SEVENTH IEEE SYMPOSIUM ON MASS STORAGE SYSTEMS

Curtis V. Canada

Fermilab

Introduction

This paper is a somewhat overdue trip report from the IEEE Symposium on Mass Storage Systems held November 4 through 7, 1985 in Tucson. It is also a submission for the Workshop on Triggering, Data Acquisition and Computing for High Energy/High Luminosity Hadron-Hadron Colliders. (Please forgive the length of this paper. I didn't have time to make it short. Readers may want to skim most of the paper and read the sections headed 'IBM 3840 Technology' and 'A Rotary Digital Cassette Recorder for MSS'.) The symposium was divided into a special tutorial session entitled 'Site Architecture and Mass Storage Systems: A Comparative Analysis' and a series of invited sessions grouped under the title 'Toward Automated Mass Storage Systems.'

Tutorial Notes

The tutorial was directed towards information systems directors, capacity planners and storage managers who are concerned about the increasing demands placed on storage and retrieval of digital data at computing centers by the scientific computing community and modern business. It was said (by the organizers) to be 'the most comprehensive, comparative analysis of mass storage systems yet presented to a public audience'.

The tutorial began with a presentation of the Mass Storage Systems (MSS) Reference Model as a basis for comparing mass storage systems and plans. Immediately following that were descriptions of four widely differing mass storage systems, including two from government agencies and two from industry. They were 1) the Los Alamos Common File System - a very centralized system, 2) the Lawrence Livermore National Laboratory's planned distributed filing system - a distributed hierarchy of file servers, 3) the Shell Development Company's Central Computing Facility - a large, seismic database, and 4) the IBM mass storage system in use at Travelers Insurance Corporation. A presentation on related IBM products was slipped into the schedule before the description of the Travelers' system to get the tutorial participants partially acclamated to IBM abbreviations. Finally, the tutorial was concluded with a panel discussion and question/answer period.

MSS Reference Model

The MSS Reference Model had its genesis in an Executive Committee meeting of the IEEE Technical Committee on Mass Storage Systems and Technology. Having observed the utility of the Open Systems Interconnect (OSI) Reference Model in the data communications field, they desired to provide a similar service in the MSS field. Special workshops were held over the last two years resulting in the model described in the July 1985 issue of Computer magazine. The model aims to define modules needed by many sites thus speeding development and broadening the market.

The talk on the MSS Reference Model ended with an overview of a few commercially available MSS elements. Three volume handlers were described. The Calcomp/Braegan Automated Tape Library (ATL), originally developed by Xytex, Inc., automates the physical handling of 1/2-inch magnetic tape reels as the physical volumes. The ATL Model 7110B can handle from 947 to 8351 reels, yielding a unit capacity of roughly 112 to 1000 Gbytes when using 6250 bpi 9 track magnetic tape. (The Fermilab ATL houses about 4000 reels and thus has nearly a 500 Gbyte capacity.) Access time to an unmounted reel is 15 to 20 seconds (minus auto-load time on the tape drive).

The largest-selling automated volume handler is the IBM 3850 MSS, which was introduced in 1975 and recently dropped from the product line. It uses a data cartridge with 3-inch wide, 770-inch long, magnetic tape which stores 50 Mbytes. System capacities range from 706 to 4720 cartridges or 35.3 to 236 Gbytes. The cartridges are stored in a planar honeycomb wall; a robotic mechanism fetches them to read-write stations at one or both ends of the wall. Access time to an unmounted cartridge is approximately 15 seconds.

Manufactured by Fijitsu and marketed in this country by MASSTOR and CDC, the M861 storage module uses the same data cartridge as the IBM 3850, formatted to hold 175 Mbytes per cartridge. The M861 holds up to 316 cartridges, provides unit capacity of 55.3 Gbytes, and achieves about 12-second access time. The cartridges are stored on a periphery of a cylinder, where a robotic mechanism picks them for the read-write station.

