EOSDIS Petabyte Archives: Tenth Anniversary

Jeanne Behnke, Tonjua Hines Watts, Ben Kobler, Dawn Lowe
NASA Goddard Space Flight Center
Jeanne.Behnke@nasa.gov
Tonjua.M.Hines-Watts@nasa.gov
Ben.Kobler@nasa.gov
Dawn.R.Lowe@nasa.gov

Steve Fox, Richard Meyer
Raytheon Information Solutions
Stephen_A_Fox@Raytheon.com
Richard_Meyer@Raytheon.com

Abstract

One of the world's largest scientific data systems, NASA's Earth Observing System Data and Information System (EOSDIS) has stored over three petabytes of earth science data in a geographically distributed mass storage system. Design for this system began in the early 1990s and included a presentation of the design of the mass storage system at this conference in 1995. Many changes have occurred in the ten years since that presentation, much of it performed while the system was operational. In its first operational year (2000), the EOSDIS system had increased NASA's collection of earth science data holdings eight-fold. Today, EOSDIS collects over 7,000 gigabytes of data per week, almost 60 times more than the Hubble Space Telescope. This load represents major challenges for ingest into the mass storage system , as well as for timely and balanced data distribution out of the mass storage system. This paper discusses the evolution of the EOSDIS archives focusing primarily on the mass storage system component of the archive. We present the lessons that were learned over the years and some directions that we are taking for the future.

1. Introduction

In the 1990's, NASA devised the Earth Science Enterprise as a long-term research mission to study the processes leading to global climate change. Key to the twenty-year mission is the Earth Observing System, a NASA campaign of satellite observatories. These satellite observatories are individual projects in charge of a spacecraft flying several scientific instruments from which NASA collects data about the Earth. The idea behind EOS is to have a set of climate data records that span twenty years or more. The Earth Observing System Data and Information System (EOSDIS) is the principal component of the mission that provides the earth science community with easy, affordable, and reliable access to earth science data. The EOSDIS software architecture has been designed to receive, process, archive, and distribute several terabytes of science data on a daily basis. As shown in the EOSDIS context diagram in Figure 1, the EOSDIS is complex and involves many steps from satellite data collection to dissemination of data to end users. Beginning with the launch of the Terra spacecraft in 1999, the EOSDIS supports the storage of data from 28 separate NASA missions supporting a total of 66 distinct scientific instruments. As a measure of its success, data from the last spacecraft Aura, launched in 2004, was quickly and easily included in the archive system. As of 2005, EOSDIS has stored in excess of three petabytes of earth science data. Thousands of science users and non-science users access the data through several interfaces on a regular basis. This paper describes the development and operation of the EOSDIS archives from 1995 through 2005, principally discussing the design of the archive systems and how these systems have evolved to incorporate changing technology.

EOSDIS is a distributed system, with major facilities at eight data centers located throughout the United States. The principal EOS archive and distribution centers are at four Distributed Active Archive Centers (DAACs). These data centers are located at Goddard Space Flight Center (Greenbelt, MD), Langley Research Center (Hampton, VA), EROS Data Center (Sioux Falls, SD), and the National Snow and Ice Data Center (Boulder, CO). Other installations are located in Alaska, New York, Tennessee and California. The Earth Science Data Information System (ESDIS) Office at NASA Goddard Space Flight Center is responsible for EOSDIS science operations, which include data center management and operations, science data processing, and data archival and distribution. ESDIS has the responsibility for managing unique system development to further science operations objectives.

Proceedings of the 22nd IEEE/13th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST 2005)
0-7695-2318-8/05 $20.00 © 2005 IEEE

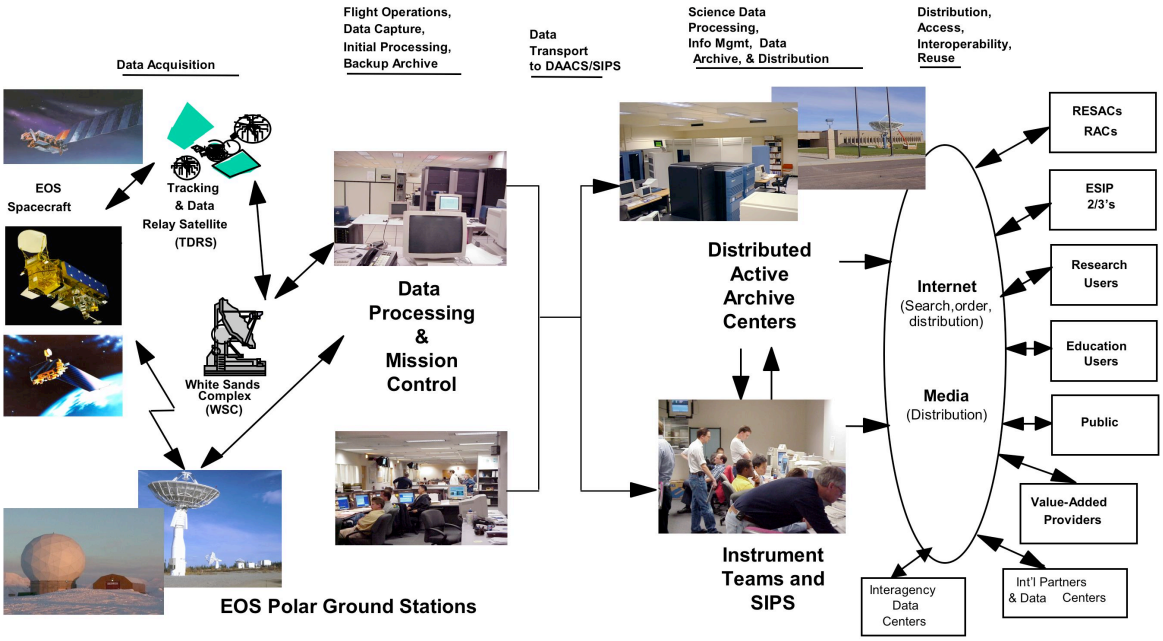

Figure 1. EOSDIS context diagram

A key subcomponent of the information system (EOSDIS) is what is commonly referred to as the EOSDIS Core System (ECS). ECS provides the science community with easy, affordable and reliable access to the data collected by the missions. A contractor team, led by Raytheon Information Solutions, developed the ECS for NASA. Without the successful teaming of NASA, Raytheon and the DAACs, the EOSDIS would never be able to accomplish this enormous mass storage effort.

In fiscal year 2004, over 1.9 million distinct users accessed the EOSDIS systems and 208,000 distinct users received data and information products from the archives. EOSDIS users received more than 36 million data and information products from the EOSDIS Core System in fiscal year 2004. This is 7 million more products than was distributed in the previous fiscal year. As the system continues to archive and manage this large collection of science data, EOSDIS can expect more users and increases in data access and download. Many lessons were learned about managing large archives, migrating to new technology and operating large, distributed data systems. But the future also holds change as many interesting technologies are becoming available that address moving terabytes and even petabytes over the network, such as grid computing, storage brokering and service enhancements. Hardware developments in tape and disk storage will allow us to migrate ever more data to inexpensive high performance storage and archive systems, and new developments will allow us to move to a more service oriented architecture. However, issues of data reliability and long term maintainability of data archives will need to be addressed.

