EOSDIS: Archive and Distribution Systems in the Year 2000

Jeanne Behnke
Goddard Space Flight Center - Code 423
Greenbelt, MD 20771
jeanne.behnke@gsfc.nasa.gov
tel: +301-614-5326
Alla Lake
Lockheed Martin Corp
1616 McCormick Drive, Upper Marlboro, MD 20774
alake@eos.hitc.com
tel: +1-301-925-0626
fax: +1-301-925-0651

Abstract

Earth Science Enterprise (ESE) is a long-term NASA research mission to study the processes leading to global climate change. The Earth Observing System (EOS) is a NASA campaign of satellite observatories that are a major component of ESE. The EOS Data and Information System (EOSDIS) is another component of ESE that will provide the Earth science community with easy, affordable, and reliable access to Earth science data. EOSDIS is a distributed system, with major facilities at seven Distributed Active Archive Centers (DAACs) located throughout the United States. The EOSDIS software architecture is being designed to receive, process, and archive several terabytes of science data on a daily basis. Thousands of science users and perhaps several hundred thousands of non-science users are expected to access the system. The first major set of data to be archived in the EOSDIS is from Landsat-7. Another EOS satellite, Terra, was launched on December 18, 1999. With the Terra launch, the EOSDIS will be required to support approximately one terabyte of data into and out of the archives per day. Since EOS is a multi-mission program, including the launch of more satellites and many other missions, the role of the archive systems becomes larger and more critical. In 1995, at the fourth convening of NASA Mass Storage Systems and Technologies Conference, the development plans for the EOSDIS information system and archive were described [1]. Five years later, many changes have occurred in the effort to field an operational system. It is interesting to reflect on some of the changes driving the archive technology and system development for EOSDIS. This paper principally describes the Data Server subsystem including how the other subsystems access the archive, the nature of the data repository, and the mass-storage I/O management. The paper reviews the system architecture (both hardware and software) of the basic components of the archive. It discusses the operations concept, code development, and testing phase of the system. Finally, it describes the future plans for the archive.

Introduction

Earth Science Enterprise (ESE) is a long-term NASA research mission to study the processes leading to global climate change. The Earth Observing System (EOS) is a NASA campaign of satellite observatories that are a major component of ESE. The EOS Data and Information System (EOSDIS) is another component of ESE that will provide the Earth science community with easy, affordable, and reliable access to Earth science data. EOSDIS is a distributed system, with major facilities at data centers located throughout the United States.

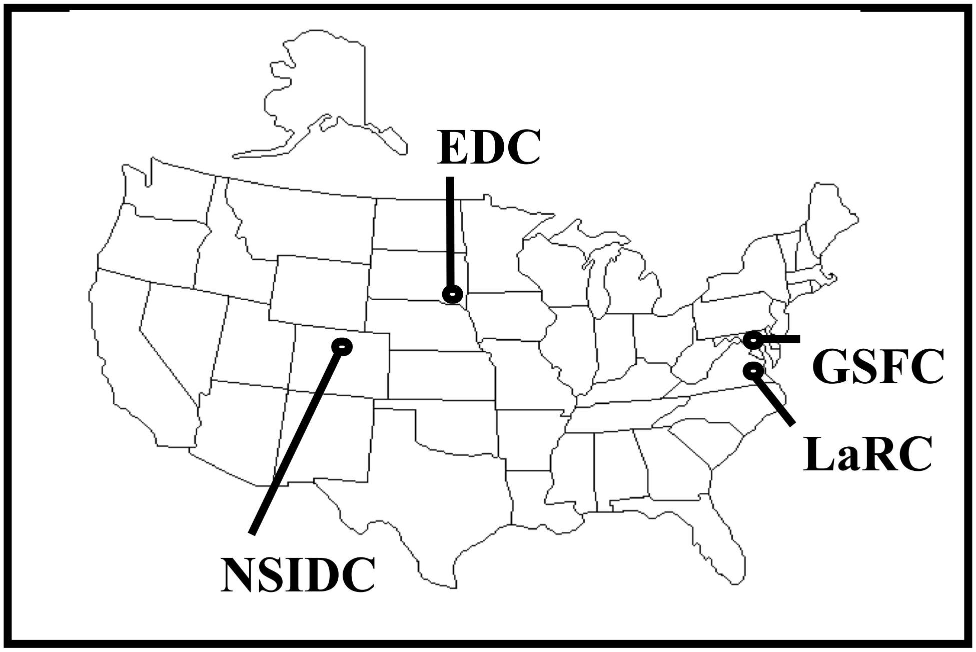

Figure 1. Locations of EOSDIS Distributed Active Archive Centers

In this paper, we describe the archive and distribution operations at four Distributed Active Archive Centers (DAACs). These DAACs 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). The EOSDIS software architecture is being designed to receive, process, and archive several terabytes of science data on a daily basis. Thousands of science users and perhaps several hundred thousands of non-science users are expected to access the system.