Los Alamos Common File System

The Common File System (CFS), operational since June 1979, is a centralized, hierarchical network file system for the Los Alamos Integrated Computing Network (ICN). The ICN is composed of over 70 computers running 5 different timesharing operating systems in a scientific computing environment (primarily interactive but including batch) using the latest supercomputers. In general, active files are stored online on magnetic disk, less active files are stored online on a IBM 3850 Mass Storage System, and inactive files are stored offline on magnetic tape. Files are automatically moved between these media based on the time since last accessed. The CFS currently serves timesharing and batch users, worker operating systems, and network serves. The user is responsible for all file manipulation via the following commands:

CREATE a root directory node
ADD a directory node
REMOVE a directory node
MODIFY directory information
LIST directory information
STATUS of system or user request
ABORT a request

SAVE a file
GET a file
REPLACE a file
COPY a file
DELETE a file
MOVE a subtree

A proposed future (there are many proposed futures) is for the CFS to supply automatic space management. About 10 sites are using (or trying to use) CFS. Several of the sites quickly point out that CFS is not a turn-key system and that any change to the nearly 120,000 lines of PL1 code in CFS is hard.

Invited Sessions

The two days of invited sessions (17 in all) were segmented into three groups; Mss Topics, Applications, and Automated Libraries. It is not the author's intent to summarize each of the sessions in this paper as that would be time consuming and contain uninteresting material. Rather it is hoped that this paper can summarize and compare some of the more absorbing (in the context of high energy physics computing) topics discussed. If more information is desired on a subject please contact the author.

MSS Topics

A High Data Rate, High Capacity Optical Disk Buffer

Martin Levene of RCA/Aerospace and Defense/Advanced Technology Laboratories described an optical disk buffer concept (under development) that is projected to provide gigabit-per-second data rates and terabit capacity through the use of arrays of solid state lasers applied to an optical disk stack. Twelve double-sided, 14-inch diameter disks are fixed to a common shaft and each disk surface is served by an independent electo-optics module for record, playback, and erasure of data. A magneto-optic technique is used to provide erasable-reusable recording and playback.

Projected performance:

	capacity
number of user bits/revolution/track	1,081,344
track density	8 tracks/13.3E-6 m
number of tracks/surface	37,813
total buffer user data capacity	122.5 Gbytes
	data rate
disk rotation rate	15.413 rps
user data rate per track	2.08 Mbytes/s
user data rate per module	16.67 Mbytes/s
user data rate per 12 modules	200 Mbytes/s
	access time
access time	100 ms

It is interesting to note that little was said of the projected data error rate. Several symposium attendees speculated that it was much higher than is achieved by magnetic recording.

The development of the Optical Disk Buffer is being funded by a consortium of Government sponsors that includes NASA, the US Air Force, and the Defense Mapping Agency. RCA plans a demo by the end of 1986.

Applications

A Data Management System with Optical Disk Storage

Douglas Thomas of the George C. Marshall Space Flight Center described the application of a high rate Data Base Management System (DBMS) with an on-line archive under development as part of the NASA Data Systems and Technology Program. (Joseph Calabria of RCA Advanced Technology Laboratories later described the optical disk "jukebox" mass storage system used here. It is one of two such specially designed and built units, each estimated to cost about $2 million.) The design goals are to provide a capability to accept data at high rates, archive the data, and make the directory and data available to the end user in near real time. The key design elements and components being used are a 100 Mbits/s fiber optic data bus and optical disk unit capable of recording and reading data at a sustained rate of 50 Mbits/s (burst rate 100 Mbits/s), a distributive processing system, and an autonomous data packet format wherein data from multi sources and disciplines can be interleaved onto an archive. The system supports a network of on-line users that have near real time access to the directory of the data, which resides on one of three VAX 11/780 computers in the system, and to the data stored on the optical disk archive. The optical disk system consists of 125 disks in an on-line "Jukebox," each of which has a capacity of 10 Gbytes, with a total on-line capacity of 1.25 terabytes of user data.