2. Original design constraints and decisions

The original design of the ECS archive storage subsystem was completed in 1995. The main concerns at that time were reliability of tape archives, local network throughput, platform I/O throughput, and tape archive throughput. While ECS is not considered a real-time system, it does have to sustain 24 x 7 data streams coming from various data providers and distributing data to the local processing system, external processing systems, and end users. Daily ingest and distribution rates at a single site can exceed 3 TB each. In the paragraphs below, the original concepts for the data flow, the data processing chain, and related archive constraints are discussed.

Primary input from satellite ground systems and ancillary data sources is archived; that same data is provided to science processing. Output from processing constitutes the secondary input. It flows from the science processing systems into the archive; that same data is also distributed in bulk to select users (via subscriptions); and over the life of the program on demand to a large science community. Reprocessing data also presents challenges to the design of the ECS system. During a mission, instruments often need to be recalibrated and the

algorithms that process the data need to be updated to reflect changes in scientific understanding. Consequently, we need to pull all original (termed Level 0) data from the archive and process it into new and better products. Reprocessing constitutes an additional load on the archive systems.

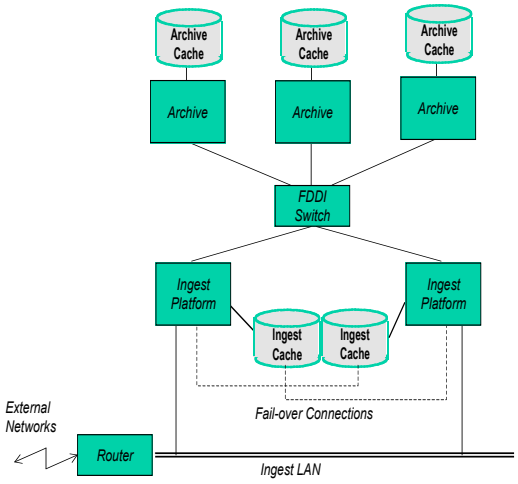

Figure 2: Ingest Architecture

The highest priority mission goal is to archive the satellite data. Requirements called for a high degree of system availability to prevent data from being lost. There was concern that a tape archive could be down for extended time periods due to problems with robotics and drives. The solution (see Figure 2) was to configure the Ingest Subsystem with two platforms for fall back (i.e., no real time fail over) with a large ingest disk cache capable of buffering 48 hours of data. Data is either pulled from the data provider or the data provider delivers (push) the data directly into the ingest cache. Data is then copied from the ingest cache into the archive cache. This decouples the ingest step from the archive step. Fluctuations in ingest bandwidth then can be smoothed out by the ingest disk cache. However, it also increased the number of I/O and network hops. I/O bandwidth considerations eventually led us to choose SGI platforms as ingest, archive and processing hosts. Network bandwidth considerations led to an elaborate network architecture, which divided network loads among subnetworks that eventually included FDDI switches, and was augmented by a switched HiPPI network.

We felt it was important to collocate data on tape that are likely to be demanded coincidentally. For example, for reprocessing, it is advantageous to collocate the inputs needed by processing and arrange them by observation time. Based on this consideration, we chose to organize the data products into archive tape groups. One important role of the archive cache was to buffer inputs destined for tapes not currently mounted for later writing to avoid frequent tape swapping.

Among the data subscribers are our own processing systems. This ensures that freshly ingested data are processed with little delay. Another role of the archive cache was to hold this data for a short time to avoid having to retrieve it from the tape. The role could not be filled by the ingest cache; it needed to be cleaned up promptly to ensure that under normal operating conditions, there is plenty of disk space in the ingest cache to cover unforeseen problems.

Disk space cost was at a premium in the mid 1990s. Caches were expensive and their size limited – 256 GB of cache space was considered generous back then. Careful planning of cache usage was essential. For example, spreading the archive inputs over too many tape groups would require buffering and would exceed the available disk space, which would then have led to tape thrashing. Spreading the archive inputs over too few tape groups would lead to extraneous tape mounts later during data access.

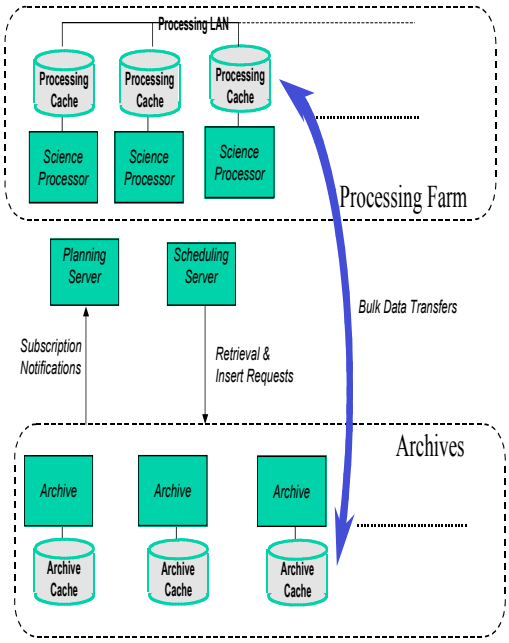

Figure 3. Original processing architecture

To reduce the demand on the archive cache, the platforms in the processing farm were equipped with their own caches (see Figure 3). Bulk data transfers were routed via a high performance switched network (HiPPI). Planning processing chains based on data availability and production plans was the task of a planning server. Dispatching the chains to specific platforms and then routing the inputs and outputs between the caches on the processing farm and the archives was allocated to a

