Architecture & Evolution of Goddard Space Flight Center Distributed Active Archive Center

Jean-Jacques Bedet, Lee Bodden, Wayne Rosen, Mark Sherman
Hughes STX
7701 Greenbelt Road, suite 400
Greenbelt, MD 20770
301-441-4285 Fax (301) 441-2392
{bedet, bodden, rosen, sherman}@daac.gsfc.nasa.gov

Phil Pease
NASA/GSFC
Greenbelt Road
Greenbelt, MD 20771
301-286-4418
pease@daac.gsfc.nasa.gov

Abstract

The Goddard Space Flight Center (GSFC) Distributed Active Archive Center (DAAC) has been developed to enhance Earth Science research by improved access to remote sensor earth science data. Building and operating an archive, even one of a moderate size (a few Terabytes), is a challenging task. One of the critical components of this system is Unitree, the Hierarchical File Storage Management System. Unitree, selected two years ago as the best available solution, requires constant system administrative support. It is not always suitable as an archive and distribution data center, and has moderate performance. The Data Archive and Distribution System (DADS) software developed to monitor, manage, and automate the ingestion, archive, and distribution functions turned out to be more challenging than anticipated. Having the software and tools is not sufficient to succeed. Human interaction within the system must be fully understood to improve efficiency and ensure that the right tools are developed. One of the lessons learned is that the operability, reliability, and performance aspects should be thoroughly addressed in the initial design. However, the GSFC DAAC has demonstrated that it is capable of distributing over 40 GB per day. A backup system to archive a second copy of all data ingested is under development. This backup system will be used not only for disaster recovery but will also replace the main archive when it is unavailable during maintenance or hardware replacement. The GSFC DAAC has put a strong emphasis on quality at all level of its organization. A Quality team has also been formed to identify quality issues and to propose improvements. The DAAC has conducted numerous tests to benchmark the performance of the system. These tests proved to be extremely useful in identifying bottlenecks and deficiencies in operational procedures.

Introduction

The GSFC DAAC is being developed in several phases with Version 0 (V0) being developed to support existing and pre-Earth Observing System (EOS) Earth science data sets, facilitate the scientific research, and test EOS Data and Information System (EOSDIS) concepts. This paper presents the GSFC DAAC V0 missions and requirements, and describes its architecture at the software and hardware level. The ingest, archive, and distribution processes are explained and a walk-through of these functions is presented. Numerous tests have also been conducted to benchmark the performance of storage devices, specific functions (e.g., ingestion), and the overall system. The tests which helped identified deficiencies in operational procedures and software are described. The Hierarchical File Storage Management System, Unitree, is a critical component of the DAAC. Some major issues were discovered during the integration of Unitree with the GSFC DAAC hardware and software, and a list of lessons learned has been compiled. There are some issues that were identified during the development, integration, and operational support of this system which are also discussed. Another topic presented in this paper is the focus and pursuit of quality by the GSFC DAAC.

GSFC DAAC V0 Mission

The initial version of NASA's EOSDIS is Version 0 (V0). This system consists of eight DAACs disseminated across the United States. Each DAAC is generally specialized in Scientific disciplines. The DAAC role is to enhance and improve scientific research and productivity by consolidating access and distribution of Earth science data. The evolutionary approach of building a Version 0 system is intended to demonstrate the concept of an interoperable set of distributed archive centers and to prototype various aspects of the system prior to the first EOS satellite launch.

The Goddard DAAC has defined its mission "to maximize the investment benefit of the Mission to Planet Earth by providing data and services to enable the realization of the potential of global climate data by the science and education communities".

GSFC DAAC V0 Requirements

The GSFC DAAC is being developed in response to EOSDIS functional requirements as well as requirements generated from Science projects such as Sea-viewing Wide Field-of-view Sensor (SeaWiFS), Coastal Zone Color Scanner (CZCS), Total Ozone Mapping Spectrometer (TOMS), Advanced Very High Resolution Radiometer (AVHRR), Tiros Operational Vertical Sounder (TOVS), and Upper Atmospheric Research Satellite (UARS).

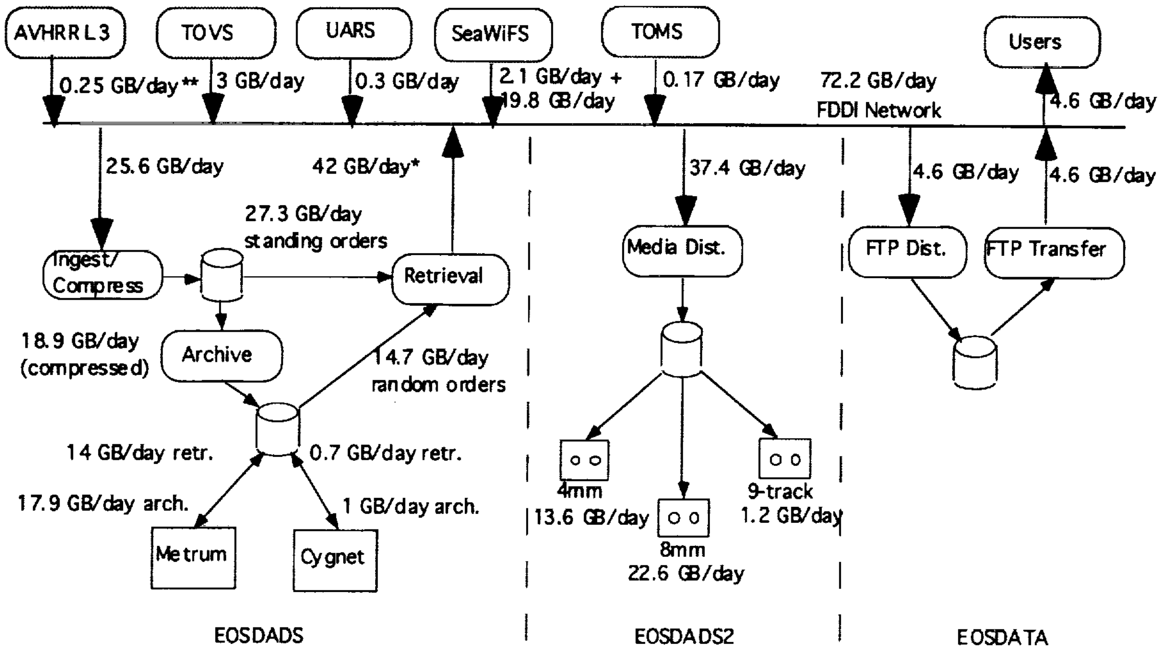
The GSFC DAAC has currently 731 GB of data archived (Table 1). This number is expected to increase to about 18 Terabytes by FY97 [1]. In 1995 the daily ingestion workload is estimated to be 26.4 GB/day (Table 2). All ingested data (except AVHRR) are compressed to reduce storage needs. This results in 18.9 GB/day of data being archived on the Metrum RSS-600 ATL (95%) and the Cygnet Jukebox (5%). The volume of data distributed is anticipated to be 40 GB/day of SeaWiFS data and 20 GB/day of non-SeaWiFS data, for a total of 60 GB/day. Two types of distribution orders have been identified: standing orders and random orders. The standing orders, by definition, are requests by users for some or all of the data as it is being received at the DAAC. The random orders are interactive requests by users for data that has been previously archived and is available for order. A significant proportion of orders are expected to be standing orders (65%) and most of the data ordered (89%) are assumed to be distributed on physical media (e.g., 8 mm) with the remaining being sent over the network (ftp orders). The distribution media supported currently at the GSFC DAAC are 8mm, 4mm, 9 track- 6250 bpi. The estimated V0 DAAC workload is illustrated in Figure 1.