The first major set of data to be archived in the EOSDIS is from Landsat-7. Landsat-7, an Earth imaging satellite launched on April 15, 1999, provides repetitive, synoptic coverage of continental surfaces; spectral bands in the visible, near-infrared, short-wave, and thermal infrared regions of the electromagnetic spectrum; average spatial resolution of 30 meters (98-feet); and absolute radiometric calibration (http://landsat.gsfc.nasa.gov).

Following Landsat-7, the Terra satellite (formerly known as EOS AM-1) was launched on December 18, 1999. Terra is uniquely designed for "comprehensive" Earth observations and scientific analysis, covering science priorities as land cover change and global productivity, seasonal-to-interannual climate predictions, natural hazards, long-term climate variability, and atmospheric ozone (http://terra.nasa.gov). With the Terra launch, the EOSDIS will be required to support approximately one terabyte of data into and out of the archives per day. The next big mission is the Aqua satellite to be launched in December 2000 (http://aqua.gsfc.nasa.gov). Moreover, EOS is a multi-mission program that includes several more Earth study campaigns and satellites through the year 2011. Given this extended time frame, the role of the archive systems becomes larger and more critical. In 1995, at the fourth convening of NASA Mass Storage Systems and Technologies Conference, the development plans for the EOSDIS information system and archive were described [1]. Five years later, many changes have occurred in the effort to field an operational system. It is interesting to reflect on some of the changes driving the archive technology and system development for EOSDIS.

The focus of this paper is the description of the Science Data Processing System (SDPS) segment of EOSDIS, with particular attention to the ingest, archive and distribution processes and components. The SDPS system will be required to manage, store, retrieve, and process more than a terabyte of data per day at its data centers. As Table 1 illustrates, the projected capacity required by the project during 2000 is quite formidable. Across the data centers, the expectation is to archive on the order of 1.5 TB per day and 16,500 granules. A granule is the smallest package of data made available by EOSDIS. A granule can contain 1 or many files. Another important distinction is made between a full dataset and a “browse” dataset. Browse datasets can be thought of as small examples of the full resolution data. They are used by scientists to quickly determine whether a particular dataset is useful without having to look at its entire contents.

Data Center	Archive Volumes GB/Day	# of granules per day	Archive Volumes In TB per year	# of Granules cumulative per year	Distribution via Network GB/day	Distribution via tape GB/day
EDC	522	6886	190	2,513,390	194	159
GSFC	688	5545	251	2,023,925	226	226
LaRC	312	2945	114	1,074,925	102	102
NSIDC	22	1083	8	395,295	6	6
Total	**1544**	**16459**	**563**	**6,007,535**	**528**	**493**

Table 1. Projected capacity through the end of 2000

In addition to designing and providing a comprehensive data retrieval and processing system, the SDPS is tasked to be a flexible, scaleable and reliable system. The architecture should be capable of supporting:

- new data types with minimal software modifications
- new data centers that will not require new code and software agreements
- standard interfaces (HDF-EOS) enabling coordinated data analysis
- data access from a wide variety of users (e.g., kindergarten teachers, as well as college professors)
- technological advances and the infusion of new COTS products and techniques (e.g., new file storage management systems)
- inevitable change and new additions

To meet the challenge of the SDPS, the EOSDIS Core System (ECS) was designed under contract to NASA by the Raytheon Systems Company/Landover MD. Lockheed Martin Corporation designed the archival component of the system encompassing Ingest, Storage Management, and Distribution, under a subcontract to the Raytheon Systems. It is an enormous development effort for the entire ECS comprises 75 COTS packages, about 1 million lines of code and the efforts of approximately 220 developers. In this paper, we describe how the archive and distribution systems work for the ECS.