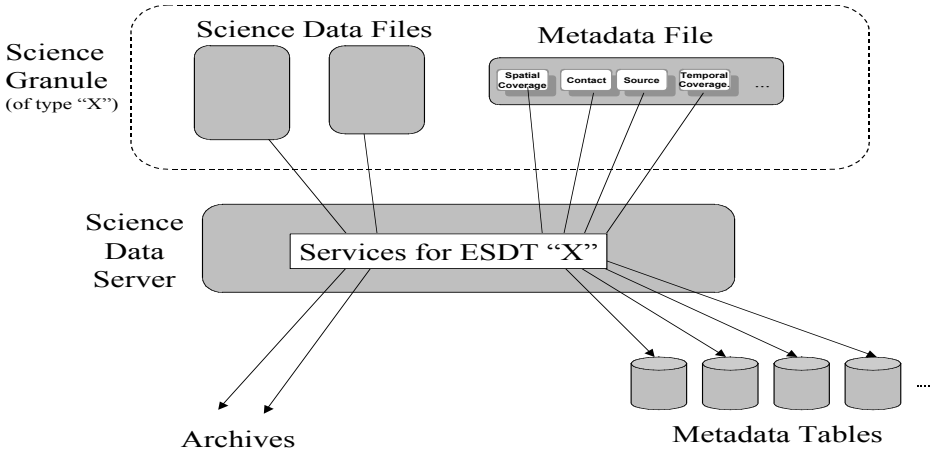

Figure 4. Granules and ESDT

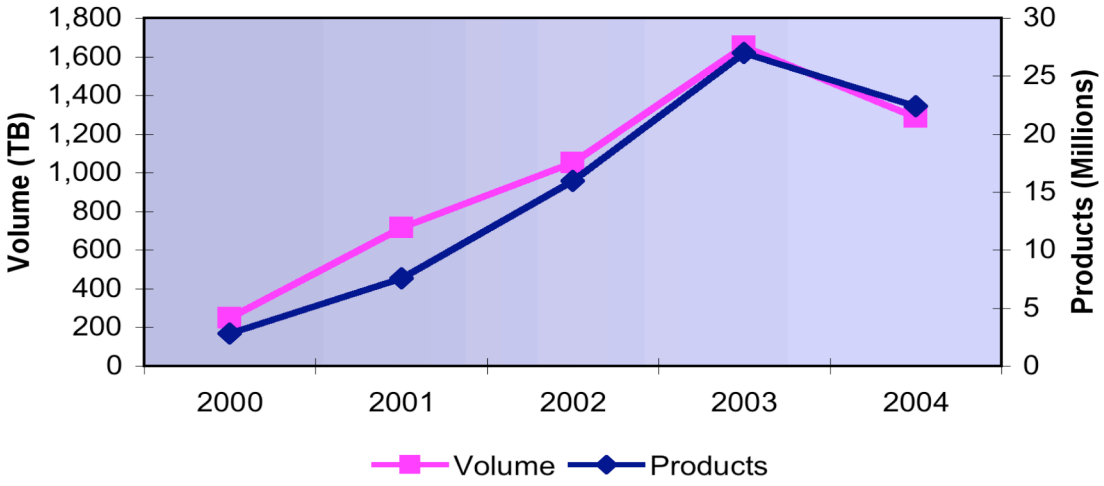

Figure 5. Historical ECS ingest rates

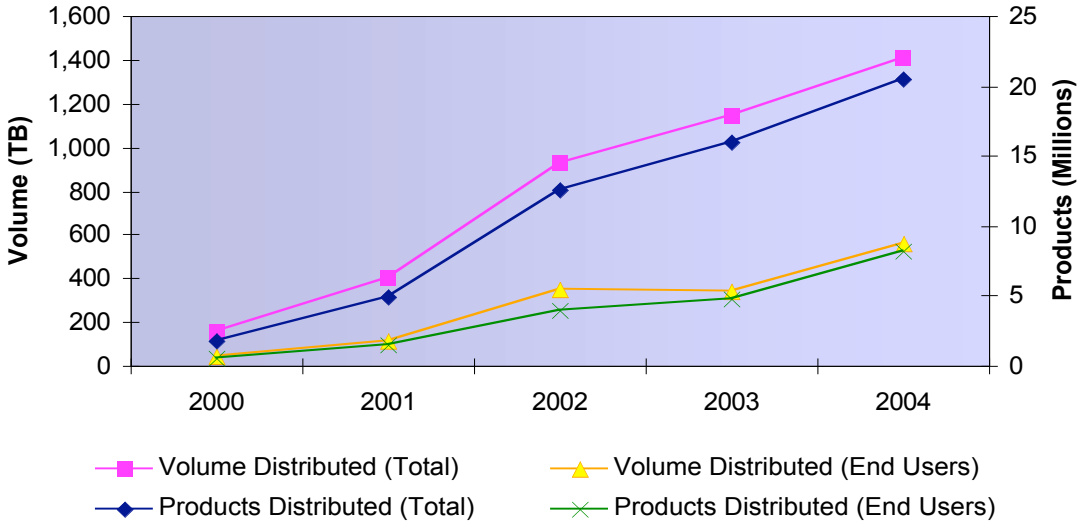

Figure 6. Historical ECS distribution rates

scheduling server. The processing caches also retain outputs from earlier steps in a processing chain until needed by later processing steps and while they are archived. However, the HiPPI network was sized to support the traffic between archive and processing caches; traffic between science processors had to occur via local area network. Network bandwidth was limited then and accessing the processing cache of another platform was very slow compared to local disk access. It was, therefore, essential to avoid situations where a science algorithm had to fetch its inputs across the network from a differennt platform. Simply limiting on which platform each type of chain can run leads to poor use of processing resources and backlogs, because processing imposes very high cpu and I/O demands; and the start of processing chains depends on the availability of inputs from that instrument, which can fluctuate significantly during the day. As a result, a lot of effort was put into the scheduling algorithm and into planning the processing chains and their resources. Typically, granules have some 50 to 100 attributes distributed over some 20 to 25 database tables. The types of attributes depend on the type of granule and vary widely, and different granules may require different representations for the same type of attribute. For example, ECS needs to accommodate eleven different types of spatial representations and several different types of temporal attributes. This has implications for ingest as well as searches. To accommodate this, we developed a metadata server (the 'Science Data Server') which supports a "type system" called Earth Science Data Types (ESDT). It coordinates storing the metadata in the database with archiving the science data files and its "ESDT Services" hide much of the complexity of the metadata from the users.

3. Current system metrics

Figures 5 and 6 show historical ingest and distribution metrics. Ingest rates increased significantly over the years as data from new missions became available and data reprocessing began. Last year, showed the first decrease in total volume ingested as a major reprocessing campaign ended in the middle of the year. Ingest rates for 2005 are expected to top 5 TB per day as full reprocessing is once again resumed. Data distribution rates continue to increase each year with over 1.4 PB of data distributed to end users and other science processing systems in 2004. Over 30,000 orders are processed each day with more than 70% of the orders generated by subscription. By subscription, we mean that users have the option to have regular deliveries of selected data as they arrive at the archive. Most data are delivered electronically via Ftp Push or Ftp Pull. Less than 5% of the data are delivered via media (8MM, DLT, CD-R, DVD-R).

Daily ingest and distribution volumes fluctuate significantly (See Figure 7). Distribution volume correlates strongly with ingest volume. This is because many users receive current data via subscriptions. Table 1 shows distribution statistics by media type from the Land Processing (LP) DAAC at the EROS Data Center. The bulk of the data is distributed via ftp, either pulled by the user from our public staging area, or pushed by us to the user's site.

Over 80% of the requested data is more than 90 days old (see Figure 8). Distribution lag time ranges from a few minutes for small requests for recent data that is still in cache to days for very large requests whose data need to be fetched from many tapes.

