MANAGED STORAGE SYSTEMS AT CERN

Ingo Augustin and Fabrizio Gagliardi
CERN, Geneva, Switzerland

Abstract
The amount of data produced by the future LHC experiments requires fundamental changes in the management of the mass storage environment. At present the contents of the CERN tape libraries are not transparently managed. The experimental data rates and volumes will grow by more than an order of magnitude in future. This implies major changes in the management of centrally stored data. Several efforts to address this challenge are described in this paper.

1. INTRODUCTION

Traditionally the majority of High-Energy Physics experiments have recorded their data locally at the experimental sites and later transferred the tapes to the computer center. Since 1995 experiments like NA48 send their data online to the computer center via dedicated fibers where it is stored centrally (Central Data Recording). The sheer amount of data (100 TB/year for NA48) and the data rates require high performance storage devices, which are too expensive for individual collaborations.

LHC will exceed the present requirements by orders of magnitude. ALICE plans to take data at more than 1GB/sec. Others will produce data at 100MB/sec. Each of these experiments will collect at least 1 PB/year (1 PB = 1 PetaByte $= 10^{15}$ Bytes ~ 1.5 million CD-ROMs or a bookshelf of a length of 5000 km). Although CERN will be one of the biggest storage facilities in the world, the storage itself of this data is not a problem. However, large storage facilities are usually used for backup or archives. This implies that the data is written once, and rarely read back. In our environment the situation is reversed. We write data once, but improved calibrations and reconstruction will require more than one pass reading the raw data. Efficient data retrieval becomes important. In order to achieve this, optimized access to resources (disks, tapes...) has to be guaranteed. This article describes some of the efforts CERN has undertaken to tackle this task.

2. THE CERN PETABYTE PROBLEM

The maximum data rate of the LHC experiments is 1-2 GB/sec. Taking into account that the data has to be read several times, a network (and of course storage) bandwidth of several GB/sec seems necessary. Current tape drives and disks operate with a bandwidth of at most tens of MB/sec, which implies that hundreds of devices are necessary to achieve the needed throughput. Even with tapes (or disks) of a size of 100 GB per piece, at least 10'000 of them are needed for each of the LHC experiments every year. Global collaborations and the more and more systematic production of analyzed data leads to round-the-clock operations of all computing systems. This, and the sheer amount of data, requires automated operations. Human resources are also scarce everywhere.

The future experiments will all used asynchronous readout models. Pipelining of data, buffering and highly sophisticated event filtering allows (or even requires) parallel streams of data. These streams can be handled and stored independently. The number and throughput of these streams can be adjusted to network and storage capabilities and is therefore highly scalable. Sufficient disk-buffers can de-couple the data acquisition performance from the central data recording, thus ensuring the highest possible performance of the tape system.

Usually the data has to be reconstructed several times due to improved calibrations and reconstruction software. The large amount of data requires centralized procedures to do this. A systematic reconstruction of the whole data can be viewed as the reversal of central data recording.

The classical HSM model features the migration from expensive, fast and small devices, such as RAM to inexpensive, slow and large devices (tapes).

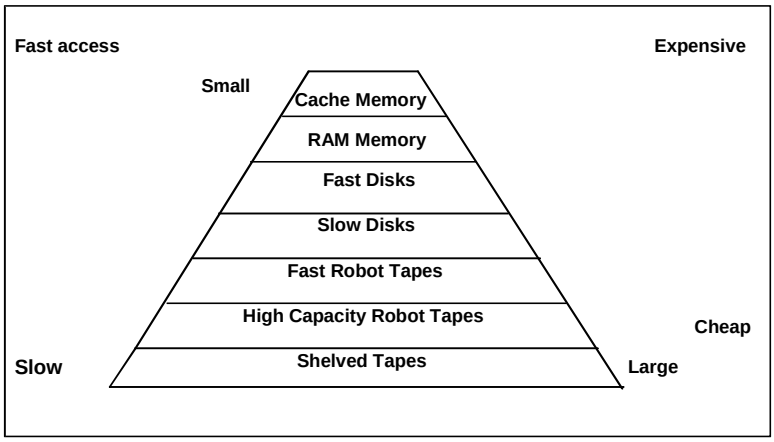

Figure 1: Classical storage model where the data migrates from high performance storage devices to cheaper and slower ones.