Product	**Volume archived on Metrum (GB)**	**Volume archived on Cygnet (GB)**	**Total volume archived (GB)**
SeaWiFS L1 A (test)		1	1
SeaWiFS L2 (test)	1		1
AVHRR L3	111		111
UARS L3	35		35
TOMS	97		97
CZCS Level 1		345	345
4D assimilation	141		141
Total	385	346	731

Table 1 Total Volume of Data Archived as of 10-31-94

Product	**Volume before compression (GB)**	**compression ratio**	**Volume after compression (GB)**
SeaWiFS (regular)	2.10	0.72	1.51
SeaWiFS (reprocessing)	19.80	0.72	14.26
AVHRR	1.00	0.25	0.25
TOVS	3.00	0.80	2.40
UARS	0.30	1.00	0.30
TOMS	0.17	1.00	0.17
Total	26.37		18.89

Table 2 Estimated 1995 Daily Ingestion Workload



* Data compressed (equivalent to 60 GB if uncompressed)

** Data compressed by AVHRR Pathfinder PGS (equivalent to 1 GB if uncompressed)

Figure 1 Estimated DAAC Workload (Volume/day)

GSFC DAAC V0 Hardware Architecture

GSFC DAAC consists of three components, a Product Generation System (PGS), an Information Management System (IMS), and a Data Archive and Distribution System (DADS). The PGS receives low level data products (raw data requiring processing) and generates higher level data products. The IMS serves as a catalog of the data holdings which can be searched and browsed by researchers to help them identify and order data of interest. All data are archived within the DADS where they are available for on-line retrieval to fill researchers' orders for data.

A strategy was initially developed [1] to identify the best cost effective hardware and software configuration, and to measure the performance of the selected system [2]. Based upon the latest requirements, and projected workloads, the GSFC DAAC Fiscal year 1995 hardware configuration for the IMS and DADS is illustrated in Figure 2. The following are the points of the strategy.

- An SGI 4D/440 S (DADS) runs Unitree and the DADS software. To reduce the load, the DADS software is planned to be moved to a SGI Challenge L. The Unitree cache has 40 GB of disk space.
- Near-line data are archived on either a Cygnet 1803 jukebox (1179 MB) with 2 ATG WORM optical drives or an RSS-600 Metrum Automated Tape Library (ATL) (8700 MB) with 4 RSP 2150 VHS drives.
- A secondary archive is planned with a Challenge S (Backup) to keep a backup copy of all data ingested at the DAAC. The primary copy is archived by Unitree on an SGI 4D/440 S.
- The SGI Challenge L (DADS2) which has a larger number of I/O ports and fast internal bus, has all the distribution tape drives attached to it. The GSFC DAAC has nine 8 mm

drives, four 4 mm drives, and two 9 track drives. Additional drives may be added to satisfy future needs. To receive ingested data and copy data to tapes (e.g. 8mm) 40 GB and 72 GB respectively of disk space is available. Requests for FTP transfers are kept on-line on 40 GB of disks.

• An SGI 4D/440 VGX (DATA) computer runs the IMS software and Oracle. This machine has also the client which provides interoperability with other DAACs through a high-level Information Management System.

• The DAAC's distributed environment includes several ethernet Local Area Networks and an FDDI network connected to the EOSWAN.

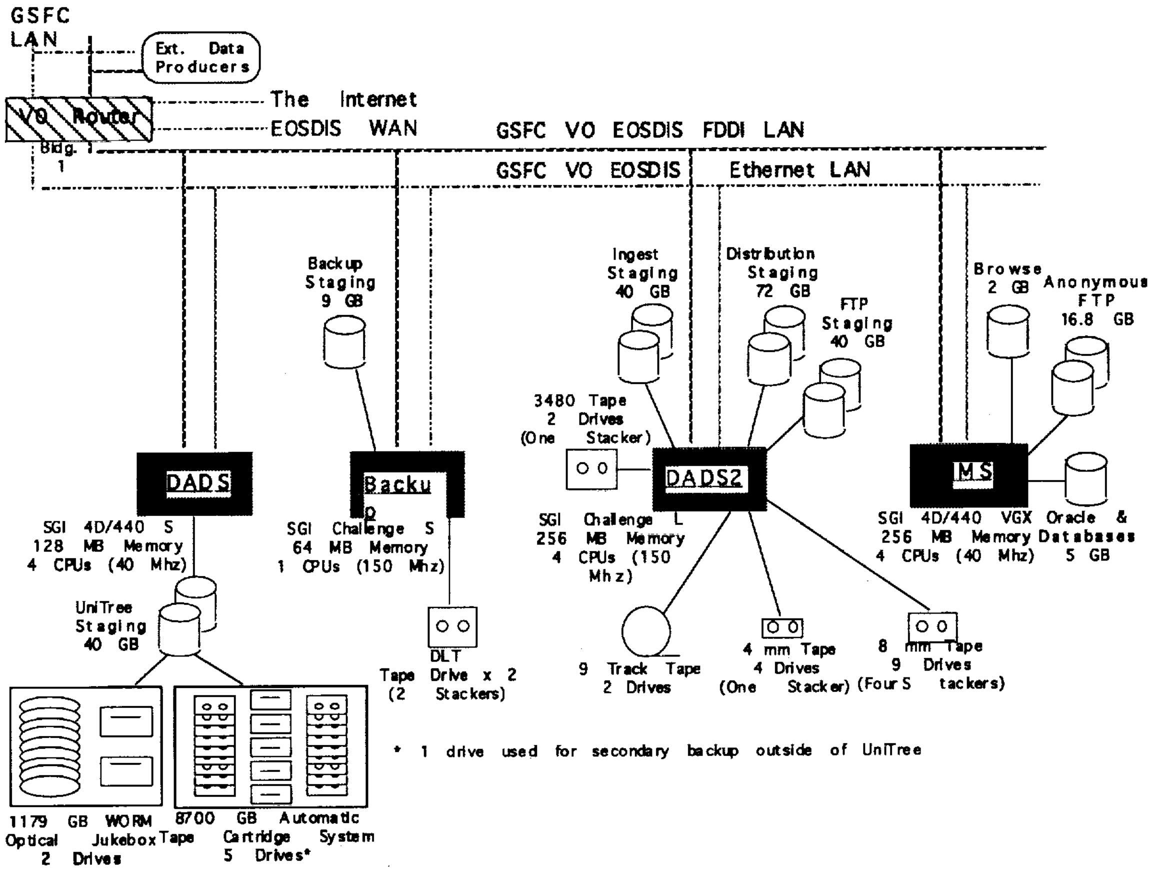

GSFC LAN
Ext. Data Producers
VO Router
Bldg. 1
The Internet
EOSDIS WAN
GSFC VO EOSDIS FDDI LAN
GSFC VO EOSDIS Ethernet LAN
Backup Staging 9 GB
Ingest Staging 40 GB
Distribution Staging 72 GB
FTP Staging 40 GB
Browse 2 GB
Anonymous FTP 16.8 GB
3480 Tape 2 Drives (One Stacker)
DADS
Backu p
DADS2
IMS
SGI 4D/440 S 128 MB Memory 4 CPUs (40 Mhz)
SGI Challenge S 64 MB Memory 1 CPUs (150 Mhz)
SGI Challenge L 256 MB Memory 4 CPUs (150 Mhz)
SGI 4D/440 VGX 256 MB Memory 4 CPUs (40 Mhz)
Oracle & Databases 5 GB
UniTree Staging 40 GB
DLT Tape Drive x 2 (2 Stackers)
9 Track Tape 2 Drives
4 mm Tape 4 Drives (One Stacker)
8 mm Tape 9 Drives (Four Stackers)
* 1 drive used for secondary backup outside of UniTree
1179 GB WORM Optical Jukebox 2 Drives
8700 GB Automatic Tape Cartridge System 5 Drives*

Figure 2 GSFC DAAC FY 95 Configuration

GSFC DADS Functional Design

This paper will now focus on the DADS and the mass storage issues. The GSFC DADS has three main functions: Ingest & Archive, Distribution, and Management. The ingest & archive function consists of accepting data products from outside the system, extracting or creating metadata, validating files, storing the files in the primary and backup archives, and updating the database. The distribution function retrieves files from archives, stages them to a distribution staging area, reformats the data if necessary (e.g. tar is the normal format for orders), and then writes the data to tapes or to an FTP staging disk. The DADS management software handles the scheduling, tracks DADS activities, and controls allocation and deallocation of resources. The DADS functional design is illustrated in Figure 3.