The ECS system is designed at a central development location and then distributed and installed at the various DAAC sites. Each of the DAACs has a different area of science emphasis and the system to be deployed is adapted to that need. For example, not all DAACs will have the same archive size requirements. However, the software system works the same way at all DAAC sites. The science datasets supported by this SDPS also vary in size and type. The SDPS is composed of six major subsystems shown in Figure 2, ECS Context Diagram.

1. INGEST subsystem - receives data from external and internal sources and submits them for storage into the archive
2. DATASERVER subsystem - archives and distributes data
3. PLANNING subsystem - develops plans for producing data products (level 0 to level 1)
4. DATA PROCESSING subsystem - manages, queues and executes processes for the generation of data products
5. INTEROPERABILITY subsystem - provides the software infrastructure for the communications between clients and servers in the system
6. DATA MANAGEMENT subsystem - supports the location, search, and access of data and services.

This paper will principally describe the Data Server subsystem including the description of how the other subsystems access the archive, the design of the data repository, the mass-storage I/O management, and archive operations. The paper will review the system architecture (both hardware and software) of the basic components of the archive

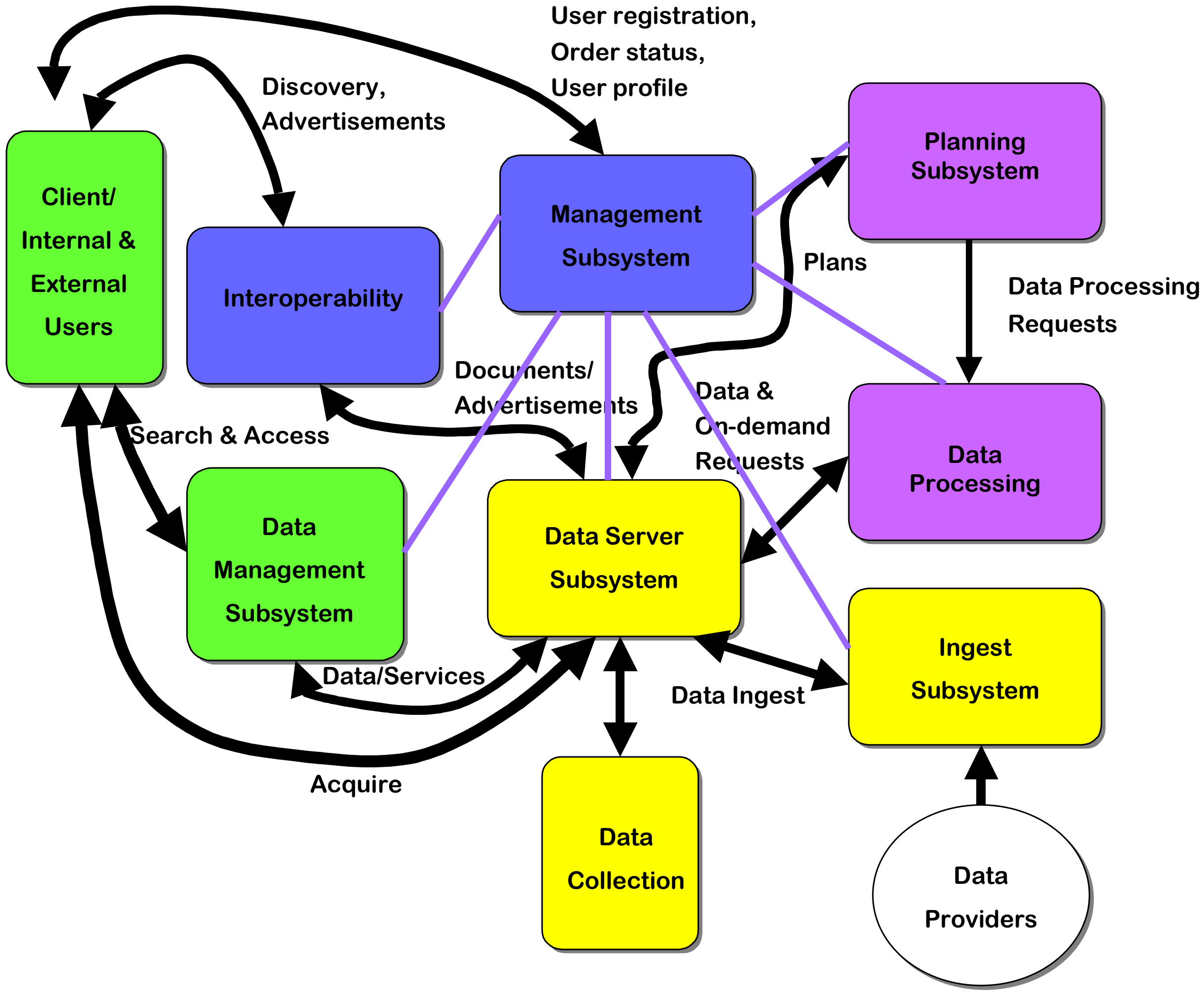

Figure 2. ECS context diagram

It will discuss the operations concept, code development, and testing phase of the system. Finally, it will describe the future plans for the archive.

