NASA Conference Publication 3165, Vol. III

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

Volume III

Proceedings of a conference held at
NASA Goddard Space Flight Center
Greenbelt, Maryland
1991

NASA Conference Publication 3165, Vol. III

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

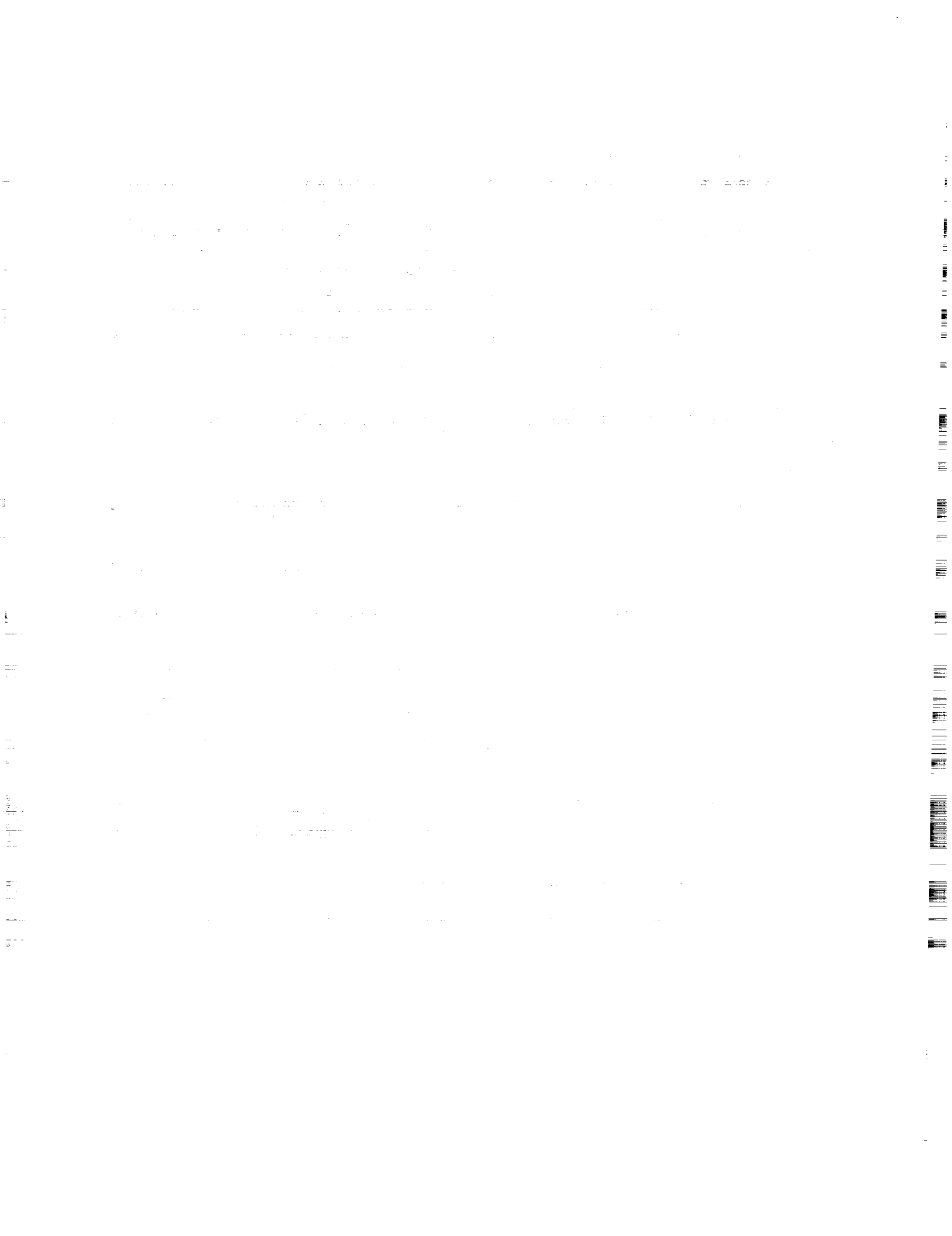
Volume III

Edited by
Ben Kobler
Goddard Space Flight Center
Greenbelt, Maryland

P. C. Hariharan
and L. G. Blasso
STX Corporation
Lanham, Maryland

Proceedings of a conference held at
NASA Goddard Space Flight Center
Greenbelt, Maryland
July 23-25, 1991

NASA
National Aeronautics and
Space Administration
Office of Management
Scientific and Technical
Information Program
1992



Preface

The National Space Science Data Center (NSSDC) at NASA's Goddard Space Flight Center has been charged with the archiving of data collected from NASA's scientific spaceflight missions flown over the past 30 years. During this time NSSDC has accumulated an archive of several terabytes of data. In the coming years NASA will be generating this volume of data every few days or less. Thus, data storage media and systems become critically important to NASA if it is to successfully manage this data volume and to have a chance to transform these data into scientific knowledge.

NSSDC will play an important role in NASA's awareness of and exploitation of emerging mass storage systems, both at NSSDC and in the increasingly distributed NASA scientific data environment. For this reason, NSSDC organized a conference at Goddard in the summer of 1991 to review the status of and the outlook for data storage media and systems. Leading experts in each of several areas were invited to make presentations, and a highly informative conference transpired. In order that the record of that conference be preserved, this set of presentations is being published.

The Proceedings of the NSSDC Conference on Mass Storage Systems and Technologies for Space and Earth Science applications are published in four volumes, with each of the first three volumes containing the talks and presentations for that particular day. Discussions following some of the talks are collected in the fourth volume along with introductory biographical material on the speakers. Despite our best efforts, the questions and answers were sometimes inaudible to the transcriptionist. An effort was made to contact the participants to clarify the transcript, and we are grateful to the speakers who cooperated.

The success of an endeavor of this magnitude depends on the generous help and cooperation of the participants. We would like to record our gratitude to the speakers, the audience, and in particular, to the following individuals and organizations for their assistance:

The Program committee whose membership is listed in the front of each volume

The session chairs who kept the schedule on track:

Professor Bharat Bhushan of Ohio State University
Dr. Barbara Reagor of Bellcore
Dr. Robert Freese of Alphatronix
Mr. Patric Savage of Shell
Professor John C. Mallinson of Mallinson Magnetics
Dr. Kenneth Thibodeau of the National Archives and Records Administration

The members of the Panel Discussion and Professor Mark Kryder of Carnegie Mellon University who moderated the discussion

The hard-working crew from Westover Consulting

Dr. Dennis E. Speliotis of Advanced Development Corporation for his help with the transcript of the Panel Discussion

Dr. James L. Green and the National Space Science Data Center

Dr. J. H. King
Dr. P. C. Hariharan
Benjamin Kobler

NSSDC Mass Storage Conference Committee

Ben Kobler, *NASA/GSFC (Chair)*
John Berbert, *NASA/GSFC*
Sue Kelly, *DOE/Sandia*
Elizabeth Williams, *SRC/Bowie*
Al Dwyer, *STX Corporation* *
P. C. Hariharan, *STX Corporation* *
Sanjay Ranade, *STX Corporation* *

Conference Coordinator

Kim Blackwell, *STX Corporation* *

* Hughes STX Corporation as of October 1, 1991

MASS STORAGE CONFERENCE PROCEEDINGS

CONTENTS

VOLUME I *- FIRST DAY - TUESDAY, JULY 23, 1991*

CONTENTS (Continued)

VOLUME II *- SECOND DAY - WEDNESDAY, JULY 24, 1991*

CONTENTS (Continued)

VOLUME III *- THIRD DAY - THURSDAY, JULY 25, 1991*

CONTENTS (Continued)

VOLUME III - *TRANSCRIPTIONS*

TRIBOLOGY OF MAGNETIC STORAGE SYSTEMS

Bharat Bhushan

Ohio Eminent Scholar Professor

Director, Computer Microtribology and Contamination Laboratory

Department of Mechanical Engineering

Ohio State University

Columbus, Ohio 43210

1.0 INTRODUCTION

1.1 Physics of Magnetic Recording

Recording technology is founded on magnetism and on electromagnetic induction. It is well known that currents in electric wires produce a magnetic field that can magnetize a hard magnetic material permanently. This permanent magnet can, in turn, generate an electric voltage: if it is dropped through a coil, it will generate a voltage pulse. We have just described the basic principle of magnetic recording (writing) and playback (reading), and it is shown schematically in Fig. 1, which consists of the relative motion between a magnetic medium and a read/write ring head (Lowman, 1962; Hoagland, 1963; Camras, 1988; Jorgensen, 1988; Mee and Daniel, 1990). Read/write heads (inductive type) consist of a ring of high-permeability magnetic material with an electrical winding and a gap in the magnetic material at or near the surface of the storage medium. Writing is accomplished by passing a current through the coil. The flux is confined to the magnetic core, except in the region of the small nonmagnetic gap. The fringe field in the vicinity of the gap, when sufficiently strong, magnetizes the medium, moving past the write head. The magnetic medium consists of high-coercivity magnetic material that retains its magnetization after it has passed through the field from the write head gap. The medium passes over the read head, which, like the write head, is a ring core with an air gap. Each particle in the medium is a miniature magnet, and its flux lines will add up with those of the other particles to provide an external medium flux, proportional in magnitude to the medium magnetization. The flux lines from the medium permeate the core and induce a voltage in the head winding. This voltage, after suitable amplification, reproduces the original signal. A single head can be used for both read and write functions.

More recently, some read heads are of the magnetoresistive (MR) type in which a strip of a ferromagnetic alloy (for example, $Ni_{80}Fe_{20}$) is mounted vertically. The variation of the magnetic-field component in the magnetic medium (perpendicular to the plane of medium) causes the variation in the electrical resistance of the MR stripe which can be readily measured (Van Gestel et al., 1977). MR-type read heads are attractive because they can be miniaturized without reducing the sensitivity to an unacceptably low value. One disadvantage of the MR-type head over the inductive-type head is that both read and write functions cannot be combined in one head (Bhushan, 1990).

So far, we have described longitudinal (horizontal) recording. In 1977, perpendicular (vertical) recording was proposed for ultrahigh density magnetic recording by Iwasaki and Nakamura (1977). In vertical recording, magnetization is oriented perpendicular to the plane of medium rather than in its plane. Single-pole heads are generally used for recording on vertical media. Ring-shaped heads, previously discussed, have also been used with some design changes. Vertical recording has the advantage of reduced self-demagnetization (Bhushan, 1990).

For high areal recording density, the linear flux density (number of flux reversals per unit distance) and the track density (number of tracks per unit distance) should be as high as possible. Reproduced (read) signal amplitude decreases with a decrease in the recording wavelength and/or the track width. The signal loss occurs from the magnetic coating thickness, head-to-medium spacing (clearance or flying height) and read gap length. The spacing loss of interest here was first described by Wallace (1951),

$$S(\lambda) = \exp(-2\pi d/\lambda) \text{ or } -54.6\, d/\lambda \text{ dB} \quad (1)$$

where λ is the recording wavelength and d is the head-medium separation. We note from Eq. 1 that the spacing loss can be reduced exponentially by reducing the separation between the head and the medium. The noise in the reproduced signal needs to be minimized. Thus signal-to-noise ratio (SNR) must be as high as possible. The wide-band SNR is also dependent upon head-to-medium spacing.

1.2 Importance of Tribology

We have just stated that the magnetic recording process is accomplished by relative motion between magnetic media against a stationary (audio and data processing) or rotating (audio, video and data processing) read/write magnetic head. The heads in modern magnetic storage systems are designed so that they develop hydrodynamic (self-acting) air bearing under steady operating condition. Formation of air bearing minimizes the head-medium contact. Physical contact between the medium and the head occurs during starts and stops. In modern high-end data-storage tape and disk drives, the head-to-media separation is on the order of 0.1 to 0.3 μm, and the head and medium surfaces have roughness on the order of 2 to 10 nm rms (Bhushan, 1990, 1992a). The need for higher and higher recording densities requires that surfaces be as smooth as possible and flying height be as low as possible. Smoother surfaces lead to increase in adhesion, friction, and interface temperatures and closer flying heights lead to occasional rubbing of high asperities and increased wear. Friction and wear issues are resolved by appropriate selection of interface materials and lubricants, by controlling the dynamics of the head and medium, and the environment. A fundamental understanding of the tribiology of magnetic head medium interface becomes crucial for the continued growth of the magnetic storage industry which is currently a $50 billion a year industry.

This chapter starts out to define the construction and the materials used in different magnetic storage devices. It then presents the theories of friction and adhesion, interface temperatures, wear, and solid-liquid lubrication relevant to magnetic storage systems. Experimental data are presented wherever possible to support the relevant theories advanced in this chapter.

2.0 MAGNETIC STORAGE SYSTEMS

2.1 Examples of Modern Systems

Magnetic Storage devices used for information (audio, video and data processing) storage and retrieval are: tape, flexible disk and rigid disk drives (Bhushan, 1990, 1992a).

2.1.1 Tape Drives

Linear analog technique is most commonly used for most domestic audio recorders which use typically a tape with a 4 mm width. There is always a physical contact between the head and the tape. Rotary single-track heads developed for video recording are also used for high recording density capabilities. For professional recording, dedicated digital technologies have emerged based, again, on either a multi-track stationary-head (S-DAT) approach or a rotating-head (R-DAT) approach.

A video recorder uses a helical-scanning rotating-head configuration. Because a rotating head can be moved at a greater speed than a heavy roll of tape, much higher data rates can be achieved in a rotating head drive than a linear tape drive. A 12.7-mm or 8-mm wide tapes are most commonly used. For professional video recording, digital video recorders are used for high signal-to-noise ratio.

Digital tape drives are used for data-processing (computer) applications. Figure 2 shows a schematic of the high-density, high-data-rate, IBM 3480/3490 tape drive for mainframe computers. For this drive, a 165-m long, 12.7-mm wide, and roughly 26-μm thick particulate tape wound on a reel is housed inside a rectangular cartridge (100 x 125 x 25 mm). A schematic of the 18-track read-write thin-film head is shown in Fig. 3. The write head is an inductive type and the read head is a MR type. The bleed slots are provided to reduce the flying height for maximum reproduced amplitude. Edge slots are provided for flying uniformity.

Helical scanning rotary-head configuration in a 8-mm tape format similar to video recorders and in a 4-mm tape format same as R-DAT audio recorders are used in tape drives; 8-mm format is used for very high volumetric density for mid-range computers (work stations). Drives using 130-mm full height form factors are commonly used. Belt-driven (longitudinal) data cartridges are commonly used for relatively high capacity in mid-range computers. A data cartridge tape is a self-contained reel-to-reel tape deck without a motor or read-write head. The tape width is 6.35-mm. The overall dimensions of the cartridge are 152 mm x 102 mm x 17 mm and 80 mm x 61 mm x 15 mm for 130-mm half-height and 95-mm half-height form factors, respectively. The 4-mm reel-to-reel data cassettes similar to audio cassette recorders are used for relatively small computers. These are also packaged in 95-mm and 130-mm half-height form factors.

2.1.2 Flexible Disk Drives

Flexible disk, also called a floppy disk or diskette is a magnetic recording medium, which is physically a thin (~82-μm thickness) and pliable disk and functionally a removable, random-access cartridge. When mounted in the drive, the disk is clamped at its center and rotated at a relatively low speed while the read-write head accesses the disk through a slot in the jacket. In most designs the accessing arms traverse above and below the disk and with read-write head elements mounted on spring suspensions. Head positioning is usually accomplished by a stepping motor, Fig. 4. The flexible disk heads are either spherically contoured or are flat in shape. Most commonly used disk drives today are in 90-mm (3.5 in.) and 130-mm (5.25 in.) form factors which use disk diameters, of 85.8 mm and 130.2 mm, respectively. The 50-mm (2 in.) form factor drives use 47-mm or 50.8-mm diameter disks and these drives are used in "notebook" computers. The 90-mm form factor disks use a metal hub and are encased in a hard plastic jacket which does not bend like the 130-mm form factor soft jackets and incorporates a shutter to protect the disk surface. A so-called "Bernoullie" disk drive has been developed by Iomega Corp., Roy, Utah to obtain the stable interface at higher operating speeds. In this design, the flexible disk is rotated at high speed in close proximity to a fixed flat plate called the Bernoullie plate. A foundation stiffness results from hydrodynamic loading (air bearing) created by spinning the disk in close proximity to the base plate. A foundation stiffness results from hydrodynamic loading (air bearing) created by spinning the disk in close proximity to the base plate. The high disk stiffness pushes up the first critical speed. Therefore, these disks can he operated at higher speeds (> 1500 rpm) without disk flutter.

2.1.3 Rigid Disk Drives

The rigid disk drive technology commonly referred to as "Winchester" technology utilizes the rigid disks as a nonremovable stack in a drive. The disks rotate at a constant angular speed, with concentric data tracks recorded on their surfaces. The heads are moved by an actuator that positions each head over a desired data track. Typically one or two heads (for fast accessing) are moving over each disk surface. While all heads are actuated together, only one head is selected at a time to read or write. The schematic of the head-disk interface of a high density, high-data rate, IBM 3390-type disk drive with a linear actuator driven by a voice coil motor is shown in Fig. 5.

Fairly large motors are used to drive the precision spindle holding the disks. A high starting torque is needed in order to overcome the static friction from often many heads resting on the disks. Hence brushless DC motors are used, mounted inside the spindle or underneath the base plate casting. Preloaded and sealed ball bearings are used to support the spindle shaft in both ends.

Conventionally, a slider is mounted on a flexure in the orientation optimal for linear actuators. The longitudinal axis of flexure points is in the direction of carriage actuation, with

the slider mounted at a right angle. However, current trends are toward smaller, more compact disk storage devices, especially in the low-end applications. The compact, low mass, low-cost rotary actuators are used to save space in the drive. In rotary actuators, the slider is mounted along the rotary arm. The actuator is operated by a stepping motor or voice-coil motor (VCM), Figs. 4 and 5. The VCM is very much like a loudspeaker coil/magnet mechanism. The VCM has the advantage of providing the desired linear or the rotary motion directly whereas circular motion provided by the stepping motor needs to be converted to the desired linear or rotary motion generally by a lead screw or capstan-band (Fig. 4) method. Furthermore, the VCM provided faster and smaller stepping than that by stepping motor. The small drives use both linear and rotary actuators driven by either stepping or VCM motor. The large drives use a linear actuator driven by a VCM. The actuator connected to the VCM, rides on a set of ball hearings on the tracks as shown in fig. 5.

The head slider used in high-end rigid disk drive (IBM 3380K/ 3390) is a two-shaped rail, taper-flat design supported by a leaf spring (flexure) suspension, made of nonmagnetic steel to allow motion along the vertical, pitch and roll axes, Fig. 6. The front taper serves to pressurize the air lubricant, while some of the air is lost through leakage to the side boundaries of the rail resulting in a pitch angle. In the shaped-rail design, each side rail has a widened leading-edge rail width which is flared down to a smaller rail width towards the trailing end, Fig. 6(a). Shaped-rail design is used to attain increased pitch angles, independent of airfilm thickness. The inductive-type thin-film read-write elements, located at the trailing edge of each rail, are an integral part of the slider where the lowest flying height occurs. Suspension supplies a vertical load of either 100 mN (10 g) or 150 mN (15 g) which is balanced by the hydrodynamic load when the disk is spinning. The stiffness of the suspension (~25 mN mm^{-1}) is several orders of magnitude lower than that of the air bearing (~0.5 kN mm^{-1}) developed during use so that most dynamic variations are taken up by the suspension without degrading the air bearing.

Small disk drives use inductive-coil type heads. Two types of head sliders are most commonly used: minimonolithic (mini-Winchester) and minicomposite. A minimonolithic head slider consists of a slider body and a core piece carrying the coil, both consisting of monolithic magnetic material. It is a tri-rail design. The taper-flat bearing area is provided by the outer two rails of the tri-rail design. The center rail defines the width of the magnetic element in the trailing edge where a ferrite core is formed. A minimonolithic head slider consists of a Mn-Zn ferrite core with read-write gap, glass bonded into air-bearing surface of a nonmagnetic, wear-resistant slider (typically calcium titanate) of approximately the same size as minimonolithic slider. The 3380-type suspensions are normally used for heads in small drives and apply a 95 mN (9.5-g) load onto the slider.

The 3380-K/3390 type sliders are about 4.045-mm long by 3.200-mm wide by 0.850 mm high with a mass of 0.45 mN (45 mg), as opposed to about 4.1 mm long by 3.1-nun wide by 1.4 mm high and 0.7 mN (70 mg) for the mini Winchester. The surface roughness of the air-bearing rails is typically 1.5-2.5 nm rms.

Schematic representations of head-medium interfaces for tape, flexible, and rigid-disk drives are shown in Fig. 7. We note that the environment, usage time, and contamination (external and wear debris) play a significant role in the reliability of the interface.

2.2 Magnetic Head and Medium Materials

2.2.1 Magnetic Heads

Magnetic heads used to date are either conventional inductive or thin-film inductive and magnetoresistive (MR) heads. Trends to film-head design have been driven by the desire to capitalize on semi-conductor-like processing technology to reduce fabrication costs. In addition, thin-film technology allows the production of high-track density heads with

accurate positioning control of the tracks and high reading sensitivity. Conventional heads are combination of a body forming the air bearing surface and a magnetic ring core carrying the wound coil with a read-write gap. In the film heads, the core and coils or MR stripes are deposited by thin-film technology. The air-bearing surfaces of tape heads are cylindrical in shape. The tape is slightly underwrapped over the head surface to generate hydrodynamic lift during read-write operation. For inductive-coil tape heads, the core materials that have been typically used are permalloy and Sendust. However, since these alloys are good conductors, it is sometimes necessary to laminate the core structure to minimize losses due to eddy currents. The air-bearing surfaces of most inductive-coil type heads consist of plasma sprayed coatings of hard materials such as Al_2O_3-TiO_2 and ZrO_2. Read and write heads in modern tape drives (such as IBM 3480/3490) are miniaturized using thin-film technology, Fig. 3. Film heads are generally deposited on Ni-Zn ferrite (11 wt % NiO, 22 wt % ZnO, 67 wt % Fe_2O_3) substrates. Flexible-disk heads are inductive-coil type composite heads which are spherically contoured or are flat in shape. Mn-Zn ferrite (30 wt % MnO, 17 wt % ZnO, 53 wt % Fe_2O_3) is generally used for head cores and barium titanate is generally selected for the magnetically inert structures which support the cores.

The material used in the construction of thin-film (Winchester-type) head used in large disk drives is generally (nonmagnetic) Al_2O_3-TiC (70-30 wt %). Some manufacturers use yttria-stabilized zirconia/alumina — titanium carbide composite. Small rigid-disk drives for low-end applications use heads with magnetic ring core and a wound coil. Two types of head sliders are most commonly used: minimonolithic (or mini-Winchester) and minicomposite. A minimonolithic head slider consists of a slider body and a core piece carrying the coil, both consisting of monolithic magnetic material, typically Mn-Zn ferrite. A minicomposite head slider consists of a Mn-Zn ferrite core with read-write gap, glass bonded into the airbearing surface of a nonmagnetic, wear-resistant slider (typically calcium titanate). Typical physical properties of hard head materials are presented in Table 1.

2.2.2 Magnetic Media

Magnetic media fall into two categories: particulate media, where magnetic particles are dispersed in a polymeric matrix and coated onto the polymeric substrate for flexible media (tape and flexible disks) or onto the rigid substrate (typically aluminum, more recently introduced glass for rigid-disks); thin-film media, where continuous films of magnetic materials are deposited onto the substrate by vacuum techniques. Requirements of higher recording densities with low error rates have resulted in an increased use of thin films which are smoother and considerably thinner than the particulate media. The thin-film media is extensively used for rigid-disks and has begun to be used for high.density audio/video and data processing tape applications.

Sectional views of a particulate and a thin-film (evaporated) metal tape are shown in Fig. 8. Flexible disks are similar to tapes in construction except these are coated with magnetic coating on both sides and the substrate is generally about 76.2 μm in thickness. The base film for flexible media is almost exclusively poly(ethylene terephthalate) (PET) film. Other substrate materials such as polyimides and polyamide-polyimide copolymers have been explored for better dimensional stability of flexible disks. Typically a 6.35 to 36.07-μm thick (25, 35, 57, 88, 92, and 142 gage) PET substrate with an rms roughness of about 1.5 to 2.5 nm for particulate media and even smoother (1.5 to 2 nm rms) for thin-films media is used for tapes (14.48-μm or thicker for data processing applications) and 76.2-μm thick (300 gage) for flexible disks. The base film is coated with a magnetic coating, typically 2 to 4 μm thick (coated on both sides with a magnetic coating, in case of flexible disks), composed by acicular magnetic particles [such as γ-Fe_2O_3, Co-modified γ-Fe_2O_3, CrO_2 (only for tapes), and metal particles for horizontal recording or hexagonal platelets of barium ferrite for both horizontal and vertical recording], polymeric binders [such as polyester-polyurethane, polyether-polyurethane, nitrocellulose, poly(vinyl chloride), poly(vinyl alcohol-vinyl acetate), poly(vinylidene chloride), VAGH, phenoxy, and epoxy, lubricants (mostly fatty acid esters, e.g., tridecyl stearate, butyl stearate, butyl palmitate, butyl myristate, stearic acid, myrstic acid), a cross

linker or curing agent (such as functional isocyanates), a dispersant or wetting agent (such as lecithin), and solvents (such as tetrahydrofuran and methyl isobutylketone). In some media, carbon black is added for antistatic protection if the magnetic particles are highly insulating and abrasive particles (such as Al_2O_3) are added as a head cleaning agent and to improve media's wear resistance. Magnetic particles are typically 70-80% by weight (or 43-50% by volume), lubricants are typically 1-3% by weight of the total solid coating. The coating is calendered to a surface roughness of 8 to 15 nm rms.

Most magnetic tapes have a 1 to 3-μm thick backcoating for antistatic protection and for improved tracking. The backcoat is generally a polyester-polyurethane coating containing conductive carbon black and TiO_2. The TiO_2 and carbon contents are typically 50% and 10% by weight, respectively.

Flexible disks are packaged inside a soft polyvinyl chloride (PVC) jacket or a acrylonitrile-butadiene-styrene (ABS) hard jacket (for smaller 90-mm form factor). Inside the jacket, a soft liner, a protective fabric is used to minimize wear or abrasion of the media. The wiping action of the liner on the medium coating removes and entraps particulate contaminants which may originate from the diskette manufacturing process, the jacket, head-disk contact (wear debris), and the external environment. The liner is made of nonwoven fibers of polyester (PET), rayon, polypropylene or nylon. The liner fibers are thermally or fusion bonded to the plastic jacket at spots. The soft jacket near the data window is pressed with a sponge pad to create a slight friction and hence stabilize the disk motion under the heads. The hard cartridge is provided with an internal plastic leaf spring for the same purpose.

Thin-film (also called metal-film) flexible media consist of polymer substrate (PET or polyimide) with an evaporated film of Co-Ni (with about 18% Ni) and less commonly evaporated/sputtered Co-Cr (with about 17% Cr) for vertical recording, which is typically 100-300 nm thick. Electroplated Co and electroless plated Co-P, Co-Ni-P and Co-Ni-Re-P have also been explored but are not commercially used. Since the thickness of the magnetic layer is only 100-300 nm, the surface of the thin-film medium is greatly influenced by the surface of the substrate film. Therefore, an ultra-smooth PET substrate film (rms roughness ~1.5-2 nm) is used to obtain a smooth tape surface (rms rougliness ~5-6 nm) for high-density recording. Several undercoatings, overcoatings and oxidation treatments (by adding oxygen into the vacuum chamber during evaporation) are used to increase corrosion resistance and durability. An undercoating layer has been employed consisting of very fine particles, each protruding few nanometers. Various organic (such as acrylic polymer in 50-80 nm thickness) and inorganic overcoats (such as diamondlike carbon in about 10-20 nm thickness, SiO_2 and ZrO_2) have been used to protect against corrosion and wear. Alumina (0.2 μm) particles are also added in some polymeric overcoats. Yet in another approach oxygen is leaked into the vacuum chamber at a controlled rate during the Co-Ni evaporation. This has the effect of forming an oxide top layer of 10 to 30 nm, and probably introduces oxidized grain boundaries. This oxidation improves durability and corrosion resistance and decreases magnetic noise. Surface oxidation of evaporated Co-Cr coating by annealing in air has also been reported to improve the durability of the coating. In addition to a solid overcoat, generally a thin layer (2-10 nm in thickness or 4-20 mg/m^2 in weight) liquid lubricant is applied on the medium surface to further reduce friction, wear and corrosion. Fatty acid esters or fluorine-based compounds (such as perfluoropolyethers) are most commonly used (Bhushan, 1992a).

Figure 9 shows sectional views of two types of rigid disks - a particulate disk and a thin-film disk. The substrate for rigid disks is generally a non heat-treatable aluminum-magnesium alloy AlSi 5086 (95.4% Al, 4% Mg, 0.4% Mn, and 0.15% Cr) with an rms surface roughness of about 15-25 nm rms and a Vickers hardness of about 90 kg/mm^2. For particulate disk, the Al-Mg substrate with a thickness of 1.3 to 1.9 mm, is sometimes passivated with a very thin (< 100 nm) chromium conversion coating based on chromium phosphate. The finished substrate is spin coated with the magnetic coating about 0.75 to 2 μm thick and burnished to a surface roughness of about 7.5 to 15 nm rms. The binder system generally used is a hard copolymer of epoxy, phenolic and polyurethane constituents. About 30 to 35% by

volume of acicular magnetic particles of γ-Fe_2O_3 are interdispersed in the binder. A small percentage (2 to 8% by volume) of Al_2O_3 particles 0.2-0.5 μm in size are added to improve the wear resistance of the magnetic coating. A thin film of perfluoropolyether lubricant (3-6 nm thick) is applied topically. The magnetic coating is made porous for lubricant retention (Bhushan, 1990).

For high-density recording, trends are to use thin-film disks. For thin-film disks with a metallic magnetic layer, the Al-Mg substrate with a thickness of 0.78 to 1.3 mm, is electroless plated with a 10- to 20-μm thick nickel-phosphorous (90-10 wt %) layer to improve its surface hardness to 600-800 kg/mm^2 (Knoop) and smoothness. The coated surface is polished with an abrasive slurry to a surface roughness of about 2 to 4 nm rms. For a thin-film disk with an oxide magnetic layer, a 2- to 20-μm thick alumite layer is formed on the Al-Mg substrate through anodic oxidation in a CrO_3 bath. In this application, Ni-P cannot be used because it becomes magnetic when exposed to high temperature during preparation of γ-Fe_2O_3 film. Start-stop zone of substrates for thin-film are generally textured mechanically in the circumferential or random orientation to a rms roughness typically ranging from 6-8 nm rms in order to minimize static friction at the head-disk interface. For convenience, the entire disk is generally textured. Circumferential direction of texturing in the data zone (if textured) is preferred in order to keep the magnetic orientation ratio as high as possible. The finished substrate is coated with a magnetic film 25-150 nm thick. Some metal films require a Cr undercoat (10 to 50 nm thick) as a nucleation layer to improve magnetic properties, such as coercivity. Typically, magnetic films that have been explored are metal films of cobalt-based alloys, with sputtered iron oxide being the principal exception. Magnetic films that are used to achieve the high recording density have weak durability and poor corrosion resistance. Protective overcoats with a liquid lubricant overlay are generally used to provide low friction, low wear, and corrosion resistance. Protective coatings, ranging in thickness from 20-40 nm, generally used are sputtered diamondlike carbon, spin coated or sputtered SiO_2, sputtered yttria-stabilized zirconia and plasma-polymerized protective (PPP) films. In most cases, a thin layer of perfluoropolyether lubricant (1-4 nm thick) is used (Bhushan, 1990).

Typical materials used for various magnetic media and operating conditions for data-processing applications are shown in Table 2. Typical physical properties of components of magnetic media are presented in Table 3.

3.0 FRICTION AND ADHESION

When two surfaces come in contact under load, the contact takes place at the tips of the asperities and the load is supported by the deformation of the contacting asperities, Fig. 10. The proximity of the asperities results in adhesive contacts caused by interatomic attractions. In a broad sense, adhesion is considered to be either physical or chemical in nature. Experimental data suggest that adhesion is primarily due to weak van der Walls forces. When the two surfaces (in contact) move relative to each other, frictional force, commonly referred to as "intrinsic" or "conventional" frictional force, is contributed by adhesion and deformation (or hysteresis). For most practical cases, adhesional friction is the primary contributor (Bhushan, 1990).

In addition, "stiction" can occur due to meniscus/viscous effects, microcapillary evacuation, and changes in surface chemistry (Bhushan et al., 1984a, 1984b; Bradshaw and Bhushan, 1984; Bradshaw et al., 1986). Here we will concentrate on the meniscus/viscous effects only. Generally, any liquid that wets or has a small contact angle on surfaces will condense from vapor in the form of an annular-shaped capillary condensate in the contact zone. The pressure of the liquid films of the capillary condensates or preexisting film of lubricant can significantly increase the adhesion between solid bodies. Liquid-mediated adhesive forces can be divided into two components: meniscus force (F_M) due to surface tension and a rate-dependent viscous force (F_V). The total tangential force F required to

separate the surfaces by sliding is equal to an intrinsic force (F_A) and stiction force (F_S) (combination of friction force due to meniscus effect and the peak viscous force)

$$F = F_A + F_S = f_r (W + F_M) + F_V \quad (1)$$

where f_r is true static coefficient of friction and W is the normal load.

Our analysis shows that normal force required to move two flat, well polished surfaces (such as magnetic head and medium surfaces) apart in the presence of liquid medium and/or sticky substance, can be large (up to several N in extreme cases). Therefore, we define the difference between stiction and conventional static and kinetic friction being that stiction requires a measurable normal force (normally several mN or higher) to pull the two surfaces apart from the static conditions.

3.1 Conventional Friction

From Tabor's classical theory of adhesion, frictional force due to adhesion (F_A) is defined as follows (Bowden and Tabor, 1950):

$$\text{for dry contact, } F_A = A_r \tau_a \quad (2a)$$

$$\text{for lubricated contact, } F_A = A_r[\alpha\tau_a + (1 - \alpha)\tau_\ell] \quad (2b)$$

and

$$\tau_\ell = \eta_\ell V/h \quad (2c)$$

where A_r is the real area of contact, α is the fraction of unlubricated area, τ_a and τ_ℓ are the shear strengths of the dry contact and of the lubricant film, respectively, η_ℓ is the absolute viscosity of the lubricant, V is the relative sliding velocity, and h is the lubricant film thickness.

3.1.1 *Greenwood and Williamson's Contact Model*

The contacts can be either elastic or plastic which primarily depend on the surface topography and the mechanical properties of the mating surfaces. The classical model of elastic-plastic contact between rough surfaces is that of Greenwood and Williamson (1966) (G&W), which assumed the surface to be composed of hemispherical asperities of uniform size with their heights following a Gaussian distribution about a mean plane. The radius of these asperities is assumed to be equal to the mean radius of curvature that is obtained from roughness measurements.The expression for real area of contact for elastic (e) and plastic (p) contacts are (Bhushan, 1984),

$$A_{re} / A_a p_a \sim 3.2 / E_c (\sigma_p / R_p)^{1/2}$$
$$\text{or } \psi_p < 1.8 \text{ or } \psi < 0.6, \text{ elastic contact} \quad (3a)$$

$$A_{re} / p_a A_a = 1 / H$$
$$\text{or } \psi_p > 2.6 \text{ or } \psi > 1, \text{ plastic contact} \quad (3b)$$

and

$$\psi_p = (E_c/Y)\,(\sigma_p R_p)^{1/2}, \text{ for polymers} \quad (3c)$$

$$\psi = (E_c/H)\,(\sigma_p/R_p)^{1/2}, \text{ for metals/ceramics} \quad (3d)$$

where A_a is the apparent area of contact; p_a is the apparent pressure (W/A_a); E_c is the composite modulus of elasticity, H and Y are the hardness and yield strength of the softer material, σ_p and R_p are the composite standard deviation and radius of curvature of the surface summits, and ψ is the plasticity index. σ_p and R_p depend on the instrument resolution and hence are not unique.

Equation 3 for elastic and plastic contacts in the case of metals/ ceramics is plotted in Fig. 11 for better visualization of dependence of A_r on ψ. We note that the plastic contact results in a minimum contact area. However, repeated plastic contact would lead to an undesirable permanent deformation and smoothening resulting in elastic contacts (and a higher real area of contact). Wear is more probable when asperities touch plastically than in pure elastic contacts. Therefore, it is desirable to design components in the elastic contact regime with ψ close to the elastic contact limit ($\psi \sim 0.6$) or $E_c/(\sigma_p/R_p)^{1/2}$ to be as high as possible.

Bhushan (1984, 1985a), Bhushan and Doerner (1989), Bhushan and Blackman (1991) and Oden et al. (1992) have measured the mechanical properties and surface roughnesses of various particulate tapes and particulate and thin-film (metal and oxide) disks. Mechanical properties of magnetic coatings of tapes are measured by dynamic mechanical analysis (DMA) system and of rigid disks were measured by a nanoindentation hardness apparatus. Surface roughness parameters of tapes and disks were measured by a noncontact optical profiler (NOP) (Bhushan et al., 1988) and an atomic force microscope (AFM) (Rugar et al., 1989). Measured values are used in the Greenwood and Williamson's contact model. The lateral resolutions for the surface topographs spanned the range of 1 μm for NOP down to 2 nm for AFM. AFM can measure topographic features which cannot be measured with conventional profilers, Fig. 12. The NOP image of the magnetic tape in Fig. 12(a) does not show any distinctive feature. In contrast AFM image clearly shows magnetic particles. The hole in the center of disk A is the pore which is created in particulate disk which acts as reservoir for the lubricant, Fig. 12(b). The AFM has sufficiently high spatial resolution to image the grain structure of the sputtered coatings, Figs. 12(c) and 12(d). Topograph shown in Fig. 12(e) for a lapped slider shows that the surface is very smooth with grooves less than 1 nm deep. Bhushan and Blackman (1991) and Oden et al. (1992) found that topography and contact statistics predictions are a strong function of the lateral resolution of the roughness measurement tool, Table 4. Here η is the density of summits per unit area, n is the number of contact spots and p_r is the real pressure. The surface topography statistics calculated for the AFM and NOP data shows that the average summit radius (R_p) for the AFM data is two to four orders of magnitude smaller than that for the NOP data whereas summit density for the AFM data is two to four orders of magnitude larger than that for the NOP data. We note that the plasticity index (ψ) calculated using the AFM data suggests that all contacts made of nanoasperities are plastic, while ψ calculated with NOP data suggest that all contacts made of microasperities are elastic (see Fig. 16 to be described later).

3.1.2 *Fractal Model of Elastic-Plastic Contact*

The contact analyses developed over a last quarter century, consider only an averaged surface with a single scale of roughness to be in contact with another surface. However, due to the multiscale nature of surfaces it is found that the surface roughness parameters depend strongly on the resolution of the roughness-measuring instrument or any other form of filter, hence not unique for a surface (Bhushan et al., 1988). The dependence of variances of surface height, slope and curvature are shown in Fig. 13. The scale dependence in Fig. 13 suggests that

instruments with different resolutions and scan lengths yield different values of these statistical parameters for the same surface. Therefore the predictions of the contact models based on these parameters may not be unique to a pair of rough surfaces. However, if a rough surface is characterized in a way such that the structural information of roughness of all scales is retained, then it will be more logical to use such a characterization in a contact theory. In order to develop a contact theory based on this motivation, it is first necessary to quantify the multiscale nature of surface roughness.

A unique property of rough surfaces is that if a surface is repeatedly magnified, increasing details of roughness are observed right down to nanoscales. In addition, the roughness at all magnifications appear quite similar in structure as qualitatively shown in Fig. 14. Such a behavior can be characterized by fractal geometry (Majumdar and Bhushan, 1990, 1991b). The main conclusions from these studies were that a fractal characterization of surface roughness is scale-independent and provides information of the roughness structure at all the length scales that exhibit the fractal behavior. Based on this observation, Majumdar and Bhushan (1991a) and Bhushan and Majumdar (1991) developed a new fractal theory of contact between rough surfaces.

Consider a surface profile z(x) as shown in Fig. 14 which appears random, multiscale, and disordered. The mathematical properties of such a profile are that it is continuous, nondifferentiable and statistically self-affine. The nondifferentiability arises from the fact that a tangent or a tangent plane cannot be drawn at any point on the surface since more and more details of roughness will appear at the point. In short, a rough surface is never smooth at any length scale. The statistical self-affinity is due to similarity in appearance of a profile under different magnifications. It was found that the Weierstrass-Mandelbrot (W-M) function satisfies all these properties and is given as (Berry and Lewis, 1980),

$$z(x) = G^{(D-1)} \sum_{n=n_\ell}^{\infty} \frac{\cos(2\pi\gamma^n x)}{\gamma^{(2-D)n}}; \quad \gamma > 1, \quad 1 < D < 2 \tag{4a}$$

$$\frac{dz}{dx} \longrightarrow \infty \text{ for all } x; z(\gamma x) = \gamma^{(2-D)} z(x) \tag{4b}$$

where the parameter D is the fractal dimension, G is a characteristic length scale of the surface and γ^n are the discrete frequency modes of the surface roughness. The theory of fractal geometry and the concept of fractional dimension is well described by Mandelbrot (1982) and in the recent paper of Majumdar and Bhushan (1990). The parameters which characterize the W-M function are G, D. and η_ℓ where $\gamma = 1.5$ was found to be suitable value for high spectral density and for phase randomization. Since a rough surface is a nonstationary random process the lowest cut off frequency is related to the length L of the sample as $\gamma\eta_\ell = 1/L$. The parameters G and D can be found from the power spectrum of the W-M function

$$S(\omega) = \frac{G^{2(D-1)}}{2\ell n\ \gamma} \frac{1}{\omega^{(5-2D)}} \tag{5}$$

where S(ω) is the power of the spectrum and ω is the frequency which is the reciprocal of the wavelength of roughness. Note that if S(ω) is plotted as a function of ω on a log-log plot, then the power law behavior would result in a straight line. The slope of the line is related to the fractal dimension D of the surface roughness and the parameter G is related to the location of the spectrum along the power axis.

To verify whether surfaces do follow a power-law fractal behavior and to obtain the parameters D and G of a surface, one needs to compare the power spectrum of a real surface profile with that of the W-M function as given in Eq. (5). Figures 15(a) and 15(b) show the averaged spectrum of a surface profile of an untextured thin-film magnetic rigid disk of type C. The spectrum in Fig. 15(a) corresponds to the surface that was measured by an optical profiler and the spectrum in Fig. 15(b) corresponds to the surface which was measured by a scanning tunneling microscope (STM). The spectrum in Fig. 15(a) follows $S(\omega) = \omega^{-2.24}$ corresponding to $D = 1.38$ whereas the spectrum in Fig. 15(b) follows $S(\omega) \sim \omega^{-2.35}$ corresponding to $D = 1.33$ and $G \sim 10^{-16}$ m. This data show that the surface of the rigid disk follows a fractal structure for three decades of length scales.

Bhushan et al. (1988) and Oden et al. (1992) measured the surface roughness of magnetic tape A at different resolutions by NOP and AFM. Fractal analysis of the tape surface reveals two regimes of roughness demarcated by a scale of 0.1 μm corresponding to the size of magnetic particles.

The fractal model of elastic-plastic contact has been developed by Majumdar and Bhushan (1991a) and Bhushan and Majumdar (1991). An interface between a statistically isotropic rough surface and a flat plane was considered. The contact spots will be of different sizes and spread randomly over the interface. Depending on the radius of curvature and height (or deformation) of the asperity, the contact spot will be either in elastic or plastic deformation. Limit of elastic deformation (propensity of yielding) is governed by the Tresca or Huber-Mises yield criteria in which the plastic flow will occur when maximum shear stress is equal to half of the tensile yield strength of the material. Whether contacts go through elastic or plastic deformation is determined by a critical area

$$a_c = \frac{G^2}{(H / 2E)^{2/(D-1)}}$$

$$\text{If } a < a_c, \text{ plastic content}$$

$$a > a_c, \text{ elastic content} \qquad (6)$$

This shows that small contact spots ($a < a_c$) are in plastic contact, whereas large spots are in elastic contact. This result is in contrast with that of the G & W model where small ones are in elastic deformation — — a prediction that is a direct implication of the assumption of uniform asperity radii. For magnetic tape A, typical values of the parameters are $D = 1.97$, $G = 5.15 \times 10^{-9}$m and $H/E = 0.14$. The critical contact area for inception of plastic flow is $a_c = 10^{-14}$m^2 (contact diameter ~100 nm). Therefore, all contact spots larger than 100 nm would deform elastically. For a smooth magnetic thin-film rigid disk of type C, typical values of the parameters are $D = 1.38$, $G = 10^{-16}$m and $H/E = 0.06$. The critical contact area for inception of plastic deformation is $a_c = 10^{-27}$m^2, which is practically zero. Therefore, all contact spots can be assumed to be in elastic contact at moderate loads.

The question now remains as to how do large spots become elastic when they must have been small plastic spots in their history of deformation. The possible explanation is graphically shown in Fig. 16. As two surfaces touch, the nanoasperities (detected by AFM type of instruments) are first to come in contact. As the load is applied, the smaller asperities have smaller radii of curvature and are therefore plastically deformed instantly and the contact area increases. When the load is increased, the nanoasperities in the contact zone merge and the load is supported by elastic deformation of the larger scale asperities or microasperities (detected by NOP type of instruments).

It is assumed that cumulative size distribution of the contact spots follow the power law relation of the form (Majumdar and Bhushan, 1991a)

$$N(A>a) = (a_\ell / a)^{D/2} \tag{7}$$

where the distribution is normalized by the area of the largest contact spot a_ℓ. Since the power spectra of surface indicate that a surface can be fractal even at nanoscales, the assumption of $a_s \to 0$ is valid. Note that in the distribution of Eq. (7), the number of the largest spot is unity, whereas the number of spots of area $a \to 0$ would tend to infinity.

The real area of contact, A_r is given as

$$A_r = \frac{D}{2-D} a_\ell \quad \text{for } D < 2 \tag{8}$$

For the case $a_\ell > a_c$, the portion of the real area of contact in elastic deformation can be evaluated as

$$A_{re} / A_r = 1 - \left[\frac{D a_c}{(2-D)A_r}\right]^{(2-D)/2} \tag{9}$$

The total elastic-plastic load $W(a_\ell > a_c)$ is related to the real area of contact as

$$W / EA_a \sim G^{D-1} A_r^{D/2} \left\{\left[\frac{(2-D)A_r}{D}\right]^{(3-2D)/2} - a_c^{(3-2D)/2}\right\} + (H/E) A_r^{D/2} a_c^{(2-D)/2} \quad \text{for } D \neq 1.5 \tag{10a}$$

and

$$W / EA_a \sim G^{1/2} A_r^{3/4} \ln(A_r / a_c) + (H / E) A_r^{3/4} a_c^{1/4} \quad \text{for } D = 1.5 \tag{10b}$$

and the total plastic load $(a_\ell < a_c)$

$$W/HA_a \sim A_r/A_a \tag{11}$$

We note that for the special case of $a_c \to 0$ (e.g., in thin-film disk C), in the elastic-plastic regime

$$W \sim A_r^{(3-D)/2} \tag{10c}$$

Here, the load-area relationship depends on the fractal dimension whereas G&W predict a linear relationship. Fractal model verifies the load-area relationship observed by Bhushan (1985a) for the magnetic tape A and by Bhushan and Dugger (1990a) for the thin-film disk C (Majumdar and Bhushan, 1991a).

3.1.3 *Measurement of Contact Area*

The real area of contact of magnetic tapes and rigid disks have been measured using the optical-interference technique by Bhushan (1985a) and Bhushan and Dugger (1990a). A loading-unloading experiment was conducted to determine if the majority of the contacts in the measurement range (>0.7 μm in diameter) were elastic (Bhushan, 1984). Photographs of tape contacts were taken at 28 kPa; then higher pressure (1.38 MPa) was applied for short durations and the tape contact was brought back to 28 kPa and rephotographed, Fig. 17. There were no changes in the real area of contact after unloading to 28 kPa which suggests that the contacts which can be detected, are elastic. This observation is in agreement with the predictions from the fractal model.

If the contacts are elasic, then the real area of contact and friction is governed by the E_c and σ_p/R_p of the magnetic medium surface. Figure 18a shows an example that the friction of various magnetic tapes depend significantly upon the complex modulus. Stable frictional behavior was exhibited only by those tapes which displayed a complex modulus of greater than 1.2 to 1.5 GPa. Figures 18b and 19 show the examples that the friction also strongly depends on the surface roughness. Typical contact diameters for tapes and rigid disks were found to be about 6 μm and 1.5 μm, respectively.

Bhushan and Dugger (1990a) reported a significant increase in the contact diameter, number of contacts and total real area of contact of the thin-film rigid disk as a function of loading time, Fig. 20. Asperities under load viscoelastically and viscoplastically deform which not only increase the size of the existing asperities but also brings the two surfaces closer to allow contact of additional asperities. We expect the rate of increase in the real area of contact as a function of loading time to be dependent on the rate-dependent mechanical properties and the normal stress. Therefore, attempts should be made to select materials for disk coatings with low creep compliance, to reduce the normal stress at the head-disk interface, and to explore methods (such as load/unload mechanisms) to minimize or avoid altogether the storage of head in contact with the disk. In the case of magnetic tapes, creep compliance and hydrolytic degradation characterisitcs of the binder also need to be optimized for sustained low friction after storage at high pressure (e.g., near end-of-tape on a reel) and high temperature/humidity (Bhushan, 1990).

3.2 Liquid-Mediated Adhesion (Stiction)

A rather smooth magnetic medium (especially thin-film disk) has a tendency to adhere or stick strongly to the smooth magnetic head. The liquid-mediated adhesion commonly referred to as "stiction" in the computer industry, is especially pronounced when liquid lubricants and adsorbed moisture are present at the interface. Liquid-mediated adhesion can be divided into two components — a meniscus term and rate-dependent viscous term. Both components can contribute significantly to the adhesion or stiction. The meniscus term depends on the surface tension of the liquid and viscous term depends on the viscosity of the liquid. Viscous term does not depend on the surface tension and can be observed even when the surfaces are completely surrounded by the liquid. If the surfaces are submerged in the liquid they may be separated easily, provided the separation is carried out very slowly. However, if the rate of separation is rapid, the viscosity of the liquid will be the determining factor. It is easy to see that when the surfaces are pulled apart, the liquid must flow into the space between them, and if the separation is rapid, the viscosity of the liquid will be the determining factor.

For analysis purposes, we consider a model of contact region between smooth surfaces with different level of "fills" of the interface and it depends on the mean interplanar separation and the liquid levels, Fig. 21 (Matthewson and Mamin, 1988). There are two extreme regimes in which either a small quantity of liquid bridges the surfaces around the tip of a contacting asperity (the "toedipping" regime) or the liquid bridges the entire surface (the "flooded" regime); and in the third regime, the liquid bridges around from few asperities to

large fraction of the apparent area. The different regimes can be modelled and the expressions for F_M and F_V can be obtained (Bhushan, 1990).

In the toe-dipping regime, the effect of the liquid condensate on the adhesion force between a single asperity and a surface can be modelled by a sphere of composite radius of curvature in contact with a flat surface with a liquid bridge in between. The total meniscus and viscous forces of all wetted asperity contacts can be calculated by multiplying the number of contacts by the meniscus and viscous forces at a typical contact. The flooded regime can be modelled by a liquid bridge between the two flat surfaces. The pill box regime can be modelled by two flat surfaces. If we assume that all surface asperity radii are constant and their heights follow a Gaussian distribution, the true coefficient of friction is given as follows:

For toe-dipping regime:

$$F \sim \frac{f_r W}{\dfrac{1 - 16.6\gamma_l(\cos\theta_1 + \theta_2)}{E' \sigma_p (\sigma_p / R_p)^{1/2}}} \tag{12}$$

For flooded regime:

$$F = f_r\left[W + \frac{A_a \gamma_\ell}{h}(\cos\theta_1 + \cos\theta_2\right] + \frac{\eta_\ell A_a}{h}(L\alpha)^{1/2} e^{-1/2} \tag{13a}$$

where

$$\frac{h}{\sigma_p} \sim 1.4\left\{\log\left[0.57\eta R_p \sigma_p E'(\sigma_p / R_p)^{1/2} / p_a\right]\right\}^{0.65} \tag{13b}$$

where F is the friction force, γ_ℓ and η_ℓ are the surface tension and viscosity of the liquid, θ_1 and θ_2 are the contact angles of the liquid on the two surfaces, h is the average thickness of the liquid bridge, L is the distance surfaces need to slide to become unstuck, and α is the start-up linear acceleration. We note that the liquid-mediated adhesive forces decrease with an increase in roughness.

We make an important observation that in the toe-dipping regime, the adhesion force is independent of the apparent area and proportional to the normal load (i.e., number of asperity contacts). However, the flooded regime shows the opposite tendencies. The pillbox regime is intermediate and can exhibit either behavior at the extremes. In all three regimes, adhesion force decreases with an increase in σ_p and a decrease in R_p and is independent of η.

The relative humidity of the environment, rest period, head-slider area, surface roughness, lubricant viscosity and its thickness and relative velocity affect the liquid-mediated adhesion (Liu and Mee, 1983; Bradshaw and Bhushan, 1984; Yanagisawa, 1985; Miyoshi et al. (1988); Bhushan, 1990; Bhushan and Dugger, 1990b; Streator, 1990a, 1990b). Miyoshi et al. (1988) have measured the effect of water vapor on adhesion of a Ni-Zn ferrite pin in contact with a flat of Ni-Zn ferrite or of magnetic tape A, Fig. 22. They found that the adhesive force (normal pull-off force) of ferrite-ferrite or ferrite-tape A contact remained low below 40% RH, the adhesion increased greatly with increasing relative humidity above 40%. Changes in the adhesion of contacts were reversible on humidifying and dehumidifying. The adhesive forces for a liquid bridge between a spherical surface with radius same as of the pin and a flat surface were calculated using surface tension and contact angle values for water. The

calculated values compared well with the measured values. They concluded that ferrites adhere to ferrites or tapes in a saturated atmosphere primarily from the surface tension (or meniscus effects of a thin-film of water absorbed on the interface).

Bhushan and Dugger (1990b) measured the effect of water vapor (relative humidity) and lubricant film on adhesion of a 3380-type Al_2O_3-TiC slider in contact with a thin-film (metal) disk C. The effect of exposure time on the adhesive force at 90% RH for an unlubricated disk is shown in Fig. 23. Measurable adhesion (>0.1 mN) was observed only after 90 minutes of exposure. The adhesive force increased with the exposure time up to about 5 hours, after which there was no significant increase in adhesive force with exposure time. Effect of humidity on the lubricated and unlubricated disks is shown in Fig. 24. They found that the adhesive force (normal pull-off force) of head-disk contact remained low (below resolution of the measurement technique ~50 mN) below 75% RH, the adhesion increased greatly with increasing relative humidity above 75%. The increase in the adhesive force was slightly larger for the unlubricated disk than for the lubricated one. Since the disk lubricant is hydrophobic, it repels some of the water condensation. However, water can replace some of the topical PFPE at concentrated asperity contacts during long exposures to water vapor or if the lubricant is squeezed out of the local contact by high pressure or displaced by sliding of the two surfaces. Studies of the penetration of lubricant layers by water (Baker et al., 1962) suggest that water may diffuse through the lubricant and condense into droplets around nuclei on the solid surfaces. Spreading of the water drop will be controlled by the energy difference between the water/disk and lubricant/disk interfaces. At this time, water with a surface tension on the order of 3 to 4 times that of typical magnetic medium lubricants, wets the magnetic-medium surface creating a meniscus at the asperity contacts. Attempting to separate the surfaces against this water film gives rise to the observed adhesion. Bhushan and Dugger also measured the adhesive force as a function of separation rate (proportional to sliding velocity), Fig. 25. They found that the adhesive force increases approximately linearly with an increase of the square root of the loading rate. This is attributed to viscous effects.

The effect of lubricant thickness and its functionality on the static and kinetic coefficients of friction in particulate disks was studied by Scarati and Caporiccio (1987), Fig. 26. We note that static friction increases with an increase in the lubricant film thickness; however, the reverse is true for kinetic friction. Increase in static friction with an increase in the lubricant thickness occurs at a lower thickness for a nonpolar lubricant than for the functional lubricant. Kinetic friction decreases with an increase in the lubricant film thickness.

The effect of lubricant viscosity and its thickness, lubricant functionality, and disk surface roughness on the static and kinetic coefficients of friction on thin-film disks was studied by Yanagisawa (1985) and Streator et al. (1991a). Figures 27 and 28 show the data for three disks with different roughnesses coated with four perfluoropolyether lubricants—L1 to L3 and F having different film thicknesses. Lubricants L1, L2 and L3 are nonpolar liquid lubricants, while F is a polar liquid lubricant with dihydroxyl functional end groups. In lubricant L3 the carbon atoms form a branched bonding structure, while lubricants L1, L2 and F contain carbon atoms in a linear arrangements. The static friction does not show a monotonic increase at the higher values of lubricant thickness as seen for kinetic friction. The coefficients of static and kinetic friction are essentially independent of lubricant thickness below a "critical" lubricant thickness. All lubricants show a sharp increase in friction at the critical thickness, which is different for each lubricant. The polar lubricant F exhibits the smallest critical thickness for dramatic friction increase. In the case of nonpolar lubricants (L1, L2, and L3) the critical thickness was lower for the lower viscosity lubricants. Note that the polar lubricant F does not fit within the viscosity trend exhibited among the nonpolar lubricants. The buildup of the friction is believed to be governed by micro-flow capabilities of the liquid on the disk. The lower viscosity lubricants among a class of nonpolar lubricants flow more readily to develop the menisci bridges, resulting in higher friction than that of lubricants with higher viscosity. The polar lubricant exhibits high friction compared to the nonpolar lubricants and does not follow the viscosity trend. The increase in friction

with increasing lubricant-film thickness above the critical thickness can be attributed to strong adhesive forces in the interface (Bhushan, 1990).

The critical thickness is correlated with disk surface roughness, which roughly corresponds to the rms roughness. The rougher surface would correspond to a larger mean separation of the surfaces, and thus a higher value of lubricant-film thickness when the meniscii are formed. The friction values above critical-film thickness are higher for smoother disks. However, these results of effect of roughness on the critical film thickness are only applicable to short contact times on the order of seconds or minutes. With longer rest times (hours or days), the adhesion usually reaches much higher value even if the lubricant thickness is well below the critical value. This time effect can be explained by the slow diffusion of the lubricant molecules towards contact points driven by the Laplace pressure and deformation of the interacting asperities with the corresponding increase in the real area of contact.

Streator et al. (1991a) also reported that kinetic coefficient of friction decreases with increasing sliding speed.

4.0 INTERFACE TEMPERATURES

In a sliding operation, almost all of the frictional energy input is directly converted to heat in the material close to the interface. During a sliding situation, asperity interactions result in numerous high temperature flashes. Bhushan (1987a) presented a detailed thermal analysis to predict the interface temperatures and applied it to predict temperatures in a head-medium interface (Bhushan, 1987b; 1992b). The head-medium interface can be modelled as the case of sliding two rough surfaces, Fig. 29(a). The surface profile is measured and surface asperities are modelled with a series of spherically-topped asperities. The degree of interaction between the two surfaces at any time during sliding depends on the average contact stress. The interaction problem at an asperity contact reduces to a sphere sliding against another sphere, assuming the distance to the center of the two spheres is fixed. When one sphere comes in contact with the other, the real area of contact starts to grow; when one sphere is directly above the other, the area is at maximum; as one sphere moves away, that area starts to get smaller, Fig. 29(b). The center of the contact moves at approximately half the relative sliding speed with respect to each asperity. The real area of contact is a source of frictional heat and the heat intensity is proportional to the real area. Bhushan showed that the total flash temperature consists of temperature of an individual asperity contact and effect of other asperity contacts on an individual asperity temperature (interaction).

The relevant equations for the average and maximum asperity temperature rise of the interface are given as (Bhushan, 1987a)

$$\bar{\theta} = r_1\left[0.65\, fp_a (A_a / A_r)(V d_{max} / K_1)^{1/2} / \rho_1 Cp_1 + 1.5\, fp_a (V \ell / K_2)^{1/2} / \rho_1 Cp_1\right] \quad (14a)$$

$$\theta_{max} = r_1\left[0.95\, fp_a (A_a / A_r)(V d_{max} / K_1)^{1/2} / \rho_1 Cp_1 + 1.5\, fp_a (V \ell / K_2)^{1/2} / \rho_1 Cp_1\right] \quad (14b)$$

and

$$r_1 \sim 1 / \left[1 + (k_2 \rho_2 Cp_2 / k_1 \rho_1 Cp_1)^{1/2}\right] \quad (14c)$$

where ρCp is the volumetric heat capacity, κ is the thermal diffusivity, k is the thermal conductivity, d_{max} is the maximum contact diameter, ℓ and is the half length of the slider.

Average and maximum transient temperatures predicted for a typical particulate tape-head interface were 7° and 10°C, respectively (Bhushan, 1987b). These predictions compared

fairly well with the infrared measurements conducted at head-tape interface by Gulino et al. (1986). The asperity-contact temperatures at head-tape interface are relatively low because of its high real area of contact, as compared to that of metal-metal or ceramic-ceramic contacts. The transient temperature of 7-10°C rise can lead to high friction in some tapes because the transition temperature of some tapes' mechanical properties is within 5°C above the ambient temperature. In isolated cases, if the magnetic particles are exposed (or get exposed in a high-speed rub) and contact the head surface, the average and maximum transient temperature rise could be about 600°C and 900°C, respectively. These temperatures potentially will cause a breakdown of the medium lubricant and a degradation of the medium binder leading to excessive friction and seizure of medium motion.

Average and maximum transient temperatures predicted for a typical particulate rigid-disk-slider interface are 34 and 44°C, respectively (Bhushan, 1992b). If the exposed magnetic particles or alumina particles contact the slider surface, the transient temperature rise could be more than 1000°C. Average and maximum transient temperature rises for a typical thin-film disk-slider interface are 56° and 81°C, respectively for Al_2O_3-TiC slider and 77 and 110°C, respectively for Mn-Zn ferrite slider. These predictions compared fairly with the infrared measurements conducted at rigid disk-slider interface by Bair et al. (1991) and Suzuki and Kennedy (1991). The size of an asperity contact is on the order of 1.5 μm and duration of asperity contact at the full operating speed is less than 100 ns. The thermal gradients perpendicular to the sliding surfaces are very large (a temperature drop of 90% in a depth of typically less than a contact diameter or less than a micron).

5.0 WEAR

5.1 Head-(Particulate) Tab Interface

The wear of oxide magnetic particles and ceramic head body materials is different from metallic wear because of the inherent brittleness and the relatively low surface energy of ceramics. The first sign of ferrite head wear with a magnetic tape is the appearance of very small scratches on the head surface (Fig. 30a). The physical scale of scratches is usually very fine and scratches as small as 25 nm have been reported (Bhushan, 1985b). Ferrite surface is microscopically removed in a brittle manner as stripes or islands, depending on the smoothness of the tape surface. Wear generally occurs by microfragmentation of the oxide crystals in the ceramic surface. Fragmentation is the result of cleavage and transgranular fractures, one dominated by intergranular fracture. We note that the worn head surface is work hardened which reflects a shift from a mechanism dominated by transgranular fracture to one dominated by intergranular fracture. Figure 30b shows a region where fracture and rupture have occurred. Such regions are commonly called "pullouts". Debris originating from these regions causes additional small scale plastic deformation and grooves (three-body abrasion). We also note that the ferrite head surface is work hardened with a large compressive stress field after wear which is detrimental to magnetic signal amplitude (Chandrasekar et al., 1987a, 1987b, 1988, 1990).

Head wear depends on the physical properties of head and medium materials, drive operating parameters, and environmental conditions. Wear data of common head materials against a γ-Fe_2O_3 tape is shown in Fig. 31. We observe,a linear relationship between wear rate and material hardness for abrasive wear model. Head wear also depends on the grain size of the head material, magnetic particles, tape-surface roughness, isolated asperities on the tape surface, tape tension and tape sliding speed. Head wear as a function of surface roughness of tapes and isolated asperities on the tape surface are shown in Figs. 32 and 33. We note that head wear increases with an increase in the surface roughness or number of isolated asperities of the tape surface. Wear rate also increases with humidity, above about 40-60% relative humidity, Fig. 34. An increase in abrasive wear at high humidities is believed to be due to moisture-assisted fracture (or static fatigue) of the grains to yield finer particles (Bhushan, 1985b; Bhushan, 1990).

Head sliders after usage sometimes become coated with thin layers of a new organic material of high molecular weight called "friction polymers" or "tribopolymers." Friction is essential for the formation of these materials. Another requirement for the formation of friction polymers in a rubbing contact is that one of the surfaces, lubricant, or even a material nearby should be organic (Lauer and Jones, 1986). It seems clear that all friction polymers are products of chemical reaction, whether they derive initially from solid polymers or from organic liquids or vapors. Friction polymers are found on head surfaces. These result in discoloration of the head surface and give an appearance of brown or blue color, and, therefore are sometimes called brown or blue stains, respectively.

During contact of particulate tape with the head in contact start/stops or during partial contact in streaming, binder and magnetic particle debris is generated primarily by adhesive wear mode. The debris can be either loose or adherent (Bhushan and Phelan, 1986). Tape debris, loose magnetic particles, worn head material or foreign contaminants are introduced between the sliding surfaces and abrade material off each. The debris that adheres to drive components lead to polymer-polymer contact, whose friction is higher than that of rigid material-polymer contact and can lead to magnetic errors and sometimes to catastrophic failures (Bhushan, 1990).

Calabrese et al. (1989) conducted *in situ* low speed sliding experiments in which Ni-Zn ferrite pin was slid against CrO_2 tape A. During wear, the tape particles consisting of binder resin and magnetic particles were generated at the contact and were literally thrown out of the interface landing within a radius of 0.8 mm from the pin. If the particles landed in the path of the pins, they would be drawn through the contact as the tape moved into the interface. Hence, a three-body wear scenario would be set up, which generated more particles. Wear particles were generally of a block-type or a flake type. The block-type particles were in the form of blocks of about 5 μm and the flake type particles were typically much smaller in size.

Bhushan et al. (1986) used autoradiographic technique to conduct wear studies at head-tape interface. An irradiated Ni-Zn ferrite head was run against various tapes. They found that measurable transfer of ferrite on the CrO_2-tape was observed after 5,000 passes. A Co-γFe_2O_3 tape did not show significant transfer even after 20,000 passes. Wear was represented in four distinct patterns: smeared areas, gray areas, dots or specks, and streaks or lines running in the direction of the tape. In a test, the amount of ferrite deposited on the CrO_2-tape after 20,000 passes across the head was about 0.6 ng/cm^2. In this test, about 0.6% of the generated ferrite debris was transferred to the tape, about 0.2% was transferred to the tape-drive component surfaces, and the rest was believed to be airborne.

5.2 Head-(Particulate) Rigid Disk Interface

During asperity contacts, the disk debris can be generated by adhesive, abrasive and impact wear. The wear debris, generated during the manufacturing process (burnishing or buffing) of the disk, and foreign contaminants present can get trapped at the interface, which results in three-body abrasive wear, or they can get transferred to the slider, resulting in performance degradation of the air bearing. The flash temperatures generated at asperity contacts can render any boundary lubricants ineffective and can degrade the mechanical properties of the disk binder increasing the real area of contact; this results in high friction and high disk wear. Any of these mechanisms can lead to head crash. The environment (humidity and temperature) has a significant effect on the head-disk friction and wear or debris generation.