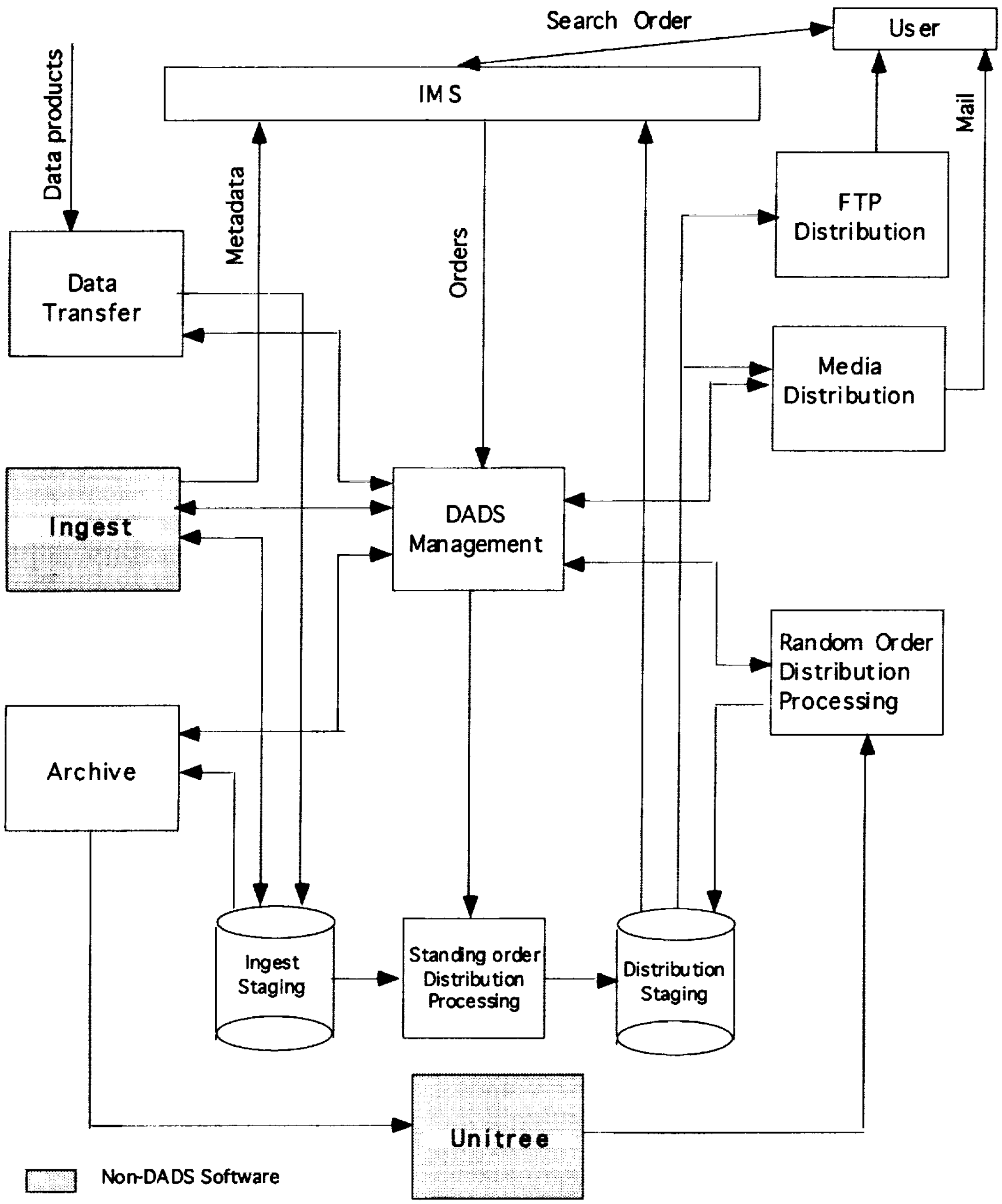


Figure 3 DADS Functional Design

GSFC DADS Ingestion, Archive, and Distribution Functions

The GSFC DADS currently ingests through network interfaces or directly from media datasets produced by the following scientific projects: AVHRR, TOVS, TOMS, 4 D Assimilation, CZCS, and UARS. The SeaWiFs project will be added to that list after the launch of its satellite scheduled in Spring/Summer of 1995. Ingestion of data over the network is usually triggered when a scientific project invokes a client hosted on their computer called Data Transfer Program (DTP) (Figure 4).