During the recent years this pyramid has begun to be distorted. Disks are getting bigger and bigger, and their prices have dropped, but the data rates did not change accordingly. At the same time the tapes became bigger and faster (with still more progress announced by the vendors), but the price is relative stable. Therefore, the relative financial share of tape storage in a HSM installation increases. The steady accumulation of large amounts of data makes it to costly to keep the data on disks for a long time. On the other hand, the data will be reprocessed several times during the first few years. Therefore access to the mass storage system is required. Optimization of these accesses is mandatory. Nowhere in the whole storage chain can performance be more easily degraded than at the tape level.

For example: the current workhorses of the CERN tape system are StorageTek *Redwood* drives. Transfer rates of about 10 MB/s and 50 GB capacity per tape volume are their principal characteristics. In terms of operations the time which is required to mount, load and position the tape is also important. If one transfers a file of 50GB to (or from) a tape, one achieves 10MB/s reduced slightly (2%) by the operational tape handling overhead (typically 100 sec for this type of drives). During the time of the overhead the drive can not be used for transfers. Many of the experiments use Linux as their standard operating system. At present Linux restricts the filesize to 2GB. This filesize appears reasonably large, but the impact on the performance of tape operations is now dramatic: 100 sec for tape loading and positioning, 200 sec for the transfer of 2 GB. This means the effective tape throughput came down from 98% to 66% of its design maximum. One third of the capacity is lost. With hundreds of drives being necessary to operate at the LHC data rate, this means hundreds of thousands of Swiss Francs that have to be invested additionally.

In this example the impact of the user (experiments) data model on the mass storage was shown. Access patterns like random access to the tapes are potentially even worse as the overhead easily becomes the dominant constraint of the input and output of the data. Therefore, the sequential mass storage on tapes has to be de-coupled from the user by using large disk pools (staging) or the users (experiments) have to get involved closely into the operations of the mass storage system. It is unlikely that the experiments will be keen to adjust their data models to the needs of hardware, which will certainly change during the decades of LHC operations.

The current model of LHC computing is built around computing farms, consisting of thousands of PCs. Although the assembly of such an amount of computers in a networking environment is not a problem per se (CERN already has this amount), the fact that they are working on the same data in a coordinated way provokes severe problems. Maintenance, configuration management and monitoring will be challenges. The computing farms can be used for online filtering or reconstruction during central data recording and for systematic reconstruction or analysis at later stages. In either way they will have to access the mass storage system, either to act as a data source or as a sink. Without optimization this kind of access will present itself as random access to the mass storage system. As described before this immediately introduces problems.

CERN investigated several routes to overcome this mismatch between the volatile PC-farm environment and the relatively inflexible mass storage environment. These will be described in a later section.

3. CERN STORAGE REQUIREMENTS

The CERN storage requirements can be separated in two general areas. First there are the objective requirements, which have to be fulfilled in order to do the job:

- Aggregate transfer rate at least tens of gigabytes/second with hundreds of streams in the 10-100 MB/s range.
- Continuous operation (operator attendance < 8 hrs/day)
- Storage capacity in the 10 - 100 PB range
- Thousand of simultaneous clients
- Large total number of files (~2^{64})
- File sizes only limited by operating system
- Reliable and error free transactions
- All this has to be achieved within the available budget (not yet defined)

It is standard CERN policy not to rely on a single provider for a system, if possible. Therefore a storage system should support different computer platforms and storage hardware. It is very likely that most of the computing for LHC is done on PC-farms, hence a support of these (at least as clients) is mandatory. Ideally the storage system itself would be running on PCs.

This last point leads to the second set of requirements. They are more a result of the need to achieve the goals within the given financial framework and the available human resources.

The experience with previous and current systems shows that easy manageability and efficient monitoring are key issues for the operation of a large storage system. Especially when users are accessing data on tape, resources like tape drives are easily wasted due to inefficient scheduling. Priorities, quotas, user groups fall in the same category.

The storage system should operate in a fully distributed environment.