Data Server Subsystem

The Data Server Subsystem (DSS) provides the data storage and management functions including archiving EOS data, managing and searching the archive, and resource staging. It stores, searches, retrieves and distributes EOS data. The DSS interfaces with virtually all ECS subsystems and components. It is composed of several internal subsystems and constitutes the largest software and hardware segment of the entire ECS. The subsystems internal to DSS include the Storage Management Configuration Item (CI), Data Distribution CI, and the Science Data Server CI.

As with all major systems in ECS, the DSS was written in C++ using an object-oriented software methodology. The DSS uses the Distributed Computing Environment (DCE) for its infrastructure and ClearCase to manage the software configuration. Many other commercial packages are used to develop, build and operate the system. The DSS contains 252,000 lines of custom code. Ingest, which is a separate subsystem designed to load the archive and enter the metadata into the inventory tables, is composed of 83,000 lines of code. The system is extensively tested prior to being fielded. The ECS was originally developed on small workstations that didn't adequately emulate the hardware being fielded at the DAACS. During the course of the five years, it became clear that the development team would require a complete archive system that duplicates the configuration of the archive systems at the larger DAACs in order to develop and test the software systems effectively. This archive system, called the Performance Verification Center, was created to not only field new versions of the software but also to provide the ability to troubleshoot and tune the system to maximize performance.

The Science Data Server CI subsystem in the DSS provides the entire ECS system with a catalog of data holdings organized by Earth Science Data Types (ESDT). The ESDT includes not only the data type definitions but also service functions that can be performed on that specific data. The Science Data Server manages and provides user access to data collections through its catalog of metadata, principally using the Sybase database management system. When another subsystem (for example, the Data Processing Subsystem) requests data from the archive, the request is sent to the Science Data Server subsystem. Science Data Server then initiates a request to the Storage Management subsystem, to allocate magnetic disk space for staging of that data and a request to the Data Distribution subsystem to stage and distribute the data appropriately to the requestor. The Data Distribution subsystem requests the data from the Storage Management subsystem. The Storage Management subsystem initiates the acquisition of the data from the physical storage in a robotic silo and stages the data in the appropriate disk space that it manages.

The Storage Management CI stores, manages, and retrieves data files on behalf of the other subsystems. It also manages all the magnetic disk space within the DSS. An archive software server is used to manage requests from the other subsystems to store or retrieve the archive data. A staging disk server is used to manage the files in the magnetic disk storage area and a pull monitor server is used to manage the files that are in the 'user pull' area. The magnetic disk is also used as a Storage Management disk cache, which is managed by a staging monitor server. The 'user pull' disk space is disk space allocated for users to FTP their requested data from the EOSDIS. A resource manager server has also been developed to manage the peripheral devices available at each data center. These include 8mm tape drives in stackers, D3 tape drives at EDC, and, in the near future, CD-ROM drives and DLT tape drives in stackers.

The Data Distribution CI formats and distributes data to users, either electronically or on physical media (e.g. 8mm tapes). It directs the Storage Management subsystem to place data in the desired location. The request could be to place the data on magnetic disk for another subsystem to retrieve it, to copy the data to tape, or to push the data via FTP to the user's workstation. This CI sends distribution notifications as the action is completed. The data distribution server provides control and coordination for data distribution through request processing. More on the request processing design can be found in the poster paper by J. Crawford presented at the Eighth NASA Mass Storage Systems and Technologies Conference. [3]

Data Repository

The Data Repository component includes the nearline storage system, cache magnetic disks, and servers required for storing and retrieving the EOS data, see Figure 3, Hardware Architecture Diagram. The architecture is specialized to insure high availability [4], high-speed access to the data in the nearline system. Because the amount of data to be stored in the EOS system is so large, the system had to be designed to use high storage density at low cost and, therefore, is a tape-based archive. In 1995, ECS was planning to purchase AML multi-media robotics from the EMASS corporation. As the requirements for the project evolved, the decision was made in 1996 to have a multi-vendor solution, with the EMASS Corporation supplying the AML multi-media robotics for the small file size Browse collection and Storage Tek Powderhorn silos for the large data archive. AML robotics are attractive, based on their support for drives and media from different vendors. After several months of evaluation, it was decided that it would be a cost benefit to trade the four AMLs for Storage Tek Powderhorns, which is our current configuration.