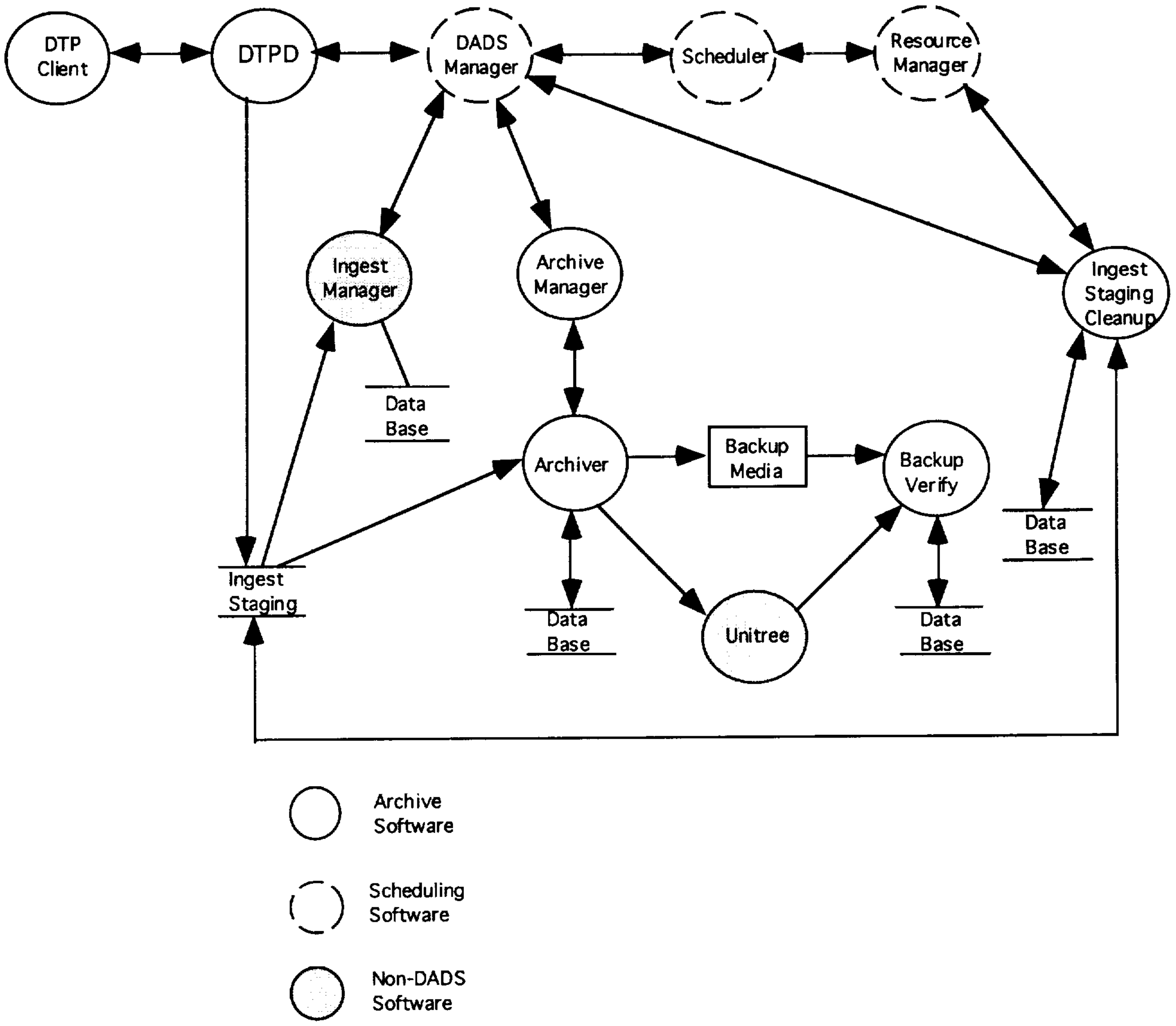


Figure 4 Archive Architecture

The transfer of data begins after DTP receives authorization from the DADS, which ensures the availability of resources to satisfy the ingest. The migration operations between the near-line devices (Cygnet jukebox and the Metrum ATL) are handled by the Hierarchical File Storage Management (HFSM) Unitree. The processing schedule and the resource allocation/deallocation are performed by the DADS modules: DADS manager, scheduler, and resource manager. A second archive copy is generated and handled by the archive manager, archiver, and backup verify. The ingestion and archive processes are described in detail in Table 3. In addition, Table 4 summarizes an Ingest/Archive walk-through.

Process	Description
DTP	Requests Ingest staging disk space from DADS Manager and Transfers files from the client system to the ingest staging area
DADS Manager	Sequences transfer, ingest, archive, verify, and staging cleanup
Scheduler	Interacts with the resource manager to allocate disk space, and Starts activities when resources are available
Resource Manager	Manages disk space in the ingest and distribution staging areas
Ingest Manager	Starts the correct processing script for each transferred file Script validates file, extracts metadata, and loads granule level database tables
Archive Manager	Batches archive requests Initiates archiving activities on a size or time basis
Archiver	Performs primary and backup archiving activities Computes and stores checksum values Exposes granules
Ingest Staging Cleanup	Checks successfully archived files against standing orders Copies files required by standing orders to distribution staging and adds items to open standing orders Removes successfully archived files from ingest staging area
Backup Verify	Run as chron job Retrieves backup archives files and recomputes checksum Compares checksum to value computed by archiver Sends E-mail to data producer on success

Table 3 DADS Ingest/Archive Processes

Step	Description
Transfer	1. DTP client and server establish connection 2. DTPD sends a request for disk space to Scheduler via DADS Manager 3. Scheduler, using Resource Manager, determines when to initiate the transfer and sends message to DTPD via DADS Manager to start transfer. 4. DTP Client and Server perform transfer 5. DTPD sends file completion message to the DADS Manager as each file completes transfer 6. DTPD sends termination message to the DADS Manager after all files are transferred
Ingest	1. DADS Manager sends ingest request to Ingest Manager for each transferred file 2. Ingest Manager starts appropriate processing script for each file 3. Ingest script extracts metadata, validates data, updates Data Base granule table, and sometimes does compression 4. Ingest Manager sends ingest complete message to DADS Manager for each file
Archive	1. DADS Manager sends archive request to Archive Manager for each transferred file. 2. Archive Manager adds file pending archive list 3. When archive list reaches a size threshold, the Archive Manager sends a batch archive message to the Scheduler via the DADS manager 4. The scheduler determines when to initiate the archiving activity and sends a message to the Archive Manager via the DADS Manager to start the Archiver 5. The Archiver copies the files to Unitree and to a backup tape, then sends an archive complete message to the Scheduler via the Archive Manager and the DADS.
Ingest Staging Cleanup	1. The DADS Manager starts the Staging Cleanup process 2. Ingest Staging Cleanup determines which files need to be staged for standing order distribution 3. Ingest Staging Cleanup requests disk space from the Resource Manager. 4. If the distribution space is available, Ingest Staging Cleanup copies the files and notifies the Resource Manager of the space used. 5. Ingest Staging Cleanup then adds the requested items to the standing order request. 6. Ingest Staging Cleanup removes the successfully archived files from the ingest staging area, notifying the Resource Manager of space made available.
Backup Verification	Runs as a chron job periodically Retrieves backup archive files and verifies using checksum Sends E-mail Notification to data producer that archive was successful

Table 4 DADS Ingest/Archive Steps

Another major function of the DADS software is the distribution of archived data to users. New order requests are generated by the user using the IMS and are then automatically submitted by the IMS to the DADS. Requests that are initially delayed are obtained later by the DADS by scanning the database using a program called pollreq (see Figure 5). Any known request can also be submitted manually for processing using ureproc. The staging operations between the near-line devices (Cygnet jukebox and the Metrum ATL) are handled by the HFSM Unitree. The processing schedule and the resource allocation/deallocation are performed by the Scheduler, Resource Manager, and Tape Manager. The DADS modules developed for the distribution function are summarized in Table 5. To clarify the distribution process, a walk-through is described in Table 6.