These operational aspects are weaker in the sense that these topics can be used to compare and evaluate different systems, but are not 'show-stoppers'.

4. THE IEEE MASS STORAGE MODELS

In the early nineties the need for large storage systems became noticeable. The IEEE Storage System Standards Working Group (SSSWG) tried to identify the high-level abstractions that underlie modern storage systems [1,2]. The model should be applicable in a heterogeneous distributed environment. Many aspects of a distributed system are irrelevant to a user of the system. As a result, transparency

became the key point of the model. In this context, transparency means that the user accesses the data always in the same way: he does not know where it is or whether other users are using it. Behavior of operations and parameters should be always the same, regardless where the data is physically stored.

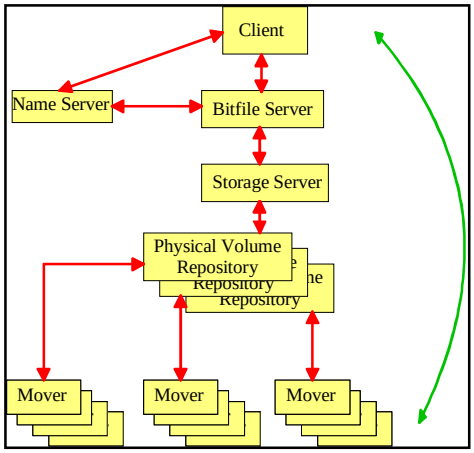

Figure 2: Components of the IEEE Mass Storage Systems Reference Model (V4)

4.1 Mass Storage System Reference Model Version 4

4.1.1 Mover

The Mover changes or monitors the read/write state of a device (e.g., positioning within the physical volume or performing loading and unloading operations).

4.1.2 Physical Volume Repository (PVR)

The PVR is a module that manages individual objects (cartridges, slots and transfer agents such as robots). It locates, mounts and dismounts cartridges on drives.

4.1.3 Storage Server

The Storage Server maps the physical volumes (cartridges, disks) to logical ones. The Bitfile Server as its client sees only a uniform logical space for each of the user or storage groups. The Storage Server consists of several sub-modules that allocate a PVR for a given data transfer, administrate the user/storage groups, enforce quotas. These modules become quite complicated if optimized tape, tape-drive or network usage is desired.

4.1.4 Bitfile Server

A bitfile server handles the logical aspects of bitfiles that are stored in one or more storage servers of the storage system. It creates the appropriate attributes of the bitfile like account-ID, access-control information, storage class, priority, backup information… Additionally the bitfile server maintains access statistics and records the real location of the file.

4.1.5 Name Server

If a file is submitted into a storage system, the bitfile server creates a unique bitfile-ID. The human-readable ID (/usr/xys/public/blabla/phd.tex) is convenient for people, but ambiguities can easily occur. The name server maps these filenames to the unique bitfile-IDs. This allows a storage system to be viewed as a global space rather than as space of host computers containing locally identified files.

4.2 Reference Model for Open Storage System Interconnection (MSS Version 5)

It is the successor of the previously described MSS model version 4. It is much more abstract than the old version. Mover and PVR are maintained, but the Physical Volume Library (PVL) and the Virtual Storage Service (VSS) have replaced everything else. The PVL manages physical volumes, such as tapes and disks. The VSS takes over the remaining functionality in order to present an abstract view of storage. Internally the old bitfile is now seen as composition of transfer units. This allows composition schemes, such as concatenation, replication, striping and various RAID methods. The big advantage is the possible redundancy and throughput during data migration. A file can be stored much faster when it is striped over several tapes (of course using the same number of drives). Unfortunately this involves the same number of tape mounts when the file has to be read back. In most tape libraries the data is read never or only occasionally. The situation in High-Energy Physics is different. Although one of the systems at CERN (HPSS) is capable of striping over tapes, this feature is not used.

5. CURRENT CERN PROJECTS

CERN investigated possible commercial solutions for quite some time and HPSS, a joint development by US DoE and IBM seemed the only potential "product" with enough performance and scalability to fulfill HEP requirements for the future generation of experiments. An HPSS evaluation project was therefore started in the fall of 1997. In parallel an internal project to develop further the in-house STAGER product was started. This was felt essential given the time scale of the COMPASS experiment scheduled to take data in 2000-2001. It was also decided to collaborate with DESY and a consortium of research and industrial partners in the EU supported Eurostore project.

