well as system-wide measurements are needed. (3) Accurate time stamps or

other timing information is necessary so that various independent measurements can be correlated with each other as a system is observed over time.

3.0 Measurement Collection Techniques

It is also understood that taking measurements and collecting them cause overhead and may in extreme cases affect the performance of the systems measured; this is not specifically addressed in this paper. However, data can be collected at various levels of detail depending on how much overhead is involved. The most complete level of measurement is a **log** or **trace** of each transaction or event. The next level of measurement is a set of counters that produce a histogram, which is an approximation to the distribution, of the metric of interest. The least detailed level of measurement is a simple counter from which the average, variance, maximum and minimum of the metric of interest can be derived. The level of measurement for any component depends on the overhead associated with the workload. When possible, the ability to selectively choose a different measurement level allows users of a system to manage how much overhead is given to measurement activities. Another way of managing the overhead associated with measuring a system is to sample a measurement at some interval that is frequent enough to observe interesting behavior but with reduced overhead. The sampling rate should be adjustable.

For measurements to be useful, they must be well documented It must be clear exactly what is being measured. The documentation should specify how much overhead is involved, what technique is being used to generate the measurement, and if there are user selectable parameters such as a sample rate or an enable/disable switch. Information about how a system is configured must either be statically defined or recorded along with a set of measurements.

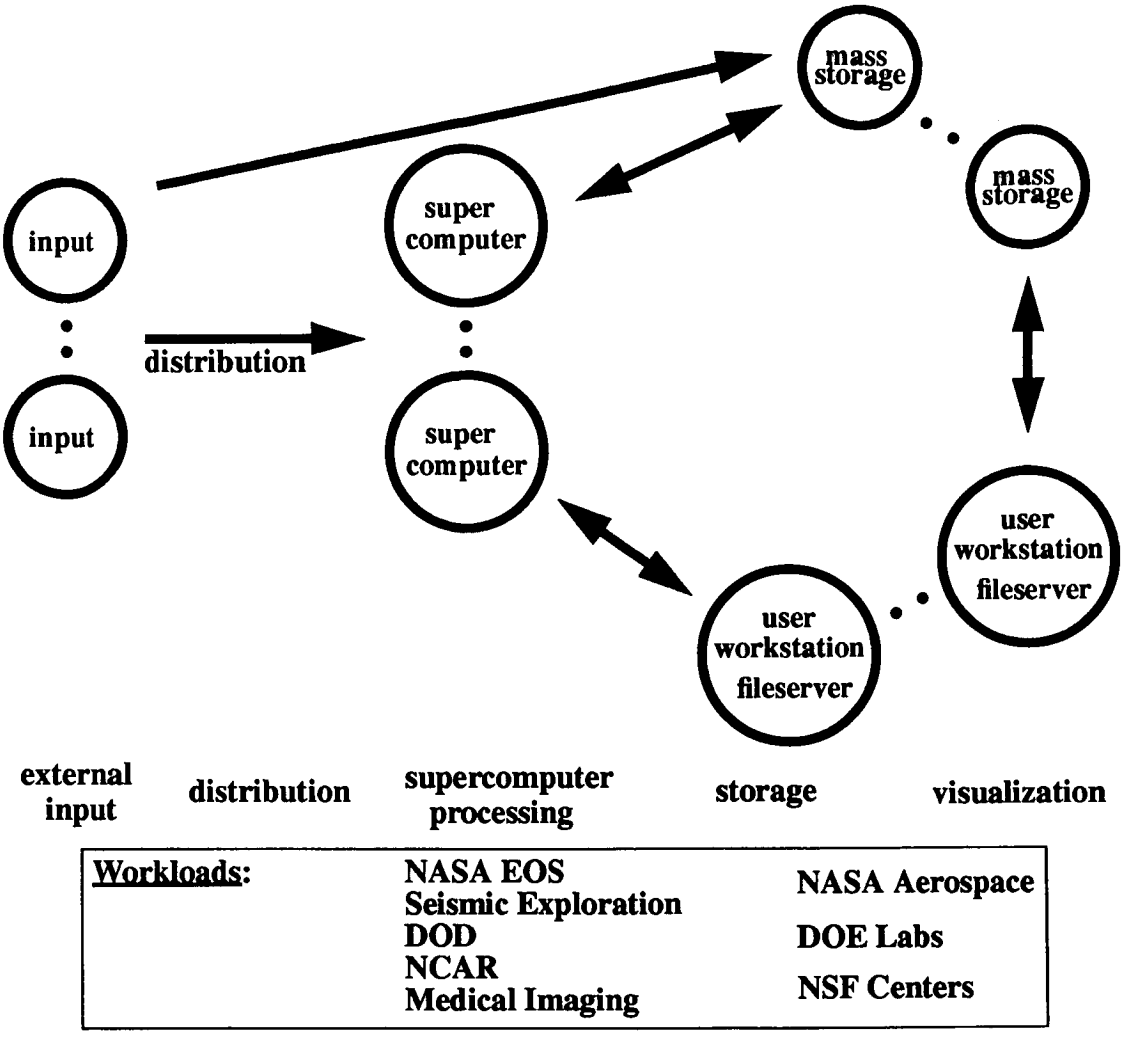


Figure 1: Model of Network Computing System

4.0 Model of Distributed High Performance Computing Systems

In Figure 1 we present a model of a distributed high performance computing system. The model identifies the five highest level functions of **external input** sources to indicate the collection of data for processing in the system, **distribution** for the network among components, **supercomputer processing** for high performance computing, **storage** for distributed mass storage, and **visualization** for user support processing. The distributed characteristics of this model are not depicted specifically but one can think of NASA's EOS system as the basis for this model. The other high performance computing systems listed at the bottom of the figure will all have similar models.

The five model functions are made up of various hardware and software system components. The hardware system components include supercomputers, mainframes, workstations, mass storage devices, file servers, networks, input machines and other network devices such as disk arrays. The

software system components include operating systems (**OS**) (includes file system and protocols), mass storage systems (**MSS**), database systems (**DBMS**), production control systems, third party software and user applications. Below the system component level are lower level building blocks to measure. These are the hardware building blocks such as CPU, memory, memory interconnection, disk, tape, terminal I/O, recorder/drive, robotics box, channel/controller, network interface, router and external I/O. The software building blocks are dependent on the particular system software component. For an operating system there are process management (scheduler/queues, context switches), I/O system (buffers, cache, queues), memory management (allocation, swapping, queues, paging, caches), file system, protocols, interprocess communications and other operating system services. For a mass storage system there is each module in the Mass Storage Reference Model (**MSRM**). For an application there are user defined modules, operating system components and various hardware building blocks used by the application. For a database there are indexes, tables, stored procedures, logs, locks, transactions and users. The software building blocks are not yet completely identified.

Figure 2 illustrates this three level hierarchy of metrics. The abstract metrics at the base of the pyramid are a list of generic metrics that are used at all three levels. The eye represents the need to have comprehensive and uniform observations at all levels.

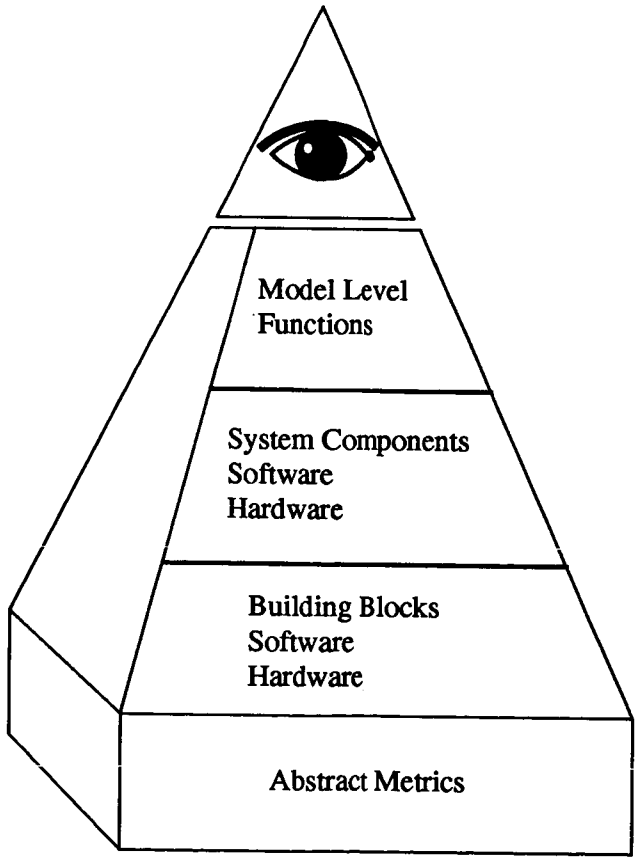


Figure 2. Hierarchical Levels

In Figure 3, the levels are expanded to show how higher level objects are composed of lower level objects. Three model functions are connected to system components and some of the system components are connected to lists of building blocks. Space does not permit a full expansion of all functions and components in this figure.

The common metrics for objects at the lowest level are derived from vendor measurements. And the metrics for higher level objects are derived from combinations of metrics for lower level objects. Thus we expect that the derivation of the lowest level metrics will be vendor specific but that higher level metrics will be vendor independent. We have not dealt with the issue of how to attach derivations to the metrics. We have not evaluated the difficulty for vendors to implement measurements to support the framework, nor have we tried to gauge the relative importance of the various metrics.

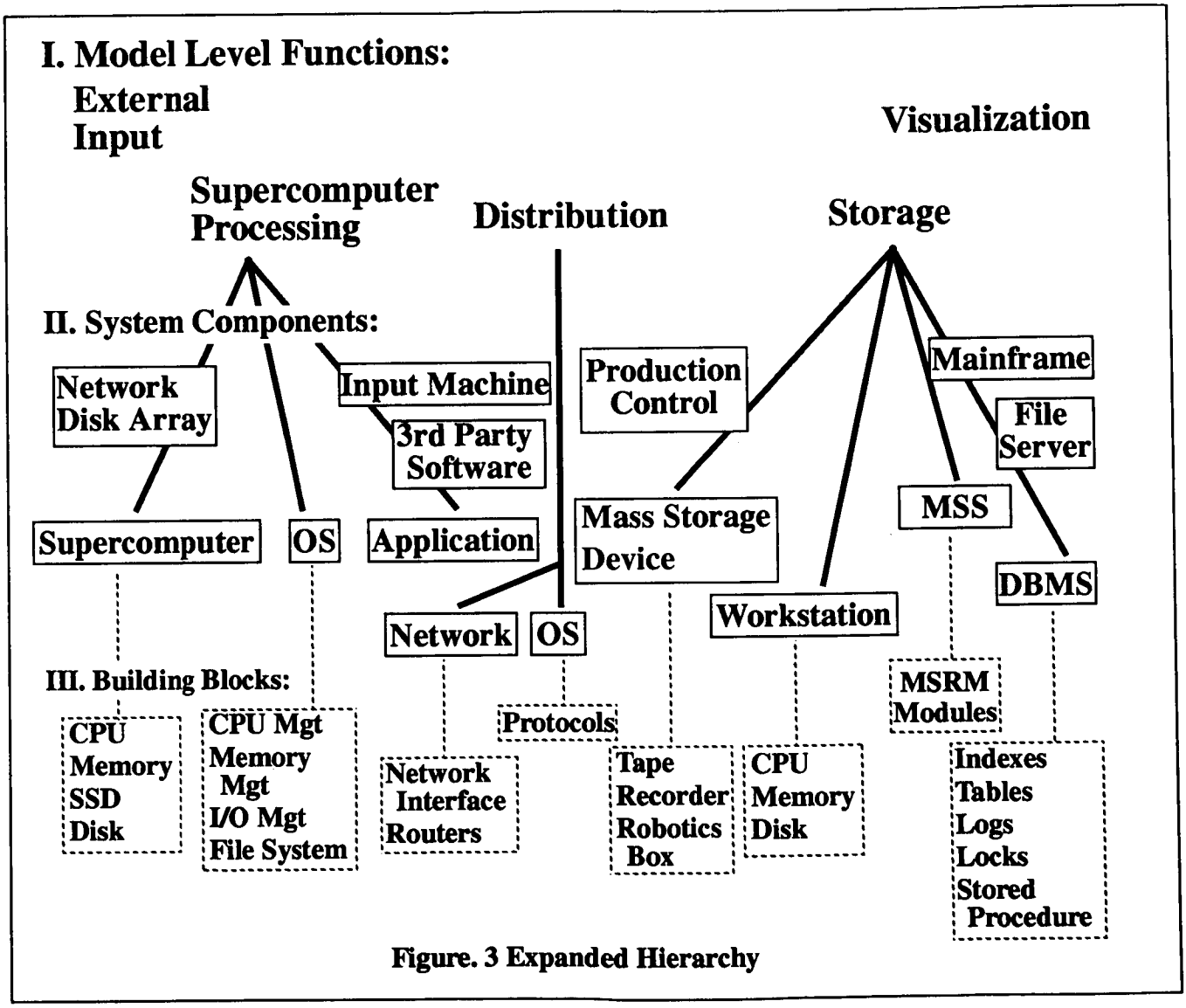


Figure. 3 Expanded Hierarchy

The use of common metrics is intended to provide vendors and other system developers a framework that can be used to design measurements as an integral part of the system that they deliver.

Too often measurements are used only to verify that a system is operating correctly and are insufficient for understanding the performance of the system especially when it is interconnected as a component of a larger system.

5.0 Abstract Metrics

The following list of abstract metrics are used to observe any **object** in the hierarchy by specifying an instance of the metric for the object:

1 Utilization, Capacity, Idle

2 Throughput

3 Response Time[1], Delay, Expansion Factor[2]

4 Waiting Time

5 Service Time

   a. Bitfile Size, Packet Size, Computation Requirement

   b. Speed of Device

6 Queue Length

7 Number of Jobs, Bitfiles, Packets

8 Routing, Branching Probabilities for Jobs Paths, Reuse, Age

9 Hit Ratios, Effectiveness of Strategies
(data migration, locality of reference)

10 Error Rates

All of these metrics are commonly used except for 8 and 9. Branching probabilities are useful for modeling systems.

At the bottom of the hierarchy the specific metric for each object is given in terms of characteristics of the object, such as mips and mflops for CPU throughput metrics. In addition at higher levels, users will want to specify metrics in terms of the workload of the system, e.g. satellite images processed per second through all model level functions for the NASA EOS.

6.0 Metric Tables

The following pages contain tables of metrics for the objects within the hierarchy. The left hand column in each table has the list of abstract metrics. The other columns have instances of the corresponding metric for the object at the top of the column. The tables are generally sparse since this

---

1. Response Time = Service Time + Wait Time + Other Time

2. Expansion Factor: wall clock time in shared system / wall clock time in dedicated system, which often can be approximated by wall clock time in shared system / CPU time

work is still ongoing and we invite help in completing the tables, adding more objects to the hierarchy and adding more abstract metrics.

abstract metric	Processing and Storage	Processing and Input	Input and Storage	Input, Processing, Storage
utilization, capacity, idle				
throughput	Mops per bit stored Mops per bitfile stored	Mops per bit input	Input bits per bit stored	Bitfiles processed through all functions per second
response time, delay, expansion factor				Response time through all functional components
waiting time				
service time: job size, device speed				
queue length				
number of jobs				
routing paths, reuse, age, branching probabilities				Bitfile routes through all functional components
hit ratios, effectiveness of strategies				
error rates				

Table 1: Model Level - Across Functions

Table 1 has metrics for the overall system where the objects being observed are combinations of model level functions. At this level, the metric instances are suggestions since they will depend on what the system does and will be defined by the users of the system.

Tables 2, 3 and 4 have the metrics for storage and some of its lower level components and building blocks. Tables 5 through 9 have the metrics for supercomputer processing and some of its lower level objects. Table 10 has the metrics for distribution (networks) and some of its lower level objects

In conclusion, we have presented a strawman proposal for a framework for presenting a common set of metrics across many systems and we have listed some of the metrics. This work is ongoing and we invite participation from users, vendors and system developers.

abstract metric	storage	mass storage device	mass storage reference model
utilization, capacity, idle	% space used # B[a], # O, # M total by class[b] or storage device	% space used % fragmentation	# bitfiles by class bitfiles/media bits/bitfile by class
throughput	{B O M}/sec access[c] by class or storage device	bits/sec accessed media/sec accessed	bitfiles/sec accessed
response time, delay, expansion factor	{B O M} response time by class or storage device or overall	{B M} response time by class	Bitfile response time by class
waiting time	*[d]		*
service time: job size, device speed	*		*
queue length			length at various model modules
number of jobs			
routing paths, reuse, age, branching probabilities	# accesses vs. storage device {B O M} vs. age vs. # accesses	#media vs.#accesses #media vs. # age #media vs. age vs. # accesses	# bitfile vs. # accesses # bitfile vs. # age # bitfile vs. age vs. # accesses
hit ratios, effectiveness of strategies			migration policy metric hit ratios
error rates	BER overall, by device failure by device	BER, failures	

Table 2: Storage - System ComponentS

a. B= Bits; O = Bitfiles; M = Media
b. class = {media type, bitfile size, access type, user, user process, user defined}
c. accesses = reads, writes, deletes
d. Asterisk implies that the metric is the obvious one in this context.

System Components not included in this configuration:

1. workstation or mainframe for controlling mass storage device
2. database for meta-data about the stored bitfiles

abstract metric	tape	recorder, tape drive	disk arm/ platters	robot
utilization, capacity, idle	% space used/tape % free tapes	% time reading % time writing % time scanning % idle	% time reading % time writing % time seeking % free space or fragmented (int/ext) for platters	% time in use
throughput		bits read/sec bits write/sec mounts/sec	bits read/sec bits write/sec seeks/sec	# requests/sec
response time[a], delay, expansion factor		includes (positioning) start, stop, scan, read/write delays	includes read/ write, seek, rotation delays	includes start, stop (positioning) delays
waiting time				
service time: job size, device speed				
queue length				
number of jobs				
routing paths, reuse, age, branching probabilities	# tapes vs. # accesses[b] # tapes vs. age # tapes vs. age vs. # accesses			
hit ratios, effectiveness of strategies			arm movement distance/seek	distance/request
error rates	BER each tape BER for all tapes	failures/time int.	failures/int. BER for platters	robot failures/int.