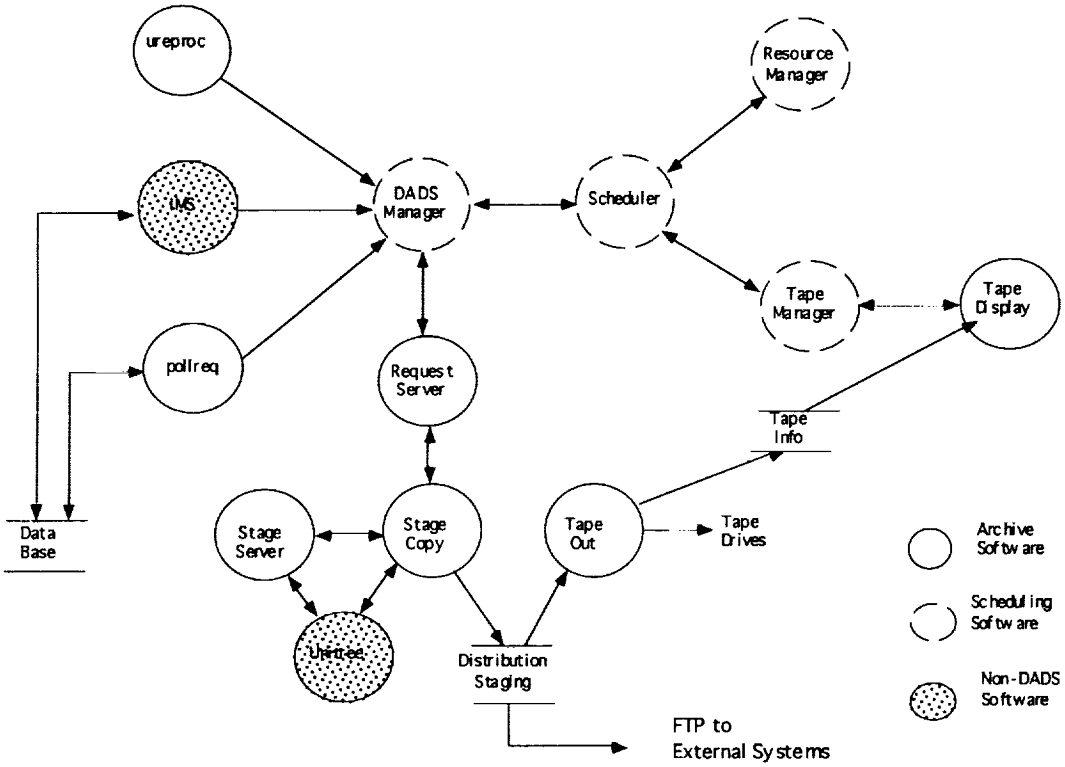


Figure 5 DADS Distribution Architecture

Process	Description
Request Poller (pollreq)	Scans data base for requests that have not been initiated Sends request ID to dadsmgr for each request found
DADS Manager (dadsmgr)	Sequences archive and distribution activities
Scheduler (schedsrvr)	Maintains queues of processing activities Interacts with resource & tape managers to allocate resources Starts activities when resources are available
Resource Manager (rsmansrvr)	Manages disk space in ingest and distribution staging areas
Tape Manager	Controls allocation and deallocation of tapes Controls automated (not manual) tape mounts for distribution
Tape Display	Show status of all tape drives Prompts operators to mount/dismount tapes
Request Sever (reqserver)	Locates all items in request and requests disk space Starts Stage Copy when disk resources are available Requests tapes required for request Starts Tape Out process when tapes are available
Stage Server	"Batches" Unitree staging requests
Stage Copy	Ask Unitree to stage files, and copies staged files to distribution staging area
Tape Out	Writes header and staged files to distribution tape

Table 5 DADS Distribution Processes

Step	Description
Staging	1. IMS or GenAutoOrder generates request in database 2. IMS or pollreq sends messages to DADS Manager to start processing request 3. DADS Manager sends request to Request Server 4. Request Server requests disk space from Resource Manager via DADS manager and Scheduler 5. Scheduler, using Resource Manager, determines when to process request 6. Scheduler sends message to Request Server via DADS Manager to start request processing 7. Request Server stages all files not already staged and creates symbolic links for all files 8. If no output tapes are required then Request Server signal completion of request
Tape Output	1. Request Server sends a message to Tape Manager via DADS Manager and Scheduler for tape drive 2. Scheduler determines when to write output tape, using Tape Manager (and Tape Display) to mount tape 3. Scheduler sends messages to Request Server via DADS Manager to write tape 4. Request Server creates child process to write tape header and files. Request Server signal completion of tape to Tape Manager via DADS Manager and Scheduler Tape Manager and Tape Display handle dismount of tape and bar-code label generation

Table 6 DADS Distribution Steps

Ingestion and Archive Functions

Files are ingested at the DAAC using DTP which incorporates a modified version of ftp. The regular ftp is not suited for background tasks and does not return error codes. The DAAC had to develop their own ftp that can be executed via a call routine and that returned error codes. The overhead associated with opening a connection and getting a response back via the DADS Manager turned out to be long (30 s). With small files (< 5 MB), the transfer time is much smaller than the opening connection time. It is therefore necessary to transfer a large number of small files with a single connection in order minimize overhead.

The DADS manager is a central point by which each message is received and sent. This design adds overhead and with a heavy load, this might become a bottleneck. Another alternative architecture would be to send messages directly to the recipient without passing through the DADS manager.

Scheduling the DADS activities efficiently is a difficult problem. The scheduler must dynamically schedule all the DAAC activities based on resource utilization and task priorities and some general policies. A resource can represent, for example, disk space, tape drives, or the number of concurrent ftp sessions. The scheduler must also prevent deadlock situations which would halt the system. In the first phase, the DAAC has developed its scheduler using a very simple scheme First In First Out (FIFO). This approach works fine when the resources are abundant. However, when there are contentions for resources, the schedule using a FIFO algorithm becomes extremely inefficient and slow. The granularity of the task is very important. Treating each process as a task is not a good solution because of the large number of processes involved. On the other hand, a task such as a distribution function has several sub-tasks that are to be scheduled separately while maintaining the order in which each subtask should be submitted. For example, a distribution function is composed of at least of a stage operation, a copy to the distribution area, and a copy to tape. It would be inefficient to allocate all resources needed at the beginning of the task. For instance a distribution tape drive should not be allocated until the data is staged to the distribution staging area. By dividing a task in a series of sub-tasks and by scheduling each individual subtask, the system resources can be better used and the overall performance can be improved. Each sub-task must allocate its own resources and the predecessors and successors of each sub-task must also be preserved. A general-purpose constraint-based scheduling engine based on the Time Map Manager (TMM) that uses a multi-level of tasks/subtasks is being studied for integration in the DADS software.

The DADS software was based on a client/server configuration. In the current architecture, each main function is a server that can be distributed over several platforms. The implementation of a client/server configuration turned out to be more complicated than expected. It is critical in this kind of environment to capture all errors and provide a mechanism to recover from these errors. It is also imperative to ensure that no single message is lost and that the communication protocol is very reliable. In the early stage of the development of the DADS software, messages were lost and processes were hanging. This could lock valuable resources indefinitely. One of the key problems with a client/server configuration is that when a server crashes, it takes many jobs along with it. A one process/one job philosophy would be better. Testing client/server software can also be a very difficult task because it is not always easy to reproduce errors that had occurred previously. With a client/server architecture, it is also important to limit the traffic of messages in order to achieve a good performance of the system.

Backup system

All V0 data are archived on several copies. The primary copy is on near-line storage (WORM platters or VHS tapes) using the HFSM Unitree. This implies that the data are stored with the Unitree Proprietary format. Relying on a single copy is prone for disaster sooner or later. During the first year of being operational the GSFC DAAC experienced unrecoverable errors on VHS tapes on six occasions, even though the life expectancy of the media was 10 years. Most of the problems were linked with a bad tape drive. In conjunction, the firmware of the Metrum drives used at that time did not limit the number of retries in search mode, and the media was damaged by an excessive number of passes. Unitree does not currently provide a mechanism to detect the number of soft errors or even the number of times a given tape is mounted/dismounted. With large archives it is imperative to detect such soft errors in order to predict when it is time to make another copy before the media is permanently damaged. The cost of creating a duplicate copy of a tape that has unrecoverable I/O errors can be a very expensive and time consuming task. Some data sets are in high demand and are used extensively. For instance one tape was mounted more than 2000 times in one year. With each mount, there are several passes and this exceeded the maximum number of passes (3000-6000 for the VHS tapes) provided by the manufacturers. Whenever possible, it is recommended to keep these highly requested datasets on magnetic disks or optical media, not only to minimize the response time but also to prevent such media degradation. It is not always easy to predict which datasets are going to be in high demand and the use of media such as VHS tapes must be closely monitored for high usage of individual tapes and a procedure put in place to copy these tapes to new tapes as needed.

Currently, the second copy of the data in the archive is done using the standard tar format on a VHS tape. This should facilitate the migration of the data to the EOS V1 system. A new backup system is under development. The plan is to copy all data by families (data set and level) on a VHS tape and on a DLT tape. The DLT media seems promising. It has a higher level of passes, stores a large volume of data and is relatively inexpensive. However DLT is a new media and because of its low cost, the project decided to make backup copies on both VHS and DLT until more is known about DLT drive and media reliability. On several occasions Unitree was unavailable for several days and the operations came to halt. The GSFC DAAC workload is going to increase several times with the SeaWiFs data sets, and another occurrence of Unitree unavailability for a long time would create difficulty in recovering from such long outage. To alleviate this problem, the DAAC has a contingency plan to use the backup system as an ingestion and distribution system. The backup is on a different machine, has its own drives and robotics, and is being designed to handle such eventuality.