Microscopic observations of disk and head surfaces after head crash in a CSS test show circumferential wear grooves and support either of the two-body or three-body abrasive wear, Fig. 35. Karis et al. (1990) and Novotny et al. (1991a, 1991b) reported the complete removal of lubricant from the start/stop track at least in one region, and degradation of lubricant

preceded the final stage of disk failure. Lubrication of the disks increased the number of sliding cycles until the coating began to wear through by up to 1000 times. The number of sliding cycles until frictional failure occurred was proportional to the areal density of lubricant.

Scarati and Caporiccio (1987) studied the effect of lubricant thickness and its functionality on the wear life of particulate media. Figure 36 shows the wear life of a particulate disk lubricated with two grades of perflunropolyether (PFPE) - Fomblin Z-25 (non-polar) and Fomblin AM2001 (polar with reactive end groups or functional lubricant) as a function of lubricant thickness. We note that relative wear life increases with an increase in the lubricant thickness (also see Karis et al., 1990), and that polar lubricants have longer wear life than nonpolar lubricants.

The mechanism of interface failure or head crash was studied by Kawakubo et al. (1984, 1991) for particulate disks in contact start-stop (CSS) test. Friction force, acoustic emission (AE) signal (to monitor head-disk contact), and read-back magnetic signal were measured during the test. The changes in read-back magnetic signal, friction force, and AE signal are shown in Fig. 37. At the point of interface failure, the read-back signal decreased to almost zero and friction force and AE signal rose significantly. This implies that the head was virtually in contact even at full speed. Kawakubo et al. (1984) also videotaped the wear process through a transparent sapphire slider. They found that the disk debris transferred to the rail surfaces preceding the interface failure.

Using submicron, fluorescent polystyrene-latex particles, Hiller and Singh (1991) studied the interaction of contaminant particles with a flying slider. On the slider, the particles were deposited mainly in two regions: on the tapered region of the air-bearing rails and in the form of whiskers along the trailing ends of the rails, Fig. 38. The whiskers contained only deformed particles, which is evidence of strong interaction between the particles and the interface, Fig. 39. After flying for sometime, large agglomerates of particles were occasionally found on the tapers. Since they contained mainly deformed particles, they could be identified to be whiskers which had detached from the trailing end, Fig. 40. No whiskers were grown on unlubricated disks, Fig. 38. This shows that liquid lubricant promotes adhesion between particles and surfaces. Liquid lubricant may thus have an adverse effect on reliability when large amounts of contaminant particles are present.

5.3 Head-(Thin-Film) Rigid Disk Interface

Magnetic films used in the construction of thin-film disks are soft and have poorer wear resistance than the particulate disks which are loaded with hard magnetic particles and load-bearing alumina particles. Thin-film disks have smoother surfaces and are designed to fly at a lower flying height than particulate rigid disks, which result in higher friction and an increased potential of head to disk interactions. During normal drive operation, the isolated asperity contacts of head and disk surfaces result in wearout of the disk surfaces by adhesive and impact wear model, and generate the debris. In addition, any asperity contacts would result in the maximum wear stress at the disk subsurface, which may initiate a crack. Repeated contacts would result in crack propagation (subsurface fatigue) leading to delamination of the overcoat and the magnetic layer. Isolated contacts in a clean environment generate very fine wear debris (primarily made of magnetic film and overcoat) which subsequently results in rather uniform disk wear from light burnishing (three-body abrasion). The disk wear results in high friction and wear in subsequent contacts, resulting in head crash (Bhushan, 1990).

External contaminations abrade the overcoat readily and result in localized damage of the disk surface by three-body abrasion. Wear debris generated at the interface invades the spacing between the head slider and the disk and/or transfers to the head slider making the head slider unstable, which leads to additional debris, and results in head crash both in the start/stop and flyability modes. Koka and Kumaran (1991) studied the effect of alumina

contaminant particles on the flyability of a drive. A significant buildup was observed in the leading-edge taper area of the slider, and abrasive wear on the disk resulted from particles trapped in the leading-edge taper region.

Engel and Bhushan (1990) have developed a head-disk interface failure model for thin-film disks. The principal physical variables include the sliding speed, surface topography, mechanical properties, coefficient of friction and wear rate. Surface protrusions, such as asperities and debris particles, induce impact and sliding encounters, which represent a damage rate. Failure occurs when a specific damage rate, a characteristic for the system, is reached. Modeling uses a set of topographic parameters describing the changing, wearing surface.

5.3.1 Role of Slider and Overcoat Materials

Figure 41 shows the coefficient of friction as a function of number of passes for a thin-film disk with carbon-overcoat (disk B_1) sliding against various slider materials (Chu et al., 1990; Chandrasekar and Bhushan, 1991). Amongst the ceramics tested, single-crystal diamond has the lowest coefficient of friction (~0.12) followed by partially-stabilized zirconia (~0.15), the remaining ceramics all have an initial coefficient of friction close to 0.2 when in contact with the disk. We note that coefficient of friction increases with number of passes. This increase was, however, small for single crystal diamond even after it had been in contact with the disk for about 5500 passes. The rate of increase in the coefficient of friction is highest, in the case of calcium titanate and Al_2O_3-TiC sliders and Mn-Zn ferrite and ZrO_2-Y_2O_3 sliders exhibited a smaller increase. Calcium titanate (1200 kg/mm^2) probably showed poorest durability because it cracks readily. Al_2O_3-TiC is hardest (2300 kg/mm^2), it burnishes the disk surface more than Mn-Zn ferrite (600 kg/mm^2) and ZrO_2-Y_2O_3 (1300 kg/mm^2). Examination of Mn-Zn ferrite slider shows scratches along the air-bearing surface, which suggests that the Mn-7n ferrite is slightly softer than the disk structure, and that, therefore ferrite is gentle to the disk surface. We believe that matching of ceramic slider and disk hardnesses is essential for low wear.

An increase in overcoat hardness improves the wear resistance of the thin-film disks (Yanagisawa, 1985b). Increase in hardness of SiO_2 overcoat by baking at various temperatures results in improvement in wear resistance as shown in Fig. 42, where the normal load was 185 mN and the sliding velocity was 1.12 m/s. Khan et al. (1988) have reported that hard carbon overcoats with better wear performance consist of homogeneous grain size and uniform grain distribution across the surface and higher percentage of sp^3 bonded carbon atoms (diamond structure) as compared to the carbon films with poor wear performance. Yamashita et al. (1988) reported that wear performance of unlubricated disks with ZrO_2-Y_2O_3 overcoat is superior to that of carbon, Fig. 43. Wear performance of ZrO_2-Y_2O_3 overcoat is comparable to carbon overcoat for lubricated disks. In a ceramic-ceramic contact, yttria-stabilized zirconia is known to have excellent friction and wear performance (Bhushan and Gupta, 1991). They also reported that ZrO_2 overcoat (with 30 nm thickness) exhibited superior corrosion resistance than hard carbon of coated thin-film (metal) disk, when exposed to 80°C/90% RH for 7 days. Ceramic overcoats with low porosity and high electrical resistivity are known to have better electrochemical corrosion resistance (Bhushan, 1990).

Calabrese and Bhushan (1990) conducted in-situ sliding experiments of various head/thin-film disk (95-mm dia.) combinations in the scanning electron microscope (SEM) (also see Hedenqvist et al., 1991). The purpose was to identify the initiation of particle removal during the sliding process. For example, after sliding at 50 mm/min. for a few minutes of a Al_2O_3-TiC slider on a thin-film disk with a zirconia overcoat and perfluoropolyether as the topical lubricant (disk B2), microscopic particles were removed from the edges of the rails and these particles were deposited from the head to the disk surfaces, Fig. 44. In Fig. 44, we also notice little disk debris deposited on the rail edges. Microscopic examination of the head and disk surfaces after the 20 minute test at 50 mm/min. showed that

there was some damage to the slider edges and the disk surface was very lightly burnished with only one scratch. Minute disk debris was found on the rail edges and rail surfaces including leading taper of the head slider, Fig. 45. Continued sliding led to the increased surface change of the disk followed by catastropic failure. There appears to be several parameters that influence the initiation of particle removal. The most significant is the condition of the rail edges which contact the disk. During the start of motion between the head and the disk, the head moves with the disk until the force on the spring suspension overcomes the adhesion between the head and the disk, it will spring back in an unstable manner, causing contact with the rail edges and the disk. This could result in transfer of material from the disk or chipping of the rail edge. Calabrese and Bhushan (1990) reported head slider and disk wear to be strongly dependent on the slider and the disk overcoat materials, Table 5. Zirconia overcoat generally exhibited less wear than the carbon overcoat. Mn-Zn ferrite slider was less aggressive to the disk than the Al_2O_3-TiC slider. The calcium titanate slider cracked early on in a sliding test, hot processing of this material appears to be a problem.

5.3.2. Role of Lubricant Film

The effect of lubricant viscosity and its thickness, lubricant functionality and disk surface roughness on the durability was studied by Miyamoto et al. (1988) and Streator et al. (1991b). The data in Fig. 46 show the friction histories of unlubricated and lubricated disks. The disk failure is defined by the advent of a relatively sharp rise in the friction as a result of repeated sliding cycles. The effect of surface roughness is summarized in Fig. 47. The disk X_2 demonstrates the lowest durability and indicates the effect of texturing as compared to high durability of the untextured disk, X_1. Disk B_1 is also textured, but is from a different disk manufacturer, and cannot be compared to the other disks on the basis of roughness alone. As can be seen in Fig. 47, the presence of the lubricant improved the durability of the disk over that of dry sliding in all cases. A trend exists indicating that durability increases with decreasing viscosity. The polar lubricant has significantly higher durability than the nonpolar lubricant with comparable viscosity. The greater durability of the less viscous lubricants can be attributed to their greater mobility on the disk surface. Figure 48 shows the durability results conducted on disk X_2, where the lubricant film thickness is being varied. There is a general increase in durability with lubricant film thickness, as expected.

Figure 49 shows the effect of storage time on static friction for disks with a nonpolar lubricant (PFPE) and disks with dual lubricant consisting of polar (aminosilane) and nonpolar (PFPE) fraction (Hoshino et al., 1988). Increase in friction from aging the disks with dual lubricant film was found to be less than that for a disk with only nonpolar lubricant. Lubricant is also spun off with disk rotation during use. Yanagisawa (1985a) and others have shown that polar lubricants spin off less than nonpolar lubricants (Bhushan, 1990). The concept of a dual layer consisting of an unbonded layer over a bonded layer is very useful because unbonded (mobile) top layer would heal any worn areas on the disk surface where the lubricant may have been removed, and the bonded layer provides lubricant persistence.

5.3.3. Role of Evironment

Unlubricated Disk

Marchon et al. (1990) and Strom et al. (1991) studied the wear behavior of unlubricated thin-film disk with a carbon overcoat sliding against ceramic sliders in various environments. Strom et al. (1991) tested unlubricated carbon-coated disks (disks B_1 with no lubricant) and ZrO_2-Y_2O_3 coated disks (disks B_2 with no lubricant) in a sliding test against commercial Al_2O_3-TiC slider, Figs. 50(a) and 50(b). The average friction of carbon-coated disk increased smoothly only in the oxygen environment which indicates poor durability, also see Dimigen and Hubsch, 1983-84; Memming et al. (1986) and Miyoshi et al. (1989). At sustained high friction, debris was generated which reduced the real area of contact and the friction

dropped. In the case of zirconia-coated disk, no consistent difference in the tribological behavior between various gas environments was observed.

The sliding test on carbon-coated disk was also conducted in the presence of humid gases. In this case, little difference in the friction behavior for these different gases was observed, Fig. 50(c). In all gases, the coefficient of friction increased smoothly to about 1.4 during the course of about 50 revolutions. Increase in friction in the oxygen or humid environment for the carbon-coated disk can be explained by the oxidation of the carbon film under rubbing, a tribochemical reaction. In the case of sliding on the zirconia-coated disk, no material is removed through oxidation. Wear occurs through mechanical means only, regardless of the concentration of oxygen in environment.

Marchon et al. (1990) have also reported the carbon oxidation as a key process that contributes significantly to wear and friction increase with test cycles. In sliding experiments with Mn-Zn ferrite or calcium titanate sliders and unlubricated thin-film disks with carbon overcoats, Marchon et al. (1990) reported that there is gradual increase in the coefficient of friction with repeated sliding contacts performed in air, however, in pure nitrogen, no friction increase is observed, the coefficient of friction remaining constant at 0.2. The alternate introduction of oxygen and nitrogen elegantly showed the role of these gases, Fig. 51. Contact start/stop tests exhibited same effect. Marchon et al. (1990) suggested that wear process in the oxygen environment involved oxygen chemisorption on the carbon surface and a gradual loss of carbon through the formation of CO/CO_2 due to the action of the slider.

Lubricated Disk

Dugger et al. (1990) conducted a wear study of hemispherical pins* (with a radius of curvature of 50 mm) of Mn-Zn ferrite sliding against a thin-film disk with carbon overcoat and perfluoropolyether as the topical lubricant (disk B_1) in controlled environments. They found that the contact life, as marked by the total distance slid to the point at which the coefficient of friction increases rapidly over the steady state value, is much larger in air with 50% RH than in dry air or vacuum, Fig. 52(a). SEM examination of the wear scars on the pin and disk revealed morphology markedly dependent on the testing environment, Fig. 53. Characteristic of post-failure surfaces in vacuum and dry air are severe damage to the disk surface, with the pin from the vacuum test also exhibiting extensive damage, including intergranular fracture and grain pull-out. In both cases there is also material transferred from the disk surface to the pin. In humid air, however, the contact area on the pin is covered with very fine debris (about 1 μm) particles in a dark film (low atomic number), with fine particles on either side of the worn area. Figure 54 illustrates the topography on the submicron scale (by STM) of the untested as well as wear track regions on disks tested to failure in vacuum and humid air. The general observation is that the surfaces of vacuum and dry air tested disks become rougher. In humid air, on the other hand, finer grooves are seen in the surface topography, leaving a surface roughness of longer wavelength. Further surface analyses showed that the wear debris generated in humid air is much finer and is enriched with cobalt (from the magnetic layer) on its surface. In dry air and vacuum, the debris is substantially larger than one micron (Fig. 55) and tends to be enriched with nickel (probably from magnetic layer and Ni-P underlayer) on its surface. We propose that two mechanisms contribute to the observed durability differences: oxidation of metallic wear debris generated at isolated asperity contacts and alteration of the coating surface by interaction with vapor. The rate of debris oxidation depends upon the test ambient and affects the tendency for metallic debris to agglomerate through sintering or mechanical compaction into larger particles which are more damaging (Dugger et al., 1991). Significant adhesion in vacuum and less so in dry oxygen probably results in significant wear debris generation and by mechanical compaction or otherwise, small wear particles may

* Dugger et al. (1992) have reported that tests with actual slider from commercial rigid disk files yield relative contact lives that are comparable to those observed with hemispherical pins despite the apparent contact stress differences. Therefore, pins were used for acceleration in wear.

agglomerate to produce large wear fragments (Fig. 55) that lead to catastrophic failure in the case of vacuum or dry air. We believe that rapid agglomeration to particle sizes of greater than a few microns is responsible for the reduction in contact life in vacuum and dry air. However, in the 50% RH air, less adhesion at the interface and oxidation of metallic debris result in increased wear life.

Wahl et al. (1991a) and Dugger et al. (1992) further conducted tests in ultra-high-purity (>99.999%) nitrogen, ultra-high purity (>99.995%) helium and 50% RH nitrogen environments, Fig. 56. The contact life is short for nitrogen and helium environments and fall in the same range as the durability in vacuum. This data suggest that water vapor and oxygen in the humid air tests and oxygen in the dry air tests are responsible for the greater durabilities in these environments, while neither helium nor nitrogen plays a beneficial role for the durability of the rigid disks studied. The data for the carbon overcoat in humid nitrogen indicate that water vapor alone has a large effect on the observed improvement in durability in humid air compared to vacuum or dry air, although both oxygen and water vapor contribute to increased durability. To explore the effects of water vapor introduction, Wahl et al. (1991a) performed durability testing using a nitrogen ambient with humidities ranging from 0.2 to 80% (these correspond to partial pressures of water vapor from about 5 Pa to 2600 Pa at room temperature), Fig. 57(a). It appears that introduction of as little as 0.2% RH to the ambient dry nitrogen resulted in over two orders of magnitude increase in the contact life. Figure 57(b) shows a comparison of coefficients of friction for steady state sliding in the nitrogen ambient as a function of relative humidity. The friction is essentially independent of humidity up to a certain value of humidity and the increase in friction at very high humidity (>80%) is believed to be due to liquid-mediated adhesion as discussed earlier, also see Dimigen and Hubsch (1983-84). The radical change in durability with the introduction of water vapor or oxygen is believed to be due to oxidation of wear debris in an oxidizing environment (either water vapor or oxygen) before it has a chance to agglomerate into larger and more destructive wear particles. The optimum humidity for maximum durability may depend on the stress at the interface; the interface with a small stress (e.g., small slider) may show more sensitivity to the humidity than the interface with high stress (large slider).

Dugger et al. (1992) conducted a wear study on the thin-film disk with ZrO_2-Y_2O_3 overcoat and perfluoropolyether lubricant disk (B_2), Figs. 52(b) and 56. The coefficient of friction increased throughout the test* to steady-state values between 0.6 and 1.4, depending upon the environment. Even at this high coefficient of friction, no visible wear track could be observed initially on the disk surface. When a wear track did become visible, an abrupt drop in the coefficient of friction occurred. This abrupt decrease is attributed to the generation of debris from the zirconia overcoat which, when present between the surfaces, reduced the real area of contact and hence, the friction force. The similarity of the contact lives of the zirconia overcoat in vacuum and dry air suggest that oxygen does not significantly affect the wear rate of this material. In vacuum and dry air, the disk surface exhibited extensive damage, with complete removal of the zirconia layer in some locations and transfer of metals to the pin surface. Wear particles are frequently larger than 10 μm. The surface morphology of the damaged area is comparable to that of carbon overcoat. In humid air, the contact life is long; the disk surface appears polished in the wear track compared to the surrounding regions, with only isolated areas of damage. The wear study on the zirconia overcoat suggest that the contact life is sensitive to the presence of water vapor.

Figure 58 shows the Auger depth profile of the unworn surface and from the wear track formed on the thin-film disk (B_1) in humid air and stopped at 90% of the anticipated contact life (Dugger et al., 1992). These data indicate that the average carbon film thickness on the wear track is not very different from that on the untested region of the disk. Therefore, the majority of the film remains intact until very near the point of which the coefficient of friction increases dramatically above the steady-state value. Similar results have been

* In contrast, the rapid increase in coefficient of friction for carbon overcoat occurred alter significant sliding and was accompanied by a wear track on the disk visible to the naked eye.

reported by Wahl et al. (1991b) for vacuum environment. Thus, the precursor to failure is the catastrophic failure of carbon overcoat rather than uniform thinning of the overcoat. It is believed that debris is generated at isolated points in the contact zone where the largest asperities meet. The debris accumulates until a critical debris size or volume is generated, which results in catastrophic removal of the protective carbon film.

6.0 LUBRICATION

Mechanical interactions between the head and the medium is minimized by the lubrication of the magnetic medium. The primary function of the lubricant is to reduce the wear of the magnetic medium and to ensure that friction remains low throughout the operation of the drive. The main challenge, though, in selecting the best candidate for a specific surface is to find a material that provides an acceptable wear protection for the entire life of the product, which can be of several years in duration. There are many requirements that a lubricant must satisfy in order to guarantee an acceptable life performance. An optimum lubricant thickness is one of these requirements. If the lubricant film is too thick, excessive stiction and mechanical failure of the head/disk is observed. On the other hand, if the film is too thin, protection of the interface is compromised and high friction and excessive wear will result in catastrophic failure. An acceptable lubricant must exhibit the properties such as chemical inertness, low volatility, high thermal, oxidative and hydrolytic stability, shear stability and good affinity to the magnetic medium surface.

Fatty acid esters are excellent boundary lubricants and the esters such as tridecyl stearate, butyl stearate, butyl palmitate, butyl myristate, stearic acid and myristic acid, are commonly used as internal lubricants roughly 1 to 3% by weight of the magnetic coating, in tapes and flexible disks. The fatty acids involved include those with acid groups with an even number between C_{12} and C_{22} with alcohols ranging from C_3 to C_{13}. These acids are all solids with melting points above the normal surface operating temperature of the magnetic media. This suggests that the decomposition products of the ester via lubrication chemistry during a head-flexible medium contact may be the key to lubrication.

Topical lubrication is used to reduce the wear of rigid disks. Perfluoropolyethers (PFPEs) are chemically most stable lubricants with some boundary lubrication capability, and are most commonly used for topical lubrication of rigid disks (Bhushan, 1990). PFPEs commonly used include Fomblin Z and Fomblin Y lubricants made by Montiedison, Italy, Krytox 143 AD made by Dupont, U.S.A. and Demnum made by Diakin, Japan and their difunctional derivatives containing various reactive end groups, e.g., hydroxyl (Fomblin Z-Dol), piperonyl (Fomblin AM 2001), and isocyanate (Fomblin Z-Disoc), all manufacturered by Montiedison. The difunctional derivatives are referred to as reactive (polar) fluoroether lubricants. The chemical structures, molecular weights and viscosities of various types of PFPE, lubricants are given in Table 6 (Cantow et al., 1986; Corti and Savelli, 1989). We note that rheological properties of thin-films of lubricants are expected to be different from their bulk properties (Israelachvili et al., 1988; Homola et al., 1991). Fomblin Z is a linear PFPE; and Fomblin Y and Krytox 143 AD are branched PFPE where the regularity of the chain is perturbed by $-CF_3$ side groups. The bulk viscosity of Fomblin Y and Krytox 143 AD is almost an order of magnitude higher than the Z type. The molecular coil thickness is about 0.8 nm for these lubricant molecules. The monolayer thickness of these molecular depend on the molecular conformations of the polymer chain on the surface.

Fomblin Y and Z are most commonly used for particulate and thin-film rigid disks. Usually, lubricants with lower viscosity (such as Fomblin Z types) are used in thin-film disks in order to minimize stiction.

6.1 Measurement of Localized Lubricant-Film Thickness

The local lubricant thickness is measured by Fourier transform infrared spectroscopy (FTIR), ellipsometry, angle-resolved X-ray photon spectroscopy (XPS), scanning tunneling microscopy (STM), and atomic force microscopy (AFM) (Kimachi et al., 1987; Mate et al., 1989; Bhushan, 1990; Dugger et al., 1990; Sriram et al., 1991). Ellipsometry and angle-resolved XPS have excellent vertical resolution on the order of 0.1 nm but lateral resolution is on the order of 1 μm and 0.2 mm, respectively. STM and AFM can measure the thickness of the liquid film with a lateral resolution on the order of their tip radius of about 100 nm which is not possible to achieve by other techniques.

The schematic of AFM used for measurement of localized lubricant-film thickness by Mate et al. (1989) is shown in Fig. 59. The lubricant thickness is obtained by measuring the forces on the tip as it approaches, contacts and pushes through the liquid film. In the top part of Fig. 59 is a schematic diagram of an AFM up interacting with a lubricant covered particulate disk. A typical force versus distance curve for a tungsten tip of radius ~100 nm dipped into a disk surface coated with a perfluorinated polyether lubricant is shown in Fig. 60. As the surface approaches the tip, the liquid wicks up causing a sharp onset of attractive force. The so-called meniscus force experienced by the tip is $\sim 4\pi r\gamma$ where r is the radius of the tip and γ is the surface tension of the liquid. In Fig. 60, the attractive force measured is about 5×10^{-8} Newtons. While in the liquid film the forces on the lever remain constant, until repulsive contact with the disk surface occurs. The distance between the sharp snap-in at the liquid surface and the "hard-wall" of the substrate is proportional to the lubricant thickness at that point. [The measured thickness is about 2 nm larger than the actual thickness due to a thin layer of lubricant wetting the tip (Mate et al., 1989).] When the sample is withdrawn, the forces on the tip slowly decrease to zero as a long meniscus of liquid is drawn out from the surface.

Particulate disks were mapped by Mate et al. (1990) and Bhushan and Blackman (1991). The distribution of lubricant across an asperity was mapped by collecting force versus distance curves with the AFM in a line across the surface. Disk A was coated with nominally 20 to 30 nm of lubricant, but by AFM the average thickness is 2.6 ± 1.2 nm. (The large standard deviation reflects the huge variation of lubricant thickness across the disk). A large percentage of the lubricant is expected to reside below the surface in the pores. In Fig. 61, we show histograms of lubricant thickness across three regions on disk A. The light part of the bar represents the hard wall of the substrate and the striped part on the top is the thickness of the lubricant. Each point on the histrogram is from a single force versus distance measurement separated by 25 nm steps. The lubricant is not evenly distributed across the surface. In regions 1 and 2 there is over twice as much lubricant than there is on the asperity (region 3). There are some points on the top and the side of the asperity which have no lubricant coating at all (Bhushan and Blackman, 1991).

6.2 Lubricant-Disk Surface Interactions

The adsorption of the lubricant molecules is due to van der Waals forces, which is too weak to offset the spin off losses or to arrest displacement of the lubricant by water or other ambient contaminants. Considering that these lubricating films are on the order of a monolayer thick and are required to function satisfactorily for the duration of several years, the task of developing a workable interface is quite formidable.

An approach aiming at alleviating the above shortcomings is to enhance the attachment of the molecules to the overcoat, which, for most cases, is sputtered carbon. There are basically two approaches which have been shown to be successful in bonding the monolayer to the carbon. The first relies on exposure of the disk lubricated with neutral PFPE to various forms of radiation, such as low-energy X-ray (Heidemann and Wirth, 1984), nitrogen plasma (Homola et al., 1990) or far ultraviolet (e.g., 185 nm) (Saperstein and Lin, 1990). Another approach is to use chemically-active PFPE molecule, where the various functional (reactive) end groups offer the opportunity of strong attachments to specific interface. These

functional groups can react with respective surfaces and bond the lubricant to the disk surface which reduces its loss due to spin off and evaporation. Their main advantage, however, is their ability to enhance durability without the problem of stiction usually associated with weakly bonded lubricants (Scarati and Caporiccio, 1987; Miyamoto et al., 1988; 1990; Streator et al., 1991a, 1991b). The effect of bonded lubricant was demonstrated elegantly in recent AFM experiments by Blackman et al. (1990). They found that when a AFM tip is brought into contact with a molecularly thin-film of a non-reactive lubricant a sudden jump into adhesive contact is observed. The adhesion was initiated by the formation of a lubricant meniscus surrounding the tip pulling the surfaces together by Laplace pressure. However, when the tip was brought into contact with a lubricant film which was firmly bonded to the surface, only a marginal adhesion, mostly due to van der Waals forces, was measured (Fig. 62).

6.3 Lubricant Degradation

Contacts between the slider and the lubricated disk lead to lubricant loss, Fig. 63 (Hu and Talke, 1988 and Novotny and Karis, 1991a; Novotny et al., 1991b). The lubricant polymer chain is scissioned during slider-disk contacts. The transient interface temperatures may be high enough (Bhushan, 1992b) to lead to the direct evaporation or desorption of the original lubricant molecules. Novotny and Karis (1991a) studied the difference between mechanisms by comparison of the Fomblin Y and Z lubricants with different chemical structure but approximately the same molecular weight:

1. In sliding, Y is removed from the surface more rapidly than Z.
2. During flying, Z is removed more rapidly from the surface than Y.
3. Z is thermally decomposed more rapidly than Y (Paciorek et al., 1979; Kasai et al., 1991a, 1991b).
4. Migration rates of Z are faster than Y.

Polymer chain scission can be driven by mechanical, triboelectric, or thermal mechanisms. (Carre, 1986; Kimachi et al., 1988; Kasai et al., 1991a; Novonty and Karis, 1991a). The lubricant thickness is typically 1 to 5 nm, and at 1 to 5 m/s sliding velocity, the shear rate is 0.2 to 5 x $10^9 s^{-e}$. At such high shear rates, there can be much energy imparted to the polymer chain, inducing the chains to slide over one another. If the molecules are strongly interacting with the surface, the sliding may be hindered. Additional hindrance to interchain sliding can be the presence of a bulky side group such as the $-CF_3$ group on the Y lubricant. The additional intermolecular friction of the chains with the side group can hinder the rapid configurational adjustments required to support such high shear rates. The chains can be broken by tearing bonds apart along the polymer backbone and reduced to volatile products which provide the route for the observed lubricant loss when contacts occur in sliding or flying. It also follows that the loss of Y with the side group should be more rapid than that of the linear Z lubricant (Novotny and Karis, 1991a).

Moreover, potential differences up to 0.1 V and 0.5 V are measured on thin-film and particulate media surfaces, respectively, between areas on and off the sliding tracks. Corresponding electric fields for slider-disk separations lead to alteration of organic materials when the localized currents pass between the disk and slider asperities (Novotny and Karis, 1991a).

Thermal decomposition of perfluoropolyethers can proceed by a free radical mechanism which involves initiation, propagation, and termination. The relative rates of initiation and propagation should be different for the two lubricants because of their chemical structure. From thermogravimetric analysis, for an equivalent rate of thermal decomposition on iron oxide, the temperature of Y must be held about 100°C higher than that of Z. However, the loss of lubricant by thermal decomposition may also depend on the relative displacement

and migration rates. Thus, the higher loss rate of Y than Z in sliding can also be consistent with the thermal decomposition pathway. Degradation of Z lubes faster than Y lubes is believed to be due to the catalytic effect of high concentration of acetal units (CF_2 - O) in Z lubes which results in chain scission (Kasai et al., 1991a; Novotny and Karis, 1991a).

In flying, there is a displacement of lubricant to the outside of the track and a decrease of lubricant in the track. The displacement can be attributed to the repetitive application of pressure in the slider air bearing and intermittent contacts between the slider rails and the disk. Taking into account both the displacement and decrease in the lubricant level during flying, there is a net loss of lubricant from the disk surface. One possible mechanism for the loss is by aerosol droplet formation. This mechanism is reasonable given the tremendous negative pressure gradient at the trailing edge of the slider. The typical pressure increase under the air bearing rail is about 10^5 Pa, and this pressure drop occurs over about 10 μm, yielding a trailing edge pressure gradient of 10^{10} Pa/m. The rate of aerosol generation can depend on the lubricant surface tension (which is about the same for Z and Y) and the lubricant molecular configuration (effectively a flow property). The -CF_3 side group can act to hinder chain slippage (flow) required for efficient aerosol generation, lowering the loss rate of Y below that of Z in flying (Novonty and Karis, 1991a).

REFERENCES

1. Blair, S., Green, I., and Bhushan, B. (1991), "Measurements of Asperity Temperatures of a Read/Write Head Slider Bearing in Hard Magnetic Recording Disks," *J. Trib., Trans.* ASME (in press).

2. Baker, H. R., Bascom, W. D., and Singleterry, C. R. (1962), "The Adhesion of Ice to Lubricated Surfaces," *J. Coll. Sci.*, **17**, 447-491.

3. Berry, M. V. and Lewis, Z. V. (1980), "On the Weierstrass-Mandelbrot Fractal Function," *Proc. Royal Soc.*, **A370**, pp. 459-484.

4. Bhushan, B. (1984), "Analysis of the Real Area of Contact Between a Polymeric Magnetic Medium and a Rigid Surface," *J. Lub. Tech.*, Trans. ASME, Vol. 106, pp. 26-34.

5. Bhushan, B. (1985a), "The Real Area of Contact of Polymeric Magnetic Media—II: Experimental Data and Analysis," *ASLE Trans.*, Vol. 28, pp. 181-197.

6. Bhushan, B. (1985b), "Assessment of Accelerated Head-Wear Test Methods and Wear Mechanisms," *Tribology and Mechanics of Magnetic Storage Systems*, Vol. 2 (B. Bhushan and N. S. Eiss, eds.), SP-19, pp. 101-111, ASLE, Park Ridge, Illinois.

7. Bhushan, B. (1987a), "Magnetic Head-Media Interface Temperatures, Part I—Analysis," *J. Trib.*, Trans. ASME, Vol. 109, pp. 243-251.

8. Bhushan, B. (1987b), "Magnetic Head-Media Interface Temperatures, Part II—Application to Magnetic Tapes," *J. Trib.*, Trans ASME, Vol. 109, pp. 252-256.

9. Bhushan, B. (1990), *Tribology and Mechanics of Magnetic Storage Devices*, Springer-Verlag, New York.

10. Bhushan, B. (1992a), *Mechanics and Reliability of Flexible Magnetic Media*, Springer-Verlag, New York.

11. Bhushan, B. (1992b), "Magnetic Head-Media Interface temperatures, Part 3—Application to Rigid Disks," *J. Trib.*, Trans. ASME (in press).

12. Bhushan, B. and Blackman, G. S. (1991), "Atomic Force Microscopy of Magnetic Rigid Disks and Sliders and its Application to Tribology," *J. Trib.*, Trans. ASME (in press).

13. Bhushan, B., Bradshaw, R. L., and Sharma, B. S. (1984a), "Friction in Magnetic Tapes II: Role of Physical Properties," *ASLE Trans.*, Vol. 27, pp. 89-100.

14. Bhushan, B. and Doerner, M. F. (1989), "Role of Mechanical Properties and Surface Texture in the Real Area of Contact of Magnetic Rigid Disks," *J. Trib.*, Trans. ASME, Vol. 111, pp. 452-458.

15. Bhushan, B. and Dugger, M. T. (1990a), "Real Contact Area Measurements on Magnetic Rigid Disks," *Wear*, Vol. 137, pp. 41-50.

16. Bhushan, B. and Dugger M. T. (1990b), "Liquid-Mediated Adhesion Measurements at the Thin-Film Magnetic Disk/Head Slider Interface," *J. Trib.*, Trans. ASME, Vol. 112, pp. 217-223.

17. Bhushan, B. and Gupta, B. K. (1991), Handbook of Tribology: Materials, Coatings, and Surface Treatments, McGraw-Hill, New York.

18. Bhushan, B. and Majumdar, A. (1991), "Elastic-Plastic Contact Model of Bifractal Surfaces," *Wear* (in press).

19. Bhushan, B., Nelson, G. W., and Wacks, M. E. (1986), "Head-Wear Measurements by Autoradiography of the Worn Magnetic Tapes," *J. Trib.*, Trans. ASME, Vol. 108, pp. 241-255.

20. Bhushan, B. and Phelan, R. M. (1986), "Frictional Properties as a Function of Physical and Chemical Changes in Magnetic Tapes During Wear," *ASLE Trans.*, Vol. 20, pp. 402-413.

21. Bhushan, B., Sharma, B. S., and Bradshaw, R. L. (1984b), "Friction in Magnetic Tapes I: Assessment of Relevant Theory," *ASLE Trans.*, Vol 27, pp. 33-44.

22. Bhushan, B., Wyant, J. C., and Meiling, J. (1988), "A New Three-Dimensional Digital Optical Profiler," *Wear*, Vol. 122, pp. 301-312.

23. Blackman, G. S., Mate, C. M., and Philpott, M. R. (1990), "Interaction Forces of a Sharp Tungsten Tip with Molecular Films on Silicon Substrate," *Phys. Rev. Lett.*, Vol. 65, pp. 2270-2273.

24. Bowden, F. P. and Tabor, D. (1950), *Friction and Lubrication of Solids*, Part I, Clarendon Press, Oxford, U. K.

25. Bradshaw, R. L. and Bhushan, B. (1984), "Friction in Magnetic Tapes Part III: Role of Chemical Properties," *ASLE Trans.*, Vol. 27, pp. 207-219.

26. Bradshaw, R. L. and Bhushan, B., Kalthoff, C., and Warne, M. (1986), "Chemical and Mechanical Performance of Flexible Magnetic Media Containing Chromium Dioxide," *IBM J. Res. Develop.*, Vol. 30, pp. 203-216.

27. Calabrese, S. J., Bhushan, B., and Davis, R. E. (1989), "A Study of Scanning Electron Microscopy of Magnetic Head-Tape Interface Sliding," *Wear*, Vol. 131, pp. 123-133.

28. Calabrese, S. J., Bhushan, B. (1990), "A Study of Scanning Electron Microscopy of Magnetic Head-Disk Interface Sliding," *Wear*, Vol. 139, pp. 367-381.

29. Camras, M. (1988), *Magnetic Recording Handbook*, Van Nostrand Reinhold, New York.

30. Cantow, M. J. R., Larrabee, R. B., Barrall, E. M., Butner, R. S., Cotts, P., Levy, F., and Ting, T. Y., (1986), "Molecular Weights and Molecular Dimensions of Perfluoropolyether Fluids," *Makromol. Chem.* Vol. 187, pp. 2475-2481.

31. Carre, D. J. (1986), "Perfluoropolyether Oil Degradation—Inference of FeF_3 Formation on Steel Surfaces Under Boundary Lubrication," *ASLE Trans.*, Vol. 29, pp. 121-125.

32. Chandrasekar, S., Shaw, M. C., and Bhushan, B. (1987a), "Comparison of Grinding and Lapping of Ferrites and Metals," *J. Eng. for Indus.*, Trans. ASME, Vol. 109, pp. 76-82.

33. Chandrasekar, S., Shaw, M. C., and Bhushan, B. (1987b), "Morphology of Ground and Lapped Surfaces of Ferrite and Metal," *J. Eng. for Indus.*, Trans. ASME, Vol. 109, pp. 83-86.

34 Chandrasekar, S. and Bhushan, B. (1988), "Control of Surface Finishing Residual Stresses in Magnetic Recording Head Materials," *J. Eng. for Indus.*, Trans. ASME, Vol. 110, pp. 87-92.

35. Chandrasekar, S., Kokini, K., and Bhushan, B. (1990), "Influence of Abrasive Properties on Residual Stresses in Lapped Ferrite and Alumina," *J. Amer. Ceramic Soc.*, Vol. 73, pp. 1907-1911.

36. Chandrasekar, S. and Bhushan, B. (1991), "Friction and Wear of Ceramics for Magnetic Recording Applications—Part II: Friction Measurements," *J. Trib.*, Trans. ASME, Vol. 113, pp. 313-317.

37. Chu, M. Y., De Jonghe, L., and Bhushan, B. (1990), "Wear Behavior of Ceramic Sliders in Sliding Contact with Rigid Magnetic Thin-Film Disks," *Tribology and Mechanics of Magnetic Storage Systems* (B. Bhushan, Ed.), Vol. 7, pp. 9-16, STLE, Park Ridge, Illinois.

38. Corti, C. and Savelli, P. (1989), "Perfluoropolyether Lubricants: Physical and Tribology Performances and Applications," *Proc. 5th Int. Congress on Tribology* (K. Holmberg and I. Nieminen, eds.), Vol. 5, pp. 155-164, Finnish Society for Tribology, Helsinki, Finland.

39. Dimigen, H. and Hubsch, H. (1983-84), "Applying Low-Friction Wear-Resistant Thin Solid Films by Physical Vapour Deposition," *Philips Tech. Rev.*, Vol. 41, pp. 186-197.

40. Doan, T. Q., and Mackintosh, N. D. (1988), "The Frictional Behaviour of Rigid-Disk Carbon Overcoats," *Tribology and Mechanics of Magnetic Storage Systems*, Vol. 5, (B. Bhushan, and N. S. Eiss, eds.), SP-25, pp. 6-11, STLE, Park Ridge, Illinois.

41. Dugger, M. T., Chung, Y. W., Bhushan, B., and Rothschild, W. (1990), "Friction, Wear, and Interfacial Chemistry in Thin-Film Magnetic Rigid Disk Files," *J. Trib.*, Trans. ASME, Vol. 112, pp. 238-245.

42. Dugger, M. T., Wahl, K. J., Chung, Y. W., Bhushan, B., and Rothschild, W. (1991), "An Investigation of Environmental Effects on the Wear and Surface Composition of Thin-Film Magnetic Disks," *Advances in Engineering Tribology*, (Y. W. Chung and H. S. Cheng, eds.), STLE, Park Ridge, Illinois (in press).

43. Dugger, M. T., Chung, Y. W., Bhushan, B., and Rothschild, W. (1992), "Wear Mechanisms of Amorphous Carbon and Zirconia Coatings on Rigid Disk Magnetic Recording Media," *Tribology Trans.* (in press).

44. Engel, P. A. and Bhushan, B. (1990), "Sliding Failure Model for Magnetic Head-Disk Interface," *J. Trib.*, Trans. ASME, Vol. 112, pp. 299-303.

45. Greenwood, J. A. and Williamson, J. B. P. (1966), "Contact of Nominally Flat Surfaces," *Proc. Roy. Soc. (Lond.)*, Vol. A295, pp. 300-319.

46. Gulino, R., Bair, S., Winer, W. O., and Bhushan, B. (1986), "Temperature Measurement of Microscopic Areas within a Simulated Head/Tape Interface Using Infrared Radiometric Technique," *J. Trib.*, Trans. ASME, Vol. 108, pp. 29-34.

47. Hahn, F. W. (1984), "Head Wear as a Function of Isolated Asperities on the Surface of Magnetic Tape," *IEEE Trans. on Magn.*, Vol. Mag-20, pp. 918-920.

48. Hedenqvist, P., Olsson, M., Hogmark, S. and Bhushan, B. (1991), "Tribological Studies of Various Magnetic Heads and Thin-Film Rigid Disks," *Wear* (in press).

49. Heidemann, R. and Wirth, M. (1984), "Transforming the Lubricant on a Magnetic Disk into a Solid Fluorine Compound," *IBM Tech. Disclosure Bull.*, Vol. 27, pp. 3199-3205.

50. Hiller, B. and Singh, G. P. (1991), "Interaction of Contaminant Particles with the Particular Slider/Disk Interface," *Adv. Info. Storage Syst.*, Vol. 2, pp. 173-180.

51. Hoagland, A. S. (1963), *Digital Magnetic Recording*, Wiley, New York.

52. Homola, A. M., Lin, L. J., and Saperstein, D. D. (1990), "Process for Bonding Lubricant to a Thin-Film Magnetic Recording Disk," U. S. Patent 4, 960, 609, Oct. 2.

53. Homola, A. M., Nguyen, H. V., and Hadzilоannous, G. (1991), "Influence of Monomer Architecture on the Shear Properties of Molecularly Thin Polymer Melts." *J. Chem. Phys.* (in press).

54. Hoshino, M., Kimachi, Y., Yoshimura, F., and Terada, A. (1988), "Lubrication Layer Using Perfluoropolyether and Aminosilane for Magnetic Recording Media," *Tribology and Mechanics of Magnetic Storage Systems*, Vol. 5 (B. Bhushan and N. W. Eiss, eds.), SP-25, pp. 37-42, STLE, Park Ridge, Illinois.

55. Hu, Y. and Talke, F. E. (1988), "A Study of Lubricant Loss in the Rail Region of a Magnetic Recording Slider Using Ellipsometry," *Tribology and Mechanics of Magnetic Storage Systems*, Vol. 5 (B. Bhushan and N. W. Eiss, eds.), SP-25, pp. 43-48, STLE, Park Ridge, Illinois.

56. Israelachvili, J. N., McGuiggan, P. M., and Homola, A. M. (1988), "Dynamic Properties of Molecularly Thin Liquid Films," *Science*, Vol. 240, pp. 189-191.

57. Iwasaki, S. and Nakamura, Y. (1977), "An Analysis for the Magnetization Mode for High Density Magnetic Recording," *IEEE Trans. Magn.*, Vol. Mag-13, pp. 1272-1277.

58. Jorgensen, F. (1988), *The Complete Handbook of Magnetic Recording*, Third ed., Tab Books Inc., Blue Ridge Summit, Pennsylvania.

59. Karis, T. E., Novotny, V. J., and Crone, R. M. (1990), "Sliding Wear Mechanism of Particulate Magnetic Recording Media," *Tribology and Mechanics of Magnetic Storage Systems*, Vol. 7 (B. Bhushan, ed.), SP-29, pp. 35-42, STLE, Park Ridge, Illinois.

60. Kasai, P. H., Tang, W. T., and Wheeler, P. (1991), "Degradation of Perfluoropolyethers Catalyzed by Aluminium Oxide," *Appl. Surf. Sci.* (in press).

61. Khan, M. R., Heiman, N., Fisher, R. D., Smith, S., Smallen, M., Hughes, G. F., Veirs, K., Marchon, B., Ogletree, D. F., Salmeron, M., and Siekhaus, W. (1988), "Carbon Overcoat

and the Process Dependence on its Microstructure and Wear Characteristics," *IEEE Trans. on Magn.*, Vol. 24, pp. 2647-2649.

62. Kimachi, Y., Yoshimura, F., Hoshino, M., and Terada, A. (1987), "Uniformity Quantification of Lubricant Layer on Magnetic Recording Media," *IEEE Trans. Magn.*, Vol. Mag-23, pp. 2392-2394.

63. Koka, R. and Kumaran, A. R. (1991), "Visualization and Analysis of Particulate Buildup on the Leading Edge Tapers of Sliders," *Adv. Info. Storage Syst.*, Vol. 2, pp. 161-171.

64. Lauer, J. L. and Jones, W. R. (1986), "Friction Polymers," *Tribology and Mechanics of Magnetic Storage Systems*, Vol. 3 (B. Bhushan, and N. S. Eiss, eds.), SP-21, pp. 14-23, ASLE, Park Ridge, Illinois.

65. Liu, C. C. and Mee, P. B. (1983), "Stiction at the Winchester Head-Disk Interface," *IEEE Trans. Magn.*, Vol. Mag-19, pp. 1659-1661.

66. Lowman, C. E. (1972), *Magnetic Recording*, McGraw Hill, New York.

67. Majumdar, A. and Bhushan, B. (1990), "Role of Fractal Geometry in Roughness Characterization and Contact Mechanics of Surfaces," *J. Trib.*, Trans. ASME, Vol. 112, pp. 205-216.

68. Majumdar, A. and Bhushan, B. (1991a), "Fractal Model of Elastic-Plastic Contact Between Rough Surfaces," *J. Trib.*, Trans. ASME, Vol. 113, pp. 1-11.

69. Majumdar, A., Bhushan, B., and Tien, C. L. (1991b), "Role of Fractal Geometry in Tribology," *Adv. Info. Storage Syst.*, Vol. 1, pp. 231-266.

70. Mandelbrot, B. B. (1982), *The Fractal Geometry of Nature*, W. H. Freeman, New York.

71. Marchon, B., Heiman, N., and Khan, M. R. (1990), "Evidence for Tribochemical Wear on Amorphous Carbon Thin Films," *IEEE Trans. Magn.*, Vol. 26, pp. 168-170.

72. Mate, C. M., Lorenz, M. R., and Novotny, V. J. (1989), "Atomic Force Microscopy of Polymeric Liquid Films," *J. Chem. Phys.*, Vol. 90, pp. 7550-7555.

73. Mate, C. M., Lorenz, M. R., and Novotny, V. J. (1990), "Determination of Lubricant Film Thickness on a Particulate Disk by Atomic Force Microscopy," *IEEE Trans. Magn.*, Vol. Mag-26, pp. 1225-1228.

74. Matthewson, M. J. and Mamin, H. J. (1988), "Liquid-Mediated Adhesion of Ultra-Flat Solid Surfaces," *Proc. Mat. Res. Soc. Symp.*, Vol 119, pp. 87-92.

75. Mee, C. D. and Daniel, E. D. eds. (1990), *Magnetic Recording Handbook*, McGraw Hill, New York.

76. Memming, R., Tolle, H. J., and Wierenga, P. E. (1986), "Properties of Polymeric Layers of Hydrogenated Amorphous Carbon Produced by a Plasma-Activated Chemical Vapor Deposition Process II: Tribological and Mechanical Properties," *Thin Solid Films*, Vol. 143, pp. 31-41.

77. Miyamoto, T., Sato, I., and Ando, Y. (1988), "Friction and Wear Characteristics of Thin-Film Disk Media in Boundary Lubrication," *Tribology and Mechanics of Magnetic Storage Systems*, Vol. 5 (B. Bhushan, and N. S. Eiss, eds.), SP-25, pp. 55-61, STLE, Park Ridge, Illinois.

78. Miyamoto, T., Sato, I., and Ando, Y. (1990), "Interaction Force Between Thin-Film Disk Media and Elastic Solids Investigated by Atomic Force Microscopy," *J. Trib.*, Trans. ASME, Vol. 112, pp. 567-572.

79. Miyoshi, K., Buckley, D. H., Kusaka, T., Maeda, C., and Bhushan, B. (1988), "Effect of Water Vapor on Adhesion of Ceramic Oxide in Contact with Polymeric Magnetic Medium and Itself," *Tribology and Mechanics of Magnetic Storage Systems*, Vol. 5 (B. Bhushan, and N. S. Eiss, eds.), SP-25, pp. 12-16, STLE, Park Ridge, Illinois.

80. Miyoshi, K., Pouch, J. J., and Alterovitz, S. A. (1989), "Plasma-Deposited Amorphous Hydrogenated Carbon Films and Their Tribological Properties," Tech. Memo 102379, NASA Lewis Research Center, Cleveland, Ohio.

81. Nishihara, H. S., Dorius, L. K., Bolasna, S. A., and Best, G. L. (1988), "Performance Characteristics of IBM 3380K Air Bearing Design," *Tribology and Mechanics of Magnetic Storage Systems*, Vol. 5 (B. Bhushan, and N. S. Eiss, eds.), SP-25, pp. 117-123, STLE, Park Ridge, Illinois.

82. Novotny, V. J. and Karis, T. E. (1991a), "Sensitive Tribological Studies on Magnetic Recording Disks," *Adv. Info. Storage Syst.*, Vol. 2, pp. 137-152.

83. Novotny, V. J., Karis, T. E., and Johnson, N. W. (1991b), "Lubricant Removal, Degradation, and Recovery on Particulate Magnetic Recording Media," *J. Trib.*, Trans. ASME (in press).

84. Oden, P. I., Majumdar, A., Bhushan, B., Padmanabhan, A., and Graham, J. J. (1992), "AFM Imaging, Roughness Analysis and Contact Mechanics of Magnetic Tape and Head Surfaces," *J. Trib.*, Trans. ASME (in press).

85. Paciorek, K. J. L., Kratzer, R. H., Kaufman, J., and Nakahara, J. H. (1979), "Thermal Oxidative Studies of Poly (hexafluoropropene oxide) Fluids," *J. Appl. Poly. Sci.*, Vol. 24, pp. 1397-1411.

86. Rugar, D., Mamin, H. J., and Guethner, P. (1989), "Improved Fiber Optic Interferometer for Atomic Force Microscopy," *Appl. Phys. Lett.*, Vol. 55, pp. 2588-2590.

87. Saperstein, D. D. and Lin, L. J. (1990), "Improved Surface Adhesion and Coverage of Perfluoropolyeter Lubricants Following Far-UV Irradiation," *Langmuir*, Vol. 6, pp. 1522-1524.

88. Scarati, A. M. and Caporiccio, G. (1987), "Frictional Behavior and Wear Resistance of Rigid Disks Lubricated with Neutral and Functional Perfluoropolyethers," *IEEE Trans. Magn.*, Vol. Mag-23, pp. 106-108.

89. Sriram, T. S., Wahl, K. J., Chung, Y. W., Bhushan, B., and Rothschild, W. (1991), "The Application of Scanning Tunneling Microscopy to Study Lubricant Distribution of Magnetic Thin-Film Rigid Disk Surfaces," *J. Trib.*, Trans. ASME, Vol. 113, pp. 245-248.

90. Streator, J. L., Bhushan, B., and Bogy, D. B. (1991a), "Lubricant Performance in Magnetic Thin-Film Disks with Carbon Overcoat—Part I: Dynamic and Static Friction," *J. Trib.*, Trans. ASME, Vol. 113, pp. 22-31.

91. Streator, J. L., Bhushan, B., and Bogy, D. B. (1991b), "Liquid Performance in Magnetic Thin-Film Disks with Carbon Overcoat—Part II: Durability," *J. Trib.*, Trans. ASME, Vol. 113, pp. 32-37.

92. Strom, B. D., Bogy, D. B., Bhatia, C. S., and Bhushan, B. (1991), "Tribochemical Effects of Various Gases and Water Vapor on Thin-Film Magnetic Disks with Carbon Overcoats," *J. Trib.*, Trans. ASME (in press).

93. Suzuki, S. and Kennedy, F. E. (1991), "The Detection of Flash Temperatures in a Sliding Contact by the method of Tribo-Induced Thermoluminescence," *J. Trib.*, Trans. ASME, Vol. 113, pp. 120-127.

94. Tanaka, K. and Miyazaki, O. (1981), "Wear of Magnetic Materials and Audio Heads Sliding Against Magnetic Tapes," *Wear*, Vol. 66, pp. 289-306.

95 Van Gestel, W. J., Gorter, F. W., and Kuijk, K. E. (1977), "Read-out of a Magnetic Tape by the Magnetoresistive Effect," *Philips Tech. Rev.*, Vol. 37, No. 2/3, pp. 42-50.

96. Wahl, K. J., Chung, Y. W., Bhushan, B., and Rothschild, W. J. (1991a), "Durability of Magnetic Thin-Film Rigid Disks in Nitrogen and Helium Environments," *Adv. Info. Storage Syst.*, Vol. 1, pp. 327-336.

97. Wahl, K. J., Chung, Y. W., Bhushan, B., and Rothschild, W. J. (1991b), "In Situ Auger Measurements of Surface Chemical Changes of Magnetic Thin-Film Rigid Disks During Spherical Pin Sliding Tests," *Adv. Info. Storage Syst.*, Vol. 3 (in press).

98. Wallace, R. L. (1951), "The Reproduction of Magnetically Recorded Signal," *Bell Syst. Tech J.*, Vol. 30, pp. 1145-1173.

99. Yamashita, T., Chen, G. L., Shir, J., and Chen, T. (1988), "Sputtered ZrO_2 Overcoat with Superior Corrosion Protection and Mechanical Performance in Thin-Film Rigid Disk Application," *IEEE Trans. Magn.*, Vol. Mag-24, pp. 2629-2634.

100. Yanagisawa, M. (1985a), "Lubricants on Plated Magnetic Recording Disks," *Tribology and Mechanics of Magnetic Storage Systems*, Vol. 2 (B. Bhushan, and N. S. Eiss, eds.), SP-19, pp. 7-15, ASLE, Park Ridge, Illinois.

101. Yanagisawa, M. (1985b), "Tribological Properties of Spin-Coated SiO_2 Protective Film on Plated Magnetic Recording Disks," *Tribology and Mechanics of Magnetic Storage Systems*, Vol. 2 (B. Bhushan, and N. S. Eiss, eds.), SP-19, pp. 16-20, ASLE, Park Ridge, Illinois.

Table 1. Selected physical properties of hard head materials

Material	Density, kg/m^3	Young's Modulus, GPa	Knoop microhardness, GPa(kg/mm^2)	Flexural strength, MPa	Electrical resistivity, $\mu ohm \cdot cm$
Ni-Zn ferrite	4570	122	6.9(700)	150	10^{11} to 10^{13}
Mn-Zn ferrite	4570	122	5.9(600)	120	5×10^4 to 5×10^5
$Al_2O_3 - TiC(70 - 30)$	4220	450	22.6(2300)	880	2×10^3 to 3×10^3
$ZrO_2 - Y_2O_3(94 - 6)$	6360	210	12.8(1300)	500-700	10^{16}
$BaTiO_3$	4320	110	10.3(1050)		

Table 2. Typical operating conditions and typical materials used in different magnetic media for computer use

Magnetic medium	Normal pressure, kPa	Sliding speed, m/s	Flying height, μm	Substrate	Magnetic coating binder	Magnetic medium/ thickness	Solid/ liquid lubricant	Lubricant application method/quantity
Tape								
Particulate tape	7-28[a]	2-4	0.1-0.2	PET[b] 14.5-23.4 μm thick	Polyester-polyurethane[c]	γ-Fe_2O_3, Co-γ Fe_2O_3, CrO_2, Fe, or Ba0.6 Fe_2O_3/2-4 μm	Fatty-acid ester	Internal/ 1-3% by weight
Metal-evaporated tape				PET[b] 14.5 μm thick	None	Evaporated Co-Ni or Co-Cr/ 100-300 nm	Polymer or inorganic and perfluoropolyether	Solution or vacuum/10-50 nm, Topical/2-10 nm thick
Particulate flexible disk	14-70 (10-20g)	1-12 (300-1800 rpm)	Partial contact (< 0.1)	PET[b] 76.2 μm thick	Polyester-polyurethane and epoxy[c]	γ-Fe_2O_3, Co-γ Fe_2O_3, Ba0.6 Fe_2O_3, or Fe/2-4 μm	Fatty-acid ester	Internal/ 1-3% by weight
Rigid Disk								
Particulate	7-14 (9.5-15g)	10-60 2600-5000	0.15-0.3	Al-Mg 1.3-1.9 mm thick	Phenolic and epoxy[c]	γ-Fe_2O_3/ 0.75-2 μm	Perfluoropolyether	Topical/ 3-6 nm thick, 5-10 mg for 275-mm disk
Thin-film (metal)				Al-Mg 0.78-1.3 mm thick with 10-20 μm electroless Ni-P		Co-X (sputtered or plated)/ 25-100 nm	Diamondlike carbon, Y_2O_3-ZrO_2, SiO_2 and perfluoro polyether	Sputtered or spin coated/20-40 nm topical/1-4 nm thick
Thin-film (oxide)				Al-Mg 0.78-1.3 mm thick with 2-20 μm anodized layer		γ-Fe_2O_3 (sputtered)/ 100-150 nm	Diamondlike carbon, SiO_2 and perfluoro polyether	Sputtered or spin coated/20-40 nm topical/1-4 nm thick

[a]Interlayer pressure on a tape surface near the hub of a wound reel (end of tape) can be as high as 1.38 MPa.

[b]PET-Poly(ethylene terephthalate).

[c]Load-bearing alumina particles are added to increase the wear resistance of the medium.

Table 3. Selected physical properties of magnetic media and its components

Material	Density, kg/m^3	Young's modulus, GPa	Knoop microhardness, $GPa(kg/mm^2)$
$\gamma - Fe_2O_3$ particles	4800	—	11.8(1200)Est.
Magnetic flexible medium substrate	1520	2.75-4.50	0.20(20)
Particulate flexible medium coating	1660	1.25-2.25	0.25(25)
Al-Mg(96-4)	2700	70	0.88(90)
Particulate rigid disk coating	1810	10	0.50(51)
Ni-P (electroless plated)	8910	130	5.9-7.9(600-800)
CrO_3 anodized layer (alumite)	—	—	3-5(310-510)
Chemically strengthened alkali-aluminosilicate glass	2460	73	5.8(590)
Sputtered Co-Cr alloy metal film	8030	170-210	6.9-8.8(700-900)
Diamondlike carbon overcoat for metal film	2100	160-180	14.7-19.6(1500-2000)
Thin-film rigid disk structure	—	110-140	6-10(610-1020)

TABLE 4

Real area of contact calculations for magnetic tape A against Ni-Zn ferrite head and rigid disks against Al_2O_3 – TiC slider (Bhushan and Blackman, 1991 and Oden et al., 1992)

Component	E,	H,	σ, nm		σ_p, nm		$1/R_p$, 1/mm		η, $1/mm^2$		Ψ		A_r/A_aP_a, $1/GP_a$		n/A_aP_a, 1/mN		$A_r/_n$, μm^2		P_r, GPa	
Designation	GP_a	GP_a	NOP	AFM	NOP	AFM	NOP	AFM	NOP	AFM	NOP	AFM	NOP	AFM	NOP	AFM	NOP	AFM	NOP	AFM
Head-Tape Interface																				
Particulate tape A	1.75	0.25	19.5	36.3	19.0	45.4	2.20	1.4×10^5	5.7×10^3	8.0×10^6	0.05	50.17	241.6	4.0	9.6	4.1×10^{-2}	25.3	97.6	$4.23\ 10^{-3}$	0.25
Mn-Zn ferrite head	122	6.9	2.15	3.61	2.51	5.47	0.23	7.8×10^3	1.2×10^3	6.2×10^6										
Head-Disk Interface																				
Particulate disk A	9.4	0.53	9.39	13.o	9	10.5	4.79	6.0×10^3	5.9×10^3	2.4×10^6	0.27	5.6	22.6	1.9	9.9	0.13	2.3	14.2	0.05	0.53
Textured thin-film disk B_1	113	6.0	7.33	6.33	7	6.7	4.90	4.0×10^3	732	9.1×10^6	0.12	3.4	4.3	0.17	2.5	4.8×10^{-3}	1.7	28.2	0.24	6.0
Polished thin-film disk C	107	6.2	2.11	3.37	2	8.8	2.24	6.1×10^3	911	3.5×10^6	0.05	4.4	9.6	0.16	8.8	1.2×10^{-2}	1.1	13.9	0.11	6.2
Al_2O_3 – TiC Slider	450	22.6	1.63	1.55	2	1.4	0.53	1.2×10^3	2.4×10^3	13.3×10^6										

Table 5

Ranking of thin-film rigid disk and slider material combinations identified by *in situ* teting in the SEM (Calabrese and Bhushan, 1990)

Mn-Zn ferrite slider vs. disk B_2 (ZrO_2-Y_2O_3)	Some transfer to the edges. Low breakaway friction value. No damage to the disk surface. Large grain size particles do not appear to contribute to wear or damage during subsequent sliding. Rated best
Al_2O_3-TiC slider vs. disk B_2 (ZrO_2-Y_2O_3)	Transfer to the rail edges. Fine grain wear debris gets trapped at the interface and contributes to damage during subsequent sliding. Rated good
Mn-Zn ferrite slider vs. disk B_1 (DLC[a])	Transfer to the rails. No damage to the disk but disk is wearing. Rated fair
Al_2O_3-TiC slider vs. disk B_1 (DLC)	Transfer to the rails. Damage to the disk on breakaway. Rated fair.
Calcium titanate slider vs. disk B_1 (DLC)	Severe damage to the disk, heavy disk wear. Head slider appears to be cracking. Rated poor

[a]DLC - diamondlike carbon

Table 6

Chemical structure, molecular weight, and viscosity of perfluoropolyether lubricants

Lubricant	Formula	Molecular weight, Daltons	Kinematic viscosity, c St(mm^2/s)
Fomblin Z-25	$CF_3-O-(CF_2-CF_2-O)_m-(CF_2-O)_n-CF_3$	12,800	250
Fomblin Z-15	$CF_3-O-(CF_2-CF_2-O)_m-(CF_2-O)_n-CF_3$ (m/n ~ 2/3)	9100	150
Fomblin Z-03	$CF_3-O-(CF_2-CF_2-O)_m-(CF_2-O)_n-CF_3$	3600	30
Fomblin Z-DOL	$HO-CH_2-CF_2-O-(CF_2-CF_2-O)_m-(CF_2-O)_n-CF_2-CH_2-OH$	2000	80
Fomblin AM2001	Piperonyl-$O-CH_2-CF_2-O-(CF_2-CF_2-O)_m-(CF_2-O)_n-CF_2-CH_2-O$-piperonyl[1]	2300	80
Fomblin Z-DISOC	$O-CN-C_6H_3-(CH_3)-NH-CO-CF_2-O-(CF_2-CF_2-O)_n-(CF_2-O)_m-CF_2-CO-NH-C_6H_3-(CH_3)-N-CO$	1500	160
Fomblin YR	$CF_3-O-(C(CF_3)F-CF_2-O)_m-(CF_2-O)_n-CF_3$ (m/n ~ 40/1)	6800	1600
Demnum S-100	$CF_3-CF_2-CF_2-O-(CF_2-CF_2-CF_2-O)_m-CF_2-CF_3$	5600	250
Krytox 143AD	$CF_3-CF_2-CF_2-O-(C(CF_3)F-CF_2-O)_m-CF_2-CF_3$	2600	–

[1] 3,4 – methylenedioxybenzyl

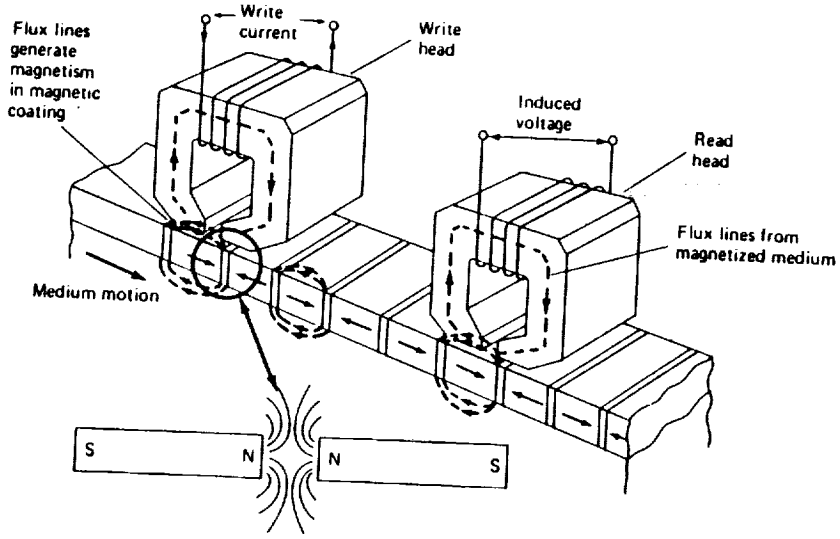


Figure 1. Principle of horizontal magnetic recording and playback (Bhushan, 1990).

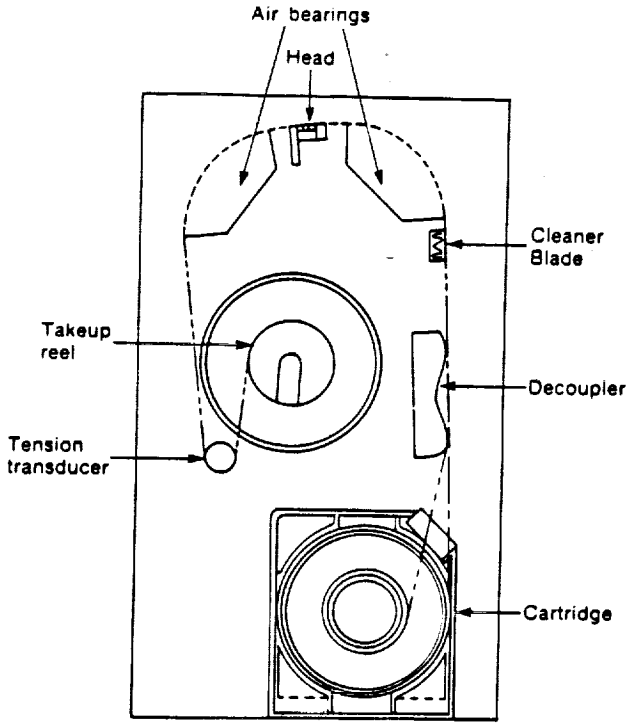


Figure 2. Schematic of tape path in an IBM 3480/3490 data-processing tape drive (Bhushan, 1990).

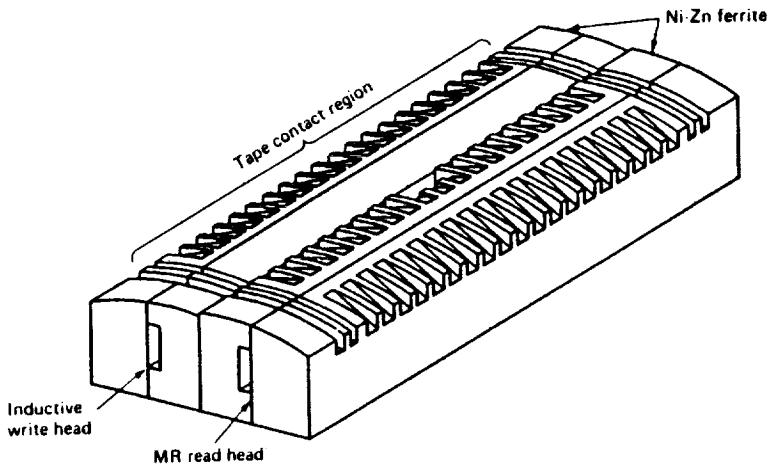


Figure 3. Schematic of a magnetic thin-film head (with a radius of cylindrical contour of about 20 mm) for an IBM 3480/3490 tape drive (Bhushan, 1990).