5.1 HPSS Evaluation

HPSS [3] (High Performance Storage System) was first installed in the fall of 1997 on a test IBM system. The original test configuration included RS6000 AIX machines with IBM 3590 tape drive drives in IBM 3494 tape robots.

For HPSS to be successfully adopted by CERN multi-vendors support was essential. High level discussions with the HPSS consortium were therefore started while a joint project with digital was initiated to develop a port to Digital Unix of the HPSS data movers. This was carried out by a joint team of CERN and Digital experts during 1997/98 and delivered to the HPSS support team in IBM Houston for inclusion in the successive base line delivery kits.

A first prototype production service was started in August 1998. This included a user tape hierarchy for the Chorus experiment and Central Data Recording storage for the NA57 experiment. More than 2TB and 3000 files were managed.

An interesting test for the Objectivity/HPSS Interface was also implemented. This was particularly important given the strategic interest for Objectivity [4] and somehow the conflictual nature of the two products. This test stored 1TB for Atlas in 2000 files. Small testbeam setups for LHCB & Alice were also implemented.

Data is separated by *Class of Service.* The class of services (COS) defined were:

- COS 1 (User Tapes): 230GB AIX disk on 3590 tapes
- COS 2 (Raw Data on Redwood): 243GB DUX disk on Redwood tapes
- COS 4 (Testbeam): 32GB AIX disk on 2 Redwood copies (25GB, 50GB)
- COS 5 (Atlas 1TB Milestone): COS 1 disks on Redwood tapes
- COS 6 (Compass Test): COS 2 disks on Redwood tapes

The current tape mount rate (without NA57 data acquisition running) is about 100 per day.

5.1.1 Ongoing Work

A joint IT/Alice Mass Storage Project was started to investigate the use of a commercial HSM to manage their high acquisition rate of sequential data.

Milestones for recording “raw data” to tertiary storage were agreed:

- 30MB/s sustained in 1Q1999 (achieved with 4 sources & 2 sinks - IBM only)
- 100MB/s by 4Q1999 (need to “borrow” ~12 Redwoods when Alpha performance improved) - interesting to compare with the in house CASTOR project.

5.1.2 Experience and problems

- Architecture: random data access performance is slow via the standard provided API.
- Networking: it was difficult at the beginning to get HIPPI working, but it is mostly stable by now. HPSS software reliability is high but the overall service reliability is much dependent on network and media problems.
- Manageability: hard in the CERN computing environment. It assumes experienced operators to be available and to this we must add the cost of maintaining a DCE environment, which is otherwise not needed at CERN. The future of DCE is also questionable in the open commercial market.
- Operation: frequent reconfigurations caused service interruptions.
- Portability: this is very critical for CERN and HEP. The first implementation of the Digital Alpha 4100 data mover was slow (HIPPI to Redwood) in SMP configurations.

5.1.3 Successes

Excellent support from the HPSS IBM team in Houston was verified. The HPSS product itself has been stable, no data were lost because of HPSS. Sequential performance using the complex interface is only limited by hardware performance.

“Retirement” of old media to new products can migrate data gracefully. This is very important now that the lifetime of tape technology is not more than 4-5 years.

A single name space for all files is a good concept and works well.

HPSS allowed an efficient use of the shared disk pools and tape storage. Tapes are filled completely with consequent reduction of media consumption. The interface to HPSS has been developed using CERN standard libraries such as rfcp and the CERN stager. This allows new user applications to be added quickly and with minimum effort from the users.

5.1.4 CERN HPSS Plans

The original idea was to decide by end of 1999 if to commit to HPSS (i.e. use it for Compass in 2000) or drop it. However the Alpha port support and packaging by the HPSS consortium is not complete yet.

The first components of the Sun Solaris port in development at SLAC are now in the product. The BABAR experiment has started with HPSS and Objectivity at SLAC and at IN2P3 in Lyon, therefore we will be able to learn much soon.