Distribution Function

After conducting tests with a heavy workload, it became clear that the number of new distribution requests to process concurrently had to be limited (around 10). Several factors contributed to this condition. First, with a large number of files to stage, each stage command uses 3 processes, the maximum number of processes (500) available on the DADS could be exceeded in some cases. Secondly, the data had to be staged to a distribution staging area and too many concurrent nfs copies to disks resulted in severe degradation of the nfs throughput which is notoriously slow to begin with. Some factors contributing to the nfs poor performance were due to a maximum of eight group ids that can be sent and an nfs feature that locks directories until the files are opened. Replacing nfs by FTP should improve the throughput by 2 or 3 times. As with nfs, the number of concurrent FTPs must be limited in order to achieve a good performance and scheduling becomes important.

Whenever a file is requested for distribution an Oracle database is searched to determine if it resides on the distribution staging area and to identify its physical location on the staging area. The access to this database was causing substantial delays (minutes) and the SQL code had to be optimized in the DADS software to achieve better performance. During the latest tests, the SGI 4D/440 VGX computer hosting the database was CPU bound and the DAAC is investigating the prospect of acquiring a more powerful machine as well as more optimal ways of accessing the databases.

The Stage server role is to group files belonging to the same family so that they can be submitted to Unitree as a single batch. This improves the overall performance of the system by minimizing the number of mounts/dismounts. The files selected that reside on the same tape are read with a single mount. Unitree philosophy is to have full data transparency and the users should not be aware of the physical location of the files. This concept may be fine with users but is completely inappropriate for system administrators, developers, and testers. If the physical location were known the stage server could group requests with files residing on the same media and schedule the stage from various orders to optimize the retrieval throughput.

An important parameter in designing the architecture of the system is the volume of data to be ingested and distributed. However it is also necessary to have good estimate on the size of the files. A system with many small files has more overhead than a system of the same size composed of larger files. With small files, more time can be spent searching the files on tapes than actually reading data from tapes. The size of the orders must also be well estimated in advance. Files belonging to the same orders are usually staged to distribution staging area prior to being copied to media or made available for ftp transfer. If the size of the orders are underestimated the distribution staging area may be too small creating delay and confusion at the operation level.

Orders are placed to the GSFC DAAC via the IMS. Data can be requested to be available over the network (ftp request) or distributed on media such as 4 mm, 8 mm, or 9 track (media request). With an ftp request, the data are automatically staged to disks to be copied immediately and the user is notified by E-mail. The User has 3 days to transfer the file(s) over their computer. As the number of requests increases the space needed to stage ftp may become so large that the 3 days policy may be cut to just a few hours and may not be long enough for the users. Sending data to users has other problems such as security, privileges and availability of user disk space.

Operation

One important role of the GSFC DAAC is the dissemination of the data requested to the scientific community. With respect to SeaWiFS only, 40 GB are expected to be distributed each day. To process this volume of data most functions have been automated by the DADS software. However, in this environment it is not unusual for something unexpected to occur (e.g. bad tape) and the operators must identify, and rectify these problems manually. This can be time consuming and one lesson learned was that operators needed more tools to be more productive. These tools are also used to monitor the system, its resources and the requests. The tools must be defined by the operators and developed by programmers. There is a tendency for developers to design software without fully understanding the need or operation concept. This can result in a product that is too complicated to use, too cumbersome, or does not meet the needs. Tools were part of the preliminary designed but the scope of the task may have been underestimated. Some of these tools are also difficult to identify until you have a real system in place. Without these tools the overall productivity can be greatly reduced.

Another major challenge in building a system such as the GSFC V0 DAAC, is to design it from the beginning with operability, condition monitoring, error recovery, and performance. These aspects are often neglected as a project starts with some type of prototyping where the emphasis is on functionality.

Creating the data requested by the users is not the only task. Tapes must be labeled, tape contents verified, documentation must accompany the order, and everything has to be boxed and mailed. All these steps can be manual intensive, time consuming, and must be streamlined in order to be as efficient as possible. Without the right procedures and tools, operators can spend a lot of time performing these tasks. This could result in a degradation in quality as less time is spent monitoring the system for unusual events. To minimize the risk of inadvertently switching tapes for different orders, all tapes are labeled with bar codes and scanned by bar code readers. Mailing labels are printed with identical bar codes to insure that the correct tape is sent to the researcher.

Not all the requests are entered electronically via the IMS. Some users still need to order datasets over the phone or need assistance. To support the users, the DAAC has a User Support Office (USO). The interaction between USO and the operation group is important. Lack of communication between these two groups or any other groups within the DAAC would results in deterioration of the service provided to the Scientific community. In addition information that are often needed by the researcher (e.g. status of order) should be available on-line to minimize the workload of the USO staff.

The GSFC DAAC is a service oriented organization and as such has the responsibility to provide the best product to users. To help to achieve this goal, a quality team has been created at the DAAC. Its primary role is to identify quality issues and to suggest solutions. A strong emphasis has been placed on quality issues that mostly impact external users. This group was established after discovering that blank/bad tapes had been sent to users. One of the first tasks of the quality team was to review complaints within the DAAC and by our customers. Then, starting from the operation level, the DAAC processes have been reevaluated to identify deficiencies and propose solutions. For example, to preclude GSFC DAAC from sending bad/blank tapes, a directory of the tape is compiled. This solution is time consuming because it takes the same amount of time to create the tape as to read it and generate a directory. Other alternatives are to read only the first records or get a directory of tapes randomly selected. Capturing I/O errors during the creation of the tape is another way of insuring the quality of the tapes. 8mm and 4 mm have a read/verify operations after a write operation that could guarantee the data is stored properly on the media. The problem is that the I/O errors are reported at the bus level only and when several drives are connected to the same bus it is not always possible to determine which drive had an I/O error. 8 mm stackers have also been purchased to minimize human intervention and reduce the risk of errors. As simple as these functions may be, examining the processes in details has revealed that their implementation is usually too complex, inefficient and filled with unnecessary manual steps that slows down the performance.

Testing

The GSFC DAAC has conducted numerous tests on the V0 System to measure the throughput of its peripherals running separately or concurrently. Basic functions such as ingest, stage, ftp have also been benchmarked in order to estimate the overall performance of the DAAC and to identify bottlenecks and limiting factors. These measurements have been summarized in Figure 6. The numbers listed in Figure 6 represent the best values obtained on a system that was not busy. The distribution tape drives (4mm, 8mm, and 9 tracks) transfer rates varied with the size of the files copied. Writing a large number of files on tapes with the tar format was found to be faster than copying the same data on the same drive using dd command. Currently, the only mechanism to transfer data in and out of the Unitree cache is via nfs or ftp. The best throughput of a single file transferred was measured at 1570 KB/s with ftp and 430 KB/s with nfs in local host. The ftp and nfs throughput is a function of the number of concurrent transfers as illustrated in Figure 7 and Figure 8 . Having too many ftps or nfs running at the same time can reduce considerably the overall throughput. If the files reside on the same disk, there may also be some disk contention. Compression and decompression are CPU intensive operations that may create a bottleneck. As expected these operations are executed faster on the SGI Challenge L than on the SGI power series (see Figure. 9). Several compressions or decompressions running simultaneously will contend for the CPUs and potentially the disk I/Os resulting in degradation in the overall individual compression/decompression transfer rates. A hardware solution for compression/decompression would alleviate this problem. The GSFC DAAC has investigated for such a hardware board, but in vain. The stage operations have been tested using the RSS-600 Metrum Automated Tape Library (ATL). It is difficult to measure the throughput of these operations because they depend on the size of the files retrieved and the position of the files on the tapes. Using a large file (270 MB) positioned at the beginning, in the middle, and at the end of the tape it was found that the overall effective transfer rates that include all the overheads (pickup time, load time, time for Unitree to read header, search time and read time) was respectively 545 KB/s, 604 KB/s, and 612 KB/s. These rates are roughly one third the native rates of the Metrum drives. These tests were for a large file and reflect best case scenarios. The latest tests conducted during several hours with 3 Metrum drives show that with 30-200 MB files the transfer rate was around 170 KB/s per drive. Even with multiple drives (5), this can become a bottleneck and it is important to schedule these stage operations in order to minimize the number of mounts/dismounts and therefore maximize the overall throughput.

