The Challenges Facing Science Data Archiving on Current Mass Storage Systems

Bernard Peavey and Jeanne Behnke
Earth Science Data and Information Systems Project
Code 505
Goddard Space Flight Center
Greenbelt, MD 20771
bernie.peavey@gsfc.nasa.gov
301-614-5279
jeanne.behnke@gsfc.nasa.gov
301-614-5326

Introduction

This paper discusses the desired characteristics of a tape-based petabyte science data archive and retrieval system (hereafter referred to as "archive") required to store and distribute several terabytes (TB) of data per day over an extended period of time, probably more than 15 years, in support of programs such as the Earth Observing System (EOS) Data and Information System (EOSDIS) Kobler [1]. These characteristics take into consideration not only cost-effective and affordable storage capacity, but also rapid access to selected files, and reading rates that are needed to satisfy thousands of retrieval transactions per day. It seems that where rapid random access to files is not crucial, the tape medium, magnetic or optical, continues to offer cost effective data storage and retrieval solutions, and is likely to do so for many years to come. However, in environments like EOS, these tape based archive solutions provide less than full user satisfaction. Therefore, the objective of this paper is to describe the performance and operational enhancements that need to be made to the current tape based archival systems in order to achieve greater acceptance by the EOS and similar user communities.

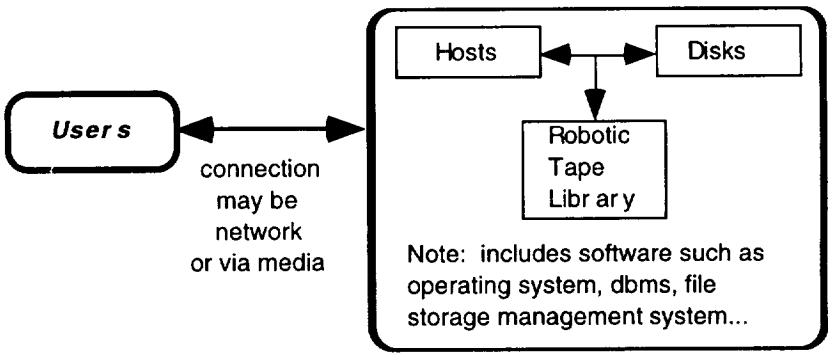

Figure 1: Generic Tape-based Archive

The archive discussed in this paper is shown in Fig.1. Its basic components - host, magnetic disk (perhaps solid state or holographic memory in the not too distant future) for

caching/staging (hereafter referred to as "disk"), robotic tape library, input/output media devices, and associated software (operating system, database, file management, resource management, network, communication protocol, operation control, etc.) are assumed to be fully integrated as an operational system, which could be centralized or distributed as appropriate to the user environment and data sources. The archive architecture and configuration are assumed to be such as to allow expansion or growth from a nominal one petabyte to 100 petabyte storage and performance capacity as data continue to be accumulated and the number of users continues to increase. The archive is expected to store and retrieve a variety of data types, the files of which may range from 1 KB (kilobyte) to 10 GB (gigabyte) in size, and handle thousands of user transactions a day. Being an operational system required to satisfy a multitude of users (vs. a laboratory facility), this archive is, therefore, characterized from a system's rather than a component's perspective. For example, the performance of a given tape drive is not addressed directly; rather, the data transfer rate from disk to tape or from tape to disk, including all overhead associated with managing each data file before it lands in a given location, is specified. Thus, the salient archive characteristics addressed in this paper are: storage density, storage organization and management, write rate, read rate, file access time, data integrity/preservation, data retrieval/distribution, data interchange or interoperability, and operation control. They are examined from an operational system's perspective to highlight their significance in realizing the archive's desired capabilities.

Given the state of current technology and available archive components as described in the literature Shields [2] and observed in the field, can the subject archive be offered by the vendor community at an affordable price? This twofold question of performance and cost is examined from the standpoint of real progress already made in this area - a reality check, and what remains to be done to reach the goal of achieving the desirable archive characteristics at an affordable price.

