The ESA Earth Observation Payload Data Long Term Storage Activities

Gian Maria Pinna [(1)], Francesco Ferrante [(2)]

*[(1)] **ESA-ESRIN***
Via G. Galilei, CP 64,00044 Frascati, Italy
EMail: GianMaria.Pinna@esa.int

*[(2)] **SERCO Italy***
Via Sciadonna24/26, 00044, Frascati, Italy
EMail: fferrante@serco.it

ABSTRACT

The ESA EO payload data archive is today exploited with a high degree of automation, provided by the MultiMission Facility Infrastructure (MMFI) concept, based on the OAIS (Open Archival Information System) Reference Model, implemented in the course of the last 4 years.

This archive today exceeds the 3PB size, is distributed across different archiving locations, also providing data backup capability, and increases steadily. The future missions with anticipated downlink data rates of more than 400Mbps, to be hosted in this archiving infrastructure, will further increase its size to more than 10PB.

In order to ensure the proper preservation of its assets, ESA has endorsed a Long Term Data Preservation Policy that provides the guidelines for its activities aimed at ensuring the proper preservation of its assets.

The MMFI being a key element for the implementation of this strategy, and with its ultimate goal of preserving the EO payload data, processing them to the desired level and distributing them to the end-users, it follows an evolution plan that ensures that the preservation process is maintained alive.

This paper presents the latest upgrades in the MMFI in the OAIS Storage functional area, showing how various technologies were evaluated and selected. Among the others:

1. Hierarchical Storage Management systems
2. Tape archiving technologies
3. Disk-based archives
4. Distributed File Systems
5. Technologies to exchange archived data among the ESA's EO distributed archive (e.g. peer-to-peer)

INTRODUCTION

The European Space Agency today manages the payload data operation of a number of Earth Observation satellites since 1975. The activity includes acquisition, archive, processing and products distribution of data from ESA and Third Parties missions, for which about 3 PetaByte of data is presently archived. The activity is performed via a network of facilities distributed in Europe and in Canada (for ERS only) mostly belonging to national and private entities, operating on behalf of the Agency via contractual agreements. The management centre of this network of facilities is located in the ESRIN ESA centre of Frascati, near Rome – Italy.

This huge archive needs to be managed and data preserved for the long term. Although the typical mandate for the missions operated by ESA are to keep the data for 10 years after the end of the missions, increased awareness of their importance for the future generations led ESA the propose a coordinated approach to the EO payload data long term preservation, which demands for much longer retention periods. This approach is presently ruled by the ESA Long term Data Preservation Strategy (ref. specific paper from Beruti et al. presented at this conference).

This paper presents the main technological advances and actions taken by ESA to implement the above-mentioned strategy and ensure a proper long term preservation of its assets.

PDGS ARCHIVES HARMONIZATION

One of the first steps deemed to be required to ensure a proper and cost effective preservation is the harmonization of the design, technologies and operational procedures used in the Payload Data Ground Segment (PDGS) developed or adapted for each mission operated.

The very nature of the ESA EO archive, heterogeneous and geographically distributed, imposes this a basic strategy for its implementation. In the course of the last 5 years a major effort was undertaken to migrate all archives and in general the PDGSs into an architecture based on a common infrastructure named MMFI (MultiMission Facility Infrastructure). This architecture and paradigm for the implementation of a PDGS (implementation by re-use and configuration) has proved to be very well suited for the harmonization and cost-effectiveness of the long term data preservation.

The MMFI was developed following the OAIS (Open Archival Information System) Reference Model, a CCSDS/ISO standard, well suited for this purpose (see ref. [1] and [2]).

The Figure 1 on the left represents the logical model of the generic ESA PDGS, as defined during the ADAR study and based on the OAIS Reference Model.