In addition to these individual tests, GSFC DAAC has conducted "mini-tests" each time a new version of the DAAC was released. The initial objective of these mini-tests was to demonstrate that the center could process 40 GB/day of SeaWiFS data. After conducting the first mini-test it became apparent that the goals of these mini-tests should be expanded. For instance, software bugs which could occur only when the system was under a heavy workload, were discovered. The mini-test was in itself an extension to thorough testing performed by an independent test team. These mini-tests also contributed to identify deficiencies in operation procedures. This resulted in increase productivity and improved the overall quality of the data ingested and distributed. The problem associated with these tests is that the operations are delayed while they are conducted. However the benefits outweigh the drawbacks.

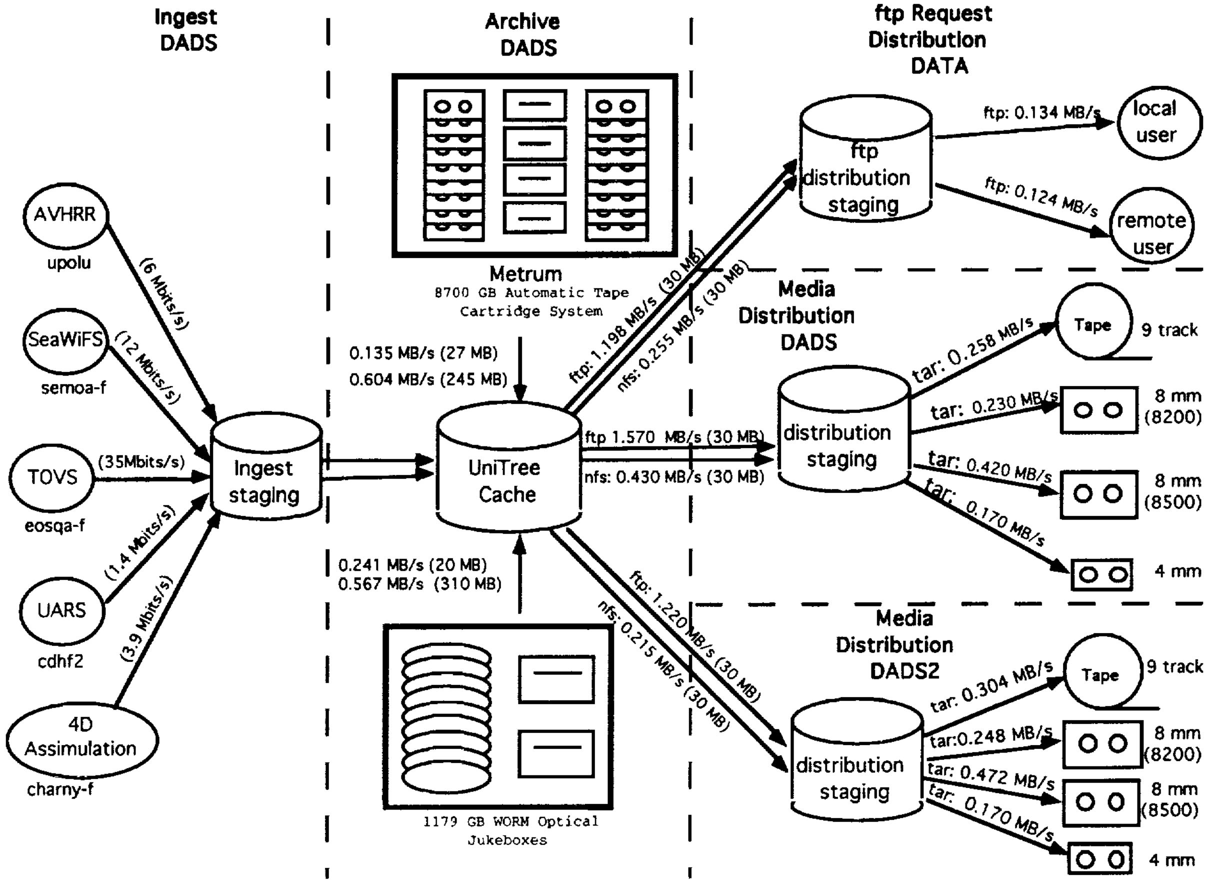


Figure 6 GSFC DAAC V0 Testing

FTP Performance on DADS machine
(Copy from Unitree Cache to Unix disks)
(3-15-94 test)

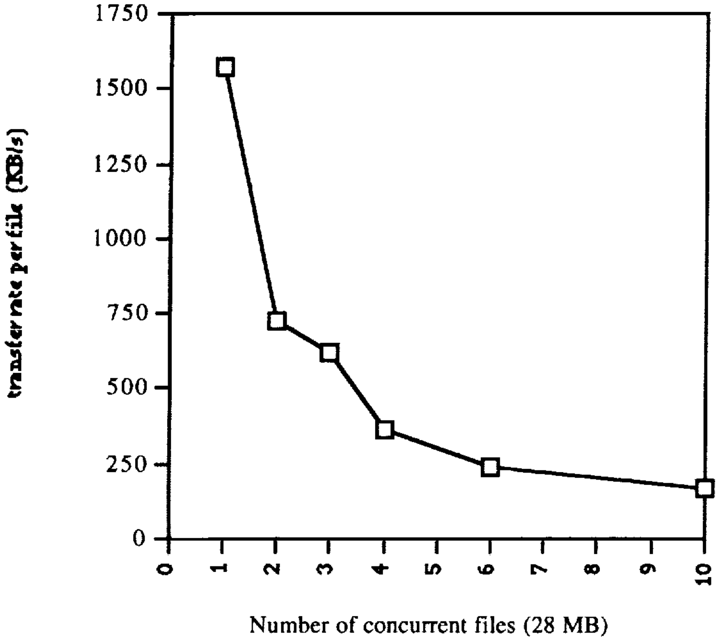


Figure 7 FTP Performance

NFS Performance on DADS machine
(Copy from Unitree Cache to Unix disks)
(3-10-94 test)