Today, the nearline storage system used by EOS is based on StorageTek Powderhorn silos as the hardware base and the AMASS (Archival Management And Storage System), a product of ADIC. StorageTek silos have been installed at the data centers: three silos at EROS Data Center (EDC), two at Langley Research Center (LaRC), one at NSIDC and four at Goddard Space Flight Center (GSFC). Each silo holds up to 6000 cartridge tapes,

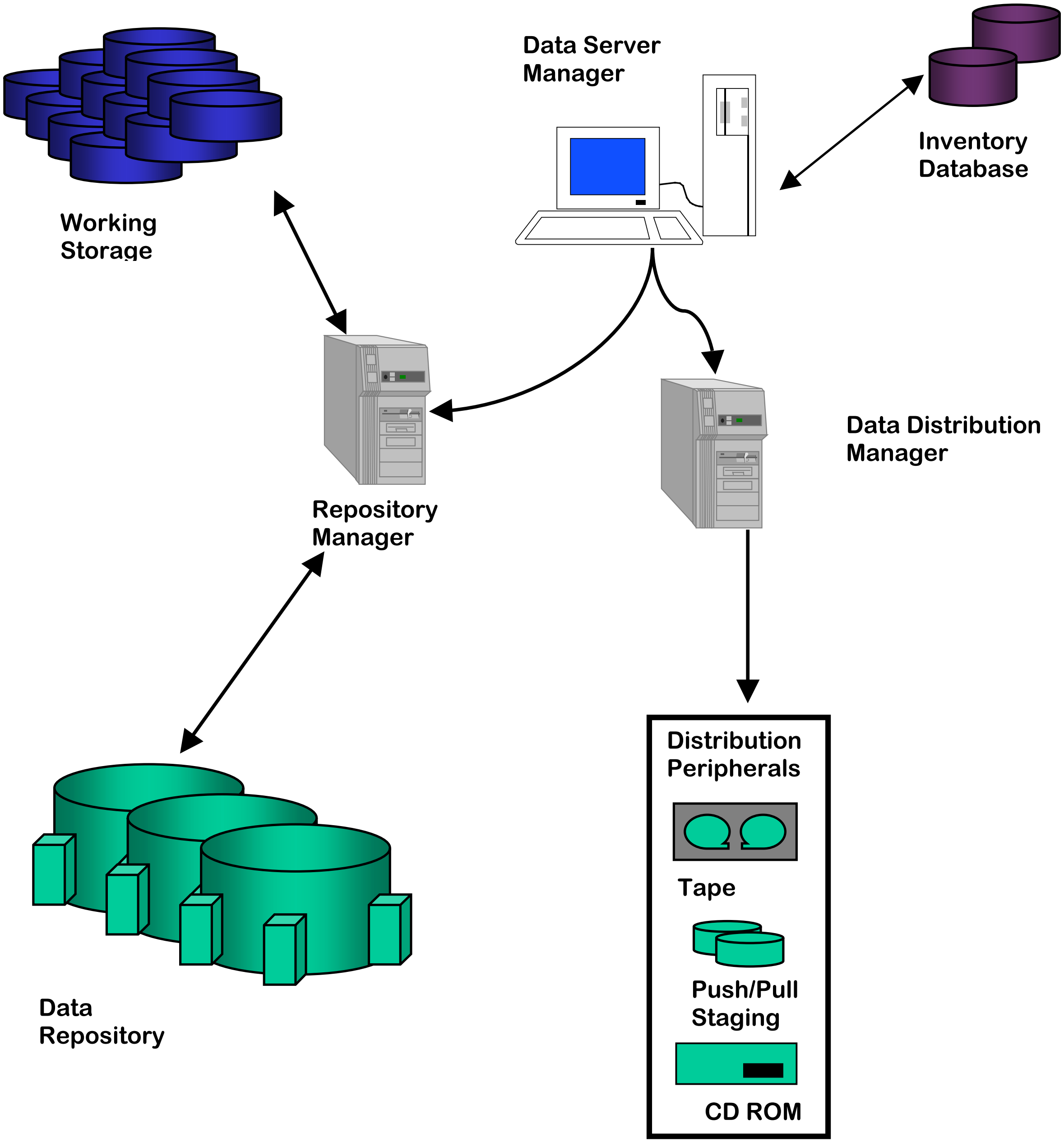

Figure 3. ECS Hardware Architecture Diagram

in our case 50GB D3 tapes, for an aggregate of approximately 300 TB in a silo. The actual number of media in a silo is derated by the number of attached tape drive enclosures and by a Plexiglas observation window, if used. On average, each EOSDIS silo filled exclusively with D3 media stores 270 TB of data without compression. Some data, however, is compressible at the drive. The Landsat collection is one example where realized data compression approximates 2:1. A silo containing Landsat data at the EDC will, when filled to capacity, store between 500 and 600 TB.

The number of tape drives attached to each of the silos varies depending on the data throughput requirements of an archival site**.** The silos at the larger sites, such as EDC and GSFC, run with eight D3 drives in each of the archive silos. A smaller site, such as NSIDC, has 3 D3 drives in its data silo. The drives are rated at a maximum sustained throughput of 11 MB/sec, but the observed effective rate with compression is near 16 MB/sec. The number of storage silos at a data center is also determined by the data center size, i.e. the cumulative size of the data holdings in storage. The physical storage of data is managed by the AMASS file storage management system.

Both the hardware and software system design supports the growth of ECS. NASA has planned to grow the system to support the archiving of future missions with additional hardware over the ECS program lifetime. For example, NASA purchased an additional silo for each of the larger data centers last summer. The silo was easily incorporated into the architecture. Another ECS design modification was to place the Browse Data Collection on STK 9840 tapes and house these tapes in the STK Powderhorn silos. That design was driven by both a very large accumulation (up to 30 TB at GSFC) of Browse data and the relatively small file size of each Browse file - in the 1 MB range. Since even at the larger sites the Browse collection will fill only part of the silo, the Browse silos may be used in a multi-media mode, outfitted with both 9840 and higher tape capacity drives and media. Just like D3, 9840 is a fast streaming drive, rated at 12 MB/sec maximum sustained throughput rate and capable of effective data rate of 17 MB/sec with compression. Unlike the Helical Scan D3, 9840 is a linear tape drive – more suitable for a start and stop operation mode associated with smaller data files.