Table 3: Storage - Building Blocks - Hardware

a. response time = service time + waiting time + other factors associated with using resource
b. accesses = reads, writes, deletes

abstract metric	physical volume repository	bit file mover	storage server	bit file server
utilization, capacity, idle	% time in use			
throughput	bitfiles/sec accessed	bitfiles/sec accessed	requests/sec	requests/sec
response time, delay, expansion factor	*[a]	*	*	*
waiting time	*	*	*	*
service time: job size, device speed	*	*	*	*
queue length	*	*	*	*
number of jobs		*	*	*
routing paths, reuse, age, branching probabilities			# {B O} vs. age vs. size vs. accesses	
hit ratios, effectiveness of strategies			migration/caching policy	
error rates				

Table 4: Storage - Building Blocks - Software

a. Asterisk implies that the metric is the obvious one in this context.

abstract metric	Supercomputer Processing	Supercomputer	Operating System	Application CPU, mem, IO
utilization, capacity, idle	% to users % to system % to idle	% to users % to system % to idle	% to system % holding on locks	% to application % to system vectorization speedup
throughput		mops, mips mflops	processes/ sec system calls/sec interrupts/sec - all by class	mflops particles/sec
response time, delay, expansion factor			response time for all processes expansion factor for all processes	response time for application
waiting time			waiting time for all processes	
service time: job size, device speed			CPU burst time vs. memory size	CPU time memory size logical reads, writes
queue length				
number of jobs				
routing paths, reuse, age, branching probabilities			process path probabilities for I/O devices	
hit ratios, effectiveness of strategies				page hit ratio swaps system calls
error rates				

Table 5: Supercomputer Processing - System Components

abstract metric	CPU	Memory	SSD[a]	disk arm/ platters
utilization, capacity, idle	% time issuing inst % time holding issue % time vect or para % vector {ops,inst} vector length	% time issuing read or write (% free space or fragmented)	% time issuing read or write (% free space or fragmented)	% time reading % time writing % time seeking % free space or fragmented (int/ ext) for platters
throughput	ops/inst mops, mips mflops	{Bytes, Words} read/s, write/s by type	reads/sec writes/sec	bits read/sec bits write/sec seeks/sec
response time, delay, expansion factor				
waiting time	% time waiting on functional units	waiting time/ref hold issue/ref contention/ref		
service time: job size, device speed	hardware specified	hardware specified	hardware specified	includes read/ write, seek, rotation delays
queue length				
number of jobs				
routing paths, reuse, age, branching probabilities	instruction mix: % {ops, inst} by instr class			
hit ratios, effectiveness of strategies	instr cache hit ratio memory cache hit ratio	page hit ratio	device cache hit ratio	arm movement distance/seek
error rates				failures/interval BER for platters

Table 6: Supercomputer - Building Blocks - Hardware

a. SSD = solid state device

abstract metric	Channel/Controller	Terminal I/O
utilization, capacity, idle	% time busy	
throughput	bits/sec by device[a] channel ops/sec	characters/sec
response time, delay, expansion factor		
waiting time		
service time: job size, device speed		
queue length		
number of jobs		
routing paths, reuse, age, branching probabilities	bits vs. device	
hit ratios, effectiveness of strategies		
error rates		

Table 7: Supercomputer - Building Blocks - Hardware

a. device = {SSD, disk}

abstract metric	CPU Management	Memory Management	I/O System
utilization, capacity, idle	% time to user % time to idle % time to system	% space used % space fragmented	% buffer space used
throughput	context switches/sec by class processes/sec	allocations per sec swaps per second by memory size pages per sec	logical & physical read/ write per sec by bits, device
response time, delay, expansion factor			
waiting time	WT	WT	WT
service time: job size, device speed	CPU burst time per process	memory size by process memory residency time	time for service by logical, physical I/O
queue length	QL of CPU queue(s)	QL of Memory queue(s)	QL of device queues
number of jobs		# jobs in memory	
routing paths, reuse, age, branching probabilities			
hit ratios, effectiveness of strategies			I/O buffer hit ratio by read, write
error rates			

Table 8: Operating System - Building Blocks - Software

abstract metric	File System	Interprocess Communication	Other OS Services
utilization, capacity, idle	% used on each I/O device		
throughput	operations/s by class		
response time, delay, expansion factor			
waiting time			
service time: job size, device speed			
queue length			
number of jobs			
routing paths, reuse, age, branching probabilities			
hit ratios, effective- ness of strategies			
error rates			

Table 9: Operating System - Building Blocks - Software

abstract metric	Distribution	Networks	Operating System: Protocols	Routers/ Network Interfaces
utilization, capacity, idle		% time used bits/packet	bits/object packets/object	
throughput	bits/s bits/s vs. path objects/s objects/s vs. path by class[a]	bits/s packets/s by class	bits/s pkt/s objects/s by class	
response time, delay, expansion factor	by class and object size	by class and object size	by class and object size	
waiting time	by class and object size	by class and object size	by class and object size	
service time: job size, device speed	by class and object size	by class and object size	by class and object size	
queue length		send/receive queues	send/receive queues	
number of jobs				
routing paths, reuse, age, branching probabilities	relative use of paths			
hit ratios, effectiveness of strategies		collisions/packet		
error rates	BER, failures	retrans/sec	timeouts failures	

Table 10: Distribution - System Components

a. class = {protocol used, path, user, process, send/receive}

7.0 References

[1] Leon Traister and Terry Flynn, "A Measurement Architecture for Unix-Based Systems", CMG Transactions, Winter, 1991, pp. 69-77.

[2] Peggy Quinn and George Preoteasa, "Reconciling Object Models for Systems and Network Management", Technical Report, UNIX System Laboratories, Inc.

Measurements Over High Performance Network Computing And Storage Systems

Elizabeth Williams
Supercomputing Research Center
17100 Science Drive
Bowie, Maryland 20715-4300

ew@super.org
(301) 805-7468

Tom Myers
Department of Defense
9800 Savage Road
Ft. Meade, Maryland 20755-6000

ctmyers@super.org
(301) 688-6507

Why such a Paper?

To include in RFPs (Request for Proposal)

- Requirements are carefully described
- Want to ensure that requirements are met over the system lifetime; not just at acceptance
- Measurements should be a planned integral part of system rather than added on later
- For Unix based systems requirements will apply to multiple vendors

We want multiple vendors to provide measurements that support a common set of metrics

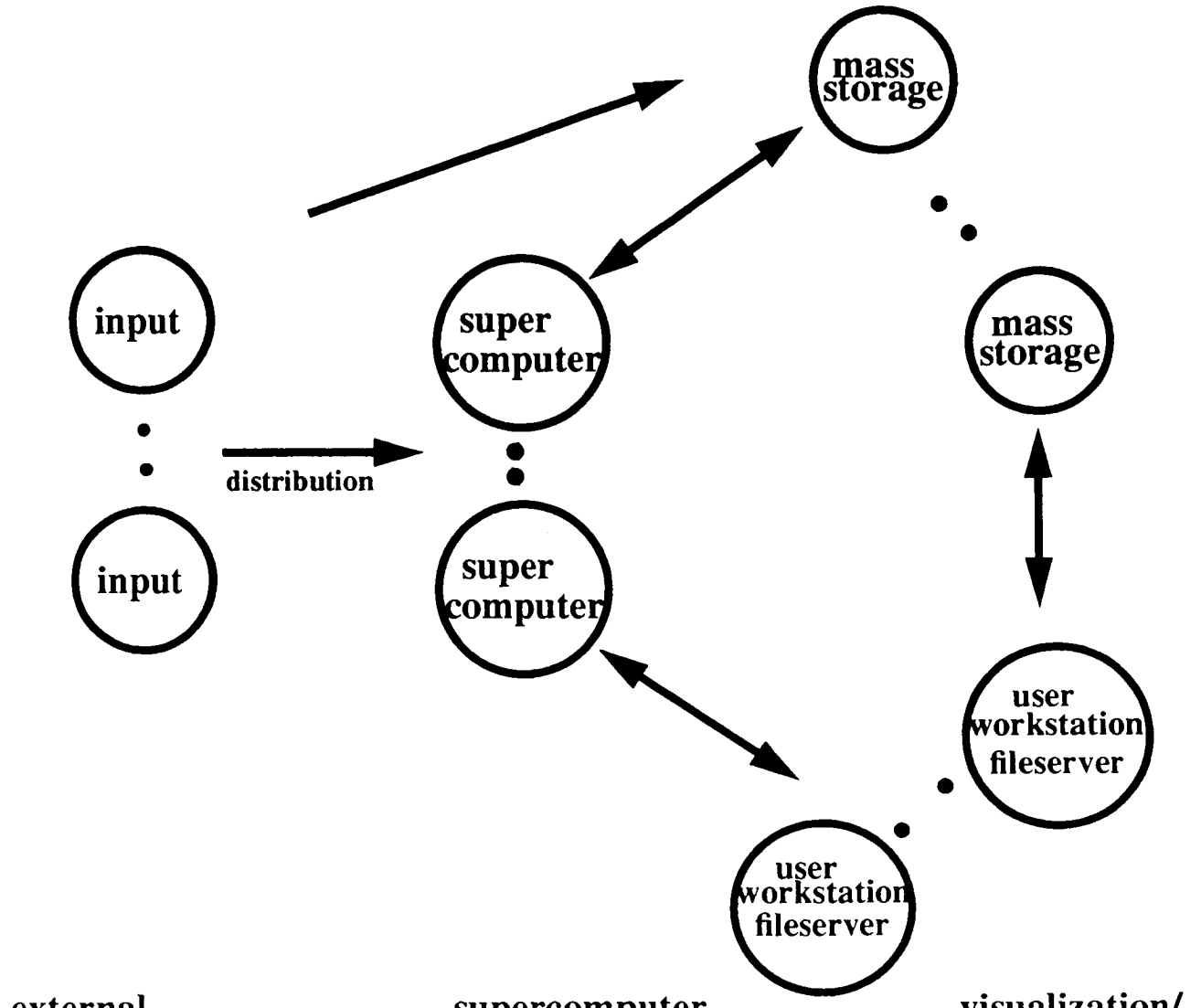
Outline

- Uses for Metrics
- Model of Network Computing System
- Hierarchical Components
- Collection Techniques
- Metrics for Supercomputers
- Metrics for Storage Systems
- Metrics for Distribution (Networks)

Uses for Metrics

- Fire-fighting - solve immediate problems to provide acceptable response time and resource allocation to all processes
- Scheduling distributed computing applications
- Validating accounting reports
- Allocating resources
- Capacity planning
- Tuning systems for current workloads
- Looking for trends and characterizing workloads
- Verifying system strategies are working or assumptions about workloads are valid,
  e.g. locality of reference

Model of Network Computing System



NASA EOS

Seismic Exploration

DOD

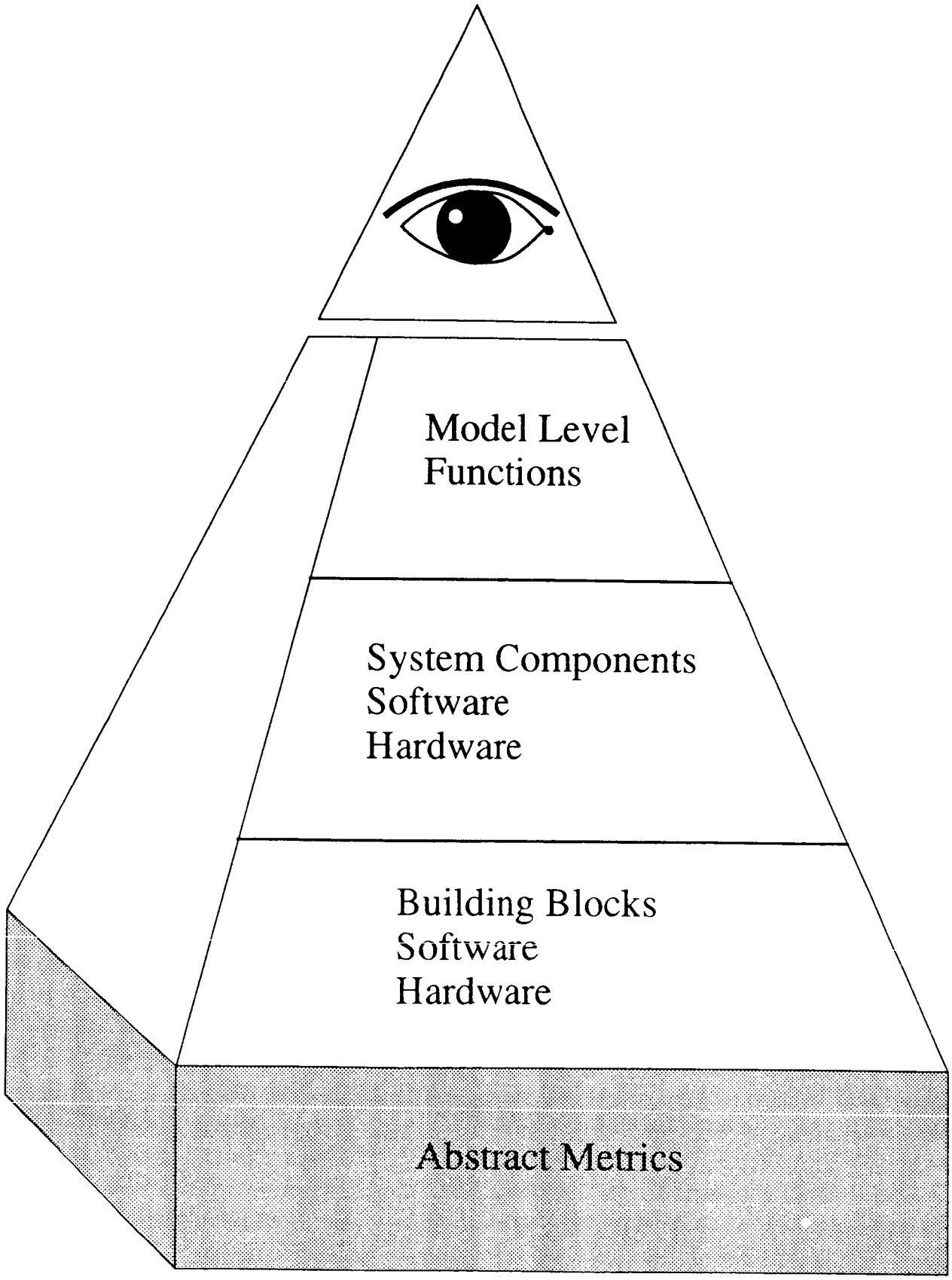
NCAR

Medical Imaging

NASA Aerospace

DOE Labs

NSF Centers


Model Level
Functions
System Components
Software
Hardware
Building Blocks
Software
Hardware
Abstract Metrics

Hardware System Components

- Input Machine
- Supercomputer
- Mainframe
- Workstation
- File Server
- Mass Storage Device
- Other Network Devices (disk array)
- Network

Software System Components

- Operating System (includes File System, Protocols)
- Mass Storage System
- Database System
- Production Control
- 3rd party Software
- Application

Building Blocks - Hardware/Physical

(combined list for all hardware system components)

- CPU(s)
- Memory
- Bus
- SSD
- Disk (magnetic, optical)
- Tape (magnetic, optical)
- Terminal I/O
- Recorders, Drives
- Mass Storage Robotics
- Channel/Controller
- Network Interface
- Routers

Building Blocks - Software

Operating System

- Process Management (scheduler/queues, context switches)
- I/O System (buffers, cache, queues)
- Memory Management (allocation, swapping, queues)
- File System
- Protocols
- Interprocess Communications
- Other OS Services

Mass Storage System

- Each Module in Storage Reference Model

Application

- Defined by User

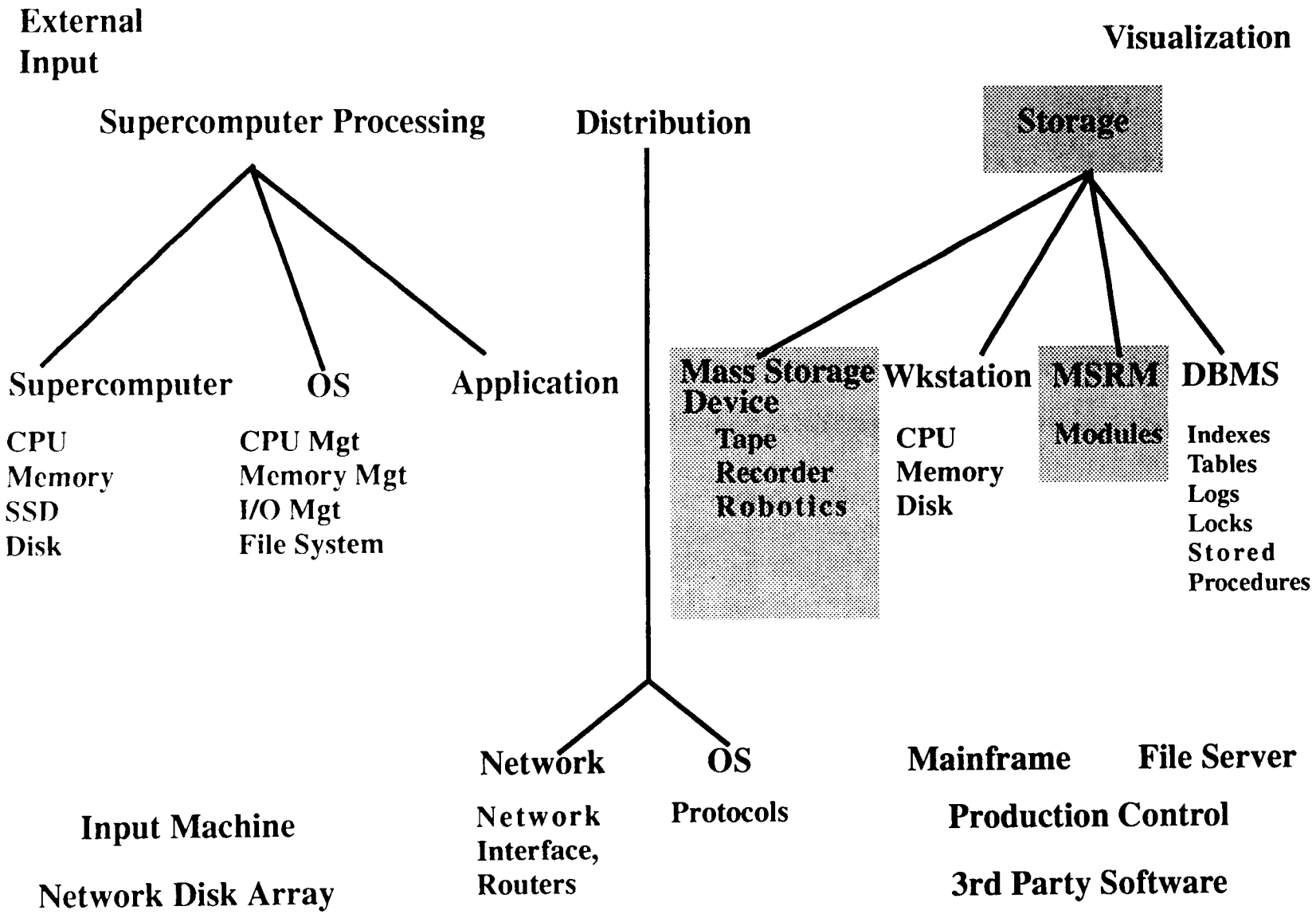
Building Blocks - Software

Database System

- Indexes
- Tables
- Stored Procedures
- Logs
- Locks
- Transactions
- Users

Production Control

- Batch
- Application Management


External
Input
Supercomputer Processing
Distribution
Visualization
Storage
Supercomputer
OS
Application
CPU
Memory
SSD
Disk
CPU Mgt
Memory Mgt
I/O Mgt
File System
Mass Storage
Device
Tape
Recorder
Robotics
Wkstation
CPU
Memory
Disk
MSRM
Modules
DBMS
Indexes
Tables
Logs
Locks
Stored
Procedures
Network
OS
Network
Interface,
Routers
Protocols
Mainframe
File Server
Production Control
3rd Party Software
Input Machine
Network Disk Array