The current strategy is therefore to continue a low level production use of HPSS to gain more experience and be ready to reconsider it as the final solution in case of positive and conclusive product developments.

5.2 CASTOR

In January 1999, CERN began to develop CASTOR ("CERN Advanced Storage Manager"). Its main goal is to be able to handle the COMPASS and NA48 data (25 + 35 MB/s) in a fully distributed environment and in a completely device independent fashion. Scalability should be good so we could also handle LHC data (100 MB/s per experiment) starting in 2005. Sequential and random access should both be supported with good performance.

CASTOR objectives are therefore:

High performance, good scalability, high modularity (to be able to easily replace components and integrate commercial products as far as they become available and show a better total cost of ownership and price performance factors).

CASTOR will provide HSM functionality with a single large name space. Migrate/recall functions are all focussed on HEP requirements, therefore keeping the design and implementation simple and less expensive.

It will be available on all Unix and NT platforms and will support most SCSI tape drives and robotics. System dependencies are grouped in few files to ease portability.

A user/operator/administrator graphical interface (GUI+WEB) is foreseen, but a command line interface will be retained for more automatic production use.

In the spirit of the CERN developed software it should remain easy to clone and deploy CASTOR outside CERN. CASTOR aims at using as much as possible commodity hardware components such as inexpensive PCs as tape servers.

The first version of CASTOR will be deployed at CERN during winter 1999/2000 and a large test (100MB/s during 7 days) will be attempted for ALICE in February 2000.

Support for Storage Area Networks (SAN) will be integrated with the goal of decreasing the number of data movers and servers. In the SAN model CPU servers are directly connected to the disks and share data. This is a move away from the traditional tape and disk server model, eliminating data copies between disk servers and CPU servers. SAN uses native filesystems, which give much better performance than NFS. It is important to acquire expertise in the area of emerging technology but even with SAN some HSM functionality will still be needed.

Support for different data storage models is being planned:

- disk pools
- local caches
- Storage Area Networks
- local disk and tape drives.

5.3 EuroStore

The third project with CERN participation is EuroStore [5], an European Union funded ESPRIT project. CERN, QSW (a supercomputer manufacturer) and DESY formed together with various smaller European enterprises a consortium to develop a scalable, reliable and easy manageable storage system almost entirely based on Java. QSW developed the Parallel File System (PFS), which is used in their high performance computer systems. DESY was the developer of the HSM system. The current storage system of DESY will reach its limits of scalability with the appearance of HERA B. The similarity of requirements for a storage system at CERN and DESY made a collaboration in this field desirable. The role of CERN and the commercial collaborators was the definition of user requirements and the assessment of the developed software according to these requirements.

The excessive requirements of LHC in terms of scalability and reliability, together with the necessity of flexible administration and maintenance, made up the bulk of the user requirements for the EuroStore software. These user requirements have been used also for HPSS, CASTOR and MONARC [6].

Similar to HPSS the EuroStore HSM is based on the IEEE mass storage standard. The complete HSM service is built out of sub-services implemented as separate processes or threads running on the same or different heterogeneous platforms. The communication mechanism between all these sub-services is done with a secure message passing environment, called Cell-Communication.

The HSM supports the notion of Storage Groups to allow a single Store to be divided into several sub-domains containing specific user groups and/or dataset types. The Store represents the root of the internal HSM object structure, which is built out of Storage Groups. The Storage Group is further subdivided into Volume Sets, which act as the source and destination for the HSM internal migration of datasets. The Volume Set is itself built out of Volume Containers defining the set of physical volumes belonging to a single physical library. To describe and control the internal HSM migration there exists an object, called Migration Path, which encloses the migration condition and the source/destination Volume Set. Each dataset stored in the HSM has a link to an existing Migration Path describing the dataset migration characteristics.

The HSM provides a simple service to the PFS (or other clients), namely storing and retrieving complete datasets (or files in the PFS nomenclature) sequentially. A future version of the EuroStore HSM might support read operations on parts of datasets (partial reads). This simplicity is mirrored in the data access API in that it contains only 3 functions: create/write a dataset, read an existing dataset and remove an existing dataset. In addition, the API will support simple query operations (ask for all files on a given volume, etc.) for its clients (like PFS). The data access API is implemented as a C based thread safe library.