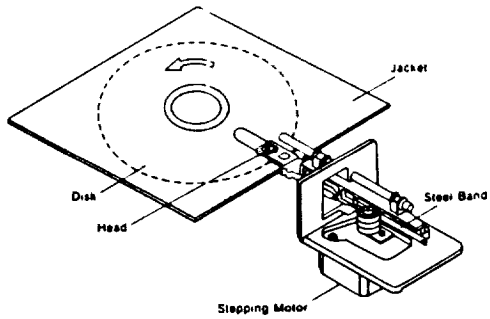


Figure 4. Schematic of the head-disk interface for flexible disk drives (Bhushan, 1990).

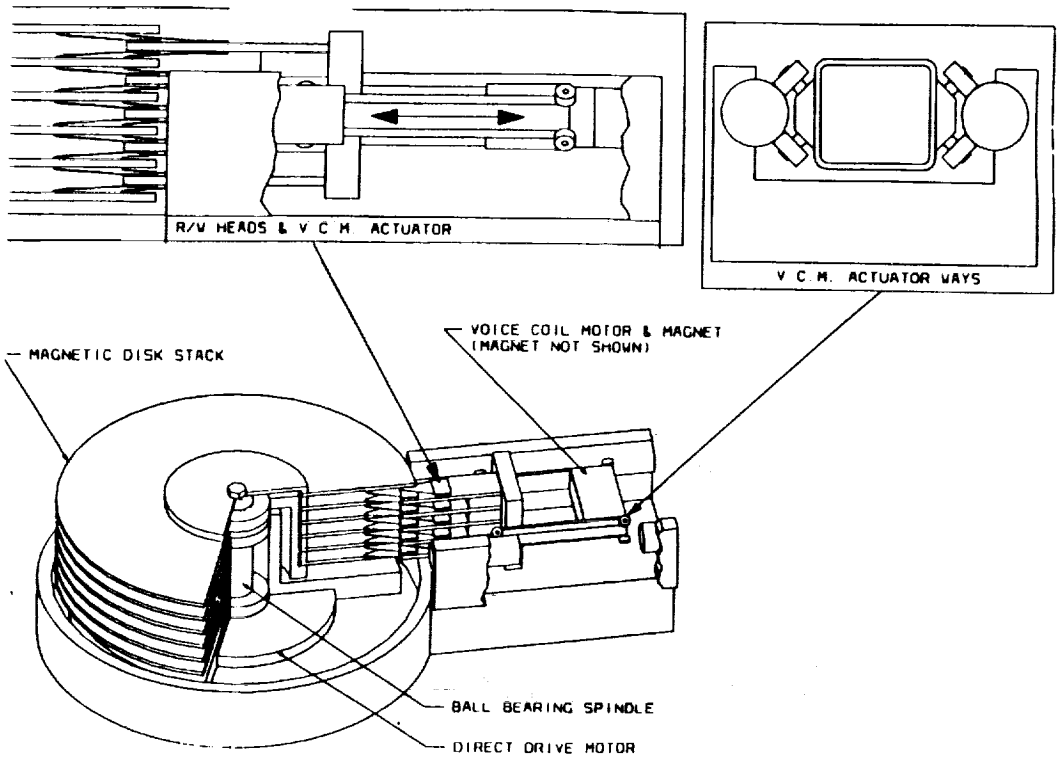


Figure 5. Schematic of the head-disk assembly in an IBM 3390-type rigid-disk drives consisting of the voice-coil-motor driven head-arm assemby and disk stack.

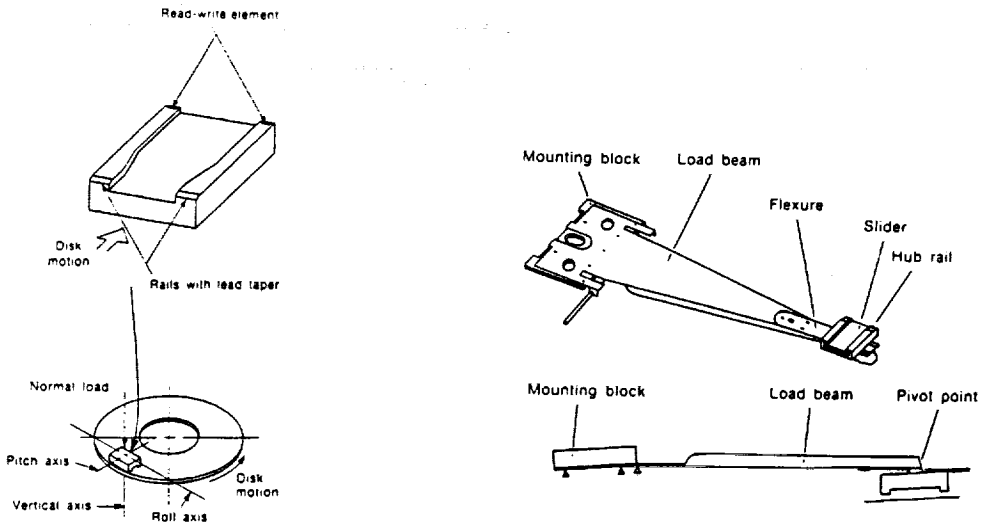


Figure 6. Schematics of (a) the self-acting IBM 3380K/3390-type head slider on a magnetic disk, (b) IBM 3370/3380/3390-type suspension-slider assembly.

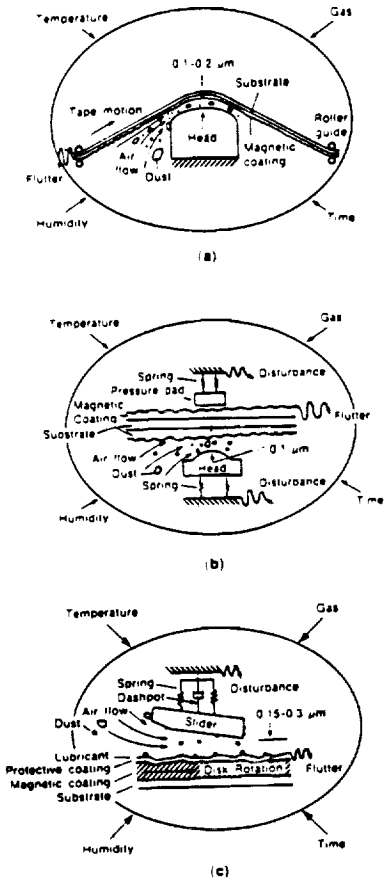
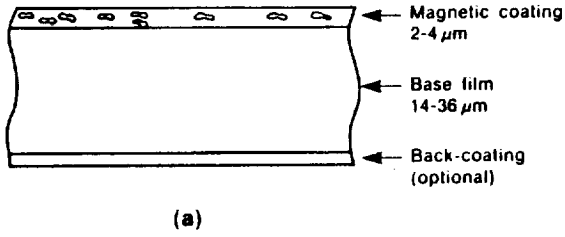


Figure 7. Schematic diagrams of various head-medium interfaces (a) head-tape interface, (b) head-flexible disk interface, and (c) head-rigid disk interface (Bhushan, 1990).



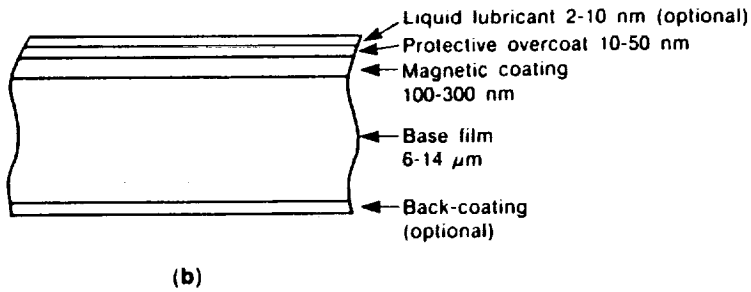


Figure 8. Sectional views of (a) a particulate, and (b) a thin-film magnetic tape.

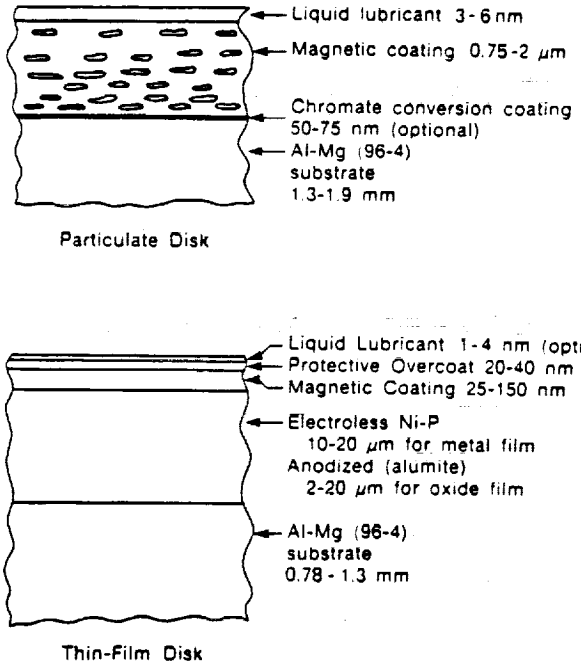


Figure 9. Sectional views of (a) a particulate, and (b) a thin-film magnetic rigid disk.

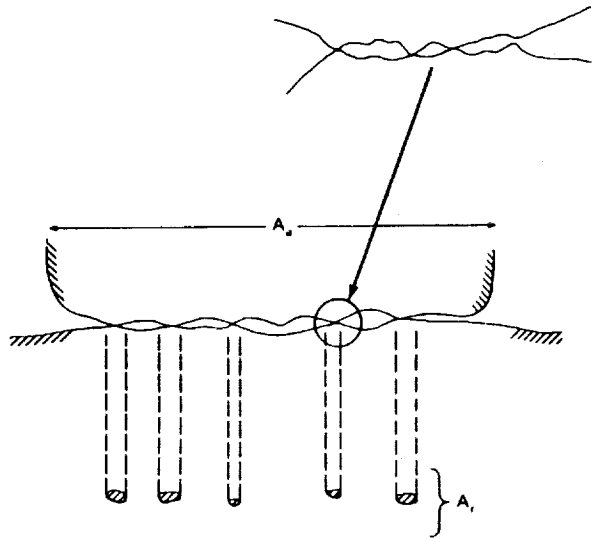


Figure 10. Schematic representation of an interface, showing the apparent and real areas of contact. Inset shows the detail of a contact on a submicron scale. Typical size of an asperity contact is from submicron to a few microns.

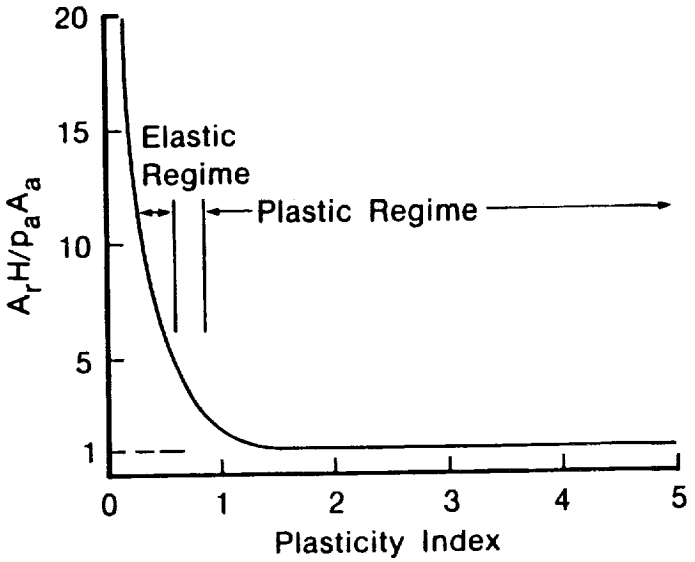


Figure 11. Influence of plasticity index on the real area of contact in metals/ceramics (Bhushan, 1990).

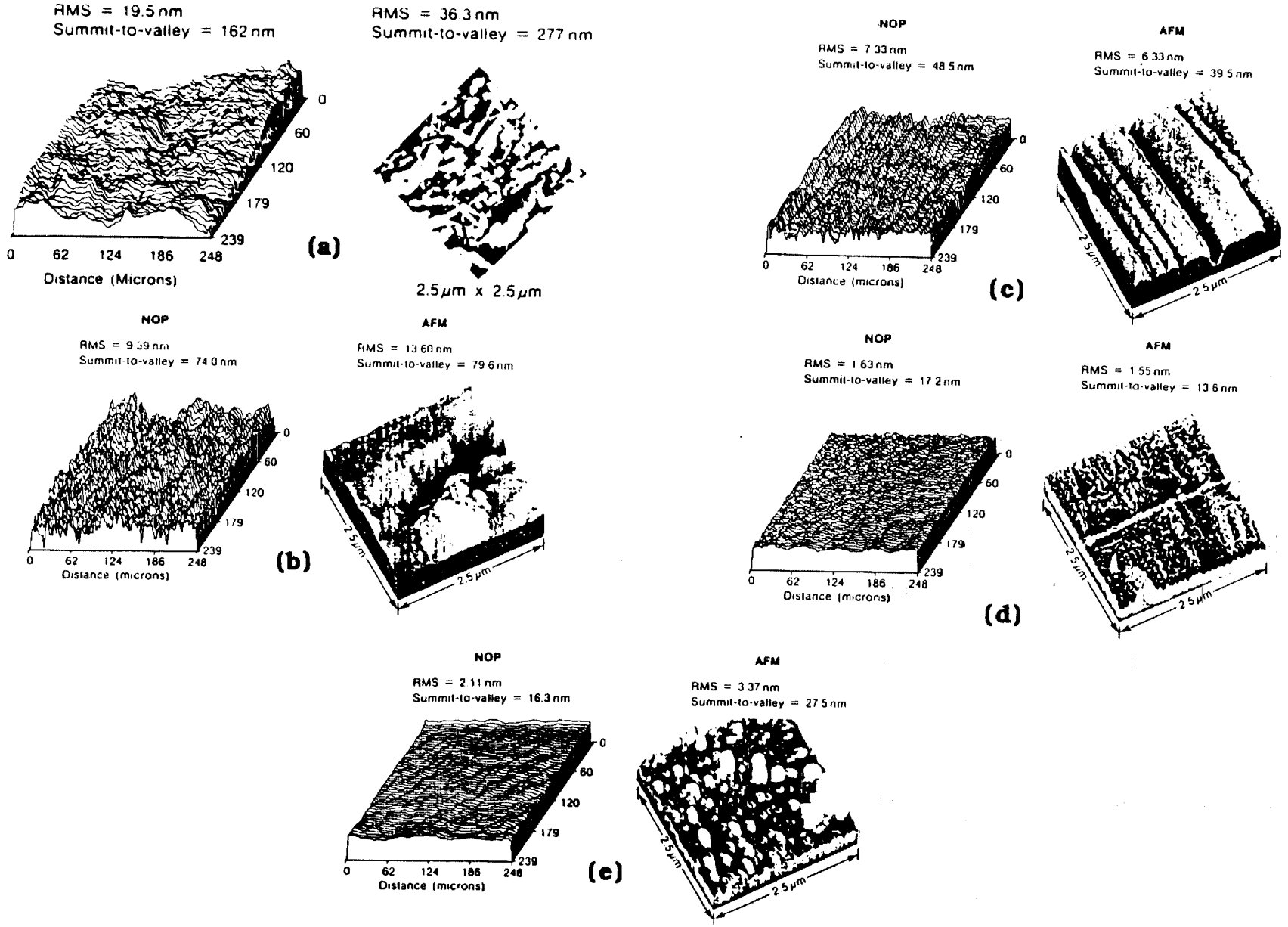


Figure 12. NOP and AFM images of (a) particulate tape A, (b) particulate disk A, (c) circumferentially-textured thin-film (sputtered Co-Pt-Ni) disk B_1, (d) polished thin-film (sputtered Co-Ni) disk C, and (e) Al_2O_3-TiC rigid-disk slider. The wireframe NOP images are of a 250 μm square region, and the AFM images are solid grey level images (white is high and black is low) of 2.5 μm regions of the same disk (Bhushan and Blackman, 1991).

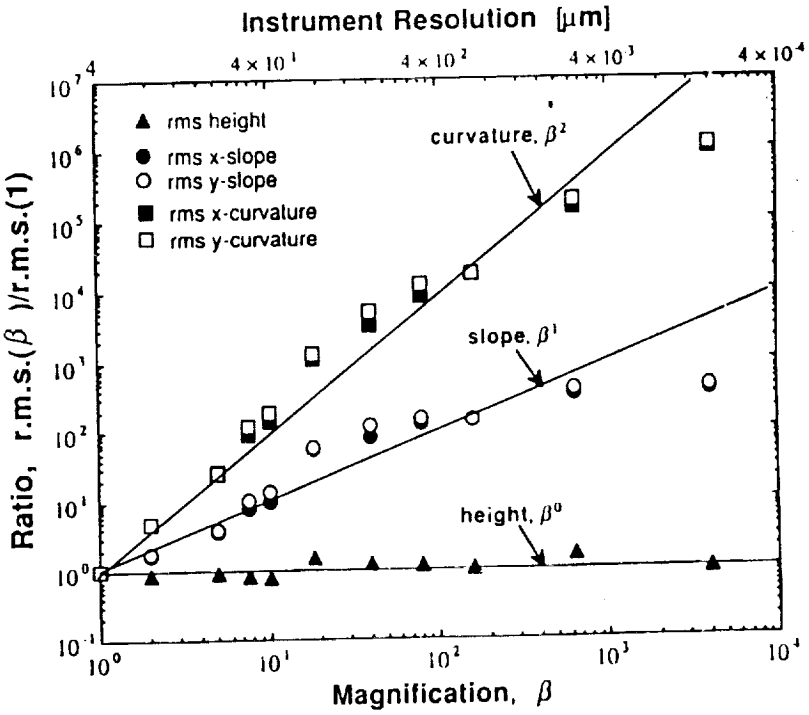


Figure 13. Variation of the ratio of the rms values at a surface magnification $\beta > 0$ to the corresponding rms values at a magnification of unity, with magnification β. The data for $\beta \leq 10$ was obtained by NOP measurements and that for $\beta \geq 10$ was obtained by AFM (Oden et al., 1992).

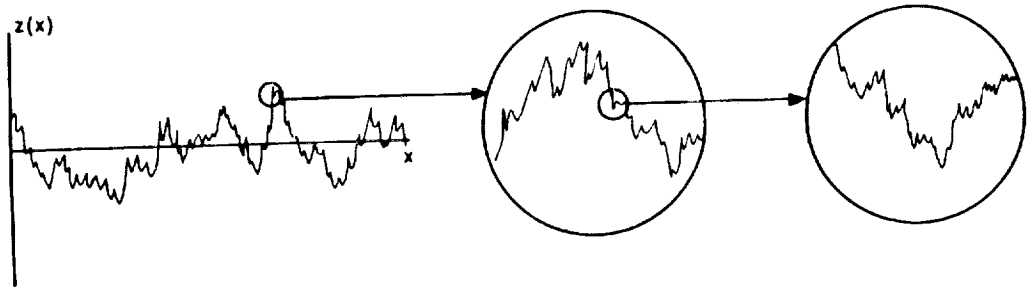


Figure 14. Quantitative description of statistical self-affinity for a surface profile.

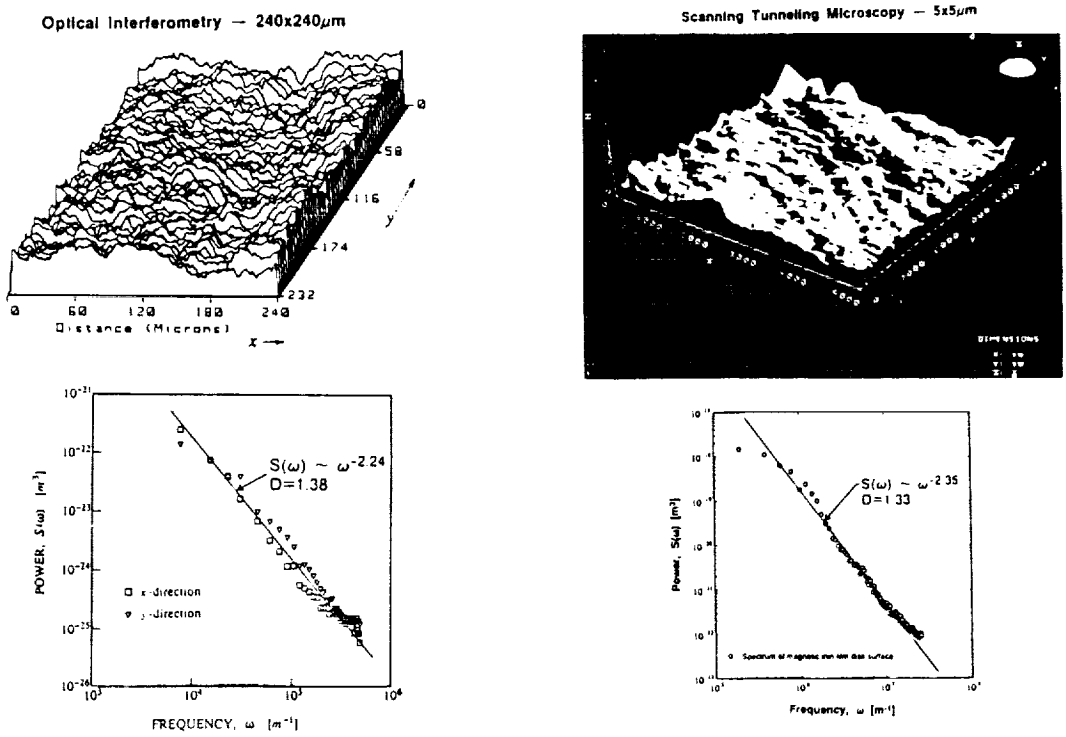


Figure 15. Surface topography and average power spectra at different length scales of a magnetic thin-film rigid disk C surface (a) NOP data, (b) STM data (Majumdar and Bhushan, 1990).

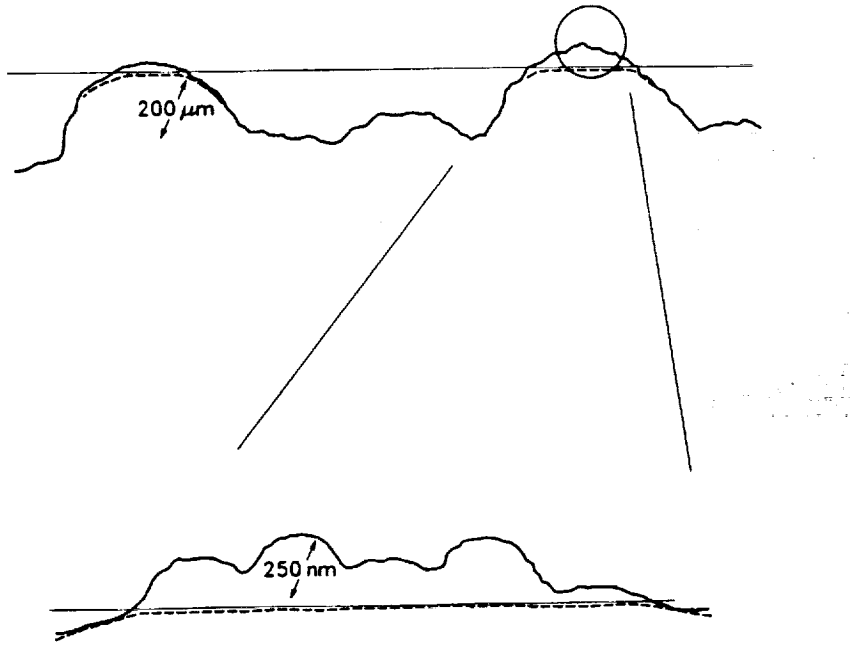


Figure 16. Schematic of local asperity deformation during contact of a rough surface (upper profile measured by NOP and lower profile measured by AFM, typical dimensions shown for a polished thin-film disk C) against a flat surface. The vertical axis is magnified for clarity. Firm lines show the surfaces before contact and dotted lines show surfaces after contact (Bhushan and Blackman, 1991).

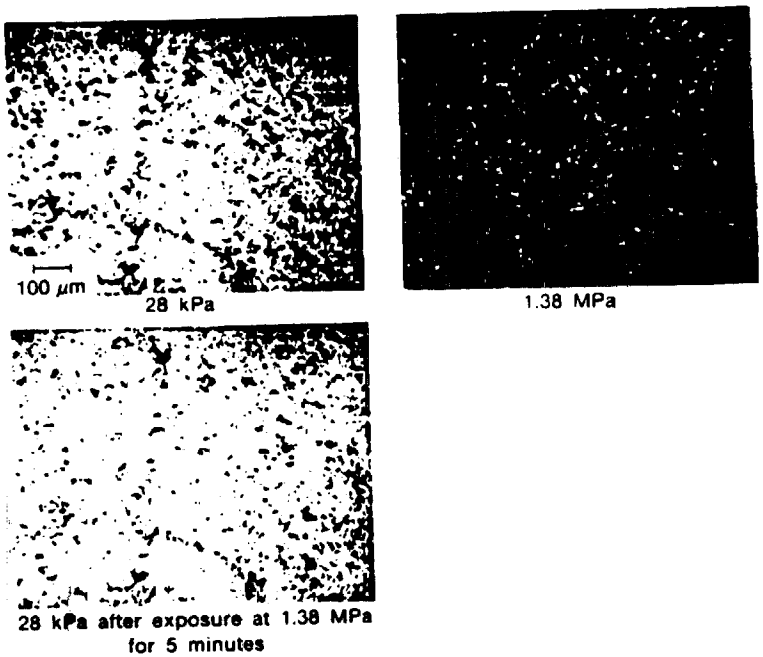


Figure 17. Optical interference photographs of tapes taken at 28 kPa; then subjected to higher pressure (1.38 MPa) for about 5 minutes and brought back to 28 kPa and rephotographed. We see no change in the real area of contact implying an elastic contact (Bhushan, 1984).

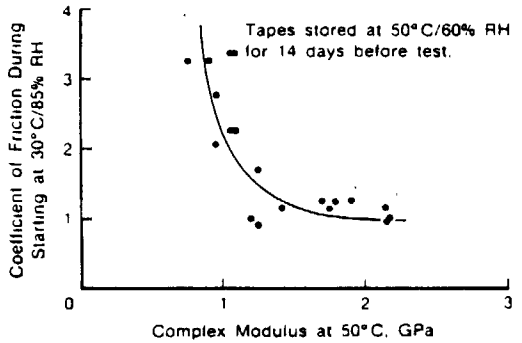


Figure 18a. Coefficient of friction during start at 30°C/85% RH measured on a commercial tape drive versus complex modulus at 50°C. The CrO_2 tapes were stored at 50°C/60% RH for 14 days before the tests (Bhushan et al., 1984a).

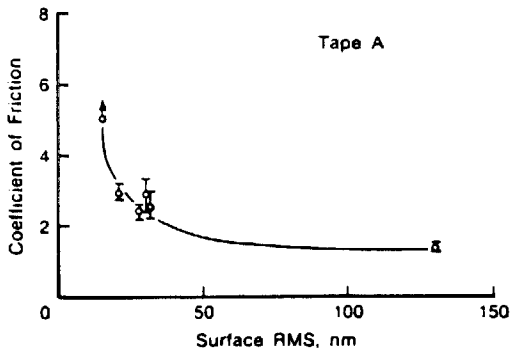


Figure 18b. Effect of surface roughness on coefficient of friction for CrO_2 particulate tapes (Bhushan et al., 1984a).

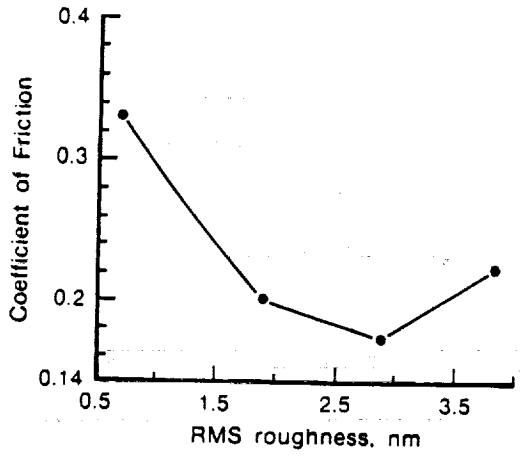


Figure 19. Coefficient of friction as a function of degree of disk texturing for a thin-film (metal) rigid disk with sputtered carbon overcoat against ferrite slider (Doan and Mackintosh, 1988).

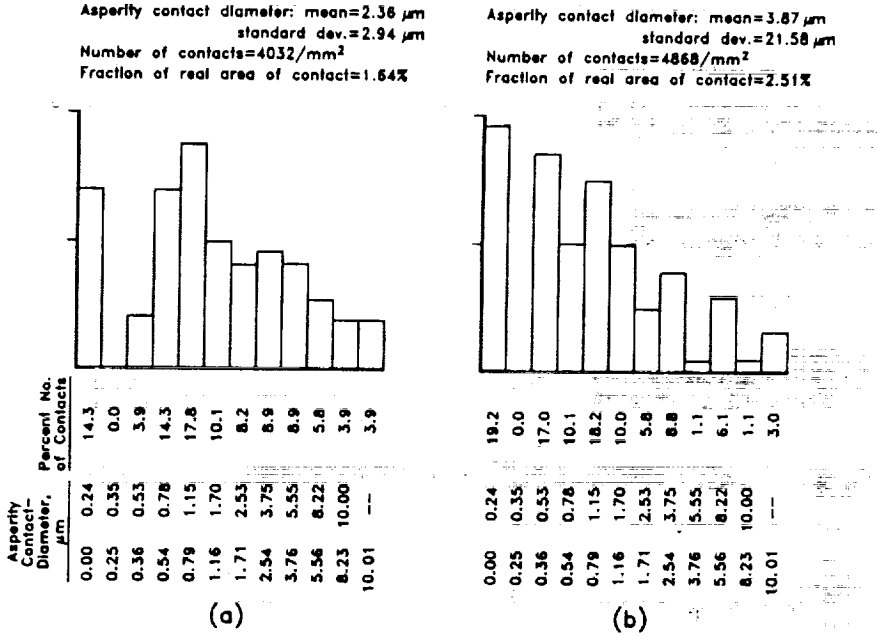


Figure 20. Log normal distribution of asperity contact diameters for a thin-film disk C loaded by 500 mN (9.27 MPa) at two loading durations (a) initial and (b) after loaded for 60 hours (Bhushan and Dugger, 1990a).

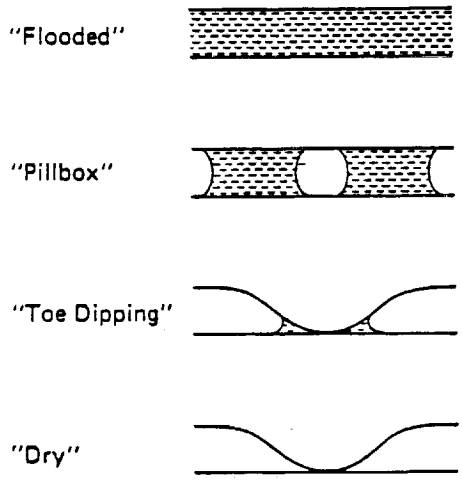


Figure 21. Regimes of different liquid levels in the head-medium interface.

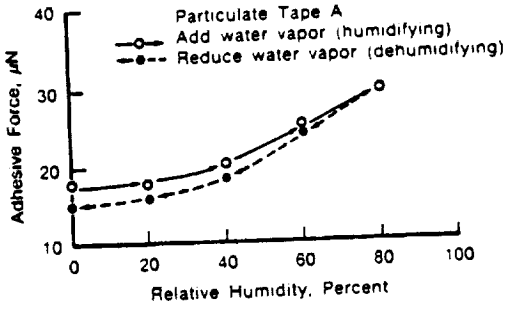


Figure 22. Effect of humidity on adhesion of CrO_2 tape A in contact with a Ni-Zn ferrite pin (Miyoshi et al., 1988).

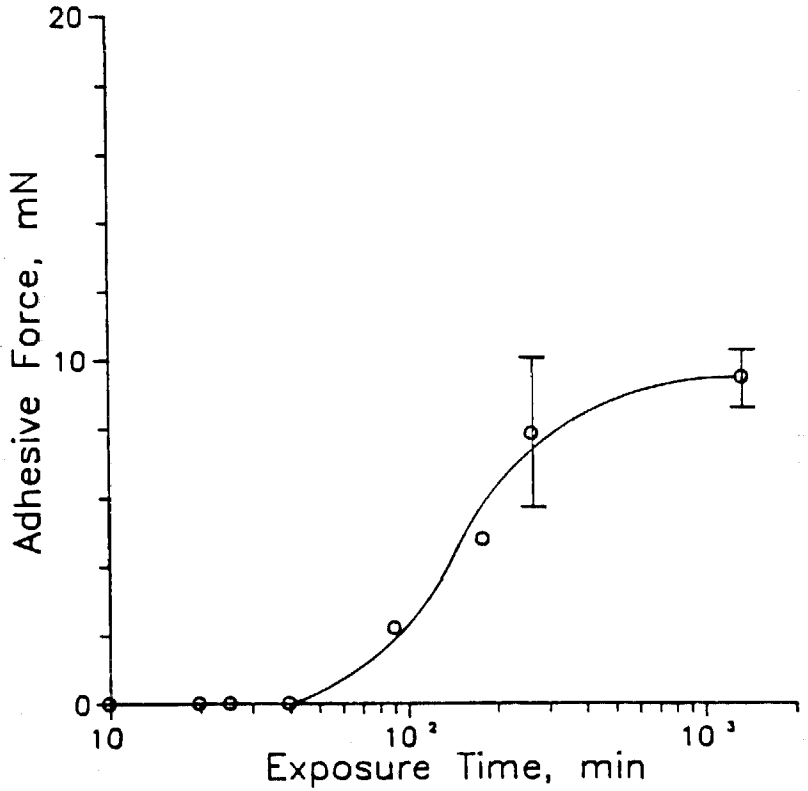


Figure 23. The effect of exposure time (before contact) of specimen surfaces to 90 percent relative humidity nitrogen on the adhesive force for an unlubricated polished thin-film disk ($\sigma = 2.11$ nm) disk. Adhesive force was measured after a 5 minute contact with Al_2O_3-TiC slider at 150 mN load, followed by separation at a normal velocity of 80 μm/s (Bhushan and Dugger, 1990b).

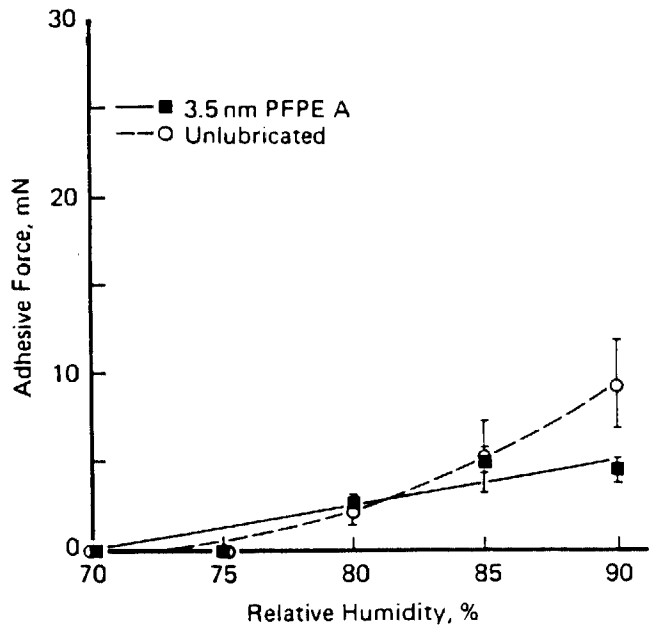


Figure 24. The dependence of the adhesive force on relative humidity for the polished thin film disk lubricated with 2 nm of Z-15 PFPE lubricant (150 cSt). Adhesive force was measured after a 5 min. contact with Al_2O_3-TiC slider at 150 mN load, followed by separation at a normal velocity of 80 μm/s (Bhushan and Dugger, 1990b).

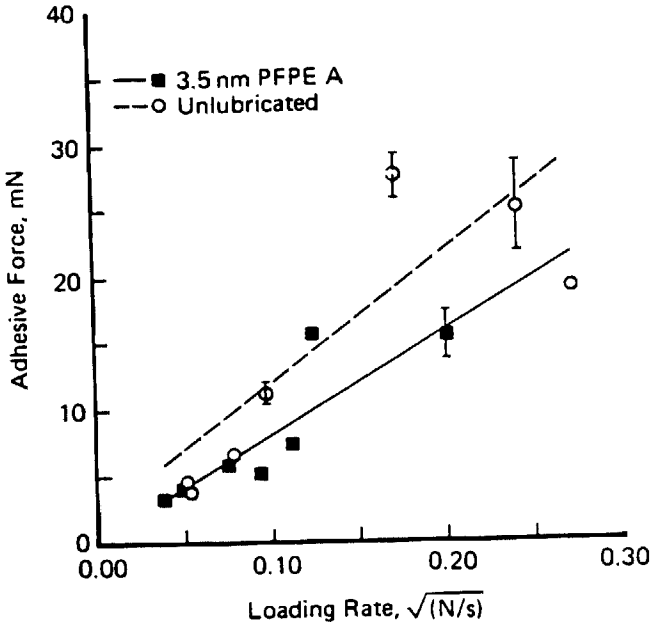


Figure 25. The dependence of adhesion on $\sqrt{(\text{loading rate})}$ for the same system as Fig. 24. 150 mN normal load applied for 5 minutes at 90 percent relative humidity (Bhushan and Dugger, 1990b).

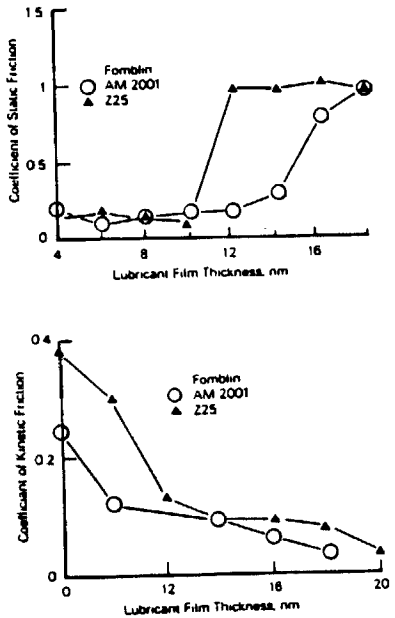


Figure 26. Coefficients of static and kinetic friction as a function of lubricant thickness for polar (AM 2001) and non-polar (Z-25) PFPE on particulate rigid disks sliding against Mn-Zn ferrite slider (Scarati and Caporiccio, 1987).

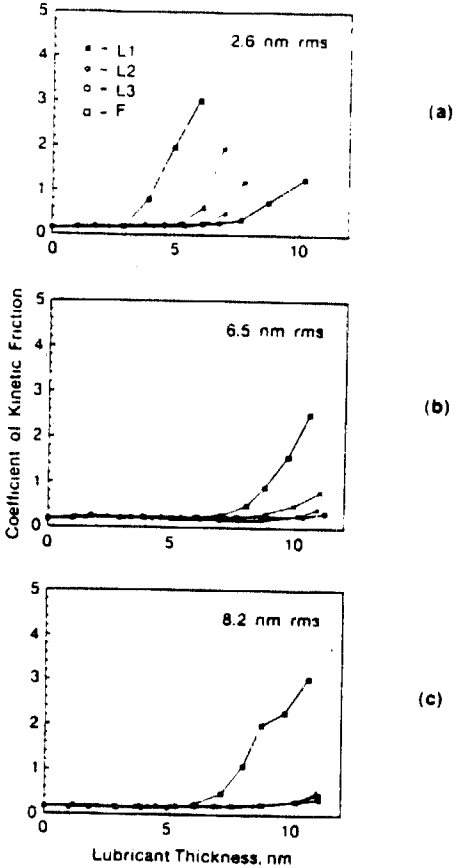


Figure 27. Coefficient of kinetic friction for thin-film disks with carbon overcoat sliding against Al_2O_3-TiC slider at 0.12 m/s as a function of lubricant thickness for the four lubricants, (a) disk X_1 roughness = 2.6 nm rms, (b) disk X_2 roughness = 4.5 nm rms, and (c) disk B_1 roughness = 5.2 nm rms (Bhushan, 1990). Lubricant L1-Z03 (viscosity = 30 cSt, molecular weight = 4250, density = 1.824 g/cc), lubricant L2-Z25 (viscosity = 255 cSt, molecular weight = 14550, density = 1.851 g/cc), lubricant L3-Krytox 143AD (viscosity = 1600 cSt, molecular weight = 8250, density = 1.91 g/cc), and lubricant F-Z-Dol (viscosity = 81 cSt, molecular weight = 2000, density = 1.81 g/cc) (Streator et al., 1991a).

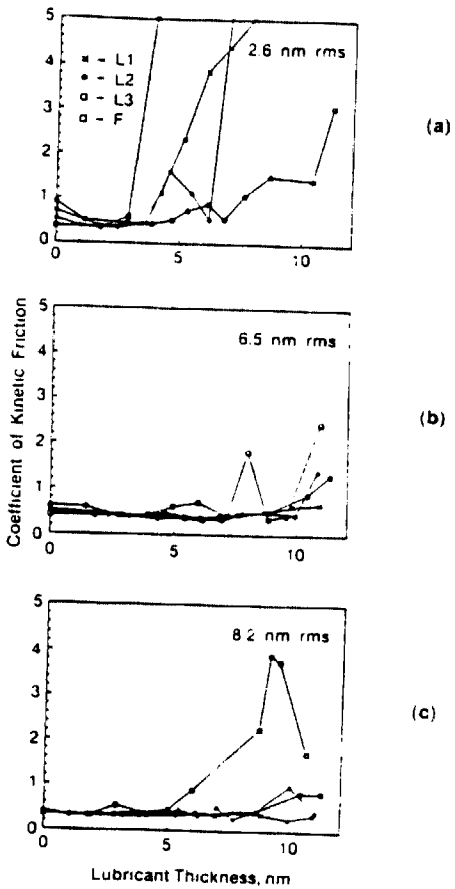


Figure 28. Coefficient of static friction as a function of lubricant thickness for the four lubricants, (a) disk X_1 roughness = 2.6 nm rms, (b) disk X_2 roughness = 4.5 nm rms, and (c) disk B_1 roughness = 5.2 nm rms (Streator et al., 1991a).

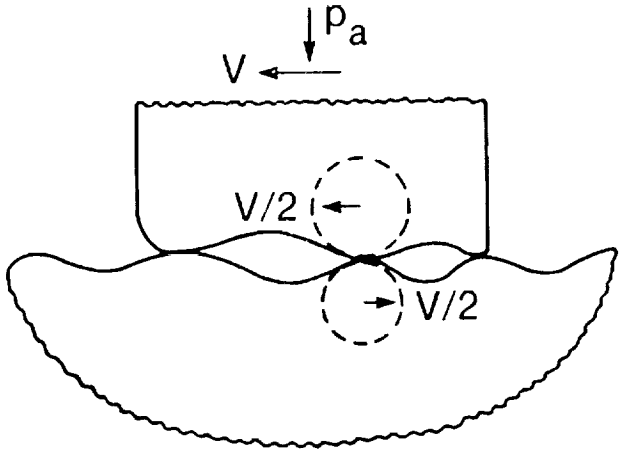


Figure 29a. Schematic of two surfaces in contact during sliding at a relative sliding speed V and a mean normal stress p_a (Bhushan, 1987a).

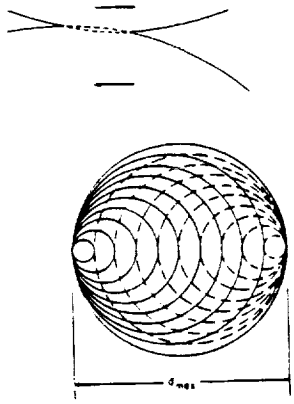


Figure 29b. The circular asperity contact grows from zero to d_{max} and then shrinks to zero. Dotted circles show the shrinking process (Bhushan, 1987a).

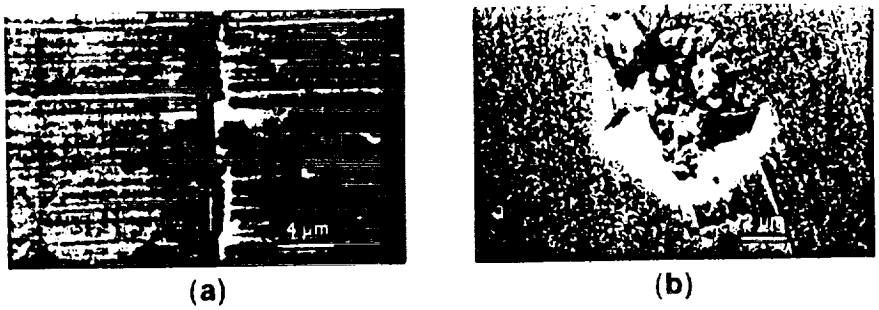


Figure 30. SEM micrographs of worn Ni-Zn ferrite head with a CrO_2 tape A (a) abrasion marks (direction of tape motion-left to right), (b) surface pull out (Bhushan, 1985b).

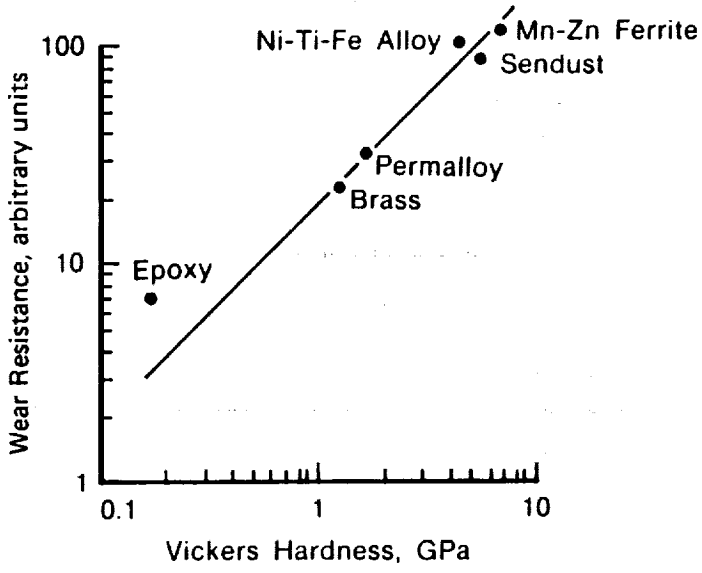


Figure 31. Wear rate of magnetic materials slid against a diamond cone as a function of Vickers hardness (Tanaka and Miyazaki, 1981).

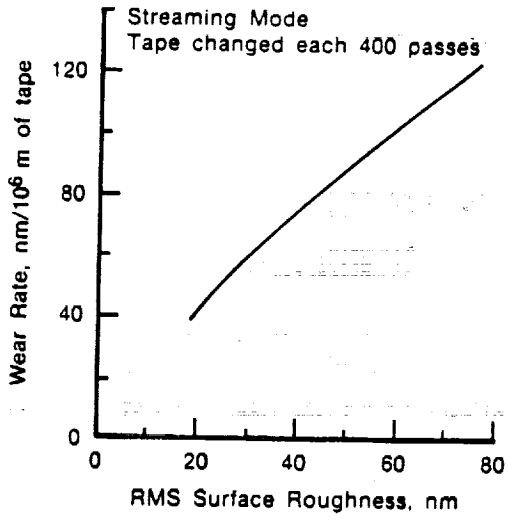


Figure 32. Ni-Zn ferrite head wear as a function of rms surface roughness of a CrO_2 tape A in streaming mode (Bhushan, 1985b).

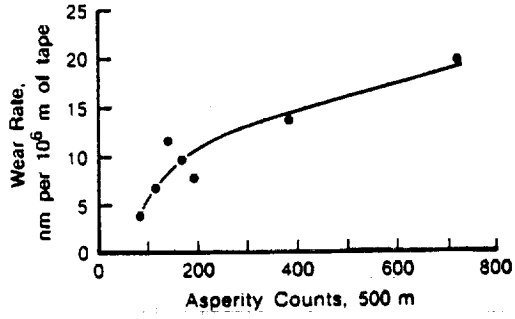


Figure 33. Ni-Zn ferrite head wear rate as a function of asperity counts on CrO_2 tape A in streaming mode (Hahn, 1984).

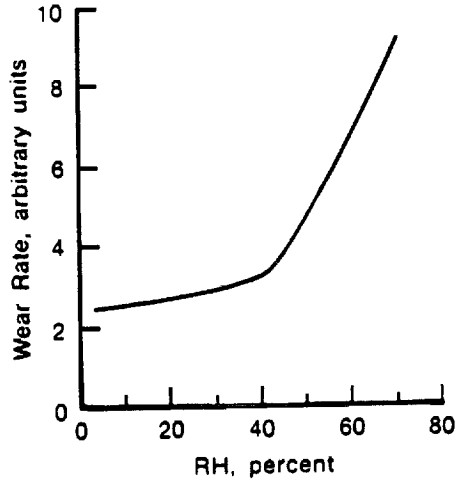


Figure 34. Head wear rate with γ-Fe_2O_3 tape as a function of relative humidity (Kelly, 1982).

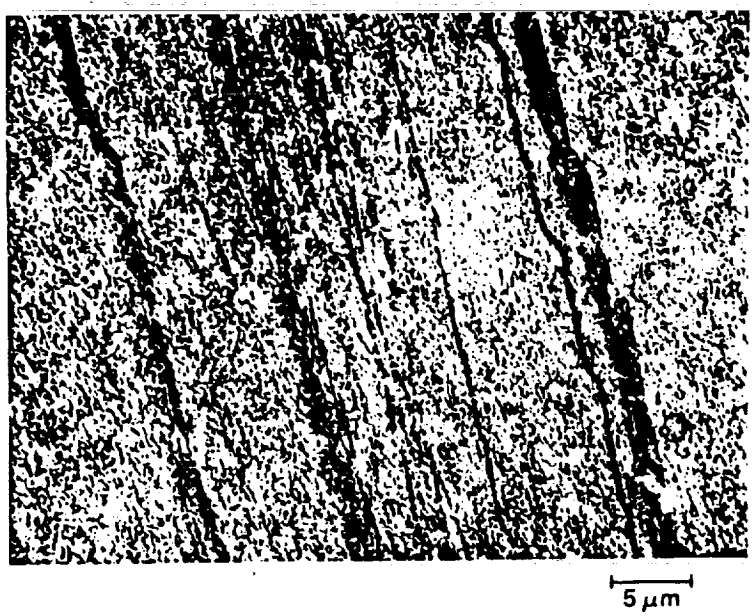


Figure 35. SEM micrograph of a worn particulate disk surface against Al_2O_3-TiC slider after extended use in CSS (Bhushan, 1990).

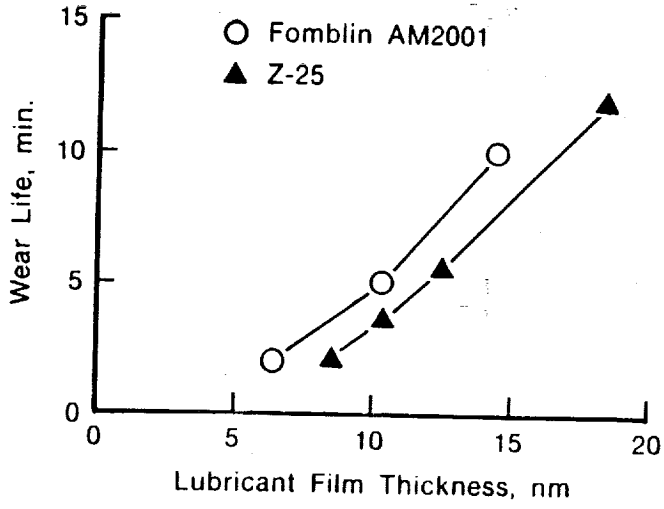


Figure 36. Wear life as a function of lubricant film thickness on a particulate rigid disk slid against Mn-Zn ferrite slider (Scarati and Caporiccio, 1987).

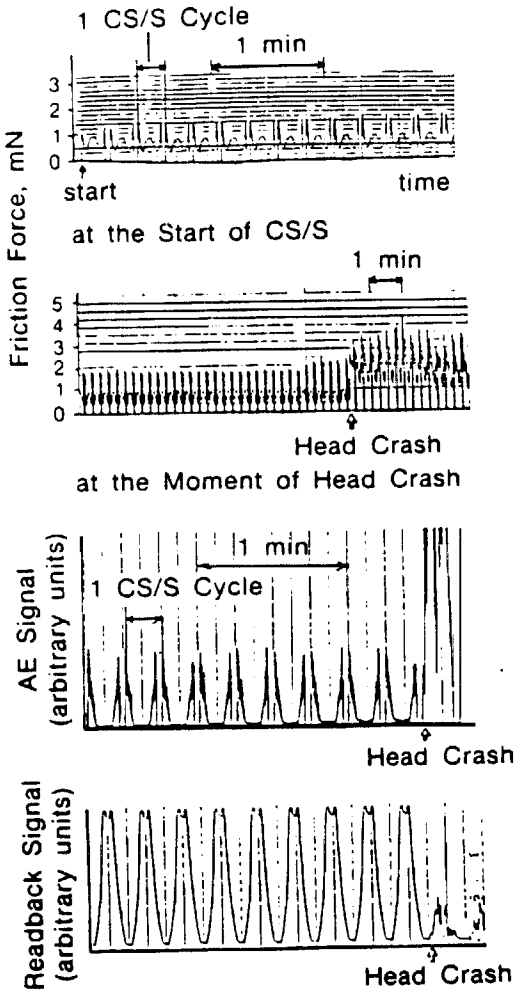


Figure 37. Friction force, AE signal and read back signal near the head crash in CSS test of a particulate disk (Kawakubo et al., 1984).

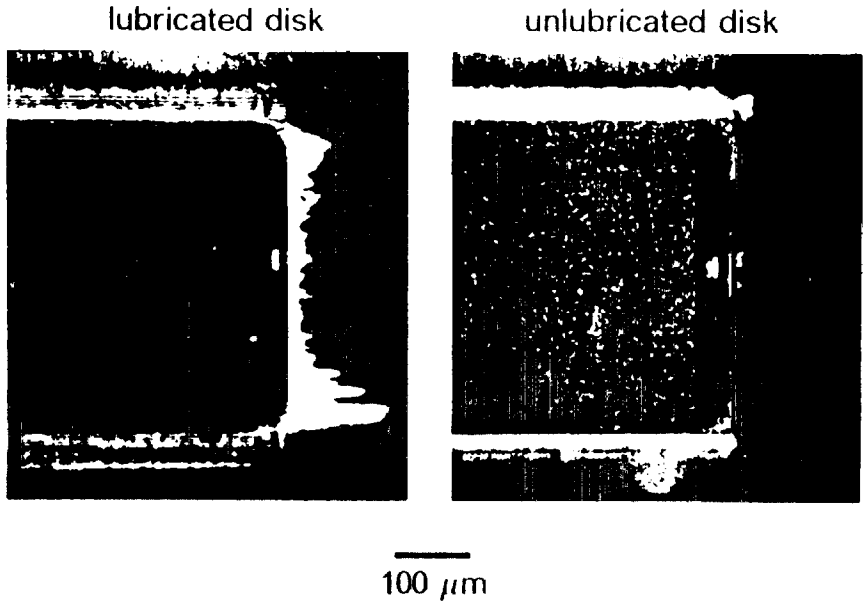


Figure 38. Trailing end of one rail of the slider: comparison between lubricated and unlubricated disk in a contamination test. Whiskers are only grown on sliders flown on disks with liquid lubricant (Hiller and Singh, 1991).

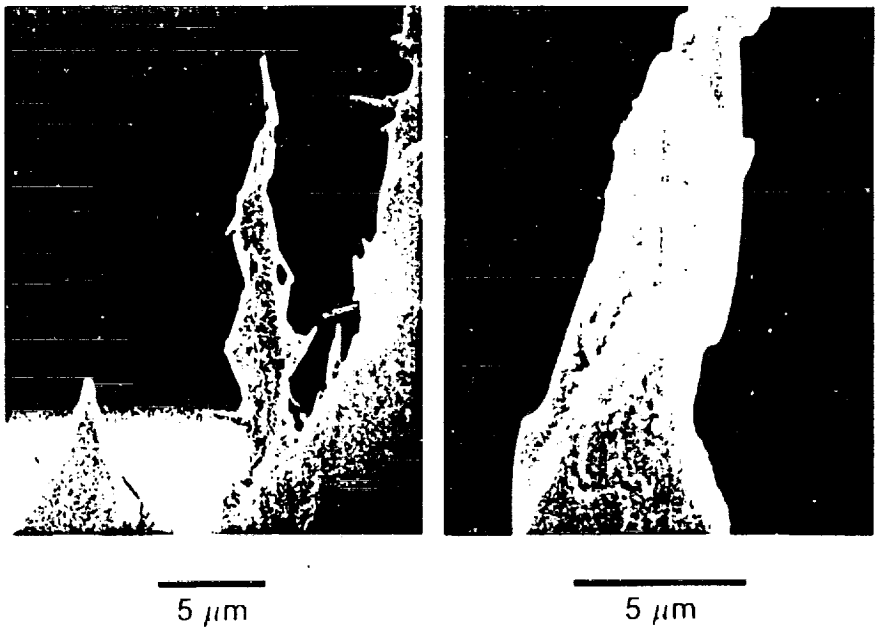


Figure 39. All polystyrene particles in the whiskers are deformed, evidence of interaction with the interface (Hiller and Singh, 1991).

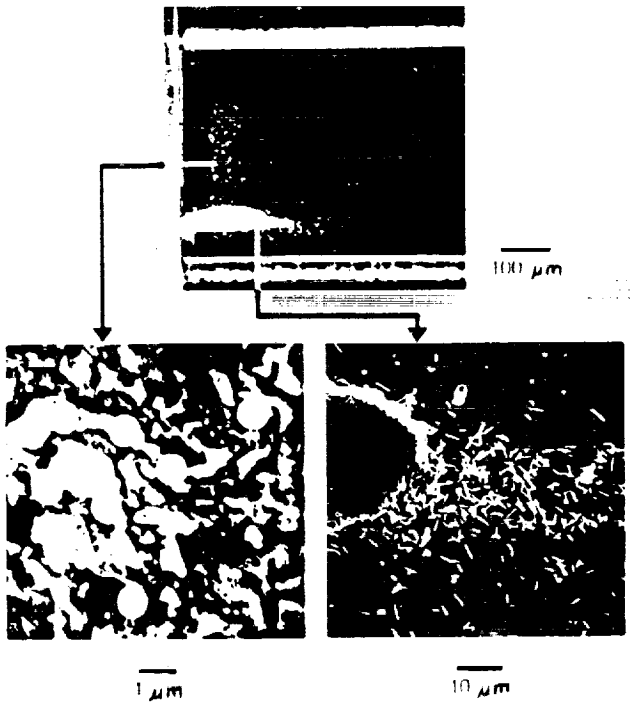


Figure 40. Dark-field photograph of the taper showing a well-defined uniform particle deposition pattern and a large agglomerate (top); SEM photographs of spherical particles within the uniform pattern (lower left) and of deformed particles in the agglomerate which is a former whisker from the trailing end (lower right) (Hiller and Singh, 1991).

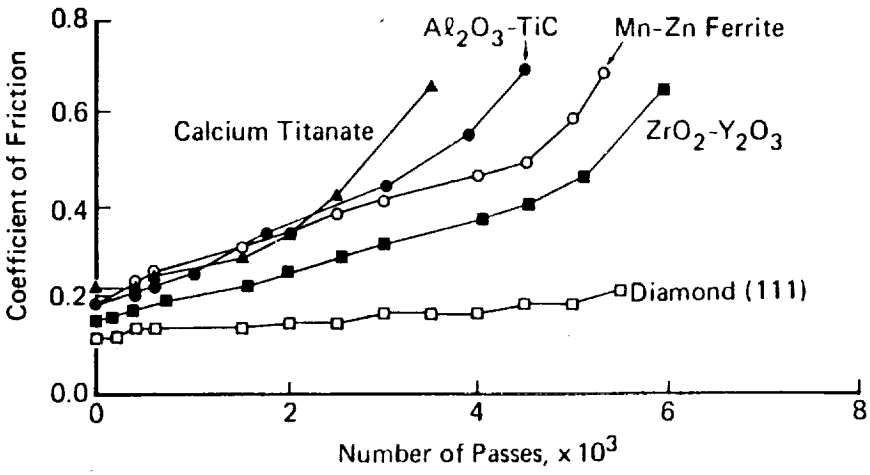


Figure 41. Change of the coefficient of friction of five ceramic sliders while sliding against a thin-disk disk with a carbon overcoat and perfluoropolyether as the topical lubricant (disk B_1) with number of sliding passes. Normal load = 150 mN, sliding velocity = 2.1 m/s (Chandrasekar and Bhushan, 1991).

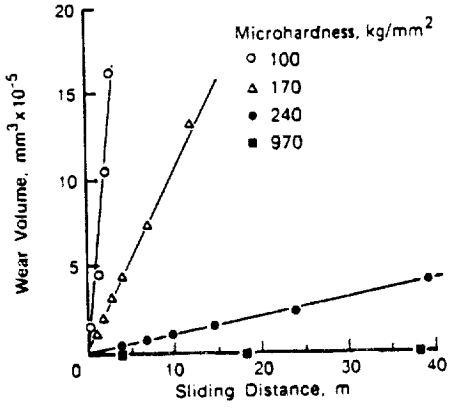


Figure 42. Relation between sliding distance and wear volume for SiO_2 films of different Vickers hardnesses, slid against Al_2O_3-TiC slider (Yanagisawa, 1985b).

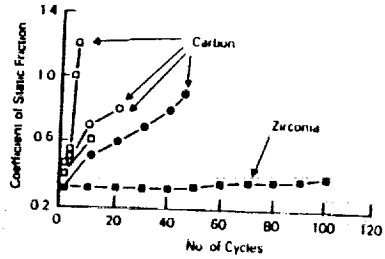


Figure 43. Coefficient of static friction as a function of number of CSS cycles for unlubricated carbon and ZrO_2–Y_2O_3 overcoated thin-film (metal) disks (disks B_1 and B_2 with no lubricant) slid against Al_2O_3-TiC slider (Yamashita et al., 1988).

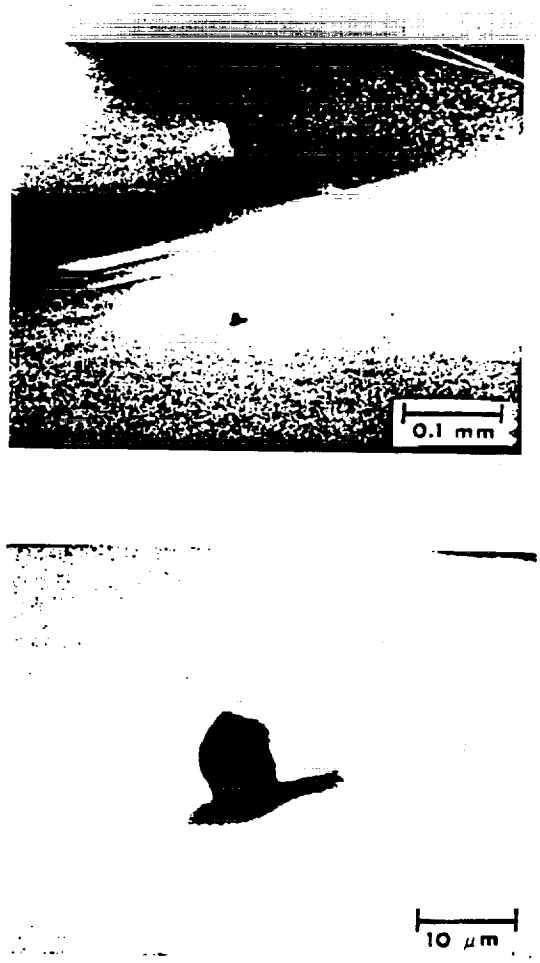


Figure 44. Appearance of the trailing edge of the Al_2O_3-TiC slider while sliding against a thin-film disk with a zirconia overcoat (disk B_2). Right micrograph shows a particle which is removed from the head and is sitting on the disk surface. This photograph was taken after 7200 mm of sliding at 50 mm/min. (Calabrese and Bhushan, 1990).

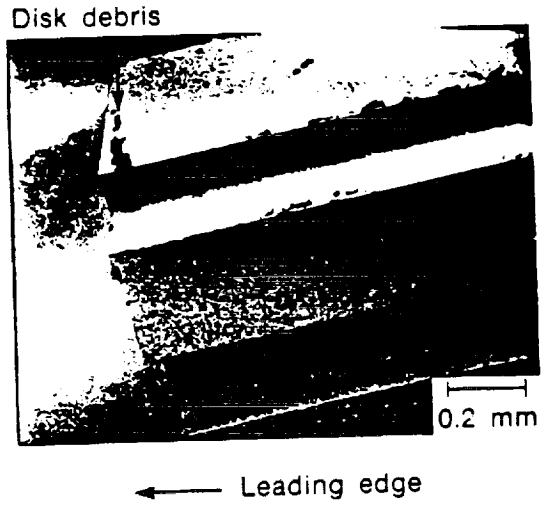


Figure 45. Appearance of leading edge of the Al_2O_3-TiC slider after sliding against the disk B_2. Wear debris is attached to the rail sides and the sides of the slider (Calabrese and Bhushan, 1990).

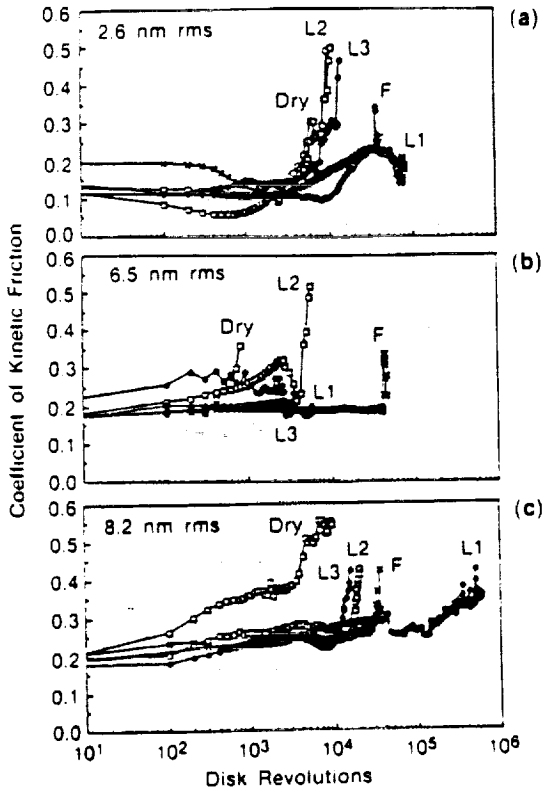


Figure 46. Friction histories during durability tests for unlubricated and lubricated thin-film disks with carbon overcoat sliding against Al_2O_3-TiC slider at 1.2 m/s with 3 nm of lubricant, (a) disk X_1 roughness = 2.6 nm rms, (b) disk X_2 roughness = 4.5 nm rms, and (c) disk B_1 roughness = 5.2 nm rms (Streator et al., 1991b).