Salient Characteristics

In discussing the archive's salient characteristics, it is assumed that the system architecture allows the use of multiple tape drives, robots, disk banks and hosts as appropriate to achieve the desired capacity and performance, and the local network bandwidth is sufficient to support this performance. As mentioned previously, these characteristics, which become specifications when they are given specific/particular values, are considered from the standpoint of a fully operational system, and their measurements are made on this basis as well. This means that for systems which utilize multiple components operating in parallel, e.g., tape drives or disk drives, characteristics such as data transfer rate (write or read) are given as aggregate values, as illustrated in Fig. 2. In general, characteristics associated with data transfer or data flow are considered to be "end-to-end", viz., for storage, data transfer begins when the data enters the host, and for retrieval, data transfer ends when data lands on the archive disk shown in Fig.1. System level characterization of the archive is key to describing the archive's capabilities in realistic terms and relating them to operational expectations. Regrettably, the practice of characterizing archives at the system level is not yet standard or even prevalent, perhaps because the vendor community does not usually offer integrated archives as products. Instead, archives are typically specified in terms of performance of their components such as tape drives, tape libraries, etc., which means that a great deal of system engineering and development effort must be applied by or provided

to the customer in order to realize the complete archive solution. From the archive customer's perspective, procuring the archive on the basis of system level characteristics presents the vendor community with an opportunity to offer fully integrated archive systems as products and, hopefully, at lower cost to the customer. In any event, what follows are the desired archive characteristics as seen by the end user. It should be noted that at this time there are no commercial-off-the-shelf (COTS) tape-based archive systems that include all of the desired characteristics. Adding new features to COTS products tends to be very costly. Thus, by examining the following characteristics, it may be possible to identify opportunities to enhance existing COTS products or to develop new products.