Silicon Graphics (SGI) workstations were chosen as the platforms for managing the data repository. The current configuration assigns a single SGI Challenge host per STK silo in order to sustain the required data rates to and from the tape drives through the attached buffer RAID. Significant effort was expended in tuning the SCSI attached RAID to produce the desired effective data rates of 120 MB/sec per RAID subsystem [2]. Over the summer, ECS will migratethe existing archive servers to SGI Origin platforms. Although the combination of Origin servers and fibre attached RAID afford much faster data rates, the same 1 server per 1 silo ratio will be preserved initially to allow for redundancy.

Mass-Storage I/O Management

The greatest challenge for the DSS is the management of the massive I/O (multiple terabytes per day) between the archive and the ECS components requesting data actions. It must handle continuous requests from

1. The Data Processing subsystem for files needed for processing Terra or Landsat data and for storing products once they are created;

2. The Ingest subsystem to store data from data providers external to ECS;
3. The client subsystems for data ordered by users from the GUI front-end;
4. The ECS subscription service to send data to users when it gets stored in the database.

The physical storage of ECS data is managed via a “commercial off the shelf” (COTS) AMASS files storage management system by ADIC. The AMASS system runs on the SGI hosts and operates the STK hardware silos. The control of the robotic mechanism of the silo (loading and unloading of the tapes) is via the STK Automated Cartridge System Library Software (ACSLS) running on a SPARC5 SUN workstation. AMASS addresses the ACSLS through a network connection. The ACSLS controls the robot directly via an RS232 line.

AMASS is a direct access file system as opposed to a Hierarchical Storage Management (HSM) product. Its cache area serves as a write-through buffer. The size of AMASS cache is set independently at each data center and each particular server. It is determined on the basis of the expected storage and retrieval profile associated with the data types handled by the server. Predominant file size, many small files or many large files for example, plays a role in choosing the exact configuration. Although AMASS supports both FTP and NFS access to the archived data, NFS is used solely by the ECS system. Unfortunately, AMASS uses an internal database that is singlethreaded in the implemented version. The performance constraints that it places on the overall system are mitigated by using custom-code manipulation of AMASS and the hardware. AMASS uses an internal database for tracking file allocations to tape. This internal database is journaled. The location of tapes in the silo slots is tracked by an Oracle database in the ACSLS. To enhance performance, AMASS allows creation of specialized volume groups of tapes in the silos. In our case, these are created for particular EOSDIS data types.