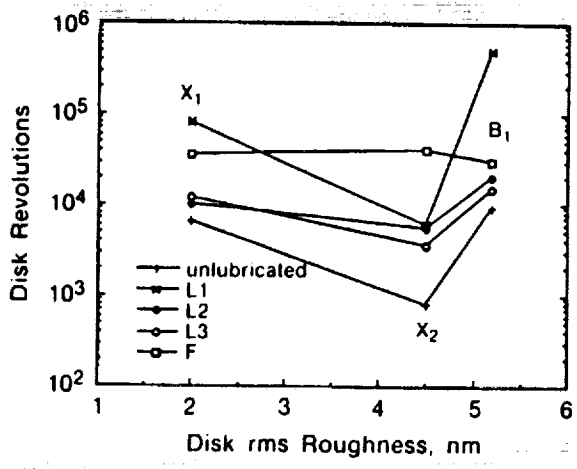


Figure 47. Disk durability summarized for each of the sliding conditions of Fig. 46 (Streator et al., 1991b).

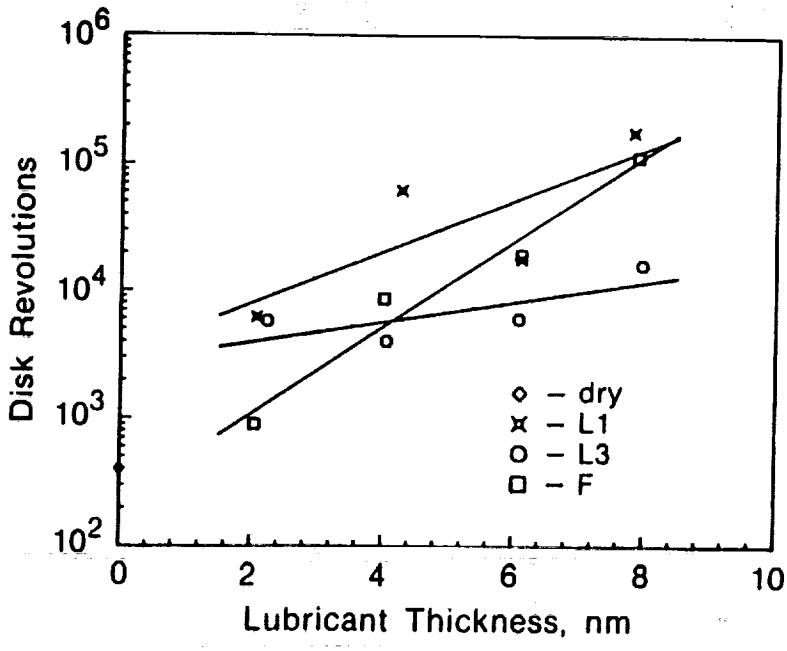


Figure 48. Disk durability as a function of lubricant film thickness on disk X_2 (roughness = 4.5 nm rms) for selected lubricants. Data point at 0 nm is the durability for dry sliding (Streator et al., 1991b).

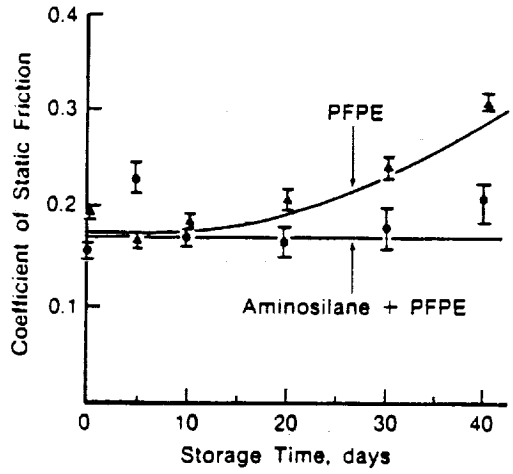


Figure 49. Coefficient of static friction as a function of exposure at 50°C and 10^{-3} torr for lubricated spin-coated SiO_2 film on a sputtered oxide magnetic film (Hoshino et al., 1988).

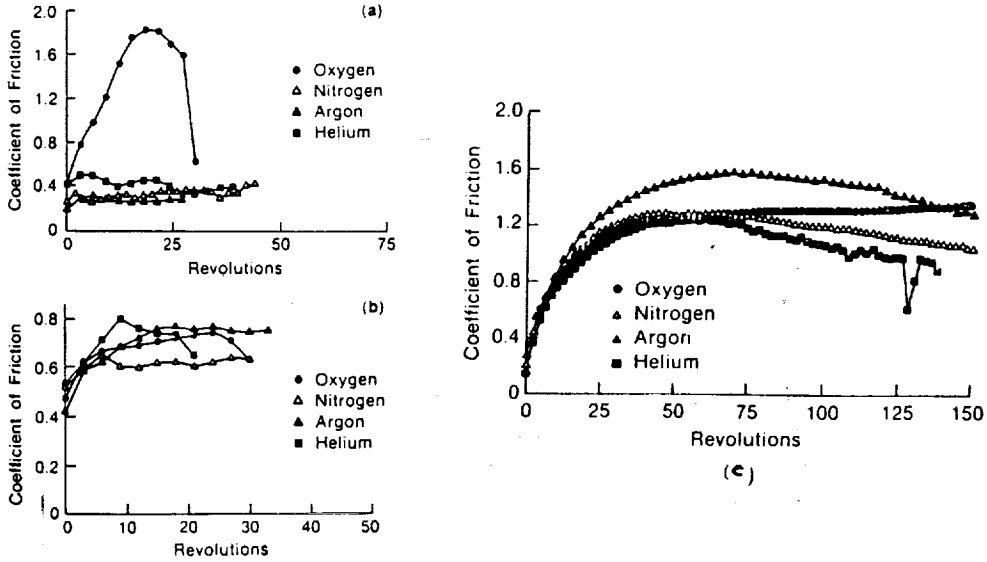


Figure 50. Coefficient of friction as a function of number of revolutions during sliding of Al_2O_3-TiC slider against an unlubricated thin-film disk at a normal load of 150 mN and a sliding speed of 60 mm/s (a) disk with carbon overcoat (B_1 with no lubricant) in dry gases (b) disk with zirconia overcoat (B_2 with no lubricant) in dry gases, and (c) disk with carbon overcoat (B_1 with no lubricant) in various gases, all at 4% RH (Strom et al., 1991).

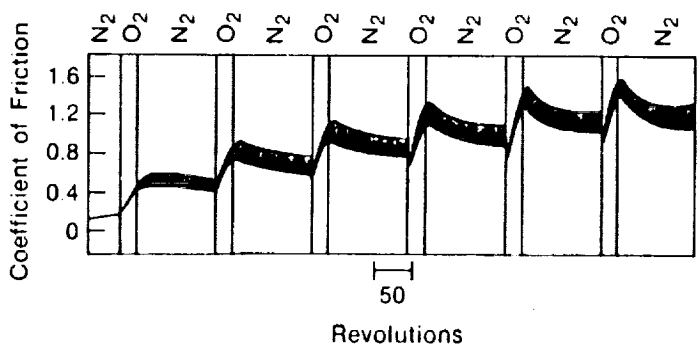


Figure 51. Coefficient of friction as a function of number of revolutions during sliding of Mn-Zn ferrite slider against an unlubricated thin-film disk at a normal load of 100 mN and a sliding speed of 60 mm/s in various dry gases. (Marchon et al., 1990).

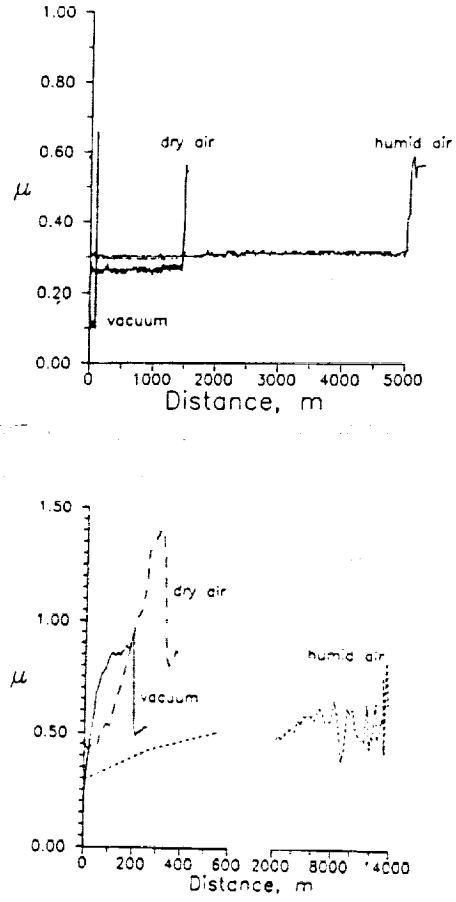


Figure 52. Coefficient of friction (μ) as a function of distance slid for hemispherical pins of Mn-Zn ferrite on lubricated thin-film disks in vacuum, dry air and air with 50% RH, at 1 m/s sliding speed and 98 mN applied load (a) lubricated thin-film disk with carbon overcoat (disk B_1), (b) lubricated thin-film disk with zirconia overcoat (disk B_2) (Dugger et al., 1992).

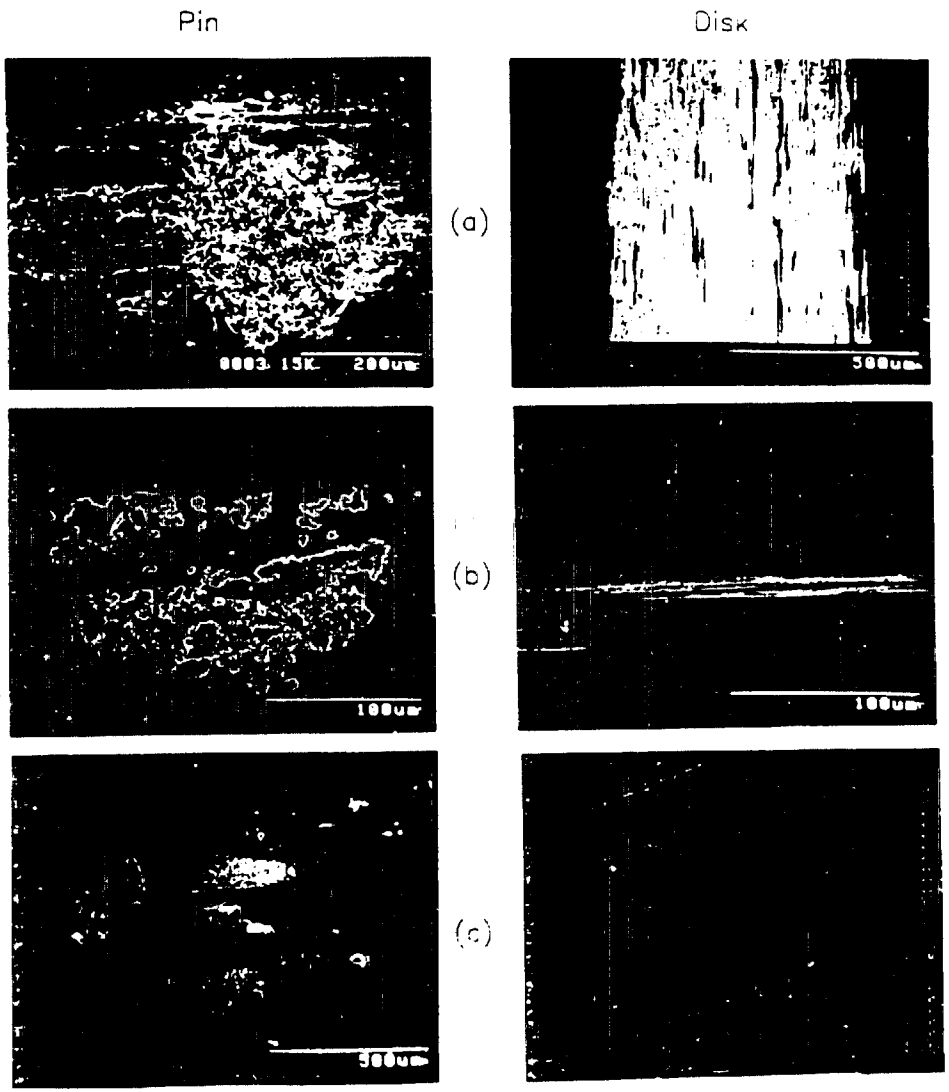


Figure 53. SEM micrographs of Mn-Zn ferrite pin and lubricated and carbon-coated thin-film disk surfaces run to failure at 98 mN normal load and 1 m/s in (a) vacuum, (b) dry air, and (c) humid air (Dugger et al., 1990).

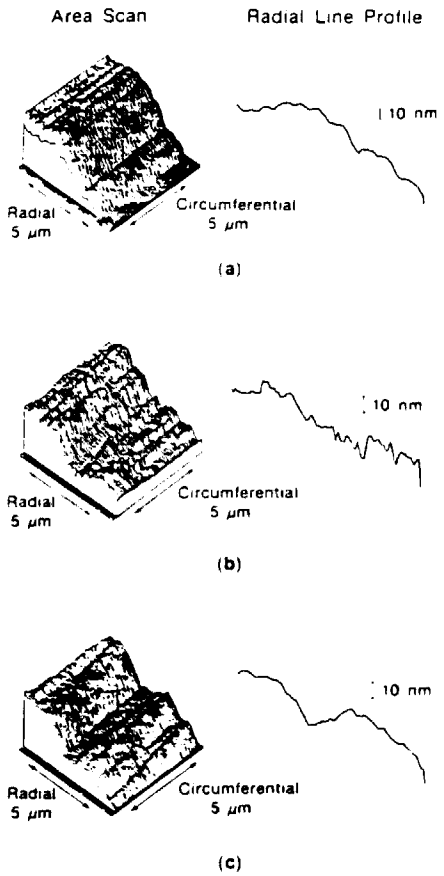


Figure 54. Two-dimensional surface profiles using the scanning tunneling microscope on (a) the untested disk surface; (b) on wear tracks after testing at 98 mN applied normal laod, 1 m/s sliding speed in vacuum and (c) in air at 50 percent relative humidity. The line scans indicate a general roughening of the surface after testing in vacuum and smoothing after testing in humid air (Dugger et al., 1990).

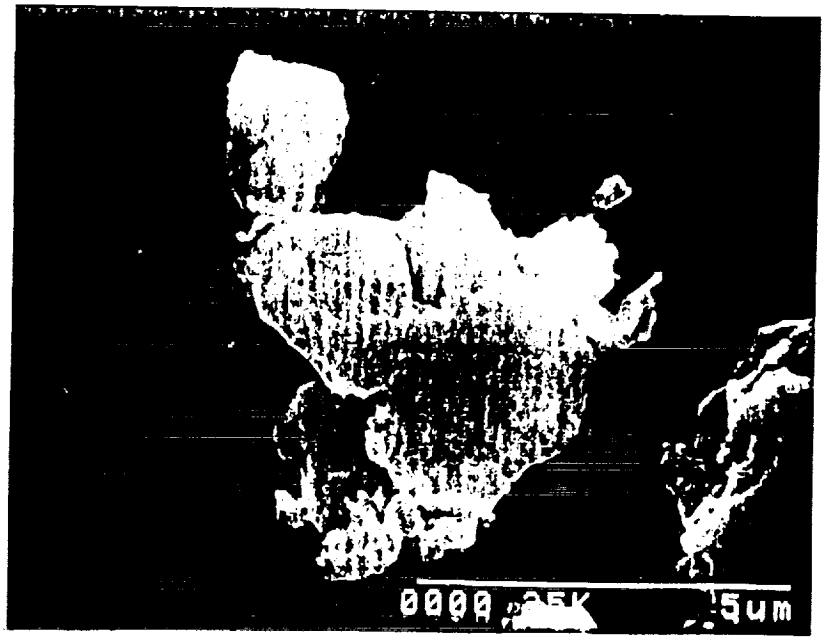


Figure 55. SEM micrograph of an isolated metallic wear particle on the pin from a test at 98 mN applied normal load, 1 m/s sliding speed in vacuum, showing evidence of agglomeration (Dugger et al., 1990).

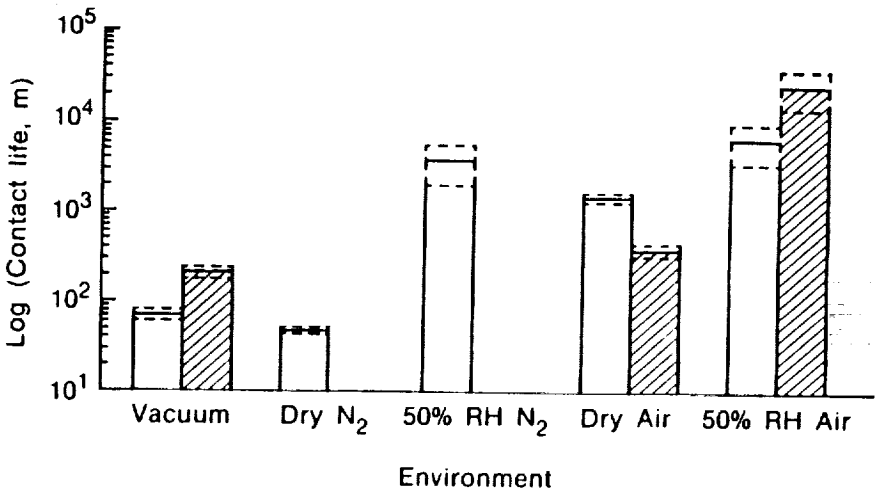


Figure 56. Summary of the contact lives as a function of environment for lubricated thin-film disks with carbon and zirconia overcoats (disks B_1 and B_2, respectively) at 1 m/s sliding and 98 mN load in a sliding test with hemispherical pins of Mn-Zn ferrite. Error bars represent the standard deviation from at least four experiments.

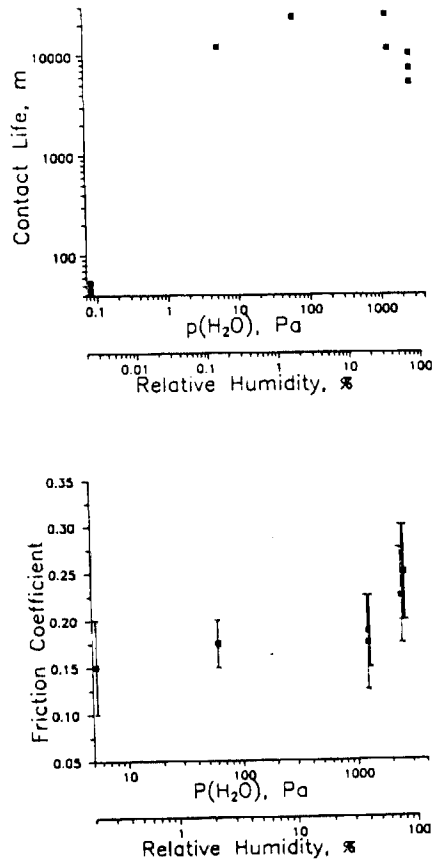


Figure 57. Contact life and coefficient of friction as a function of partial pressures of water vapor in a nitrogen ambient for lubricated and carbon-coated thin-film disk (disk B_1) at 1 m/s sliding speed and 98 mN load in a sliding test with hemisperical pins of Mn-Zn ferrite (a) durability, (b) coefficient of friction (Wahl et al., 1991a).

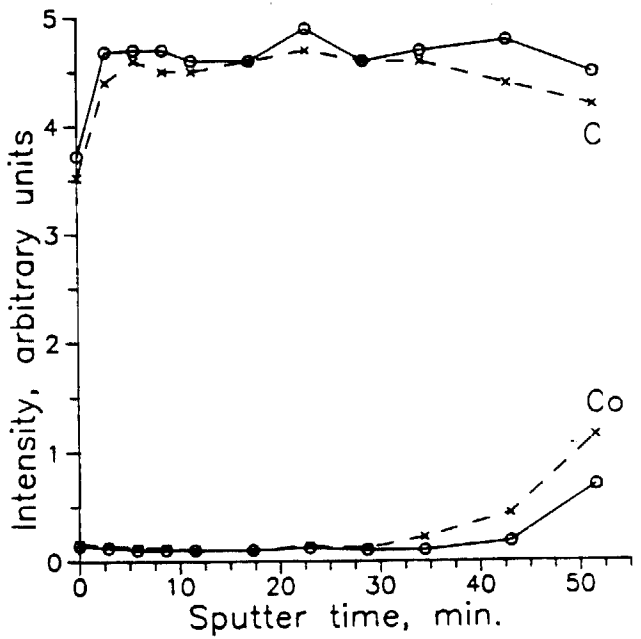


Figure 58. Auger depth profiles for carbon and cobalt from adjacent regions off (solid lines) and on (dashed lines) the wear track on a lubricated and carbon-coated thin-film disk at 90% of the anticipated contact life in humid air at 1 m/s sliding speed and 98 mN normal load in a sliding test with hemispherical pins of Mn-Zn ferrite (Dugger et al., 1992).

Network Issues for Large Mass Storage Requirements

James "Newt" Perdue

Ultra Network Technologies, Inc.
San Jose, California USA

CLIENT/SERVER NETWORK & STORAGE NEEDS

The major performance demand on today's networks in the Science and Engineering environment by far derives from mass storage requirements. The need to move massive amounts of data between the different parts of the computing environment dictate the topology and performance requirements of the local area network. This paper will explore such requirements and provide some insights into solutions which address the increased need for network performance as a result of the explosive growth of data in the science and technology area.

Data plays a key role in determining the architecture and performance needs of a computing environment. Basically, mass storage is the repository for the data and information which drive the entire computing scenario. In fact, we can think of mass storage as holding the major assets of any institution or corporation. Here is stored the "kings jewels" of the organization. In today's society information is the king and rapid access to it is the king's road. If we look at Data as the center of any organization (See fig 1), Compute Servers and Clients surround it with paths for fast access between Servers, Clients and data.

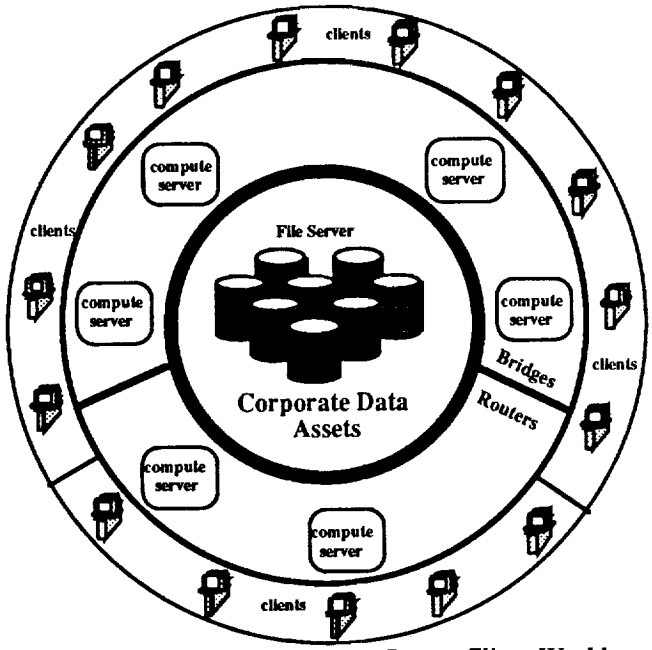


Figure 1 - File Server In The Server-Client World

These servers are usually linked to one or more data or File Servers (containing the corporate data) by a local area network with large throughput and bandwidth capabilities. Outside of a ring of such Compute Servers, lies the users of the information, the clients. They also require fast access to the data but usually in lesser amounts than the Compute Servers since they are looking at the pieces in small amounts, analyzing it in some greater detail. The clients in todays scenario are usually a variety of workstations and personal computers linked through a local area network with low to medium bandwidth and throughput capabilities. This describes the Client-Server scenario today: fast compute servers, such as supercomputers, near-supers, and high end workstations, sharing data between themselves and file servers, all working to create and manipulate the data into a form accessible by many clients, usually workstations and personal computers.

The central file server in this scenario provides for the:

- storage of corporate data in a single secure location;
- rapid access and movement of data between compute servers;
- creation of a hierarchy of storage devices for economic handling of the data;
- remote data access to workstation/pc clients in both file and record oriented (diskless operation) modes;
- temporary storage of massive data sets for distributed processing needs or common access requirements;
- storage and retrieval of large graphic images for later playback (digital VCR);
- cacheing data between Compute Servers and networks of differing speeds;
- long term reliability of the storage by implementation of archiving methods.

These are heavy demands in the supercomputing environment due to the very rapid growth of the amount of data required to feed the ever faster Compute Servers and the need to save data for both development and liability needs. In most supercomputing environments today, it is not uncommon to see existing on-line storage requirements greater than 250 Gigabytes of storage, and on-line tape storage (silo's) in the range of several terabytes of data. It's also clear that most users consider these capacities to be inadequate for the near future.

These demands create several issues for access to the storage and thus the local area networks. Because the individual file sizes can reach several gigabytes in size, many local area networks can't handle them. Often the mean time to failure of a network can be less than the time to transfer such a file making file transfer via a network not feasible. Such a single file can take over an hour to transfer between servers or servers and clients. Multiple files being transfered at once can literally stop

an ethernet from functioning due to the congestion of such transfers.

The usefulness of a local area network must be measured in terms of it's ability to provide EFFECTIVE performance for such files, not by the rated performance or speed of the bits on some part of the wire connecting hosts. Further, the need to transfer such large files places large burdens on the processing capabilities of the servers and clients involved in the transfer. On typical hosts today, protocol processing consumes between 40 to 100% of the CPU power to maintain an average of 8 megabits per second transfer capability (ref: SHIFT, CERN 2 Mar 91). Further, as the speed of the network "wires" increase the demand on the CPU increases if it is to maintain efficiency. Further, most host system structures are undersized for the efficient movement of large data files:

- buffers in both system and applications are small, causing many interupts to host I/O systems,
- disk "effective" transfer rates are not sized to the mammoth file sizes;
- copying between various system buffers creates a large amount of overhead which affects the transfer rate and the cpu utilization of the host systems;
- the "wire speeds" of the ethernet, and even FDDI systems don't measure up to the needs to move gigabytes of data.

Clearly, for the science and technology marketplace, if the file systems and networks are to maintain the pace set by the CPU industry, major improvements must occur in the integration of technologies.

Fortunately, the technologies needed to address these requirements are moving forward at an acceptable rate. But it is not enough for technology to be available, for there must also be the ability to integrate the elements of technology into products which address the specific needs. In figure 2, major technology developments affecting the computing and storage needs are listed in a relative timeline. From this diagram one can see trends that may come to our rescue. Developments such as International Standards for the definition of File Server interfaces such as the one in development by the IEEE Mass Storage Committee, standards for high-speed data connectivity such as HIPPI (ANSI X3T9.3), development of high-speed protocol processors such as UltraNet, and development of new disk architectures such as the RAID devices all contribute technology to the solution of the next generation high performance File Servers. Already such technologies are being combined to give us a taste of what's to come: DISCOS is supporting the IBM RAID product in a distributed environment with UniTree™, Cray Research has a Data Migration Facility which uses a dedicated Cray using HIPPI for I/O and striped disks for increased throughput, and NASA Ames has developed their own file system software using parallel channel connections (8) for both disk and network I/O from Amdahl systems to be able to achieve transfer rates to the Cray up to 20 MBytes/second.

The way we treat data in the supercomputing environment has certainly evolved over the last 20 years. As shown in Figure 3, file systems originally were thought as part of the host which they served. Each host owned the data it produced and networks permitted sharing by moving entire files across the network when required. Such sharing was not so important in this scenario due to the slow transfer rates possible. Next as network speeds increased, a concept of a centralized file store was introduced and is widely accepted today as the architecture most applicable to the networked environment. This permits access both at record level and file level to any host on the network.

Data produced by a host may still reside on that host, but the opportunity to move it to a central system for later retrieval by other systems or for archive of the data is now possible. This reduces the amount of disk space required on each host and has the advantage of permitting economies of scale to apply for storage purchased for the File Server. Of course, the File Server remains a point of failure for the entire system, and generally causes a large performance bottleneck for access to the data.

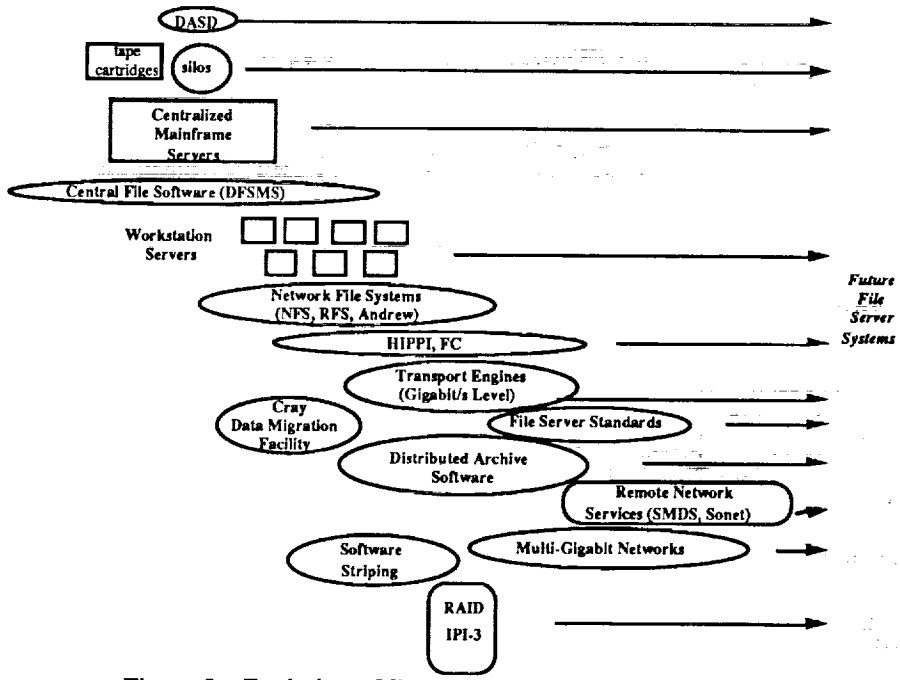


Figure 2 - Evolution of File Server Technologies

However, the most likely future scenario for File Servers is seen as the last diagram in figure 3. In the Distributed File Server scenario, data may or may not be associated directly with a specific host or even a specific File Server CPU. Software maintains knowledge of the location of the files throughout the network and manages its migration over the network from source to requestor as a third party manager rather than directly manipulating the storage itself. This scenario presents the possibility of connecting storage directly to the network itself, without a host to manage it due to the evolution of two technologies: intelligent controllers and standardized high-level command languages such as Intelligent Peripheral Interface Level 3 (IPI-3). The intelligent controllers can play the role of the low-level disk driver and the network interface. Further, with the availability of network interfaces with the effective performance of high-speed I/O channels AND the ability to run network protocols at the TRANSPORT level (such as the UltraNet Network Processors), these stand-alone disk servers are made even more practical. This scenario may provide the supercomputer owners freedom of choice in INDEPENDENTLY selecting the peripherals, the CPU's, the file software and the network. Each element can evolve it's own competitive marketplace which is sure to drive the prices for mass storage, computer systems and networks to more advantageous levels for the users.

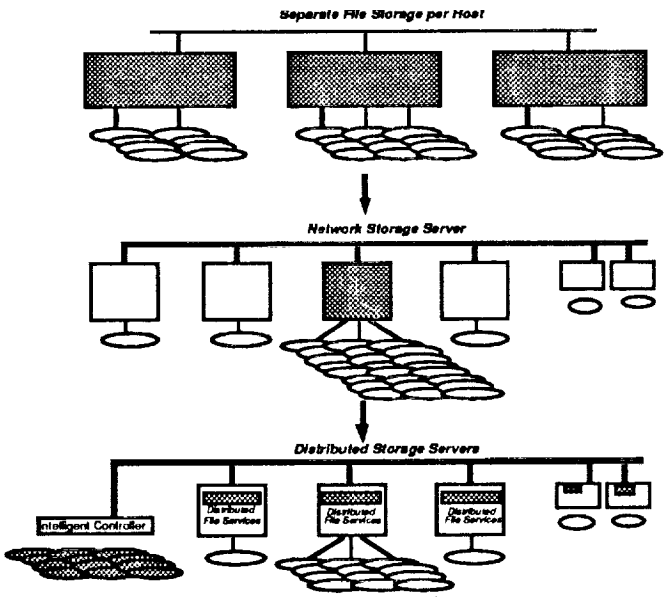


Figure 3 - Trends in File Server Architecture

Of course, performance is still a major consideration when designing file systems for the supercomputer environment. Figure 4 diagrams four different approaches to increasing performance to disk systems. Today, several vendors (including Cray Research) have implemented software to stripe the data from several disk drives in parallel to achieve raw disk transfer rates in the range of 32 to 120 MBytes per second.

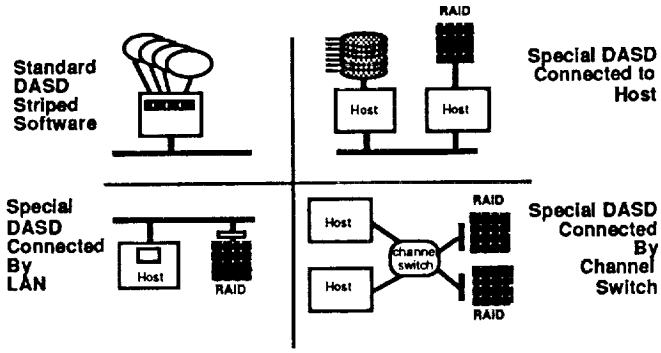


Figure 4 - Methods To Achieve High Performance Disk Throughput

Generally the drawback of this approach has been the problem of "all the eggs in one basket". If a disk fails, it is possible to lose the entire (VERY LARGE) file store. Another approach includes the use of parallelism in the disk devices themselves. Parallel heads on a single platter can increase transfer rates today up to about 20 MBytes per second, and a new area of development called RAID (Redundant Arrays of Inexpensive Disks) parallelize the data streams from a number of inexpensive disks. The advantage of this approach includes the ability to offer redundant paths to protect against most loss of data. For network access, several possibilities arise. The CPU can manage the RAID or parallel head disk devices directly and pass the data across the network. The main problem with this approach is that most networks today cannot maintain the transfer rates required. Another approach now possible with the introduction of IPI-3 and HIPPI interfaces is the connection of the RAID devices directly to the network. Although this has not been done in any operational implementation yet, it is possible and developments are in progress. Finally, with the introduction of HIPPI channel switches, these devices can be connected between hosts (which have HIPPI channels) much in the same way that multi-channel controllers permit access by more than one host.

The most promising approach for increased performance with economic rewards may prove to be the distributed file server concept with network attached disk devices and peripherals. A major advantage of this approach is the ability to move data directly from the disk device to the requestor without going through a host mainframe, which only adds to the performance overhead and the cost of the system. The data management software, if centralized can reside on a much smaller host, such as a workstation, with dramatic savings possible in both initial capitalization, maintenance costs and in-house system personnel costs. In this scenario, the network must be fast enough to maintain effective rates higher than a single host to a variety of hosts with a variety of connection capabilities (BMC, HSX, HIPPI, LSC, VME, Microchannel, etc.). A standard disk I/O command language, such as IPI-3 makes it possible to ask for block data and the intelligent controllers execute the low-level

commands required for disk I/O. Large blocks then get sent to the destination directly over the network.

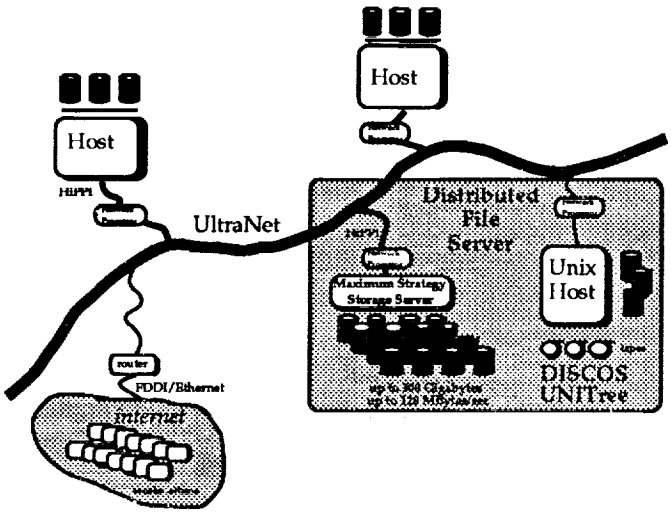


Figure 5 - Concept for Distributed File Server

Figure 5 diagrams a concept for a Distributed File Server using network attached peripherals. In this scenario, a RAID disk with HIPPI interfaces permit transfer to hosts at transfer rates between 20 to 40 MBytes/second. In this specific scenario, UltraNet is proposed for it's high performance and it's ability to do the network protocol processing, Maximum Strategies HIPPI RAID devices for maximum transfer performance and redundancy (using HIPPI channels and IPI-3 command languages), and DISCOS UniTree for it's distributed hierarchical storage management software. Although this combination must still be proven, it is an example of a system that could be constructed to provide a completely distributed file server environment with very high performance and a variety of connectivity (must greater than allowed by HIPPI only devices). Figure 6 (at the end of this paper) presents details of this implementation.

SERVER NETWORKS

From a network perspective two major requirements exist for performance-based file access. First, Server-to-Server traffic must be managed. Data from a single users job may exist on several servers, and may take several days to accumulate. Data is transfered between the servers to accomplish the integration of the task. Gigabytes of data flow between File Server and Computer Server each day. Server-to-Server traffic accounts for a large amount of the traffic in a local area network.

Second, once the data is computed, client systems such as workstations need access for data analysis, visualization and presentation. Usually transaction-oriented access, such as that provided by NFS dominates Client-Server communications. The large databases are generally not transfered to the client, but only accessed in pieces as needed. Further, this Server-Client access serves as the path for software development, not requiring major amounts of traffic but frequent access.

Therefore, in view of the Client-Server model discussed earlier, VOLUME data is required between Servers but TRANSACTION oriented traffic dominates between Client and Servers. An ideal network model for this would segment the network in such a way that permits use of multiple network technologies. For the Server-to-Server traffic utilize the highest speed network technology and for the Server to Client (and Client-Client) utilize the technology with the lowest cost, highest connectivity and most standardization.

Figure 7 contrasts two approaches using technology available today. One connects both Servers and Clients using a single network.

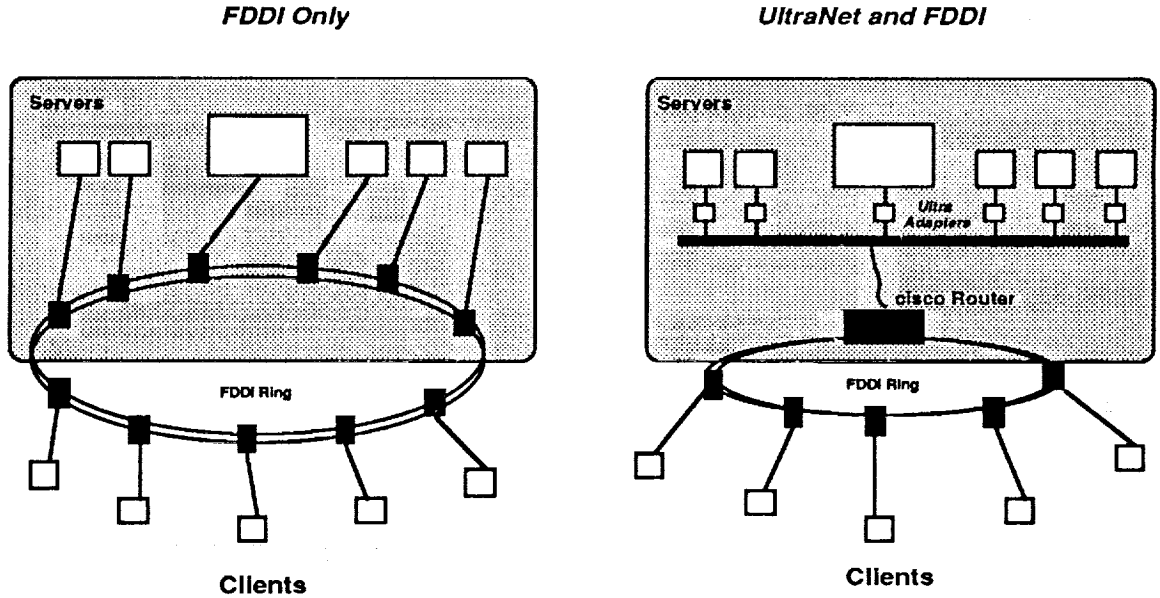


Figure 7 - Alternatives for Networking Servers

The other connects two networks via a high-speed router. In the later instance, one network excels in large aggregate performance (for backboning) and also in high task-to-task transfer rates for Server-to-Server communication. The other network would be more transaction oriented, standard, and have a lower cost point.

By measure of published EFFECTIVE performance results, the only network available today which can maintain the task-to-task data rates and the large aggregate rates needed in the supercomputing scenario is UltraNet. In figure 8, UltraNet is used as a FRONT-END network for Servers bridging (routing) to Clients connected by FDDI. One (or more) high-speed routers are needed to connect to FDDI. No Server needs a direct connection to FDDI or ethernet. This saves the cost of multiple network interfaces for the Servers. As a Server Network, UltraNet can sustain at least a gigabit/second aggregate transfer rate (for the Backbone function) and can sustain task-to-task transfer rates up to 50 MBytes/second. Connectivity to multiple hosts are available using interfaces such as HIPPI, HSX, LSC, BMC, VME and Microchannel.

As a way to explore the merits of using UltraNet as a Server Network instead of using only FDDI, a simple network model was built and then tested to confirm it's results. Figure 9 shows the results of the network model. For the FDDI (or ethernet) only solution (on the left), the same maximum transfer rate is achieved for either Server-Server traffic or Server-Client traffic. The maximum transfer rate from the host, in the demonstrated case is limited to an effective rate of ~1.5 MBytes/second for FDDI and is evenly shared between Server traffic and Client traffic. If 50% of the traffic is between Servers, the maximum available bandwidth is .7 MBytes/sec, and the other .7 MBytes/sec is available for Server-Client traffic. In the UltraNet scenario, the Server-Client traffic is still limited to the .7 MBytes/sec (due to bottlenecks in the Client systems), but between Servers, UltraNet can now

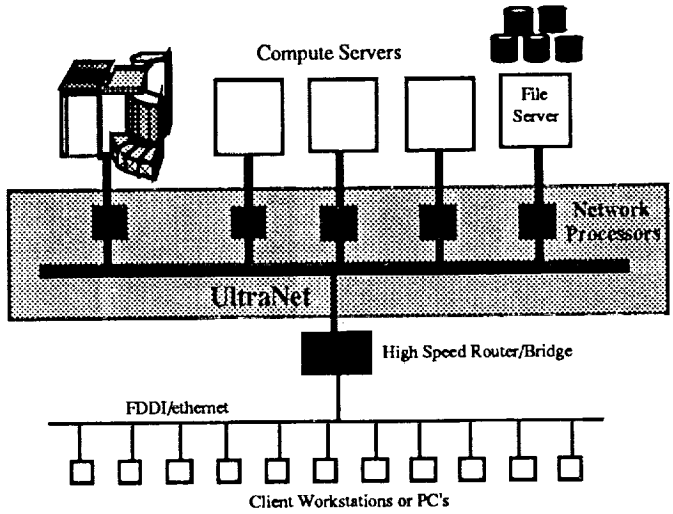


Figure 8 - UltraNet as a Server Network Solution

provide up to 4 MBytes per second while still reserving 50% of the host transaction capabilities for Server-Client traffic. When all traffic is between Servers UltraNet can provide up to 9 MBytes/second. This example features a typical low end near-supercomputer as Compute Server. For Cray based systems, the UltraNet can provide over 40 MBytes/second effective performance. This model was based on the number of transactions per second possible by the host for I/O, and the amount of data that could be transfered per transaction. In this case, I/O to Clients is limited to FDDI packet sizes of 4.5 Kbytes. For each transaction, only 4.5 Kbytes is transfered. Between Servers connected through the UltraNet, 32 Kbytes of data can be transfered, over 7 times the amount of data for a single FDDI transaction.

Figure 10 shows the results of an actual test done to demonstrate this point. Near-supercomputers (Convex C-1 and Alliant FX-80) were used as the main Compute Servers together with several workstations. Although FDDI was not actually used, it was simulated by limiting the transactions to 4.5 Kbytes. An actual test using 6 computers was run using the TSOCK test program and the

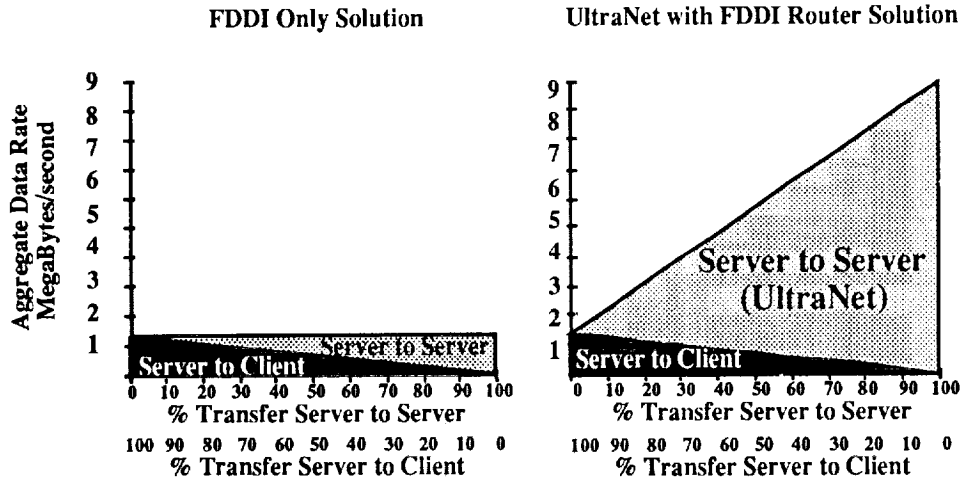


Figure 9 - Results of Modeling Server Network Scenarios - FDDI only vs UltraNet/FDDI

results show that the model in figure 9 is very closely approximated. When 50% of the traffic is to Clients (using 4.5 Kbyte transactions), about half of the FDDI sized traffic is possible (about .5 MBytes/s in this test). But at the same point, UltraNet provides over 5 MBytes per second to the other Servers.

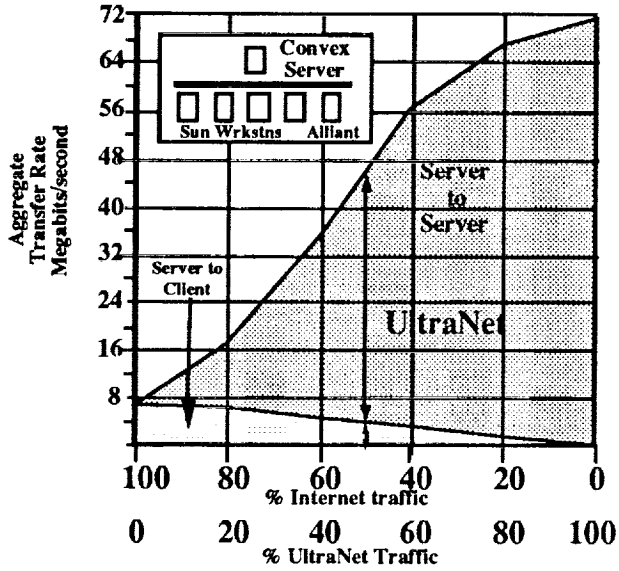


Figure 10 - UltraNet as Server Network (Actual Results)

The point of this is to demonstrate that if a network architecture can be selected which maximizes the transfer rates for Server-to-Server traffic and minimizes the cost of the Server to Client traffic then the best result is achieved for implementing high performance file systems.

Another factor which is important in evaluating networks to use for file systems is the total aggregate data rate possible. Task-to-task transfers are extremely important

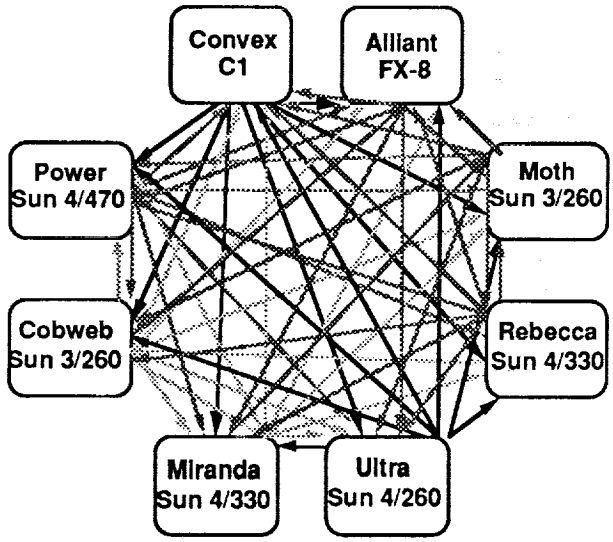


Figure 11 - Test Environment for Bandwidth Test

but more important in an busy network environment is that the aggregate transfer rate does not drop off dramatically when additional conversations are added. Figure 11 shows the basis for an experiment using UltraNet and ethernet to demonstrate this point. Eight computers, (including Convex, Alliant, and Sun Servers) were used to demonstrate this point.

Each computer transmitted and received to every other computer 114 MBytes of data. Each computer established 7 vitual circuits to each of the other computers (a total of 56 virtual circuits for the test). Each computer began the transfers within about 5 seconds of the other. A total of 3,600 MBytes of data was transfered between the computers using the TSOCK test program. Each computer had both an UltraNet connection and an ethernet connection. Two ethernet segments were utilized to increase the aggregate transfer rate on the slower network.

Figure 12 demonstrates that UltraNet delivered the entire 3.6 Gigabytes of data in less than 6 minutes. Ethernet, on two circuits complete the transfer in 26 minutes. If one segment had been used, it would have take over 52 minutes. In the UltraNet case, each computer sustained over 3 MBytes/second, generally limited by the Sun workstation transfer rates.

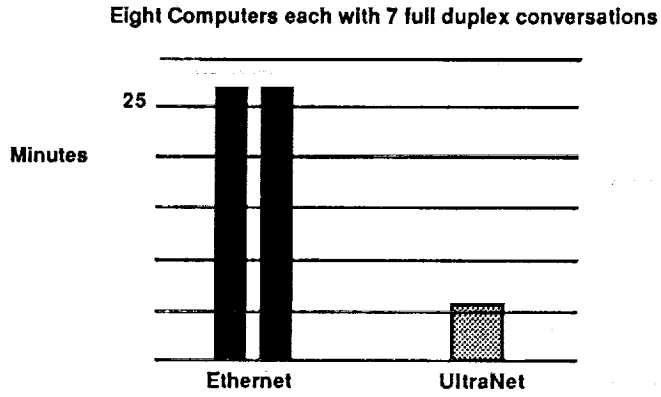


Figure 12 - Results of Aggregate Bandwidth Test

Ethernet performed well in the test from the viewpoint that each segment sustained over .83 Megabytes/second or almost 65% of the ethernet bandwidth. UltraNet sustained over 11.8 MBytes per second. However, with UltraNet, only 10% of the total bandwidth of 125 MBytes/second was utilized, leaving another 110 MBytes/second of bandwidth for additional conversations.

Although this test shows the large aggregate capability of UltraNet as a Server Network, probably more instructive is how it performs when supporting actual file applications. Disk-to-disk transfer rates are most instructive in evaluating any network to be used for a file server. Figure 13 summarizes the results of a test performed by Cray Research on the ability of FTP to transfer between two Supercomputers over UltraNet. The

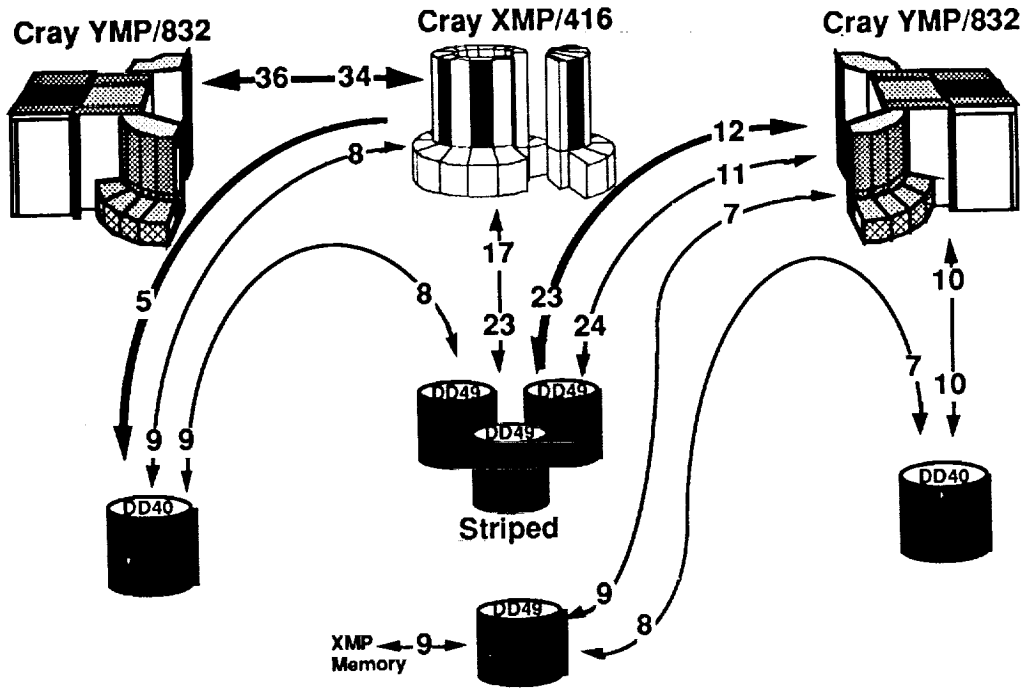


Figure 13 - FTP Test Results Using UltraNet with 2 Cray's

Cray XMP had both striped (3X) and non-striped disks available. The Cray YMP had only non-striped DD40 disks available. In this test, FTP was modified to have buffers of up to 1 MByte in size to take advantage of the transfer capabilities of the UltraNet.

Clearly, this test shows that generally speaking the UltraNet was not a bottleneck in maintaining the high disk transfer rates available on the Cray. Over the network, the Cray YMP could write the striped disks on the Cray XMP as fast as a user sitting directly on the Cray XMP (23 MBytes/sec). Disk-to-disk rates between the DD40 on the YMP and the DD49 on the XMP was very close to the non-network rates (8 MBytes per second).

Finally, it should be instructive to examine the network performance of an actual installation using a high-speed Compute Server (Cray 2 and YMP), file servers (Amdahl 5880) and workstation computer servers (SGI). At NASA Ames Research Center, UltraNet is installed as a gigabit/second backbone across several buildings connecting the supercomputers, file servers and more than 40 SGI workstations (all equipped with the Powerchannel I/O option). Figure 14 diagrams the configuration at NASA Ames Research Center.

Tests were run on a variety of these hosts to demonstrate actual performance achieved in a variety of scenarios. Each test was run in a heavily loaded system with over 100 users logged in and competing for resources. Therefore, each run was repeated several times (variation noted in the results) to give an idea of the range of results possible. Dedicated testing should prove higher effective data rates.

Figure 15 summarizes this data. For memory-to-memory tests, using the TSOCK application, it is observed that the maximum transfer rates possible approach up to 92 MBytes per second for a single graphics application. (Over 32 MBytes per second still left for other traffic.) For transfer between Computer Server (Cray YMP) and File Server (Amdahl 5880), UltraNet can sustain memory-to-memory transfer rates approaching 22 MBytes/second using striped BMC channels on the Amdahl (UTS) and HSX channels on the Cray (UNICOS). Although testing has not been performed as yet on the disk to disk rates on the Amdahl, it is expected that near 20 MBytes/sec can be achieved due to the software striping possible at the NASA site on the UTS based Amdahl 5880 system.

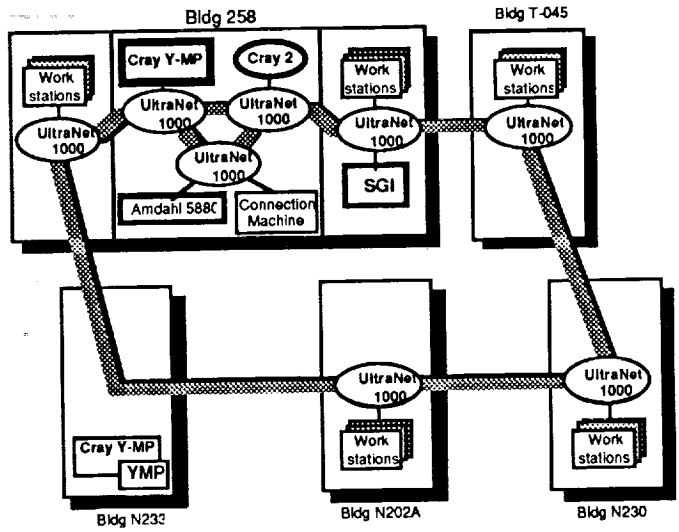


Figure 14 - NASA AmesGigabit/s Network Configuration

Perhaps notable also is the transfer rates between the disk on the Cray YMP and the memory of the SGI workstation. A user running an interactive application on the workstation can sustain data transfer at rates of over 4.5 MBytes per second from the Cray YMP disk. This would permit a user to run a large simulation on the Cray and access the results as it progresses interactively from the SGI without interrupting the Cray simulation.

For disk to disk tests, FTP was used between the various computers. FTP between the Cray and SGI disks maintained the maximum data rate possible for the SGI disks (about 1.5 MBytes/second) as demonstrated by timing the SGI disk rates (using the dd command in UNIX). Between the Cray computers, FTP transfered at somewhat lower rates than possible from a single Cray to it's own disk, in the range of 2.5 to 3.5 MBytes/ second over the network. However, it was shown by TSOCK disk to disk tests that if the FTP buffer sizes are increased to 1 or 2 MBytes the transfer rates approach that of the dd rates on a single disk (without the network).

Transaction oriented file access provided by NFS was also measured. Only NFS reads are possible at NASA Ames. Users can create the files only on the same computer they are using, but can read disks attached to the other Cray using NFS over UltraNet. Transfer rates were significantly better than what might be achieved using Ethernet, but were considerably lower than those achieved using FTP. NFS is a very transaction oriented protocol, uses the host stack instead of the network-based protocol processing provided by an UltraNet processor and therefore sees much less performance than other applications over the UltraNet. However, the Ultra was still able to provide from 1 to 2.6 MBytes/ second transfer rates, more than available with ethernet. However, UltraNet can support a much large number of NFS transactions than ethernet (not shown here).

It is expected that future modifications to NFS by Sun Microsystems and improvements in UltraNet's ability to handle the small packets of NFS transfers will improve this NFS rate.

In summary, File Servers and Supercomputing environments need high performance networks to balance the I/O requirements seen in today's demanding computing scenarios. UltraNet is one solution which permits both high aggregate transfer rates and high task-to-task transfer rates as demonstrated in actual tests.

UltraNet provides this capability as both a Server-to-Server and Server to Client access network giving the supercomputing center the following advantages:

- highest performance Transport Level connections (to 40 MBytes/sec effective rates)
- matches the throughput of the emerging high performance disk technologies, such as RAID, parallel head transfer devices and software striping;

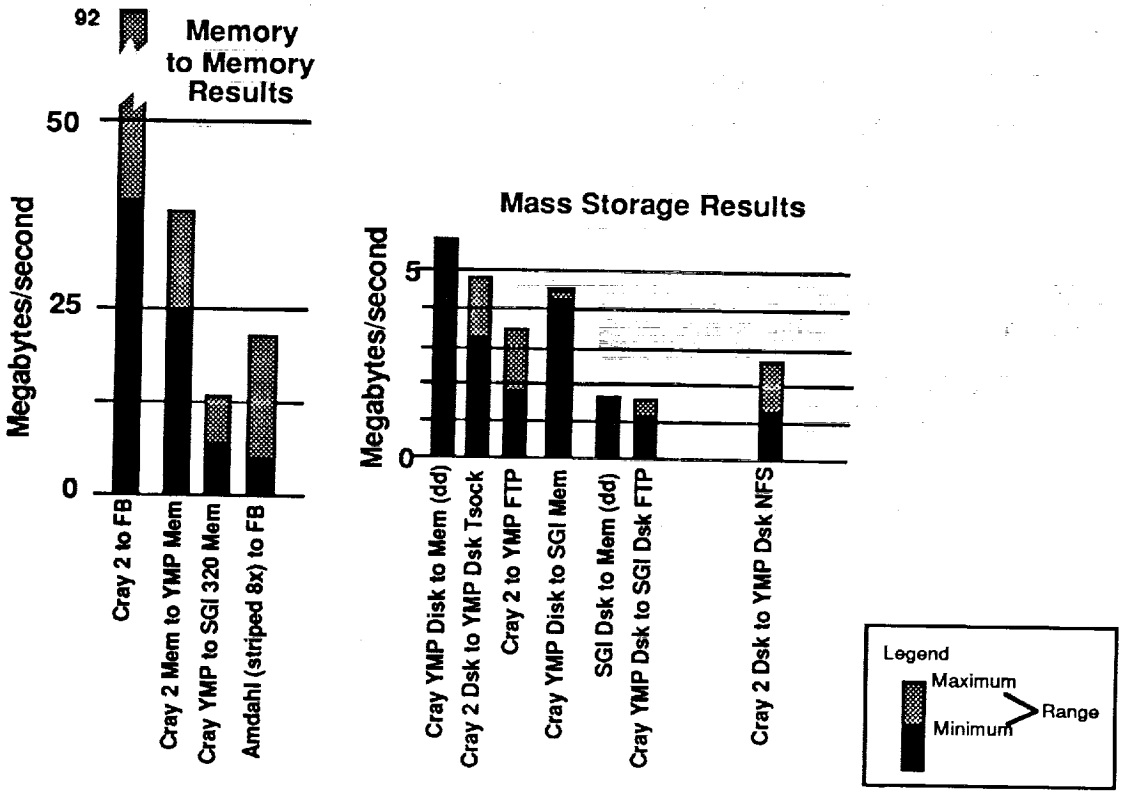


Figure 15 - Typical Application Test Results At NASA Ames Research Center (UltraNet)

- supports standard network and file system applications using SOCKET's based application program interface such as FTP, rcp, rdump, etc.
- Supports access to NFS and LARGE aggregate bandwidth for large NFS usage;
- provides access to a distributed, hierarchical data server capability using DISCOS UniTree product;
- Supports file server solutions available from multiple vendors, including Cray, Convex, Alliant, FPS, IBM, and others.

This paper appeared in the Cray User Group Spring 91 Proceedings (London, England).

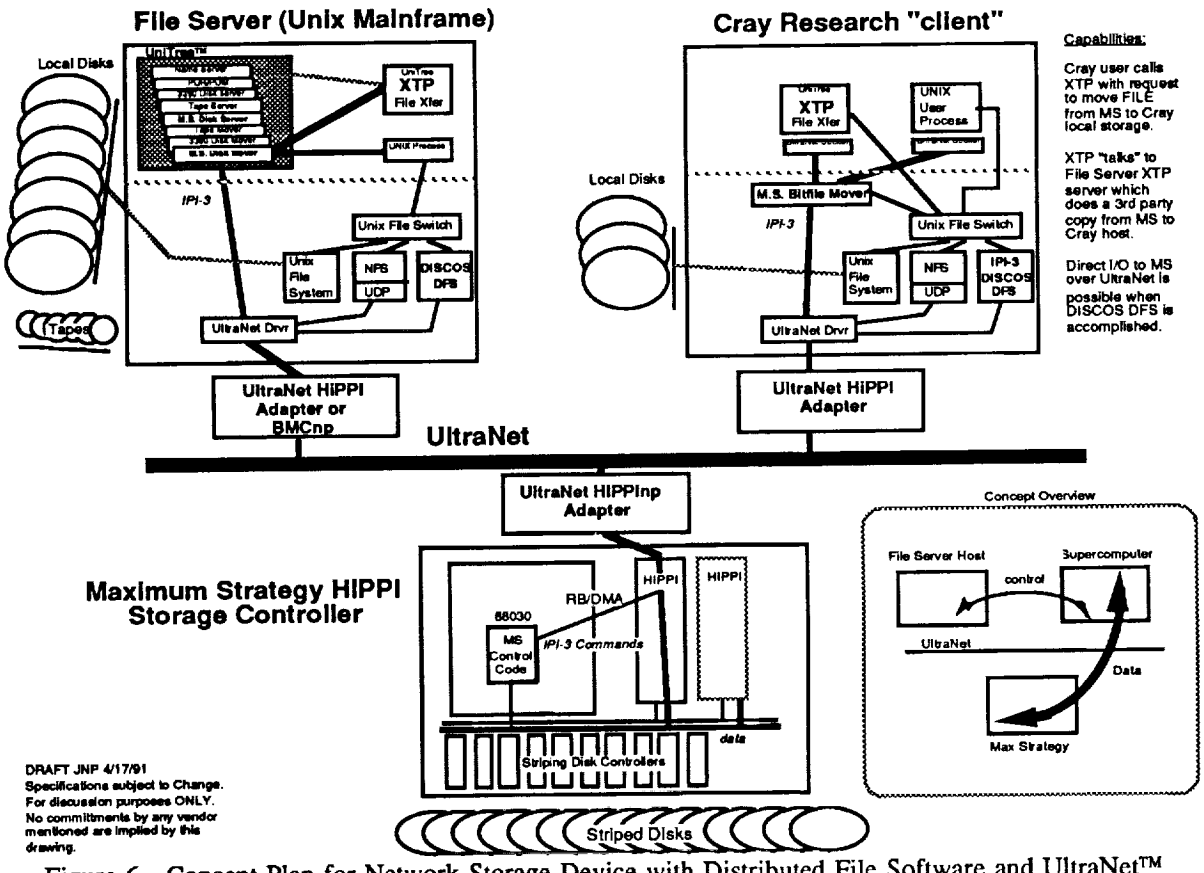


Figure 6 - Concept Plan for Network Storage Device with Distributed File Software and UltraNet™

Network Issues for Large Mass Storage Requirements

Presented to the
NSSDC Conference on
Mass Storage Systems & Technologies
for Space & Earth Science Applications

By

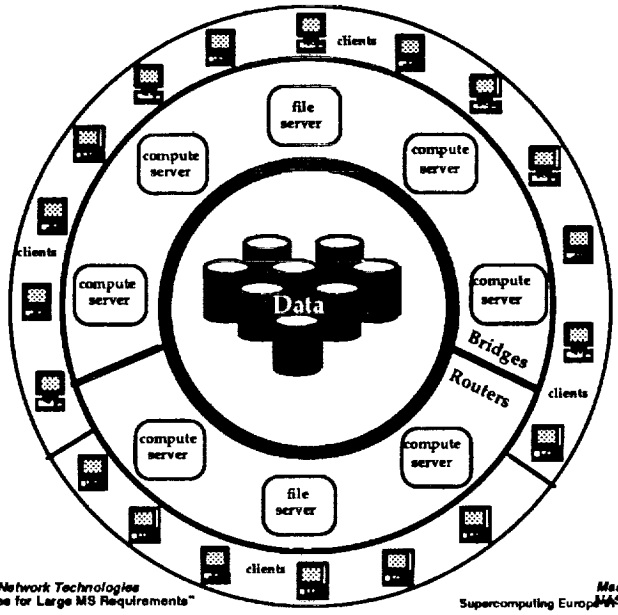
Newt Perdue
Vice President
Ultra Network Technologies
San Jose, California
U.S.A.