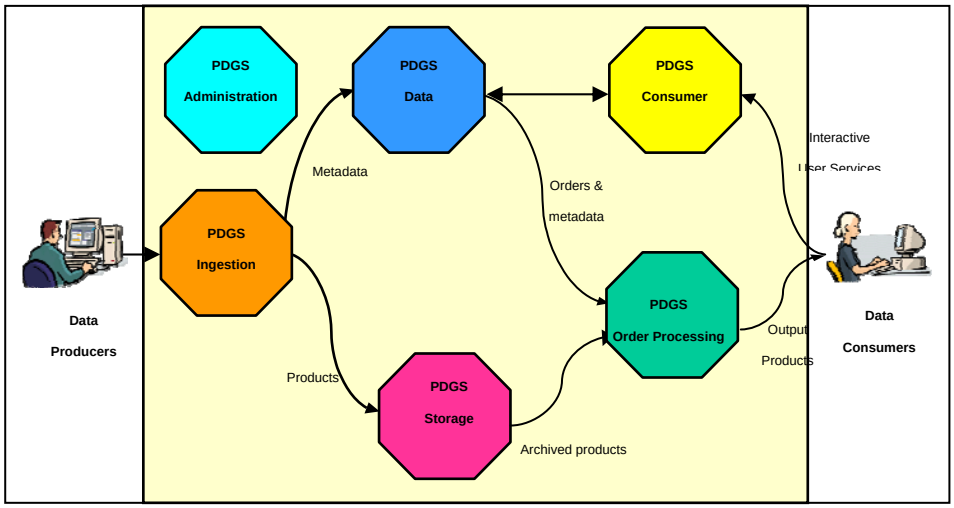

Figure 1 - Payload Data Ground Segment Logical Model (OAIS-based)

The MMFI does not only provide the long term archiving function required by a generic PDGS. Also data ingestion, processing orchestration, data dissemination to end-users, and other functions are implemented with the MMFI (see ref. [3]). In this paper we'll concentrate on the implementation and evolutions of the MMFI "archive function" and highlight where the solution has impacts over the other components.

HIERARCHICAL STORAGE MANAGEMENT

Tape storage is the basic technology utilized for the Long Term Data Preservation; on the other hand, the need for faster access, both in reading and writing, to the data requires a disk-based storage layer to be dimensioned based on the required I/O performance. It is also common in certain architectures that other levels of storage, either disk- or tape-based, are utilized to increment the performance and reduce the overall storage cost. The migration among the different levels of storage, from the faster and usually more expensive to the slower but cheaper storage levels, is normally managed by a hierarchical storage management system. ESA started to study the archive software for its EO data from 1996 and one year later two COTS were identified as archive infrastructure: AMASS and SatSTORE. AMASS, the HSM,

stores the data on tapes while the management of the data and the client interaction is demanded to SatSTORE. From the transfer to operations of AMASS and SatSTORE to fulfill new missions and new operational requirements it was needed to perform evolutions (user access control, request priority, data handling) both to SatSTORE and AMASS. The system utilized today is named AMS (Archive Management System) and represents the Long Term Archive of the ESA Multi-Mission Facility Infrastructure.

When on 2006 it was made public that AMASS was at the end of the product life and its upgrades stopped, ESA started the process to select a new HSM. The drivers of this process, together with the needs to have a different HSM, were to try to separate the data storage functions (storage layers, multiple copies, ..) as much more as possible from the data/client handling functions (data transfer to/from clients, data sub-setting, access control lists, ...). The result of the AMS evolution project, that has investigated the availability and functionalities of several HSMs, has identified Sun SAMFS as the substitute of AMASS. The SAMFS COTS (file based HSM) has all features requested:

- Storage tiering, the files can be hosted on more than one tier that can be tape or disk based.
- File multiple copies; a file can have up to four concurrent copies (on tapes or on disk).
- Different data handling based on single file or directory (based on regular expressions matching the file or directory name); it is possible to configure in a temporary or a persistent way the method used to retrieve the files from tape: by block or full file.
- File based HSM, the default SAMFS behavior is to deliver to the client the whole file in one shot. This is a basic issue considering the continue increasing of the tapes capacity.
- SAMFS tape can be read (using the Star utility provided) also if the HSM daemon is not available.

It should also be noted that Sun has released SAMFS as Open Source, the code is available to the community and consequently its life cycle can be not strictly related to the Sun business strategies.

Another important aspect driving the choice of SAMFS has been the possibility, via a special agreement with Sun, to have a SAMFS license tailored to the specific ESA needs. This was particularly important in the ESA EO archive case because the archive is distributed among various centers in Europe (9 presently) and the need exists to transfer data from one centre to another. In addition the archive operations of data from new missions are normally awarded in open competition and the need exists also to be able to extend the storage capacity of one or more centers. In order to simplify this important aspect of the archive management, ESA has negotiated and acquired from Sun a global SAMFS license volume-based (the standard license is based on the number of tape library slots) of 40PB (PetaBytes) for all its archive centers. While maintaining this overall limit, ESA has the right to move data from one archive to another, to increase the size of a specific archive and duplicate in the same centre the SAMFS systems in order to improve if necessary the archive performance.

Although the libraries are considered only hardware they are deeply involved in the Long Term Data Preservation. A library has to be flexible such that it is possible to upgrade tape drives technologies with a negligible impact in the operational infrastructure; e.g. having a mixture of tape drive technologies during the data transcription. In the course of the years, various tape libraries have been used, from GRAU AML/J (located in ESA ESRIN) to the StorageTek 9310 PowderHorn (located on six sites).