Archive Operations

With the launch of Terra, archive operations for EOSDIS are now fully established at all the data centers. Many of the data centers are operational on a 24 x 7 basis, however, the ECS system has been designed with a ‘lights out’ approach. To be able to maintain the required ingest and distribution rates, the software is designed to be highly autonomous. The only area of operations requiring direct human involvement is hard media distribution, where the distribution media must be loaded and unloaded and packaged for shipment to the user.

A Systems Monitoring Center has been built at GSFC to monitor each of the data centers as well as to provide some special, centralized functions. Locally at each data center, all areas of operation are closely monitored, especially process and log monitoring. The operations system administrators and system engineers are automatically paged if specific events or error conditions are encountered. For example, a severe error in the archive systems will trigger an event ‘page call’ to the engineers. This enables staffing to be minimal even at the largest data centers during off-hours. Routine full system backup of the software is performed using DLT tape. The AMASS database (ORACLE) in the archives is backed up at some data centers as frequently as every hour. Backup of the actual data is determined at each data center and is dependent on the data type. It can be done as simply as setting aside an additional volume group in a secondary silo for a backup copy. One of the most important concerns in operating of the mass storage system for ECS was the ability to recover from tape errors. A recovery procedure that is combination of automated scripts, custom code, operator actions and vendor actions has been designed and tested.

Each data center supports three modes in its current system environment. There are two test modes and one operational mode. The routine work of the data center is performed in the operational mode. The two test modes are used for testing and installing software patches/releases and COTS patches/releases. For the ECS system, it was important to fully test the archive systems as much as possible prior to becoming operational. We tested both the hardware and software for functionality as well as performance. Many problems and features of the systems were discovered during the test phase. The archive vendor was notified and fixes were supplied. In some cases, ECS compensated with custom code. Testing the archives required almost two full-time engineers in advance of being operational and will continue to require test engineers during the operational EOS lifetime. Each data center also requires support from very experienced archive engineers.

Conclusions

The focus of this paper is the Data Server subsystem, the archive component of the EOSDIS Core System (ECS). It comprises the data repository, the mass storage I/O management and the software configuration items needed to manage, store, retrieve, and process massive amounts of EOSDIS data on a daily basis. The Data Server Subsystem design is constrained in many areas by the schedule pressures, cost considerations, realities of implementing large and very complex system, and sheer technical limitations existing at the time of initial design. Even so, the initial performance results are satisfactory to meeting the data flow and operating requirements. In order to continue to meet the requirements of the future missions and the anticipated growth in user demands, the system must continue to evolve. The ECS design allows for such evolution in areas of both the custom implementation and replacement of outdated COTS technologies by their successors. Several evolutionary steps, such as robotic technology replacement and migration to the new SGI server model, have already been undertaken or are in the process of being undertaken. In the foreseeable future, the hardware architecture is limited to tape archives, although a plan for regular migration to alternative tape media is being considered. Some changes may be anticipated by evolving storage area networks, file storage management software and disk technology. The very design alterations that took place during the course of the project represent the systems capacity to evolve. We have described the archive and distribution systems for the EOSDIS Core System (ECS) for 2000 in this paper, however there are many other components to the system. Please feel free to contact the authors for further information.

References

[1] Caulk, P.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, College Park, Maryland, March 1995.

[2] Lake, A., “Performance Tuning of a High Capacity/High Performance Archive for the Earth Observing Systems Project”, Sixth Goddard Conference on mass Storage Systems and Technologies, College Park, Maryland, March 1998.

[3] Crawford, J.M., “A Scalable Architecture for Maximizing Concurrency”, Eighth NASA Goddard Space Flight Center Conference on Mass Storage Systems and Technologies, College Park, Maryland, March 2000.

[4]Lake, A., Crawford, J., Simanowith R., Koenig, B., “Fault Tolerant Design in the Earth Orbiting Systems Archive”, Eighth NASA Goddard Space Flight Center Conference on Mass Storage Systems and Technologies, College Park, Maryland, March 2000.