July 24, 1991

Overview

- File Server Network Requirements
- File Server Performance Trends
- UltraNet as a Performance Solution
- UltraNet File Performance Data

Emerging Client-Server Model



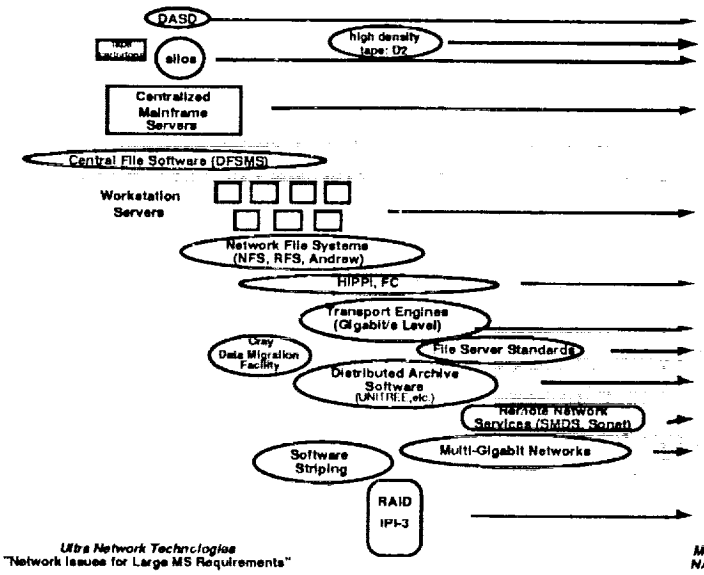
File Server Uses for Sci & Eng

- **Central File Storage**
  - Move files between Storage Hierarchies & Compute Servers
  - Support DISKLESS nodes (remote record reads)
- **Temporary or Workspace Storage**
  - Distributed Processing (provide common buffers for large data sets)
  - Image/Graphics display (digital VCR)
  - Network cache to match high speed systems to lower speed
  - Wide Area communication buffering (similar to cache)
- **Archiving to Reliable Storage Medium**
  - Very large but frequent used files
  - All files for reliable long term storage

Networked File Server Requirements

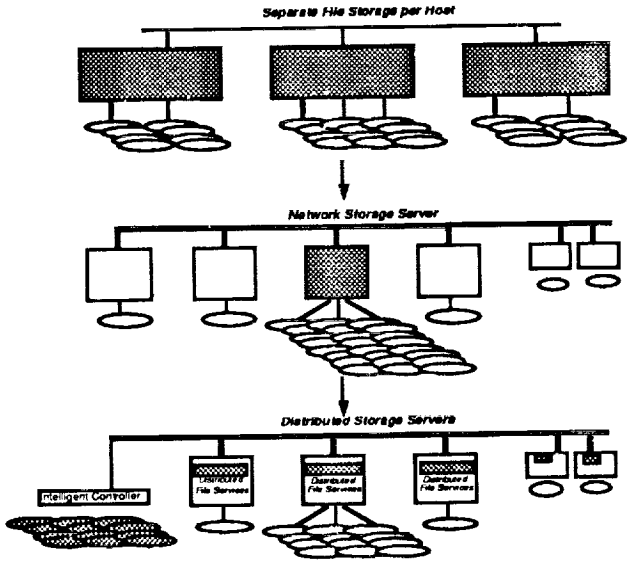
- File Management
  - archive management
  - spooling between hierarchies
  - catalog management/file scheduling
- Disk Performance Technologies
  - parallel disk head technology
  - striped disks (software)
  - striped disk controllers (hardware)
  - striped file servers
- Network Performance
  - high effective Throughput (pt to pt)
  - low latency for transaction oriented applications
  - connectivity to highest performance channels/busses
  - standard protocols for heterogenous systems
  - high aggregate bandwidth

File Server Systems
- Evolution not Revolution

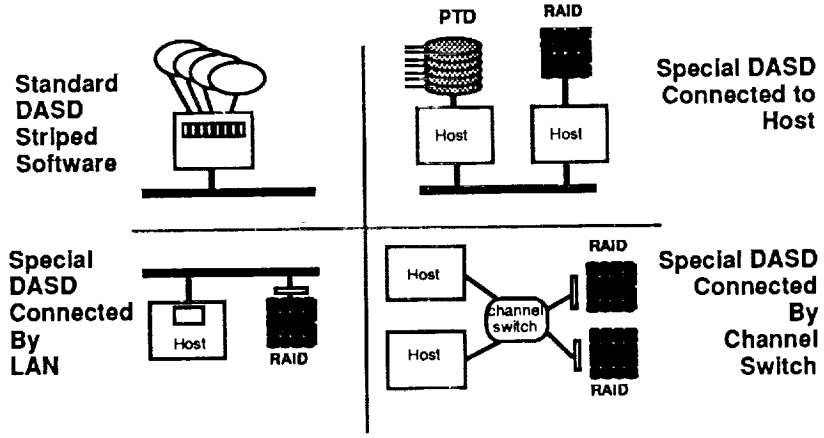


Future File Server Systems

Trends in File Servers



High Performance Disk Network Connection Strategies



Distributed File Server Trends

Factors Which Will Affect Cost/Performance Trends

- Smaller CPU's with medium I/O capabilities can control Distributed File Systems
- Transport Based Protocol Engines can provide reliable transport for network storage devices
- Standards (ala HIPPI, FC, !PI-3) create more competitive marketplace for devices
- Standards (ala IEEE MS) create more competitive marketplace for software
- Technology advancements continue in improving cost/performance of devices

Major Issues for File Transfer

- Files are getting substantially larger
  - files today range from small to several GIGAbytes in size
  - slow LAN performance limits feasibility of some file transfers
  - LAN is congested during transfers
  - MTBF of hosts/disks/LAN can be less than file xfer time
- LAN utility is determined by EFFECTIVE performance
  - EFFECTIVE Performance 7 - 100 times slower than "wire" speed
- File transfer impacts valuable host resources
  - As network "wires" get faster, it takes more CPU to be efficient
- Current system structures sized for small transfers
  - slow LAN's
  - slow effective disk transfer rates
  - application I/O buffers small
  - system buffers small & require copies

File Server Performance

Clients
- OS latencies
- Access Methods
- Network Software
- Application
- CPU Speeds
- I/O Speeds
- Device Speeds

Networks
- End to End Performance
- Aggregate Performance
- Connectivity Choices
- Interoperability
- Distance Latencies
- Host Integration

Servers
- Application
- Server Algorithms
- System Algorithms
- Network Software
- CPU Speeds
- I/O Speeds

Devices
- channel speeds
- device speeds
- device latencies
- allocation methods
- block sizes
- controller intelligence
- controller memory
- controller speeds
- controller protocols

The Performance Chain

The User Sees One Level of Performance and Reliability:

Influenced by the slowest and most unreliable in the chain

Network Performance Myths

You'd Have A Much Higher Performance Network If You Only Had:

- **Fiber Optics**
- **Switches Instead Of Busses**
- **A Lighter Weight Network Protocol**
- **A Faster Channel, ala HiPPI or ESCON**
- **Multiple Simultaneous Data Paths**
- **A New Computer**
- **Wave Division Multiplexing**
- **A Faster Disk System**

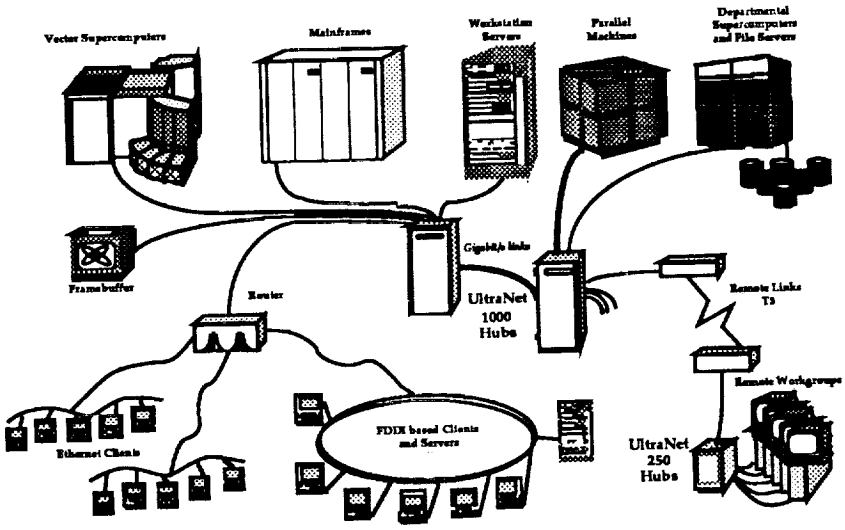
Gigabit/s Network Issues

- System Issues Dominate Performance Not Fabric
- Network Problems Dominated By Large Speed Range
- Applications Determine Realized Performance
- Higher Speeds Uncover Many Vendor/System "limits"
- Integration With Existing Network Technologies

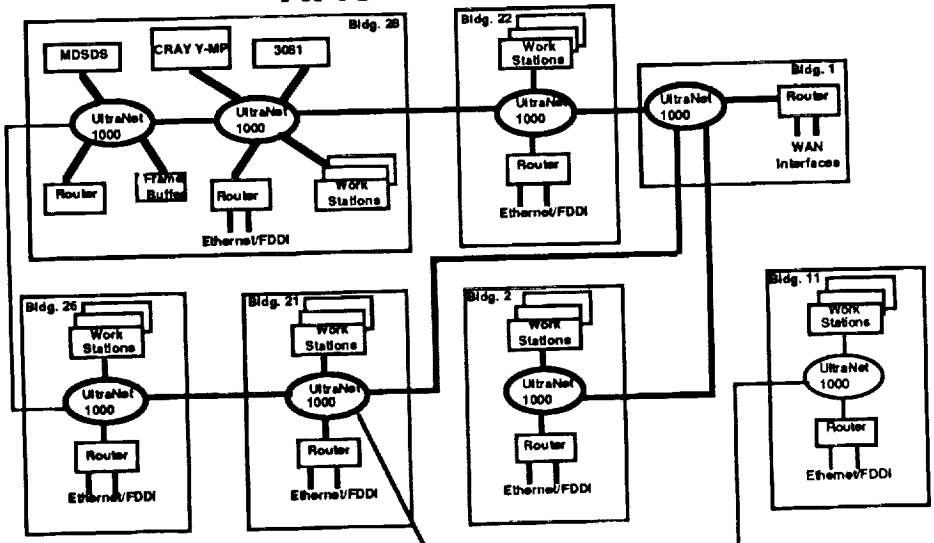
UltraNet as a System Solution

- UltraNet Is Transport Level Service To Host
- Data Delivered Directly to User Buffer From Channel
- Protocol Processing In Adapter - Reduces Host CPU cycles
- Decouples Host Transaction Sizes From Network Packet Sizes
- Uses Large Packet Sizes When Between UltraNet Connected Servers
- Uses Standard Packet Sizes To Other Networks
- Fully Participates In Internetting Environment (RIP, ARP, SNMP)

UltraNet Topology



Server Network Concept
NASA Goddard



UltraNet as Server Network

- **Front End**

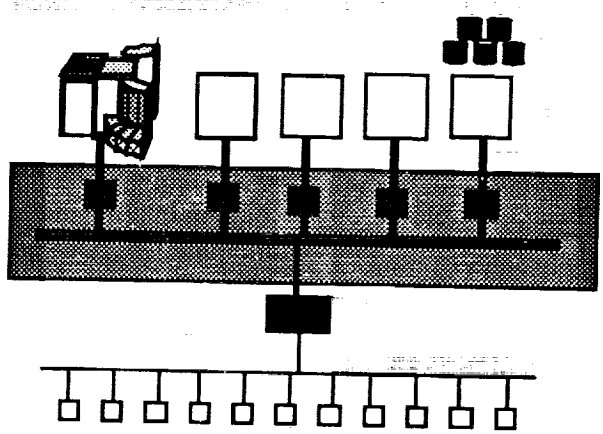
  Eliminates need for FDDI Adapters directly on Servers

- **Server Network**

  Large Aggregate and pt-pt xfer rates for direct connected servers

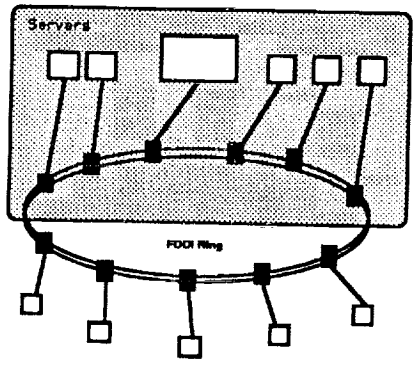
- **Backbone**

  Connect multiple buildings & other networks at gigabit/s rates

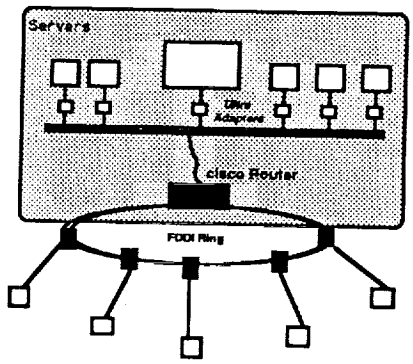


Server Network Alternatives

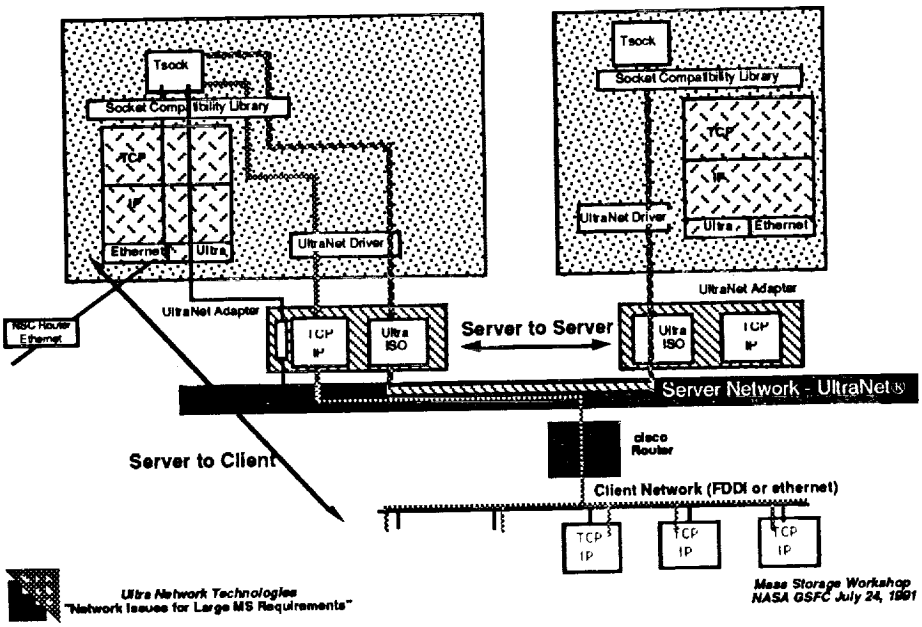
FDDI Only



UltraNet and FDDI

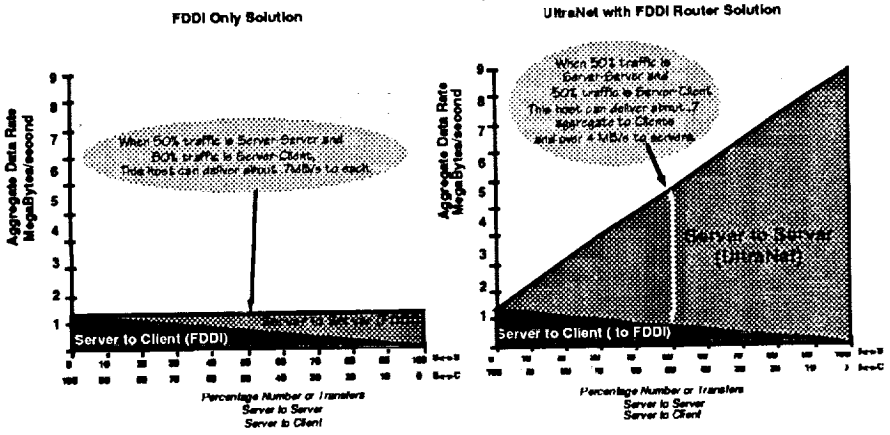


UltraNet as Server Network- the Path



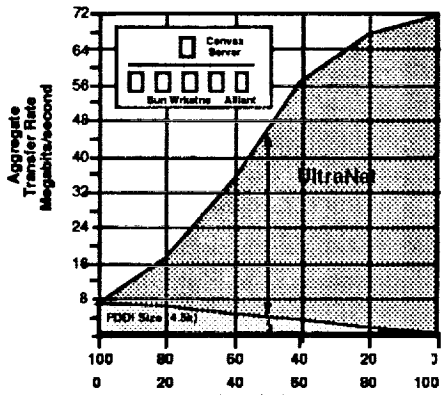
UltraNet Server Model

Application to Application Performance Simply Modeled From One Server to Multiple Server/Clients



Above data computed based on typical host transaction rates, network transfer rates, and network packet sizes.

UltraNet Server Test Results



Test Configuration

Tests Performance at Ultra Factory using TSOCK application.
Memory to Memory Xfer rates shown.

Computers used:

Sun 4, Sun 3
Sun OS 4.1.1

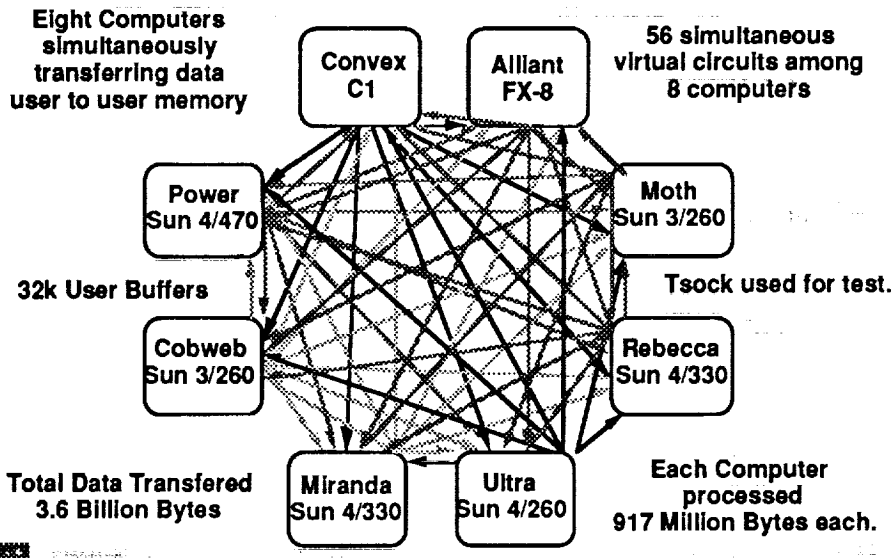
Convex C1
Convex OS 8.0.0.1

Alliant FX-8 Concentrix

UltraNet 3.50-03 software

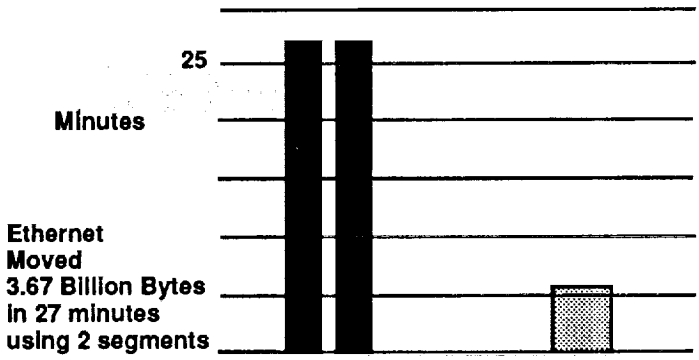
FDDI was not tested - FDDI host transaction size used to simulate the UltraNet performance from an Internet source.

UltraNet: Bandwidth Test



Bandwidth Stress Test Comparison

Eight Computers each with 7 full duplex conversations



~.83 Megabyte/second measured Aggregate per Ethernet Segment (not all ran concurrently)

11.8 MBytes/second measured Aggregate (all ran concurrently)

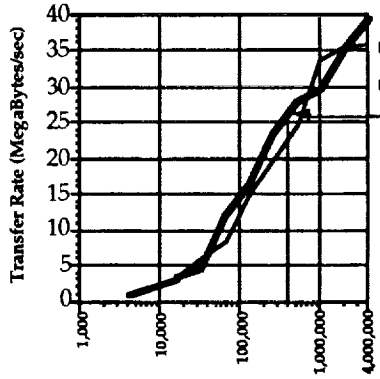
Cray File Server Networking using HIPPI Interface

Data taken at University of Stuttgart

Unicos 6.0

July 17, 1991

Data acquired during production time



Data xfer rate up to 25 MBytes/s possible

with UltraNet using 128k buffers & 3 way Striped Disks

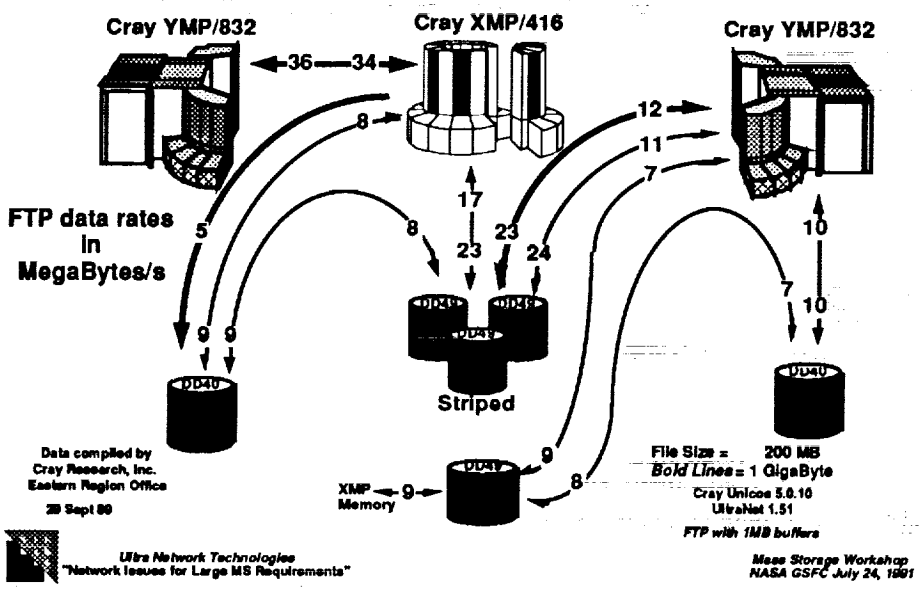
Cray 2 connected to Ultra HSXnp
Cray YMP2E connected to Ultra HIPPInp

TSOCK Test program
user buffer to user buffer

User Buffer Size

LOG Scale

UltraNet Performance - FTP Rates Between Two Crays



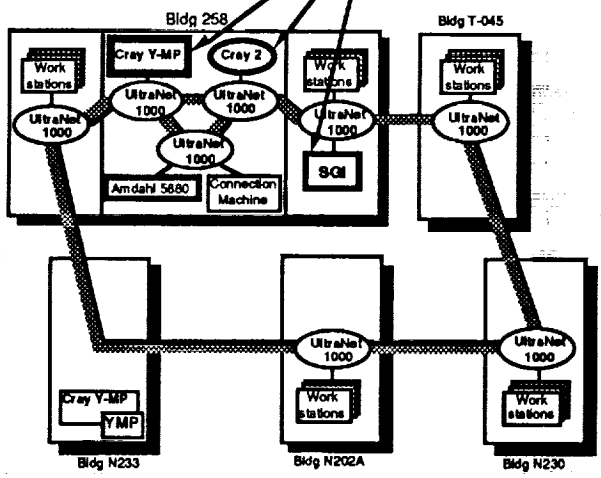
NASA Ames Research Center Performance Profile

Test Results

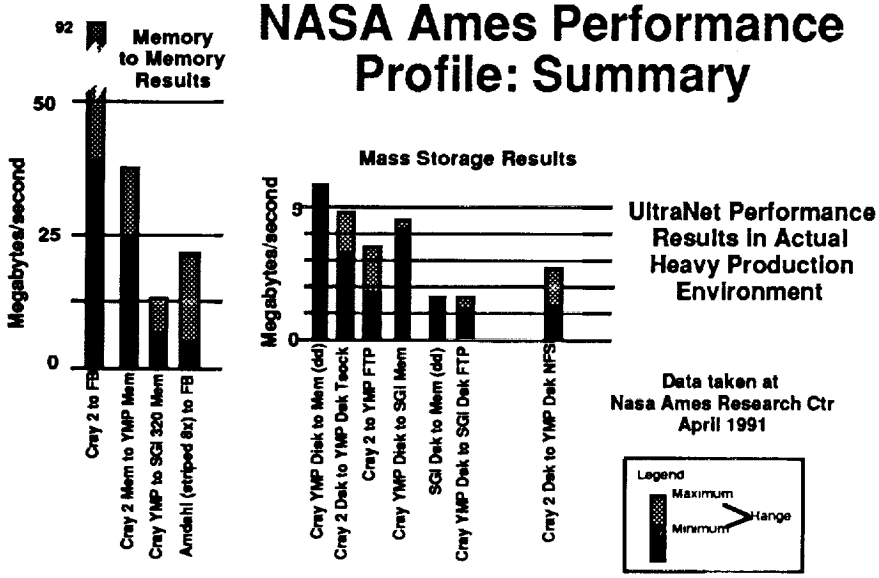
Cray 2
Unicos 5.1.11

Cray YMP
Unicos 5.1.11

SGI 4D/320 VGX
(with Powerchannel)
Irix 3.3.1



NASA Ames Performance Profile: Summary



File Server FTP Performance

	Typical FTP results Ethernet		UltraNet		*BENEFIT/Server* Single	Aggregate
Super <--> Wks	.25		.75		3.0 X	
aggregate		.35		30.0		85 X
Super <--> Mainf	.26		.60		2.5 X	
aggregate		.40		4 - 20		10-50 X
Mainf <--> Wks	.25		.60		2.5 X	
aggregate		.40		4 - 20		10-50 X
Wks <--> Wks	.25		.75		3.0 X	
aggregate		.35		4.0		11 X

Significantly More Users Can Be Supported with the Same Computing Resources for File Transfer Operations Using a Faster Network

UltraNet as File Server Transport

- Provides Highest Performance TRANSPORT LEVEL connection available 2 - 40 MBytes/second range for host to host transfers;
- Matches throughput of high performance emerging disk devices, i.e. RAID, vendor striped disks
- Supports standard SOCKET based Applications at increased speeds for FTP, rcp, rdump, user written applications
- Supports host based NFS access - improves network wide bandwidth for large NFS internets
- UNITREE application supports UltraNet for Distributed File and Archive Server Applications
- Other Applications In Test for Network Backup over UltraNet
- Supports several vendor based File Server Solutions:

  Cray Superserver; Convex, Alliant, IBM HMS, FPS

N93-14774

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

T. A. Gilbert

Network Systems Corporation
Vienna, Virginia

Introduction

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

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

The High Performance Parallel Interface

History

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

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

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

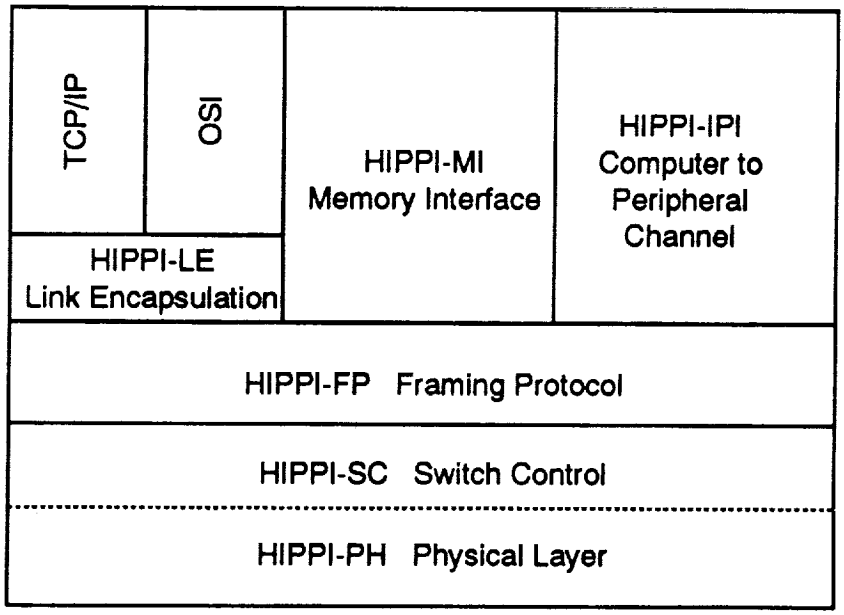
- standard interface which could be implemented by a broad range of vendors without the need for exotic or expensive technology. HiPPI physical layer interfaces can be built from off the shelf components which have been available for two decades.

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

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

The ANSI HIPPI Standards Suite

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



The six standards are:

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

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

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

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

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

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

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

HiPPI Technology

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

Computer Channels

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

Peripherals

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

Switches

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

Extenders

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

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

Gateways

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

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

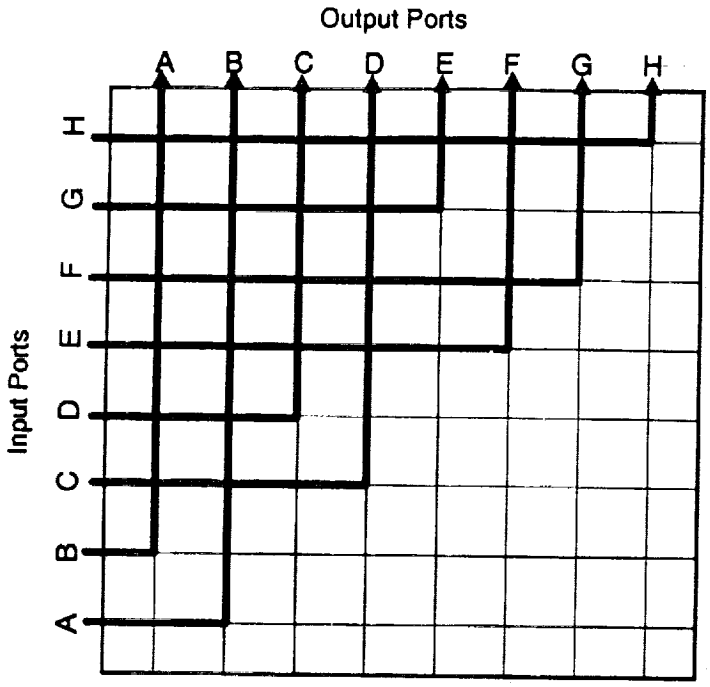
CABs will also exist within HiPPI networks to provides various types of bridging functions. For instance, where long haul extenders are inserted into a network it may be prudent to interface each end via a CAB. The CABs keep a permanent HiPPI connection up between them. Each CAB is prepared to accept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

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

Building Crosspoint Switch "Fabrics"

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

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



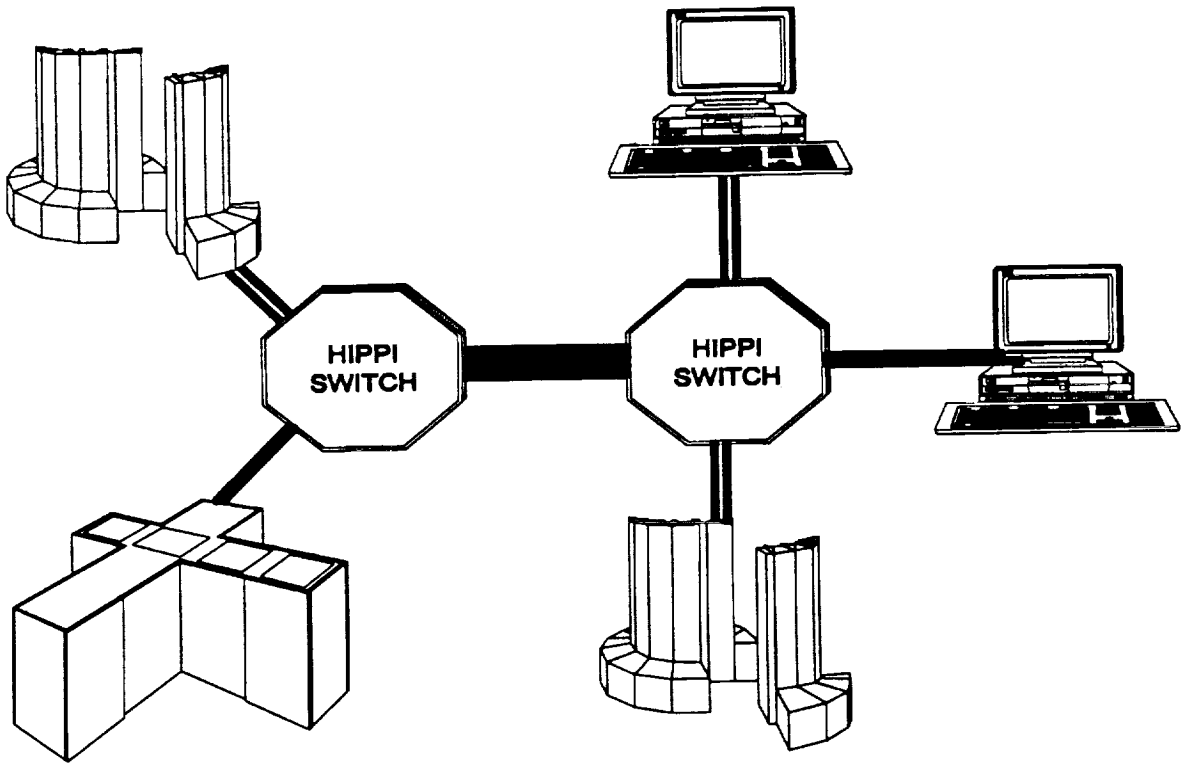
For near term applications in backend networking for supercomputers or attachment of peripherals, single stage switches with four to thirty-two port pairs are probably adequate. However, the limits of board to board connector technology means that we are rapidly approaching the limits of current switch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

Cascading HiPPI Switches

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

---

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



Addressing

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

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

Source Routed Addressing

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

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

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

Isomorphic Addressing

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

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

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

---

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

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

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

Switch Management

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

Auto-configuration

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

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

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

Additional Services

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

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

HiPPI Applications

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

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

Device Connections

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

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

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

Backend Networks

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

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

---

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

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

Backbone Networks

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

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

Isochronous Data Routing

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

Examples of digital data types considered for HiPPI channels are:

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

Potential Storage Subsystem Application

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

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

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

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

However, let's consider the following:

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

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

Related Emerging Standards

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

Fiber Channel

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

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

SONET

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

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

ATM

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

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

Conclusion

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

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

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

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

T. A. Gilbert

Network Systems Corporation
Vienna, Virginia
July 25, 1991

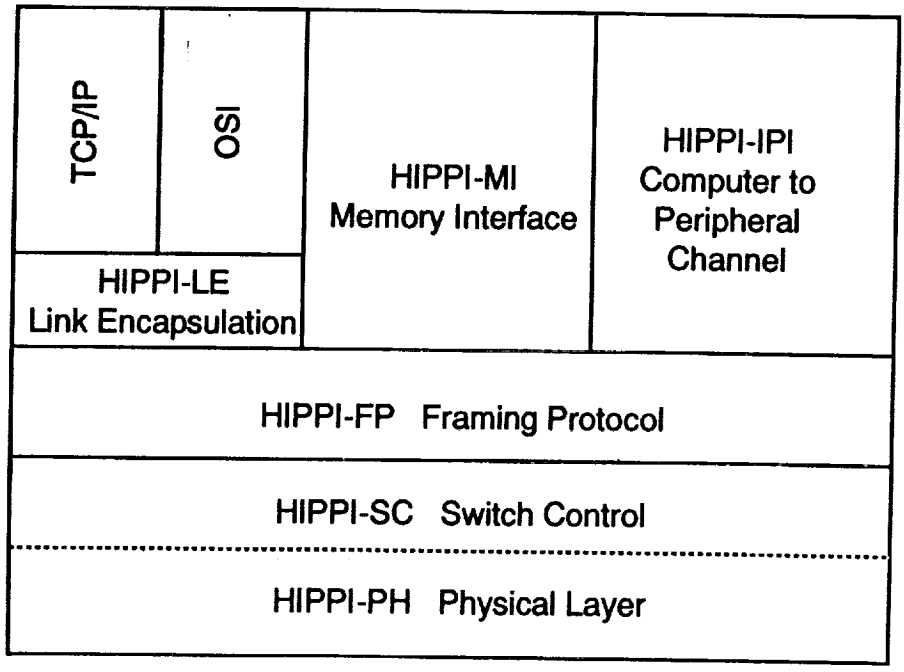
HIPPI

HIgh Performance Parallel Interface

Proposed ANSI Standard

Developed at Los Alamos National Laboratory

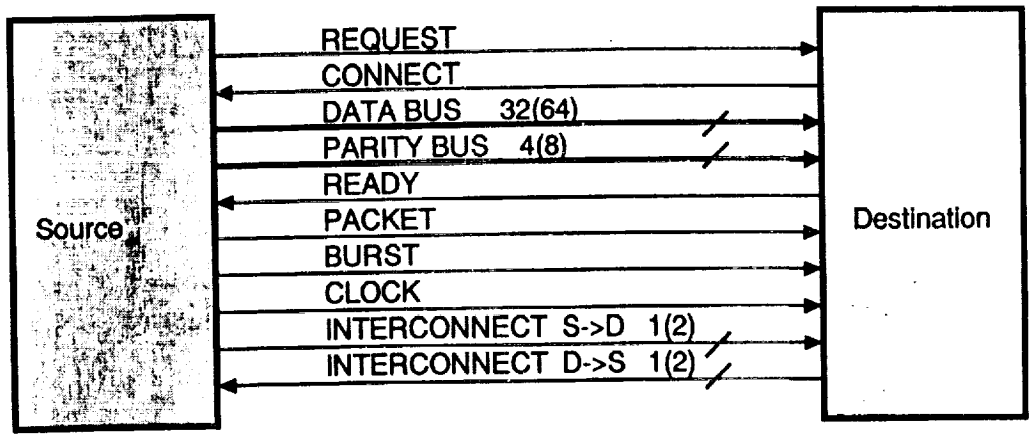
800 Mbps and 1600 Mbps Data Rates



HIPPI Physical Layer

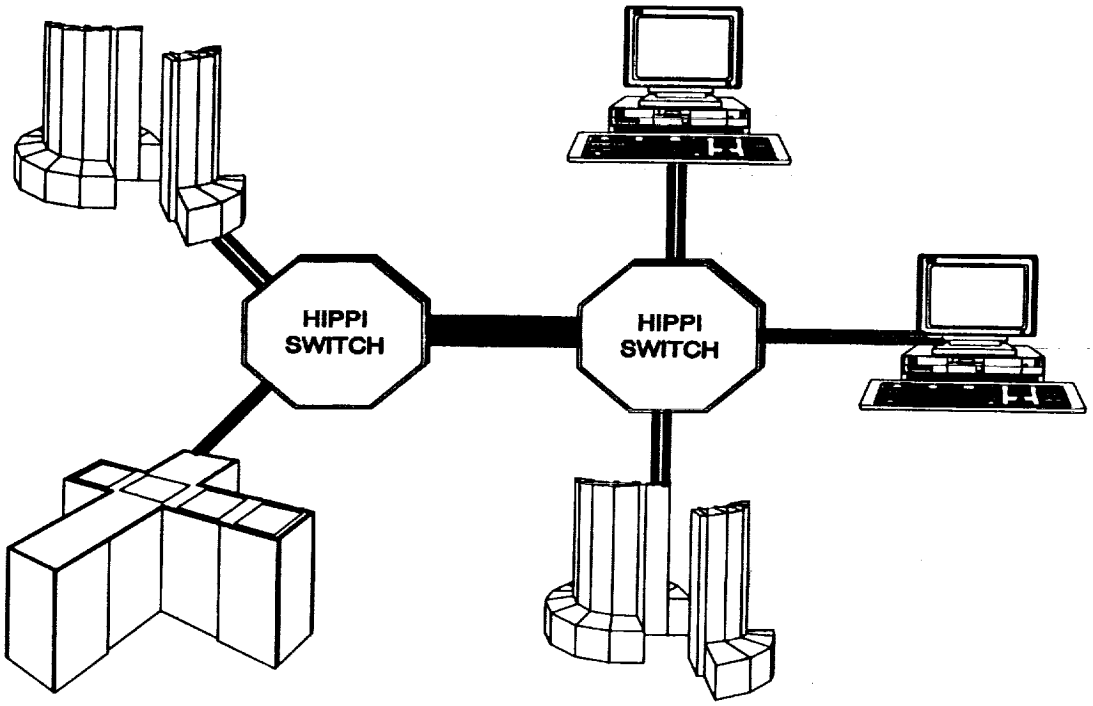
- 800 or 1600 Mbps isochronous interface
- Parallel 32 or 64 bit wide data bus
- 25 meter maximum cable length
- Simplex interface
- Parity and LRC data protection
- Ready resume flow control

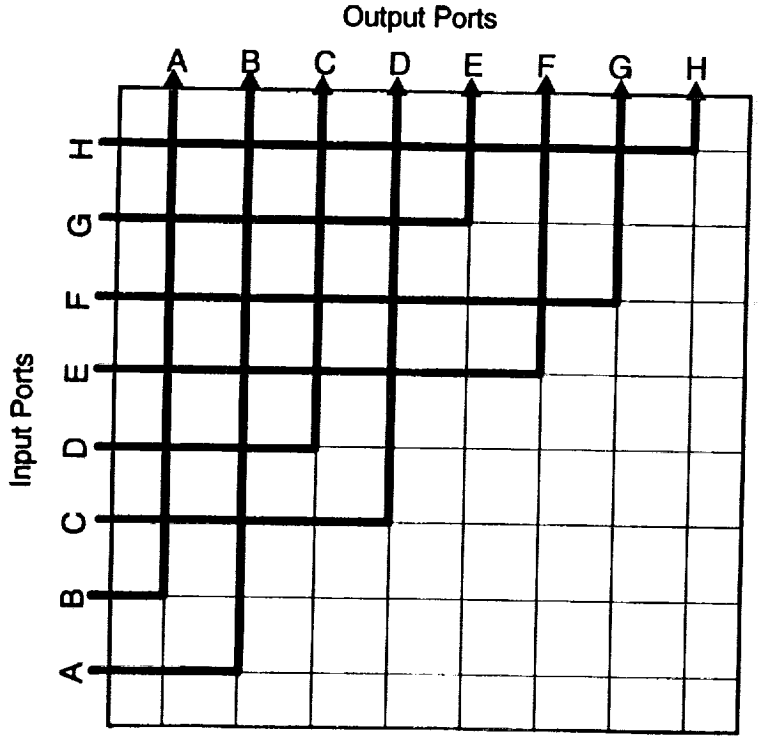
HIPPI Signal Summary



HIPPI Switch Control

- Optional extension of the physical layer standard
- Provides for inband switching of an HIPPI channel
- Uses physical layer "I Field"
- Defines two addressing modes:
  - Source routed addressing
  - Flat or isomorphic addressing
- Implemented in Network Systems HIPPI Switches





Simulatenous Data Paths

A > B
Full Duplex
B > A

C > D
Full Duplex
D > C

E > F Simplex

F > G Simplex

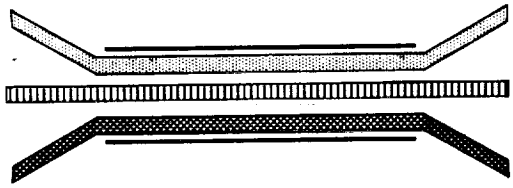
G > E Simplex

H > H Loopback

Switching Technology

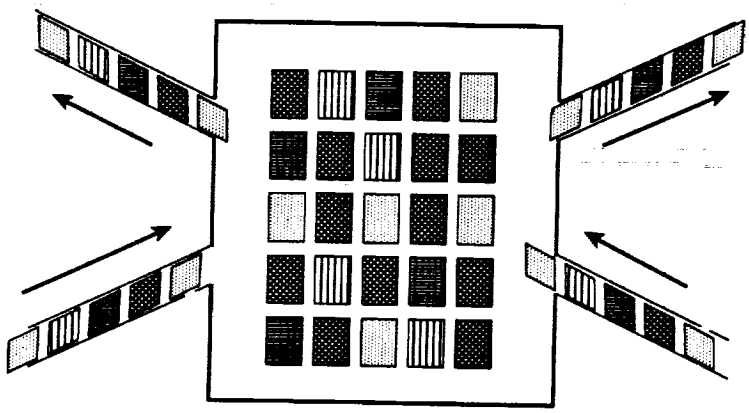
- Time Division
  - packet, frame, cell, byte, bit
  - shared media, shared memory
- Frequency Division
  - broadband CATV
  - wavelength multiplexed fiber
- Space Division
  - cross-point switches

Frequency Division



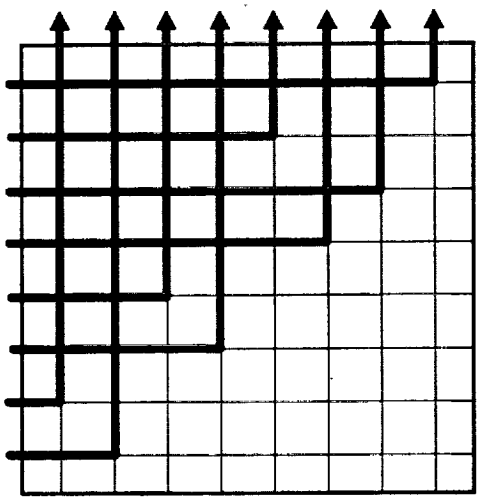
- Well suited to circuit switched networks
- Efficient use of media bandwidth
- Considered outmoded for data
- Will see renaissance in fiber wave division multiplexing

Time Division



- Two decade favorite of datacom
- Extremely flexible
- Basis for all current networking
- However, throughput limited by packet switch memory bandwidth

Space Division



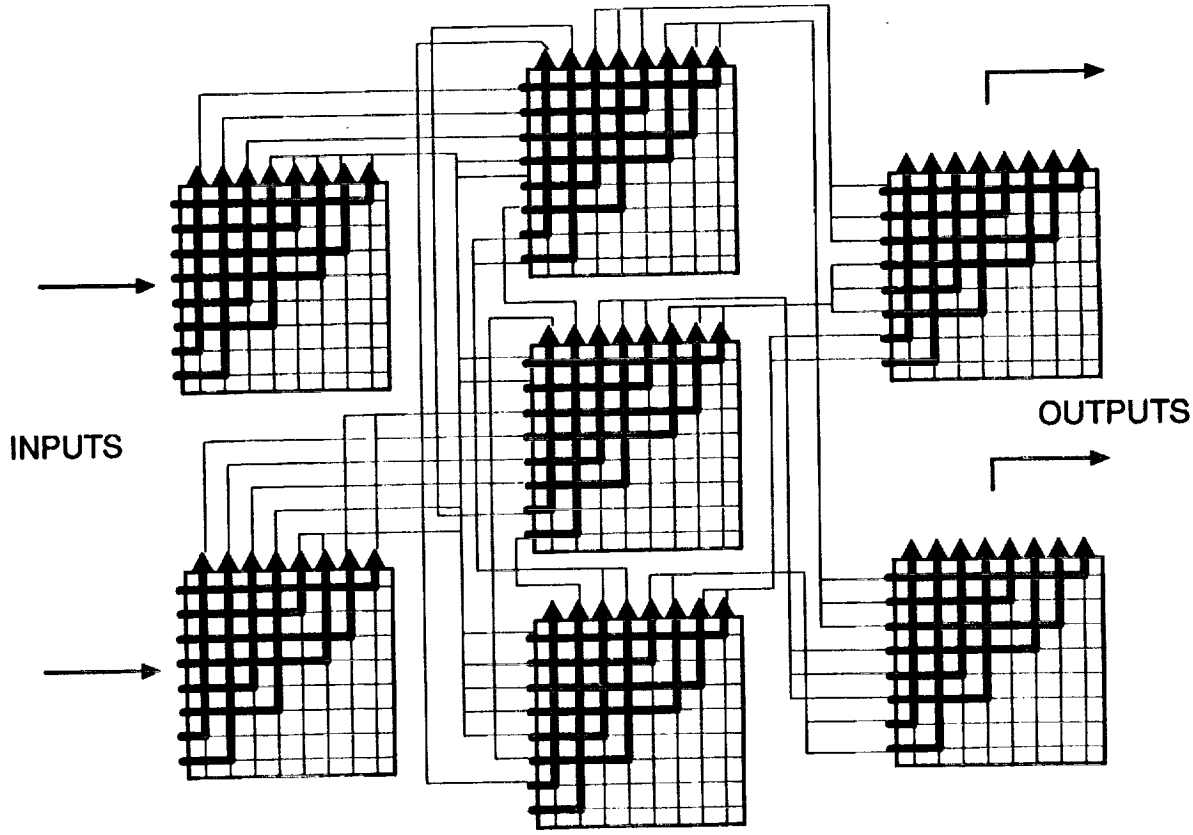
- Multipoint connectivity using patch panels or switches
- Imposes least constraint on signal type or data format
- Provides greatest aggregate bandwidth (not limited by memory or media)
- Traditionally considered ill suited for data due to switching time

HIPPI Technology

- Computer channels
  - IBM, Cray, DEC, etc.
  - workstations comming
- Peripherals
  - Disk arrays
  - Frame buffers
  - Tape cartridge drives
- Switches
  - 4X4 to 32X32 ports
- Extenders
  - Fiber up to 10 km

Developing HIPPI Technology

- Extended switch "fabrics"
  - Hundreds of ports
  - Auto-configuration
  - Multi-cast for address resolution
  - Advanced management
- HIPPI Gateways
  - SONET/ATM for wide area
  - Standard media like FDDI



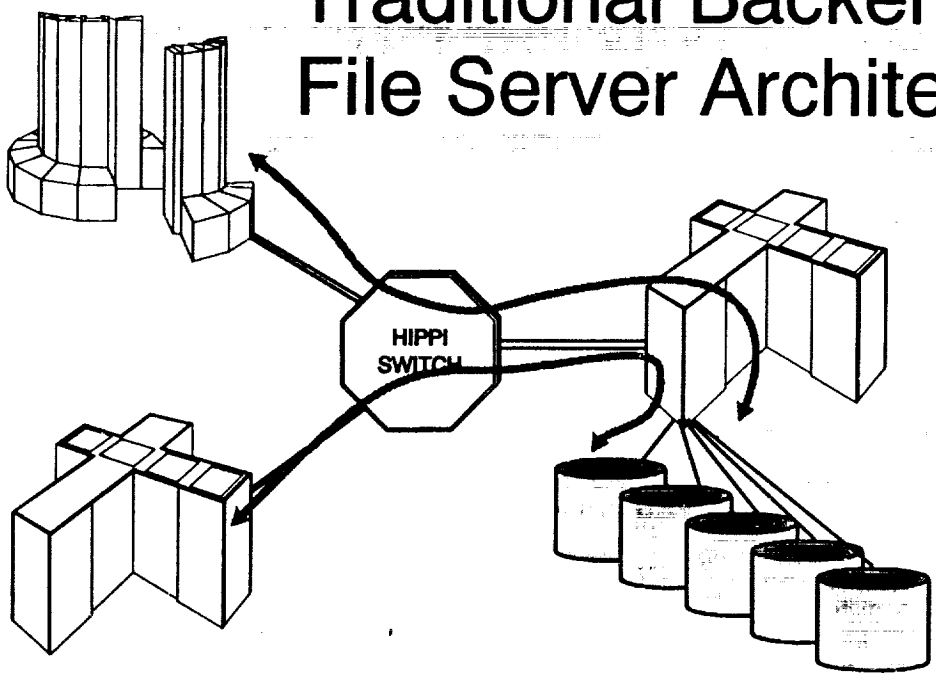
Why HIPPI?

- .8 & 1.6 Gbps data rates well matched to current high end systems requirements
- Connection oriented (as opposed to packet TDM) uses obviate elaborate media access schemes
- Flow control at the physical layer provides maximum throughput without transport layer tuning
- Sub-microsecond switching provides multi-point connectivity
- Standard protocol implementations now in the HIPPI range (CRI UNICOS TCP/IP > 500 Mbps over HIPPI)
- Channel characteristics allow device connection directly to network

HIPPI Back-end Networks

- TCP/IP is first target protocol.
- FTP and NFS upper layer protocols for file server applications.
- TCP/IP performance optimization required by almost all vendors.
- Cray UNICOS TCP/IP is expected to deliver 500-600 Mbps TCP memory to memory over HIPPI (YMP to YMP).

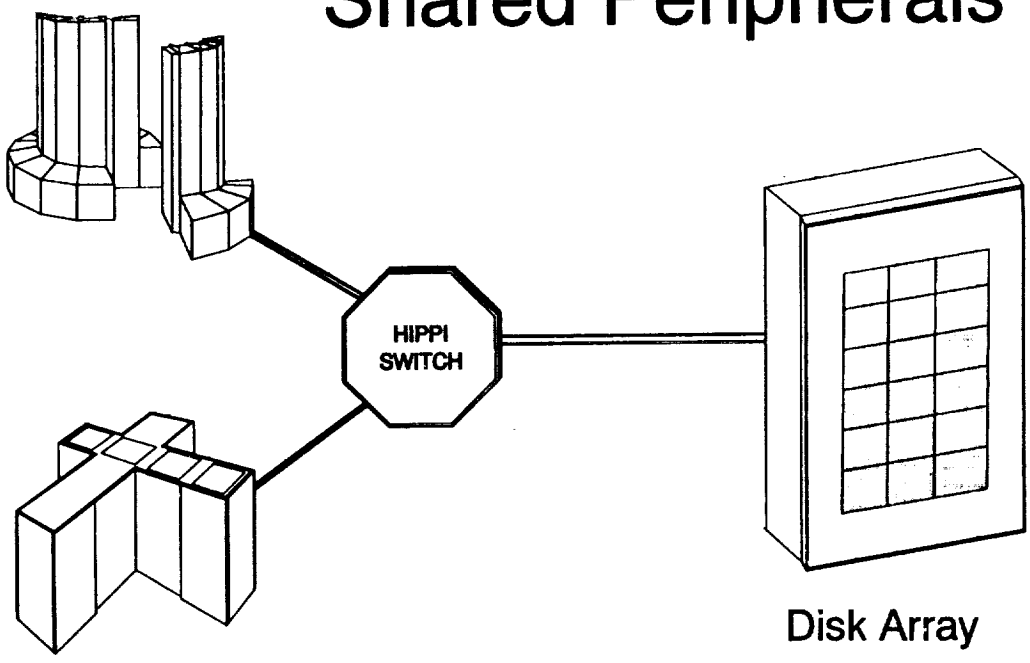
Traditional Backend File Server Architecture



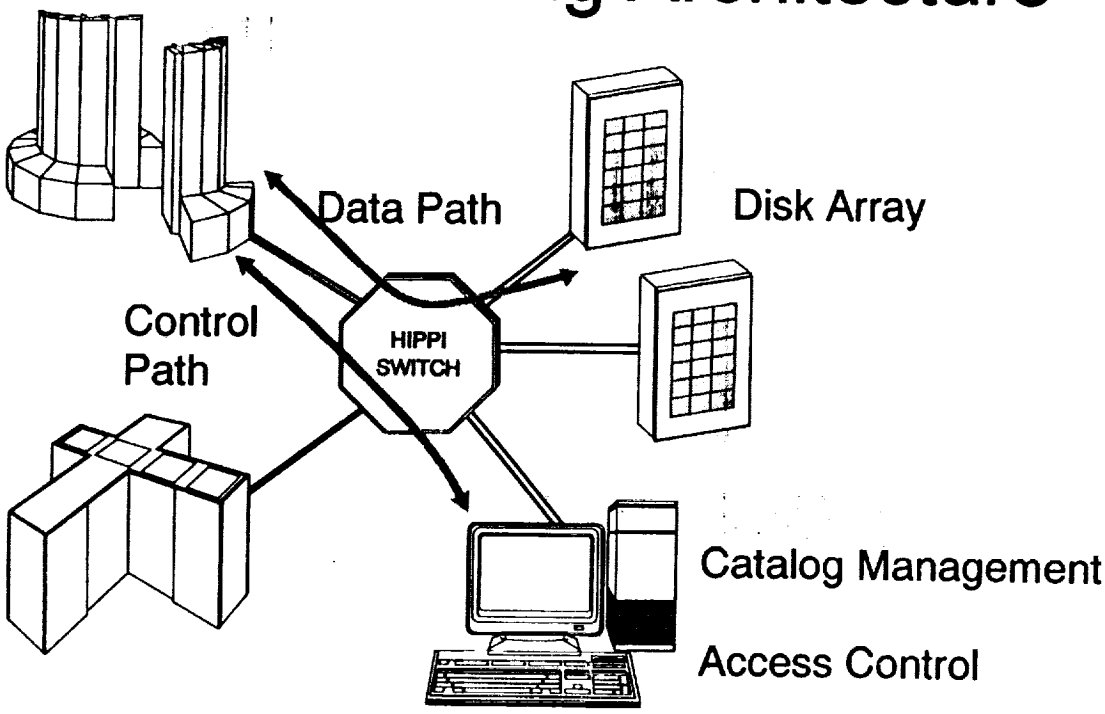
Switched Computer Device Interface

- Vendor independant channel standard.
- IPI-3 over HIPPI to support intelligent device controllers.
- HIPPI compatible disk array controller available now.
- HIPPI frame buffers available in 1H91.
- Promise of new visualization architecture.

Shared Peripherals



Evolving Architecture

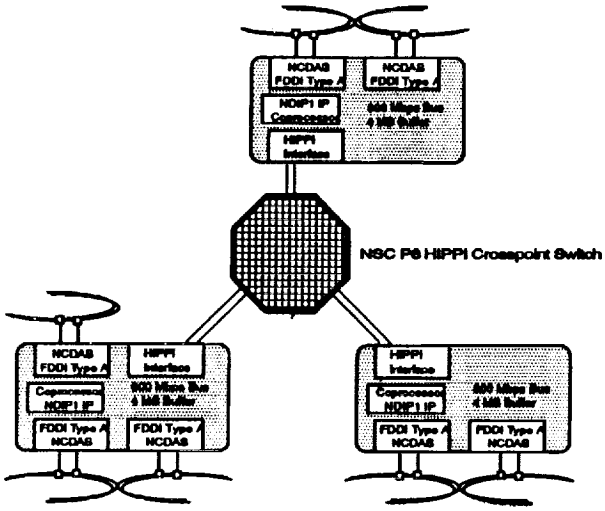


HIPPI Backbone for FDDI Networks

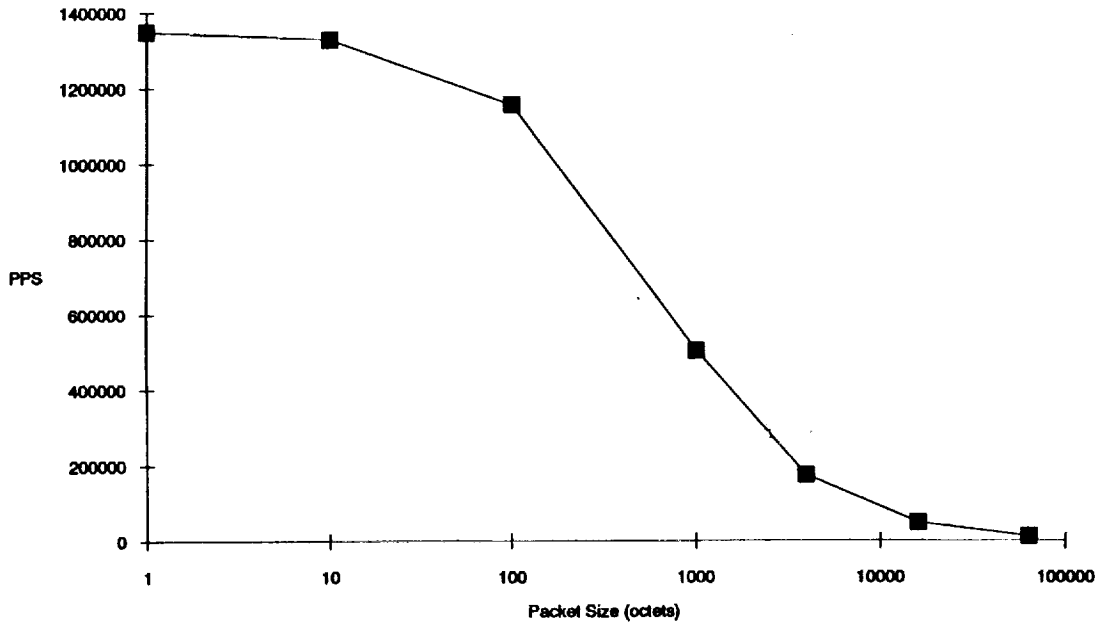
- Rapid acceptance of FDDI will create connectivity problems similar to those experienced with Ethernet.
- Multiple FDDI networks will face performance versus cost trade-offs when universal connectivity is needed.
- HIPPI "fabrics" apprear to offer a cost effective solution for the design of high performance large private inter-networks.

HIPPI Switch Backbone

- DX - P8 switch interface allows multi-gigabit per second FDDI backbone.

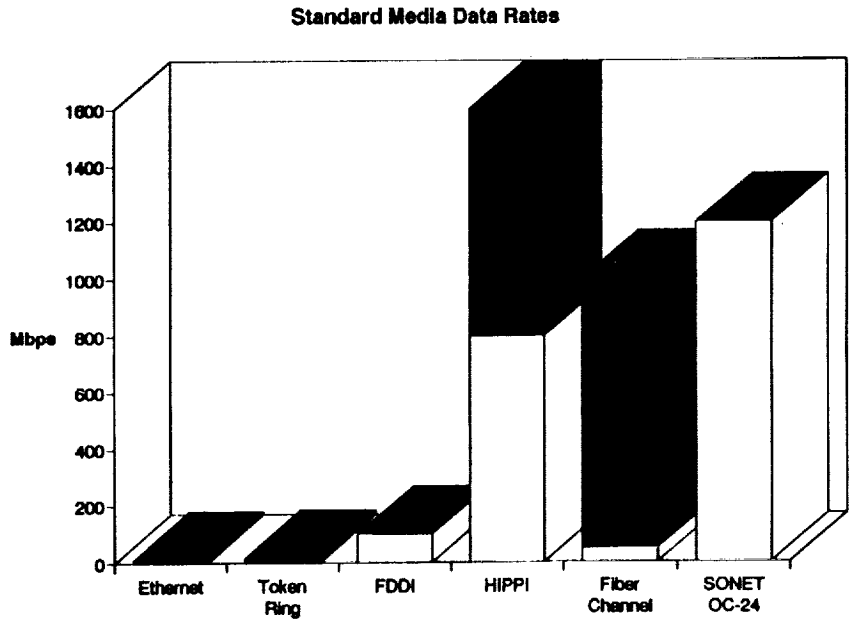


HIPPI Switch Packet Throughput

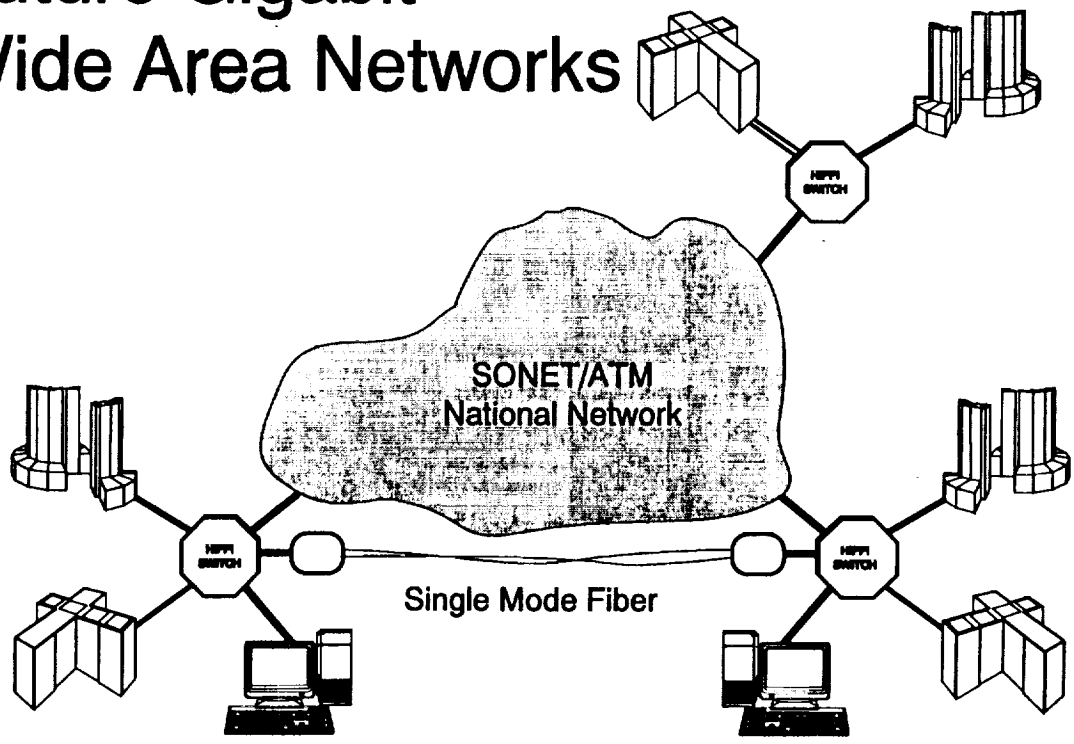


HIPPI as Interconnect Standard

- Low cost, high performance.
- Physical layer does not restrict upper layer application.
- Connection oriented approach will support "non-protocoled" data streams.
- Potential application to digital voice and imagery.



Future Gigabit Wide Area Networks



Summary

- HIPPI standards based products will begin to deliver multi-gigabit per second performance for computer applications in 1991.
- Within two years, new hardware and software will revolutionize high performance storage server architecture.
- Within five years, gigabit scale wide area access to data repositories will be technically feasible.

THE NATIONAL SPACE SCIENCE DATA CENTER
- AN OPERATIONAL PERSPECTIVE -

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

ABSTRACT

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

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

INTRODUCTION

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

INGEST AND ARCHIVE ENVIRONMENT

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

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

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

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

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

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

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

DATA RESTORATION

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

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

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

NEAR-LINE DATA ACCESS

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

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

RESEARCH TOOLS

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

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

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

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

PROACTIVE DATA MANAGEMENT FOR FUTURE MISSIONS

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

SUMMARY

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

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

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

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

The National Space Science Data Center
- An Operational Perspective -

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

July 25, 1991

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

An Operational Perspective

Introduction

Ingest and Archive Environment

Technology Pair

Data Stewardship

Data Restoration

Migration of Data to Network Accessible Media

Improvement in Research Tools

Proactive Involvement in Data Management Planning for Future Missions

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Introduction

The NSSDC is a heterogeneous data archive and distribution facility responsible for:

- Ingest
- Archive
- Distribution

Data Holdings:

- 110,000 data tapes
- 4000 data sets
- Six TBytes

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Ingest and Archive Environment

Data ingest Procedures

- Routine generation of two copies of all incoming data
- Use of precertified tapes
- Off site storage of backup

Metadata management to ensure useability of data

- Directory and catalog information
- Format information
- Inventory information

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Technology Pair

Purpose:

- The selection of two media technologies for data archival

Criteria:

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

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Technology Pair (cont.)

Current technology pair at NSSDC:

- IBM 3480 cartridge (primary)
- 9-track/6250 bpi (backup)

Migration strategy:

- Select new media technology
- Identify new technology pair
- Migrate data to new media
- Discontinue support for older media

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Data Stewardship

Charter:

- The total responsibility for data and media management lies with the primary archive data center

The NSSDC is addressing this responsibility through:

- Data restoration
- Migration of data to network accessible media
- Improvement in research tools
- Proactive involvement in data management planning for future missions

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Data Restoration

Purpose:

- The systematic migration of data from older media to current technology pair

Older media:

- 7-track
- 9-track/low density
- Many tapes are greater than 20 years old

Status:

- Twenty-five percent of media volumes completed
- Greater than 98 percent of data integrity preserved

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Data Restoration (cont.)