In the course of the study for the HSM replacement the need for a library hardware upgrade was also identified. The following tape libraries are today used for the ESA EO archive:

- Sun Storagetek SL8500, located in ESA ESRIN (Frascati, Italy).
- Sun Storagetek SL3000, located in Matera (I), Maspalomas (S), Kiruna (SE), Farnborough (UK), Oberfaffenhofen (D).
- Sun Storagetek SL1400, located in Tromsoe (N) and Sodankyla (Fi).

The new libraries are modular and extensible, in term of number of tapes and installed modules, and allow the installation of a maximum of 64 drives (except the SL1400 that have the possibility to mount up to 24 drives) enough to support the future EO missions.

TAPE ARCHIVING TECHNOLOGIES

The tape technology is directly related to the uses that the HSM makes of tapes and to what the whole operational infrastructure requires to the archive (in term of throughput, use of the data and size of the data-file).

The first tape drive technology used by the AMS was the DLT4000; the way how the drive performs read/write operations and its throughput, 3 Mb/s on average and an access time of 65 seconds, was well utilized by AMASS. AMASS performs the reading/writing operations by block, the block size being defined via a fine tuning process. Every time a read or write operation is requested AMASS divides the file in '<file size>/<configured block size>' blocks, which enables the HSM to divide a single request in multiple atomic requests and to use tape technologies that have a good positioning speed but low throughput (like the DLT4000) in a more efficient way. As a consequence of the block management, if more than five big files are requested on the same tape AMASS tries to serve all of them at the same time performing continuous repositioning on the tape (every four contiguous blocks of the same file). This results in a decreased overall I/O performance (see Figure 2). In order to resolve this problem, in the MMFI a specific function was developed in SatSTORE to serialize the requests on the same tapes when they are more than five.

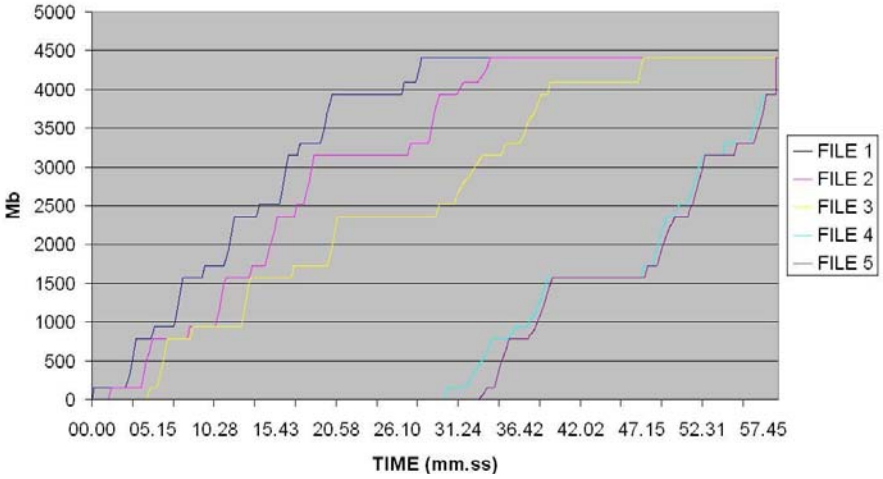

Figure 2 - AMASS: parallel retrieval of five files from same tape.

The upgrade of the DLT4000 to StorageTek T9940B increased the overall AMS performances but the increased speed of the drives also rendered not convenient any more the AMASS block size access feature. The average T9940B drive speed is 29 Mb/s while the average access time of 50 seconds, which implies that the full file download is much more efficient than tape repositioning to extract different blocks in parallel.

The requirement of more capacitive archives, throughput enhancement in read/write operations and the HSM substitution, which works by file and not by block, has also speed up the evolution of the tape drives technology. The Sun StorageTek T10000B has an uncompressed capacity of 1Tb and an average speed of 120 Mb/s. The usage of the new archive technology thus will allow the AMS to have five times the current archive size and six times the read/write drive speed compared to the previous T9940B tape technology. The T10000B tape management uses an enhanced tape mount process that does not touch the tape surface, thus increasing the tape life time and use. Furthermore, the increased performance of the T10000B in mount/dismount times and throughput permits a smaller number of tape drives in order to achieve the same global performance of the tape library.