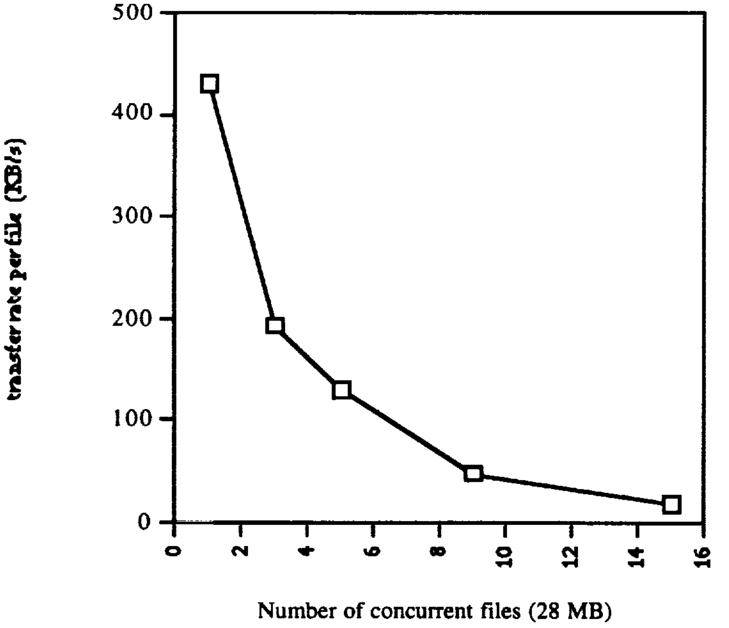


Figure 8 NFS Performance

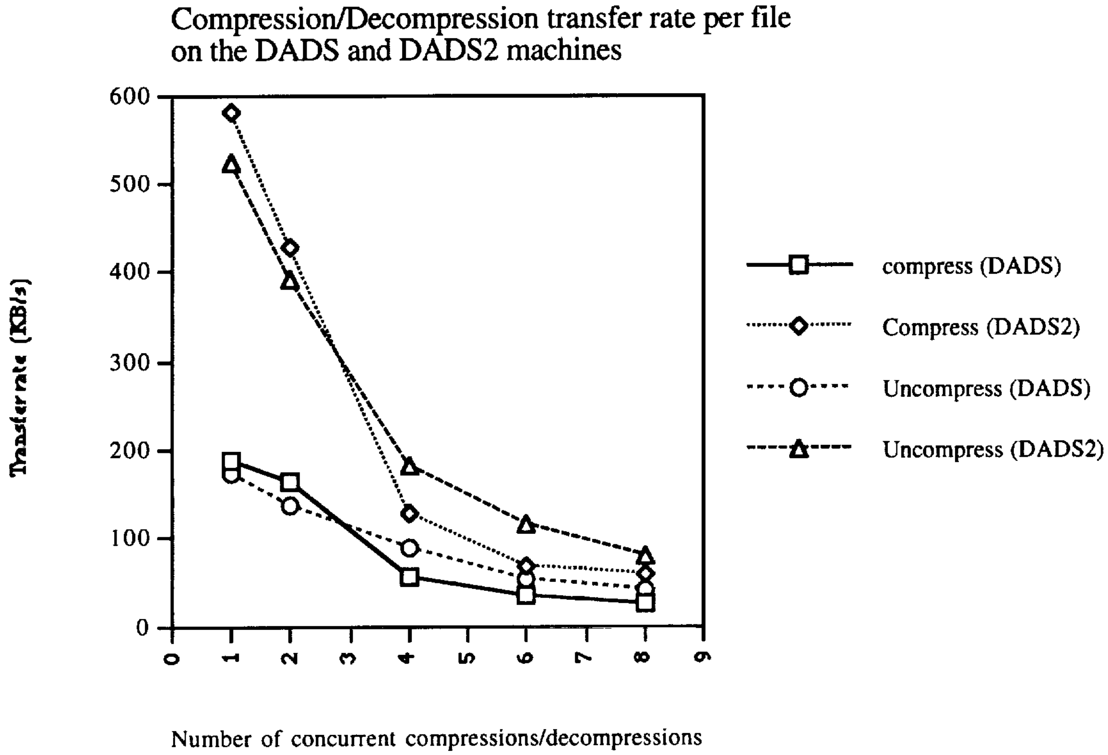


Figure 9 Compression and decompression performance

A 16 hour test was conducted on Tuesday, December 13, 1994. The primary objective of this test was to demonstrate that the GSFC DAAC could distribute 40 GB of SeaWiFS orders each day. No ingestion was processed during this test. The total number of orders and total volume of orders processed exceeded the target goal for both standing orders and random orders. During the test, the DADS software proved to be very robust. All the SeaWiFs test orders were completed more than 3 hours before the end of the test. During the test, all data copied from the Unitree cache to the distribution staging area, were transferred at the speed of the nfs because the disks were nfs mounted This is currently the main bottleneck in the system. However, preliminary tests have shown that by using ftp, the transfer rates between the Unitree cache and the distribution staging area should be 2 or 3 times higher

Hardware

GSFC DAAC bought hardware peripherals (disk & tape drives) at a discount price from third party vendors. The initial saving was not always a good investment as the DAAC system staff had to work very hard to integrate the peripherals. This distracted the system engineering from other urgent tasks, increasing system downtime, and generally caused grief to developers and operators. However, because staff time and system downtime do not get accounted directly we were able to procure significantly more disk capacity than otherwise. Another risk associated with purchasing peripherals from small third party vendors is that they are more prone to go out of business and with them go the warranty.

The GSFC DAAC experienced serious network throughput with its Science producers. After some investigation, it was discovered that older routers and bridges could not handle the load of Ethernet and FDDI, and had to be replaced.

Unitree

To automate the migration and staging operations between the robotic devices and the magnetic disks, the GSFC DAAC is using Unitree. At the time of the selection process, Unitree was the only product that fulfilled some of the requirements of the version 0 GSFC DAAC. The initial design of the DADS was to read files directly from Unitree cache and to copy them to the distribution media selected by users. On several occasions, files that were needed for distribution were purged from the disk cache by Unitree before they could be copied to tape. Another problem associated with Unitree is its poor performance in getting data in and out of its cache. The GSFC DAAC had to resort to developing and managing a second cache (i.e., various disk staging areas) to avoid the problems listed above. The duplicate cache increased the complexity of the DADS software and is expensive in terms of additional disk space needed. Having an Application Program Interface (API) would have been very useful in the development of the DADS software. Titan/Avalon recently delivered an API for Unitree but it was too late for the project to incorporate it and be ready for the SeaWiFs launch

One of the main drawbacks of Unitree is its lack of robustness. The GSFC DAAC has one person dedicated to monitoring Unitree at all times. This is not unusual as we discovered by talking to other Data Centers. This is a serious problem during the weekends as ingestion and distribution were disrupted because of problems related to Unitree and there is no one to monitor it. Unitree has come a long way, and its new version is more thoroughly tested and provides added functionality. However, it has not yet reached the maturity where it can run unattended, and it is still very expensive.

There are other issues that have been reported to the last Unitree users' group meeting held at GSFC on November 9-10, 1994. Most of them are related to inadequate documentation, cryptic error messages, lack of monitoring and administration tools, and no mechanism to capture soft errors detected by the drives during a read or write operation. This latter function is important as an increasing number of soft errors is an indication that the media might be degrading and that a new copy of a tape should be made. Because this function is not available, GSFC DAAC is currently monitoring the number of mounts/dismounts for each tape and copying tapes after a set number of mounts.

The overall performance of Unitree has been measured during the numerous tests that the DAAC conducted. In particular the stage operations were identified as a major bottleneck. The Metrum drives were benchmarked to read at 1.6 MB/s from UNIX. The same tests running with Unitree show a degradation of an individual Metrum drive to 1 MB/s in the best case scenario. When an ATG drive from the Cygnet jukebox was doing I/O at the same time as a Metrum drive the transfer rate of this later drive was reduced by at least half. All these tests were conducted with a system that had no activity.

Conclusion

The V0 System of the GSFC DAAC has gained valuable experience from building a few terabytes archive and distribution system and has demonstrated that it is capable of distributing 40 GB of data per day. Unitree needs to be more robust and easier to manage. The DADS software has turned out to be a real challenge. The difficulty being primarily in developing a reliable product that is fully automated with a good error recovery and with good performance. The operability, reliability, and performance aspects should all be major considerations in designing such a system. Special attention should be paid when buying hardware from third party vendor. It is usually cheaper, but the integration may be difficult and time consuming. Selecting the right media is very critical because of the high cost to migrate to another media. With larger and larger archives it is imperative to monitor media degradation and make new copies before unrecoverable I/O errors.

1. L. Bodden, P. Pease, JJ. Bedet, W. Rosen: Goddard Space Flight Center Version 0 Distributed Active Archive Center. In Third Conference on Mass Storage Systems and Technologies. NASA CP-3262, 1993, pp. 447-453.

2. JJ. Bedet, L. Bodden, A. Dwyer, PC. Hariharan, J. Berbert, B. Kobler, P. Pease: Simulation of a Data Archival and Distributed System at GSFC. In Third Conference on Mass Storage Systems and Technologies. NASA CP-3262, 1993, pp. 257-277.