Abstract Metrics

- Utilization, Capacity, Idle
- Throughput
- Response Time, Delay, Expansion Factor
- Waiting Time
- Service Time
- Object Size, Packet Size, Computation Requirement
- Speed of Device
- Queue Length
- Number of Jobs, Objects, Packets
- Routing, Branching Probabilities for Jobs Paths, Reuse, Age
- Hit Ratios, Effectiveness of Strategies (data migration, locality of reference)
- Error Rates

By Class as defined by customer

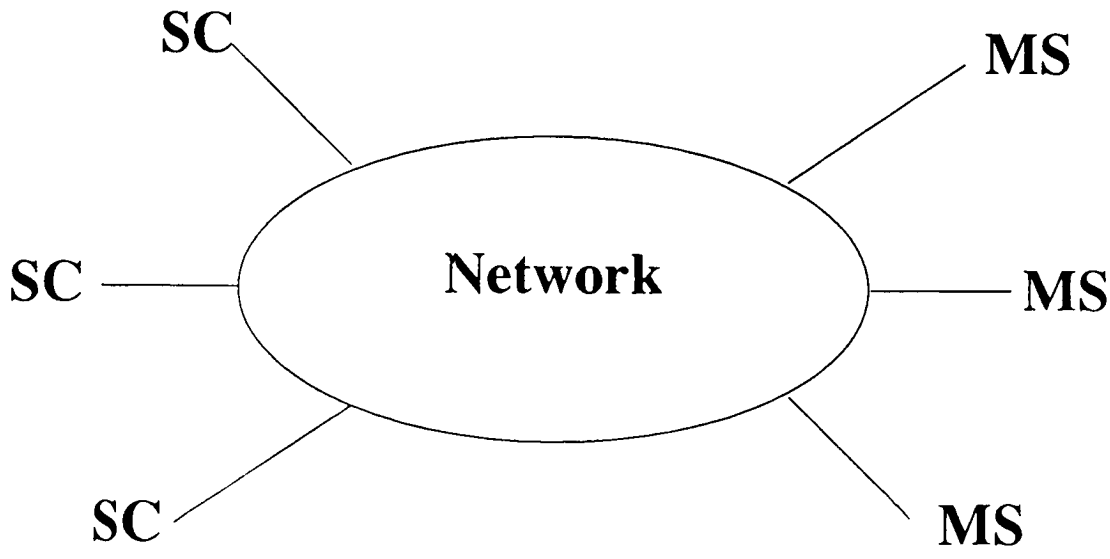
Expansion Factor: wall clock time / CPU time
Response Time = Service Time + Wait Time + Other Time

Relate abstract metrics to metric terminology for each component

An Example

{Supercomputer Processing, Storage, Distribution}

several supercomputers, several mass storage devices, and interconnecting network



What kind of measurements at different levels?

Higher Level Metrics:

- % free space on mass storage devices
- Mb/s sent over network
- Mflops per bit stored

Lower Level Metrics:

- % free space in specific robotic storage device (silo) within tapes and by empty slots
- Mb/s sent or received between pairs of systems
- Mflops per CPU; bits stored per machine

Can one always compute higher level metrics from measurements collected at lower levels?

Collection Techniques

Perturbation and Overhead to the system being measured: needs to be acceptable both in time and space

Traces and Counters

- Logs, Traces (include time stamp)
- Histograms (distribution), thresholds
- Average, Std. Dev., Max, Min (sustained/peak)

Sample Rate:
Log or count **every event** versus **sampling events**
Samples should be time stamped.

Time intervals for collection should be tunable by user.

Technique may be selectable by user.

Please document the metrics collected and collection techniques according to perturbation / overhead, trace / counter, every event / sample rate.

Performance of a Distributed Superscalar Storage Server

Arlan Finestead
University of Illinois, National Center for Supercomputing
605 East Springfield, Champaign, Il 61820
arlanf@ncsa.uius.edu

Nancy Yeager
National Center for Supercomputer Applications
152 Computer Applications Building
605 East Sprinfield, Champaign, IL 61820
nyeager@ncsa.uiuc.edu

Introduction

Traditionally, mass storage systems have been single centralized systems; however, a highly distributed mass storage server implemented on superscalar workstations may challenge the centralized model in terms of high file transfer rates and favorable price-performance characteristics. Additionally, a workstation based distributed mass storage server is scalable and may be hierarchically configured as a component of a larger more centralized mass storage system.

National Center for Supercomputing Applications offered a UniTree™ archival service to a select group of users for a trial period of time. The objectives of this trial period were to a) monitor distributed UniTree performance in a production environment under normal and high load conditions b) quantize archival transfer rates from supercomputer clients c) ascertain patterns of UniTree user access d) optimize system performance by tuning file migration from disk to tape.

The archive system architecture consisted of UniTree storage storage servers installed on an IBM RS/6000 Model 550 and an Amdahl model 5860. The UniTree archival software in conformance with the IEEE storage reference model supports a distributed architecture such that the disk operations and tape operations of the storage system may reside on physically separate hosts(see Appendix Figure I). The RS/6000 AIX machine which is fairly efficient at disk operations and protocol processing operations functioned as the Disk Server while the Amdahl UTS serviced tape operations.

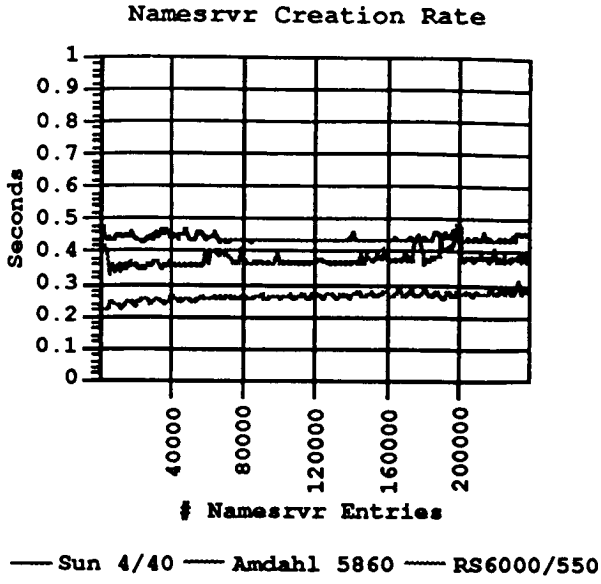
The mass storage serviced archive requests from a farm of loosely coupled IBM RS/6000 Model 550s running scalar computational chemistry codes such as Gaussian-90. Individual RS/6000's within the cluster are interconnected via ethernet; the UniTree Disk server RS/6000 is networked to the Amdahl tape server via FDDI and ethernet.

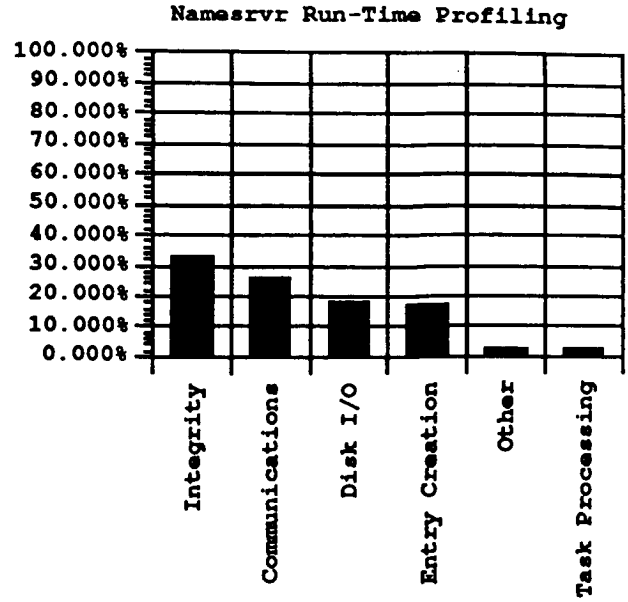
UniTree Performance Testing

Locally developed programs that interfaced directly with the various components of UniTree were used to ascertain the user-perceived performance of UniTree. The tests were varied to simulate a work load model (the load placed on a system by application users) and a system load model (the load according to system metrics such as CPU utilization, inter-server protocol processing, and network traffic). Each of the UniTree components were profiled to determine where potential bottlenecks might exist.

Name Server Performance

The UniTree Name Server daemon exhibited uniform, linear performance when directed to create 230,000 Name Server entries on a RS/6000 Model 550, on a Amdahl 5860, and on a SPARCstation IPC.





The average entry creation time on an RS/6000 Model 550 was .263 seconds, on an Amdahl 5860 was .377 seconds, and on a SPARCstation IPC was .442 seconds. Improvements in Name Server creation performance were realized when the testing program interfaced directly with the UniTree Name Server daemon, bypassing LibUnix altogether. Through optimized creates, entries could be created on the RS/6000 Model 550 in .07 seconds.

The UniTree Name Server was profiled to determine where the majority of execution time was being spent. The UniTree Name Server was categorized into six areas:

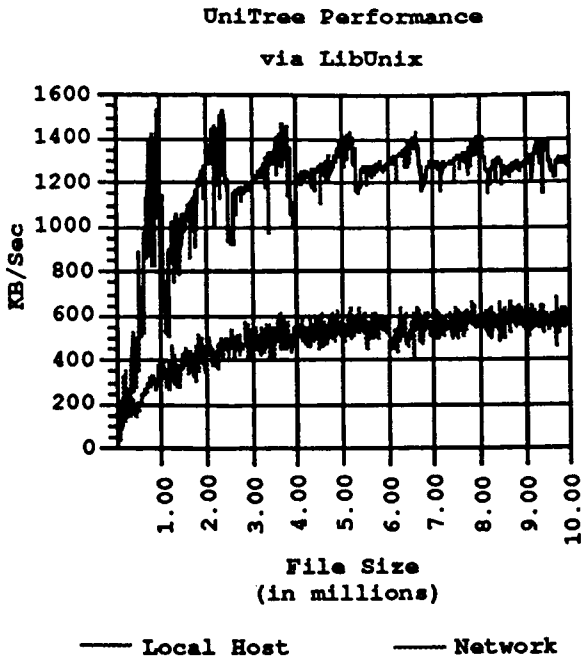
- Integrity - locking data structures, verifying Capabilities.
- Communications - sending, receiving messages via the UniTree APST communication mechanism.
- Disk I/O - performing actual disk operations such as reads and writes.
- Entry Creation - maintaining the Name Server btree structure.
- Task Processing - performing the UniTree task processing.
- Other - includes areas such as logging messages, opening configuration files.

The Name Server creation test program was used to gather the profiling data.

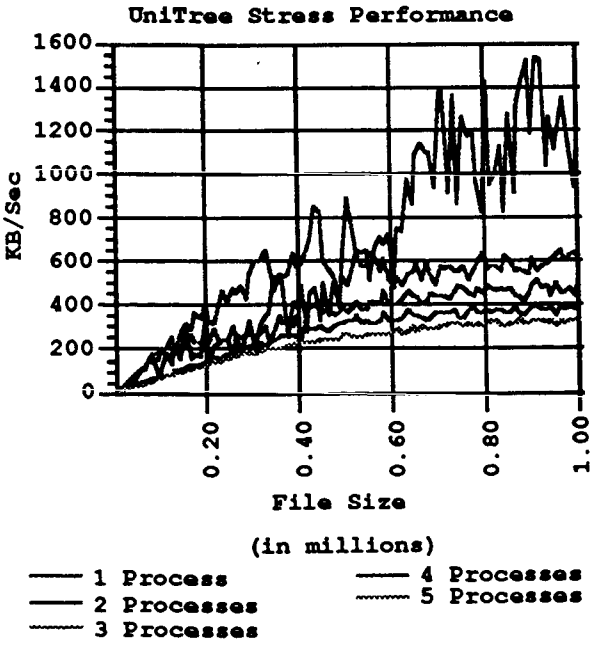
UniTree LibUnix Performance

A test program that interfaced with UniTree via the UniTree LibUnix library was used to determine the performance characteristics of the UniTree Name Server, Disk Server, and Disk Mover daemons. The test program generated increasingly large files in the UniTree archival system, recording the performance with each creation.

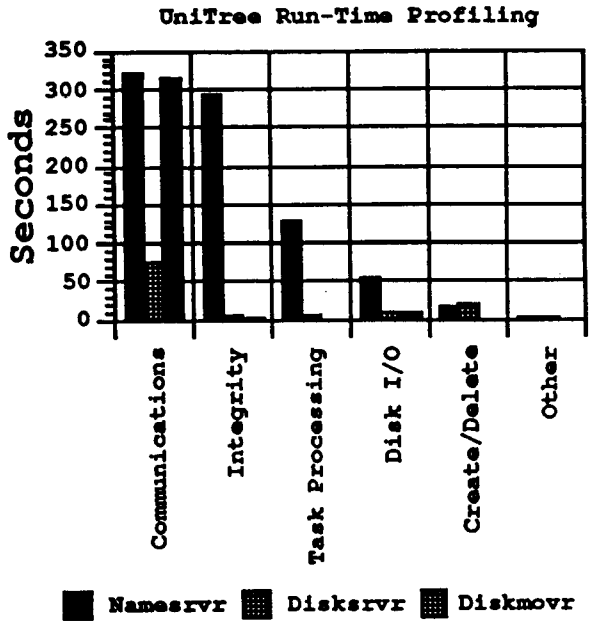
The UniTree performance of the Name Server, Disk Server, and Disk Mover daemons on the RS/6000 Model 550 showed performance at an average of 1311KB/sec when the test program was executed on the local host, and at an average of 594KB/sec when the test program executed on a remote host.



This test case was expanded further, and multiple processes were initiated on the UniTree local host to stress UniTree. Just the localhost was tested to eliminate the limitations of the network. There was a 46% drop in performance when the second process was added, and a 20% drop when each additional process was added.



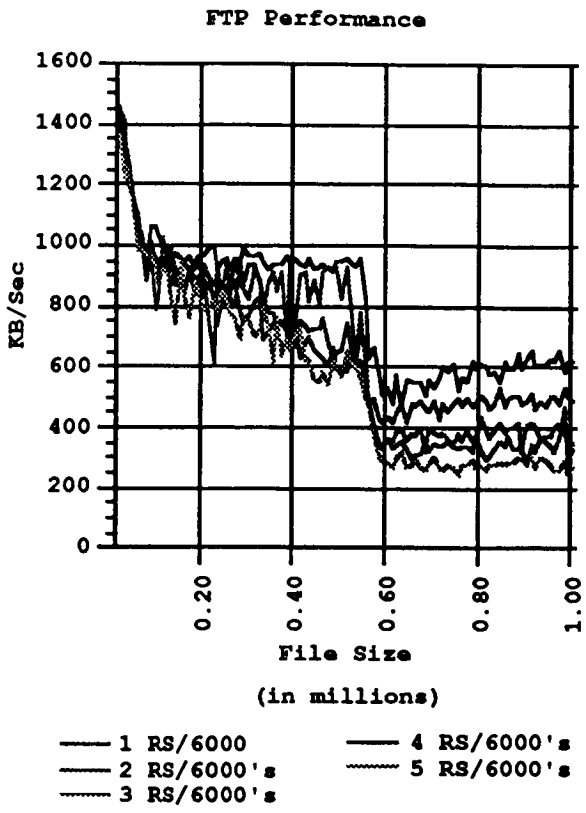
Using the above testing scenario with only one local process, the UniTree Name Server, Disk Server, and Disk Mover daemons were profiled. The daemons were categorized into the same six areas that the UniTree Name Server was categorized with the Entry Creation category broaden to include the functions the Disk Server uses to maintain the physical disk header map.