Experience:

- Storage environment of critical importance
- Frequent cleaning of drives necessary to avoid tape damage
- Problems have been manageable
- Use of precertified tapes recommended

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Migration of Data to Network Accessible Media

Increase public access retrieval on NSI wide area network

- Near-line storage - IUE example
  - Use of IBM 3850 mass store to manage over 70,000 images
  - Image size is typically one MByte
  - VAX interface to provide image ordering capability
  - Manual extraction and staging of data on FTP account
- On-line storage - NODIS
  - Used for small, often requested data
  - Captured VAX account provides browse and retrieval

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Migration of Data to Network Accessible Media (cont.)

Other Initiatives:

- NASA Data Archive and Delivery System (NDADS) optical disk juke box
- Mass Data Storage and Delivery System (MDSDS) six TByte system
- Increased network bandwidth

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Improvement in Research Tools

Through the use of tools the researcher is able to spend more time analyzing the data content, rather than the data formats

- Coordinated Data Analysis Workshops (CDAW)
- Common Data Format (CDF)
- Data browse

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Proactive Involvement In Data Management Planning for Future Missions

Adoption of problem avoidance strategies to anticipate future requirements:

- Cost model developed to estimate resource requirements for data archival
- Early participation in Project Data Management Plans (PDMP)
- Research into auto ingest of future data deliveries
- Involvement in EOS data system development/evaluation
- Sponsorship of data management conferences and committees
- Participation in the establishment of standards for data and media

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Summary

The NSSDC is a heterogeneous data archive and distribution center operating in a dramatically changing scientific and technological environment.

A conservative but forward-looking approach to data management is necessary to avoid situations that could jeopardize the integrity and availability of its data holdings.

Its 25 years experience in data and media management enable the NSSDC to proactively participate in the challenges of the EOS era.

EOSDIS

DADS

REQUIREMENTS

J. BERBERT
B. KOBLER

NSSDC Conference on Mass Storage Systems
NASA/Goddard
July 23-25, 1991

EOSDIS DADS REQUIREMENTS

by J. Berbert and B. Kobler

ABSTRACT

A brief summary of the EOSDIS Core System (ECS) DADS requirements is given, including the ECS relationship to EOSDIS Version-0, phased implementation of ECS, and data ingest, archive, and distribution daily data volumes anticipated at each of the 7 Distributed Active Archive Centers (DAACs).

EOS GOALS

The Earth Observing System Data Information System (EOSDIS) Data Archive and Distribution System (DADS) is part of the Earth Observing System (EOS) program. The EOS program goals are given in Fig. 1. In short the goals are to acquire, access, and analyze Earth Science data as NASA's contribution to the Global Change Research Program.

PHASED IMPLEMENTATION

The full capability of the EOSDIS DADS is built up in a series of steps, as indicated in Fig. 2. Version-0 (V0) implementation began in 1990 with an estimated data volume of 5 Terabytes (TB) and is expected to grow to 33 TB by 1994.

Version-1 (V1) and Version-2 (V2) are part of the separately funded EOSDIS Core System (ECS). A request for proposals (RFP) for the 10 year contract to build the ECS was released by the Government on July 1, 1991. During V1 implementation the ECS archive is expected to grow from 10 TB to 40 TB, and the number of active DADS is expected to grow from 1 to 7. the 3 DADS at GSFC, Langley, and Marshal are to be operational for the Tropical Rainfall Measurement Mission (TRMM), which is scheduled for launch in 1997. The V2 implementation of ECS is primarily to support EOS-A1 with its order of magnitude increase in data products the first year and a subsequent increase of about 330 TB, or one third Petabyte (PB), per year, thereafter.

EOSDIS V0

The contribution anticipated from V0 and the specific relationship of V0 to ECS are given in Fig. 3. It is anticipated that V0 will provide significant heritage to ECS through prototyping efforts and by working towards interoperability amongst existing data systems.

ECS SEGMENT AND DAACS

A logical system architecture for ECS is shown in Fig. 4 (taken from the ECS RFP Statement of Work (SOW)). The 3 ECS segments shown are the Flight Operations Segment (FOS), the Communications and System Management Segment (CSMS), and the Science Data Processing Segment (SDPS). The SDPS includes the Distributed Active Archive Centers (DAACs) and the Information Management System (IMS). A DAAC includes a Product Generation System (PGS) with a collocated DADS and a distributed part of the IMS.

DAAC LOCATIONS

The locations of the 7 DAACs are shown on the Fig. 5 map. They are at Goddard Space Flight Center (GSFC), Jet Propulsion Laboratory (JPL), EROS Data Center (EDC), Langley Research Center (LaRC), National Snow and Ice Data Center (NSIDC), Alaska SAR Facility (ASF), and Marshall Flight Center (MSFC).

DADS FUNCTIONS AND REQUIREMENTS

The 5 major DADS functions, namely Ingest, Archive, Process Orders, Manage System, and Distribution, are given in more detail in Fig. 6, along with some of the key performance requirements for ingest and distribution. A key performance requirement for ingest is to be capable of accepting Level-0 (L0) data from the Customer Data Operations System (CDOS) at a high data rate. Key performance requirements for Distribution are to provide data products ready for network distribution within an average of 5 minutes of receipt of product order, and ready for physical media distribution within 24 hours of receipt of product order. Also, the capabilities for both network and media daily distribution rate must be equivalent to daily ingest rate.

DADS INTERFACE

Fig. 7 is the Conceptual DADS Context Diagram taken form the RFP Requirements Specification. This illustrates the multitude of data exchange interfaces for DADS data ingest and distribution. Some entities on this diagram not previously identified are the Affiliated Data Centers (ADCs), Other Data Centers (ODCs), Earth Probe Data Systems (EPDSs), Science Computing Facilities (SCFs), and International Partners (IPs).

DATA VOLUMES PER LEVEL

In Fig. 8, the total DADS daily data volumes, for the platforms EOS-A1, TRMM, and SAR are given for the data processing levels, L0, L1A, L1B, L2, L3, and L4. SAR is the EOS Synthetic Aperture Radar (SAR) platform, which is a separately funded option on the ECS contract. These daily data volumes are taken from the ECS Requirements Specification, Appendix C. As can be seen from this figure, the amount of data to be ingested, archived, managed, and distributed expands significantly from the amount of L0 data received from CDOS. For this set of platforms, the expansion factor is 3.6.

The total daily data contribution from TRMM is 18 GB/day, or 6.6 TB/year, which is small compared to EOS-A1, but large enough to fill 6 StorageTek Near-Line Library Units (Silos) per year, each Silo containing 6000 3480 type cartridges. Moreover, the total daily data contribution form EOS-A1 is 895 GB/day, or about 50 times the TRMM contribution. SAR adds 591 GB/day, or about 33 times the TRMM contribution.

DATA VOLUME PER DAAC

In Fig. 9, the total DADS daily data volumes, for the same 3 platforms, are given for each of the 7 DAACs. The 5 DAACs at JPL, LaRC, NSIDC, ASF, and MSFC vary in size from 3.5 to 8.3 GB/day, which is equivalent to 1.3 to 3.0 TB/year, or 20 to 45 TB in the 15 year EOS data collection period. Thus, these 5 DAACs could be called Tera-DAACs.

The other 2 DAACs at GSFC and EDC are roughly 2 orders of magnitude larger, and with EOS-A alone, each grows to a size of 2 to 3 PB over the 15 year EOS-A lifetime, thereby qualifying as Peta-DAACs.

It should be noted that the data volumes given in Figs. 8 and 9 do not include additional data volume required due to backup of hard-to-replace data products and due to reprocessing of selected data sets. However, this is partially offset by the fact that CDOS provides the disaster backup for L0 data, so that it is necessary for the DAACs to archive L0 data for only a year.

MEDIA REQUIREMENTS AT GSFC FOR EOS-A1

In Fig. 10, the daily data volume of 489 GB/day at the GSFC DAAC for the EOS-A1 platform is converted into a daily media requirement for several types of physical

media, ignoring utilization efficiency. For 3480 type cartridges containing 200 MB of data, this translates into 2445 cartridges per day, enough to fill a 6000 cartridge StorageTek silo every 2.5 days. With the newer 3490 cartridges, having double the data density, it takes 5 days to fill the silo. The D1 and D2 tape technologies reduce the daily number of cartridges required by about 2 orders of magnitude. It is anticipated that technological progress toward higher density data recording will continue over the next 7 years prior to EOS-A1 launch, resulting in a physically smaller and more manageable archive at that time than would be possible with current technology.

TRANSFER RATES

A potential bottleneck for timely EOSDIS DADS operations, is in the available data transfer rates for read/write devices compatible with available storage media. Data transfer rates available for drives compatible with the types of media considered in Fig. 10 are given in Fig. 11. With a transfer rate of 3.0 MBps, as is available for 3490 type cartridges, a single image file of 327 MB requires 109 seconds to physical read, again ignoring efficiency factors. Technological progress toward faster data transfer rates may be achieved prior to EOS-A1, but progress in this area has not been as rapid as in the area of higher density data recording.

EOS (EARTH OBSERVING SYSTEM) GOALS ARE TO DEVELOP AND OPERATE:

a) An observing system to acquire essential, global Earth science data on a long-term, sustained basis and in a manner which maximizes the scientific utility of the data and simplifies data analysis.

b) A comprehensive data and information system to provide the Earth science research community with easy, affordable, and reliable access to the full suite of Earth science data from EOS and international partner observatories, NASA Earth Probes, and selected Earth science data from other sources.

c) As the cornerstone of the Mission to Planet Earth Global Change Research Program, an integrated scientific research program to investigate processes in the Earth System and improve predictive models.

I.e. TO ACQUIRE, ACCESS, ANALYZE EARTH SCIENCE DATA

Fig 1

PHASED IMPLEMENTATION

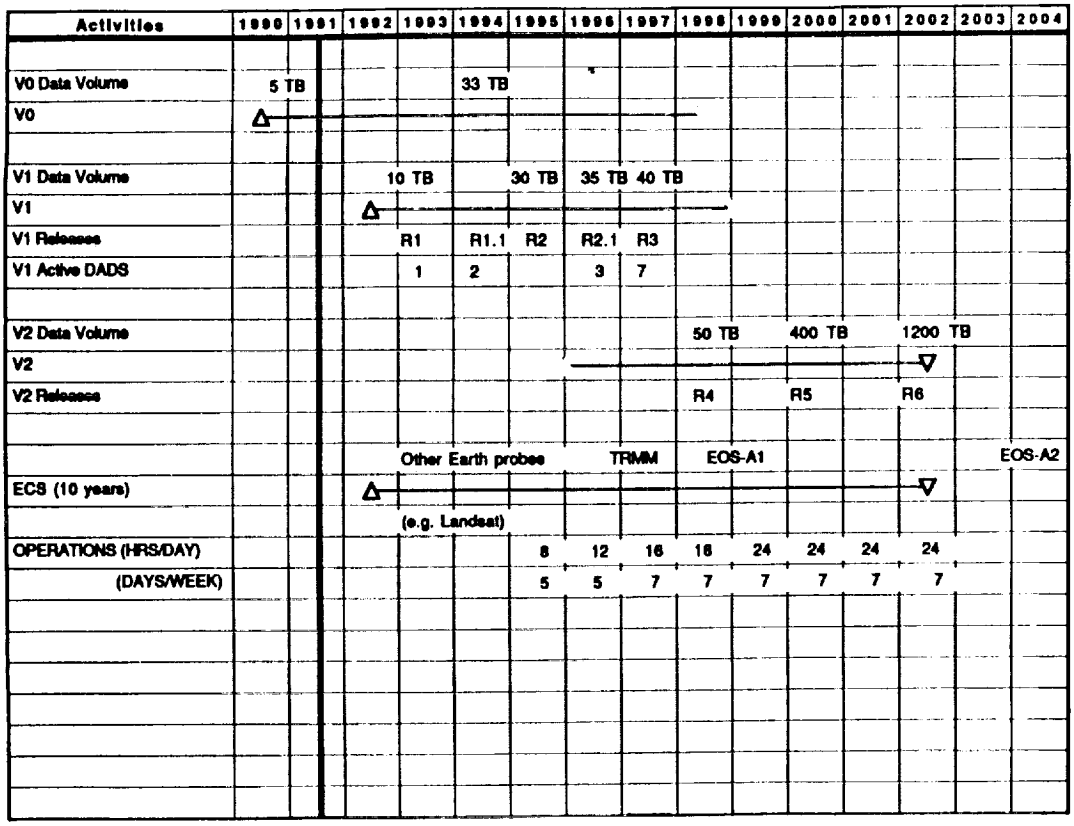


Fig 2

EOSDIS V0

V0 TO PROVIDE:

- **Interconnection of existing data systems at DAACs**
- **Prototyping of selected tasks in distributed IMS, networking, standards**
- **Some additional Earth Observation data sets to be added to the existing Data Systems under V0**

V0 RELATIONSHIP TO ECS

- **Provides early experience/information/results for potential inclusion in ECS design**
- **ECS contractor to connect ECS to V0 and provide a level of interoperability**
- **Selected data sets from V0 to be copied for inclusion into ECS**

Fig 3

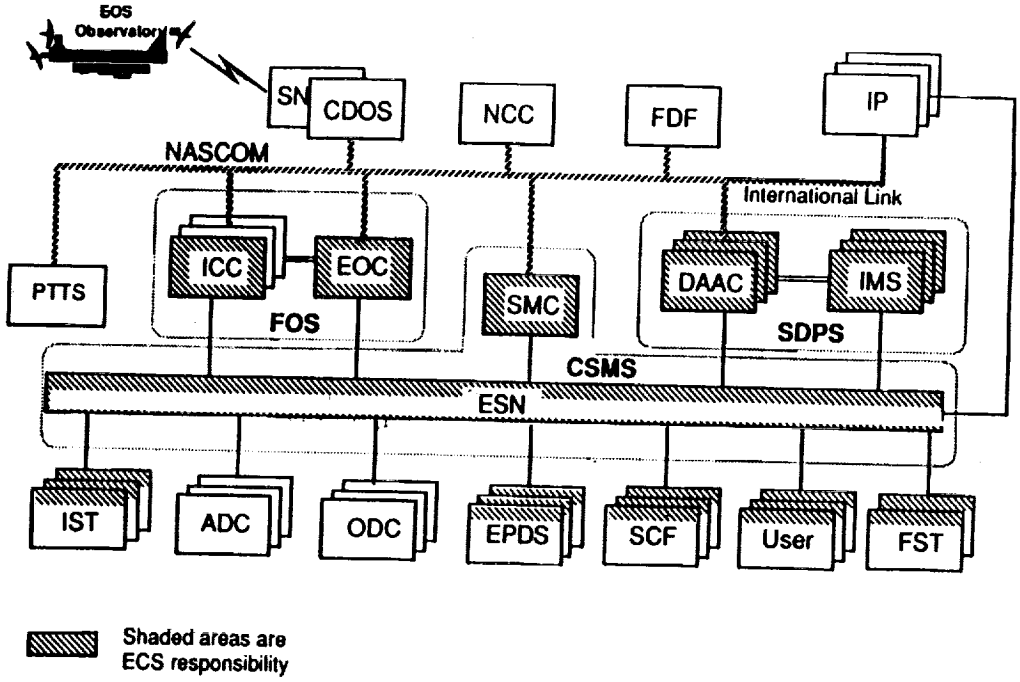


Figure 1.4.1 - 1. ECS Logical System Architecture

Fig 4

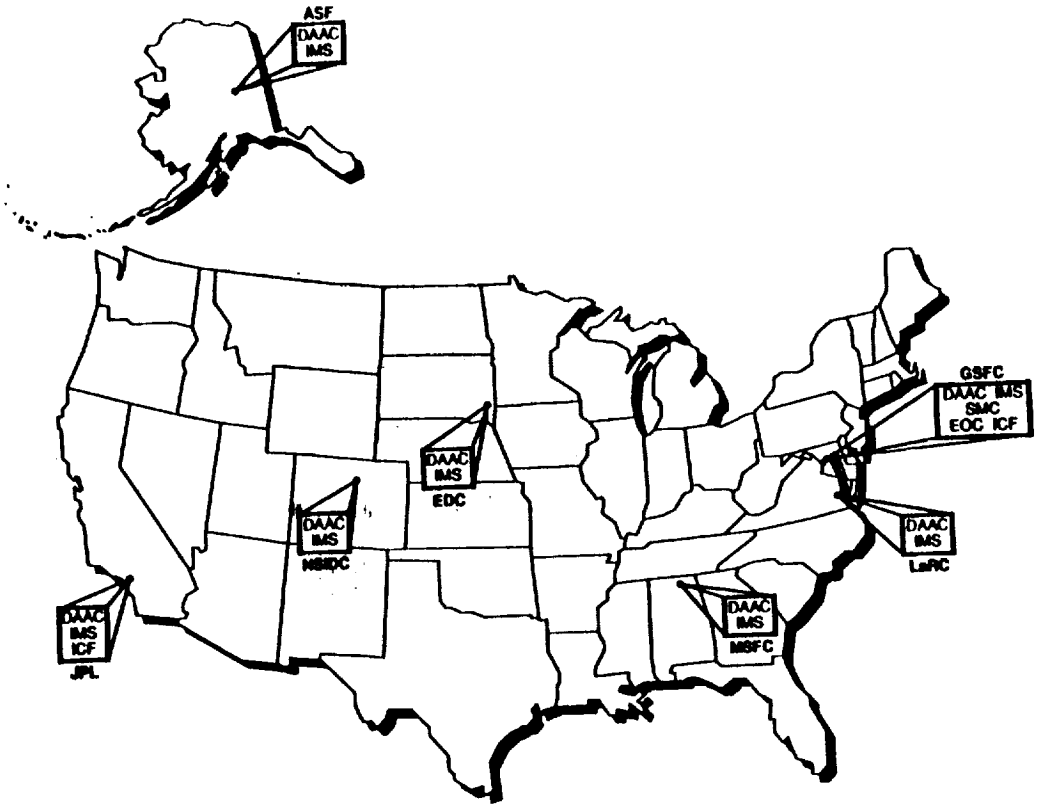


Figure 7.2.2-1 SDPS Physical Architecture

Fig 5

DADS FUNCTIONS

INGEST- Receive/Validate data products and data from CDOS , PGS, SCFs, other DAACs, ADCs, ODCs, EPDSs, IPs, Users, and others

ARCHIVE- Store data and data products on archive media

PROCESS ORDERS - Fulfill product orders provided by IMS, Retrieve data from archive, subset, reformat, stage for delivery. Support reprocessing

MANAGE SYSTEM - Monitor and report status and accounting information to SMC, operate File Storage Management System with hierarchical archive, schedule operations according to SMC directives, monitor media BER (10-12) and provide for data restoration/migration, backup selected data**

DISTRIBUTION - Distribute data and data products to PGS, SCFs, other DAACs, ADCs, ODCs, EPDSs, IPs, Users, and others via networks (5 minutes) and by Physical media (24 hours). Provide daily distribution rate capability equivalent to daily ingest rate

Fig 6

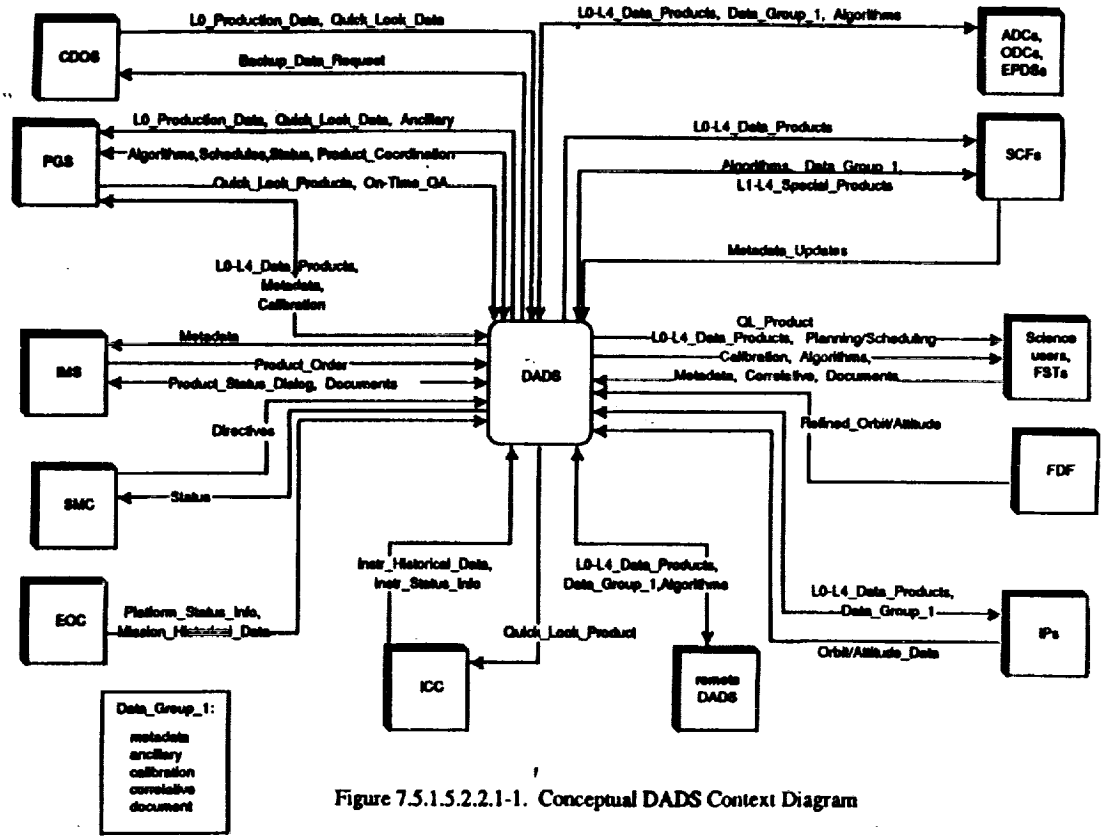


Figure 7.5.1.5.2.2.1-1. Conceptual DADS Context Diagram

Fig 7

DADS Daily Data Volume by Location (Log Scale)

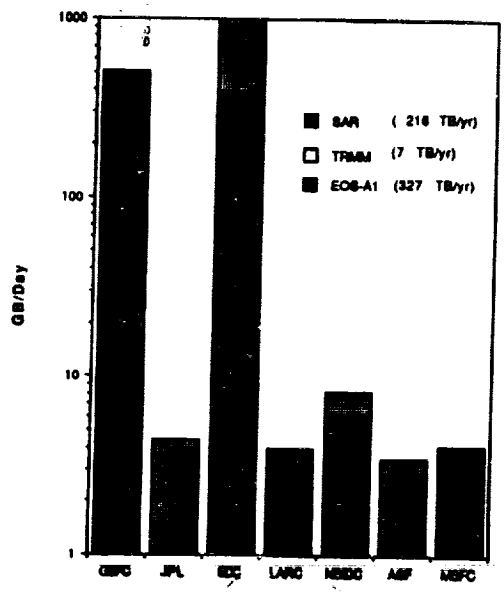


Fig 8

DADS Daily Data Volume by Location

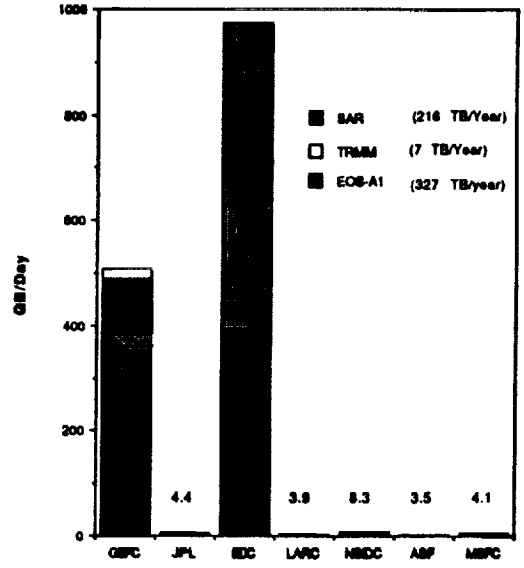


Fig 9

GSFC (EOS-A1) Daily Media Requirements

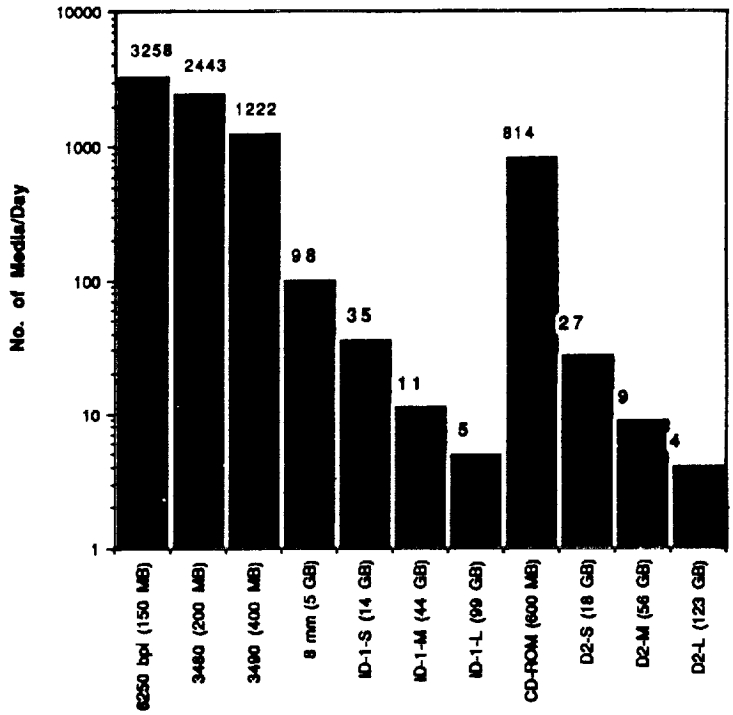


Fig 10

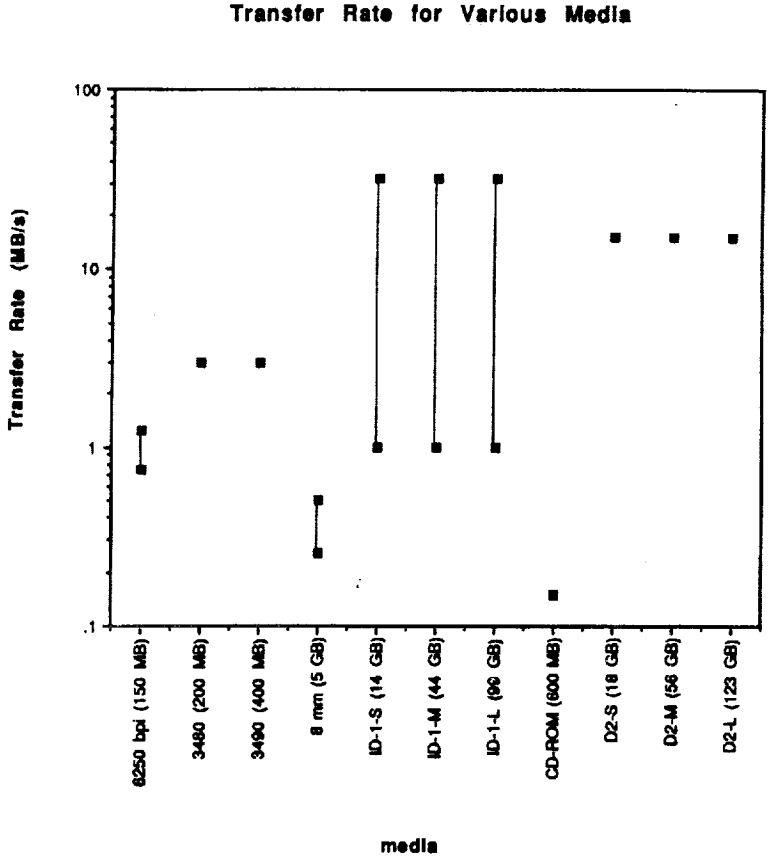

Transfer Rate for Various Media
Transfer Rate (MB/s)
100
10
1
.1
8250 bpi (150 MB)
3480 (200 MB)
3490 (400 MB)
8 mm (5 GB)
ID-1-S (14 GB)
ID-1-M (44 GB)
ID-1-L (96 GB)
CD-ROM (600 MB)
D2-S (18 GB)
D2-M (56 GB)
D2-L (123 GB)
media

Fig 11

The Preservation of Landsat Data by the National Land Remote Sensing Archive

John E. Boyd
U.S. Geological Survey
EROS Data Center
Sioux Falls, SD 57198

ABSTRACT

Digital data, acquired by the National Landsat Remote Sensing Program, document nearly two decades of global agricultural, environmental, and sociological change. The data have been widely applied and continue to be essential to a variety of geologic, hydrologic, agronomic, and strategic programs and studies by governmental, academic, and commercial researchers. Landsat data have been acquired by five observatories that use primarily two digital sensor systems. The Multispectral Scanner (MSS) has been onboard all five Landsats, which have orbited over 19 years; the higher resolution Thematic Mapper (TM) sensor has acquired data for the last 9 years on Landsats 4 and 5 only. The National Land Remote Sensing Archive preserves the 800,000 scenes, which total more than 60 terabytes of data, on master tapes that are steadily deteriorating. Data are stored at two locations (Sioux Falls, South Dakota and Landover, Maryland), in three archive formats.

The U.S. Geological Survey's EROS Data Center has initiated a project to consolidate and convert, over the next 4 years, two of the archive formats from antiquated instrumentation tape to rotary-recorded cassette magnetic tape. The third archive format, consisting of 300,000 scenes of MSS data acquired from 1972 through 1978, will not be converted because of budgetary constraints.

The data consolidation and conversion project will transcribe approximately 55,000 reels of high-density tape to 1,500 cassettes, ensuring that the data will be readable for the next 10 years. Some of these data, less than 10 percent, will not be reproducible because of the deterioration of the magnetic coating binder, the physical edge damage to the tape, or the demagnetization of the recorded data. The archive conversion activity will involve several computation and data manipulation tasks, in addition to rerecording the data. For example, to ensure that the archive inventory accurately specifies the location and quality of every retrievable scene, data will be spatially referenced after they are transcribed. Transcribed images will be assessed visually for cloud cover and data quality, and a numerical rating will be entered into a land information data base along with other catalog metadata, such as acquisition date and time, solar illumination angle, latitude and longitude, and sensor gain. A subset of each scene will be created during transcription by sampling every 64th element of the original data. The resulting archive of 500,000 browse images (325 gigabytes of data) will be made available to researchers through an interactive land information system.

This data preservation project augments EDC's experience in data archiving and information management, expertise that is critical to EDC's role as a Distributed Active Archive Center for the Earth Observing System, a new and much larger national earth science program.

LANDSAT DATA PRESERVATION

Conversion of Landsat Thematic Mapper and Multispectral Scanner Data in the National Satellite Land Remote Sensing Data Archive

Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications July 23-25, 1991

John E. Boyd
EROS Data Center
U.S. Geological Survey

EDSPO6 - 12 - 91
TMACS

USGS / EDC
TKO

EDC Land Information System Impacts

- Approximately 900,000-scene catalog of U.S.-held worldwide Landsat coverage
- Additional 2,000,000-scene catalog of international ground station TM/MSS data, compiled from contributions from 11 of 16 worldwide Landsat ground stations
- Catalog includes guide, inventory, and fairly extensive metadata
- Online digital browse data will be added to a Global Land Information System (GLIS) currently under development, as archive conversion operations progress

  ~350 GB of TM digital browse data (3 bands, ~1/64 subsampling)
  ~460 GB of MSS digital browse data (3 bands, ~1/36 subsampling)

- Currently, digital browse are planned to be archived in two optical disk jukeboxes; they will be available to IBM-compatible PC and Unix workstation inquiry and display

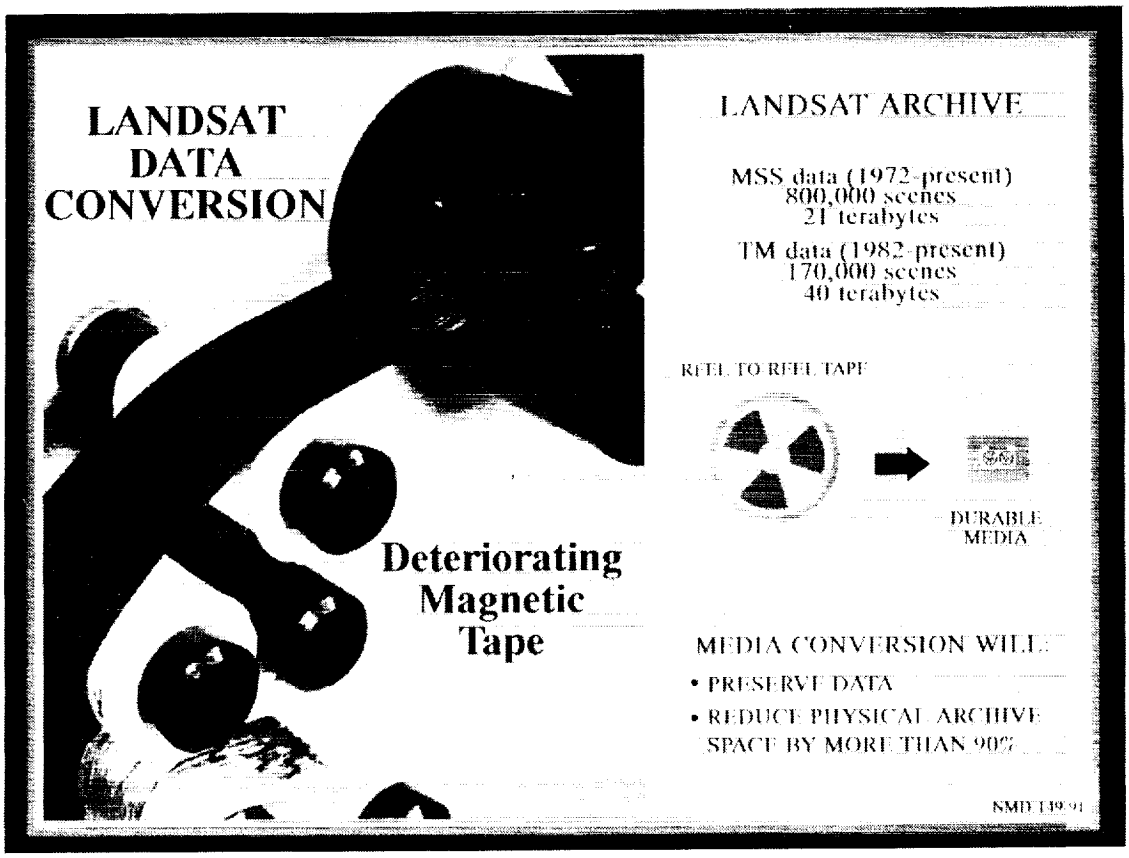
EDSPO6 - 12 - 91
TMACS

USGS/EDC
TIO

Center Pivot Irrigation, Saudi Arabia

Landsat MSS Image
December 25, 1972

Landsat MSS Image
February 15, 1986



Conversion Objectives

- **Transcribe data from high density tapes (HDTs), readable only by scarce obsolete longitudinal-track instrumentation recorders, to cassette-format rotary-head recorders**
- **Transcribe HDT data to new media before known physical degradation of master tapes results in significant (and intolerable) loss of irreplaceable scene data**
- **Maintain data integrity during transcription to 1 bit error per 100,000,000 bits recorded, by incorporating deeply interleaved Reed-Solomon error correction code**
- **Maintain long-term data integrity by using a fully enclosed cassette to protect the 1-inch-wide tape from edge scalloping, which too-frequently occurs on open-reel HDT**
- **Shrink the physical volume of the archive by a factor of approximately 100, as a result of the high media density and the stacking of multiple HDTs on one cassette**
- **Record 3 physical HDT formats (14- and 28-track) in a single rotary-head format**

Project Goals

- **Ensure availability of the 20-year Landsat data archive to:**
  - **— traditional Federal & State agency programs, researchers, and commercial users**
  - **— international global change research community**
  - **— Earth Observing Systems science investigators**
- **Maintain data integrity at the highest possible level through conservative implementation of recording technologies and meticulous archival procedures**
- **Commit to a program of periodic archive transcription to new media, at 5-7 year intervals, to avoid future technological obsolescence and data degradation**

Archives to be Converted or Transcribed

Data Type:	Acquired:	Volume:	Scenes:	Tapes:
MSS-WBVT	'72 - '78	9.5 TB	310,000	26,000
MSS-P	'79 - '80	2.8 TB	65,000	2,500
MSS-A	'81 - '91	7.2 TB	240,000	5,700
TM-A	'84 - '85	3.5 TB	15,000	1,400
TM-R	'82 - '91	40.0 TB	175,000	16,600
MSS-X-CCT	'72 - '78	1.3 TB	43,500	12,500
TM-P-CCT	'82 - '91	2.6 TB	8,700	35,000

Comparison of Storage Technologies

COMPARISON OF HIGH-DENSITY STORAGE TECHNOLOGIES						
Technology	(GB)	(¢ / MB)	Cost Imprv	Volume Impr	Archival	Exchange
CCT	0.15	5.0	Reference	Reference	Excel	Excel
3480	0.20	2.5	2 X	5 X	Excel	Excel
14-Tk HDT	2.3	7.0	0.7 X	8 X	Good	None
Optical Disk	6.0	5.0	1 X	25 X	Excel	Fair
DCRSi	48.0	0.3	17 X	300 X	Excel	Poor
Digital D2	75.0	0.1	50 X	475 X	Good??	Poor
Optical Tape	1,000.	0.7	7 X	2500 X	Good??	Poor

Capacity Comparison of Familiar Media

The following number of copies of the M/W Dictionary can be stored on these media:

- 2 on a standard 10.5-inch, 6250-bpi computer-compatible magnetic tape
- 9 on one 5.25-inch CD-ROM optical disk
- 35 on one 8-mm Exabyte-type cartridge magnetic tape
- 100 on one 12-inch double-density WORM optical disk
- 800 on one DCRSI-type or medium-sized DD-2 cassette magnetic tape
- (Merriam-Webster 9th New Collegiate Dictionary: 50-MB text + 10-MB illustrations)

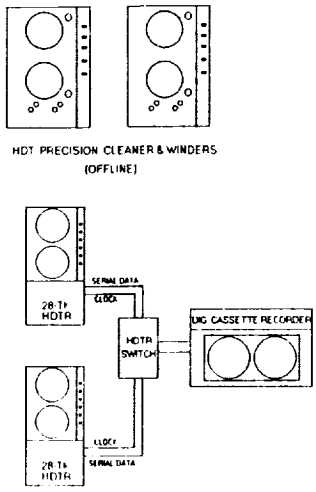
Desirable Characteristics of a Digital Cassette Recorder System

- High data density — 48 gigabytes per cassette
- High data transfer rate — 108 megabits per second
- Acceptable bit-error rate — 1 error in 10^{-8} bits recorded
- Low tape stress & quick rewind — 1,700 ft. rather than 9,200 ft. of tape
- Easy tape handling and no edge scalloping — strong, light polycarbonate cassette
- Archival quality tape formulation — 10-year-life minimum w/ gamma ferric oxide
- Incremental recording and playback — 8-MB I/O buffer permits variable-rate operation
- Rapid data identification and retrieval — IRIG timecode or data-block addressing

TMACS
TM/MSS ARCHIVE CONVERSION SYSTEM

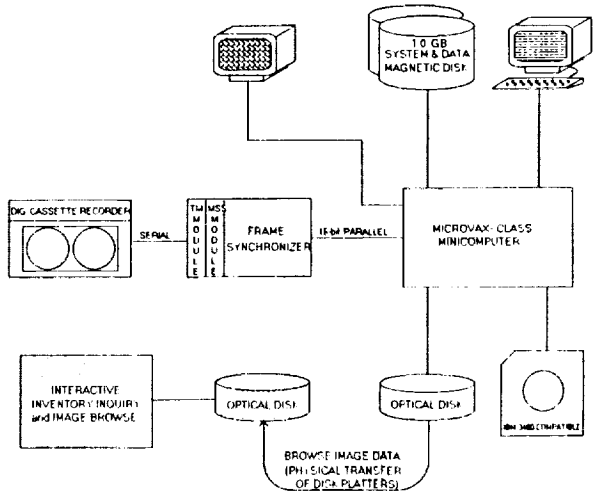
LANHAM HDT DATA TRANSCRIPTION

1) HDT ARCHIVE MANAGEMENT
2) HDT CLEANING AND PRECISION WINDING
3) DATA TRANSCRIPTION TO DIG CASSETTE
4) HDT DATA ERROR MONITORING & REWORK



EDC ARCHIVE / BROWSE GENERATION

1) SCENE FRAMING and DATA BASE CREATION
2) ANALYTICAL and VISUAL DATA VERIFICATION
3) SUBSAMPLED IMAGE BROWSE GENERATION
4) ONLINE USER INQUIRY TO OPTICAL DISK LIBRARY



Space Compression Resulting from Data Conversion

Sensor	Data Volume	Current No. Tapes	Converted No. Tapes	Current Sq. Ft.	Converted Sq. Ft.
TM	43,500 GB	18,000	1,050	~10,000	50
MSS	10,000 GB	8,200	250	1,200	15
MSS-WBVT	9,500 GB	26,000	225	~5,000	15
	63,000 GB	52,200	1,525	~16,200	80

- **Note: Conversion of MSS Wideband Video Tape (WBVT) currently is neither funded nor scheduled, but USGS vigorously has sought a supplemental FY92 appropriation.**

Archive Conversion — Rates and Volumes

- Assume:
- TM archive growth of ~25 scenes per day during the next 12 months
- Final TM archive size to be ~200,000 scenes contained on ~19,000 HDTs
- 250 workdays per year X 2 years required to complete the archive conversion
- HDT playback to be ½ realtime rate — 42.5 Mbit/sec, or 5.3 MB/sec

- Then:
- A production goal of 40 HDTs to be processed per 2-shift day
- ~420 scenes and 97,000 MB to be transcribed to DCRSi media, daily
- ~2.2 DCRSi cassettes to be recorded, daily
- HDT transcription requires: 18,300 seconds (5.0 hours) of capstan time, daily
- HDT precision clean and rewind requires: 10 hours daily, at 15 minutes per tape

Data Conversion Milestone Chart

TASK / ACTIVITY	FY91				FY92				FY93				FY94			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
ARCHIVE CONVERSION SYSTEM PROCUREMENT			CA													
SYSTEM DEVELOPMENT / TEST / INSTALLATION							I/T									
TM ARCHIVE CONVERSION OPERATIONS - LANHAM, MD							R-	—	—	—	—	—	-R	-A		
TM DATA QA & BROWSE GENERATION AT EDC							R-	—	—	—	—	—	-R	-A		
MSS ARCHIVE CONV., DATA QA, & BROWSE GEN. - EDC									A-	—	-A	-P				
1972 - 1978 MSS WBVT ARCHIVE CONVERSION ???																

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

by

Fernando L. Podio

National Institute of Standards and Technology
Computer Systems Laboratory

Abstract

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

by

Fernando L. Podio

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

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

NIST has developed a methodology for determining the life expectancy of optical disks. NIST is developing care and handling procedures for optical digital data disks and is involved in a program to investigate error reporting capabilities of optical disk drives.

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

Keywords: computer storage media, interchange standards; magnetic tapes, media interchange standards; magnetic tapes, Standard Reference Materials; optical disks, life expectancy; optical disks, media interchange standards; optical disks, error rate; test methods standards.

07/91

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

Fernando Podio
Project Leader for
Optical Storage Research
301/975-2947

National Institute of Standards and Technology
Computer Systems Laboratory

Scope of Talk

This overview is on the status of emerging interchange standards for computer peripheral storage technologies which utilize removable optical disks and magnetic tapes

Computer Storage Media Interchange Standards

Media - optical, magnetic, mechanical, etc.

Physical format - track location, data code, etc.

Logical format - volume labels, files structures, etc.

Test Methods Standards

Why do we need standard testing methods for media characteristics?

When media interchange standards are adopted, standard test methods for testing media characteristics are needed for conformance testing.

Data Permanence Standards

A general data permanence testing methodology does not exist

There are no standards regarding longevity that can assist managers in planning how long the information may be stored on removable computer storage media.

NIST Participation in Standards Committees for Removable Computer Storage Media

* Technical Committee (TC) X3B11 - Optical Digital Data Disks
  Member and Chair, Test Methods Project Group
* Joint Technical Commission on Data Permanence
  Member and Technical Liaison with TC X3B11
* NIST/NASA Working Group (Test Methods and Specifications for 356 mm Ruggedized Rewritable Optical Disk Media)
  Chair
* Technical Committee (TC) X3B5 - Digital Magnetic Tapes
  Member and Supply Standard Reference Materials
* ISO/IEC JTC1/SC23 Optical Digital Data Disks
* ISO/IEC JTC1/SC11 Digital Magnetic Tapes

Technical Committee X3B11
Optical Digital Data Disks

* Media Interchange Standards
* Test Methods Standards
* Label and File Structure Standards (X3B11.1)
* 46 principal members
* Meetings every two months
* Work coordinated with ISO/IEC JTC 1/SC23

Status of X3B11 Projects

* **356 mm (14 inch) WORM**

  Letter ballot passed TC X3B11.

  TC X3B11 resolving comments on first Committee Document 10885 (CD 10885).

  Next step is DIS (Draft International Standard) status

Status of X3B11 Projects

* **300 mm (12 inch) WORM**

Next generation standard.

There are two proposed incompatible formats:

CCS (Continuous composite servo).
ECMA/TC31/TG300/91/8 (first draft 300 mmm, CCS, 05/91).

SSF (Sampled servo). LMSI proposed the MASS format.

TC X3B11 working in ISO towards an ISO standard (the ISO document will be used as the ANSI standard).

Status of X3B11 Projects

* **130 mm (5 1/4 inch) WORM**

3 draft standards approved by X3B11. One new project using MO media:

- Continuous Composite Servo (CCS).	ISO/IEC 9171 - A X3B11/90-125
- Sampled Servo 4/15 Modulation (SSF).	ISO/IEC 9171 - B
- Sampled Servo High Capacity.	ANSI X3.191-1991
- CCS format using MO media, (CCW).	New approved project

Status of X3B11 Projects

* **130 mm (5 1/4 inch) Rewritable (MO media)**

One draft standard

- CCS format — ISO/IEC 10089 - A
- The ANSI document will be aligned with ISO/IEC 10089 - A — X3B11/90-165

Status of X3B11 Projects

* **86 mm (3 1/2 inch) Rewritable CCS**
  - CCS similar to 130 mm Rewritable
  - The document was aligned with ECMA/TC31/91/32.
  - The ISO document will be used as the ANSI standard (ISO CD 10090). Expected release of the ISO standard: June 1992.

* **86 mm (3 1/2 inch) Rewritable DBF**
  - X3B11/91-050.

TC X3B11 Test Methods Projects

Size	Status
356 mm (14 inch) WORM	Approved as an ANSI Standard: X3.199-1991
300 mm (12 inch) WORM	Call for editors
130 mm (5 1/4 inch) WORM - CCS	Document in preparation.
130 mm (5 1/4 inch) WORM - SSF	Document in preparation.
130 mm (5 1/4 inch), MO, CCS	X3B11/91-032 will go for public review.
90 mm (3 1/2 inch), MO, CCS	A new project will be requested shortly.
90 mm (3 1/2 inch). MO, DBF	New project. A Base Working Document is expected by August.

Test Methods for Media Characteristics of
356 mm Optical Disk Cartridge - Write-Once

ANSI Standard: X3.199-1991

Imbalance
Apparent Axial Runout
Residual Focus Error
Drop Test
Dead Weight Strength
Double Pass Retardation
Signal Characteristics
Optical Disk Write Power
Read Power
Narrow-Band Signal-to-Noise Ratio
Cross-Talk
Radial Runout
Residual Tracking Error Signal

Proposed American National Standards

Test Methods for Media Characteristics of 90 mm and 130 mm Rewritable Optical Disk Data Storage Cartridges with Continuous Composite Servo (CCS)

Moment of Inertia
Imbalance
Axial Runout and Acceleration
Radial Runout and Acceleration
Drop Test
Refractive Index
Thickness
Baseline Reflectance and
Reflectance Uniformity
Resolution (only for 130 mm)
Signal Imbalance

Erase Characteristics
Write Characteristics
Read Characteristics
Figure of Merit
Narrow-Band Signal-to-Noise Ratio
Cross-Talk
Prerecorded Characteristics
PEP Cross Track Loss

Approved ISO/IEC Standards for Optical Digital Data Disks

*	ISO 9660	Volume and File Structure of CD-ROM
*	ISO/IEC 10149	120 mm (CD-ROM)
*	ISO/IEC 9171, Part A	130 mm WORM CCS
*	ISO/IEC 9171, Part B	130 mm WORM SS 4/15
*	ISO/IEC 10089, Part A	130 mm RWRT CCS
*	ISO/IEC 10089, Part B	130 mm RWRT SSF, 4/15 Modulation

Data Permanence Standards

Joint Technical Commission on Data Permanence
IT9-5 (ANSI) and Subcommittee 84 (AES)

Task Groups

TG I	Definitions
TG II	Storage and Handling
TG III	Transfer
TG IV	Optical Systems and Media
TG V	Magnetic Systems and Media

TC X3B11 Technical Liaison : F. Podio

Joint Technical Commision on Data Permanence Task Group IV Optical Systems and Media

* Standards, test methods, recommended practices and specifications pertaining to the life expectancy and retrieval of information recorded on optical systems.
* Currently developing a standard for life expectancy of CD's.
* Call for editors for similar documents for other types of optical media.
* Next meeting: October 21 and 22, 1991 - Tucson, Arizona
* Technical discussions on:
  - Quality parameters and prediction methods

NIST/NASA Working Group

Scope of Tasks

Development of a set of Test Methods and Specifications for 356 mm (14 inch) Ruggedized Rewritable Media (sponsored by NASA, Langley Research Center).

Participation

Government and Industry.

Status

NIST Special Publication SP500 - 191: "Test Methods for Optical Disk Media Characteristics (for 356 mm Ruggedized Magneto-optic Media".

NIST/NASA Working Group

Test Methods for Media Characteristics of 356 mm Ruggedized Rewritable Media

Operational Environment Test
Non-operational Environment Test
Storage Environment Test
Environmental Qualification
Mechanical and Physical Characteristics
Substrate Characteristics
Recording Layer Characteristics
Preformat Characteristics
Media Lifetime

Technical Committee X3B5
Digital Magnetic Tape

* Media Interchange Standards
* Label and File Structure Standards
* About 50 principal members
* Meetings every three months
* Work coordinated with ISO/IEC JTC 1/SC11

Status of X3B5 Projects

12.65 mm (0.5 in) Wide Open Reel Magnetic Tape

Category	ANSI	ISO	ECMA
Logical Format (Label and File Structures)	X3.27-1987	1001:1986	13-1985
Media Standard (Unrecorded)	X3.40-1983	1864:1985 (CD 1864)	62-1985

Status of X3B5 Projects

12.65 mm (0.5 in) Wide Open Reel Magnetic Tape

Physical Format Standards

	ANSI	ISO	ECMA
200 cpi	X3.14-1983	1862:1975 (R 1986)	-
800 cpi	X3.22-1983	1963:1990	62-1985
1600 cpi	X3.39-1986	3788:1990	62-1985
3200 cpi	X3.157-1987	-	-
6250 cpi	X3.54-1986	5652:1984	62-1985

Status of X3B5 Projects

12.65 mm (0.5 in) Wide, 1491 CPMM (37871 cpi) Magnetic Tape Cartridge "3480 Technology"

Category	ANSI	ISO	ECMA
Logical Format (Label and File Structure)	X3.27-1987	1001:1986	13-1985
Media (Unrecorded)	X3.180-1990	9661:1988	120-1987

Status of X3B5 Projects

12.65 mm (0.5 in) Wide, 1491 cpmm (37871 cpi), Magnetic Tape Cartridge "3480/3490/3490E Technology"

Physical Format Standards

	ANSI	ISO	ECMA
GCR, 18 tracks, 200 MBytes	X3.180-1990	9661:1988	120-1987
GCR, extended format, 18 tracks	X3B5/90-342 (3rd draft)	-	TC17/91/6 (2nd draft)
GCR, 36 tracks, parallel serpentine	-	-	-

Status of X3B5 Projects

8 mm (0.315 in) Wide, Helical Scanned Magnetic Tape Cartridge

Category	ANSI	ISO	ECMA
Physical format	X3.202-199X (dpANS)	DIS 11319	145-1990
Media (unrecorded)	X3.202-199X (dpANS)	DIS 11319	145-1990

Status of X3B5 Projects

3.81 mm (0.150 in) Wide, Helical Scanned Magnetic Tape Cartridge

Physical Format Standards	ANSI	ISO	ECMA
DDS-DC Format	X3B5/90-323 (1st draft)	-	TC17/91/5
DDS Format	X3.203-199X (dpANS)	DIS 10777	139-1990
DATA DAT Format	X3.205-199X (dpANS)	DIS 11321	146-1990
Media Standard (Unrecorded)	X3.201-199X (dpANS)	DIS 11321	146-1990 TC17/91/5

Status of X3B5 Projects

19 mm (0.748 in) Wide, Helical Scanned Magnetic Tape Cartridge

Category	ANSI	ISO	ECMA
1st pdp, Unrecorded	X3B5/91-178 (1st draft)	-	-

Status of X3B5 Projects

New Projects

*	Media and physical standard for 12.65 mm (0.5 in) Helical Scanned Cartridge, 20 GBytes capacity, 3480 form factor	Approved
*	Media standard for D-2, 19 mm Helical Scanned Cartridge (Metal T-film)	-
*	Media standard for DD-2, 19 mm Helical Scanned Cartridge (Metal T-film)	-
*	Physical format standard for 12.65 mm (0.5 in), 36 track, parallel serpentine (400 MBytes capacity). "3490E technology" (CrO_2)	-

Approved ISO/IEC Standards for Magnetic Tapes

*	ISO/IEC 1864	0.5 in Open Reel Magnetic Tape, (Unrecorded)
*	ISO/IEC 1862	0.5 in Open Reel Magnetic Tape, 200 cpi, (Physical Format)
*	ISO/IEC 1863	0.5 in Open Reel Magnetic Tape, 800 cpi, (Physical Format)
*	ISO/IEC 3788	0.5 in Open Reel Magnetic Tape, 1600 cpi (Physical Format)

Approved ISO/IEC Standards for Magnetic Tapes

*	ISO/IEC 5652	0.5 in Open Reel Magnetic Tape, 6250 cpi (Physical Format)
*	ISO/IEC 9661	0.5 Magnetic Tape Cartridge, "3480 Technology", (Unrecorded and GRC, 18 Tracks, Physical Format)
*	ISO/IEC 1001	0.5 Magnetic Tape Cartridge, "3480 Technology", (Label and File Structure)

Related Contributions of NIST

* Active Participation in Standards Committes
* Optical Media Research Program
* Digital Magnetic Tape Program

NIST Optical Media Research Program

Standards

Data Interchange (Participation in TC X3B11 and ISO/SC23)

Test Methods (Leadership in TC X3B11 Test Methods Project Group)

Data Permanence (Technical Liaison between JTC and TC X3B11)

NIST/NASA Working Group (Chair and technical editor)

NIST Optical Media Research Program

Research and Development

Methodology to determine the life expectancy of optical media.

Care and handling of optical media.

Test methods for media characteristics.

Program for investigating error reporting capabilities of optical disk drives.

Optical Disk Laboratory

- Optical disk drives
- Electronic Instrumentation
- Automatic measuring systems for media characteristics
- Optical inspection systems
- T/H chambers and ovens with computer control
- Access to scientific computers (modelling)

Aging Tests

Aging tests were run on 300 mm (12 inch) WORM media.

Over 5000 hours of testing

Three groups of disks:

* 80°C, 90% RH
* 70°C, 90% RH
* 60°C, 90% RH

NIST Testing Methodology for Determining the Life Expectancy of Optical Disk Media

Some conclusions:

Life expectancy values strongly depend on several factors:

* the test methods for the quality parameter(s)
* the aging methodology
* data patterns written
* the areas measured
* amount of data tested
* statistical methodology (i.e. number of disks, number of samples, etc)
* End-of-life definition

A complete specification of these conditions is essential in any comparison of life expectancy values

Program for Investigating Error Reporting Capabilities of Optical Disk Drives

* Workshop held August 5th, 1991 in Colorado Springs, CO
* Creation of an Government/industry Working Group
* Demonstration platform

Optical Disk Drives Error Reporting Considerations

* Most optical disk systems do not report sufficient information to the host on error rates, error distributions, etc.

* Error rate information would include parameters such as the total number of correctable errors, maximum number of errors per interleave, and the location of these errors.

* This information would provide data managers with a better understanding of the status of their data.

* Monitoring the status of data recorded on optical media is particularly important because of its anticipated use for the long-term storage of valuable data.

NIST Government/Industry Working Group

* Technical representatives from Government and industry.

* Document standardized method(s) of reporting error rate information (NIST guideline/standard).

* Contributions to relevant standards committees.

* Meetings concurrently with TC X3B11 (Optical Digital Data Disks).

* First meeting: October 7, 1991, Phoenix, AZ.

* Target date for NIST publication: end of FY92.

Demonstration Platform

* Demonstration of the state of the art in error rate monitoring and reporting techniques as well as the interface capabilities to U.S. Government managers.

* Two hosts:

  SUN SPARCstation IPC
  Compaq 386/25e

* Optical disk drives.

* Adapt the demonstration platform to accommodate next generation drives with enhanced error reporting methodologies.

NIST Digital Magnetic Tape Programs

World source for digital magnetic tape Standard Reference Materials (SRMs)

Appraisal of potential risks associated with storing data on 12,65 mm (0.5 inch) CrO_2 "3480 technology"

* Literature search

* Consultation with users and manufacturers

* Development of reasonable measures for the proper use of the 3480 tape cartridge

Federal agencies technical support for the care and handling of digital magnetic tapes

Standard Reference Materials for Digital Magnetic Tapes (SRMs)

* SRM are materials that NIST produces that are calibrated against a generally accepted Master Standard

* For magnetic tapes, some properties cannot be directly specified in standards (e.g., output signal amplitude is dependent on head used in drive)

* Therefore reliable interchange of data requires that the tapes be designed and manufactured on the basis of a comparison to a known and accepted SRM. Currently, NIST is the world source for these SRMs.

Standard Reference Materials for Digital Magnetic Tapes (SRMs)

Parameters measured by NIST

* Output signal amplitude

* Typical current

* Resolution

* Overwrite

* Peak shift

SRM Tapes Provided by NIST

SRM 1600	3.8 mm, 63 ftpmm cassette
SRM 3216	6.3 mm, 126 ftpmm cartridge
SRM 3217	6.3 mm, 126 ftpmm cartridge
SRM 3200	12.65 mm, 8/32/126 ftpmm open reel
SRM 6250	12.65 mm, 356 ftpmm open reel
SRM 3201	12.65 mm, 262/394 ftpmm cartridge
Proposed SRM 3202	12.65 mm, 972 ftpmm cartridge
Under development	6.3 mm, 394 ftpmm cartridge
Under development	6.3 mm, 492 ftpmm cartridge

DATA MANAGEMENT IN NOAA

William M. Callicott

July 25, 1991

ABSTRACT

NOAA has 11 terabytes of digital data stored on 240,000 computer tapes. There are an additional 100 terabytes (TB) of geostationary satellite data stored in digital form on specially configured SONY U-Matic video tapes at the University of Wisconsin. There are over 90,000,000 non-digital form records in manuscript, film, printed and chart form which are not easily accessible. The three NOAA Data Centers service 6,000 requests per year and publish 5,000 bulletins which are distributed to 40,000 subscribers. Seventeen CD-ROMs have been produced. Thirty thousand computer tapes containing polar satellite data are being copied to 12 inch WORM optical disks for research applications. The present annual data accumulation rate of 10 TB will grow to 30 TB in 1994 and to 100 TB by the year 2000. The present storage and distribution technologies with their attendant support systems will be overwhelmed by these increases if not improved. Increased user sophistication coupled with more precise measurement technologies will demand better quality control mechanisms, especially for those data maintained in an indefinite archive. There is optimism that the future will offer improved media technologies to accommodate the volumes of data. With the advanced technologies, storage and performance monitoring tools will be pivotal to the successful long-term management of data and information.

TABLE OF CONTENTS

1.0. Data Management in NOAA

Management of environmental data and information resources is becoming an increasingly visible and important issue for the scientific community. This is of particular importance to the National Oceanic and Atmospheric Administration (NOAA), which routinely measures and collects large amounts of environmental data and information in its own work, and is also officially charged with maintaining environmental records for the Nation. Through its activities over time, NOAA has become the steward of a treasury of Earth systems data and information--the most comprehensive, long-term, and up-to-date environmental description of the Earth that exists today. This treasury contains answers to urgent environmental questions facing the Nation. The success of all NOAA's scientific work, and the national priorities which it supports, depends on the accountability and accessibility of environmental data and information. These data and information must be accurate, complete, stable and fully-integrated across the spectrum of NOAA's organizational functions, and they must be made easily accessible, in a timely and cost-efficient way. Throughout this effort to provide resources to meet the NOAA missions of managing data for global change research and for enhancing warning and forecast services, there will be a continual inherent process to migrate and protect data held in the NOAA archives.

2.0. NOAA Data Management Operations:

The NOAA Centers respond to requests from a broad community of research, legal, engineering, individuals, insurance, business, consultants, and manufacturing. About 6,000 requests per year are received for digital data. Within the next two years as global change research increases, this should grow to 9,000. The current data files stored on computer readable media have a volume of 11 terabytes stored primarily on 240,000 computer tapes. There are 100 TB of GOES data at the University of Wisconsin copied in digital form on specially recorded SONY U-matic video tapes. There are analog, i.e., non-digital, holdings having a volume equivalent to over 50 TB in a digital domain. With the conversion of some of the analog data to digital format, and with new sources of digital data, the digital holdings will grow by 1996 to about 35 TB not counting source level data from GOES. The rapidly expanding quantity of data will force a different approach to managing data and information within the centers. The use of an integrated mass storage systems with appropriate file management will be essential to manage the archive, storage hierarchy and data migration processes. New mass storage hardware and software technologies will have to be developed and adopted and the current archives copied. If a mass storage system is not developed, the data centers will require immense tape storage areas and reduce the access mechanism from data granules to physical volumes of data.

In accordance with the NASA/NOAA Memorandum of Understanding for remotely sensed earth observations, data processing, distribution, archiving and related science support, EOS data is intended to be archived at the NOAA data centers. The EOS

data from prototype operational instruments used for NOAA operational purposes, will be an inherent part of NOAA's data archives. As part of the EOS pathfinder activity, NOAA is migrating the environmental satellite data from the operational polar orbiting satellites from computer tapes to 12 inch SONY optical write once, read many (WORM) disks. Also, there are selected special sensor microwave data from the Air Force Defense Meteorological Satellite Program (DMSP) of interest to EOS scientists. Initially, 8 terabytes of data will be migrated to optical disk.

The computing capacity at the centers is provided by mainframes, workstations, and personal computers (Pcs). In the aggregate, the systems do not have the capacity to handle the anticipated archive, the growth in data ingest and dissemination, and expanded quality control, analysis and reprocessing requirements in the coming years. The computational capacity will need to grow from 10 MFLOPS at the beginning of 1991 to over 300 MFLOPS by 1996, i.e., a factor of 30:1. On-line disk storage should grow from 60 GB to over 500 GB during this period. The configurations are evolving from a central processor surrounded by dedicated terminals to a fully distributed client-server architecture which can expand in response to workload demands.

2.0. NOAA Data Centers: There are three National Data Centers and over a dozen centers of data in NOAA. The data centers are structured as formal archive centers and serve at this Nation's world data centers for their respective disciplines. The following provides a description of the centers and their activities.

2.1. The National Climate Data Center (NCDC) in Asheville, North Carolina was established in 1950. The National Archives and Records Service, in compliance with the Federal Records Act of 1950, specified that NCDC's climatological records be permanently retained. It has been designated as a World Data Center (WDC)-A for Meteorology. It also operates the Satellite Data Service Division which manages the high volume satellite data. The Center is responsible for ingest, quality control, archiving, managing, providing user access, and performing analysis of data which describes the global climate system. The NCDC also supports major new programs such as the National Weather Service Modernization, Climate and Global Change, the Coast Watch Initiative, and the Level-0 Earth Observing Systems (EOS) path finder effort. In 1992, new data sources from foreign satellites will be introduced. The center has a staff of 290 full time employees (FTEs), 100 contract and 190 federal employees. Each working day results in 160 orders from 360 user contacts. Annually, about 5,000 bulletins are prepared and supplied to 40,000 subscribers. Much of NCDC's data (by physical volume) is in the form of paper records such as ship logs, and although manually accessible, is not readily usable, and they appear to be deteriorating. There are fifty thousand cubic feet of paper records at NCDC. There are also film and other non-machine readable information stored in the National Archives.

NOAA has a contract with the University of Wisconsin to collect and archive data from the NOAA geostationary operational satellites (GOES). The GOES data collected from

1978 to the present, is recorded and stored at the University of Wisconsin's Space Science and Engineering Center. To date, the GOES data are stored on 25,000 Sony U-matic beta video (19mm commercial video standard) video cassettes in 4 GB increments. Access to the data is not highly efficient because some of the cartridges are offsite in the state records office, and because the data must be reingested as a satellite readout. The GOES archive represents the largest amount of data anywhere in the NOAA system to be rescued and be made more readily accessible. To improve access, the center is engaged in a pathfinder study on mechanisms to improve the access to the data.

2.2. The National Geophysical Data Center (NGDC) in Boulder, Colorado was established in 1972. Its mission is to manage solid earth and marine geophysical data as well as ionospheric, solar and other space environmental data; and to provide facilities for World Data Center-A for Geophysics which encompasses Solid Earth Geophysics, Solar-Terrestrial Physics, Marine Geology and Geophysics, and Glaciology. The center has a staff of 60 FTEs.

NGDC has over 300 databases including 54 million (M) Ionograms, 2.5M magnetograms, 12 million flight miles of aeromagnetic data and 10 million miles of ship track data. There are 25,000 magnetic tapes partly at NGDC in Boulder and partly in Asheville at NCDC. About 2000 tapes per year arrive from originators outside of NGDC. Their goal is to keep all of the ingest tapes and maintain two additional copies for use in normal center activities (3 copies total). About 14,000 tapes have no backup. There is a requirement to copy 12,000 tapes a year for routine migration. NGDC relies on the error checking provided by the tape copying system and supplements this with printouts of beginning/end of record data and record counts.

NGDC began the NOAA CD-ROM program four years ago. Its first CD-ROM title was "Geophysics of North America". The center now has a total of 12 CD-ROMs completed or underway. This effort should change the distribution system for data from its tape orientation and probably deflect some use of the network to obtain data. During 1990, distribution of data by CD-ROMs has increased both the number of requests for digital data and the annual amount of data distributed by the center. The distribution of data by CD-ROM puts the data into a form highly convenient and useful to Pcs and workstations.