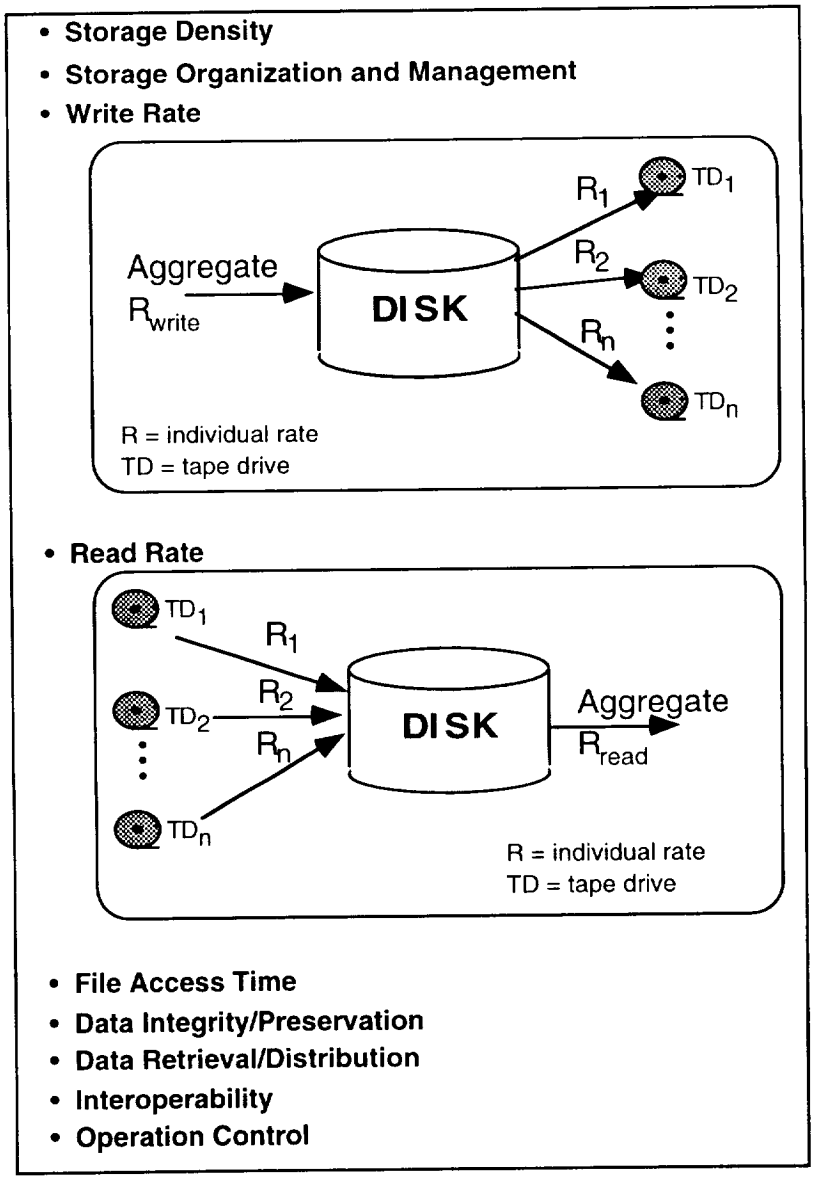

Figure 2: Salient Characteristics of a Tape-based Archive

Storage Density

Storage Density, given in terms of bytes/in, bytes/cm, or bytes/tape (with known tape dimensions, i.e., width and length), is directly related to the archive's storage capacity. For example, the D3 tape cartridge is advertised to hold 50 GB. Actually, from a system's perspective, the effective storage density is lower due to the associated file management overhead, which increases with the number of files. In addition, data compression, if used, must also be taken into account. Therefore, this characteristic should be given in terms of effective storage density. A petabyte (PB) archive using 50 GB tape cartridges requires 20,000 cartridges which, at $50/unit, amounts to $1,000,000! Both numbers are prohibitive, especially when extended to a 100 PB archive. Clearly, a tenfold increase in storage density would be welcome within the next few years, and a 1 TB per tape capacity would be required in the near future. But, increased capacity (at the same cost and overall size, of course) alone is not enough without higher read/write rates, and shorter file access time to sustain a reasonable performance level. Is this a technological challenge, economic (commercial demand) challenge, or both? The likely answer is that the challenge is economic, but time will tell.

Storage Organization And Management

Storage Organization And Management (SOM) provides the capability to control the way in which data files (hereafter referred to as "files") are stored on and retrieved from tape. For example, SOM selects tape drives (hereafter referred to as "drives") and tapes, directs the flow of files to/from selected drives, provides logical and physical file organization, maintains knowledge of file location and status, causes the robotics to load or unload selected tapes (volume mounting/dismounting), controls access to each file, and keeps statistics on file access frequency. In discussing this characteristic, it is assumed that SOM also controls the availability of the cache/staging disk (Fig. 1), which is part of the archive. The criticality of this system level characteristic cannot be overstated with regard to system performance, especially when ordered files such as those arriving from Landsat USGS [3] are requested to be retrieved in random subsets, and the system has to manage a continuously increasing file inventory on the order of 1-10 billion files.

In order to allow system performance tuning, the SOM should include, among others, the following selectable options for writing files onto tapes:

(1) Chronological order

(2) No file splitting across tapes

(3) File continuation on second tape (The first tape must identify the existence of a partial file and provide the identification of the second tape. The second tape must identify the existence of a continuation file and provide the identification of the first tape. Note that no more than 2 partial files may exist on a given tape: the beginning part of one, and the continuation part of another.)

(4) Unique file grouping (Writing a uniquely identifiable file collection on the same tape, e.g., files from a certain scientific instrument)

(5) Superfiles (Writing a collection of files as a super file so as to be retrieved as one super file or as individual files)

(6) Data compression (Per whole tape)

(7) Maximum tape utilization (Random collection of files to minimize unused tape)

(8) File replication (Writing the same file to different tapes or to the same tape)

(9) Tape duplication (Writing multiple tapes of same files simultaneously)

(10) Simultaneous file recording (Writing multiple files to multiple tapes simultaneously, see Fig. 2)

For data retrieval, the SOM must provide the following read options:

(1) Ordered files from a single tape (Per requested sequence)

(2) Ordered files from multiple tapes (e.g., k files from tape 1, m files from tape 2, n files from tape 3, etc.)