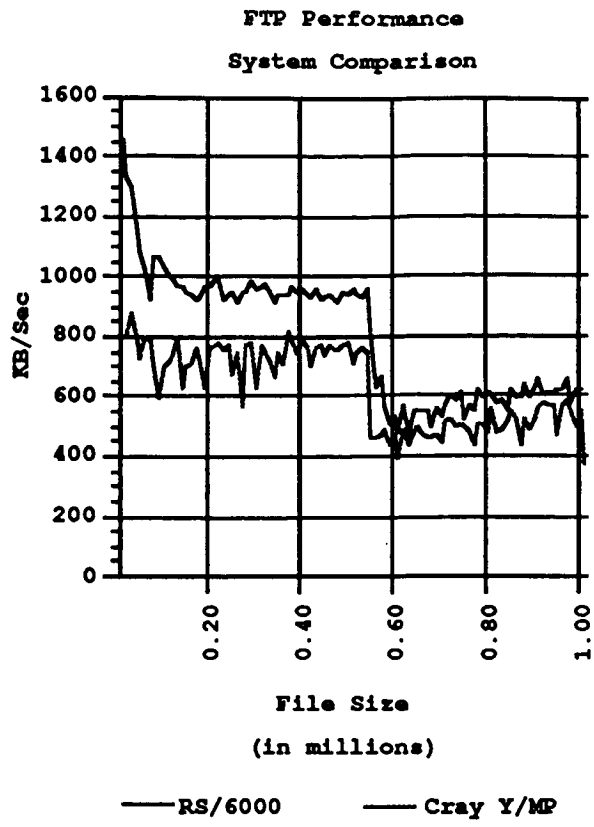


As with the UniTree Name Server profiling data, the categories Integrity and Communications show the highest execution usage.

FTP Performance

FTP clients were initiated on several RS/6000 Model 550 systems (connected via ethernet and on the same subnet) and directed to transfer increasingly large files into the UniTree system. Multiple instances of the FTP test programs were initiated and synchronized on separate systems to eliminate contention for system resources.





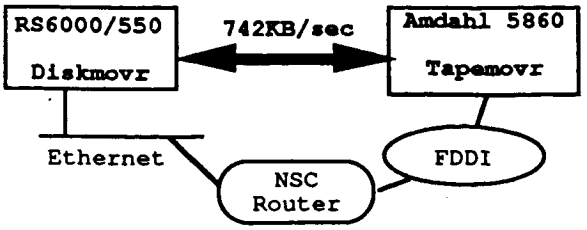
The performance data as cited by the FTP clients shows that there is a 15% degradation in performance as each additional client is added. However, the overall aggregate performance increases almost linearly with each additional client.

An FTP session was initiated on a Cray Y/MP to allow for a performance comparison between the RS/6000 and the Cray Y/MP.

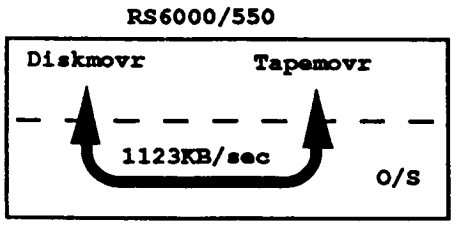
The Cray Y-MP shows comparable performance to the RS/6000 Model 550. The Cray Y-MP was tested while in a production, while the RS/6000 Model 550 was in a dedicated mode. The Cray Y-MP FTP session interfaced with UniTree through an FDDI and ethernet network.

Distributed UniTree Performance

In the NCSA distributed environment, the tape and the disk daemons of UniTree reside on physically separate hosts. The observed performance of the caching and migration of files between the disk daemons on the RS/6000 Model 550 and the tape daemons on the Amdahl 5860 was 742KB/sec.



Observed performance between the tape and the disk daemons when both reside on the same host was significantly faster - 1123KB/sec.



System Scalability

How well does the departmental server scale? Installation of multiple instantiations of the departmental Disk Server as seen in Appendix Figure II result in a disjoint namespace problem. Users do not have location independent file access capabilities under such a configuration. A user creating a file "foo" on archive server A would not be able to access "foo" if he or she were presently using server B for their archiving service. One method by which this problem could be circumvented would be to configure a global nameserver for use by both Disk Server A and Disk Server B (Appendix, Figure III). This configuration has been tested and was deemed functional. However, the UniTree servers lack some necessary intelligence when performing FTP operations and file attribute fetches. For example, the client must pass the address of it's Disk Server to the name server when requesting file attribute data such that the name server could fetch the information from the appropriate Disk Server. Disk Server addresses could be registered in a system configuration file. In summary, the servers would need non-trivial customized addressing enhancements in order to make this distributed system fully functional.

These customized enhancements are not, however, the correct approach to resolving the deficiencies in scalability. A scalable filesystem interface tightly integrated with the archive filesystem would be an effective way to solve the system scalability problem. This integration effort will be the focus of ongoing studies and software development efforts at NCSA.

Summary

The RS/6000 performed well in our test environment. The potential exists for the RS/6000 to act as a departmental server for a small number of users, rather than as a high speed archival server. Multiple UniTree Disk Server's utilizing one UniTree Name Server could be developed that would allow for a cost effective archival system.

Our performance tests were clearly limited by the network bandwidth. The performance gathered by the LibUnix testing shows that UniTree is capable of exceeding ethernet speeds on an RS/6000 Model 550. The performance of FTP might be significantly faster if asked to perform across a higher bandwidth network.

The UniTree Name Server also showed signs of being a potential bottleneck. UniTree sites that would require a high ratio of file creations and deletions to reads and writes would run into this bottleneck. It is possible to improve the UniTree Name Server performance by bypassing the UniTree LibUnix library altogether and communicating directly with the UniTree Name Server and optimizing creations.

Although testing was performed in a less than ideal environment, hopefully the performance statistics stated in this paper will give end-users a realistic idea as to what performance they can expect in this type of setup.

UniTree Archive Server

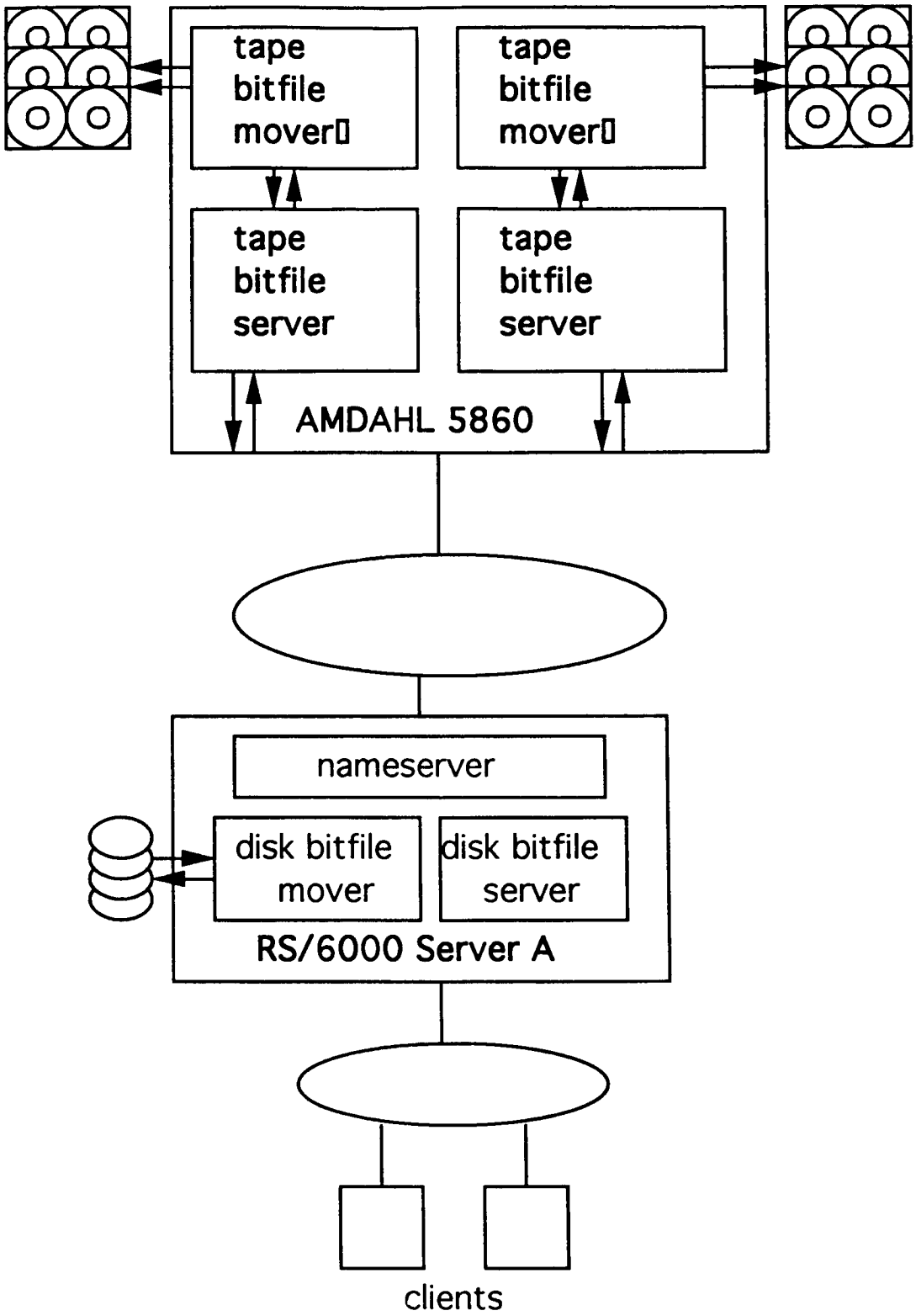


Figure 1

UniTree Archive Server

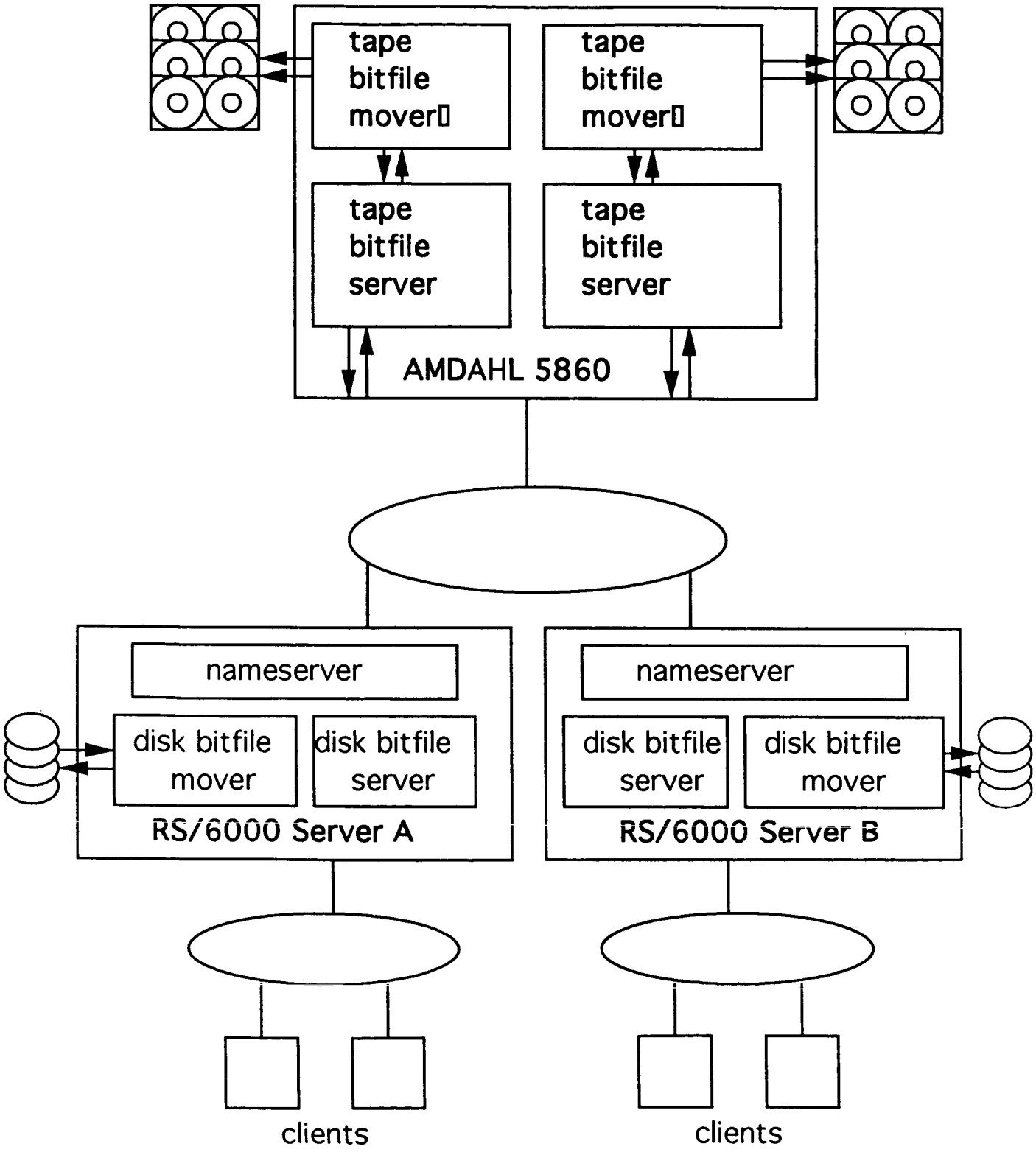


Figure II

UniTree Archive Server

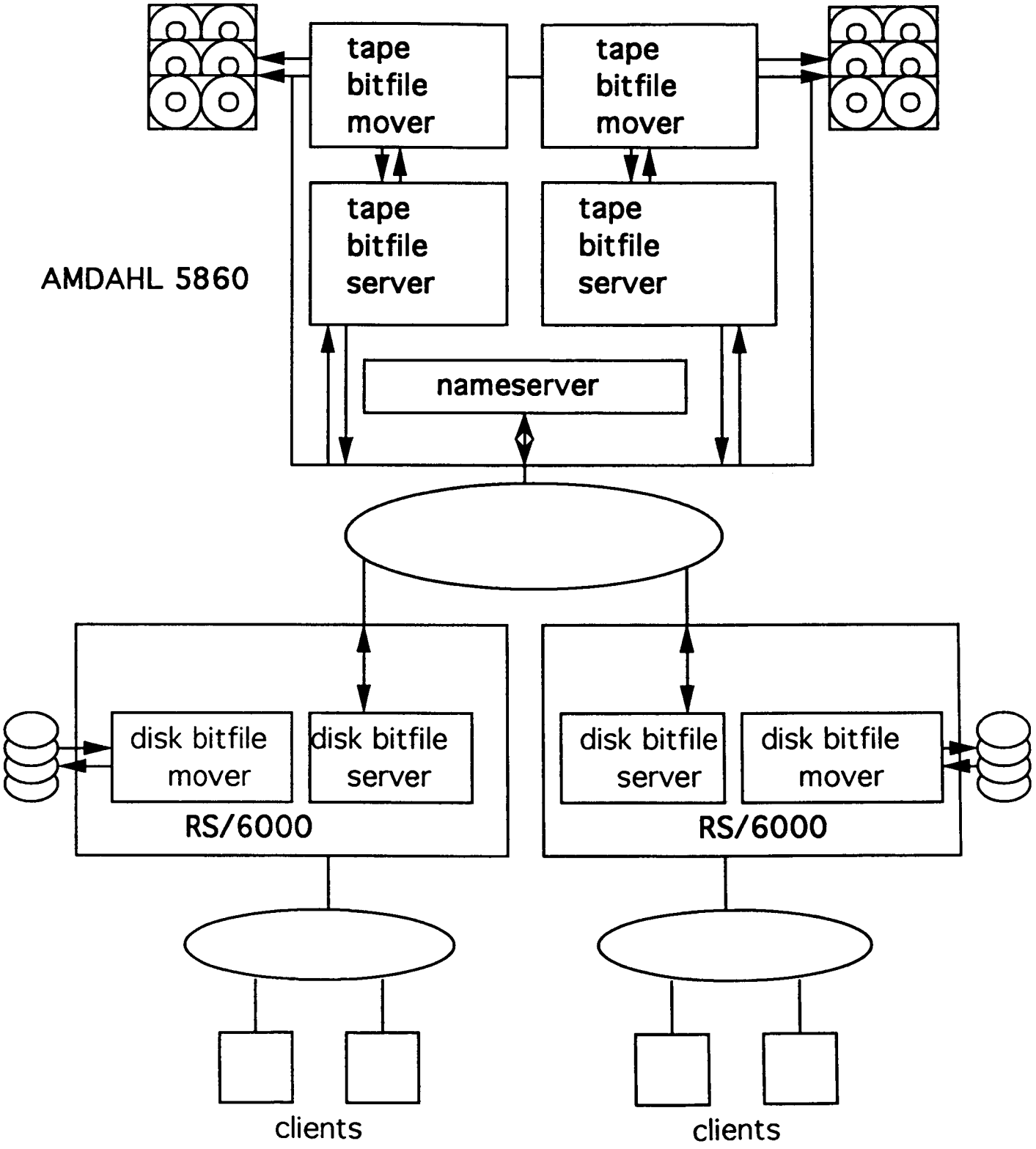


Figure III

The Redwood Project: An Overview

Sam Cheatham

Storage Technology Corporation
Vice President
Tape and Library Systems Development
VC Technical Committee X3B5
2270 South 88th Street
MS 0275
Louisville, CO 80028

StorageTek

REDWOOD™

AN

OVERVIEW

REDWOOD

- REDWOOD IS A NEW GENERATION TAPE SUBSYSTEM NOW UNDER DEVELOPMENT AT STORAGETEK USING HELICAL SCAN TECHNOLOGY.

- THIS LIBRARY BASED STORAGE SUBSYSTEM IS DESIGNED FOR THE HIGH PERFORMANCE, DEEP ARCHIVAL MARKET.

REDWOOD SUBSYSTEM OVERVIEW

- RedWood is the outgrowth of a series of internal strategic planning and customer advisory board meetings.
- The RedWood Project, combined with the StorageTek Library Systems, is StorageTek's strategy in satisfying our customer requirements.
- RedWood consists of combination of:
  - High-performance 36-track StorageTek tape subsystem
  - State-of-the-art digital video system as used in broadcast studios

REDWOOD SUBSYSTEM OVERVIEW

- The architecture of the RedWood tape subsystem takes best advantage of the formats and operational parameters defined for video 'D3' devices.
- Capacity per meter of as much as 50 times more information than 3490E cartridges.
- State-of-the-art media formulation integral factor in deck's design.