DISK-BASED ARCHIVES

Although a true Long Term Data Preservation policy requires to archive safely on tapes the data, the operational infrastructures often needs to access the data in the fastest way possible to perform data

retrieving and production. The use of an HSM, with its storage tiering capability (like SAMFS), helps the current infrastructure to cope with this requirement. However, not all the disk based archive can be, or is convenient that it is, managed by the HSM. The MMFI has four disk-based archives:

- MMFI central cache: this is the disk archive used inside the MMFI to exchange data among the various components.
- AMS disk archive: this is the disk archive to store the data when a faster access to them is needed (e.g. auxiliary data files or frequently accessed products).
- HSM disk cache, the cache used by the HSM as first archiving tier.
- MMFI On-line Archive.

The MMFI has started its operational activities using host based NAS and NFS protocol to share the disk storage space, solution mainly dictated by costs. These MMFI disk archives are in the course of upgrade to more efficient and reliable hardware.

In order to rationalize the various disk archives in term of hardware and allow synergies between them, it was decided to evolve, where possible, the disk archive infrastructure in a unique enterprise level disk array that can serve the whole infrastructure (see Figure 3). This will simplify the management and reduce the operational costs of the complex archive infrastructure of the MMFI.

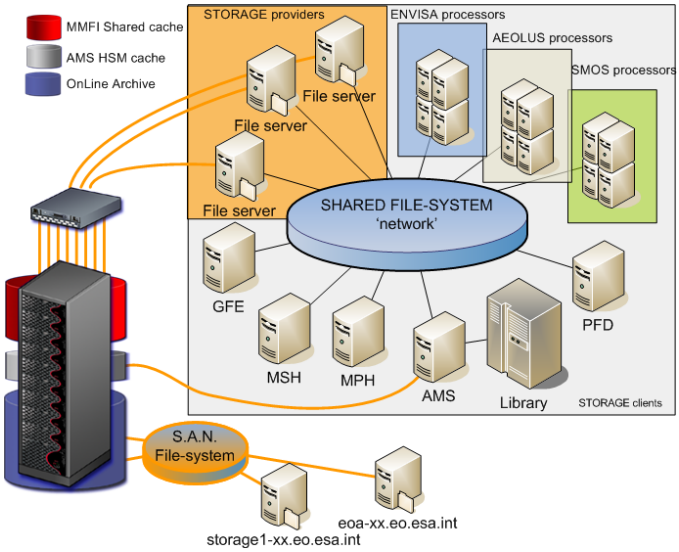

Figure 3 - MMFI future disk-based archive infrastructure

The On-Line Archive storage is located in the MMFI DMZ; it was upgraded to a mid-range disk array that shares its space among all the MMFI components located in the DMZ using the Red Hat Global File System. The connection of the storage between the MMFI components are done using FC cables that permit a theoretic speed of up to 4x4Gb/s (if the disk storage is properly configured in term of disk number and types). In consideration of its peculiar location and consequently its security policies, the On-Line Archive can not be directly reachable by the MMFI components that are in the MMFI private network. It can be only accessed using ftp and ssh protocols. The new disk archive for the On-line archive was designed to favor capacity rather than performance, using a high number of SATA disks (tuned on centers needs), however due to the disk array features a mixture of disks technologies is possible, consequently it is possible to enhance the infrastructure with high performance disks.

The MMFI Central Cache and the AMS disk archive is today a NAS that shares two partitions using NFS; they will be both upgraded to an enterprise range disk archive that will share the space among the MMFI components and the AMS via a high speed network (e.g. 10 Gb link infrastructure, InfiniBand). The new central cache will have a direct connection to the HSM in order to implement the 2nd copy functionality (the primary copy of the data, mainly auxiliary data files, is hosted on disk in order to be accessed in the fastest possible way) while a shared file system will enable all the MMFI components to access part of it to exchange data among them.

The HSM disk cache is the first archive tier of the HSM. Currently it is based on a Storagetek D240 disk array that will be substituted by a portion of the enterprise level disk array that will be chosen. Its requirements are mainly to be capacitive in order to speed up the ingestion and retrieval process of data-files and fast enough to archive the data on the n T10000B tape drives (average speed 120 Mb/s) installed in the centre's library.

DISTRIBUTED FILE SYSTEM

The introduction of an enterprise and centralized disk archive unlocks the door to the introduction of a shared file system to replace the NFS protocol. The study for the evolution towards a distributed file system has highlighted two different types:

- Direct attached storage: all the systems are interconnected via FC cables (SCSI over FC) to a storage array and have concurrent access to the same shared block storage.
- Distributed storage: all the systems are interconnected via FC, InfiniBand or copper cables (IP protocols) between each others but the storage disk is only attached to one or more of them, not to all.

The studies have highlighted that the current distributed file system used in most of the HPC environment is Lustre. It is an open source, scalable (10000 clients and more), widely used SW and has enterprise support if needed.