(3) Interleaved files from multiple tapes (e.g., file A from tape 1, file B from tape 2, File C from tape 3, etc.)

(4) Superfiles (Collecting multiple files from a single tape or multiple tapes into one file as requested)

(5) Compression/decompression

(6) Tape quality information

The SOM must also include the capability to produce or write tapes that are self describing so as to be read on any compatible drive external to this archive.

As a file manager of a growing archive, the SOM must be scalable to accommodate a 100 fold (from 100 million to 10 billion) increase in the number of files. In addition, it should be applicable to centralized as well as distributed archive architectures. It would be nice if the disk shown in Fig. 1 could be eliminated while still providing the desired SOM, since by so doing the scalability problem could easily be solved, and one data flow hop could be eliminated as well. However, barring that possibility, separating file management and volume management should be considered as part of the scalability problem solution. In addition, advantage should be taken of this disk to improve the efficiency of file storage management and retrieval (e.g., executing the various writing options stated above, distributing a given file to multiple users, collecting files located on multiple tapes to satisfy a single data request). In general, the SOM should have the necessary features to optimize the overall file storage and retrieval performance, while being independent of any operating system (OS) as much as possible. This independence is crucial for the SOM software to be able to run on any hardware platform, present or future, which is key to evolvability.

Although a number of SOM versions such as UniTree, AMASS, FileServ, which are known as File Storage Management Systems (FSMS), are presently in use, they incorporate only a few of the SOM options, and are strongly dependent on the platform's OS. Also, these FSMS do not conform to any standard since none exists yet. To achieve plug and play COTS FSMS (or SOM) products, the vendor community must support the development and adoption of a FSMS standard. It appears that the efforts made by the IEEE and ISO over the years to develop an open systems standard have not borne fruit yet. However, some activity in this area has been afoot which provides an opportunity to revitalize this effort. Kobler [4], Jones [5].

Write Rate

This characteristic defines the time required to read incoming files from the disk and write (store) them to tape so that they can be retrieved upon request. As a system level characteristic, it includes the time to uniquely identify each file, append location metadata, select the drives, load the tapes, perform compression (when required) perform error protection for error detection and correction, write the files, update the catalog/database, and return status. The write rate is given for a single or a multiple drive configuration. For a multiple drive configuration, the write rate is the aggregate rate, viz., $R(w) = R(1) + R(2) + ... + R(n)$, where R() are the individual write rates with all n drives writing simultaneously (See Fig. 2). For example, if the incoming data rate is 10 MB/sec (as expected from EOS), the system could handle it with one drive, which must be capable of writing at a rate greater than 10 MB/sec in order to compensate for delays due to FSMS overhead, and physical tape handling functions such as robotics, loading and unloading. Alternatively, the system could accommodate this incoming data rate with multiple drives writing simultaneously at an individual drive write rate lower than 10 MB/sec. Therefore, the write rate (which could also be referred to as "storage rate") is the effective end-to-end system rate at which files can be stored in the archive. It is assumed that in cases where unique file grouping is required, the disk provides sufficient staging and buffering capacity to feed the drives. To write files onto a 50 GB D3 tape cartridge at 10 MB/sec requires the use of drives that cost $150,000 each, which is expensive. It appears that the drive write rate needs to be increased by a factor of 2 or more, and the drive cost needs to be reduced considerably to make a petabyte archive more affordable.