(In a NSC HYPERChannel connection between a VAX 780 with a NSC A400 adaptor and the jukebox with a NSC A300 adaptor they have been measuring a transfer rate of only 6.5 Mbits/s. The protocol used is based on the ISO OSI Reference Model. The two adaptors are rated at 11 Mbits/s and 44 Mbits/s respectively.)

IBM 3840 Technology

Andy Gaudet of IBM/Tucson described the 3840 tape cartridge. A few numbers (and comparisons to the 1973 tape technology) are:

IBM 3840	1973 tape technology
200 Mbytes/cartridge	156 Mbytes/0.5 inch reel
38000 bpi	6250 bpi
3.02 Mbytes/s	1.25 Mbytes/s
2 m/s	5 m/s
1000 Gbytes/error	10 Gbytes/error
3% head replacement/5 yrs	100% head replacement/5 yrs
1 clean/week (cartridge)	1 clean/shift (manual)
18 tracks	9 tracks
4 redundant tracks	2 redundant tracks
123 Gbytes/m**3 in racks	----
$14 per cartridge	----

Additionally, four points were stressed. The first is that areal density will probably go up by a factor of 10 during the next few years. Secondly, reliability is great! Thirdly, compatability with future tape drives was almost guaranteed. Lastly, the lower cost per byte was highlighted.

Although many people felt a robot/Jukebox for the 3840 was in development IBM refused to comment.

Automated Libraries

Hitachi Optical Disk Subsystem

Michio Miyazaki of Hitachi presented an overview of the Hitachi Optical Disk Subsystem consisting of optical disk drives, library units, and controllers. Although the hardware is available today the necessary software is not. A few numbers:

disk capacity (double sided)	2.6 Gbytes
library capacity	320 Gbytes
access time (once mounted)	250 ms
transfer rate	0.44 Mbytes/s
mount time	8 seconds
dismount time	7 seconds
raw data error rate	1E-6 error/bit
cost per platter (double sided)	$300
cost per platter (single sided)	$250

A System Approach to Solve the Mass Storage Problem

John Burgess of FileTek described how they have used their IBM 3851-based Sperry mainframe backend mass storage system and experience to develop a mass storage system using optical disk. Their current configuration (which is quite flexible) includes a FileNet Optical Storage and Retrieval Unit (OSAR), Hitachi optical drives, NSC HYPERChannel and NETEX, and a Digital VAX running VMS. This hardware, used in a hierarchical management system concept, provides (at least on paper) an elegant solution to many user-oriented file management needs. Much of their software for storage management is written C. They have had a prototype since July and will be in beta testing for the first half of 1986.

(FileTek has measured the data transfer rate from VAX to VAX across the HYPERChannel at 400 KBytes/s. They characterize that rate as unacceptable and hope for a 1 Mbytes/s rate with the Intelligent Product Interface (IPF).)

(OSAR is a family of systems for providing on-line storage of up to 530 Gbytes per unit using either Hitachi or

OSI optical disk drives. The average time to move a cartridge, select it, move it to a drive and insert it into a drive is 4.7 seconds. The time to exchange cartridges at a drive is 2.7 seconds. The average time to move to a cartridge, select it, move to a drive, insert it, start spindle, position laser, position data read including average seek and latency is 9.3 seconds. The average time between cartridges when servicing a queue of requests including retrieval of next cartridge, spindle start, stop, and read of 1 Mbyte of data on a two-drive system is 7.5 seconds. All of these timings are for the OSI drives with the Hitachi timings slightly greater. FileNet has shipped about 40 units and is producing 5 units per month.)

A Rotary Digital Cassette Recorder for MSS

Arthur Strahm of Datatape (a Kodak company) described a rotary digital cassette system, currently under development, with user data throughput of 12.5 Mbytes/s and storage density of roughly 30 Gbytes per cassette. The system uses a jukebox housing 32 cassettes storing nearly 1 terabyte of data and is microprocessor controlled. The system can be configured for a number of user requirements. Configurations range from a single read/write transport to three read/write transports with a jukebox. The system is designed to provide mass storage of archival data and to be used with fast access, low storage-density disk systems.