The current MMFI infrastructure is quickly evolving, integrating new/historical missions, increasing the product file size and performing, in the near future, massive data reprocessing campaigns. Its evolution forces to identify the most suitable solution to exchange data between MMFI as the current solutions (GB network, NFS and ftp) cannot satisfy the future requirements. Considering the MMFI as central infrastructure for the future evolutions and, consequently, that the systems interconnected can reach easily (in case of massive and fast reprocessing campaign) the number of 1000 systems, Lustre was chosen as distributed file system.

TECHNOLOGIES TO EXCHANGE ARCHIVED DATA AMONG THE ESA'S EO ARCHIVE

The current MMFI WAN infrastructure uses the GEANT network to interconnect the acquisition ground stations with the archiving centers, the PACs, using a VPN. The network is full duplex and provides accesses today between 64 and 200 Mbit/s. The ground stations transfer the data to the PACs using ftp, which implies that the majority of the network traffic is outbound, from the ground stations, and inbound, to the PACs. The current scenario does not widely use the PACs outbound network. In order to use all the available bandwidth of each centre, inbound and outbound, in a more efficient and complete way, a study project has been run in 2008. The aim of the project was to understand the feasibility of using the BitTorrent technologies to exchange data between centers. The project used the current open source BitTorrent client/server to build a prototype and test the performances of data exchange among more than three centers (see Figure 4).

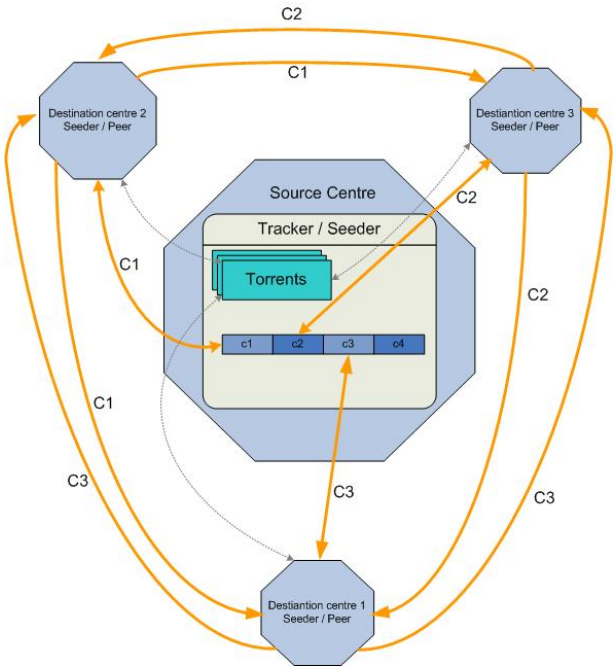

Figure 4 - BitTorrent distribution schema

The tests highlighted that the current network topology (GEANT) does not allow high speed traffic with the BitTorrent default 'network block size' (about 16kb). The modification of the 'network block size', up to 1 Mb, has shown an increase of performances that can be compared with the current ftp throughput. However it was not possible to achieve better performances as the BitTorrent protocol seems to decrease the throughput if the 'network block size' is larger than 1 Mb.

The study conclusions have highlighted that the development of an ad-hoc BitTorrent-like protocol, with a network packet size compatible with the operational network, and an ad-hoc application (to take into

account a proper monitoring and control) can enhance the distribution of data between the ESA centers (Ground Station vs. PACs and vice versa).

CONCLUSIONS

The MMFI is in operations at ESA since several years now and has proven to be very effective on one side in improving the performance of the PDGS of all missions, and on the other to decrease the overall operational costs for the ESA EO payload data exploitation.

The evolution of the MMFI is continuing to either implement new requirements coming from new missions plugged in or improve the overall performance or finally decrease the operational costs to a more affordable level. This is normally done by upgrading specific MMFI components or further standardizing and harmonizing the infrastructure.

The archive function being one of the most important carried out by the MMFI, its upgrade toward better and cheaper architectures and systems is an essential part of the MMFI evolution plan for the future.

REFERENCES

[1] ISO 14721:2003, “Space Data and Information Transfer Systems - Open Archival Information System - Reference Model”, Edition 1, February 2003

[2] CCSDS 650.0-B-1., “Reference Model for an Open Archival Information System (OAIS)” – CCSDS Blue Book, Issue 1, January 2002

[3] G.M. Pinna, E. Mikusch, M. Bollner, B. Pruin – “Earth Observation Payload Data Long Term Archiving. The ESA’s Multi-Mission Facility Infrastructure.” – PV2005 Edinburgh, 21-23 November 2005