Read Rate

This characteristic defines the time required to read (retrieve) files from tape and write them to the disk for distribution. As a system level characteristic, it includes the time to read the data request, identify and locate the tapes of the requested files, access the files, read the files and write them to the disk with error detection and correction (EDAC) applied, append the metadata, and return status. This read time is comprised of 2 components: file access time, and the time to read the file. The file access time is described in the next paragraph as a separate characteristic, although it is included here as part of the read rate definition for completeness. The read rate is given for a single or a multiple drive configuration. For a multiple drive configuration, the read rate is the aggregate rate, viz., $R(r) = R(1) + R(2) + ... + R(n)$, where R() are the individual read rates with all n drives

reading simultaneously (See Fig. 2). For example, if the required outgoing data rate is 30 MB/sec (as expected for EOS), the system could support it with one drive, which must be capable of reading at a rate greater than 30 MB/sec in order to compensate for the delay due to file access time. Alternatively, the system could accommodate this outgoing data rate with multiple drives reading simultaneously at individual drive read rates lower than 30 MB/sec. (Of course, if all requested files were to be located on the same tape, the multiple drive configuration would not meet the 30 MB/sec output rate). Therefore, the read rate (which could also be referred to as "retrieval rate") is the effective end-to-end system rate at which files can be retrieved from the archive. It should be noted that, based on current technology, the file access time can become so significant when many files have to be accessed on many tapes as to require additional drives to compensate for it. The requirement for multiple drives should also be considered in light of the user response requirements, namely, the number of users that need to be served simultaneously. This aspect is discussed later as part of the Data Retrieval/Distribution Characteristic. Generally, the read rate requirement is significantly more stringent than that for the write rate, not only because more data is going out of the archive to users, but also due to the need to minimize waiting time for non-uniform data request distributions. Therefore, to accommodate thousands of transactions a day, the archive may have to utilize 10-20 drives which, on the current market, may cost $1.5 million to $3 million. This is prohibitive, and points to the need for improved drive performance and cost reduction.

File Access Time

File Access Time (FAT) which is part the previous read rate characteristic, is the total system time required to locate a given file in a tape-based archive following the issuance of the request to retrieve it. This time includes file identification, drive and tape selection, robotic motion/travel, loading the tape, reaching the desired file in a position ready to be read, unloading and returning the tape to its bin. The current technology achieves a FAT of 1-2 minutes, depending on the tape length and file location. Clearly, this lowers the effective retrieval rate, especially when many files have to be retrieved from many tapes. To cope with such a delay, today's archives must utilize multiple drives, with attendant cost increases. Therefore, the FAT must be reduced by a factor of 3 or more to improve the cost performance ratio, and allow the on-line user to start receiving data within less than one minute from the time of having made the request.

Data Integrity/Preservation

A persistent archive requires that files stored on tape be entirely preserved with no degradation of their content during the archive's life (30 years). Therefore, the system must be capable of monitoring the state of data quality (e.g., BER), and the physical condition of the medium to determine when to refresh (transcribe to a new tape) or just rewind a given tape, and do so automatically or under operator control. These actions should be based on frequent checks of the BER, which should not exceed 1 in 10 to the 12th bits (each time a file is read or at specified time intervals), file access frequency, and time in storage. In addition, a backup capability is needed to make and manage copies of selected tapes or files. Since in today's systems the capability of this characteristic seems to be limited to manual intervention, this capability should be enhanced to the fullest level.

Data Retrieval/Distribution

This characteristic defines the manner in which files are to be retrieved and distributed to users electronically or on media (tape, CD-ROM, the drives of which are assumed to be included in the archive). For example, it should be possible to retrieve and distribute files in whole or in part, in specified order (e.g., chronological - oldest file first, or most recent file first; per list specified in the request; or other), grouped by category (e.g., instrument; science discipline; product type), random file collections, file interleaved by tape (a given file from tape 1 followed by a given file from tape 2, etc.), and compressed or uncompressed format. Format conversion is a separate service which may be included in the archive system. This system level characteristic applies to both software (FSMS or SOM, DBMS, request processing) and hardware components' performance in order to achieve the desired data outflow rate. It is assumed that an appropriate DBMS is available and is included in the archive to serve the file catalog and file search functions, however, the schema design and implementation is a user provided application. It is also assumed that the disk capacity and speed (data transfer rate), the number of drives and their read rates are sufficiently high to support the required data distribution rate and the number of simultaneous data requesters.