Table 1. Distribution volumes by media type

LP DAAC, October 2004			
Media	**Requests**	**Granules**	**Volume (GB)**
8MM	2	3	1
CDROM	35	2,908	337
DLT	74	32,445	4,377
DVD	108	31,883	2,286
Ftp Pull	3,310	232,516	10,035
Ftp Push	5,034	137,134	17,903

Figure 9 illustrates how the total workload (inserts plus reads) varies by archive at our largest site. The workload depends very much on what data products an archive holds and for which time periods, and what portion of these products are currently being produced and the user demand for these products, i.e., the workload for a given archive can fluctuate over time. In this and subsequent figures, each 'archive' represents one of the tape silos at the Goddard Space Flight Center DAAC.

Total tape mount rates (reads+write) never exceed 50% drive capacity, and on average, are around 10% of the robotics capacity. Mount rates vary by time of day and are highest during working hours (see Figure 10), presumably because user accesses are highest during that time. Figure 10 again shows the wide fluctuation of tape workload across archives. The average size of the products stored in an archive varies by archive; and the ratio of ingest to access load varies as well. As a result, mount rates are not necessarily proportional to data volumes.

Figures 11 and 12 show archive read and write statistics over a 6-month period at Goddard in terms of the number of files retrieved and inserted. The average file size is about 60 MB. On busy days, reads and writes

combined exceed 100,000 files, or 6 TB. The figures again illustrate the daily fluctuations in access patterns and how the workload varies across archives.