REDWOOD SUBSYSTEMS OVERVIEW

MEDIA

- This current generation of MP media is an ideal candidate for reliable data storage.
- Significant improvements have been achieved in durability and stability over earlier generations.
- The RedWood MP media along with its improved cartridge will meet or exceed 3480-class media lifetimes.
- MP media technology will continue to benefit from extensive R & D expenditures in the commercial broadcasting sector and from work now in process within the data storage sector.

MEDIA STANDARDS

3rd Draft

PROPOSED

AMERICAN NATIONAL STANDARD

HELICAL-SCAN DIGITAL COMPUTER TAPE CARTRIDGE

12.65 mm (0.50 in)

FOR INFORMATION INTERCHANGE

13 May 1992

(ASC X3 Project No. 850-D)

Prepared by

Technical Committee X3B5

of Accredited Standards Committee X3

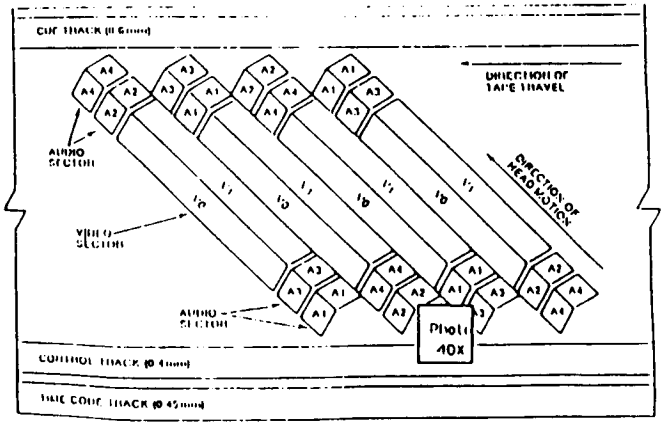
Revision History

1st Draft:	X3B5/91-228	14 August 1991
1st Draft:	X3B5/91-228A	14 November 1991
2nd Draft:	X3B5/91-466	12 February 1992
3rd Draft:	X3B5/92-068	13 May 1992

REDWOOD DEVELOPED TAPE

- Using a ferromagnetic fluid, the magnetic domains of a recorded tape can be viewed under magnification to show track alignment, transition spacing, defects, and data format.
- Scale factor is 125X or about 2mm of the 12.65mm tape width is represented by this slide.
- Bottom transitions are the longitudinal time code track, which are used for high speed searching and location verification.
- Middle transition are servo sync pulse used to align the reel motors, capstan, and scanner motors for precise positioning of the tape.
- Helical tracks in the upper area are written at opposing 20 ° azimuth angles to reduce cross talk and allow gapless recording. The tracks are written at an helix angle of 4.92 ° with a track spacing of 20 uM (1270 TPI).

D3 HELICAL RECORDING FORMAT



REDWOOD CARTRIDGE

- Same media as D3 with video format extended for data.
- Storage reel only - permanent take-up reel; in drive.
- Packaged in a 3480 form-factor cartridge.
- Accommodates same range of tape length as 3480.
- Meets or exceeds 3480-class media lifetimes.
- ANSI format includes data compression.

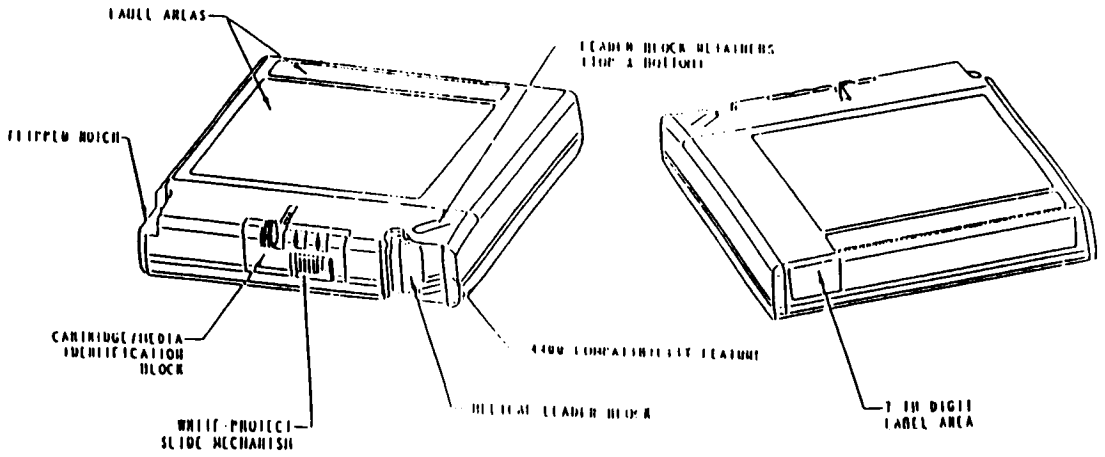
REDWOOD CARTRIDGE FEATURES

- Tape pulled from opposite corner to 3480
  - Straighter path for loading arm
- Improved leader block design over 3480
  - Field-replaceable without special tools
  - More reliable latching mechanism
- Notch to ensure no damage if inserted into 3480 drive
- Design ensures no damage if 3480 cartridge inserted in helical scan device

REDWOOD CARTRIDGE FEATURES

- Improved write-protect switch, length and machine type recognition scheme
  - Separate cartridge/media identification block for ease of manufacturing
- Same label areas as 3480 cartridge
  - Additional area on trailing-edge

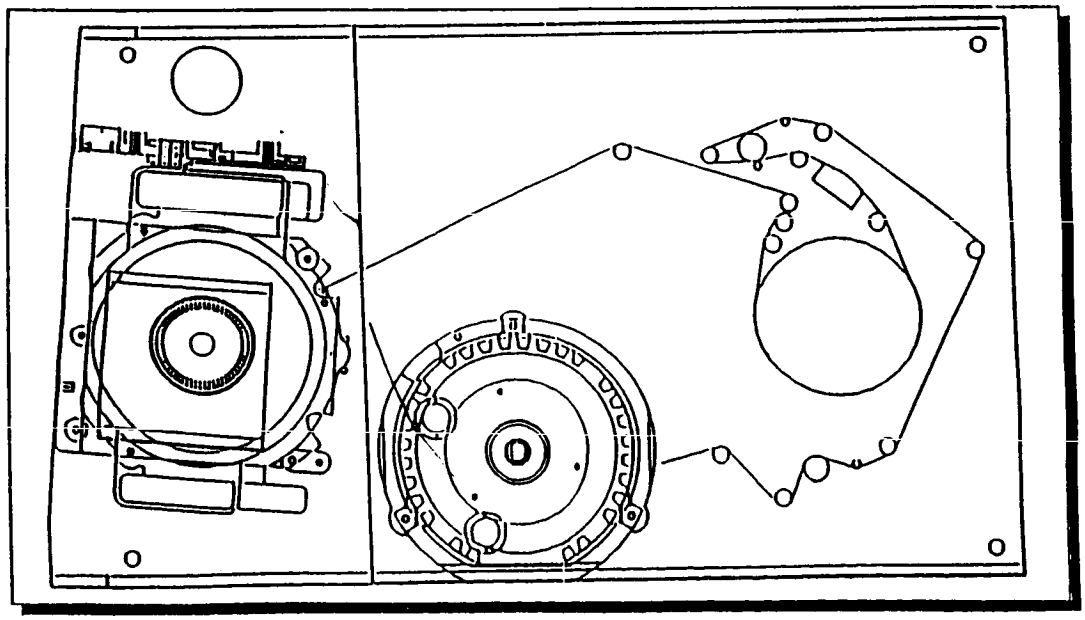
REDWOOD CARTRIDGE



REDWOOD SUBSYSTEM OVERVIEW

- Head and tape wear characteristics differ from linear tape devices since the relative head-to-tape speed for RedWood is quite high, on the order of 1000 ips, in support of available data rates.
- Life expectations for media and heads will far exceed those currently thought possible in all modes of use.
- StorageTek has defined a means to keep track of customer tape and head usage facilitating preventive maintenance, timeliness and convenience to the customer.

STK HELICAL DECK



REDWOOD 1 FEATURES

- o CAPACITY PER METER OF AS MUCH AS 50 TIMES MORE INFORMATION THAN 3490E CARTRIDGES
- o DEVICE DATA RATES COMPLEMENTARY TO CAPACITY
- o 18 MB/S CHANNEL DATA RATE/ESCON
- o 10 MB/S CHANNEL DATA RATE/SCSI II
- o FIBER CHANNEL
- o BIT ERROR RATE OF 10^{-15}

REDWOOD
STORAGETEK PROPRIETARY

REDWOOD 1 FEATURES

- o High speed search: 60 to 100 times
  - Position of key records retained by for future searches
- o File Safe™
  - Allows tape to be written once and only once
  - Information can be appended, but existing records cannot be overwritten
  - Emulates optical Write Once Read Mostly (WORM)

REDWOOD
STORAGETEK PROPRIETARY

INTERFACE OVERVIEW

- Higher device data rate requires fresh approach
  - Original ESCON announcement by IBM barely fast enough
  - Need ESCON performing at limit (approx. 18 megabytes/s)
- Given the push towards open-systems and standards
  - New versions of SCSI (SCSI-2 fast and wide) will see use for workstation market with RedWood
  - HIPPI established in supercomputer systems, and thus is addressed in RedWood architecture
  - Fiber Channel expected to become interface of choice for medium and high-performance RedWood systems users

HELICAL VERSUS LONGITUDINAL

- Helical has longer mechanical latencies
  - Not a problem - uses large buffers
  - Very short records may lead to non-optimal performance and capacity utilization
- Helical has lower inherent BER than longitudinal
  - Add 3rd Level ECC to achieve 10^{-15}
  - Both 3rd Level ECC and write retry can be disabled by system

REDWOOD NEARLINE OVERVIEW

- All RedWood products will operate in a library environment. The architecture also allows the customer to use stand-alone drives with or without stacker-loaders.
- All StorageTek libraries are capable of storage and management of the new helical scan cartridge.
- General availability features will include mixed media in the StorageTek library family (both helical and 3480 type media).

LIBRARY COMPARISONS

LIBRARY		TIMBERWOLF	WOLFCREEK	4400/PH
Estimated Floor space (sq ft)		23	33	100
3490E (36 track)	Capacity	0.2	0.4	2.4
RedWood 1	(terabytes)	10	20	120

- Capacity comparison does not include compression
- Floor space does not include access for manual loading, servicing

HOST SOFTWARE FOR REDWOOD LIBRARIES

- o IBM ARENA
  - MIXED MEDIA SUPPORT AT GA
- o OPEN SYSTEMS
  - CUSTOMER REQUIREMENTS WILL BE USED TO REFINE HOST SOFTWARE AND DEVICE CONNECTIVITY FOR EACH SPECIFIC CASE.

PRODUCT EMPHASIS

IBM ARENA

OPEN SYSTEMS

FEDERAL AGENCIES

MID RANGE PRODUCTS

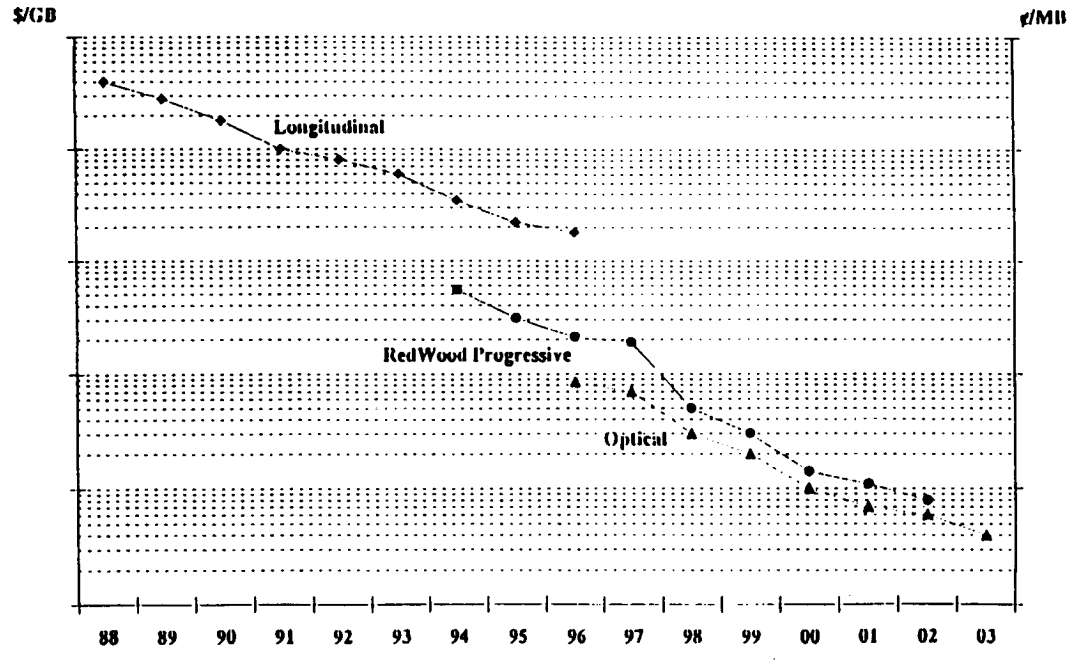
MARKET OPPORTUNITIES

- Continued steep growth in capacity requirements
- Driven by new applications
  - Imaging, seismic
  - High-definition full-color video stored digitally
  - Archives formerly on fiche, etc.
- Only 1% of business data stored in digital form in 1990, growing to 3-5% by year 2000 (source AIM report)
  - Every company will have a true "mass storage" problem as percentage grows

PRODUCT AVAILABILITY

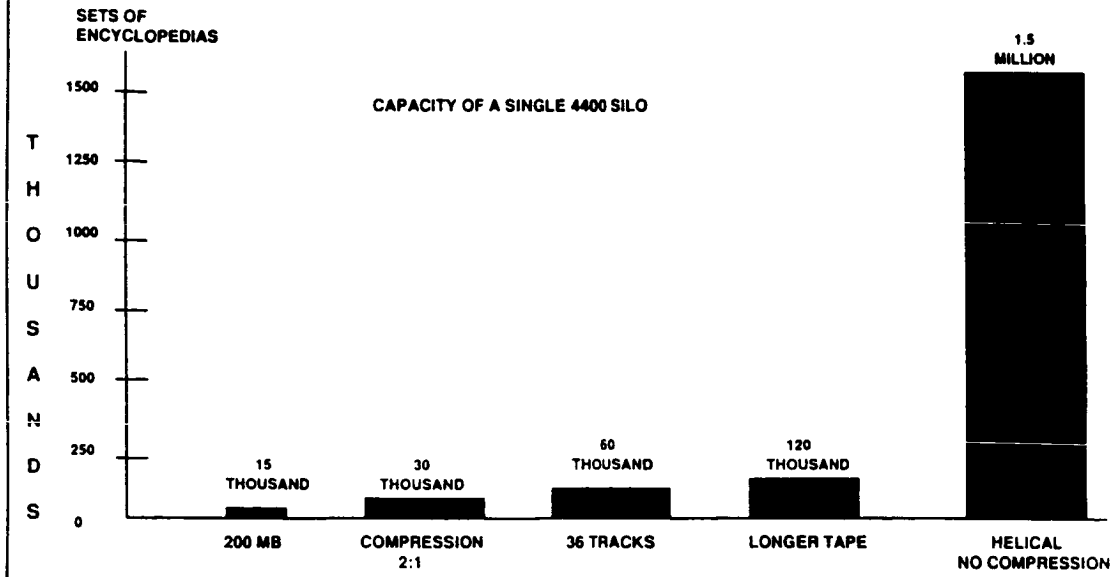
1994

TAPE SUBSYSTEM PRICE TRENDS



REDWOOD
STORAGETEK PROPRIETARY

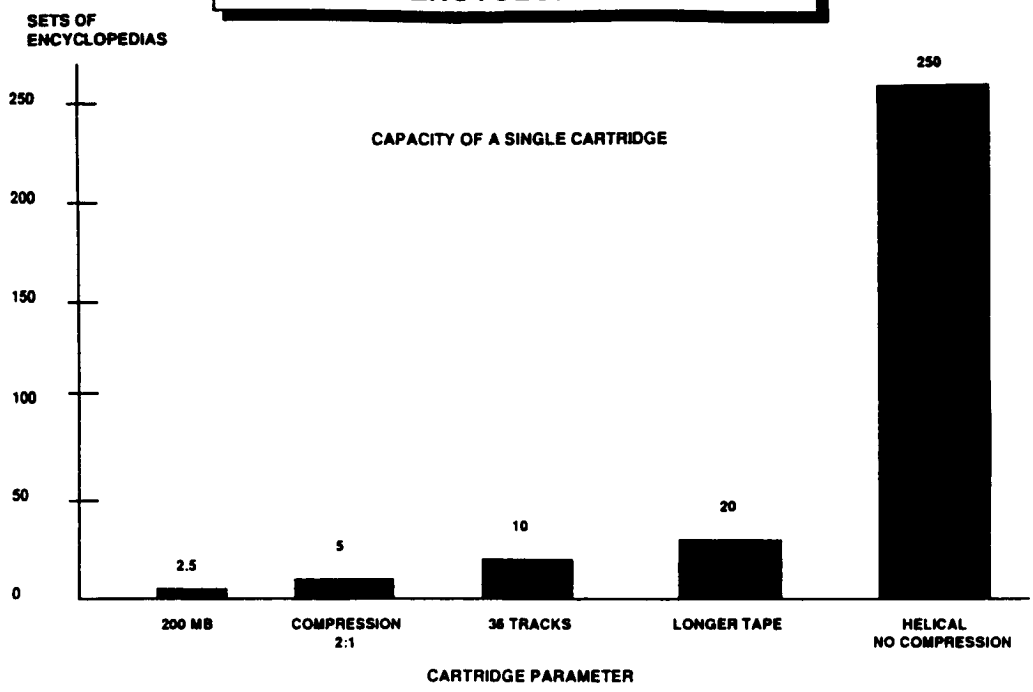
REDWOOD POTENTIAL IN A 4400 LIBRARY IN SETS OF ENCYCLOPEDIAS



SILO CAPACITY PER CARTRIDGE CAPACITY

REDWOOD
STORAGETEK PROPRIETARY

CARTRIDGE CAPACITIES IN SETS OF ENCYCLOPEDIAS



SUMMARY

- RedWood allows investment in current-generation Nearline technology to be preserved in defining next-generation mass-storage systems.
- Helical scan technology offers order-of-magnitude improvement in capacity and density, cost/GB for all Libraries.
- Use of existing broadcast technology in RedWood significantly lowers risk.