As mentioned previously in the Read Rate paragraph, to support the requirement to retrieve and distribute several terabytes of data per day in response to thousands of transaction requests is very demanding of software (FSMS, DBMS, NFS) and hardware performance. With today's technology available on the market, this requirement can be met only by using lots of expensive hardware. Therefore, it is imperative that the hardware performance and reliability be greatly improved to make petabyte archives less costly.

Interoperability

This characteristic is intended to allow the archive components to be changed out in a "plug and play" manner without affecting the archive's functionality, and to support media-based data interchange (providing data to and distributing data from archives and users) among archives and users. In addition, the archive architecture must provide for the application software and user interface software to be independent of a given hardware platform and its OS. Thus, this characteristic, allows the archive to be scalable and evolvable as capacity and performance requirements continue to grow, and superior technology becomes available. To realize such a characteristic, COTS products (hardware and software) must comply with appropriate standards which are yet to emerge. Regrettably, today's products do not lend themselves to open interchanges. For example, tape formats are unique to the systems, FSMS are tailored to specific platforms and OS, and information describing their implementation is proprietary.

With regard to developing archive system standards, it should be mentioned that the work begun under the IEEE and ISO sponsorship has not progressed as far as was expected. Perhaps this slow progress can be attributed to the approach undertaken by these groups, without realizing that advances in archive and Internet technology are occurring at a much more rapid pace than anticipated, thus diminishing the desire of system developers and

vendors to wait for these standards before participating in the market and application opportunities. A better approach to developing archive system standards would be the model of the IETF. As Dave Clark of the IETF said in 1992: The IETF (Internet Engineering Task Force) credo is:

> "We reject kings, presidents, and voting.
> We believe in rough consensus and running code."

Perhaps this nontraditional approach taken by the IETF group should be followed in developing the standards for Mass Storage Systems (MSS) and FSMS. Rather than following a top-down approach to include "all or nothing", it might be more productive and effective to pursue the incremental and less rigorous approach with the notion that "having a standard is better than none". The EOSDIS Project at the Goddard Space Flight Center is participating in the effort to develop these standards, and is committed to using them.

Operation Control

This characteristic describes the extent to which system operation should be controlled automatically. The most desirable feature would be full automation or "lights out" mode of operation, where the only required interface is the user, while the operator/technician performs maintenance, or user services type functions. To achieve a high degree of automatic control, the system must be capable of self checking, monitoring ongoing activities, sensing critical conditions and reacting to them, controlling resources, balancing workloads, managing request queues, tracking user requests to the file level, accounting for resource utilization per user request, helping users, monitoring system performance and quality, collecting production statistics, reporting and logging events, issuing remedial instructions, etc. (Also, it would be nice to have the system repair itself, but for now this must remain a dream to come true). Unfortunately, today's systems require considerable operator intervention in running an archive. Therefore, such intervention should be minimized at best in order to control the operation cost.

Discussion

A growing tape-based petabyte archive for science data, which is the subject of this paper, is described in terms of its salient characteristics, and their implication on the architecture, implementation, acquisition, and cost thereof. Ideally, these functional and performance characteristics should be sufficient to specify the desired archive (large or small) so that it could be procured at a reasonable price from a given vendor as a COTS product, consisting of COTS components which the vendor would select, integrate, test, demonstrate, and turn over to the customer as a fully operational archive. The customer's involvement in this process would be minimal except for a fixed price proposal/bid evaluation and acceptance testing. To use the archive acquisition approach described above, which is expected to result in considerable cost savings, the customer must know what is needed, the technology must be mature, suitable components must be available as COTS products that are compliant with industry standards, and there must be a market for these components. By examining these characteristics in light of available COTS products, the aforementioned premises are not all satisfied at this time. The most critical of these premises are technology