A few numbers:

```
----------------------------------- transport and cassette
28.75 Gbytes per cassette user data capacity
12.5 Mbytes/s data rate (read/write)
manual or automatic load
38 minute record time
200 ips search speed
6.5" x 6.5" x 3" cassette of 1 inch 3M5198 700 Oe tape
1030 feet tape
62 second rewind
70E-6 second start scan
290E-6 second start
1E-9 bit error rate
200 to 250 read/write cycle (design point 1500)
1000 hour head life
about $400 per cassette (or $13 per Gbyte)

------------------------------------------------- jukebox
32 cassettes per jukebox capacity (nearly 1 terabyte)
7 second maximum access time (including threading)
```

The error rate is without a sophisticated controller or great effort and could easily be 1E-11 or 1E-12. A simple reduction in the error rate is achievable if the data rate and capacity were halfed. The bit error rate would then be around 1E-14.

A prototype system will be available in May 1986 with an estimated cost of $1 million. The tapes are currently being bought from Bosch but Kodak is building a manufacturing plant with a capacity to 30,000 cassettes per year. Although this system is new, Datatape has been building similar transports and cassettes for some time. They have about 200 systems of the same basic tape technology installed worldwide (22 in Europe). (They seem to have been quietly supplying the military industry with advanced technology tape units for years.) Datatape is eager to talk to prospective customers about possible applications of their products.

Another tape system that Datatape has been marketing for about 5 years is their System 600. During an evening chat the system was described as follows:

-- System 600
37.5 Gbytes per reel user data capacity
450 Mbits/s (56 Mbytes/s) data rate (read/write)
45 Kbits/track/inch (5.6 Kbytes/track/inch)
36 bit parallel interface
2" tape on 14" reel
84 tracks (72 data tracks/12 redundant tracks)
16 speeds (150 ips or 56 Mbytes/s to 8 ips or 3 Mbytes/s)

Datatape reports about a 1 year order time and that a system will be on the test floor in Pasadena during the first half of January.

An Automated MSS with Magnetic Tape Cartridges

Kiyoshi Itao and Shigefumi Hosokawa of NTT and later Shizuo Abe of Fijitsu desribed a joint project which produced an automated mass storage system with magnetic tape cartridges (18 track VHS tapes) as follows:

system capacity	60-670 Gbytes
cartridge capacity	280 Gbytes
data transfer rate	2.5 Mbytes/s
recording density	32,000 bpi
tape length	807 feet
tracks	18
tape speed	2 m/s
average tape handling time	10 seconds
head cleaning	automatic

The 670 Gbyte system requires an installation area of 10 square meters (10m x 1m x 1.7m).

Design of a Optical Disk Subsystem

David Bonini of CYGNET Systems described the design goals and elegant design implementation of a modular, high-performance, high-capacity storage system for single or double-sided 12-inch removable optical storage media. Four points were stressed during design: 1) high capacity with flexible growth, 2) media protection and data availability, 3) throughput, and 4) vendor adaptability. The design consists of three modules: an optical disk storage module; an elevator/transport module; and a mixed optical disk storage and disk drive module. This module also as a 19-inch rack to house in-board computing hardware. The subsystem can hold up to 141 cartridges and up to 7 drives. (There is a trade-off between these numbers. More drives means less room for cartridges.) The subsystem currently uses Thomson, Optimem, OSI, Hitachi, or Sony drives and may be easily modified to accomodate other brands. At two

Gbytes per disk, the maximum capacity is 282 Gbytes yielding a floor density of 18 Gbytes per square foot. Timings for a complete search, pick, move, and insert range between 3 and 3.5 seconds for most brands of drives. (The Thomson drive requires 5 seconds.) A complete slot to slot action takes 8 to 12 seconds. The command interface is a RS232 line. The estimated price for a full 3 module unit is $115,000 without drives. ($80,000 to OEM customers.)