The National Snow and Ice Data Center (NSIDC) under contract to NGDC operates the World Data Center-A for Glaciology. The role of NSIDC is to acquire, archive and disseminate data relating to all forms of snow and ice. It provides data to about 500 requesters per year from a digital archive data base of about 15 GB (300 standard tapes); 7GB are from the NIMBUS-7 Scanning Multi-channel Microwave Radiometer, and 3 GB are Special Sensor Microwave/Image (SSM/I) data. Many of the NSIDC datasets are redundantly held at other NOAA data centers. A daily volume of 1 GB of data from the Defense Meteorological Satellite Program (DMSP) Operational Linescan

System (managed by NGDC) will be forwarded from the DMSP readout site to NSDIC on EXABYTE tape. A similar means is being developed to distribute a weekly data volume of 800MB between the Joint Ice Center (JIC), in Suitland and the National Snow and Ice Data Center (NSIDC) in Boulder. The NSIDC has been designated as an EOSDIS Distributed Active Archive Center (DAAC). As such, it will build up its computational and archival ability to meet EOSDIS defined mission goals. As a DAAC, the NSIDC will relocate to the University of Colorado campus.

2.3. The National Oceanographic Data Center (NODC) has been in operation since 1961 as an interagency, facility under the U.S. Navy, and became a part of NOAA in 1970. Its mission is to manage oceanographic data. It has operated as a World Data Center-A (WDC-A) for Oceanography since 1962. NODC has a staff of 85 FTEs. NODC's files include data collected by U.S. federal agencies; state and local government agencies; universities and research institutions; foreign government agencies and institutions; and private industry. Currently, NODC maintains a digital archive of both in-situ and satellite-sensed ocean data in excess of 30 GB. A potential equivalent of 10 GB of digital data are currently maintained in analog form such as data reports, manuscripts, and analog instrument recordings. With the establishment of NODC data management responsibilities for ocean observing satellites, including non-NOAA geostationary and orbiting platforms; new global collection efforts, including the Global Ocean Flux Study (GOFS), the World Ocean Climate Experiment (WOCE), and the Climate and Global Change Project; and new U. S. coastal ocean studies, including CoastWatch and the Coastal Ocean Program, the archive is expected to increase twenty fold between FY90 and FY95.

NODC has a large amount of analog data. A tablet digitizer is used to annually process 10,000 expendable bathythermograph traces (XBT) to 2 MB of data. NODC has a contract with the Navy to annually process 100,000 similar traces. Mechanical bathythermographs (MBT) on glass slides (300,000) await conversion and are expected to result in 1.5 GB of data. Acoustic Doppler Profiling, done from University and NOAA ships, is expected to be a new source of data. There are perhaps 20 ships that may be equipped with these profilers. At the present, only a few are capturing the data for archiving purposes. Each profiler should provide 10 MB per ship month. In the future, 100 ship months per year of these data could be archived.

The NOAA Coastwatch Data Management, Archive, and Access System (NCAAS) now under development will result in expansion of the archive, related quality control, and retrieval and distribution activities based on SONY 12 inch WORM Optical Disks. NODC also is responsible for the NOAA Library and there is interest in digitizing some of its data and metadata holdings.

NODC is responsible for the NOAA Earth System Data Directory, and interfacing it with the larger NASA based master directory effort. This is part of the catalog interoperability effort which is underway among other government agencies and

foreign data centers. The master directory is also needed to fit with the EOSDIS version 0 effort where the catalog interoperability will perform relevant IMS (Information Management System) functions. NOAA's master directory is VAX based, with specially written software, and the ORACLE Data Base Management System.

A prototype database system has been developed to provide NODC users with direct access to an on-line, interactive data archive. It maintains a data base of over 23 million marine observations from 310,000 ocean stations. Access to the data is obtainable through spatial or temporal searches with arbitrary combinations of instruments, platforms, and parameters. By FY 1993, NODC plans to add all of its vertical profile data (Nansen, Bathythermograph, C/STD, etc.) to the POSEIDON system.

2.4. NOAA Centers of Data include those data collection and operations elements performing observations and monitoring services as part of NOAA's recurring mission responsibilities. Listed below are the principal centers:

Discipline	Title	Location
Bathymetry, Nautical charts, Geodesy	Charting and Geodetic Services	Rockville, Maryland
Climate	Climate Analysis Center	Camp Springs, Maryland
Fisheries	National Marine Fisheries Service	Seattle, WA; Woods Hole, MA; Miami,FL; Bay St. Louis,MS; San Diego, CA
Ice	Navy Joint Ice Center	Suitland, Maryland
Lake Data	Great Lakes Environmental Research Lab	Ann Arbor, MI
Oceanography	Center for Ocean Analysis and Prediction	Monterey, CA
Oceanography	Ocean Products Center	Suitland, Maryland
Pacific Ocean Data	Equatorial Pacific Information Collection	Seattle, WA

Particle Deposition Data	Air Resources Lab	Silver Spring, MD
Sea Level	University of Hawaii	Honolulu, HI
Snow and Ice	National Snow and Ice Data Center	Boulder, CO
Tides	National Tide and and Water Level Data Base	Rockville, MD
Trace Gases	Global Monitoring for Climate Change	Boulder, CO

3.0. Digital Data Request History and Projection

The support for global change by the NSF has increased about 35% per year since 1987 and is expected to increase for FY 1992. NOAA's global change funding has roughly doubled each year since 1989. Other agencies are also increasing their global change funding. The overall funding for all agencies has increased seven times from FY89 to 1991. From this, one could expect a large increase in the number of data requests at the centers. However, in the aggregate, there has only been a modest increase in the number of requests for each of the last two years ('89 and '90), and the projections are, therefore, based on this modest rate of increase. Another view is that the secondary distribution of data from scientist to scientist may be on the increase because of the ease of transmission over networks, coupled with a desire to obtain a dataset that has had use in a familiar science project. Possibly this secondary distribution masks the size of the actual data dissemination.

The biggest impact on the number of requests and volume of data distributed has been from the introduction of CD-ROMs. This indicates that the increased use of CD-ROMs for data distribution provides a means for rapid deployment of the data among members of the research community. Secondary distribution of data from scientist to scientist may be on the increase because of the ease of transmission over networks, coupled with a desire to obtain a dataset that has had use in a familiar science project.

4.0. Mass Storage/File Management Requirements

As the amount of data increases and the NOAA mission focuses on improving accessibility of data for global change research, there is an urgent requirement to develop mass storage systems with file management software at the centers to

improve archive management, provide vastly improved access mechanisms, and to reduce the amount of media and associated space requirements. Moreover, the mass storage system is the heart of a file management system. For the immediate purpose at hand, a mass storage system should include the following:

4.1. Hierarchical File Director

A hierarchical file directory is needed that permits, as a minimum, the acceptance of UNIX file names. The directory needs to maintain the access and update history information for the file. The directory should allow for handling mixed media within a single search, for cross indexing between devices, and for recording data set utilization records for future knowledge based system applications. This directory must interoperate with a number of different data base systems passing query information through during interoperable data searches.

4.2. A hierarchy of storage levels

The mass storage system should support a hierarchy of storage levels with increasing physical capacity and decreasing performance at the bottom, and decreasing physical capacity and high transfer rates at the top (as viewed from the user client processes). At the top, this permits the evolution to direct access electronic storage, so that the mass storage becomes a truly integrated part of a computing environment.

4.3. File Management Software

The file management software should offer options for data compression. It should permit the use of checksums as an overall error control mechanism. Data conversion software should be available to migrate the data from one physical media to another, as generic files, without disturbing the data content. The software would sample files on a statistical basis reading them to verify that they were still intact and that the media had not deteriorated beyond the point where only soft data checks were obtained. In the event that sufficient degradation was detected during this sampling process, the files would be migrated to new media with a corresponding directory update. Migration would also be triggered during normal accesses whenever too many soft errors occur.

Migration of files from archival or working media to a buffer storage area would occur following the initial access to permit data to be more rapidly retrieved from the faster devices in the storage hierarchy. The migration and actual location of the data should be presented to the user/application programs in a transparent manner including, as an option, presentation to client processes in a manner simulating direct retrieval from the ingest media if desired. To accomplish this transparency, the file management software should provide for the ingest of data from existing media and distributed to: standard half-inch magnetic tapes in all densities and formats, EXABYTE, DAT, CD-

ROMs, optical disks, video cassette recording technologies, digital optical media, etc. An encapsulation of the ingested media's data should record the presence and location of permanent data errors, physical record lengths in bytes, the presence and location of ingest media specific flags, such as tape marks, end of tape flags, etc., so that upon access, the data can be handed to processing programs that need to be aware of the different media. The file management software should be able to handle any of the existing data formats and to invoke conversion routines to a standard if one is adopted. The ability to move files from the mass storage to a requester's media in the original format should be provided.

4.4. Network Access and Networking

The NOAA centers should be coupled to the internet, at internet backbone rates, and eventually to the National Research and Education Network (NREN). With 740 universities, laboratories and industrial sites now on the network, and 75 more expected in the 2nd half of 1991, the internet is widely available to the scientists involved in global change and EOS. There are 16 NSFnet backbone sites. Two of these, the University of Maryland and NCAR, are in close proximity to NOAA data centers. Where large data volume data transfers are required, conventional conveyance services would probably suffice with economic considerations determining the mode of conveyance.

5.0. Considerations

The system life under the NOAA mandate to manage data for long-term global change research purposes is open ended. The value of data increases with age for use in performing long term environmental change research. Global patterns are known to be subtle over time, even when viewed in a rapidly changing environment. Today's collection of environmental data is pitifully small and of too short a duration compared to the amount required to filter out the statistical noise over an extended time domain.

The operational requirements are influenced by incremental science requirements established as the knowledge of the relationships between instrument responses and conversion to physical units became better known from the results of research and development of more sophisticated processing algorithms. Because the development of sufficient knowledge to fully understand the earth observation responses is an accretive and repetitive process, the entire data set will require repeated reprocessing to adequately describe the data for long-term documentation and preservation.

Aggregation of similar but different and sometimes disparate data types is also an important feature to include. A well understood aggregation principle will allow for compartmenting the data across the media domain in a "most" convenient form for vastly improving the access mechanisms. This will become increasingly important as

the longevity of the archive increases. Aggregation implies some degree of redundancy, but in a positive sense, in that redundancy of particularly critical data sets reduces the risk of data loss over time.

Volume management may require compression mechanisms to reduce the over-volume of data as it ages. Critical data and information will require the application of lossless data compression where data sets are reduced in volume. When permitted, other means can be used to reduce data volume with controlled data loss as achieved through sampling, or through small-loss data compression techniques or a combination of the techniques to yield much higher compression ratios. This may particularly attractive for managing very high volume image data sampled in the visible spectra.

The technological gallop of the last several years continues to accelerate and new storage and processing technologies are obviating the need to consider destruction of cumbersome data through full scene sampling, scan sampling or reduction to gross descriptive parameters which describe the sum of the parts in abbreviated form. In order to take advantage of improved and less costly storage technologies, there is a plan to migrate the data periodically as the volumes dictate and the technology allows and with each migration to yet developed capabilities, it becomes even more feasible to consider placing all of the data in a near-line access environment. The migration process will require content processing to re-develop the cross reference inventory information to include additional content description information as part of the inventory metadata file to increasingly document the data as it ages. Data migration is an essential element in developing a never-ending data life for the sake of offering an extended time baseline data set essential for detecting global scale changes.

A wide variety of media will be used for distributing data and information to users. The large number of formats used by the NOAA data centers means that many conversion procedures will be needed. It would be better not to convert the data in the archives themselves, but to write procedures which can be invoked in a demand fashion. In this way, the data can be left in its original form while confidence is gained in the accuracy of the conversion. If any problems arise in the converted data, the original data will not be contaminated. The problem becomes one of reworking the conversion routine and alerting previous users of the defect rather than trying to fix a partially scrambled dataset. The downside is that there will be an additional processing cost when the data is requested. Another problem with the standard data format concept is that many researchers who submit data to the data centers will probably never conform to a standard format. Insistence on a format may become an impediment to releasing the data to a center for distribution.

6.0. Assumptions and Constraints

Factoring today's technology advancement timescale for the purpose of being both realistic and conservative, the period of migration to exploit new technologies and avoid system obsolescence to extend the validity of the data and information is established at no less than every 10 years. A suggestion was recently made that the migration rate be a function of the expected half life of the medium used. For the 3480 tapes, the manufacturers agree that 10 years of full performance life should be expected, thus the half life for migration purposes would be five years. The criteria for accepting a new technology as a candidate for data migration is; improved archival qualities, the per data byte storage cost must be one half the previous, the physical storage requirements be at least five times less, and the data transfer rate to move the data from the media be no slower than that of the older media. And, finally, the data migration step will include data processing to derive or extend content description data to be used for the purpose of validating the preservation state of the data and for reinventoring the data to add content information developed through the accretion of user knowledge and experience.

Another assumption is that all of the data will be reprocessed three times during a 25 year cycle. This reprocessing cycle may coincide with a data migration step since all of the migration will include a content review during the passage of data from one medium to another. The development of reprocessing algorithms will not be charged as a data management system task but will require that the data management system be able to put significant quantities of data on-line or at-hand for "live" ingest mode processing.

7.0. Conclusion

To continue managing data as we do today would eventually require a facility to accommodate an enormous number of media units to hold the data volumes projected for the future. As the new data continues, the added function of migrating data from degenerating media (from a systems as well as physical viewpoint), will compound the storage requirements as the annual data volume accumulates by the hundreds of trillions of data samples each year. If acceptable mass store systems are not continually developed to match the data growth and data management requirements, the logistics would be overwhelmed and the system would fall apart never to be recovered again because of the enormous cost to recover an inevitable backlog.

The data only has value to the research community if it is conveniently and efficiently accessible. If the data were placed in a warehouse environment, which would ultimately have to happen if nothing was done, it would soon lose its value and possibly its identity because of the cost to acquire it and eventually would be lost because of the huge cost to locate, ship back, copy and return the data copies as the data volume grows beyond manageable proportions. This is aside to the issue of data

loss due to media deterioration. The only acceptable solution would be through development of a system capability to provide highly efficient and sophisticated data management capabilities which would accommodate online data sources. In order for this to happen, advanced media technologies have to be employed along with advanced sophisticated data management software to eliminate the manual interfaces where possible to provide the data in a ready mode for user interaction at the subsetting level. The data value increases dramatically when placed in this type of environment as the access to the system can provide instant gratification and encourages repeated and expanded data query activities. This, in turn, accelerates the research progress and enhances the research results thus broadening the value of the data to the advancement of science and knowledge.

To physically compress the data through the implementation of high capacity media coupled with the systems capability to control and index these data in an online or near-line environment offers a significant reduction in the requirements to house the archives both in terms of physical space and recurring energy and labor expenses. The closer on line the data are placed, the less labor is required to physically mount data either in the appropriate archive slot or onto the processing system. As electronic access become more fully integrated into the system, direct labor service categories will be eliminated. A major cost avoidance to be reckoned with is the cost of adding increasing large physical facilities as the data volume grows at the projected rates. The pace of implementation of new technologies should allow shrinking of the space requirements to match the increase of data accumulating in the facilities. An indirect benefit of space compression through improved storage technologies would be realized form compressing the facilities requirements sufficiently to consider replicating the data in distributed locations as a risk reduction measure.

The broadest benefits are in terms of what the value of the data is to the world of environmental change. Without a responsibly managed data record of scientific measurements over time, there would be no baseline to objectively determine if the environment we live in is really changing, how much, and at what rate. Without these data, economies would be based on subjective opinions and in some cases, hear say. Public policy would more often than not be misguided and consequences of enormous proportions could occur to our physical well being either through economic collapse or through direct physical changes. Even today, global change scenarios portend potential devastating effects to our coastal cities and this country's agro-economies. But, do we build dikes and seek alternate water sources if we are not really sure what, if any, impacts there are? Without the data, no one knows, so any investment in mitigating a potential problem is an economic risk. The other question is; even if we know, can we afford to take avoidance action? Or better yet, is the cause due to environmental causes or due to a much broader cyclic processes. One thing is certain, there is a great potential for change based on the knowledge at hand today, and sound economic planning based on knowledge may be sufficient to avoid economic collapse. In a Nation with a trillion dollar economy, the risks are too great

not to invest an insignificantly small percent of this economy to acquire the maximum amount of knowledge possible and establish this knowledge base as soon as possible. In the case of environmental data, data is knowledge; there can never be enough data, and the data record can never be long enough. But where there is data, it must be systematically managed to be of any value at all.

DATA MANAGEMENT

WILLIAM M. CALLICOTT

OFFICE OF SYSTEMS DEVELOPMENT
DATA MANAGEMENT SYSTEMS DIVISION
NATIONAL ENVIRONMENTAL SATELLITE,
DATA AND INFORMATION SERVICE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

NSSDC CONFERENCE ON MASS STORAGE SYSTEMS AND TECHNOLOGIES
FOR SPACE AND EARTH SCIENCE APPLICATIONS

NASA/GODDARD SPACE FLIGHT CENTER
GREENBELT, MARYLAND
JULY 25, 1991

DATA MANAGEMENT PROCESSES

- INGEST
- QUALITY CONTROL
- CATALOG
- ACCESS
- PRESERVATION

INGEST

- Three discipline centers and U.S. World Data Centers:

  National Climate Data Center - Asheville, North Carolina
  National Oceanographic Data Center - Washington, D.C.
  National Geophysical Data Center - Boulder, Colorado

- Sixteen centers of data:

  Satellite Data Processing and Distribution - Suitland, MD
  National Weather Service - Silver Spring, MD
  National Meteorological Center - Camp Springs, MD
  Charting and Geodetic Services - Rockville, MD
  Climate Analysis Center - Camp Springs, MD
  National Marine Fisheries Service - Seattle, WA; Woods Hole, MA; Miami, FL; Bay St Louis, MS; San Diego, CA
  Navy Joint Ice Center - Suitland, MD
  Great Lakes Environmental Research Lab - Ann Arbor, MI
  Center for Ocean Analysis and Prediction - Monterey, CA
  Ocean Products Center - Suitland, MD
  Equatorial Pacific Information Collection - Seattle, WA
  Air Resources Laboratory - Silver Spring, MD
  University of Hawaii - Honolulu, HI
  National Snow and Ice Data Center - Boulder, CO
  National Tide and Water Level Data Base - Rockville, MD
  Global Monitoring for Global Change - Boulder, CO

NOAA CENTERS

- DIGITAL HOLDINGS INCLUDE: 11 TB ON 240,000 COMPUTER TAPES
- IN-SITU DATA INCLUDED IN ABOVE: 2 TB
- GOES DATA HELD AT THE UNIVERSITY OF WISCONSIN: 100 TB ON 25,000 SONY U-MATIC BETA TAPES
- SERVICE OVER 7,000 REQUESTS FOR DATA AND INFORMATION PER YEAR
- ANNUALLY PRODUCE 5,000 BULLETINS TO 40,000 SUBSCRIBERS
- OVER 90,000,000 PAGES OF NON-DIGITAL DATA AND INFORMATION

IMMEDIATE DIGITAL VOLUME GROWTH
(VOLUMES IN BILLIONS OF BYTES)

	1991	1992	1993	1994	1995	1996
CLIMATE DATA CENTER	520	570	840	1,300	2,130	3,420
OCEANOGRAPHIC DATA CENTER	30	160	200	260	280	370
GEOPHYSICAL DATA CENTER	450	540	650	780	930	1,110
SATELLITE DATA SERVICES	10,400	14,200	18,000	21,800	25,600	29,400
GOES DATA ARCHIVE	107,000	113,000	120,000	134,600	149,200	164,000
ACCUMULATIVE TOTAL:	118,400	128,470	139,690	158,740	178,140	198,300

ALL ENVIRONMENTAL SATELLITE SOURCES

Includes level–1 (L0+10%) plus levels 2–4 (40% of L0)

VOLUMES IN TRILLIONS OF BYTES PER YEAR

SATVOLS 6/25/91 Disk #8

SATELLITE SYSTEM	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Conventinal OBS	2.00	1.00	1.00	2.00	3.00	4.00	5.00	5.00	5.00	5.00	6.00	6.00	6.00	6.00	6.00	8.00	8.00	8.00	8.00	8.00	8.00	10.00	10.00	10.00
NOAA–D (5/91)	0.55	1.10	1.10	0.55																				
ERS–1 (9/91) (ESA)	0.03	0.11	0.11	0.11																				
NOAA–I (12/91)		1.10	1.10	1.10																				
JERS–1 (3/92) (NASDA)		0.08	0.16	0.16	0.08																			
TOPEX (6/92) (NASA)		0.08	0.16	0.16	0.16	0.08																		
GOES–I (10/92)		0.93	9.27	9.27	9.27	9.27	9.27	9.27																
GOES–J (6/93)			4.64	9.27	9.27	9.27	9.27	9.27	4.64															
NOAA–J (12/93)				1.10	1.10	1.10																		
NOAA–K (7/94)				0.48	1.20	1.20	0.72																	
RADARSAT (7/94) (CAN)				0.05	0.16	0.16	0.16																	
ADEOS (3/95) (NASDA)					0.08	0.16	0.16	0.16																
NOAA–L (7/96)					0.48	1.20	1.20	0.72																
NOAA–M (2/97)							0.90	1.20	1.20	0.30														
TRIMM (3/97) (NASA)							0.05	0.11	0.11	0.11														
GOES–K (10/97)							1.85	9.27	9.27	9.27	9.27	9.27	4.64											
GOES–L (6/98)								4.64	9.27	9.27	9.27	9.27	9.27	4.64										
EPOP–A (9/98) (ESA)								5.06	20.26	20.26	20.26	15.19												
EOS–A (12/98) (NASA)									58.20	58.20	58.20	58.20												
JPOP (12/98) (NASDA)									0.18	0.18	0.18	0.18												
NOAA–N (9/99)									0.30	1.20	1.20	1.20												
NOAA–O (4/01)											0.60	1.20	1.20	0.60										
EOS–B (6/01) (NASA)											29.10	58.20	58.20	58.20	29.10									
EPOP–B (9/02) (ESA)												4.05	20.26	20.26	20.26	16.21								
GOES–M (10/02)												1.85	9.27	9.27	9.27	9.27	9.27	7.42						
EOS–C (12/02) (NASA)													58.20	58.20	58.20	58.20	58.20							
GOES–NEXT (6/03)													5.37	10.73	10.73	10.73	10.73	10.73	5.37					
JPOP–B (12/03) (NASDA)														0.18	0.18	0.18	0.18							
NOAA–P (4/04)														10.13	20.26	20.26	20.26							
EPOP–C (9/06) (ESA)																5.06	20.26	20.26	20.26	10.13				
EOS–D (9/06) (NASA)																14.55	58.20	58.20	58.20	58.20	29.10			
NOAA–Q (4/07)																	10.13	20.26	20.26	20.26				
GOES–NEXT–P (10/07)																	2.15	10.73	10.73	10.73	10.73	10.73		
EOS–E (12/07) (NASA)																		58.20	58.20	58.20	58.20	58.20		
GOES–NEXT–PP (6/08)																		5.37	10.73	10.73	10.73	10.73	10.73	
NOAA–R (4/10)																				10.13	20.26	20.26	20.26	
EPOP–D (9/10) (ESA)																				14.55	58.20	58.20	58.20	58.20
EOS–F (9/11) (NASA)																					14.55	58.20	58.20	58.20
NOAA–S (4/13)																							10.13	20.26
EPOP–E (9/14) (ESA)																								14.55
TERA BYTES / YEAR	2.57	4.39	17.53	24.25	24.82	26.46	28.61	44.71	108.43	103.80	134.09	164.63	172.41	178.21	154.00	142.46	197.38	199.16	191.75	200.93	209.77	226.32	167.52	161.21

- TIROS N: 1.10 Tera Bytes/Year
- NOAA–KLM: 1.20 Tera Bytes/Year
- GOES Composite: 1.97 Tera Bytes/Year
- GOES–I/M GVAR: 7.30 Tera Bytes/Year
- GOES–NEXT GVAR 8.76 Tera Bytes/Year
- NOAAPOP: 20.26 Tera Bytes/Year
- EPOP: 20.26 Tera Bytes/Year
- EOS–A Platforms: 58.20 Tera Bytes/Year
- EOS–B Platforms: 58.20 Tera Bytes/Year
- NOAA EOS Reqts 11.50 Tera Bytes/Year
- NOAA JPOP Reqt's: 11.83 Tera Bytes/Year
- Eos SAR Platform 118.26 Tera Bytes/Year
- Space Sta Freedom 0.18 Tera Bytes/Year

Accumulated in 20 years: From 1991 to 2011 — 2.3 Peta Bytes

Daily Totals in MBytes for Level 1 plus Levels 2 – 4 Data (1.5 x L0):

			GBytes/Day Archive Vol
– Composite Mapped GOES:	3800		5.400
– GOES–I/M bulk GVAR:	20000		20.000
– GOES Next bulk GVAR:	24000		24.000
– TIROS–N:	2000		3.000
– NOAA KLM:	2200		3.300
– NOAA Platform:	37000	(3.5 Mb/s)	55.500
– European Platform:	37000	(3.5 Mb/s)	55.500
– EOS–A Research Platform:	106304	(9.84 Mb/s)	159.456
– EOS–B Research Platform:	106304	(9.84 Mb/s)	159.456
– NOAA Reqt's off EOS:	21000	(2 Mb/s rate)	31.500
– JPOP	21600	(2.0 Mb/s)	32.400
– SAR (Special SAR S/C)	216000	(20 Mb/s rate)	324.000
– Space Station Freedom	324	(.03 Mb/s)	0.486

PRESERVATION

- NOAA MISSION TO MANAGE THE ARCHIVES ON A PERMANENT BASIS AS A NATIONAL TRUST
- ALTERNATIVES FOR PRESERVING DATA ON AN INDEFINITE BASES:
  - FIND A MEDIA THAT IS INDELIBLE INDEFINITELY
    - ... MANAGE ACCESS SYSTEM INDEFINITELY
    - ... ENSURE MEDIA LONGEVITY (ENTOMB MEDIA AND SITE)
    - ... RECURRING QUALITY CONTROL
    - ... LIFETIME SYSTEM MAINTENANCE
    - ... MIGRATE ON DEMAND
    - ... OPERATE ARCHIVE CENTER(S) AS DEEP ARCHIVE
  - ASSUME A 10 YEAR SYSTEM AND TECHNOLOGY CYCLE
    - ... RECURRING MIGRATION OF 10 YEAR OLD MEDIA CONTINUALLY LOOKING AT DEVELOPING TECHNOLOGY ADVANTAGES
    - ... KEEP TWO COPIES, ONE ENTIRE DATA SET ENTOMBED, ONE IN ACTIVE ARCHIVES AT DISTRIBUTED DISCIPLINE ARCHIVE CENTERS
    - ... MAINTAIN PORTIONS OF THE DATA AT THE ACTIVE CENTERS ON-LINE, THE REST NEAR-LINE
    - ... THE MASTER COPY KEPT NEAR-LIE WITH SUFFICIENT ON-LINE CAPABILITY TO SERVICE ACCESS REQUESTS AND FOR MIGRATION PROCESSING

ARCHIVE INTEGRITY

- MEDIA PERFORMANCE CONTINUALLY MONITORED THROUGH ERROR DETECTION AND CORRECTION INFORMATION PASS BACK
  - PROCESSING ON-DEMAND
  - SCHEDULED MEDIA MAINTENANCE
- MIGRATION OFFERS OPPORTUNITIES TO:
  - UP-TO-DATA CATALOG FACILITIES
  - INCLUDE LOW-LEVEL DATA INVENTORY DESCRIPTIONS WITHIN CATALOG INVENTORY FILE
  - IMPROVE DATA AGGREGATION TO MEET CURRENT SCIENCE REQUIREMENTS
  - COMPACT STORAGE AND DATA TRANSFER THROUGH THE USE OF ADVANCED TECHNOLOGIES
  - REGENERATION OF SYSTEMS AND DATA EXERCISES THE DATA FOR ITS HEALTH
  - ENABLES INCREASED ON-LINE ACCESS FACILITIES
  - OPENS NEW DOORS FOR DATA ACCESS AND TRANSFER
- QUALITY CONTROL:
  - ANALYZE DATA TO ENSURE CREDIBILITY DURING INGEST
  - MONITOR MEDIA PERFORMANCE TO ENSURE RELIABILITY OVER TIME
  - MAINTAIN LOG OF ACCESS ACTIVITIES TO BUILD DECISION HISTORY

Storage Needs in Future Supercomputer Environments

Notes for the presentation by:

Sam Coleman
Lawrence Livermore National Laboratory

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

Introduction

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

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

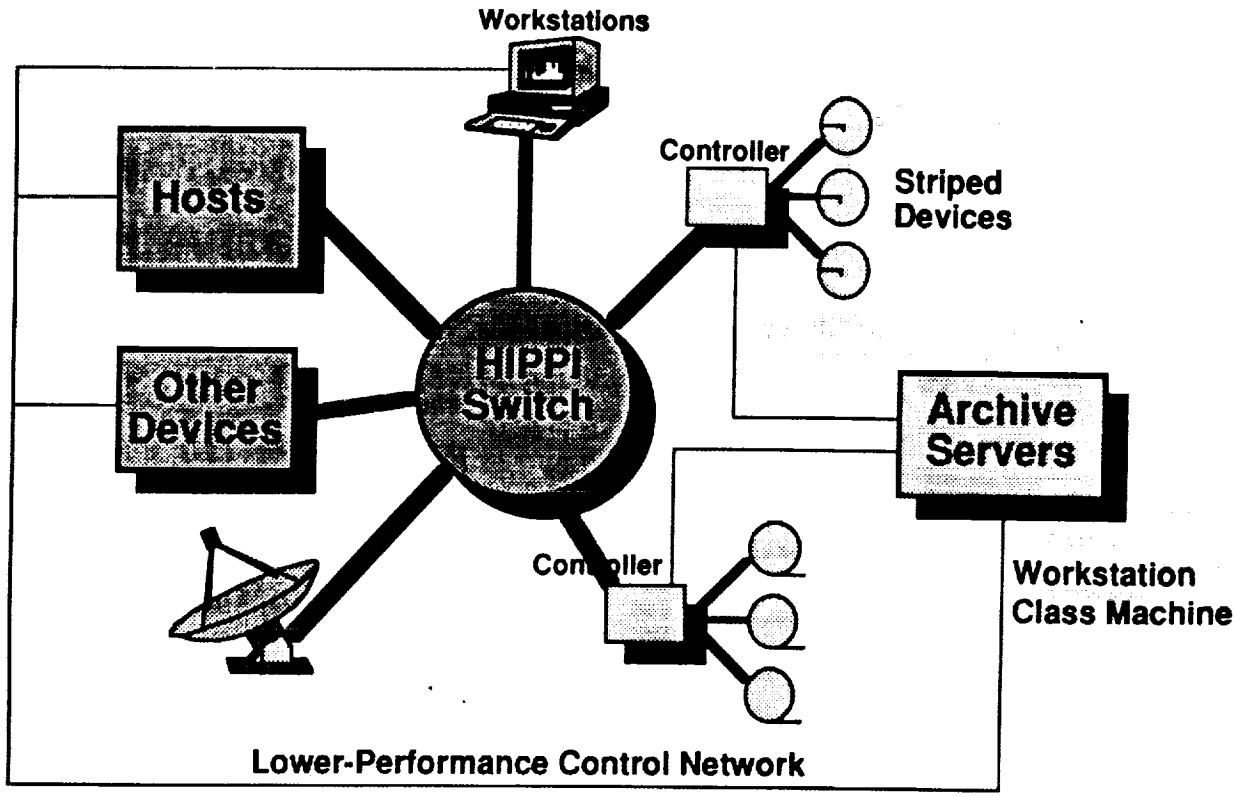
Storage Architectures

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

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

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

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



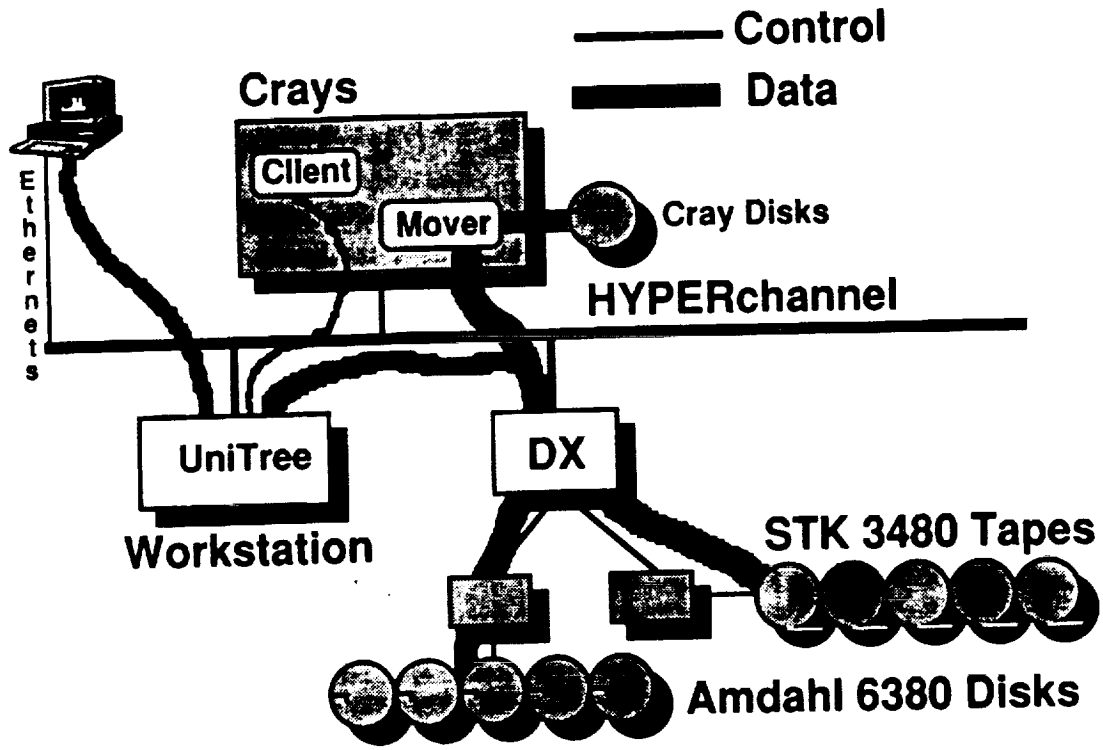
A High-Performance Storage Architecture
Figure 1

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

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

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

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



An Interim Storage Architecture at LLNL
Figure 2

available today, the necessary software is not. In particular, there is no high-level file-transport software in the NSC DX HYPERchannel adapter. As an interim solution, we will put IEEE movers[1] on our host machines, allowing direct file-transport to and from the storage devices over the HYPERchannel. The UniTree workstation will provide service to client workstations and other network machines. This is acceptable, in the near term, because most of the archival load comes from the larger host machines. This architecture will replace the Amdahl mainframes that we use to control the current archive.

Software Needs

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

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

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

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

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

Execution
Programs can execute in any location without being changed.

Fault
Clients are not aware that certain faults have occurred.

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

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

Migration
Clients are not aware that services have moved.

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

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

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

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

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

The IEEE Reference Model

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

Application
Normal client applications codes.

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

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

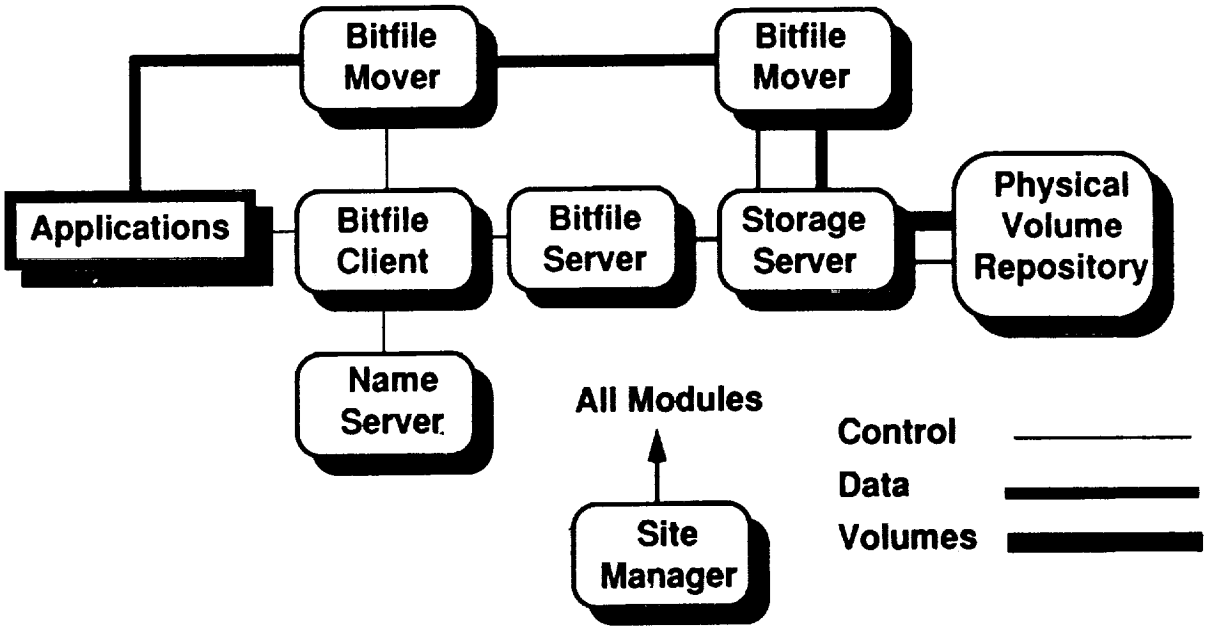
Storage Server
The module that manages the actual storage of bitfiles, allocating media extents, scheduling drives, requesting

volume mounts, and initiating data transfers.

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

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

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



The IEEE Mass Storage System Reference Model
Figure 3

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

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

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

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

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

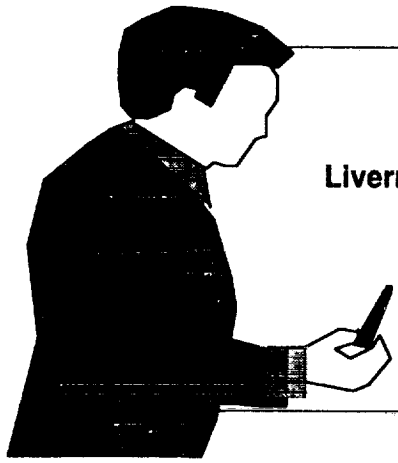
To learn more about all of the storage issues that I have mentioned, I would encourage you to attend the 11th IEEE Mass Storage Symposium in Monterey, California October 7-10, 1991. For details, contact:

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

Reference

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

Storage Needs in Future Supercomputer Environments



Sam Coleman
Livermore Computer Center

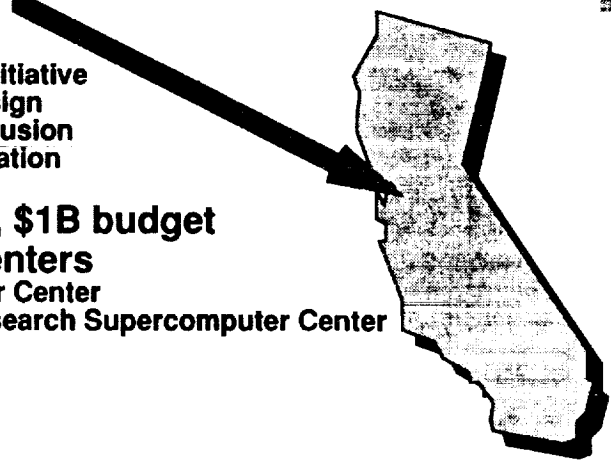
July 25, 1991
NASA Goddard

University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

The Lawrence Livermore National Laboratory

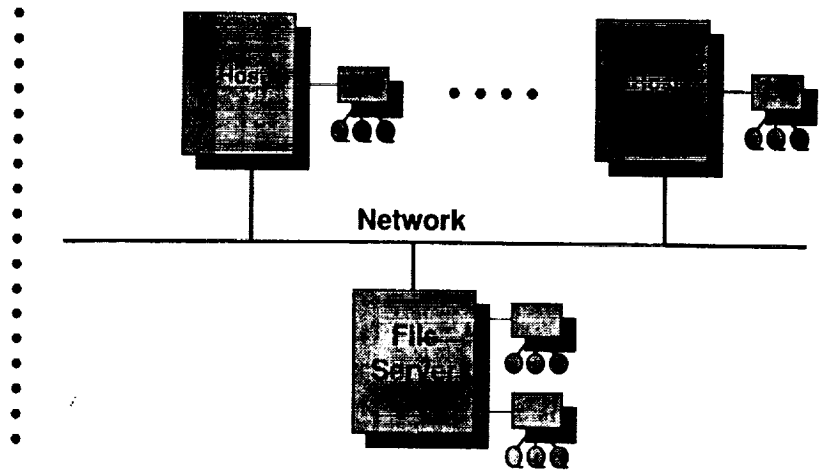
- Department of Energy contractor
- Managed by the University of California
- Founded in 1952
- Major projects
  - Strategic Defense Initiative
  - Nuclear weapon design
  - Magnetic and laser fusion
  - Laser isotope separation
  - Weather modeling
- 8,000 employees, $1B budget
- Two computer centers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



University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

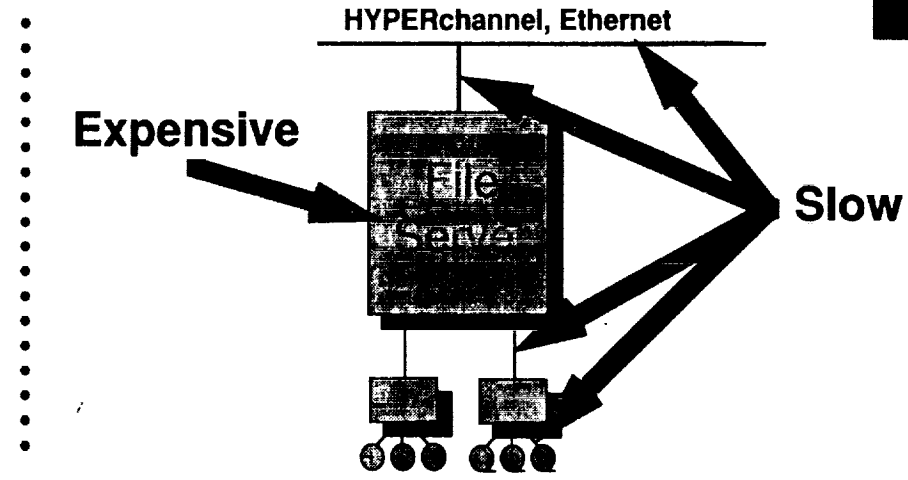
Traditional Supercomputer Storage Architecture



University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

Problems with the Traditional Architecture



University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

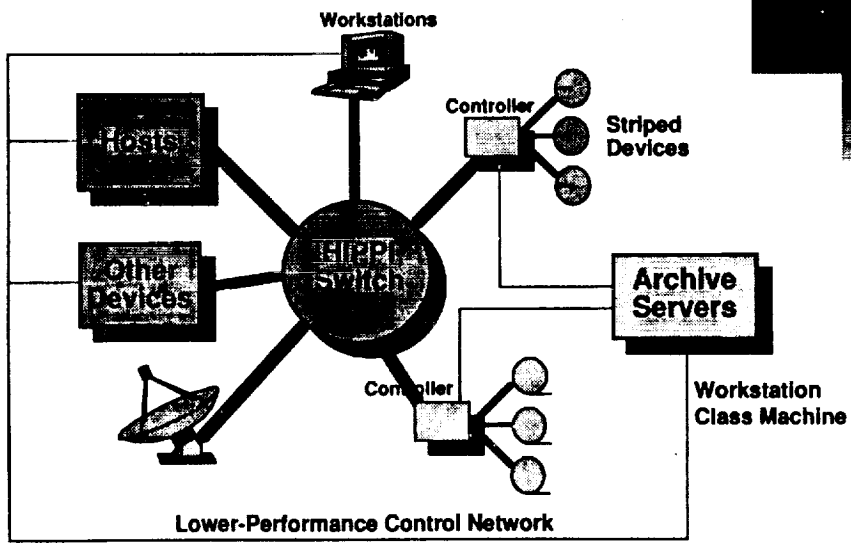
The Need for Higher-Performance Storage

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

University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

A High-Performance Storage Architecture



University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

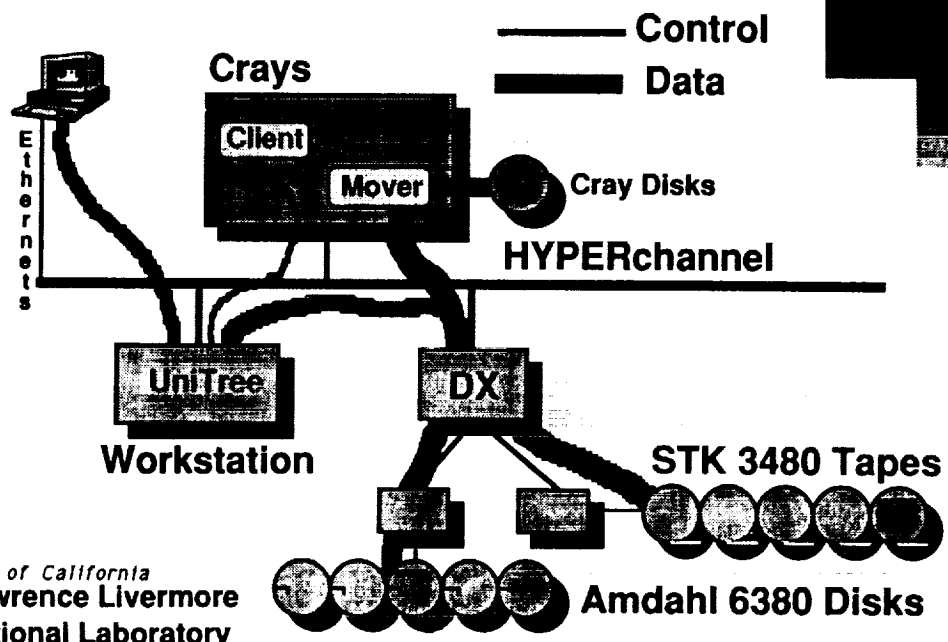
What is Needed

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

University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

In the Meantime.....



University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

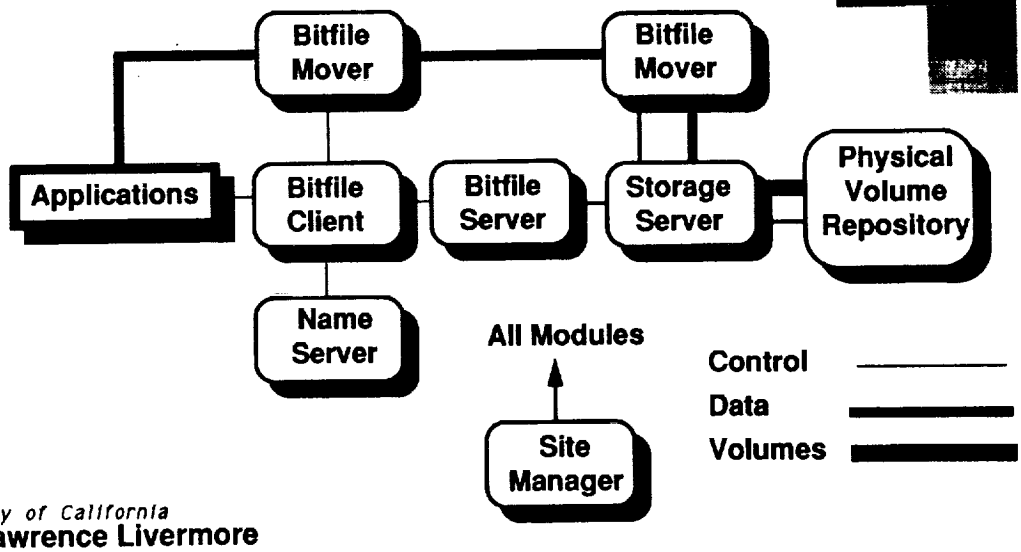
Software Needs

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

University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

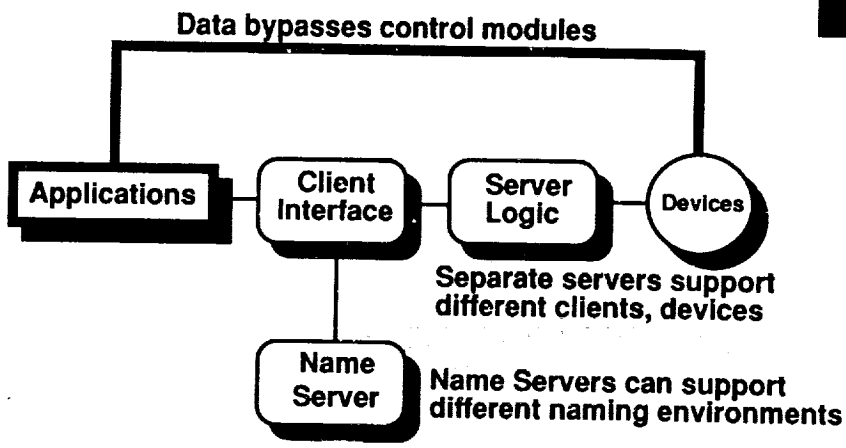
The IEEE Mass Storage System Reference Model



University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

The Significant Modularity



University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

In the Meantime.....

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

University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

Summary of Important Issues for Future Storage Systems

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

University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

To Learn More

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

University of California
Lawrence Livermore
National Laboratory

NASA SC 7/25/91

Requirements for a Network Storage Service

Suzanne M. Kelly and Rena A. Haynes

Sandia National Laboratories
Albuquerque, NM

INTRODUCTION

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

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

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

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

APPLICATION DESCRIPTION

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

NSS Perspective

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

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

NSS Functions

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

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

End-User Functional Requirements

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Operations Functional Requirements

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

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

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

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

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

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

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

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

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

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

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

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

NSS Characteristics

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

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

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

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

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

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

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

SYSTEM DESIGN CONSTRAINTS

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

Network Interface

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

Performance

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

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

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

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

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

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

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

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

Capacity

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

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

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

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

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

Security

The NSS must enforce the following security rules.

1. Users must be authenticated before accessing the system.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

ACKNOWLEDGEMENT

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

TRADEMARKS

UNIX is a registered trademark of AT&T.

NFS is a trademark of Sun Microsystems, Inc.

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

CLOSING REMARKS

Ben Kobler

MR. KOBLER: That concludes our conference. I thank you all for attending and for sticking it through to the end.

I am not going to try to summarize three days of proceedings in just a few minutes; but I did want to leave you with just a couple of thoughts.

One is that I think the understanding of the underlying chemistry and physics of media is important; and that is one of the things we have tried to emphasize here at this conference. Let's not forget about that as we select the media that we will be putting into our deep archives.

The second point is that as we build some of these media solutions into systems -- first into local systems and then into wide-area systems -- I am hopeful that we will be getting to the point where we can manage the wide-area networks under much tighter controls than we are able to do today, so that we can, hopefully, build and take advantage of some of the distributed RAID technologies that were mentioned during this conference.

And then the third point, which wasn't really emphasized very much during this conference, but which is important, is that as we put this data into the archive we want to be sure that we will be able to get to the data later; and as the data volumes increase, we want to be intelligent about how we manage that data, so the concept of building intelligent front ends is going to be important in the future.

With that, let me conclude by just thanking the program committee for putting together an excellent program. I would like to just mention their names: John Berbert, Sue Kelly, Elizabeth Williams, Al Dwyer, P.C. Hariharan, and Sanjay Ranade. Thank you for attending.

(Applause)

(Whereupon, at 3:30 p.m., the conference was adjourned.)

DR. BHUSHAN: Thank you, Dr. Halem. The keynote address this morning will be given by Mr. Fred Moore. Fred Moore has been a Corporate Vice President of Strategic Planning since May 1990. Mr. Moore came to Storage Technology Corporation 14 years ago as a systems engineer and has since held positions including Vice President of Systems Marketing and Director of Worldwide Product Marketing and Director of Marketing Systems Engineering. Mr. Moore has published numerous articles and has spoken on five continents at industry conferences. Before joining Storage Technology, he worked for the Public National Bank in Dallas and the University of Missouri as a systems programmer.

Mr. Moore received a Bachelor Degree in Mathematics and a Master's Degree in Computer Science from the University of Missouri in 1989, and his alma mater honored him with an Arts and Science Distinguished Alumni Award. Mr. Moore?

DR. BHUSHAN: Thanks again for a really interesting presentation. The next talk is entitled "Optical Disk and Tape Technology," and it will be given by Dr. Robert Freese, who is the President and co-founder of Alphatronix, Inc. After earning a Ph.D. in Optical Engineering from the Institute of Optics at the University of Rochester and a B.A. in Physics from Kalamazoo College, Dr. Freese spent six years leading 3M's erasable optical recording efforts, where he built and operated the world's first computer-controlled erasable optical media.

Before his work in erasable optical technology, Dr. Freese supervised the research activities of 3M's erasable optical technology. Dr. Freese received the 1989 Distinguished Inventor Award from the Intellectual Property Owners, Inc. Foundation for his contribution to the development of erasable optical media.

In April 1991, Dr. Freese and his founding partners were honored with an Entrepreneur Excellence Award for Outstanding Achievement in 1990-1991 from the North Carolina Council for Development. He has received numerous other awards.

Without further ado, Dr. Freese.

DR. FREESE: Thank you. Everybody who promises to read the paper, please applaud now.

(Applause)

(Laughter)

DR. FREESE: We have some time for any questions or comments.

PARTICIPANT: (Inaudible)

MR. SAVAGE: Forever.

PARTICIPANT: (Inaudible)

DR. FREESE: Let me take a minute and see if I can paraphrase your question. The subject of your question dealt with: What is the definition of "forever" for the various types of data which might exist in the world? What is your opinion of forever? Did I get the question right?

MR. SAVAGE: I think that's the way I interpreted the question. In my company and in my industry, the word "forever" is forever. As long as the oil business lasts, as long as the mining business lasts, we will consider that data to be extremely valuable because it's stuff that we gathered out of the earth. It's good stuff; it isn't going to change. The scientific world is always going to be interested in that data.

Forever means forever; and to the National Archives, forever means forever. I mean, that's not a question of 50 years; it's forever. And I think that the data that you are going to be collecting with EOS--a great majority of it, especially the first level of the data that you store before you start jacking around with it and interpreting it--the uninterpreted first clean copy of the data that you acquire should be your archive.

You can take it and do anything you want with it; you know, process it and make interpretations, but I think that is forever. We would never want to lose the logs that we got from sailing ships of hundreds of years ago. We want to keep that stuff alive forever. Really, forever; and it is possible to do this.

But as I say, if you want to keep it forever, you have got to protect it. You have really got to protect it.

DR. FREESE: A question in the back?

PARTICIPANT: How large is your archive at Shell

MR. SAVAGE: I'll say that it is verging on going over two million reels and cartridges.

PARTICIPANT: How many bytes or terabytes or exabytes?

MR. SAVAGE: I estimate that it is--when I wash out all the empty parts of the tapes and everything--I estimate that there are 300 terabytes there. But you guys are talking about single satellites that will be accumulating data; so, that's the accumulation of 25 or 26 years.

You guys are talking about single satellites that will collect that much data in a year. So, I've got a big problem; I don't manage this stuff myself. Shell has a big problem of managing what I call this very large digital data archive.

You guys are going to have a problem that is much bigger. You know, you have got to approach it systematically. You have got to understand the magnitude and the horror of what your mission is. That's the word I'm trying to get over to you.

Don't belittle this problem. It is an extremely difficult problem, and it is going to be a very costly one. If you really want to keep that data forever, you are going to have to pay the piper.

DR. FREESE: Additional questions?

(No response)

DR. FREESE: Very good. Thank you very much.

(Applause)

DR. BHUSHAN: So far this morning, we have heard about magnetic tapes, optical disks, and optical tapes. Our final presentation this morning is on magnetic disk; it will be given by Dr. John Mallinson, who received his M.A. Degree in Physics from University College at Oxford. He joined AMP in Harrisburg, Pennsylvania in 1954 to work on the theory and design of all magnetic lodging elements. In 1962, he joined the Ampex Corporation in California, where he held many positions concerned with the understanding and development of magnetic recording systems. From 1976 to 1978, as a manager of hybrid recording in the Data System Division, he was concerned with initial design of 750 megabit-per-second digital recorders.

From 1978 to 1984, he supervised the Magnetic Recording Technology Department, High Density Head Publication, Coding and -- Theory, and Exploration of Advanced Concepts of various areas of recording.

In 1984, he was appointed as the Founding Director of the Center for Magnetic Recording Research at the University of California, San Diego. Since 1990, he has been the President of Mallinson Magnetics, Inc. He has published over 60 papers on a wide variety of topics in the magnetic recording field.

DR. BHUSHAN: Let's take a moment for questions

PARTICIPANT: (Inaudible)

DR. MALLINSON: The question is regarding my not making any reference to the increasing density of large-scale integration silicon memory devices like DRAMs and things. And Mr. Moore had a table on that, I think, didn't you? It showed it all depended on the line size that you can get from diffraction; and he remarked that when the line size gets below a quarter of a micron, extensive retooling to something--I don't think he said what--but let's call it X-ray lithography or something is required.

The figures that I have seen on these projection of areal density of LSI devices and magnetic recording show that in the LSI devices, the areal density is doubling every two years. And if one assumes that it continues to double every two years, without regard to this glitch about running out of ultraviolet diffraction and having to move to X-ray or synchrotron light or something like that, if it continues like that, then the figures that I have seen show that the crossover occurs in the year 2007.

Why is this as far out as that? It's basically because LSI is so expensive and disk recording is so cheap at the moment. At the moment, you can have 2 gigabytes of storage for $2,000; 1 gigabyte of LSI at the moment costs $.5 million; 2 gigabytes of LSI costs $1 million.

So, there is an enormous leeway from $1 million to $2,000 to be made up. And I personally, having been trained as a physicist, am more skeptical that the LSI people--the semiconductor people--can go on through this diffraction limit without change in slope than I am that magnetic recording can continue to do what I've just been talking about because recording doesn't depend on diffraction. It's that change of philosophy at the end that I was mentioning.

PARTICIPANT: (Inaudible)

DR. MALLINSON: No. The question is: Was there any indication in the IBM demonstration that they were doing optical track following? The servo in that demonstration was one man with his hands on a micrometer. In fact, it's the only part--the track following was the only part of the IBM demonstration which was, as they say, "flaky".

(Laughter)

DR. MALLINSON: They did everything else. They were recording 50 tracks, one after the other; they were going back and changing data and tracks. But the track following servo was a man.

PARTICIPANT: This was demonstrated in 1989; where do you think they are now?

DR. MALLINSON: Well, I don't know. IBM is a very closed-mouth company. I don't know the answer to that question. I do know that in the spring of 1989, I went up there to give a talk on areal densities in magnetic recording and told them that I reckoned that thin film media of today should support 1.6 gigabits to the square inch and not a man-jack of them blinked.

(Laughter)

DR. MALLINSON: But I guess I now know that half a dozen of them were actually doing it at that time. The Hitachi experiment that was reported at the INTERMAG Conference five weeks ago was 2 gigabits to the square inch; and they didn't do it all. They did the writing of many tracks at 17,000 tpi; the inserting of several different data tracks, the measuring of the signal-to-noise ratio, the measuring of the crosstalk.

But they didn't implement the signal processing channel; they did not do the Viterbi or the EDAC. They just looked at an oscilloscope eye pattern and said: Gosh, that looks like a detectable eye pattern.

(Laughter)

MR. YEAGER: I'm Tim Yeager from NOAA. (Inaudible)

DR. MALLINSON: The question is: Do I have any idea about supercomputer disks? And the answer is no; I have no idea about supercomputer disks. What's special about supercomputer disks?

PARTICIPANT: (Inaudible)

DR. MALLINSON: All I know about supercomputer disks is that they run at 6 megabytes a second--48 megabits. They run at twice the normal data rate.

PARTICIPANT: (Inaudible)

DR. MALLINSON: Oh. Well, parallel transfer disks; it seems to me to be just a matter of doing it. Parallel transfer disks have been made--and I said it in the written version of the paper--that will support the full CCI component digital video. That is the 216 megabit per second.

I'm not really answering your question; I don't know about supercomputer disks.

PARTICIPANT: (Inaudible)

DR. MALLINSON: The question is: As the flying height decreases, is there some implication on the head crashes and the stability of the disk? Yes, you are exactly right; that is the whole issue. How low dare you run the head and still avoid nondestructive crashes?

On the question of head crashes, I might add that it seems to me that, once upon a time in the late 1950s or early 1960s, the name "head crash" was a very appropriate name because you did, indeed, go through the disk drive; and there's a mess inside it. The head had crashed.

But I'm told now that the name "head crash" is a generic term now, or "disk crash" is a generic term, that refers to anything that has happened to the disk drive. And over 90 percent

of the time, what has happened to the disk drive is, as you might guess, the high power component of the drive motor or the smoothing capacitor on the drive motor has failed.

Disk drives are sold at the moment--for instance, a Maxtor disk drive says on it: Mean time to failure: 100,000 hours. Mean time to repair: 22 minutes. Well, in 22 minutes, I submit that all the service man can do is run the diagnostic routine, find that the capacitor on the drive motor has failed, and replace that, or he can replace one PC card. I don't think in 22 minutes he can even replace a head, let alone the disk.

So, I'm told that actual head disk crashes are a very rare event; and most of the disk crashes are other failures, to do with selected circuitry.

DR. BHUSHAN: Any other questions? Yes?

PARTICIPANT: (Inaudible)

DR. MALLINSON: The question is: What about the life of the magnetic storage material? It's a long, long story. The most stable magnetic storage materials are those that are based on iron oxide because iron oxide, after all, is the highest or the lowest oxidation state of iron. It can't do anything.

So, the next question is: What happens about the plastic binder system in tapes? The answer on that seems to be that if you keep the tapes at the correct temperature and humidity; and the correct temperature and humidity is something like it ought to be in this room, which means 40% humidity and 65^0 Fahrenheit, that the binder system will stay in equilibrium. And in that case, all you need do is rewind the tapes occasionally because tapes relax their tension with storage. With the disks, people worried a great deal, with the onset of thin metallic disks, which started in 1986 or 1987, that these thin metallic disks would corrode.

But they are overcoated with materials; I listed some of them: carbons, zirconia. And I think the general opinion is that--as Dr. Freese was saying about optical disks--thin metallic disks have only been out since 1987. I think the general opinion is that if you keep them in a reasonable environment, they will last for a long time.

A disk drive is not a hermetically sealed enclosure; it is almost hermetically sealed. There is a little pressure-equalizing hole in it somewhere which might be ten-thousandths of an inch in diameter, with a filter behind it. And that is just to stop changes in barometric pressure flexing the box.

That's all I can tell you about it.

DR. BHUSHAN: All right. Thank you.

DR. REAGOR: Good afternoon. I would like to welcome you back to the mass storage conference this afternoon. Our first speaker for the program is Mr. Harriss Robinson. Harriss has a long, distinguished career that I will share a little bit with you. He received his B.A. Degree from the University of Illinois at Urbana/ Champaign in 1941, serving in the U.S. Army Signal Corps during World War II, having graduated from the Command and General Staff School in 1944, and then going on to the University of Chicago School of Business and Marketing.

Mr. Robinson has held executive positions in Motorola, RCA, Westex, User, Kraft, Aerojet Electrosystems and Datatape. With Datatape for the last 11 years, he has managed several of their major magnetic tape mass storage projects.

Today, he is with us to share information on magnetic tapes. Please welcome Harriss.

DR. REAGOR: Thank you very much, Mr. Robinson. Do we have any questions from the audience?

(No response)

DR. REAGOR: Okay.

MR. ROBINSON: What? No questions? Thank you.

PARTICIPANT: Actually, may I ask you one question?

DR. REAGOR: Here we go.

PARTICIPANT: (Inaudible)

MR. ROBINSON: No. D-2 is a video tape, not a Datatape.

PARTICIPANT: And R-90 is a visual --

MR. ROBINSON: R-90 is a digital data tape which they want to call a DD-2, digital data 2, as opposed to video digital recorder. There's a difference between a video recorder and a digital data recorder; and Ampex is taking their video digital recorder and making a data recorder out of it. And I'm sure they'll do a good job with it.

DR. MILLER: I'm Stephen Miller from SRI. (Inaudible)

MR. ROBINSON: Yes.

DR. MILLER: In the digital DD-2 or DD-1, you have to get up to error rates that are usable in computers -- (inaudible)

MR. ROBINSON: Yes. Let me talk about that. Most of our companies have a problem. That is, in order to introduce a new machine and get it on the market in a reasonable period of time at a reasonable cost--only millions, not hundreds of millions--we generally use broadcast video recorders as our starting point for a laboratory machine.

So, the video recorders that have been available in our company are the D-1 video and the D-2 video recorders, both of which use error concealment in their processing. So, they may make a fix out of place once in a while, but you can't see it because your eye integrates the whole picture. You would never know there is a problem there.

Now, we can't use error concealment in computer machines; we have to do the best error correcting we can, and then we have to do read after write and rewrite in order to get very good data.

Now, in the case of the D-1 video machine, it had a composite video; they broke up the scanning track. I mean, a component video; they broke up the scanning track. So, there were four audio tracks and two video tracks on the same path, each slot.

Well, we changed that in the ID-1 Committee to a continuous track on the ID-1; and we introduced Reed-Solomon error correction so that we could handle computer data. In the case of the D-2 video recorder by Ampex, it's a composite digital machine with error concealment.

Now, Ampex has modified that machine to incorporate their digital data electronics. They put in Reed-Solomon encoding and decoding, and they have done other things to make it a digital machine. There are probably three levels of error correction--RS 4/5 and something else.

Our only difference with Ampex is that it's just their machine; they are the only one that makes it. And if I read Dr. Mallinson's writings earlier, the biggest problems with obsolescence in media is to have a machine to play it on.

So, one of the ways you get that is to have several manufacturers of the same machine to the same standards.

DR. REAGOR: Thank you very much. You had mentioned earlier about users that may have large data systems on file. The company I work for, Bell Communications Research, which is the research arm for the Bell regional companies, basically probably has one of the largest data files in existence in terms of every telephone call that has been made in this country--the to, the from, the date, the time--are stored for ten years minimally. And all that is on magnetic tape right now.

MR. ROBINSON: I'll be around to see you.

(Laughter)

DR. REAGOR: It's many, many hundreds of thousands. I'll get into that tomorrow.

Our next speaker is Ms. Carole Hogan. She's the Director of Technical Development for the Distributed Computing Solutions Division of General Automics. Before joining DISCOS, Ms. Hogan was employed at the Lawrence Livermore National Laboratories, where she directed the original development of a portion of the software that became Unitree Storage Systems.

In addition to her career within the computing environment, she is also a member of the State Bar of California and practices law. I find it interesting that the title of her talk is "Storage Management and File Systems." From a lawyer's standpoint, that is one of the major things we have in our firm in terms of the volume of data that we want to access from time to time.

Carole, would you like to join me?

MS. HOGAN: I think the only thing that was missing from the introduction is the fact that the title is actually Storage Systems Solutions.

(Showing of viewgraphs)

MS. HOGAN: There are several. Maybe I should say that before I answer questions, someone said we could have a break until 2:00 if there were no questions.

PARTICIPANT: (Inaudible)

MS. HOGAN: No, I'm talking about file data as well as directories. The storage system should provide the capability to do what a site wants it to do. If you want to back up copies of your files, then, for example as an implementation choice, provide the interface to the user to write multiple copies of the same file to the storage system and guarantee that those separate copies reside on separate media.

And I don't mean disk versus tape; I mean on different tapes.

MR. MILLER: If that is a high-speed data path, why not have the control -- (inaudible)

MS. HOGAN: You can. You can. It may be a cost choice at a particular site. They may choose, because they have got an existing Ethernet and because the bandwidth of their high speed channel is not that wide, to simply send the control over the Ethernet.

MR. MILLER: So, again, you're saying that the software will allow you to do it either way?