The PVL supports additional functions:

- Priorities, specified by the client application. This was an important requirement of the EuroStore collaborators of the Health sector.
- Configurable numbers of write operations on a given Volume Set. This allows the choice between storage in chronological order, as in Central Data Recording, and the policy based selection of available resources (the PVL would choose a volume and a drive according to the current situation).
- Regular expression assigned to a storage device (drive). The PVL will manage a defined set of mainly request dependent variables that can be used to construct a regular expression. For example, a drive might be available during the time between *3:00* and *4:00* only for a user called *oracle_backup* on the host *oracle_server.cern.ch*. During all other times other users could use the drive.
- Virtual library partitioning allows dynamic resource allocations like "*20% of the tape drives are given to a certain client/user-group*".

The modular design of the EuroStore HSM provides the necessary scalability. Every component (e.g. movers, PVRs, PVLs, Bitfile servers) can be located on a different computer. The implementation in Java will provide the necessary portability to cover a wide range of heterogeneous hardware platforms.

The EuroStore prototype was deployed at CERN during April 1999. The hardware platform consists of four dual processor SUN Enterprise 450 servers. Each of the servers is equipped with four 8 GB hard disks, which build the components of one or more Parallel File Systems. The PFS can be

striped over several nodes of the cluster. The data is transferred between the nodes via a switched ELAN/ELITE network (max. 250 MB/s). Each of the E450s is connected to the CERN LAN with Gigabit Ethernet. At present the prototype uses two StorageTek 9840 tape drives, located in one of the automated tape libraries of CERN.

During the initial assessments many configurations have been tested and, except for the usual programming bugs, no conceptual problem could be found. The GUI based administration and management of the HSM system proved to be very effective and flexible. The implementation of an HSM in Java has been successfully demonstrated, although the issue of performance and reliability could not be really addressed yet, due to the ongoing development of the prototype. The EuroStore project will continue until summer 2000. DESY intends to deploy the EuroStore HSM as a production system at the end of the project.

6. FUTURE DEVELOPMENTS

The current plans at CERN are to continue the lines of development described above while exploring ways to increase the interoperability of HEP HSM systems.

7. CONCLUSION

It is clear that while commodity components computing seems to offer scalable and affordable solutions for LHC computing, the management of the data will remain a difficult challenge to tackle.

Disk storage is the only component which seems so far to follow the Moore price evolution curve of PCs. Tape and robotics seem to stagnate or have a very slow evolution at best.

The HSM commercial market doesn't seem to match HEP requirements, although some analysts predict tremendous growth in this field in a near future.

Until this happens probably we need to take a conservative approach and develop simple in house solutions in collaboration with other major HEP centres which share our needs.

We should in parallel continue to monitor technology and market evolution to spot trends, which we could exploit, and commercial products, which we could acquire.

The impact of large distributed data access models such as the ones investigated by MONARC should be taken into appropriate consideration.

ACKNOWLEDGEMENTS

The authors want to acknowledge the contributions of the Data Management section of the PDP group of the CERN IT-division. In particular J.-P. Baud, C. Curran and H. Renshall provided valuable input by reviewing the manuscript.

REFERENCES

[1] S. Coleman and S. Miller, ed., Mass Storage System Reference Model Version 4, May 1990, Technical Committee on Mass Storage Systems and Technology, Institute of Electrical and Electronics Engineers.

[2] A. Abbott et al., Reference Model for Open Systems Interconnection, September 1994, IEEE Storage Systems Standards Working Group (P1244), Institute of Electrical and Electronics Engineers.

[3] http://www.sdsc.edu/hpss/hpss.html

[4] http://www.objectivity.com

[5] http://www.quadrics.com/eurostore

[6] http://www.cern.ch/MONARC

RELATED WEB PAGES

http://nicewww.cern.ch/~fab/default.htm

http://www.cern.ch/eurostore/

http://www.ssswg.org/

http://home.cern.ch/~hepmss/

http://www.nsic.org/

http://www.snia.org/