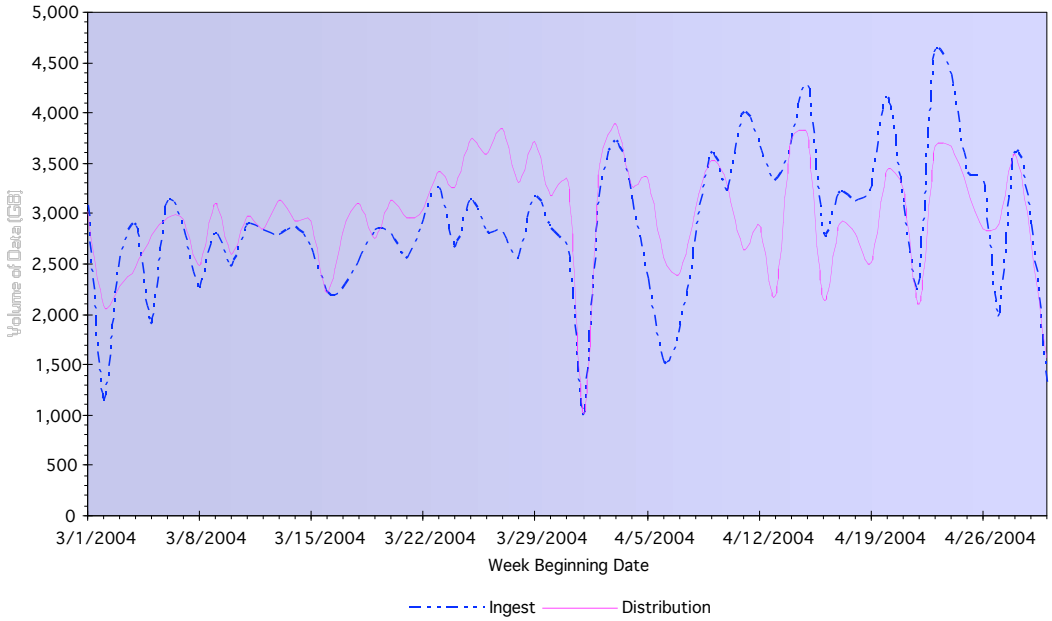

Figure 7. Ingest and distribution volumes

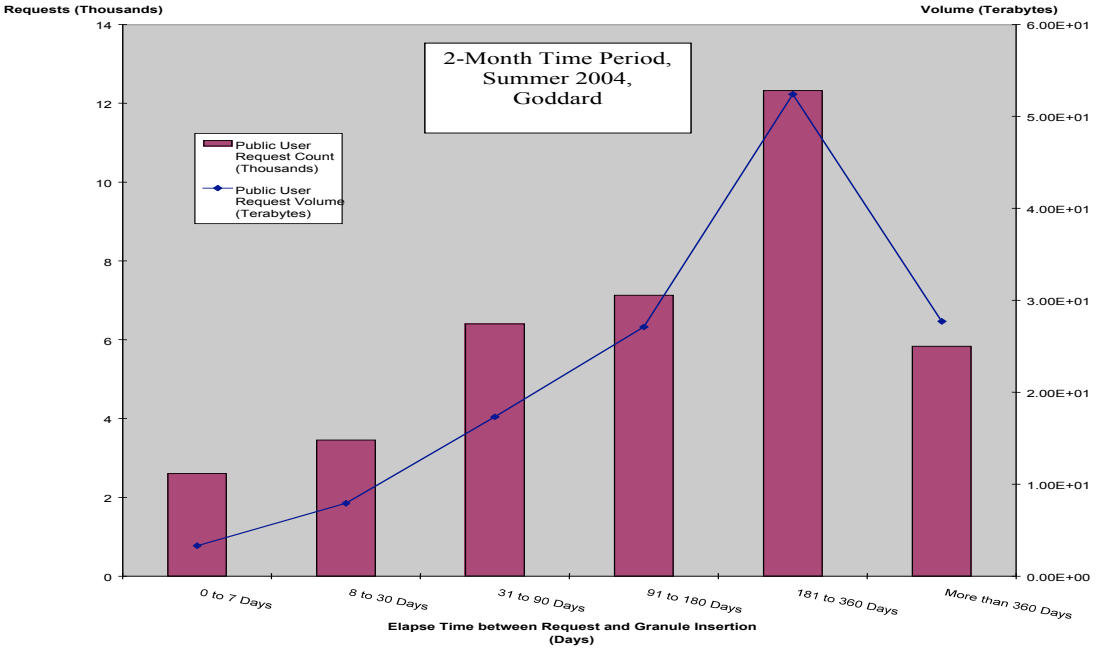

Figure 8. Requested data by data age

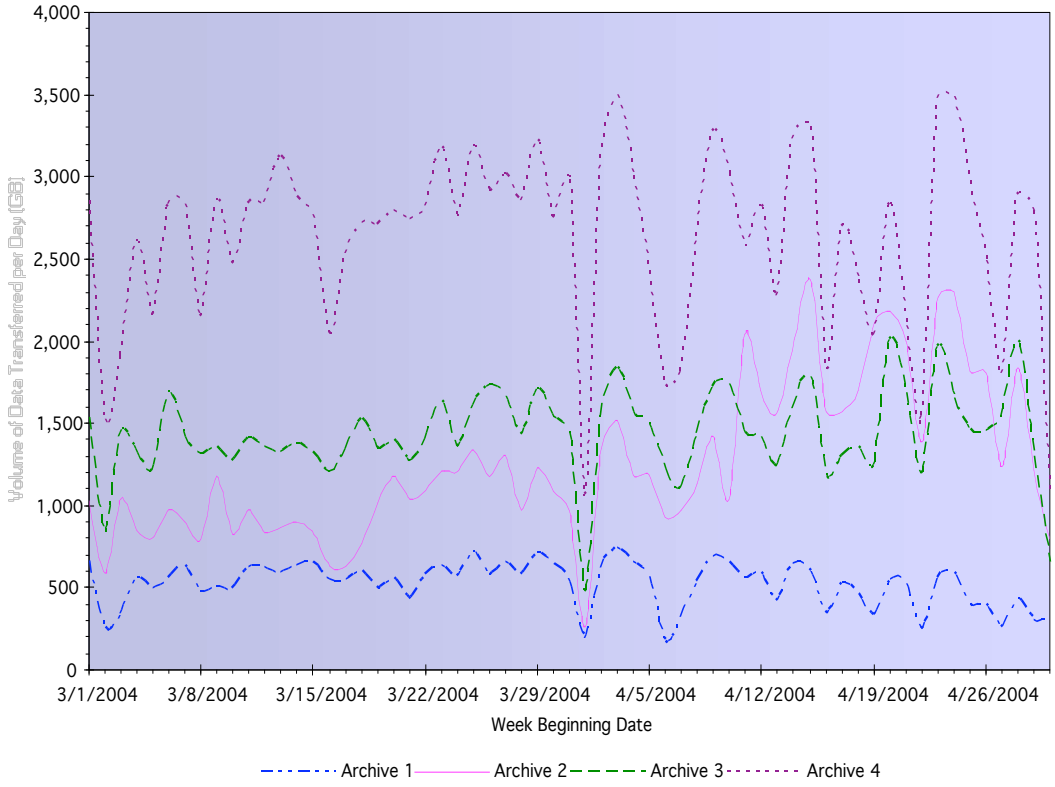

Figure 9. Daily archive throughput

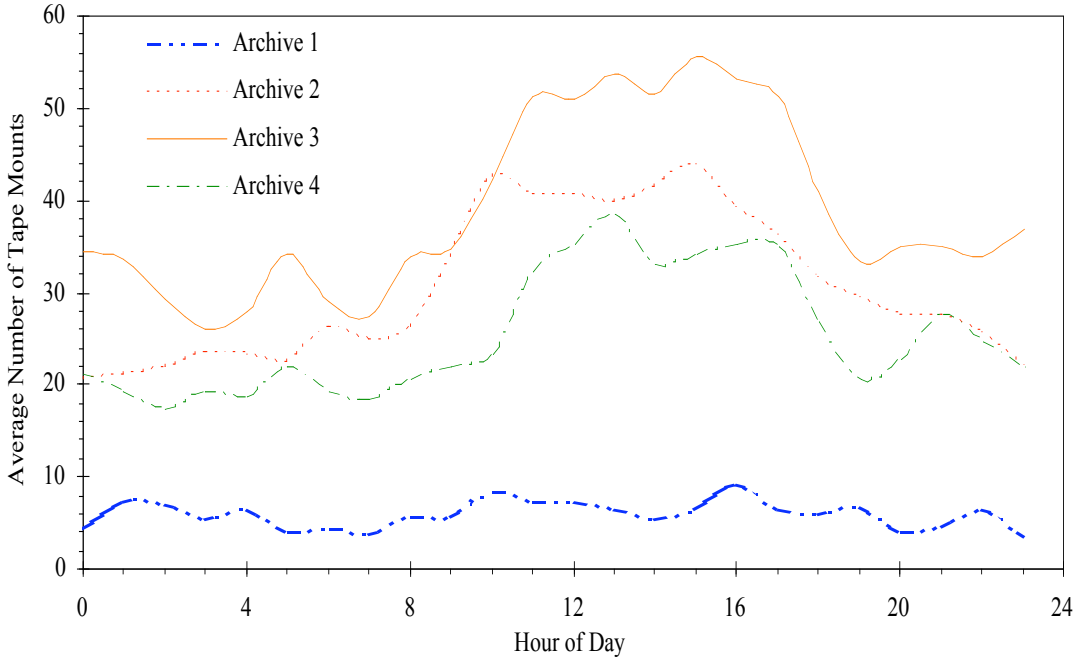

Figure 10. 24-Hour distribution of tape mounts

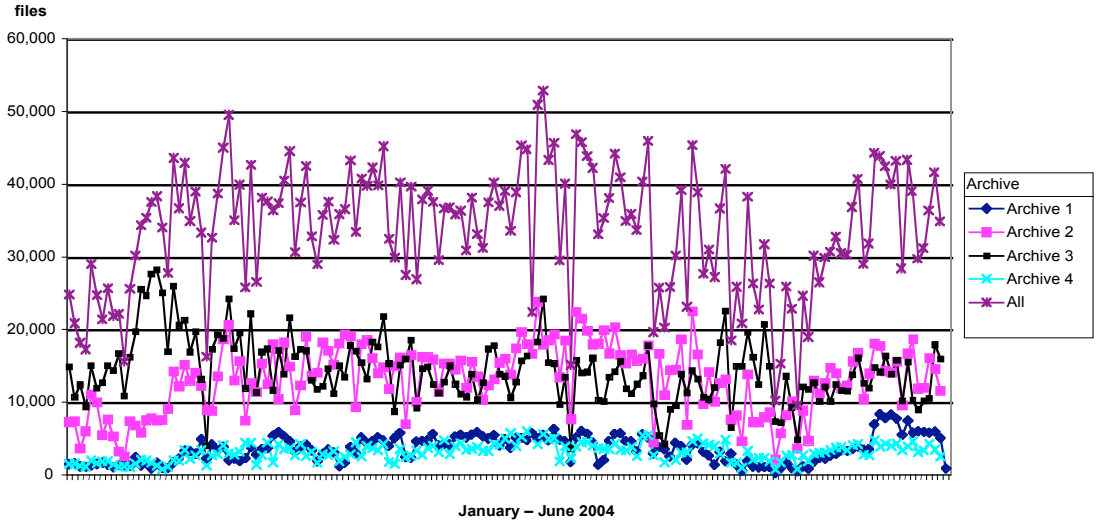

Figure 11. Daily file retrieval, January to June 2004

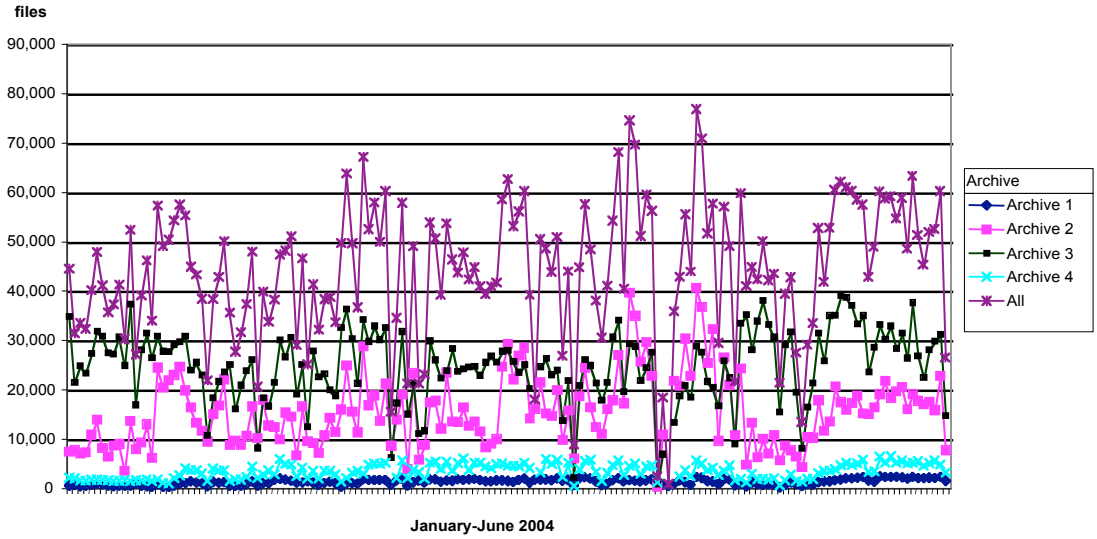

Figure 12. Daily file inserts, January to June 2004

Figure 13 shows the average transaction rate per second on our main database server at Goddard. Most of the database transactions are related to these five activities:

- Inserting the metadata
- Managing the inserts into the archive and the staging files
- Managing data retrieval from the archive for distribution purposes
- Managing the distribution activities and related staging files
- Responding to user searches

As was shown earlier (fig. 7), distribution volumes follow ingest volume closely. Since the first four types of transactions are directly related to ingest and distribution, the database workload should be closely related to ingest volumes. This is borne out by Figure 14 where we plotted the weekly CPU utilization and compared it to the insert rates.