and standard COTS products that would satisfy the desired functionality and performance requirements at a reasonable cost. Historically, not much has happened until 1995, when new tape drives and cartridges were introduced that boosted the read/write rates to 10 MB/sec, and increased the storage capacity to 20 GB per 3480 type cartridge (higher capacities are on the way, e.g., the D3 cartridge). However, more work is needed to produce a 1 TB cartridge, and a 30 MB/sec read rate drive with a file search time of less than 20 seconds anywhere on the tape. In the DBMS and FSMS areas, plug and play products are not yet available. Perhaps there will be an opportunity to develop a standard modular (to allow for incremental addition of features and scalability) SOM product which can be plugged into a microkernel type OS. Of particular interest and concern are the scalability and evolvability aspects of FSMS and DBMS COTS products in the absence of open system standards. The promises made in 1991 Rybczynski [6], McLean [7] toward the realization of petabyte archives have been slow in coming. It seems that the challenge to do so is still up for grabs.

The salient characteristics approach describes and specifies the archive at a system level because these characteristics are directly related to the user's needs or expectations, and can be measured on that basis. By so doing, the vendor is offered the opportunity to be creative and cost-effective in producing the optimum archive system in terms of functionality and performance. For example, selection of the type and number of tape drives should be a key consideration for a petabyte tape-based archive to achieve the required storage and retrieval rates, and to satisfy the required number of simultaneous user requests. Similarly, the vendor has the choice of selecting the hardware platforms and disks, as well as the appropriate software components. (Please note the emphasis on the vendor rather than the customer). Thus, vendors have the opportunity to offer standard archive components, or fully integrated, scalable turn-key archives. At this time, it is still necessary to stage files on disk as part of the storage and retrieval operation. (How nice it would be if disks could be eliminated from this operation). Therefore, adequate disk capacity and speed (data transfer rate) must also be a key consideration.

In conclusion, it appears that affordable (less than $10 million) tape-based petabyte archives for science data are difficult to find on today's market. However, it might be possible to find them in the near future with the help of enhanced technology, standard COTS products supporting plug and play system architectures, system level procurement specifications, integrated archive system products, turn-key system acquisition, and open storage system standards. The time must come when a 1 petabyte archive could be expanded or scaled up 100 times by simply replacing (plugging in) existing components with new more powerful components as they become available, in a manner completely transparent to the user, and at reasonable cost. That is still a challenge.

Acknowledgements:

We wish to acknowledge the invaluable assistance of P.C. Hariharan of Systems Engineering and Security, Inc in the preparation of this document.

References:

1. Kobler, B. J. Berbert, P. Caulk, P.C. Hariharan, "Architecture and Design of Storage and Data Management for the NASA Earth Observing System Data and Information System (EOSDIS)," *Fourteenth IEEE Symposium on Mass Storage Systems*, Monterey, CA, November 1995.

2. Shields, M., "Toward a Heterogeneous Common/Shared Storage System Architecture," *Fourteenth IEEE Symposium on Mass Storage Systems*, Monterey, CA, November 1995.

3. US Geological Survey and National Oceanic and Atmospheric Administration, "Landsat User's Guide," available from the EROS Data Center at URL: http://edcwww.cr.usgs.gov/glis/hyper/guide/landsat_tm

4. Kobler, B. and J. Williams, "A Straw Man Proposal for a Standard Tape Format," *AIIM International Conference (IEEE)*, Chicago, IL. 1996.

5. Jones, M., J. Williams, and R. Wrenn, "A Proposed Application Programming Interface for a Physical Volume Repository," *Fifth NASA Goddard Conference on Mass Storage Systems and Technologies*, September 1996.

6. Rybczynski, F., "Network Accessible Multi-Terabyte Archive," *Proceedings of the NSSDC Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications*, Goddard Space Flight Center, July 1991.

7. McLean, R. and J. Duffy, "ICI Optical Data Storage Tape," *Proceedings of the NSSDC Conference on Mass Storage Systems and Technologies for Space and Earth Science Applications*, Goddard Space Flight Center, July 1991.