  Extensive R&D expenditures in the commercial broadcast sector will facilitate future generations of StorageTek helical-scan products

SUMMARY

- Proposed standards-based helical-scan cartridge
  - Allows existing Nearline products to be upgraded by mixing new media with the existing cartridge set
  - Provides improvement in data rate over existing Nearline products in archival applications
- New helical-scan features, e.g., high-speed search, File Safe will be application enablers.
- RedWood facilitates use of Nearline technology in next generation mass-storage systems.

Architectural Assessment of Mass Storage Systems at GSFC

M. Halem, Code 930.0
J. Behnke, Code 633.1
P. Pease, Code 935.0
N. Palm, Code 931.0

NASA/Goddard Space Flight Center
Greenbelt, MD 20771

Architectural Assessment of

Mass Storage Systems at GSFC

NDADS: National Space Science Data Center
Data Archive and Distribution Service

GDAAC V.0: Earth Observing System Data Information System
Goddard Distributed Active Archive Center

M(DS)^2: NASA's Center for Computational Science
Mass Data Storage and Delivery System

by

Dr. M. Halem, J. Behnke, P. Pease and N. Palm

Space Data and Computing Division
NASA/Goddard Space Flight Center

Presentation for
Goddard Conference on
Mass Storage Systems &
Technologies
NASA/GSFC, Bldg. 8
September 24, 1992

OVERVIEW

- Background
- System Functionality
- Characteristics
- Data Sources
- Hardware/Software Systems
- Performance Assessments
- Conclusions

BACKGROUND OF MASS DATA STORAGE SYSTEMS

NDADS:

Prototype of the Hubble Space Telescope Data Archive and Distribution Service (HST-DADS) contracted to Loral AeroSys in 1989. Evolved as the Astrophysics and Space Physics archiving system for the National Space Science Data Center to maintain a mix of near and on-line data and manage a deeper data storage archive

GDAAC/V.0:

EOS prototype archive and distribution systems initiated in FY91 and planned for operational availability in FY94. One of nine geographically distributed discipline-oriented interoperable DAAC's

M(DS)^2:

A mass storage and delivery system serving more than 1400 users within the NASA Computational Science Center at Goddard that has to manage both the high-speed computer-generated simulation data, as well as space-borne observational data

SYSTEM REQUIREMENTS

	NDADS	GDAAC V.0	M(DS)^2
NEAR-ONLINE STORAGE /DEEP	2.6 TB/6 TB 16GB DASD	10TB/3TB 16GB/DASD	7TB/35TB 240 GB/DASD
SCALABLE UP TO	10 TB/50TB 100 GB DASD	18 TB 100 GB/DASD	225 TB/500 TB 3 TB/DASD
INGEST (RATE)	13 GB/DAY	30 GB/DAY	90 GB/DAY
DISTRIBUTE (RATE)	1050 MB/DAY - NET 700 MB/DAY - TAPE 100 PHOTOS/DAY	150 GB/DAY	100 GB/DAY
PEAK CONCURRENT USERS	146 240 CATALOG QUERIES/HR	100	MIN. 128 MIN. 32 SIMULTANEOUS FTP TRANSFERS

SYSTEM FUNCTIONAL CHARACTERISTICS

Data and Metadata Functions	NDADS	GDAAC	MDSDS
Network Access (Ethernet, FDDI, DecNet, UltraNet)	x	x	x
Security	x Barrier	x Barrier	x RACF/C2
Integrity and Quality Control	x	x	
Automated Data Migration and Compaction	Partial		x
User Ingest and Retrieval	Partial	Partial	x
Remote Ordering and Delivery Service	x	x	
Catalogues and Inventories	x	x	User
Browse - On-Line	x	x	
Interoperability		x	Partial
Database Queries and Subsetting	x	x	
Portable Software Operating Systems		x	x
Incorruptible Archive	x	Partial	Partial
Remote Back-Up/Safe Store	x		———
Data Compression	x	x	———
Redundancy (NSPOF)	Partial	Partial	x
Scalable Upgrade	x	x	x
Accounting & Monitoring	x	Partial	x

NDADS Data Sources:

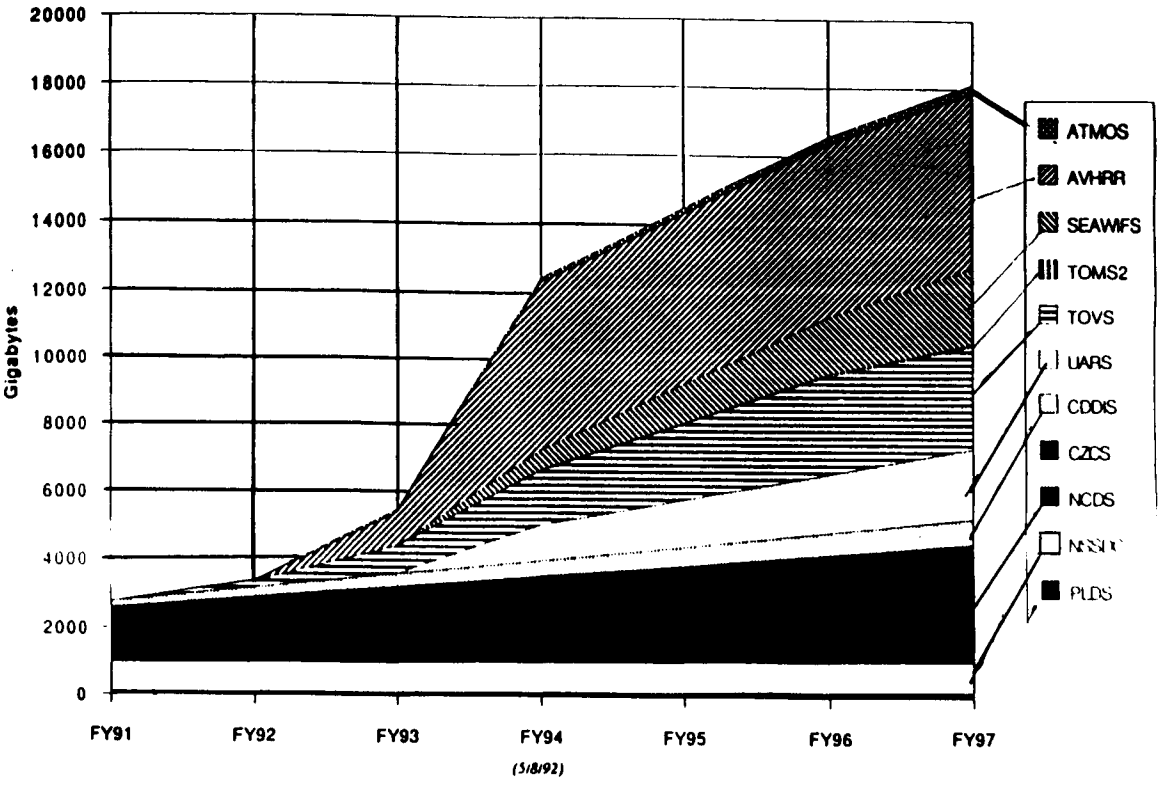
Astrophysics:

Wave Length	Project	DataTypes	Granules	Archived	Total Size
High Energy	EXOSAT	10	34K		150 GB
	HEAO-3	2	10K	100%	4 GB
	HEASARC	200	100K		20GB
	EINSTEIN	20	15K	30%	120 GB
	ASTRO-D	launch 1993	—	—	1650GB
	XTE	launch 1996	—	—	
	GRO	100	—	none	50GB/year
	ROSAT	60	> 100K	30%	100 GB
	VELA 5B	1	1K	100%	3GB
Ultra Violet	IUE	6/8	80K	100%	200 GB/750GB
Copernicus	EUVE	launch 1992	—	—	50 GB
Optical	HST	4			1GB/4GB
Infrared	COBE	6	100K	0	30 GB
	IRAS	6	150K	100%/10%	5GB+78GB
Radio	VLA	1	—	60%	1000 GB
Miscellaneous	ADC Catalogs	Various	1K	100%	1GB
TOTAL					2.41 + (1.75)

Space Physics:

Mission	Data Types	Granules	Archived	Total Size
SKYLAB	1	3500	100%	10 GB
ISTP-GOES	8			.4 GB
ISTP-IMP8	8			.4 GB
ISTP-GEOTAIL	8			1.0 GB
Atomic Physics	2	100	10%	1 GB
DE-1				100 GB
VOYAGER				2 GB
TOTAL				114 GB

GSFC V0 DAAC Data Volume Requirements (by Project)



Science Project Data Products

Project	Product Description
UARS	• Profiles of 15 trace species, temperature, and wind • Solar UV irradiance measurements (115 - 400 nm)
SeaWiFS	• Ocean pigment, chlorophyll a concentrations • 5 water leaving radiances, 3 aerosol radiances • Diffuse attenuation coefficient
Atlas / ATMOS	• Profiles of 30+ upper atmosphere trace species • Upper atmospheric temperature profiles
TOMS2	• Total ozone, effective tropospheric reflectivity • 6 backscattered UV radiances (313 - 340 nm)
AVHRR Pathfinder	• Binned 5 channel clear sky radiances • Daily cloud fraction, height, and reflectivity at 9 km and 1 degree spatial resolutions • Daily, weekly and seasonal surface reflectance NDVI at 9 km resolution • Daily Surface albedo at 9 km resolution • Aerosol optical thickness, longwave surface flux
TOVS Pathfinder	• Profiles of atmospheric temperature, humidity, and geopotential height • Precipitable water in 6 tropospheric layers, total ozone, and tropopause pressure • Surface air and skin temperatures, 3.7 micron bidirectional surface reflectance, and 50 GHz surface microwave emissivity • Cloud fraction, cloud top pressure, precipitation estimate, visible reflectance, outgoing longwave radiation, and longwave cloud forcing

M(DS)2 DATA SOURCES

PROJECTS: NIMBUS/TOMS
ISTP
IUE
GRO

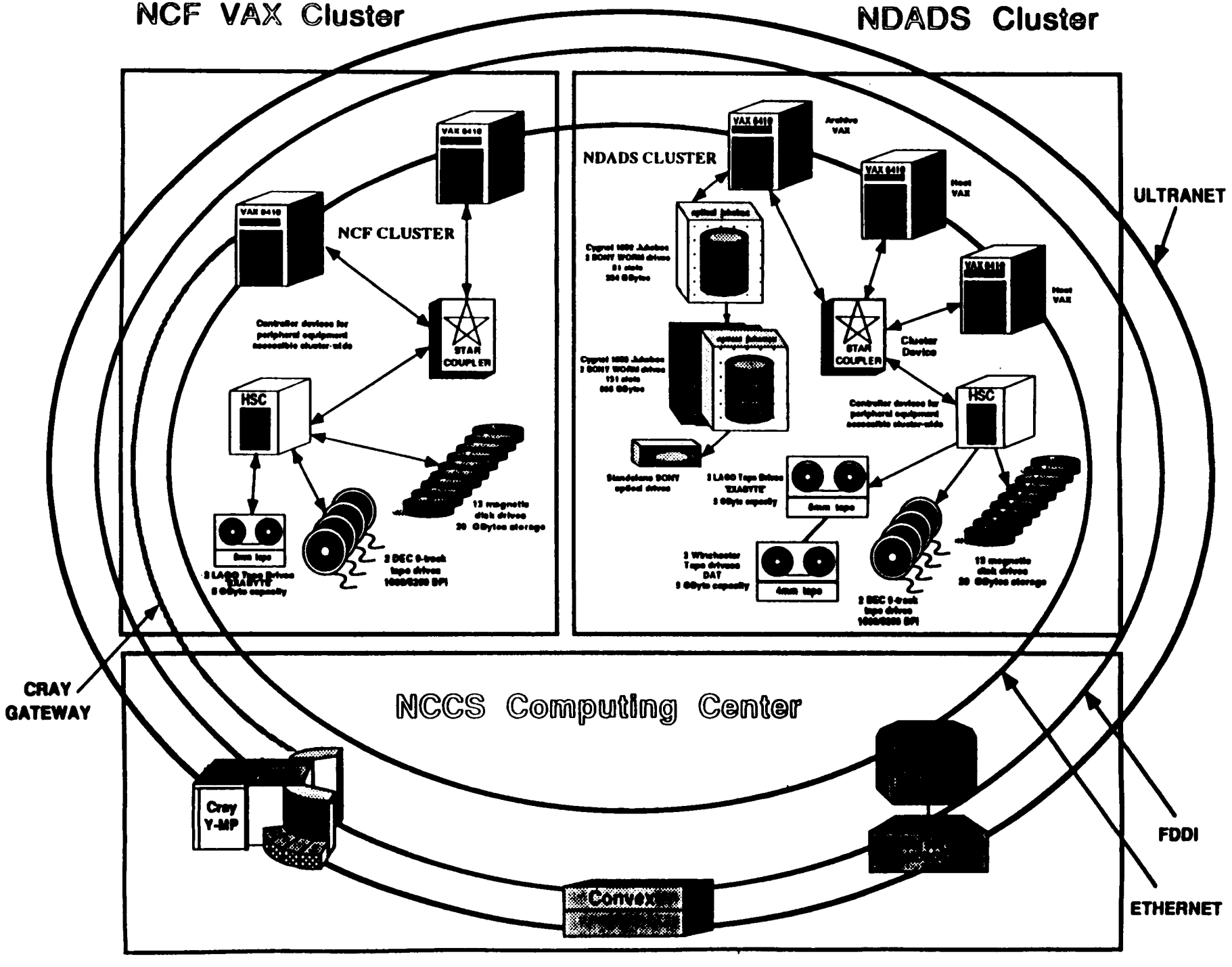
MODELING: DATA ASSIMILATION
COUPLED OCEAN/ATMOSPHERE/STRATOSPHERE
GEODYNAMICS
SPACE PHYSICS PLASMA MODELING

ANALYSIS: TOVS PATHFINDER
ALGORITHM DEVELOPMENT (TRMM, MODIS...)
HST IMAGE DEBLURRING

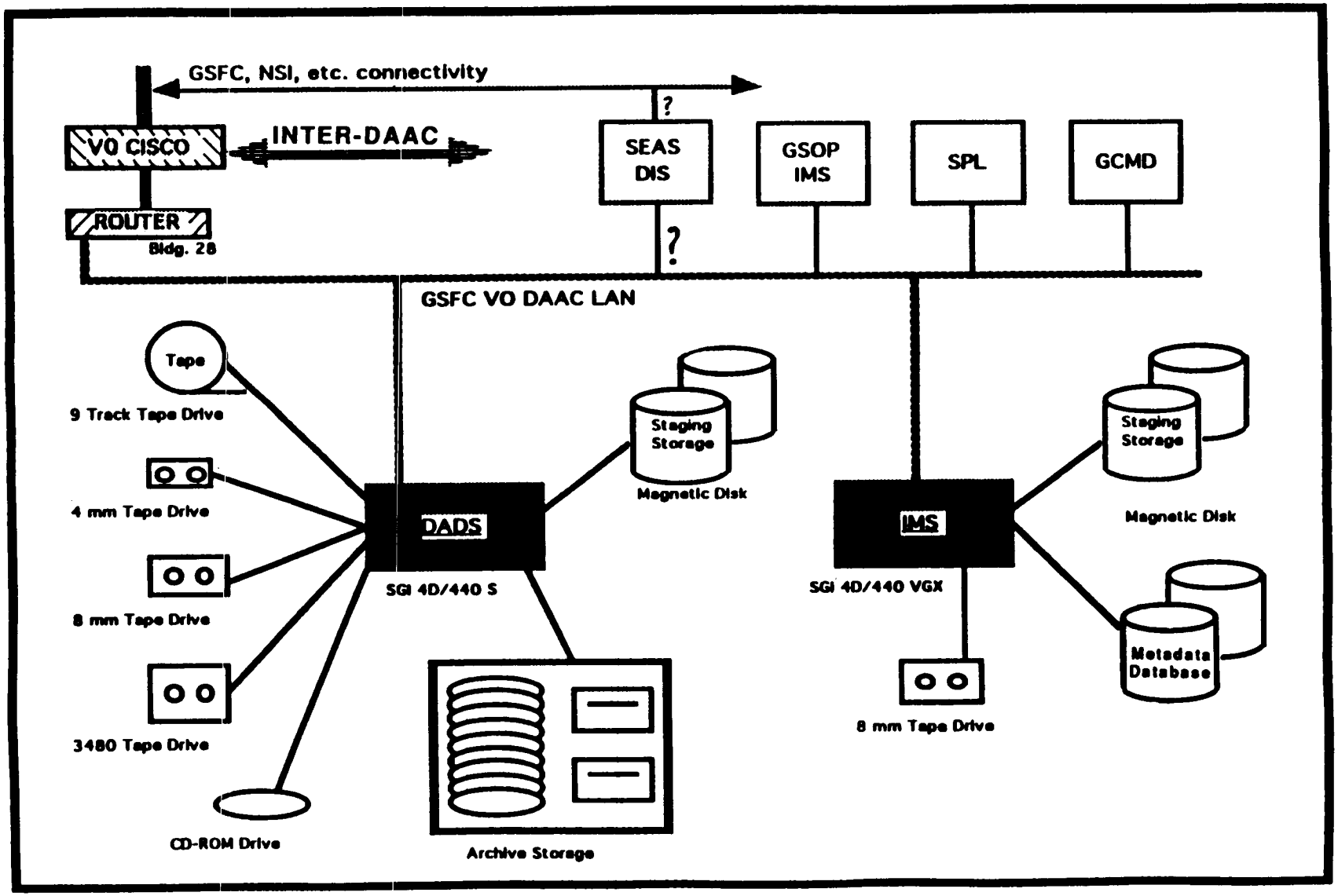
HPCC: EARTH AND SPACE SCIENCE TESTBEDS

Mass Storage Hardware Systems

NDADS	GDAAC	MDSDS
• 3 VAX 6410's (14GB/DASD)	• 2 SGI 4D/440 (16 GB DASD)	• IBM ES 9021/500 (56 ch, 128 MB • IBM 3980 - (240 GB/DASD) • Convex 3240 - (512 MB)
• 2 CYGNET WORM Jukeboxes with 4 SONY drives	• 2 CYGNET WORM Jukebox with 4 ATG 9001 drives (24 TB) • 1 Metrum ACS (8.7TB) w/ 4 drives	• 3 STK 440 (4.8TB) • 1 Dataware WORM Jukebox 3 34/850 (1.2 TB) --- • B-Test Helical E-Systems Tower (8.2 TB)
• 2 9-track Dec tape drives • 2 8mm Exabyte tape drives • 2 4mm Winchester tape drives • CD-ROM pre-mastering	•1 I/O power channel 80MB/s •5 6250 9 track tape •4 Exabyte 8 mm tape drives •2 3480 tape cartridge • CD-Rom pre-mastering • 2 4mm DAT tape drives	• 8 ESCON Channel • Ultranet • 40 Memorex 3480 compatible tape drives