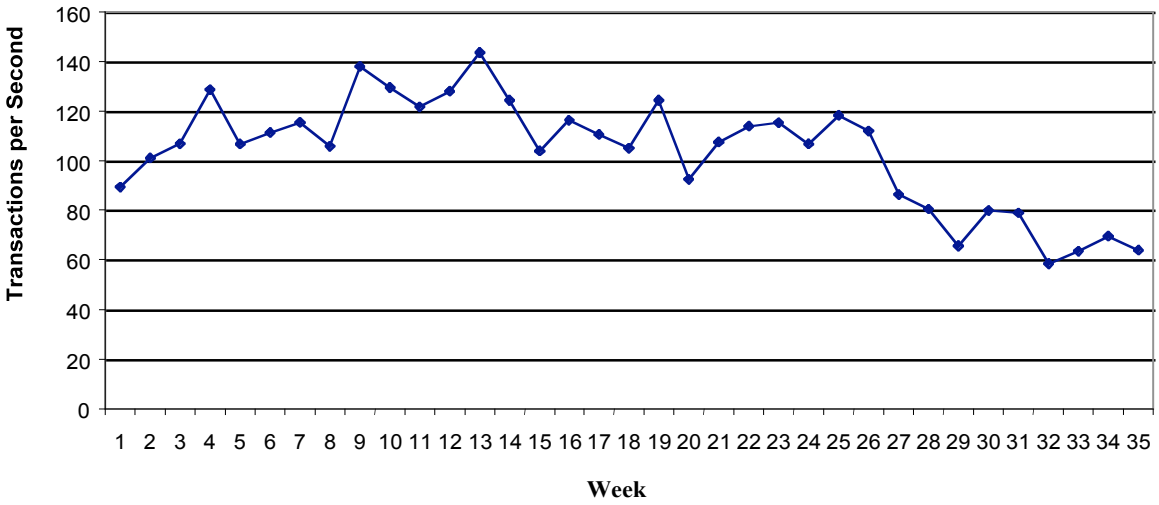

Figure 13. Average transaction rates at the metadata server during 2004

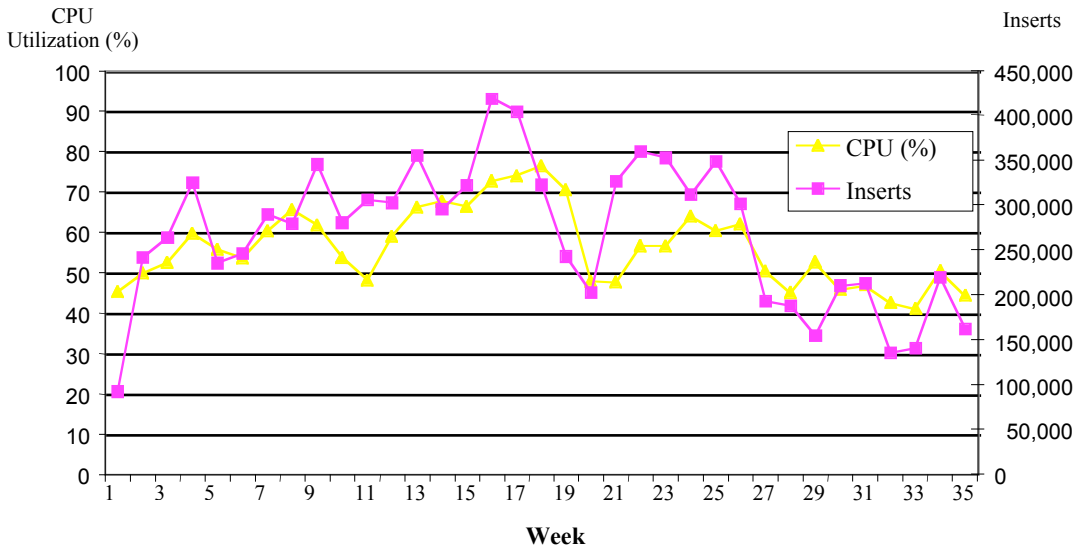

Figure 14. CPU utilization compared to insert rate during 2004

4. Technology change

There have been many changes in technology over the years that have affected the design of the system as discussed in section 2. In a project this large, technology migration must be carefully planned. Changes often take several months and sometimes years. Cost analysis is performed for both upgrades and migrations. Using current system metrics and careful future requirements analysis, the project determines when to migrate to new technology or whether current systems should be upgraded. For example, the need to migrate from 9940A to 9940B tape drives has been justified by the need for more storage capacity in the silos as well as a need for increased access rates and the vendor's plans for drive maintenance. Typically, migration and upgrades are performed at one data center at a time. Whenever possible, sharing and/or transferring hardware between data centers is performed to drive down costs. Over the years, the project has invested a lot of money in prototyping and benchmarking new technologies. In many cases, this has served to enable easier migration.