MS. HOGAN: The software will allow you to do it either way. The software should be configurable to your site in that way.

MR. MILLER: Is it necessary to come through all this or can you redirect around --

MS. HOGAN: That's an implementation choice. I've had discussions with a lot of users on that issue. Many prefer--in fact, some prefer--that if the data is actually on on-line tape, they want it cached directly to the highest layer--for example, magnetic disk in the examples I have been using--to bypass the intermediate layers and to just land on the disk cache, and then to go from there to migrate back down through the layers.

But again, that's an implementation choice.

PARTICIPANT: (Inaudible)

MS. HOGAN: Yes. Somebody today does.

PARTICIPANT: (Inaudible)

MS. HOGAN: Is there someone in the audience who will admit to providing a network-attached disk array?

(No response)

MS. HOGAN: I guess not.

(Pause)

MS. HOGAN: Another question?

PARTICIPANT: That will be available?

MS. HOGAN: That will be available. There was another question?

PARTICIPANT: (Inaudible)

MS. HOGAN: It's coming to a hard --

(Laughter)

MS. HOGAN: I'm not trying to be cute. Other questions?

PARTICIPANT: (Inaudible)

MS. HOGAN: Yes. The question was: Is transparency important to users as they use the storage system? Yes, transparency is. There are several types of transparency. There is location transparency, for example. Users generally should not need to know--although many want to know--on which layer in the hierarchy their file may be located at any given moment.

That information should be transparent, that is, invisible to them. There's such a thing as naming transparency, that a file is given the same name and accessed by the same name, regardless of what point in the network you name the file.

Performance transparency. We talked about performance from the storage system approximating that of local disk. If it could equal that of local disk, then you would have performance transparency. Users would not know that the file was stored remotely. We don't have that, not today.

So, if users are today accessing files that are stored not on their local disk but on a remote storage facility, they will know it because they will see a lag time in accessing that file, in reading the file, from storage. So, performance is not transparent; but yes, transparency is a storage need today.

But the concept of transparency--and I didn't mean to denigrate it in my talk--is incorporated by inference in several of the storage needs of a particular need identified, such as performance. Other questions?

PARTICIPANT: (Inaudible)

MS. HOGAN: That you won't want to?

PARTICIPANT: (Inaudible)

MS. HOGAN: Many users request the ability to control where their data enters the system, how fast it migrates down. Generally, that is not a good idea. Generally, you don't want users controlling your storage resources; but you may want them to influence the storage system's migration policies.

So, for example, if you write a file to the system--you are storing it on the system--you may want to send an option in that says: I'm not going to use this for a while. The storage system then recognizes that and may migrate that file first. Right?

But that should be a choice of the storage system because you want the system to control allocation of resources, allocation of space because it does it better than a collection of users with competing interests trying to do the same thing.

But you do want to give them the ability to influence. Did you have an example that that didn't cover? (No response)

MS. HOGAN: Okay.

PARTICIPANT: (Inaudible)

MS. HOGAN: You don't want to pay for disk storage --

PARTICIPANT: (Inaudible)

MS. HOGAN: Not everybody is in agreement, though. The gentleman right behind you is shaking his head vigorously no. Do you want to speak to that?

PARTICIPANT: (Inaudible)

MS. HOGAN: Yes.

PARTICIPANT: (Inaudible)

MS. HOGAN: The question is: If a single file spans, for example, disk partitions and one of the partitions goes down and you lose that partition, does that affect the reliability of the system?

It depends on whether you are operating in standby mode, that we talked about, and you have a redundant copy--you have a standby partition with a redundant copy of the data, and it can come up, or if you are using an array and you don't have to worry about that. But if you are not, if you don't incorporate redundancy, then you have lost data.

PARTICIPANT: (Inaudible)

MS. HOGAN: But that is an implementation choice. Right. A site may not have enough money to provide redundant copies, redundant processors, redundant storage media, in which case they risk data loss. Right.

But that is not dictated to them by the storage system software. The storage system software gives them the flexibility for the redundancy; it is up to them or their budget whether they can take advantage of it.

PARTICIPANT: (Inaudible)

MS. HOGAN: Compression? Data compression?

PARTICIPANT: (Inaudible)

MS. HOGAN: Yes.

PARTICIPANT: (Inaudible)

MS. HOGAN: That is a service certainly that a storage system should provide--the ability to compact or repack tape. Whether you want to provide the service, for example, to compact disk, if magnetic disk is your top layer, is arguable because frankly your top layer is used as a cache. It is not a permanent storage media.

So, files are migrated off, purged off; and new files take their place. So, compaction is not really an issue. However, repacking would be with tape, and that's a service that a storage system should provide. Certainly, it is a need that users have.

MS. HOGAN: One last question?

PARTICIPANT: (Inaudible)

MS. HOGAN: They have a layered approach with respect to, for example, naming. The question is: Do conventional file systems have the concept of a layered approach to data management?

They do, of course, with respect to, for example, naming. Right? At the highest layer, the highest extraction for naming is a directory path name. But that is not how the file system knows the resource.

A path name is mapped, for example in the Unix system, to an inode number. Right? But that is not how the disk management system knows a file, not as an inode, but as a series of physical blocks.

So, if you are going to look at it in that way, that's a mapping, a layered naming mechanism, yes. But to store data hierarchically on multiple media, that is not a concept that is known to a conventional file system.

DR. REAGOR: Thank you very much, Carole.

(Applause)

DR. REAGOR: With the concurrence of Dr. Moore, we thought that, rather than hold you here until 4:00 o'clock to take a little break, we might take the 15 minute break now. It has been about two hours since lunch; we probably all need a cup of coffee or a drink.

So, if you could join me back here in about 15 minutes; and we will start with Dr. Moore's presentation.

(Whereupon, at 3:05 p.m., the conference was recessed.)

DR. REAGOR: Okay. As we continue this afternoon, I would like to introduce our next speaker, Dr. Reagan Moore.

Dr. Moore is currently the Manager of Programming and Software Services at the San Diego Supercomputing Center. There, he is responsible for all software systems.

Dr. Moore received his Ph.D. in Plasma Physics from the University of California-San Diego in 1977 and followed that by working as a computational plasma physicist at General Atomics for the next ten years.

In 1985, Dr. Moore joined the staff of San Diego Supercomputer Center, and there his fields of interest have included supercomputer performance evaluations, tuning and software sharing.

The title of his talk today is "File Servers, Networking, and Supercomputers." Please welcome Dr. Moore.

PARTICIPANT: (Inaudible)

DR. MOORE: In this case, the user only is forced to go through the hierarchy at one point, when they go from Cray disk to the archival storage system. Once it is in the archival storage system, data is automatically migrated to the most cost-effective storage medium. So, the user, at the moment, is only in that particular system at one point.

PARTICIPANT: (Inaudible)

DR. MOORE: Very much so, yes.

Any other questions?

DR. REAGOR: When you showed the limitations for the -- (Inaudible)

DR. MOORE: The question is: What is the cause of the latency communicating from Los Alamos to San Diego? That is just the physical speed of light through a fiberoptic link on the ground. So, there are 10 milliseconds to go speed of light from Los Alamos to San Diego, another 10 milliseconds to go back.

PARTICIPANT: (Inaudible)

DR. MOORE: Okay. The limitation for the .6 megabytes per second from the Cray to archival storage across the FDDI ring is limited by the capabilities of the Amdahl 5860. If we just do native FTP file access from the Amdahl to the Cray, you can push files at that rate; and they don't go any faster.

So, if we increase the power of the CPU out there, we would increase the amount of the data rate we could sustain on the FDDI ring.

PARTICIPANT: (Inaudible)

DR. MOORE: At the moment? No.

PARTICIPANT: (Inaudible)

DR. MOORE: The data that is stored is stored by project. So, each project can either provide access to another project for data that is stored on the shelf tape; or they can keep it proprietary to their own project.

So, in practice, most of the data that goes out to shelf is data that has been written once and read never, about 60 percent.

PARTICIPANT: (Inaudible)

DR. MOORE: The question is: How long does it take to access tapes in the robot?

PARTICIPANT: How long does it take for a person to go in and take it off the shelf?

DR. MOORE: Okay. Versus accessing tape on shelf.

PARTICIPANT: Right.

DR. MOORE: The tape access in the robot is 15 seconds. The time it takes to spin that tape to the end is on the order of 2 minutes. So, if your data is at the end of the tape, the retrieval time is dominated by the time to spin the tape.

The average time for a person to go get a tape off a shelf and mount it manually is 3 minutes. So, you could add 2 minutes or half a minute or whatever for the tape spin. The real advantage of having the tape robot is the fact that there are 14,000 tape mounts per month; and if you have 95 percent of them being done by the robot, then the operators have much less work to do in mounting tapes.

PARTICIPANT: (Inaudible)

DR. MOORE: The question is basically whether it is better to have a supercomputer or a lot of work stations. The answer is, in our case, the applications that are being run are generating more data than I can store on a work station.

So, the environment that we are providing is both the I/O environment as well as the CPU power. Then, if you go to a more powerful supercomputer, the question is: What are you not doing if you can no longer support the I/O requirements of the applications?

So, if you run a 100 gigaflop supercomputer and say that there really is no way to store the I/O from it, you are throwing away the utility of the supercomputer at that point. And that's a major challenge. A 100 gigaflop supercomputer only makes sense if you can really store the I/O that it generates.

PARTICIPANT: (Inaudible)

DR. MOORE: The question is: If you used optical disk, could you get better packing of information than the 60 percent we currently see on the archival tape, both shelf tape and robot tape? It's possible.

The question then is whether the optical disk is only two-thirds the cost or whatever of the archival storage tape. We have been trying to go for the most cost- effective storage media.

DR. REAGOR: Thank you very much.

DR. REAGOR: Our next speaker is Mr. Bernard T. O'Lear. He is the Manager of the Scientific Computing Division, Systems Department, at the National Center for Atmospheric Research. He has been with NCAR since 1964; and in 1989 and 1990, he was a member of the Congressional Office of Technology Assessment/Advisory Panel on Information Technology and Research.

Mr. O'Lear is currently a member of the Advisory Panel for the National Center for Supercomputing Applications. Mr. O'Lear?

DR. REAGOR: I will comment on his last slide, that that is how the telephone network works in terms of individual users simultaneously accessing a central switch. And there is work going on in Bellcore with some other consortiums to develop optical switching that can help to enhance computing.

MR. O'LEAR: Don't forget these. There are a few up here. Thank you.

DR. REAGOR: Okay. Are there any questions?

PARTICIPANT: What's the name of that bar again?

(Laughter)

MR. O'LEAR: It's called the mass storage bar.

DR. REAGOR: You did have a question?

PARTICIPANT: (Inaudible)

DR. REAGOR: Bernie, can you repeat the question?

MR. O'LEAR: Okay. How is Data ooze implemented at NCAR? It's implemented at each one of the points of the hierarchy. We have 120 gigabytes; the 3380 is our equivalent to Reagan's archive disk. So, we have programs that we run inside of the Cray which are what we call scrubbers.

So, it starts up at the Cray disk level; and the data is aged off there by the following criteria. There are five partitions on the 3380: 120 gigabytes for datasets that are less than 30 megabytes; and the Storage Tech robot takes everything that is greater than 30 megabytes--and we had to do that because of robot arm latency versus data transmission time.

So, we had to have datasets greater than 30 megabytes for the six robot arms on that one. So, the data is split up there, unless you do it direct. The user has the option to do a direct send of the data off-line if he wants to, or he can choose a level in the hierarchy.

That's one thing; that's the ooze through the hierarchy based on either user option or system generated moves.

On the off shelf, there are background processes that run at night; we have several of them. When there is nothing else to do, there is a random sampler that goes out; and we don't make duplicate copies. We check everything as it crosses a media boundary.

So, as it is moving through that hierarchy, all the data is checked as it moves across, on the last two cartridges, I think, for the amount of time that we have had it in. So, that's one thing that we do.

We keep tables about data. For each cartridge, what the fill is on a cartridge if data gets off to another cartridge or whatnot. Once there is a certain threshold hit on a cartridge, then this compaction routine makes a list of cartridges and recompacts them, also rechecks them again at the same time.

So, that's why we don't do the double copying; we just check them across all the media boundaries.

PARTICIPANT: (Inaudible)

MR. O'LEAR: For instance, what happens in this case, when we got the 3480 cartridges that were 30 feet longer, all we had to do was set a switch in the system, and we knew what kind of cartridge; and we knew we could put more data on that cartridge. So, as we go to the 3490s, the same thing will happen.

You know what type of cartridge you are reading from, what kind of cartridge you are writing to; and we've got it set so it doesn't matter what kind of media it is. If we had an X that would hold 25 gigabytes, we would compact in some percentage field 3480s onto a 25 gigabyte cartridge; and all the logic is in there to do that.

PARTICIPANT: (Inaudible)

MR. O'LEAR: Yes, that's been growing. It started about--here again, that's a function of the type of Cray you've got in or whatever; and those datasets represent the average dataset size. And if I had that talk with me, what the curve looks like is a whole bunch of little datasets, and then the average is about 26.2.

And then there are some very large datasets in the 7 gigabyte range out at the other end of the curve; and those are spread across several cartridges.

PARTICIPANT: (Inaudible)

MR. O'LEAR: We didn't really feel there was a viable one at this time, other than that one.

PARTICIPANT: (Inaudible)

MR. O'LEAR: There's about 17 percent of it that is-- You might think about this: there is another whole group outside of the computing group which is a hybrid scientist data manager person. These guys are meteorologists, data manager people; and they are the people who worry about the real archive of the system.

That seems to stay pretty constant unless we get some influx of data like we did from Bill Callicott's shop; I saw him here a little while ago. When they shut down the Ampex system for the National Environmental Satellite System, we imported about 10 terabytes of data, I think, from them. And then, we redistributed it on 3480 cartridges. So, when we move a dataset around in this community, you know it might be 10,000 cartridges or something. So, that's kind of a measure of it.

We call it WORN data--write only/read never. That's the stuff that you try to have the aging on for someone to get rid of; but since it's user set, that's almost all the rest of the data, as near as I could tell. And the access rate on that is very low.

So, what we track is the number of users and the rate of growth per month and report that to the division heads; and you'll see rates of growth-- I'll give you an example. When we dedicated the Cray to a user who was working on a climate model in December, he generated 1,000 3480 cartridges just solving a cloud problem; and those are packed 90 percent full.

It's the same thing with this dedicated Cray. There is going to be a lot of data generated there; whether or not it is any good when we go back and look at it, I'm not sure.

PARTICIPANT: (Inaudible)

MR. O'LEAR: Yes, that's what that special group does. They worry about the cataloging; they have another whole catalog of that kind of data.

Any special data that they get, for instance, like Russian River data, the Kuwaiti oil fire data, and that sort of thing will be available to researchers.

PARTICIPANT: (Inaudible)

MR. O'LEAR: Not that I know of.

DR. REAGOR: Thank you very much. We are now at the final speaker for the day. Let me introduce him.

Ken Thibodeau is currently the Director of the Center for Electronic Records at the National Archives and Records Administration; and prior to that, he was the Chief of the Records Management Branch of the National Institute of Health.

Ken earned his Ph.D. in History and Sociology at the University of Pennsylvania and then has held fellowships in history and computer science at the University of Strasburg, France, the University of Kansas, and the Newbury Library in Chicago.

Today, he will be speaking to us on the National Archives and Records Administration. Ken?

PARTICIPANT: (Inaudible)

DR. THIBODEAU: The question has to do with how we identify and even know in reality what the internal formats of data from different fields are. That is a tremendous drain on resources; and in spite of all the good things we have heard today about what the technology can do with this, the biggest problem we have is one that was labeled in the recent Spatial Data Transfer Standard that was proposed by NIST as "truth in labeling."

You can't really second-guess the rest of the Federal Government on what the data should be; but you can at least ask anyone who is creating data and transferring it to anyone else about what their standards of accuracy and precision were, what the internal organization of the data was, what the encoding was, what the file structure or database structures were; and that's an immense problem.

We are operating in wonderland right now. We have a regulation on the books that says: If we told you we want the data, you have got to give it to us in a flat file.

There's a fundamental problem for us, as well as for the agencies, with that in that we don't preserve data. We preserve the records of the Federal Government, which means we preserve what the original agency has created true to the way they organized it, on the grounds that if we reorganize it, we are subtracting value from the information. We are potentially making it useless. We have a regulation that says you, the agency, reorganize it before you give it to us; reorganize it in a very simple file format. We are going to change that very soon. What we are working on right now is building on the strengths of the robustness of the relational database model in spite of what the theorists say that none of the standards or none of the real products actually embody the model.

It is a very nice, simple model, which is around, is fairly simple. So, we are developing a generalized utility which at least allows us to deal with rectangular--databases which consist of rectangular files. Unfortunately, we are discovering that not all of our flat files are amenable to rectangular formatting; but we think we can reformat most of the problems we have.

Quality control is something we do. We don't know what it is going to be until we get it. Even when we make the decision we want it, even if we get extensive documentation from the agency about how it's organized and coded, there tends to be a disparity between that and when it comes in, no matter what the time frame is.

There aren't many systems that survive very long without being changed. Probably the worst one we have seen is we got one dataset in--or one file, actually--that has a 1 X 6 field. And the agency told us what the codes are, except that there are five different codes in that field; we only know what two of them mean.

(Laughter)

DR. THIBODEAU: The others-- But everything we get, we go over with a fine-toothed comb so that we know what it's supposed to be, that the documentation is there at rudimentary levels to allow you to understand what it is supposed to be, and that the data actually matches the documentation.

But over the years, we have seen every possible variant, from data that comes in with no documentation whatsoever to data that comes in with beautiful documentation that doesn't match the data.

In fact, one epidemiological document that I had delivered from NIH down to the Archives, they had paid a contractor an extra $12,000 to get the documentation in shape; and it was beautiful. And the tapes were absolutely blank.

(Laughter)

DR. THIBODEAU: Last year, we had an agency's major on-line database that it uses for its major mission programs come in, very simple file structure--a flat file-- a 5,000 character record. Except when we got it, there were an extra 16 bytes in each record that weren't supposed to be there. And in spite of the numbers you have been hearing today about gigabytes and terabytes, I can assure you that 16 digits is a lot; and this had to do with the expenditure of -- dollars. So, we really hounded the agency about where those 16 extra bytes came from and what they were supposed to be.

And it was kind of embarrassing for both of us because we were over a year going back and forth to them, first denying there were an extra 16 bytes to, well, it's there, and we'll have to figure out what it is, to finally: It was a mistake in the copy job.

You know, it's incredible how many people make mistakes in running simple copy jobs.

The only thing you can do for that ultimately is to examine everything you get and examine the documentation. What the relational system will allow us to do is basically translate documentation into metadata--this is true; we do this. We take anything we can bring in, tell the computer, and even though you can read it, assume nothing. Dump it to paper.

I'll have a human sit there with a ruler and compare it to the layout and the code structure; but a relational database will allow us to just load it into the database, and it will tell us what's wrong with the data.

But it is a very big problem, and it is going to get worse when you get into object-oriented databases and GIS and stuff like that.

PARTICIPANT: (Inaudible)

DR. THIBODEAU: Yes. We're actually pretty lucky because we have regulatory authority. So, we have a regulation out that says if we want that data, you give it to us on a tape.

We are going to change that regulation very soon to allow 3480, and we are also actually bringing stuff in right now on 3480 cartridge and on CD-ROM. And within the next year, we'll go out to expand that within practical limits.

Our basic answer is: Our records are not like paper; you don't have to wait to the end of the life cycle. So, we really try to get the agency to give us a copy as soon as it is in a good state so that we don't really have this problem of obsolescence.

Our biggest problem with the media right now is that I have 1,600 rolls of microfilm images of -- cards, which is the only database of the military service personnel from World War II. And with the fiftieth anniversary going on, a lot of people would like to have that database accessible on computer.

We actually have a pattern recognition specialist up at Penn State who has the grad students figure out how to read it back into the computer. The problem now is to get someone to build us a machine that will do that; we have the algorithm.

We are still reading punched cards in. We are still reading 800 BPI tape. This year, we managed to find someone with a seven-track tape drive. So, media has not really been a big problem for us because, when we get it, we immediately copy it over onto tape that we have certified as being physically sound.

And over time, we do data ooze, that is, over ten years, we are going to copy it onto current generation media.

PARTICIPANT: Do you have an estimate of the size of your --

DR. THIBODEAU: The total right now is about 125 gigabytes of unique tape.

PARTICIPANT: (Inaudible)

DR. THIBODEAU: Yes. In fact, everything I've got right now is on open reel, in spite of everything bad one says about it --

PARTICIPANT: (Inaudible)

DR. THIBODEAU: Okay. The question was basically the issue; it is not a media issue. It's how long you can count on accessibility of drives to read any media; and what do you do then to maintain the readability of the data? Tapes have been fairly standardized, even though they are extremely fragile; you know you have to keep them around, and we are not losing data because of fragility of tape.

You know, sometimes you think you've seen it all. This year we actually had returned to us two reels that were rewound inside out.

(Laughter)

DR. THIBODEAU: You see, you never really know; but unfortunately, tape isn't going to be around that much longer. It is no longer the universal transfer medium. We get demands to output on 3480, to output onto -- , optical and CD-ROM.

Within the next year, I intend--I have to be a little hedgy about this because there is an RFP coming out; it's actually an RFI--I intend to hedge my bets by going to 3480 and optical. And my concern there is simply, you know: Find me a media that is cost-effective over a time span that I can count on that technology being viable.

And right now, it looks to me like if I can count on that stuff for between five and ten years, it will be cost-effective. I have some fond hopes and really believe it will happen; and I have some fond hopes that after a millennium, we will have a solid-state off-line memory. There is some research going on there. It looks interesting. If it ever happens, it will solve the problem.

But as long as the densities get denser and the retrieval times get faster, every time I have to migrate to a new medium, it is cheaper to do that. And not only is it cheaper, but then I can deliver it to people much more cheaply. So, I don't really mind having to copy everything over within a five to ten-year time frame, as long as we can do it cost effectively.

DR. REAGOR: Thank you very much.

DR. THIBODEAU: You're welcome.

(Applause)

DR. FREESE: Good morning, ladies and gentlemen. My name is Bob Freese, and I'll be this morning's chair.

I would like to welcome everybody to the Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications.

I have a couple of announcements first. I would remind everybody that if you have not yet submitted your paper, there will be a conference proceedings published; and please submit your paper before you leave.

I would also like to remind everybody that tonight there is a poster session and reception at the GEWA Recreation Center, right across the street. I understand there is a little map in your handout giving you directions.

We have a full lineup this morning. Our first speaker is Patric Savage from Shell Development Company. Mr. Savage is a research consultant at Shell Development Company. He has a B.A. in Mathematics from Rice University in 1952, and he has pretty much been in computing since that time.

He led Hughes' pioneering efforts in the use of computers for inventory management, production control, shop scheduling, etcetera. And following a brief stint in the aerospace industry in 1962, he joined IBM and then Shell in 1965.

He has been very active in and has promoted the development and use of array processors at Shell and has led the design development for the Univac array processing system, which is capable of some 120 megaflops.

More recently, since 1980, he has been active in parallel and distributed computing systems; and for the past year, he has been regularly involved in the HIPPI and fiber channels standards working groups.

Mr. Savage is a member of several different societies and committees. He holds a life membership in the Sigma Chi Society for Scientific Research, and he chairs the Storage and Archiving Standards Subcommittee for the Society of Exploration Geophysicists.

Without any further ado, Mr. Savage.

Our next speaker is Adrian Abineri from APTEC Computer Systems. Mr. Abineri is the local representative for APTEC and has been designing and selling real-time computing solutions for about the past 15 years. And his presentation is going to include details associated with the NASA Goddard space telescope data archive and distribution system.

DR. FREESE: Are there questions?

PARTICIPANT: (Inaudible)

DR. FREESE: The question is: Does the system bus have parity on it?

MR. ABINERI: Dr. Chen, do you want to address that?

DR. CHEN: The system bus does not have a parity on it, but -- (inaudible)

MR. ABINERI: (Inaudible) The bus itself does not have parity on it. The bus is actually comprised of three 32 bit buses--a READ bus, a WRITE bus, and a -- (inaudible) (Inaudible discussion)

DR. FREESE: As a follow-up question to that: Why don't you use --

MR. ABINERI: Why don't we use a more elaborate error-correcting scheme on the bus?

(Inaudible discussion)

DR. FREESE: Another question?

PARTICIPANT: (Inaudible)

MR. ABINERI: (Inaudible)

DR. FREESE: The question was: How do we communicate between I/O processors and, for example, the HIPPI adapters thattalk to disks?

MR. ABINERI: Okay. What we did was we took Vx Works for those -- Again, it's designed to build target software for real-time systems that would typically be single processor systems. We implemented a network interface for Vx Works based around our bus. Okay?

And then, we went beyond the standard Vx Works/ Unix protocols of using things like sockets and so forth-- internal. We built a set of multiprocessor services which allow you to communicate between I/O processors using things like flags and command ports, flags that exist in shared memory, or semiports in shared memory, or command ports that allow you to send messages between I/O processors, to coordinate the interaction between I/O processors.

We are typically using buffering schemes, where we are bringing data into shared memory and when a buffer is filled, setting a flag, and kicking off another I/O transfer to another device or another I/O processor.

Now, that set of multiprocessor services is, if you will, our added value to Vx Works because typically it's aimed at a kind of single processor implementation.

PARTICIPANT: (Inaudible)

MR. ABINERI: That's correct. That's in fact part of what Vx Works is; it includes the interface between the host and the development of the target systems.

The Vx Works looks to the user like a Unix system; it provides many of the same system calls and user interface calls that Unix itself provides. It's just that the underlying structure is a priority driven scheduler, with fast interrupt times and fast context switch times and things of that sort.

DR. FREESE: Next question?

PARTICIPANT: (Inaudible)

MR. ABINERI: No. Chen, did you want to address that?

DR. CHEN: (Inaudible)

MR. ABINERI: That's the IPI-3 framing protocol for supporting a disk connection.

DR. FREESE: Does that address your question now?

PARTICIPANT: Yes.

DR. FREESE: Additional questions?

PARTICIPANT: (Inaudible)

MR. ABINERI: Yes, that's correct.

DR. FREESE: Let me just repeat that. Could you clarify the use of D-1 and D-2 on your overheads?

MR. ABINERI: Yes. That was my error. It is the data version.

PARTICIPANT: (Inaudible)

MR. ABINERI: Chen, do you want to address that?

DR. FREESE: I think the question was: Are you active in the POSIX --

DR. CHEN: (Inaudible)

MR. ABINERI: In case everybody didn't hear that. The Wind River Systems is the owners and licensees of Vx Works. So, we are in some ways dependent on them to achieve POSIX compliance.

DR. FREESE: Question?

PARTICIPANT: (Inaudible)

DR. FREESE: Let me repeat the question: Do you have any plans for fiber channel? You didn't talk about fiber channel in your discussion.

MR. ABINERI: I think that our approach as a vendor, at least to this point, has been just to provide the HIPPI connections. And what we perhaps would expect industry to start doing is have fiber adapters that connect to HIPPI channels. Go ahead, Chen.

DR. CHEN: (Inaudible) Since we are on the HIPPI Standards Committee and we are also now participating on the fiber channel in the HIPPI -- So, we will be following that when it comes up.

And we do have the capability of doing that available through our -- 50 in the back side. We have just not been driven by the customers to say we ought to have this by now.

DR. FREESE: Any additional questions?

(No response)

DR. FREESE: If not, I would like to thank our morning speakers.

(Applause)

DR. FREESE: Let's take a coffee and doughnut break and reconvene at 11:00 a.m.
(Whereupon, at 10:45 a.m., the conference was recessed.)

AFTER RECESS

(11:05 a.m.)

DR. FREESE: May we have everybody's attention, please? Let's start the mid-morning session.

Before we begin the next series of papers, I have two announcements.

The first announcement is that we will generate from this conference an attendee list; and you are all probably already on this list. But if you want to just double-check and make sure your name is spelled correctly and that sort of thing, you might want to double-check on the attendee list on the table right outside the conference room.

The second announcement is a reminder that this evening there is a poster session and reception at the GEWA Recreation Center. Directions to the recreation center are in the handouts, and we hope everybody can attend that meeting.

Our next paper is entitled "EMASS: An Expandable Solution for NASA Space Data Store Needs." This paper will be presented by Anthony Peterson of E-Systems. Anthony has a Bachelor's Degree in Mathematics and a Master's Degree in Mathematics from North Texas State University. Anthony joined E-Systems fourteen years ago; and most recently, in the past three years, he has been the Chief Software Engineer for the EMASS Project.

Anthony is a member of the IEEE Mass Storage Working Group. Anthony?

(Applause)

DR. FREESE: We have time for a couple questions.

MR. SAVAGE: Since you are running all this on a Convex and since Convex uses a VME bus to -- (inaudible)

DR. FREESE: Let me repeat the question: Since you are on a Convex and Convex uses VME buses, how can you guarantee data integrity?

MR. PETERSON: Okay.

DR. FREESE: And infinite life, too.

MR. PETERSON: Yes. I am not a hardware person. You are addressing the question to one who has a definite software background. I'll try to field it.

As I understand the question, the Convex uses a VME for its backplane And if I recall correctly, on the connector there may be one small place that is vulnerable, that doesn't have parity against it.

PARTICIPANT: If someone is here from Convex, perhaps they could respond.

DR. FREESE: Would you like to take a shot at the answer?

MR. COLLINS: Yes.

DR. FREESE: Could you identify yourself, and come down here and speak into the microphone?

MR. COLLINS: My name is Kirby Collins; I'm from Convex. The interfaces that they are using are not VME based. The tape interface to the R-90 recorder is an IPI-3 based interface that we built basically specifically for that recorder.

The interface to the storage technology silo is a block-mux interface. Those interfaces behind are attached to a proprietary bus called the P-bus on a Convex, which does carry parity. The disk interface is an IPI-2 based interface, which carries parity.

So, all the interfaces we are talking about for this product carry parity and are not VME-based. We do have a VME bus that is usually used for low-speed kinds of connections, terminal connections. We do have VME ESD controllers, but I don't believe E-Systems is using any of those in its product.

MR. SAVAGE: Can I then interpret what you just said? That wherever the data passes, the entire data path is at least parity protected so that it never moves around raw inside the chain?

DR. FREESE: Maybe you should repeat the question.

MR. COLLINS: The question is: Is data parity protected throughout the paths of the machine? And for all the paths described for the EMASS product, I believe that is correct.

Now, depending upon which network interfaces you are using, there may be some considerations there, but -- The network generally has other checks or protections and other kinds of things for it.

DR. FREESE: Thank you. Are there still questions?

PARTICIPANT: (Inaudible)

DR. FREESE: Is it correct that you have a single bit file server and multiple storage service?

MR. PETERSON: That would be a fair interpretation.

DR. FREESE: Question?

PARTICIPANT: (Inaudible)

DR. FREESE: The question was relative to the file migration policy when you are migrating from tape to disk and disk to tape. Is the same policy employed both ways?

MR. PETERSON: Yes. The same policy is employed, with one extra feature; and that is when the policy identifies the file to be migrated, we examine to see if the file had actually changed. If the file has not changed, then it's merely placed in a staging directory without migration again to tape.

DR. FREESE: One more question?

PARTICIPANT: (Inaudible)

DR. FREESE: The question is: Can you predict the number of retrieval cycles? What is the retrieval rate--how many per hour or any given period?

MR. PETERSON: We have run some benchmarks. In fact, we ran a very exhaustive 24-hour benchmark to demonstrate the capacity of the first system. However, I do not have those figures with me at this time.

DR. FREESE: Does anybody in the audience have those figures?

(No response)

DR. FREESE: I guess we don't have the numbers for you.

MR. PETERSON: I'm sorry.

DR. FREESE: Question?

PARTICIPANT: (Inaudible)

DR. FREESE: The question is: What is the threshold for determining degradation before you put a file on the suspect list?

MR. PETERSON: The threshold is site configurable. So, you at your site decide how much correction you wish to allow for a file before it will be placed in the suspect list.

DR. FREESE: So, you determine it yourself?

MR. PETERSON: Yes.

DR. FREESE: Thank you, Anthony.

MR. PETERSON: Thank you.

(Applause)

DR. FREESE: Our next talk is "Data Storage and Retrieval System" from the Mitre Corporation. Our speaker is Mr. Glen Nakamoto.

Mr. Nakamoto has a Bachelor's Degree in Electrical Engineering from Georgia Tech and is the principal engineer with the Mitre Corporation. He has over 22 years experience in the intelligence business, of which the first 13 years were spent at the CIA and involved the research and development of collection, exploitation, and dissemination of imagery and imagery products.

He was responsible for heading up one of the first digital image laboratories in the mid-1970s; and since then, he has been actively involved in virtually all aspects of the image exploitation process. He is currently involved in studying different architectural foundations for digital image exploitation.

Please welcome Mr. Nakamoto.

MR. NAKAMOTO: Any questions?

DR. FREESE: Are there questions or comments from the audience?

PARTICIPANT: (Inaudible)

DR. FREESE: The question is: Repeat once again how to achieve the 93.7 megabit data rate--transfer rate.

MR. NAKAMOTO: Okay. What we did on that was to transfer data from the IOC's memory--the optic computer. We figured we could get data into the memory--you know, whatever data rate any peripheral--up to the theoretical maximum of the 96. The 12 megabytes per second is the maximum that the I/O processor can handle. Once we got the data in memory, we could pretty much show that whatever we could get in, we could move out pretty much at the same rate. We were then moving the data from IOC memory into memory across a VME bus; and we were clocking transfer of about --

What we actually did was we transferred the same byte of data across this five foot, the same data on output, and wrote it into high-speed VME memory on the VME chassis side. And then, we could verify at least the first 16 megabytes of that data.

Then thereafter, what happens is that after we write 16 megabytes, we scroll it back up again. And so, we clocked that at pretty much the 96 megabit rate.

DR. FREESE: A second question?

PARTICIPANT: (Inaudible)

DR. FREESE: Maybe you could just repeat that question.

MR. NAKAMOTO: Okay. The question was whether we expect to get 48 megabytes per second off the Sun-360. Now, that data relates to getting data from the image server to some work station. In other words, the chart that I had--let's see if I have that.

(Showing of viewgraph)

MR. NAKAMOTO: If you look at this guy here, we are talking about data moving these kinds of peripherals to these kinds of peripherals. You know, there is no Sun involved here. Okay?

The comparison was that if I had on this network a Sun workstation and a Sun here, we were getting these kinds of transfer rates. But if I had a device here and I am now moving it out to a workstation here, what we are hoping to achieve there is the 48, not constrained by this guy here, though.

So, that's why I said we couldn't find any commercial workstation that could accommodate those data rates. So, what we did was put a VME chassis with a single-board computer with a dedicated processor to do the TCP/IP processing.

We are using Vx Works, a real-time operating system, instead of the Unix so we can get away from the operating system overhead; and that is how we hope to be able to demonstrate that we can get end-to-end from a device like this to some kind of workstation that the server side of it can at least accommodate that 48 megabits. That's what our goal is.

DR. FREESE: So, you actually wouldn't expect then the Sun-360 --

MR. NAKAMOTO: No. On the Sun-3, we would probably still get the 2.4.

(Laughter)

MR. NAKAMOTO: It can only go as fast as the weakest link.

PARTICIPANT: (Inaudible)

DR. FREESE: Right.

MR. NAKAMOTO: Ah, okay.

PARTICIPANT: (Inaudible)

MR. NAKAMOTO: Yes, on Ultranet. Right.

DR. FREESE: Okay. A question?

PARTICIPANT: (Inaudible)

MR. NAKAMOTO: Right.

DR. FREESE: The question is: If you do a quick calculation, it looks like your data array would be filled up pretty quickly; and so, what is the solution to that? How do you solve that?

MR. NAKAMOTO: That's correct. Keep in mind that this was set up as sort of a feasibility prototype as to whether or not we could do something like this with off- the-shelf equipment quickly.

So, the idea was: Go out and find out what is commercially available and sort of try to slap it together as quickly as possible and see if we can get these kinds of throughputs. The system is actually expandable just in a disk array alone in a single chassis; I think you get up to about 120 some-odd gigabytes.

So, the limiting reason, or the reason why it is only at actually 7.5--and it has upgraded since--is mainly because of cost. And I would like to say that the intent here was to sort of demonstrate the feasibility.

In reality, what we are doing now is we are taking the image files that are 1 and 2 gigabytes up to 4; and we are actually compressing the data down to about 200 megabytes per image. And we are trying to do the decompression on the work station.

So, in reality, the 7.5 gigabytes wind up being quite insufficient for conducting the kinds of experiments and studies that they want to do.

DR. FREESE: Two final questions, and then I'm going to have to defer the questions until lunch time.

PARTICIPANT: (Inaudible)

DR. FREESE: Yes. Did he have to develop a lot of software for the I/O management?

MR. NAKAMOTO: Yes. That was predominantly the thrust of our effort; that is, to look at how-- You know, if you looked at that software chart where we were going from buffer to buffer to buffer to buffer, across all these buses, that was basically our contribution to getting this done.

Basically, the hardware was already put together and in many ways integrated for us. But all we had in terms of software was a device driver to read or write. So, the buffer management and the pipelining of this information is what we worked on.

DR. FREESE: So, Glen, you ended up doing an awful lot in software?

MR. NAKAMOTO: Right. This is almost all software.

DR. FREESE: Okay.

MR. NAKAMOTO: Unfortunately, it is all high level. We did not have to write a single device driver; and software on the workstation side is written such that, because we didn't know what sort of workstation we were going to have, we were told we are going to make it a memory structure. We have registers; we have a mailbox; and we have a FIFO.

And no matter what workstation you are using, if you can read and write to memory, which almost every workstation hopefully can do, then this software can be ported over virtually in a matter of days. And the idea is that we have developed a protocol to talk to this server that says, you know: Here's what I want to do; here is my command.

And it is very much like an FTP; you do a "get" or "put"; you file names. You can get directory information off of it, and it is all memory-oriented.

DR. FREESE: Patric, do you have a question?

MR. SAVAGE: I wonder if you would consider any of the -- (inaudible) or the optimistic protocols on Ethernet? We have been able to prototype transfers over Ethernet and workstation to workstation at 9 megabits per second sustained, reserving the Ethernet first and then using -- protocols. Have you looked at that at all?

MR. NAKAMOTO: Yes.

DR. FREESE: The question is: Have you looked at blast protocols for your Ethernet?

MR. NAKAMOTO: Yes. In fact, we actually implemented--not with this--but in the past, we have done something like that and achieved the same kind of data rates.

One of the things that we are looking at, even with this configuration, as we start to think about how this might fit into a facility architecture for distributing imagery data within a building, we are looking at how do we utilize the Ethernet connections in light of the existing work stations.

And to that end, one of the components that we have been studying is these Ethernet multiplexers. Okay. So, we used the existing Ethernet connection or interface in these workstations, but we dedicate that one Ethernet to the workstation; and these Ethernets connecting to this box, which go on to an FDDI. And so, we effectively get typically from 6 to maybe as high as 9 megabits per workstation without having to buy an FDDI interface. So, we have looked at that type of architecture as the means of utilizing existing investment with Ethernet equipment.

DR. FREESE: Thank you, Glen.

(Applause)

DR. FREESE: Our next talk is entitled "The Challenge of a Data Storage Hierarchy." It will be presented by Michael Ruderman from Mesa Archival Systems.

Michael has about 20 years of experience with the computer industry, starting out with the first 15 years with IBM. Most recently, he has spent a lot of time in the data archiving business and is presently Vice President of Marketing at Mesa Archival Systems.

Please welcome Michael Ruderman.

DR. FREESE: Questions, comments, discussion from the floor?

MR. SAVAGE: I do have one question. (Inaudible)

DR. FREESE: Could you paraphrase that?

MR. RUDERMAN: Yes. The question had to do with, I believe, an interpretation of the NCAR system, which was that the archive manager was really just directing the requesting system or telling it where the data was, as opposed to actually shipping it to it.

That is not my understanding of the NCAR system. The NCAR system, as I understand it--it could be that that was originally--but as I understand it, the NCAR system, the mass storage system, the software, from which we are derived, was really implemented to keep the Cray--right now they have two Crays--going and the Cray disk is so expensive.

So, maybe what you are referring to is the fact that there is direct data transfer from the IBM disks--from the IBM system--to the Cray disks; but the Cray, to my knowledge, does not access the IBM disks directly.

MR. SAVAGE: Well, it was doing that when -- (inaudible) The IBM machine would look it up, find out what disk it was, and ship that to the Cray. The Cray would then create a channel program and send it down over that channel to actually directly read the disk.

(Inaudible)

DR. FREESE: No. His question is whether you do that way.

MR. RUDERMAN: Well, our system is not oriented to doing that. I don't know if NCAR is doing that. It is not the architecture as I understand it, but they may be -- Labs are different than products, and a lab code is real different. We all know that; so, they could be doing something like that.

DR. FREESE: Any other questions or comments?

(No response)

MR. KOBLER: The chair this afternoon, of course, needs no introduction. Patric?

MR. SAVAGE: Our first speaker this afternoon is Fred Rybczynski. Fred is a technical specialist from Metrum on the VHS-based multi-terabyte archive products. He has a very long set of credentials, but he said that he would rather spend the time talking about his subject.

MR. SAVAGE: We'll maybe take a couple of questions. In the back?

PARTICIPANT: (Inaudible)

MR. RYBCZYNSKI: They load over data. The question was: Do -- cartridges load or unload data? Do they go to a protected area on the tape, or do they just load and unload right at that location?

They do the same thing your home video system does; they unload at that location; and they retension the cartridge. So, the cartridge tension is maintained at about 24 grams so that you don't have loose loops of anything going out in the flat and coming back down to protect the tape.

But they do not rewind or go to end of tape.

MR. SAVAGE: Any other questions?

PARTICIPANT: (Inaudible)

MR. RYBCZYNSKI: The question was: Does the virtual file system read and write to the tape directly, or does it stage through the disk?

We stage through a caching disk. This expedites the overall process, allows the user file transfer to occur before a cartridge is loaded; and in the case where files are less than 40 megabytes, it allows the entire file transfer process to complete when files are destined to the archive. It allows them to complete before the cartridge is even loaded, and the user experiences no system delay.

On the basis of retrieval, it is also cached through the disk because we find that in many cases, the remote system may not be able to accept the data as rapidly or the network might be loaded. So, we cache through the disk so that we can free that resource and begin addressing the next request in the queue.

MR. SAVAGE: One last question?

PARTICIPANT: (Inaudible)

MR. RYBCZYNSKI: Okay. Is your question: What is the function of the system administrator?

PARTICIPANT: No. (Inaudible)

MR. RYBCZYNSKI: Okay. The question is: Does the user require the presence of a system administrator in order to write to a tape? The answer is no.

The system administrator is used to establish the initial user account, just as you have today on any system, you need a system administrator or somebody to go in there and create

the user account and the home directory and, in any existing file, whether you want to use a C-shell or a K-shell or whatever it is that you might want to use. And that is the same function here.

The system administrator performs those functions and, if required, can make an association between the user's directory area and a specific cartridge or a set of cartridges. Once permission has been given to access the system, the system administrator is unnecessary to that user.

MR. SAVAGE: Okay.

PARTICIPANT: (Inaudible)

MR. RYBCZYNSKI: The life of the optical tape? In the interest of time, let me not answer that question, only because the presentation given right after me by ICI Image Data is going to talk explicitly about optical tape itself.

MR. SAVAGE: Thank you.

(Applause)

MR. SAVAGE: Our next speaker is Robert McLean. He is from ICI Image Data. I take it he is a Canadian, and he went overseas to the University of St. Andrews in Scotland, where he got a B.S. in Physics and Electronics in 1981. You can see he is a very young man. He then returned to Canada and went to the University of Western Ontario in London, Ontario, where he picked up an M.S. in Physics in 1983. And he did his research in atomic and molecular laser spectroscopy.

He then went back to Britain and joined ICI's Optical Data Storage Project just as it was starting up as an R&D project. He was in charge of media testing and physics during the development of ICI's flexible optical media.

He is now in the Marketing Department as a technical support manager based in Essex, which is in England--well, that was unnecessary to say, that it's in England --

(Laughter)

MR. SAVAGE: Mr. McLean?

(Applause)

MR. SAVAGE: Questions?

PARTICIPANT: (Inaudible)

MR. McLEAN: So, you are looking at shifts in what? Absorption?

PARTICIPANT: (Inaudible)

MR. McLEAN: Well, we haven't been able to detect a shift due to ultraviolet exposures. Is that your question?

PARTICIPANT: (Inaudible)

MR. McLEAN: No. The dye doesn't degrade at all under the normal reading power. We read a particular track for 30 million passes, and we couldn't detect any change in the signal at all, using either a time interval analyzer or a spectrum analyzer. So, there doesn't appear to be any degradation with the heating that goes on during the read process, and the material doesn't erase.

We were using a 1 milliwatt read beam, and the velocity of the read beam on the track was about 8 meters per second. That kind of read power gives you a very adequate read-back signal.

PARTICIPANT: (Inaudible)

MR. McLEAN: There aren't any that we have ever seen.

MR. SAVAGE: Any other questions?

PARTICIPANT: (Inaudible)

MR. McLEAN: All the tests that we did for the recording characteristics involved writing on the material after the exposure. So, one thing we want to do, now that we have a read/write channel on the Creo drive is to do the test both ways: by writing on the material and then looking for a change in error rate of written data.

MR. SAVAGE: Any other questions?

(No response)

MR. SAVAGE: I have a question--two questions. The first one is: At the National Media Lab when they were doing a lot of testing on various magnetic tapes, they used the same pollutants to make the test that you did; but the comment came up that they were very surprised to find that there was a very significant amount of ozone in the typical computer environment. And of course, ozone exacerbates most pollutants.

So, I would suggest that you join the bandwagon and add some ozone to those tests because it probably will have some effect. You know, if you can survive ozone in those pollutants, you probably have got some pretty good stuff.

MR. McLEAN: That's a good comment. We are constantly reviewing these things with our friends at Battelle, and I think personally it would be a good idea to do that, too.

MR. SAVAGE: The second question is more of a technical one in that optical disks, as you know, have a protective layer which also behaves as a defocusing layer. Your optical tape doesn't have any such defocusing layer, and I'm quite sure that a particle of cigarette smoke is as big as one of the spots that you are writing on now.

MR. McLEAN: Yes.

MR. SAVAGE: One would expect that, over a period of time, such tiny little particles would find their way loose in your reel as you use it and would appear as false spots. Would you like to comment on that?

MR. McLEAN: Yes, of course. The areas that we are writing bits on are about 1 micron by 1 micron in size. And so, if you have a 1 micron particle, it will obscure a bit; if you have a 10 micron particle, it might obscure 100 bits.

So, in that sense, we have the same vulnerability to contamination that any high-density tape will have. We do have one advantage in that we don't get any sort of a spacing loss due to these particles. So, we can write directly next to a 10 micron spot.

However, you don't want contamination; and all of our manufacturing is done in a clean-room environment, a Class 100 environment. Before we coat the recording layers, we put down a subbing or smoothing layer that is built a micron thick to fill in microscopic scratches on the surface of the base film.

The drive itself has to be designed with due regard to cleanliness. So, you want a filtered air supply going through a HEPA* filter into the drive under positive pressure.

* High Efficiency Particle

And finally, you have to pay attention to things like edge wear; you want to guide the tape so that you don't get edge wear. We are getting good results on the Creo drive as far as the bit error rate goes. The initial starting error rate of our media is about five times 10^{-5}; and the end of lifetime specification that we are working to is somewhere around 10^{-4}

So, those lifetime numbers I gave you of 100,000 passes equate to that error rate limit of 10^{-4}. At that level, you can still correct down to 10^{-12}.

In any case, once the tape is in the pack, it is pretty much self-protecting. The real worry is when you are moving the tape or when it is in the drive cabinet; you want things to be clean.

MR. SAVAGE: Can I make one more comment? That is, the pollutant testing would appear to be more severe if you would test it on tapes which had been exercised some reasonable number of times, giving the chance for the coating that you have there to flex a little bit and maybe crack and swell and so on.

MR. McLEAN: It's possible, I suppose; but in fact, a protective overcoat layer isn't a chemical barrier. We decided at the outset not to work with chemically reactive media like tellurium-based write-once media or magneto- optic based media because those require impermeable barriers, usually silicon oxide or silicon nitride-- basically, a very thin layer of glass. Our overcoat material is all highly cross-linked polymer, but it is permeable to oxygen and other gases, just as the back side of the polyester base film is. So, it's possible that flexing the tape might make some difference; but we don't have adhesion problems that are detectable.

MR. SAVAGE: One more question?

PARTICIPANT: (Inaudible)

MR. McLEAN: That's right.

PARTICIPANT: (Inaudible)

MR. McLEAN: Oh, yes. There are a lot of questions there, and --

PARTICIPANT: What is the question?

MR. McLEAN: The point was that the Battelle test has been calibrated to 30 years with the use of copper strips, I believe; that is what they actually used to do that correlation. And there is a big question there about the chemistry and whether or not the chemistry really is accelerated when you have a totally different material there like the metal particle tape or an optical tape.

As I say, these are lifetime projections; and I'm sure there is going to be a lot of research going on there to see exactly how realistic they are.

And in particular, the presence of other surfaces there provides protection. In the case of metal particle tape, the cartridge provides protection from that test. 30 years in an open environment might be very different as far as the attraction of active species like chlorine goes.

In our case, we have an open reel, but it still has flanges on it. And even those have been shown to give a degree of protection.

MR. SAVAGE: Thank you.

MR. SAVAGE: Our last speaker before the break is Fred Zeiler; he's the Manager of Marketing and Product Planning for Odetics, the ATL Products Division. Fred is a graduate of Case Western Reserve University; he got a B.S.E.E. in 1965, with a background in process control and computer sciences.

Fred has 25 years experience in assorted positions in engineering and marketing, covering computer graphics, CAD/CAM, computer-based engineering and manufacturing systems, and data-based management. He has been all over the territory.

In the late 1960s, he was instrumental in the engineering design of the first flatbed plotter. In the late 1970s, he was involved in marketing and product development of IBM mainframe-attached look-alike graphics displays.

After heading up a marketing consulting firm, he joined Odetics two years ago to develop product marketing and business strategies for ATL products. Fred?

DR. MALLINSON: Good morning. I apologize for the late start, but apparently, three inches of water deposited in an hour, or a half an hour or something, east of here, has kept some people from making it yet.

Nevertheless, to start on schedule, we had better get started.

This is the last morning session, and the first speaker is Dr. Bharat Bhushan. He has an M.S. in Mechanical Engineering from Massachusetts Institute of Technology, an M.S. in Mechanics, and a Ph.D. in Mechanical Engineering from the University of Colorado at Boulder, and an M.D.A., if you please, from Renssalaer Polytechnic.

He spent many years at IBM and is now at Ohio State University as an Ohio Eminent Scholar Professor.

Dr. Bhushan is going to talk on two subjects, which he will introduce himself.

DR. MALLINSON: We have time for a couple of questions. Any questions?

PARTICIPANT: (inaudible)

DR. MALLINSON: The question is: Are you concerned about static charge in tribology studies?

DR. BHUSHAN: Yes, we are and -- (inaudible)
Some work on this has been done and has been published in the literature. (inaudible) So, the answer is yes.

DR. MALLINSON: Any other questions?

(No response)

DR. MALLINSON: All right. Thank you very much. We will move on to the second paper.

The second paper is by Newt Perdue of Ultra Network Technologies; and Newt is the Vice President of Business Development and Co-Founder of the company called Ultra Network Technologies.

Prior to being with Ultra, Newt was in charge of the high-speed processor group at NASA Ames. There, he worked on a mass storage project that had a specification of 200 megabytes per second transfer rate. Prior to that, he was at the U.S. Navy Fleet Numerical Oceanography Center.

He has a degree in biology from the U.C. San Diego. Newt is going to talk on network issues for large mass storage requirements.

DR. MALLINSON: Questions?

PARTICIPANT: (Inaudible)

MR. PERDUE: Do you want me to repeat the question?

DR. MALLINSON: Yes.

MR. PERDUE: He asked for me to compare the approach our company is taking to the HIPPI switch approach, which our competitor Network System is going after at the moment.

In short, the way I would compare it is that the HIPPI switch is a technology. It's a technology for switching lines, and it has a long way to go before it is a system solution.

What our company is trying to do is build system solutions for networking. HIPPI switches happen to be a part of the technology that we will be incorporating in our future products, but it is just a technology.

Without the system solution that I was talking about and the ability to do protocol processing, the ability to solve talking to other networks, you won't have a real end-to-end user solution; but it is a great technology, and we intend to support it.

DR. MALLINSON: One more question?

PARTICIPANT: (Inaudible)

MR. PERDUE: He's asking if we plan to support the VAX. We have not supported the VAX to date mainly because our mission in life has been to support high performance systems; and we have not been able to find a place on the DEC machines to get high-performance I/O out. With the new 9000 series and DEC's commitment to a HIPPI, we absolutely are supporting that.

DR. MALLINSON: Thank you.

MR. PERDUE: Thank you very much.

DR. MALLINSON: We will move on to the next paper. The next paper is by Tom Gilbert; he is called a Sales Consultant for Network Systems Corporation, He has ten years experience at Network Systems, two years prior to that at Satellite Business Systems, ten years prior to that at IBM. He has a B.S. Degree from Renssalaer Polytechnic University. The title of the paper is "The Role of HIPPI Switches in Mass Storage Systems: A Five-Year Prospective." Tom?

DR. MALLINSON: We have time for a couple of questions. Any questions?

PARTICIPANT: (Inaudible)

MR. GILBERT: Okay. The question was regarding the tuning of the network for multiple size packets. Just let me comment in terms of the HIPPI standard.

The HIPPI standard does not impose an MTU limitation. In other words, when you make a connection, you can then transfer a HIPPI packet of indefinite length.

So, the only limitation is the end point systems design; and of course, it will take TCP/IP or OSI, as we noted, today, and if we layer it over top of HIPPI, we are going to have limitations based upon implementation restrictions and windowing restrictions and ultimately the 64K by IP datagram limitations.

So, we have a long way to go between where we are today and before we run out of steam even with those.

In the longer term, when you talk about new protocol variants, it can use of that; and it's conceivable you can dial the connection stream data for many megabytes without ever getting an acknowledgement. I mean, that's fundamentally feasible or possible. So, it's an end point systems.

Newt made a good point: We are systems bound today. All the media in the world isn't going to solve those problems. The good news is: The availability of the standard media is stimulating a lot of vendors to do something about it. Any more questions?

DR. MALLINSON: Any more questions?

(No response)

DR. MALLINSON: If not, thank you.

MR. GILBERT: Thank you.

DR. MALLINSON: We will have a break now and start again at five minutes past the hour.

(Whereupon, at 10:50 a.m., the conference was recessed.)

DR. MALLINSON: Welcome to the second part of the morning session. We will next hear about NASA and the various aspects of the system.

The first paper will be given by Ronald Blitstein. He has a Master of Science Degree in Computer Systems Management from the University of Maryland, 20 years experience in the computer and communications industry.

He was network design manager at Satellite Business Systems during its brief existence and has supported corporate implementation planning at MCI. He has been Manager of Operations at NSSDC--that's the National Space Science Data Center--for the past four years and directs the data restoration activities of the data center. The title of his paper is "National Aeronautics and Space Administration; National Space Science Data Center."

DR. MALLINSON: The paper is open for questions.

PARTICIPANT: I'm concerned with the scalability of your -- (Inaudible)

MR. BLITSTEIN: The question is: Are the methodologies that we are using right now with our current archive size of 110,000 tapes appropriate for the significant increase in data volumes that we are going to be acquiring from EOS?

John Berbert, in the next presentation, is going to address specific EOS requirements. One thing that must be identified is that we are one of many data centers. We have a charter to support particular types of data and also perhaps particular levels of those data products.

We do not traditionally support level zero or level one unprocessed data; those tend to be held by the investigators. The data products that we routinely get are higher level data products that have a broader range of research opportunities for individuals. And every time you distill something, you end up with less. So, those things work to reduce the volume of the data that you are talking about. But obviously, the new mass storage capabilities that I did describe and the auto- ingest research that are going on right now are crucial to enable us to support what we know are going to be higher volumes of data.

But at the same time, we also recognize that when new data comes in the door, current data might be more interesting. As soon as we decide whether we are ever going to go to the moon again or not, all the data that we have sitting around from Apollo days is going to start to be requested.

As soon as EOS goes up and they get a different resolution on ozone, then current data on ozone is going to be even more interesting than it is right now. So, we recognize the challenge identified in the question.

DR. MALLINSON: One more question?

PARTICIPANT: (Inaudible)

MR. BLITSTEIN: Again, the question was: Could I brief the audience on activities that we are doing in support of EOS version zero?

The next presentation is going to address that in detail. My view has been perhaps more historical than documenting the activities of these future missions. So, just be patient and wait for the next presentation.

PARTICIPANT: (Inaudible)

MR. BLITSTEIN: Yes.

PARTICIPANT: (Inaudible)

MR. BLITSTEIN: The question was: Do we keep track of the usage of the data? Yes. We have a database that tracks the activities down to the dataset that was requested, not necessarily what files, but down to the dataset level and the number of accesses to our electronic system for each file.

So, we are capturing some level of information about that.

DR. MALLINSON: Okay. Thank you very much. We will move on to the next paper, which is by John Berbert. He received his B.S. in Physics from Union College and a Master's in Physics from the University of Maryland.

He joined the Vanguard project as a member of the team which developed and operated the calibration system for the minitrack satellite tracking system. He became head of the Operations Evaluations Branch, responsible for calibration and evaluation of telemetry, command, and tracking systems in the space tracking and data network, which of course has as initials STADN.

He then became observation systems intercomparison principal investigator and GSFC science manager for the National Geodetic Satellite Program. He more recently was NOSS Algorithm Management team member, MAXET data manager, and for the last several years the EOSDIS study manager.

John Berbert's paper is entitled "National Aeronautics and Space Administration; Earth Observing System Data and Information System (EOSDIS)."

We will move on to the last paper, given by John Boyd--Boyd with a "B".

(Laughter)

DR. MALLINSON: John is currently a Systems Engineer for the EROS Data Center, the EDC project Office, for EOSDIS data systems. Over the last 12 years at EDC, he has worked in LANDSAT data production and archiving, AVHRR reception and processing, system integration, and color film recorder development and integration.

He has just completed procurement of a system to transcribe, verify, and index the LANDSAT TM and MSS data archives. The title of the paper is "United States Geological Survey, Earth Resource Observation System (EROS)." John?

DR. MALLINSON: Thank you. Are there any questions on this paper?

PARTICIPANT: (Inaudible)

MR. BOYD: The question was: Does the Geological Survey have a world-wide responsibility or U.S.? And we really are primarily a U.S. organization although, because of the interest in volcanology and earthquakes, we do get involved in cooperation internationally to support that.

PARTICIPANT: Can you show an example of -- (Inaudible)

MR. BOYD: Yes. The data is collected world- wide. In addition to the archive that we have, which is about 900,000 scenes, there are another 2 million scenes that are archived by the individual ground stations around the world.

DR. MALLINSON: Another question?

PARTICIPANT: (Inaudible)

MR. BOYD: Yes. You raised the question that since we have compressed from approximately 50 to 60,000 physical volumes to 1,500 volumes, those are going to have a lot higher access rate; and that is true.

Currently, at least, we only make about 50 products a week or 200 products a month. So, it is not a high activity. Now, that could change a lot with the fertilization of LANDSAT with EOS.

But I think that if we can trust that we can run a cassette perhaps 1,000 times, then I think we are going to be well below that threshold; and we won't wear out the tape on an individual cassette.

PARTICIPANT: A second question had to do with the browse thing. It is a fairly finite dataset, and I wonder if you have considered publishing it on CD ROM and selling it. That would take you basically out of the browse business and let everybody do their own browsing at their desk.

MR. BOYD: Yes. We have given it some thought. I'm sorry. The question was: Have we given some thought to reproducing the browse dataset on optical disk and getting out of the on-line browse business ourselves?

And that is a very good alternative. It would be a very large set of CD ROMs, given that we are dealing with almost a terabyte of data, and the current CD ROM holds only half a gigabyte.

But we have in the past distributed data in microfiche form on a subscription basis, so they can get started and then keep current. And we would probably like to move into that area ourselves with CD ROM in the future. We just haven't given that enough thought at the current time. But that is a very good suggestion.

DR. MALLINSON: We have time for another question.

PARTICIPANT: (Inaudible)

MR. BOYD: We are using it strictly as an interim solution; and I have tried to prepare our management for migrating to a new media within six to seven years at the outside. I would really prefer five.

I think that will give us time for the optical tape media and the helical scan, D-1/D-2, controversy to settle out and a clear leader emerge. So, we are really using it just to bridge a transition in technology.

Since these high-density tapes have been around for 15 years, there is really nothing that I know of that really has recently clearly supplanted them. The DCRSi comes the closest, and we will use that as a five-year bridge to a new media; and that may be just the first of many, many five-year bridges to new media actually.

DR. MALLINSON: I have a question for you. When you are transferring them to the DCRSi, what is going to be the disposition of the original tapes?

MR. BOYD: We are going to palletize and archive those in a controlled environment for a period of about five years. And then, that will give us a good sense of confidence in the new media; and I think we will do that with each successive media.

DR. MALLINSON: All right. One more question?

PARTICIPANT: Do you have any projections for response time and -- (Inaudible) How many days or weeks does it take to get that --

MR. BOYD: The recovery of the data from the tape out of the archive is a 15-minute task. Typically, we turn a product around in about a week. I hate to say that because I think we should be able to do it quicker; but you know, our track record is about seven calendar days to turn a product around.

But if we had a robot, that wouldn't particularly help us because we have a very disciplined archival staff and procedures; and with the data itself, we are not driven by time criticality.

The browse data, yes; we would like to have that on optical disk so that can be basically on line. Now, in the EOS program, there will be a lot more push to have that data in a robot and fully on-line. But given the financial constraints of the LANDSAT program, that is just not possible.

DR. MALLINSON: Okay. Thank you.

MR. BOYD: Thank you.

DR. MALLINSON: That is the end of this session.

(Applause)

DR. THIBODEAU: Good afternoon. Our first speaker this afternoon is Fernando Podio, who is a member of the Advanced Systems Division at the National Computer Systems Laboratory at NIST. He is the project leader of the Optical Storage Research Program, which incidentally is a program that does a lot of research for the National Archives.

This program includes stability studies of optical media and studies for improving the utilization of optical disks systems in the Federal Government. Mr. Podio is also chairman of the NIST-NASA Working Group for the Development of Test Methods and Specifications for 356 Millimeter Ruggedized Rewritable Optical Media.

He participates in the development of standards for optical disks as principal member for the NIST of the technological committee X3B11 optical digital data disks, where he is project manager of the (TC) X3B11 test methods project.

He is also technical liaison with the Joint Technical Commission on Data Permanence for Optical and Magnetic Systems. He has published several reports and given very many presentations, invited papers on the NIST optical disk research program and standards for optical computer storage media. Mr. Podio?

PARTICIPANT: Does industry need a reference material for optical media?

MR. PODIO: The first step is for industry to determine if there is a need for a reference material. Then industry has to define the characteristics of such reference material. These two are the technical aspects of the problem. The other problem consists in the funding for such a project. It is not possible to estimate the cost of a project like that until the characteristics of such reference material are defined. This is not the first time that somebody has raised the question of a reference material for optical disks. We had informal conversations in the test methods project group in TC X3B11 on the subject. But so far industry has not shown a strong interest in studying the need for those materials.

PARTICIPANT: Who is coordinating these information interchange standards? Are they cooordinated so that they retain some meaningful relationship with each other? Or does each group decide to do their own thing? I'm talking about the labeling and the file structure standards, which I think have a very strong interrelationship among all the different media.

MR. PODIO: The responsibility of TC X3B5 includes developing label and file structure standards. As you can see from the information presented today, there are national and international label and file structure standards for magnetic tape that cover different media standards. That TC has been active for many years. TC X3B11 has a subcommittee X3B11.1 that is documenting a label and file structure draft standard for write once media. A TC letter ballot is planned for this document after the October 21-25 meeting. A letter ballot for a rewritable document standard is expected after the January or April meeting. These documents would serve as umbrellas for different medai standards. That is, they will be size independent. We welcome these developments, but there still no label and file structure standards for WORM and rewritable optical disks..

PARTICIPANT: You are not answering my question.

MR. PODIO: Probably not.

MR. PODIO: Yes. Now I understand your question. Magnetic tape--the standards committee is very good at that. I don't sit on that committee, but I have sat in several times; and magnetic tape, for some specific technologies, they have a label and file structure standard that would serve as an umbrella of different media standards. And I agree with you; we were very happy about it.

It didn't happen in the last three or four years in optical media. Recently, however, there is a subcommittee under X3B11--B11.1--and there is activity that is being initiated at the international level on label and file structures.

remember the date--but the U.S. delegation is submitting an information paper with a consensus document that the subcommittee here under ANSI came out with.

That document is going to serve as an umbrella for different types of optical media technologies--not only one. So, it would be useful for 12-inch or 5 1/4 or 3 1/2. There are other proponents, though, at the international level they would like to have a label and file structure for one simple type of technology--very narrow--and we are not very happy about it, to tell you the truth.

We would like to see only one standard covering everything. But you have to go there and try to convince other people, and sometimes it is not very easy to do that.

PARTICIPANT: Let me just comment that the reason that I noticed this was I did a plot experiment and said: Okay. Suppose I have created a file on this optical disk, and I want to create a copy of that file; and I did not see that there was any way for me to take the disk information interchange standard and make a taped copy. There wasn't any place in the tape standard for a great many of the things that they have added.

MR. PODIO: Okay. I can tell you that at least there are two people that are members of that subcommittee that I know that are coming from the tape environment. So, they probably could comment on this. I can give you my business card and if you call me I can put you in contact with these people. And even if you want to attend the subcommittee meetings, you are welcome to do that, or serve as a member.

It is not necessary to be a member of B11 in order to be a member of X3B11.1. So, let me give you my business card.

PARTICIPANT: Thank you.

DR. THIBODEAU: I think to keep the program on time, we will have to move to the next speaker.

MR. PODIO: Okay.

DR. THIBODEAU: Thank you very much.

MR. PODIO: You're welcome.

(Applause)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1992	Conference Publication, July 23 - 25, 1991

4. TITLE AND SUBTITLE
NSSDC Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications
Volume III

5. FUNDING NUMBERS
JON 933-656-80-02-26

Ben Kobler, P. C. Hariharan, and L. G. Blasso, Editors

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NASA-Goddard Space Flight Center
Greenbelt, Maryland 20771

8. PERFORMING ORGANIZATION REPORT NUMBER
92B00093
Code 933

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, D.C. 20546-0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NASA CP-3165, Vol. III

11. SUPPLEMENTARY NOTES
Kobler: Goddard Space Flight Center, Greenbelt, MD;
Hariharan and Blasso: STX Corporation*, Lanham, MD.
*Hughes STX Corporation as of October 1, 1991

12a. DISTRIBUTION/AVAILABILITY STATEMENT
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 NSSDC Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications. This conference served as a broad forum for the discussion of a number of important issues in the field of mass storage systems. Topics include magnetic disk and tape technologies, optical disk and tape, software storage and file management systems, and experiences with the use of a large, distributed storage system. The technical presentations describe, among other things, integrated mass storage systems that are expected to be available commercially. Also included is a series of presentations from Federal Government organizations and research institutions covering their mass storage requirements for the 1990s.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Magnetic tape, magnetic disk, optical disk, mass storage, software storage	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

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18, 298-102

*U.S. GOVERNMENT PRINTING OFFICE: 1992 - 627 - 064/66003 REGION NO. 4