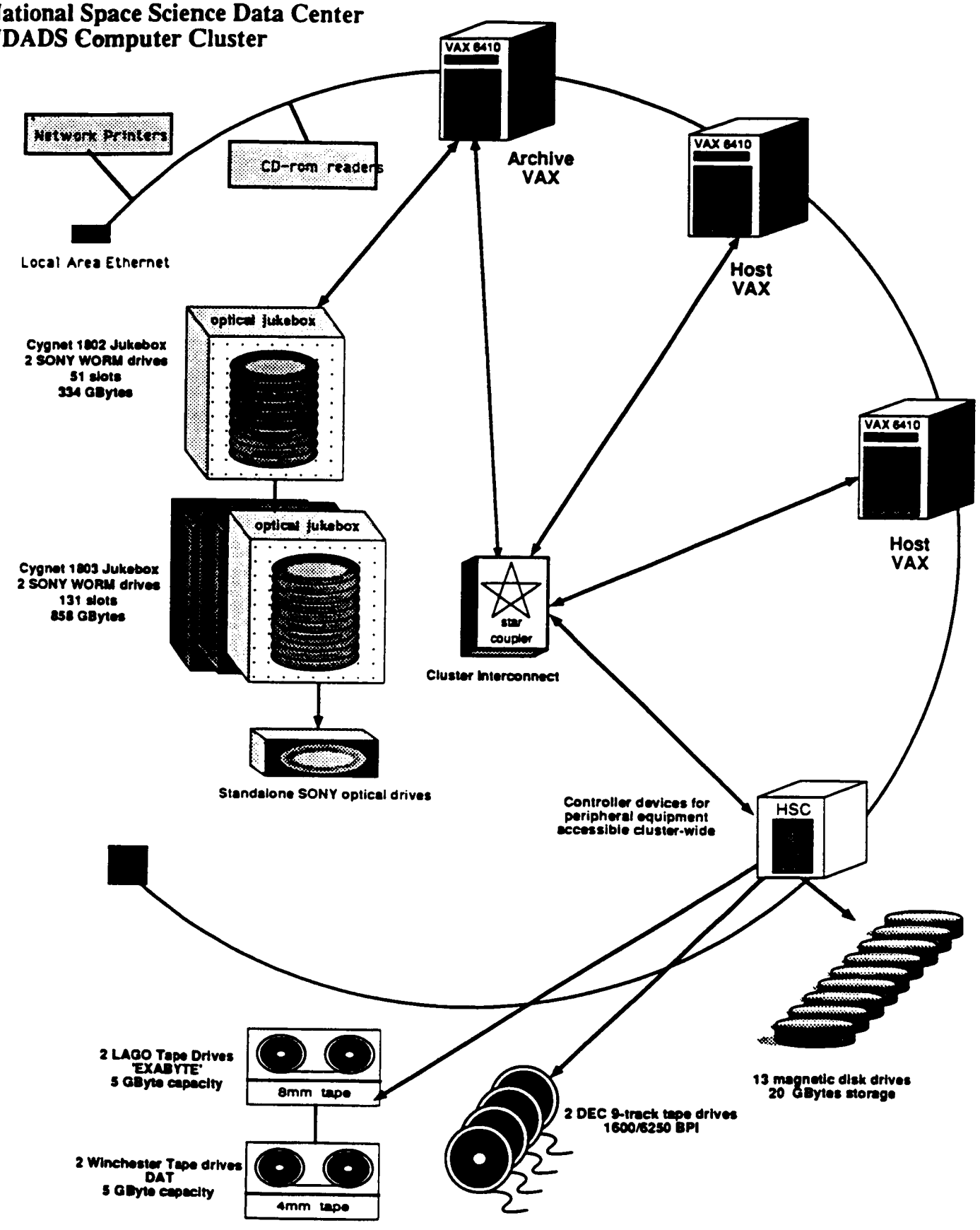

NCF VAX Cluster
NDADS Cluster
NCF CLUSTER
NDADS CLUSTER
STAR COUPLER
HSC
ULTRANET
CRAY GATEWAY
NCCS Computing Center
Cray Y-MP
FDDI
ETHERNET

Hardware Architecture Components



Goddard DAAC

National Space Science Data Center
NDADS Computer Cluster



Mass Storage Software Systems

Software System	NDADS	GDAAC	MDSDS
Client/Server Operating System	VAX VMS	IRIX	MVS, AIX
Networks Supported	DECnet TCP/IP SPRINTnet	TCP/IP	TCP/IP Ethernet BITNET NSFNET InterNet InterLink UltraNet
Database/Library Management System	FSTAGE/FSTORE SYBASE INGRES	Unitree ORACLE	HSM Unitree Oracle (Opt.)
Physical Storage Device Driver Software	JIMS SOAR	Unitree Drivers	Dataware STK Unitree Drivers
User Interface	NCDS ARMS* NSSDC ARCHIVE	NCDS/PLDS EOSDIS/IMS	USER CONTROL

* AUTOMATED RETRIEVAL MAIL SYSTEM

System Assessments

	NDADS	GDAAC	MDSDS
Strengths:	• Project customization • FTP Accessible • Distribution of archive media • Intelligent data access and optimization • Data compression • Metadata search/browse • Remote back-up	• Project customization • FTP Accessible • Interoperable • Metadata search/browse NCDS/PCDS Experience Intelligent Data Mgmt. SpatialTemp Data Fusion — • Metadata search/browse • IEEE Mass Storage Compliant • Open	• FTP Accessible worldwide • Archival/retrieval by user request • Intelligent hierarchical storage migration • Remote back-up option • IEEE Mass Storage Compliant
Weaknesses:	• Non-portable systems • In-house customization of software and hardware	• In-house customization of software • Embryonic HW/SW systems	• Embryonic HW/SW systems • Costs

PERFORMANCE CHARACTERISTICS

THESE ARE ALL PRELIMINARY ESTIMATES FOR THE NDADS M(DS)^2 AT THIS POINT!!

(DOES NOT INCLUDE PROJECT SPECIFICS)

NDADS

SONY Optical Disk Drives:	Read:	600 KB/sec
	Write:	300 KB/sec
Actual Rates:	Read:	250 KB/sec
	Write:	107.52 KB/sec (average)
		250 KB/sec (max)
Platter load speed:		11 seconds
Data Storage: Current:		120 GBytes
Growth:		240 GBytes/year
Storage Input:		5 GB/day (average), 17 GB/day (max)
Storage Output:		330 files/day, electronically
Inquiries on the archive:		24/day

M(DS)^2

DASD:	80 MB/S Throughput
Mass Storage:	18 MB/S Throughput

CONCLUSION

- Mass storage systems allow scientists to perform research previously impossible because of logistic burdens and maintain pace with rapid data growth arising from increasing computational power and observational resolution

- Mass storage hardware systems technology evolving faster than software available to integrate into system. IEEE mass storage standards model changing faster than vendors can keep up with; still need standards.

- Community needs to acquire much more performance test data reliability, stability and data access speeds across small to large mass storage systems

- Not yet clear whether many small distributed client-servers are more effective than fewer large-scale client servers

- Mass storage management systems need to become more robust and more stable

Panel Discussion on High Performance Helical Scan Recording Systems

Mr. J. F. Berry was the panel moderator. Mr. Berry is the Chief of a Processing Division in the Department of Defense operation at Ft. Meade, Maryland. For the past five years, Mr. Berry has been actively involved in the use of high capacity mass storage systems in his organization.

The panel members represent companies who are actively pursuing the development of products for the high performance mass storage market. The panel included:

Sam Cheatham, *Storage Technology Corporation - 3480 and D-3 Recorders*
Norris Huse, *Datatape Inc. - ID-1 Recorders*
Donald Morgan, *Sony Corporation of America - ID-1 Recorders*
Michael Riddle, *GE Aerospace - ID-1 Recorders*
Tracy Wood, *Ampex Corporation - DD-2 Recorders*

Question: J. F. Berry - Panel Moderator

Where can helical recording devices be used and in what type of applications?

Responses:

Tracy Wood - Ampex

Ampex has developed a very highperformance, high-data-rate device that is applicable to many of the areas that were being shown in the viewgraph. Ampex's primary marketing thrust is through OEM, rather than directly to end users. Current initial serious interest in the product is in the area of supercomputers. Both Cray and Convex have developed device drivers that will deal directly with our product. There seems to be a nice blend of Ampex D-2 technology and the particular needs of those two firms. Ampex is developing interfaces currently that will interface to both Sun and SGI and consider that a very important application for the product. At the National Storage Laboratory, the Ampex DD-2 recorder will get an opportunity to work in an RS6000 environment.

Sam Cheatham - Storage Technology Corporation

Storage Technology is also attaching its products to many of the systems described in the figure. In STK's products the library control path, which is separate from the data path, facilitates attachment to different types of platforms via either server concepts or through direct attachments. Several of the system listed in the above chart either are directly attached or available today in 3480 recording form via server concepts. STK will have a natural support path for those devices with the Redwood helical product, complemented by the server or direct attach for the libraries, whichever is chosen for the particular application.

Don Morgan - Sony Corporation

Sony made a commitment in the mideighties, to be concerned about their new entry into data recording. As a consequence, Sony participates very strongly in the standards efforts. The ID-1 media was driven by a standards committee who developed the media-based format which is now used. Sony's participation in activities and standards facilitates its interest in attaching to the systems listed in the slide. As these systems evolve they will require more and more tightly integrated storage.

Michael Riddle - GE Aerospace

GE's focus has been in high-rate, high-capacity massive storage systems. The firm's system is based on a 19mm tape line. GE is very active in the ID-1 format, because it has a product there, and its interest in DD-1 is focused toward massive data storage issues. The applications listed in the figure are addressable by GE products which support the high end of the systems.

Norris Huse - Datatape - ID-1

Datatape ships systems which operate at 25 megabytes/sec and 50 megabytes/sec. The Datatape products support the ID-1 standard, and also DoD's 2179 military standard. Datatape's traditional customer base is one of military-type customers. Datatape will be providing the Caltech Concurrent Supercomputing Consortium with a 50-megabyte/second machine with a library unit.

Question: Dr. Hariharan

How far away is the DD-1 standard?

Response: Don Morgan - Sony Corporation

We're at the point now where we see probably a release of a draft to the plenary X3B5 committee in November and submission to ISO/IEC JTC1/SC11 in the Spring of '93. The milestone progress at that point would indicate that it will take about two years.

Question: - Mr. Berry - Panel Moderator

Will a tape recorded in ID-1 would be readable in DD-1?

Response: Don Morgan - Sony Corporation

That becomes a vendor-driven option. You can implement an interface formatter/de-formatter which can be remotely commanded to do a non-DD-1 format read, that is, pass through the ID-1 data.

Question: - Kevin Howard - Exabyte

What is the panel's view of tape striping? Where would it fit into the types of devices they're talking about?

Responses: Sam Cheatham - Storage Technology Corporation

The architecture of the Redwood system is very user-friendly to tape striping for increased data rate and/or capacity in the future. There would be a cost associated with doing striping with D1, D2 or D3 devices, since these are very high- performance subsystems, but it is an available option and would provide tremendous performance if you consider the first model of the device has a 12-megabyte/second data rate.

Michael Riddle - GE Aerospace

GE Aerospace is providing units with this capability, not in the DD-1 but in the ID-1 format, which operate at composite data rates up to 3.2 gigabits/second.

Tracy Wood - Ampex Corporation

The Ampex library system does support up to four drives. We've looked at several different applications where people are interested in doing data striping. At this point, the conclusion has been that when one takes an overall, topdown systems look, one often finds it very difficult to actually, effectively and efficiently utilize the very high bandwidths that are available through data striping, especially if one takes a look at the complexities of error recovery. There are many different and fairly complex error recovery situations that do have to be taken into account. My feeling is that probably the DD-1 and DD-2 and the D-3 based solutions all can be and probably at some point will be supported in data striping operation. I doubt that any of us will be supporting that very early on.

Question: Dr. Hariharan

Is the Ampex DD-2 being put into an array by Maximum Strategy?

Response: Tracy Wood - Ampex

A controller developed by Maximum Strategy will accept up to 4 of our devices. If handled properly, we can sustain the 60-megabyte-second data stream in and out of our library.

Question: Audience

Three different standards (D-1, D-2, and D-3) are represented on the panel. Is there enough volume for all three to keep you in business in the next 3 or 4 years?

Responses: Don Morgan - Sony Corporation

The attendees of this conference have heard presentations about research in techologies which will show great promise for mass storage of data in the future. There is a very large gap between the development of research possibilities and the creation of a production-acceptable product based on that technology. In the the time-frame of that development gap, there needs to be a differentiation of applications, a definition of the detailed requirements for mass storage solutions for each of these applications. No single hardware solution can satisfy the requirements of all applications. The products based on these formats provide the tools for applications to do their business in the near future. Different applications evolve and will require different parameters of performance for mass storage of data. Over a period of time we may do so, however, currently I think it is difficult for us to predict the growth of applications.

Sam Cheatham - Storage Technology

We have a unique position in this business because of the form factors that we have selected to support compatibility with our existing and future library and the types of customers to whom we sell. In addition to this segment of the overall marketplace, we also run into many of the communications, banking industry, transportation industry, plus various other business segments. So as we view their data processing needs in the future, we see a widespread market for this type of product and this form factor. It's a natural complement as the data storage requirements increase in the future, he said. Through the compatibility approach that we have taken, we will facilitate migration of this new subsystem into the library-based applications as well as stand-alone. Based on the market research we've done so far, we don't see that as any kind of an issue.

Tracy Wood - Ampex

If you look at the professional broadcast industry where the core technology comes from, there are three major players in the high end of that business: Sony, Ampex, and Matushita, representing the D-1, D-2, and D-3 formats. If you look at this panel, with one exception, all of the technologies we're talking about here are derived from those base technologies, which are already established in a different business area. I would give a qualified yes to the original question. All companies involved here are in fact using the base technology in multiple markets.

Question: J. F. Berry- Panel Moderator

What is the panel's opinions on tape striping?

Responses: J. F. Berry - Panel Moderator

There are lots of different varieties of tape striping if one uses multiple drives. What does the panel think about the possibility of getting commodity drives for tens of thousands of dollars, not hundred of thousands of dollars?

Sam Cheatham - Storage Technology

There's a disadvantage here when we talk about tape striping, because you have a panel full of people dealing with a focused requirement area that involves what is traditionally thought of as today's environment for data striping. We will gradually see the pricing of all of our products be driven down somewhat. There's a natural evolution for data processing products. But they will not come down to the level that we think about as the generic low end. These drives will cost more because of the product segment that they're in now. The media cost will come down, to a certain degree, but then it become a "square-inch situation." The media cost starts becoming very heavily influenced by the amount of surface area in a particular medium, as opposed to the cost of a cassette or cartridge.

Question: Dr. Halem

Using the hypothetical role as a manager of a production facility serving 1500 people, how do we guarantee ourselves against brownout situations or AT&T breakdowns where the whole DC Metropolitan area can't make a phone call? In other words, what's to prevent a 100-terabyte system from going down permanently?

Response: Sam Cheatham - Storage Technology

Traditionally, UPS (uninterruptible power supply) systems are used to support large data processing centers. It's usually a two- or three-stage device that immediately will, upon sensing a power sag, kick in a motor-generator set, usually powered by a diesel. There's immediate backup by battery and inverters, then going to diesel power system. That's good only for some period of time, since you can run out of diesel fuel. For large installations where continuous availability is needed, they should all be run on a UPS system.

Question: Dr. Halem

There is a situation where the software led to problems on 2 terabytes of data. It took them something like 6 months to restore. How do we know that won't creep into this system?

Responses:

Don Morgan - Sony

I agree that this very large-scale, large-magnitude problem is likely to happen but the answer lies in a strategy. We have to assume that the system will come down. There has to be a strategy in place at the site to deal with the fact that data must move from one point in the site to another point before it is destroyed or moved from the source point. If the system goes down in the midst of moving data before it is completed, you still have the source data on a retrievable storage device. " To restore" implies that you have only one source of data and that you have lost its directory because of some power failure. You were lucky in having a restore capability that resides with that data on that large data set--that's good news. That it took six months to do the restore was unfortunate. UPS won't solve that problem but may provide for graceful system degradation.

Sam Cheatham - Storage Technology

StorageTek has several installations where they use remote library installations, remote data directories, and control data sets. They can access the control data set at the remote site. You can cross-couple these systems from hundreds of miles away, next door, or in the same installation if you choose. It depends on how much redundancy you want for offsite backup capability. You can use the Storage Tech library systems as independent systems, as a slave or just have a remote library with a control data set or together with a remote processor. The redundancy gives you the ability, if something happens locally, to pull in the backup data or dump it to another site.

Remarks:

Dr. Hariharan

When the inventories get very large (in the order of terabytes and petabytes), it is not always feasible or economical to think in terms of retaining complete backups. So, the impact of the bit-error rate on the survivability of such large holdings is still doubtful.