Table 2: Technology changes in EOSDIS Archive from 1995-2005

Technology Components	Design components of 1995	Operational components of 2005	Comments
Tape Technologies	Redwood D3 tape drives from STK/50GB cartridge for science products 9840 drives from STK/20 GB cartridge for small, frequently accessed products	9940A from STK/60 GB cartridge and at some data centers the 9940B/200 GB cartridge for science data Small frequently accessed data is now on SAN disk arrays	D3 tape drives proved to be unreliable, performance of 9940 drives has been superior. Total 9940A=101 drives Total 9940B=47 drives Total number of tapes= 45,580
Robotics	EMASS archive tape library(AML/2-Tall Quadro Tower, 2 Arms) w/HP optical drives, STK Archive tape library (Powderhorn/Wolfcreek)	STK Archive tape library (Powderhorn/Wolfcreek) at this point we have 14 silos between the 4 data centers and the contractor's facility. Silo capacity: 5528 tapes each	Original concept was to have a robotics system that could host various types of media. In the long run, it was more cost effective to have a single media type
Removable media for distribution to users	Floppies, 8mm, 4mm,CD	CD, DVD, flashdisk	Removable media is useful solely for distribution and offsite backup, not for archives.
Working Storage (the disk space that caches data to the tape silos)	SCSI attached, RAID disk arrays (4.3GB-9GB)	Fibre Channel and Storage Area Network disk arrays	Over the years, we have steadily increased the size and speed of the working storage area.
Memory used by processing hardware	64MB, 256 MB	Multiple Gigabytes	Memory management is critical to the performance of many of the COTS software, including the database management system
Distributed file systems	Monitoring NAS technologies DCE, Object-oriented DCE	Bulk Data System (BDS) , sockets	It was found that complex middleware introduced many COTS dependencies and increased system upgrade costs. DCE was replaced with a simple, high-performance socket-based middleware to resolve this.
Network	Ethernet FDDI/HiPPI	Gig-Ethernet	Continues to improve
Security	FTP for file transfers; DCE Cell configuration; Considered Kerberos	Secure File transfer High-performance firewalls	Security patches now require almost full-time support
The Web and access in general	Not a major factor in the design of the archive storage systems. Laptops, the web and EOSDIS were leading edge technology	PDA's, clusters, GOOGLE, grids, iPODS, wireless and mobile access are all leading edge technology now.	Hard to predict what will be leading edge technology in 2015

It is better for readers to find out about future technology trends in these areas from expert sources. However, there are a few specific areas of technology evolution that should be mentioned here. By 2015, we speculate that operations of our archive systems will be simpler resulting in reduced operations costs. It is an overall goal to maximize access and management of the archives at the DAACs. We are moving toward more online archives rather than nearline. The overhead associated with managing tapes and tertiary storage systems is high and a more cost effective approach would be to use tertiary storage systems solely as backup. By

2015, it is possible that the project manager will find it to be cheaper to replace hardware in the EOSDIS archives rather than pay for staff to tune hardware and software. By 2015, we hope that we will not need as many specialized staff members (systems administrator, SAN administration, network administrator, and so on) to maintain the archive systems. Hardware migration is inevitable and will occur well before 2015. However, we are looking for automatic migration solutions to enable new archive technologies. While we increase the portion of the archive available online, there are also issues with larger disk archives. First lower costs for online data including the footprint and energy utilization will be key to our environment. We are concerned with the speed with which data corruption can occur in disk based systems, and there exists limited discovery and recovery tools to resolve these problems. Solutions need to be developed that address the following four issues:

1. user errors (i.e. accidentally deleting a file or many files),
2. operator errors (i.e. accidentally initializing a disk, or rezoning a disk incorrectly)
3. malicious user problems (i.e. viruses)
4. system problems (both hardware or software).

In addition disk based solutions are generally vendor dependent, in that file systems are not transportable and that efforts need to be placed in developing file systems and object oriented storage paradigms that provide interoperability. Turnkey SAN solutions will be needed in the future to continue to make them viable online disk solutions.

5. Lessons learned

During the past sixyears, we have delivered five major software releases, numerous patches and performed over 150 COTS upgrades, including technology insertion in most areas of the system (archive tape drives, disk storage, networks, processing platforms). During that time, we have learned a number of valuable lessons about data integrity, the challenges of scaling system capacity and throughput, and how to manage change. Following are key lessons learned related to ECS data storage.

Data corruption happens

In general, one thinks of hardware and networks as reliable. That is, they work most of the time and when they do fail, they provide an error indication. However, if you move enough bytes of data around on a daily basis, you will discover over time that silent (undetected) data corruption can and will occur. Over the operational life of ECS, we have had a handful of silent data corruption issues where data has been incorrectly transmitted or stored without any error indication from the hardware or operating system. Problems have ranged from failed storage processors, an improperly seated card, to a COTS software error. Perhaps the most surprising case involved a network router that intermittently corrupted packets in a way that was not detected by the TCP protocol checksum. We would encourage anyone transmitting large volumes of data over TCP/IP to read [1, 2] on the reliability of the TCP protocol. In order to mitigate this problem, we believe that data systems must implement an end-to-end capability for verifying the integrity of each granule stored. For example, a checksum should be computed when a granule is created, travel with the granule throughout the system, and be verified when the granule is written to or read from the archive. The ECS ingest protocols now support data provider generated file checksums and will verify these checksums on ingest and re-verify the checksums each time a file is read from the archive. In addition, file checksums are included with the metadata distributed along with each granule order so that users have the option of re-verifying file integrity after receipt of data.

Data migration is a continuous operations function

When we began ECS in the mid-1990s, we knew that technology insertion would be an ongoing activity and that our architecture and processes needed to be able to accommodate it. However, we did not fully realize the operational impact of technology insertion with petabyte-scale archives. We have completed one archive tape drive technology insertion (from D3 tape to 9940A tape in our silos) and are partially through our second (9940A to 9940B). The first transition required migrating 350 TB of data from old media to new media and took over a year to complete. The current migration activity is being done on a silo-by-silo basis over a three-year period and will require migrating over 2.5 PB of data. When the next transition begins, data holdings will be large enough that by the time the data migration completes, it will be time for the next transition to begin. At this point, data migration will become a continuous operations function. This has a number of implications for petabyte-scale archive design and operation. System architectures must accommodate multiple generations of archive technology and provide highly automated support for data migration. Systems must be sized to accommodate both the expected production workload and the data migration workload. Resource management functions must allow the operations staff to easily throttle data migration workload in order to handle peaks in the production workload and the necessary operations staff must be allocated to support data migration. Finally, when data preservation activities

are considered, data migration might involve more than copying and verifying data objects. Format conversions may be required or old formats may need to be tested to ensure correct operation with current applications. This is all typically done as part of the data migration process.