J. F. Berry - Panel Moderator

It is not yet clear what the solution to that problem is.

Sam Cheatham - Storage Technology

The routine today is the practice of doing full-volume backups on an infrequent basis and transaction backups routinely. So the user does not have to depend on full- volume retention for backup of the system to minimize that problem.

Question: **Kevin Howard - Exabyte**

Is it possible to avoid doing full backups on telemetry type data?

Response: **Steve Miller - SRI International**

Telemetry data coming down, yes you must back that up. But that's not an archive. Before you archive it, you're going to do some preprocessing on it. Very often there's a lot of redundancy that's in that stream that can be taken care of when you do that first level of processing.

Question: Henry Bodzin - Ford Motor Company

Any time someone offers a storage device that is capable of storing much more data than the previous one, I have to ask whether I've just bought myself an access time problem. Is this a particular worry?

Responses:

Norris Huse - Datatape

There are multiple access time questions in the library system. In the case of tapes, the full search time, from end to end, on a large cassette is only about 3 to 4 minutes.

Michael Riddle - GE Aerospace

In the helical scan format, there are trade-offs between the size of the cassettes and all of the media we are talking about: D-1, D-2, and D-3. They're available in multiple cassette size -- small, medium, and large.

Tracy Wood - Ampex

With the Ampex DST800 product, we have attempted to optimize in terms of cost of access basis. That is, it's a relatively small library in terms of number of cassettes (only 256) and four drives. In a library like that, under right operational conditions, you could access 75% of that data (representing 6 terabytes of data) in a 24-hour period. On the small DD-2 cassette, you can search end to end in about 30 seconds. Robot times are very quick. We also provide intermediate mode and load zones on the tape, so you do not have to rewind the tape before removing it from the drive as used to be the case. A lot of thought has been put into the DD-2 product to try to put the best foot forward as far as helical technology's concerned.

Sam Cheatham - Storage Technology

If you look at the problem in terms of megabytes per second, it puts a different perspective on the issue. Think of "search" time from a physical point of view. Most operating systems only understand starting from the front of the tape. We have mechanisms built into the Redwood format to facilitate search and disconnected search out onto the length of the media, downstream.

Comment: J. F. Berry - Panel Moderator

Those of us in DOD think we're going to have to look at how to do work differently. Because one of these medium cassettes is almost equal to half of the average disk capacity of most Cray installations, it's reasonably unlikely we're going to turn around and read the data into disk and then operate on it. There's a whole new paradigm we have to develop on how to process data with this type of technology. It simply does not exist.

Question: J. F. Berry - Panel Moderator

What are each of your perspectives on how you see your drives being used?

Responses:

Don Morgan - Sony Corporation

In a format situation in DD-1, for instance, the front end directory of tape does allow a server to rapidly access locations of data on tape from a server's point of view and map that to a position on the tape and do fast search to that. So we have reduced access time in many instances.

Richard Davis - Sony Corporations

We have a 700-terabyte system and three tape recorders. You can use one of the recorders playing back into your system and, with smart software, have the other ones searching for the coincident time you're seeking. The delay will be very minimal.

Don Morgan - Sony Corporation

In a disconnected search, either the host or a file server downloads any particular record or a file ID to the subsystem. The subsystem disconnects then from a file server or the host until it can present that data, either by reading it into buffer or presenting it for a live transaction, so it doesn't keep the channel tied up.

Question: **David Owen - ICI IMAGEDATA**

Would the panel comment for each of the three formats on the implications of frequent stop/start operations?

Responses:

Tracy Wood - Ampex

I consider the question more appropriate for the implementation of the drive, not the format itself. Each of us, I think, has implemented some sort of physical blocking structure on the tape that carves the data into minimum record sizes. In terms of hardware implementation, the wear and tear on the tape is a very strong function of exactly what kind of technology is used in the tape and the mechanism itself. In the tapes of our drives (VSC600, DD-2-type drive) we use a lubricated tape path and have the situation very well controlled within the acceleration/deceleration profile.

Sam Cheatham - Storage Technology

I mentioned earlier that we had demonstrated more than 50 hours dwell time on a single stripe. This is a tremendous amount of time, far beyond what is traditionally thought of as possible in a helical environment. We've demonstrated tens of thousands of passes in the same fashion as we have with linear recorders. It is primarily a function of the hardware implementation. We do not believe it would be a problem. You would be better off, however, using a mix of hardware (between linear and helical) for high levels of start/stop.

Don Morgan - Sony Corporation

With this class of machine, we all have anticipated the use of large data-set streaming applications. I believe that is the only way to get around the wear and tear that will occur in a large number of start/stops.

Comment - Dr. Hariharan

Regarding media, Klaus Peter has reported that certain 4-mm tapes, after 200 shuttles, show a distinctive signature at the point where the shuttle stops. He has not noticed this on an 8-mm drive yet, however. The work has not been done on 19-mm drives.

Question: Dale Lancaster - Convex Computer Corporation

It appears these drives and formats are very similar in terms of order of magnitude. Are there any obvious trade-offs between the formats that a customer or end user should be aware of in terms of deciding between DD-1 and DD-2?

Responses:

Norris Huse - Datatape

Differences that would concern a customer are more in the implementation of the machine than in the formats. For example, two of the manufacturers of D-1-type machines reach 400 megabytes/second, 50 megabytes/second; the density of D-2 machines is somewhat higher than the density of D-1-type machines.

Tracy Wood-Ampex

Ampex did a fair amount of work with both the D-1 and D-2 fields before deciding to focus its efforts in the D-2 area, especially for data storage. The technology platform itself (use of metal particle tape) was one of the main reasons for using the D-2 as a place to start. To provide a compatibility path in the technology platform, we felt that D-2 had much more growth in it.

Sam Cheatham - Storage Technology

In the terms of the D-3 environment, we chose the 1/2-inch form factor so we could maintain compatibility with the library base. Other factors to be considered include significant upward growth potential in both capacity and data rate for this technology as well as very gentle tape handling. Thus, we chose D-3 technology rather than choosing D-2 and modifying it for half-inch. Our first products in the 3480 cartridge form factor, first offering will be 20 gigabytes per cartridge uncompressed, followed by 35 gigabytes uncompressed.

Don Morgan - Sony Corporation

I regard the trade-off between high data rate and high areal density as the only differentiation between these formats.

Comment: Steve Miller - SRI

With all of these, if you're interested in high start/stop rates, you need to design a storage subsystem that has other devices in it, such as the National Storage Laboratory.

Sam Cheatham - Storage Technology

I agree. That was one of the reasons we have supported full coexistence. These Redwood devices will coexist in the same library with linear tape. Thus you can direct high start/stop activities toward linear tape and direct high-capacity tasks to the helical environment.

Comment: J. F. Berry Panel Moderator

Looking at this type of tape technology, it looks like there may be a good marriage between disk arrays and this tape in which one unloads to a disk array very rapidly and then does the equivalent of start/stop under that.

Response: Panel members

All enthusiastically agreed that this was a promising direction.

Question: Dale Lancaster - Convex

In a system of this sort, will data be able to flow from tape to disk without going through the host memory in a potential implementation?

Sam Cheatham - Storage Technology

We can do anything with money and time.

Question: J. F. Berry - Panel Moderator

DOD, in a move to control operator costs, has made a policy decision not to acquire any particular tape devices that are not robotically controlled. What is the adaptability of the robotic interfaces for your devices?

Responses:

Tracy Wood - Ampex

When Ampex started its library system development program we had to deal directly with various computer manufacturers and adapt to the environment that they made available. They will not be the same until some form of win-win standardization for manufacturers and customers both is developed.

Sam Cheatham - Storage Technology

We have chosen to develop, and make available, a broad spectrum of library devices, believing that the form factor compatibility of the storage medium is of paramount importance. We have de facto standardization of that business position with the 4400 system today, offering upgrades in performance to that device that are fully compatible as well as a line of smaller libraries, also interconnected and software supported, to deal with that issue.

Don Morgan - Sony Corporation

I agree that work as a cooperative effort is needed to provide some common interface in this area.

Comments:

Dale Lancaster - Convex Corporation

It would be very helpful if there were a standard way to get to a robot that had standard functionality from a vendor's perspective. It would lower our costs to our customers as well, and would improve the time to market to implement a new robot.

Sam Cheatham - Storage Technology

In the storage server environment, StorageTek now supports about 15 different attach environments in an open systems situation, heavily influenced by software.

Question: J. F. Berry - Panel Moderator

These recorders seem to require enhancements, modifications, and the like to be useful. Please comment.

Responses:

Tracy Wood - Ampex

Ampex is supporting the Convex environment.

Sam Cheatham - Storage Tek

Required software is designed to hook into the customer's operating system but to avoid changing it.

Question: Dave Isaac

What about the separation of control and data? How does that occur in these drives? What provision is made for it? How does one deal with the robotics and that separation as well?

Responses:

Tracy Wood - Ampex

The Ampex robotic system provides clear separation of the robotic control from the normal data path. That will be the operating mode for the storage library for the National Storage Laboratory.

Sam Cheatham - Storage Technology

Storage Technology also uses separate control and data paths. Very minimal communication needs to take place between the tape subsystem and the library.

Don Morgan - Sony

Sony also has completely separate control and data I/O for drives and the robotics.

Device	Media	Width	Thickness	Areal density	Storage lifetime	Durability
Exabyte 8500	Metal particle (MP)	8 mm cassette	13 μm	74 Mbits/in^2	10 years	300 passes
DAT DDS Data DAT	MP	4mm cassette	13 μm	114 Mbits/in^2	10 years	1000 passes
ID-1	Co-doped γ-Fe_2O_3	19 mm cassette	16 μm	31 Mbits/in^2	10 years	300 passes
DD-2	MP	19 mm cassette	16 μm & 13 μm	43 Mbits/in^2	10 years	> 100 passes
D-3	MP	12.6 mm cartridge	14 μm & 11 μm	86 Mbits/in^2	10 years	> 100 passes
DCRSi	Co-doped γ-Fe_2O_3	25.2 mm cassette	25 μm	27 Mbits/in^2	10 years	200 passes
VLDS	Co-doped γ-Fe_2O_3	12.6 mm cassette (SVHS)	16 μm	22 Mbits/in^2	10 years	> 100 passes
QIC 1350	Co-doped γ-Fe_2O_3	6.3 mm cassette	11 μm	4.8 Mbits/in^2	10 years	300 passes
IBM3490E	CrO_2	12.6 mm cartridge	38 μm	1.8 Mbits/in^2	15 years	10000 passes

Device	Recording format	Capacity (Gigabits)	Transfer rate (Mbits/sec)	Vendors
Exabyte 8500	helical scan	40	4	Exabyte/Sony
DAT DDS Data DAT	helical scan	10	1.5	HP, Sony, Hitachi, Caliper, Archive
ID-1	helical scan	770	64-400	Datatape, Sony, GE
DD-2	helical scan	1320	120	Ampex
D-3	helical scan	160	96-144	StorageTek, Panasonic
DCRSi	transverse scan	380	107	Ampex
VLDS	helical scan	112	16-32	Metrum
QIC 1350	longitudinal serpentine	10.8	4.8	Archive, Tandberg, CMS
IBM 3490E	longitudinal	3.2	24	IBM

P C Hariharan 920921

		ID-1	D-1	D-2	D-3
Coercivity	Oersted	850	850	1500	1500
Wavelength	μm	0.9	0.9	0.85	0.77
Track pitch	μm	45	45	39.1	20
Track length	mm	170	170	150	117
Azimuth	degree	15	0	15	20
Drum diameter	mm	111.0	75.0	96.4	76.0
Drum rotation	rpm	6648*	9000	5400	5400
Writing speed	m/sec	39.08*	36	27.3	23.79
Wrap angle	degree	185	270	188	291
Tape speed	mm/min	423.8*	286	131.7	83.28
Channel coding		8-9	S-NRZ	M^2	8-14
Error correction		Reed-Solomon	Reed-Solomon	Reed-Solomon	Reed-Solomon
Inner		RS(163,155)	RS(64,60)	RS(95,87)	RS(95,87)
Outer		RS(128,118)	RS(32,30)	RS(68,64)	RS(136,128)

*

* For the Sony DIR 1000 operating at 256 Mbps

Notes for panel discussion on high-performance helical scan magnetic recorders

Computer interfaces for High Performance Tapes

Physical Interface	Transfer Rate (sustained)
SCSI-1	2 MB/sec
Block Mux FIPS-61	3 MB/sec
VME	6-8 MB/sec
SCSI-2	10 MB/sec
ESCON	10 MB/sec
IPI	15 MB/sec
HIPPI	50 MB/sec

Potential Platforms which can use high performance tape storage

. Powerful workstations (scalar/fp)
- RS6000
- SGI
- DEC Alpha
- HP

. Multiprocessor Workstations (scalar/graphics)
- Sun
- SGI

. Massively Parallel Systems - Vector/scalar
- NCUBE
- Thinking Machines
- Intel
- Kendall Square

. Supercomputers (vector)
- Cray
- Convex
- NEC

Device	Media	Width	Thickness	Areal density	Storage lifetime	Durability
Exabyte 8500	Metal particle (MP)	8 mm cassette	13 μm	74 Mbits/in^2	10 years	300 passes
DAT DDS Data DAT	MP	4mm cassette	13 μm	114 Mbits/in^2	10 years	1000 passes
ID-1	Co-doped γ-Fe_2O_3	19 mm cassette	16 μm	31 Mbits/in^2	10 years	300 passes
DD-2	MP	19 mm cassette	16 μm & 13 μm	43 Mbits/in^2	10 years	> 100 passes
D-3	MP	12.6 mm cartridge	14 μm & 11 μm	86 Mbits/in^2	10 years	> 100 passes
DCRSi	Co-doped γ-Fe_2O_3	25.2 mm cassette	25 μm	27 Mbits/in^2	10 years	200 passes
VLDS	Co-doped γ-Fe_2O_3	12.6 mm cassette (SVHS)	16 μm	22 Mbits/in^2	10 years	> 100 passes
QIC 1350	Co-doped γ-Fe_2O_3	6.3 mm cassette	11 μm	4.8 Mbits/in^2	10 years	300 passes
IBM3490E	CrO_2	12.6 mm cartridge	38 μm	1.8 Mbits/in^2	15 years	10000 passes

Device	Recording format	Capacity (Gigabits)	Transfer rate (Mbits/sec)	Vendors
Exabyte 8500	helical scan	40	4	Exabyte/Sony
DAT DDS Data DAT	helical scan	10	1.5	HP, Sony, Hitachi, Caliper, Archive
ID-1	helical scan	770	64-400	Datatape, Sony, GE
DD-2	helical scan	1320	120	Ampex
D-3	helical scan	160	96-144	StorageTek, Panasonic
DCRSi	transverse scan	380	107	Ampex
VLDS	helical scan	112	16-32	Metrum
QIC 1350	longitudinal serpentine	10.8	4.8	Archive, Tandberg, CMS
IBM 3490E	longitudinal	3.2	24	IBM

P C Hariharan 920921

		ID-1	D-1	D-2	D-3
Coercivity	Oersted	850	850	1500	1500
Wavelength	μm	0.9	0.9	0.85	0.77
Track pitch	μm	45	45	39.1	20
Track length	mm	170	170	150	117
Azimuth	degree	15	0	15	20
Drum diameter	mm	111.0	75.0	96.4	76.0
Drum rotation	rpm	6648*	9000	5400	5400
Writing speed	m/sec	39.08*	36	27.3	23.79
Wrap angle	degree	185	270	188	291
Tape speed	mm/min	423.8*	286	131.7	83.28
Channel coding		8-9	S-NRZ	M^2	8-14
Error correction		Reed-Solomon	Reed-Solomon	Reed-Solomon	Reed-Solomon
Inner		RS(163,155)	RS(64,60)	RS(95,87)	RS(95,87)
Outer		RS(128,118)	RS(32,30)	RS(68,64)	RS(136,128)

•

*U.S. GOVERNMENT PRINTING OFFICE: 1993-728-150/60026

* For the Sony DIR 1000 operating at 256 Mbps

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	April 1993	Conference Publication, September 22-24, 1992

4. TITLE AND SUBTITLE
Goddard Conference on Mass Storage Systems and Technologies
Volume II

5. FUNDING NUMBERS
902

6. AUTHOR(S)
Ben Kobler and P. C. Hariharan, Editors

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS (ES)
Goddard Space Flight Center
Greenbelt, Maryland 20771

8. PEFORMING ORGANIZATION REPORT NUMBER
93B00038
Code 902

9. SPONSORING / MONITORING ADGENCY NAME(S) AND ADDRESS (ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSORING / MONITORING ADGENCY REPORT NUMBER
NASA CP-3198, Vol. II

11. SUPPLEMENTARY NOTES
Kobler: Goddard Space Flight Center, Greenbelt, MD;
Hariharan: STX Corporation, Lanham, MD.

12a. DISTRIBUTION / AVAILABILITY STATMENT
Unclassified - Unlimited
Subject Category 82

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report contains copies of nearly all of the technical papers and viewgraphs presented at the Goddard Conference on Mass Storage Systems and Technologies held in September 1992. Similar to last year's conference, this year's gathering served as an informational exchange forum for topics primarily relating to the ingestion and management of massive amounts of data and the attendant problems (data ingestion rates now approach the order of terabytes per day). Discussion topics include the IEEE Mass Storage System Reference Model, data archiving standards, high-performance storage devices, magnetic and magneto-optic storage systems, magnetic and optical recording technologies, high-performance helical scan recording systems, and low end helical scan tape drives. Additional discussion topics addressed the evolution of the identifiable unit for processing purposes (file, granule, data set or some similar object) as data ingestion rates increase dramatically, and the present state of the art in mass storage technology.

14. SUBJECT TERMS
Magnetic tape, magnetic disk, optical disk, mass storage, software storage

15. NUMBER OF PAGES
296

16. PRICE CODE
A13

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

Space Administration
Code JTT
Washington, D.C.
20546-0001

Official Business
Penalty for Private Use, $300

SPECIAL FOURTH-CLASS RATE
POSTAGE & FEES PAID
NASA
PERMIT No. G27

NASA

POSTMASTER: If Undeliverable (Section 158 Postal Manual) Do Not Return