Scaling to petabytes of data

The original system design has held up well as data holdings have grown to multiple petabytes and ingest and distribution workloads have increased to multiple terabytes per day. Multi-threaded applications are used to implement ingest, metadata management, archive management, and distribution functions. Several techniques can be used to handle workload or storage increases including configuring more threads for existing application instances or adding more application instances. In the archive management area, we have found that it is important to carefully predict workload when deciding whether or not to grow existing File Storage Management Systems (FSMS) and tape silos or to add new ones. For example, if the FSMS presents data holdings as a file system, then it's important to understand how many files might ultimately be in the file system and how well the FSMS database will scale. When we first deployed, maximum tape silo capacity was 450 TB (with data compression) or about 9 million files given our average file size. Today, advances in tape drive technology have led to tape silo capacities of over 1.5 PB or about 30 million files. Having this number of files in a single file system begins to cause issues with the length of time to perform FSMS functions and has led us to configure multiple FSMS instances so that there is one per tape silo.

Balancing workload across tape libraries

In the early years after deployment, sites had only one or two tape silos and workload was light. However, as workload ramped up and additional tape silos were added (our largest site has five), we discovered that it was difficult to keep all of the silos busy all of the time. The problem was that ingest and (particularly) distribution workload was unpredictable and sometimes tended to cluster in a single silo. Our resource management scheme was first-in-first-out (FIFO) based on request priority. If we received a number of ingest or distribution requests for data types stored in a single tape silo then our application servers would get tied up processing those requests and system throughput would be limited to the throughput achievable through that silo. Recent system releases have resolved this issue on the distribution side. We have implemented cross-request optimization. This involves looking across all of the pending orders and determining the list of tapes containing all of the products being requested. With this information, the system's order management service will attempt to keep all of the tape drives in all of the tape silos busy without overloading any individual silo. This approach has two advantages. It balances the workload evenly across tape silos and it minimizes the number of tape mounts performed by pulling all requested products when a tape is mounted. Future system releases will provide the same cross-request optimization for the ingest workload.

Verifying consistency of metadata and archive data

Over time, we have discovered that it is vital to have a utility that can verify the consistency of our science metadata database, described in section 2, with the files in the archive. System anomalies can result in metadata database entries that do not have corresponding files in the archive or archive files that do not have corresponding metadata entries. Such system anomalies can occur in the course of routine operations that are interrupted by unexpected hardware failures. It is important to have a utility that can be run regularly to identify these situations so that they can be corrected. As the size of the data holdings increases, it is important that this utility be able to be run in an incremental fashion, For example, consistency should be checked on all data that was inserted over the past week.

Organizing data for efficient insertion, access and deletion

As discussed in Section 2, the original data allocation strategy was designed to minimize the number of tape mounts needed during insertion and have reasonable data collocation for data access, especially as needed to support reprocessing. This strategy involved writing all data for a given data type to the same set of tapes. However, when both routine processing and reprocessing occurred concurrently for a data type, this resulted in the intermixing of two observations times (forward processing time range and reprocessing time range) on the same tape. This strategy turned out to be inefficient when it came time to delete old versions of data, because it didn't allow tapes to be freed up for reuse in a timely fashion. The deletion process required performing a rolling delete of the old version of a data granule six months after a new version had been created via reprocessing. Since each tape contained granules from both the forward processing time range and reprocessing time range, only a portion of the granules on each tape were deleted as a result of the rolling delete process. All of the granules created by forward processing would not get deleted until months after the granules created by

reprocessing. This resulted in a large number of partially filled tapes and the only way to reclaim tapes for reuse was for the operations staff to manually copy over undeleted granules to new tapes. As a result, in June 2004, the ECS archive tape allocation strategy was changed to enable the operations staff to store data from the forward and reprocessing streams on different tape sets. While this approaches increases the number of tape mounts needed to insert data, it virtually eliminates the need to move data between tapes before they can be reused and makes the implementation of rolling delete strategies operationally viable.

Providing online access to data

As the cost of disk storage decreased, it became feasible to provide online access to a portion of ECS data holdings. In 2001, the Data Pool [3, 4] capability was introduced. Data Pool provides a large (> 60 TB at our largest site) cache of popular products that can be directly downloaded via the web or ftp. A simple web drilldown interface is provided that enables users to rapidly locate data, view metadata and browse data through their browser and then download data granules of interest. Granules and metadata files can also be downloaded directly via ftp.

6. Conclusions

In 1995, the big question facing the EOSDIS team was how to store and manage all of the data that the EOSDIS would have to handle from all of its missions. The problem seemed enormous as described in the paper that appeared at the Fourth Mass Storage Conference [5]. By 2005, our big question is how to help our users find the data that they need and how to get it to them. Further, we are now confronted with problems of how to maintain a viable archive on an operational budget. Fortunately, there are many options for the future and some clear directions that can be taken to move to the future.

As was said earlier, the EOSDIS Core System has been successfully operating for six years. The system has proved to be scalable and flexible. ECS continues to strive for an automated mass storage system and to find cost efficiencies wherever possible. We plan to increase the footprint of data pool over the next years. The use of data pool/disk technology is desirable not only to the data centers because it is cheaper to operate but also to users who acquire data more quickly. Another direction that we are continuing to build on has to do with science metadata. We are looking at several avenues in which improved metadata and access methods will help users with data discovery. There is an increasing need for access to flexible metadata not only to search for the full dataset in the inventory but also to serve as viable data itself. We are also looking at semantic enhancements to metadata that describes the science data to all types of users, not just NASA science teams. There is a clear need to support the evolution of data as it is abstracted into new types of information and models. Yet another direction that the project is examining is the need for long-term data stewardship. Stewardship ensures the quality and integrity of the data product to the users. Back in the 1990s, it was decided that after the NASA mission was complete, data would be transferred to the archives at NOAA and USGS. Ultimately, the stewardship of this invaluable data collection will be a team effort between several government agencies and contractors. However, the success of the past ten years is surely a measure of the future.

7. Acknowledgements

The authors would like to acknowledge all of the efforts of the many people who have worked on EOSDIS. Special thanks go to the NASA staff, the EOSDIS contractors, and the staff of the National Snow and Ice Data Center, the Land Processes data center, and the data centers at Langley and Goddard. We especially acknowledge the support of Mike Moore in the development of the system.

References

[1] Paxon, V. End-to-End Internet Packet Dynamics. *IEEE Transactions on Networking* 7, 3 (June 1999), 277-292.

[2] Stone, J., Partridge, C. When the CRC and TCP Checksum Disagree. *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication* (August 2000), 309-319.

[3] Moore, M., Lowe, D. "Providing rapid access to EOS data via Data Pools", *Proceedings of SPIE, Earth Observing Systems VII*, Seattle, WA, July 7-10, 2002.

[4] Carr, K., Meyer, R., Duma, C., Bryant, K., Hartranft, R., Bergman, R., Fox, Stephen. "Storage and Retrieval of Spatially-Qualified Data from NASA's EOSDIS Data Pool", *International Geoscience and Remote Sensing Symposium (IGARSS) 2003*, Toulouse, France, July 21-25, 2003.

[5] Caulk, Parris M., "The Design of a Petabyte Archive and Distribution System for the NASA ECS Project," *Fourth NASA Goddard Conference on Mass Storage Systems and Technologies*, (March 28 - 30, 1995), 7-17.