Test Results of the EuroStore Mass Storage System

I. Augustin

IT-Division, CERN, Switzerland

Abstract

Within the framework of the European Union ESPRIT program the EuroStore consortium developed a prototype of a distributed filesystem with an integrated storage management system. The design focussed on the storage requirements of CERN for the oncoming LHC experiments. This paper will describe the project and will cover the initial assessment of the EuroStore system.

Keywords: mass storage, tape storage, filesystem

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 $\approx$ 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. Several approaches have been taken to tackle this problem. One of them is the EuroStore project.
In March 1998 it was started in order to develop a modular storage system according to the requirements of the LHC environment (among others). The project will provide a prototype of the storage system. The prototype software comprises a distributed, network centric storage management system with a parallel filesystem as its main client. The aim of the partners is the assessment of the technical feasibility of the designed system according to their different requirements. The project will end in summer 2000.
The consortium consists of the following participants:

- Quadrics Supercomputer World Ltd (QSW) is a British-Italian supercomputer manufacturer and developer of the parallel filesystem (PFS).
- DESY is developer of the storage management component (HSM).
- CERN hosts the prototype installation and acts as testcenter and end-user.
- Industrial and research end-users:

- Hellenic Company for Space Applications, a Greek consulting company for the Greek aerospace industry.
- Athens Medical Center, private operator of several private hospitals in Greece and Russia.
- Hellenic National Meteorological Service
- TERA Foundation, an Italian foundation for hadron therapy

2 The EuroStore Software

The EuroStore software consists of two major parts. The parallel filesystem (PFS) is the part that can be accessed by the user (see figure 1). The PFS is the client, which interacts with the storage system (HSM) through an C-API. Using this API other clients, such as rfio or dabase systems could also access the HSM, but due to the limited scope of the project these features are left to future developments. The HSM itself interacts with the robotics and any physical storage device beyond the parallel filesystem. In the prototype these were two StorageTek 9840 tape drives in a automated robotic silo.

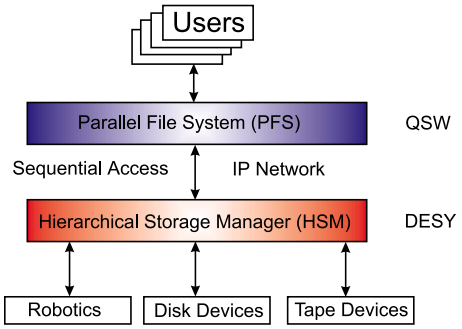

Figure 1: EuroStore basic software components

2.1 The Storage Management System

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.

3 The Test Setup

The EuroStore software has been tested on a dedicated Quadrics QM-1 system, which consists of 4 Dual-Processor SUN E450 servers. The machines are interconnected by a high-speed, low latency switched network. This QSW proprietory ELAN network is used to feed the striped PFS. The PFS consists of sixteen 8GB disks, arranged as one or more uniform filesystems. For data transfers from and into the EuroStore system each server is equipped with a Gigabit-Ethernet interface. The mass storage consists of two STK 9840 tape drives (see figure 2).

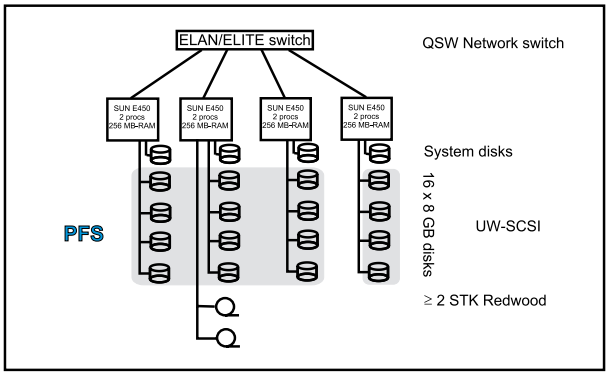

Figure 2: Standard setup for the QM-1 EuroStore harware platform

This setup has been tested in various configurations. The initial task has been to ensure the functionality of the various components, eg. movers, API, PFS. The system has been able to store and retrieve data successfully at a speed which was only limited by the according hardware setup. Instabilities of the system have been identified to originate in implementation details (vulgo: bugs) and not in the design itself.

One major concern for a large storage system is the manageability of the system. Sufficient monitoring is needed to identify bottlenecks, which limit the performance of the system. The truely modular design of the HSM in combination with the possibility of intervention and interrogation of each of the sub-modules allows an extensive monitoring tool (see figure 3. Each job can be traced throughout the whole system. These monitoring tools are provided by an http server and available to every user.

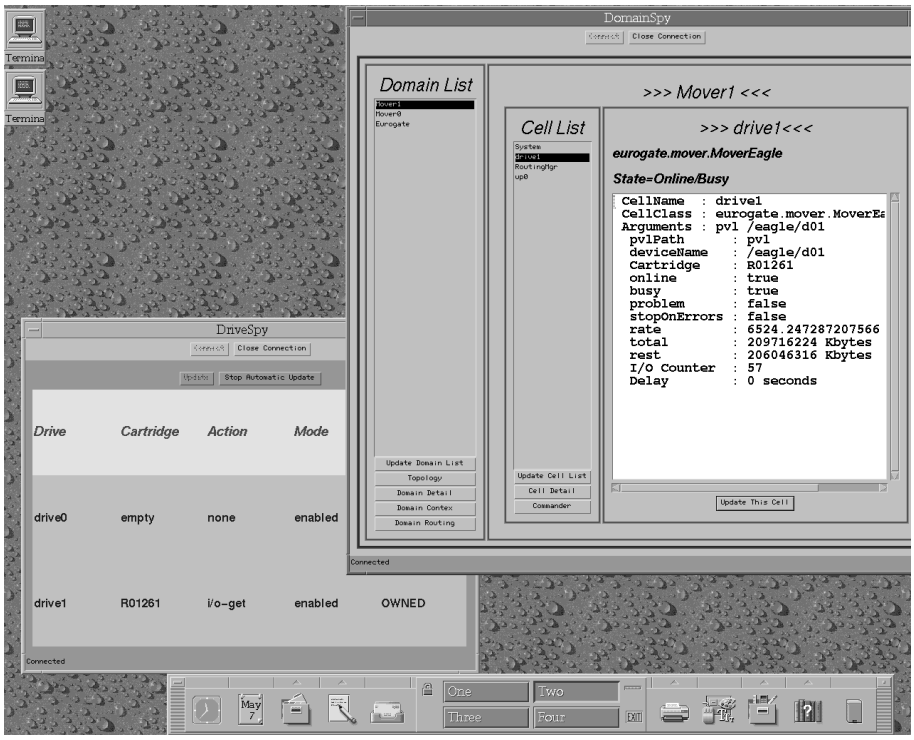

Figure 3: Example for the monitoring tools of the HSM

For administration purposes there exists a second variety of these tools. They run locally and include features that allow to interactively change the system configuration. For example tape status can be changed or parts of the system can be brought up or down. For example the power cycling of a mover will not influence the storage system (except for performance obviously).

4 Conclusions

The EuroStore prototype has been proven to operate. Not all the features are implemented yet, for example the evaluation of regular expressions is missing, but will be done during the next months. The EuroStore prototype with the PFS as the only client of the HSM system is definitely not an option for the LHC storage. The concept of one or more centralized filesytems is reasonable for industrial applications, but does not serve the thousands of LHC users with their great demand for data very well. To overcome this limitation a follow-up proposal has been submitted to the European Commission. It includes SAN support, hints to the HSM, port of all the components to Linux and use of commodity hardware among other features.

References

1 More inbformation about EuroStore at `http://www.quadrics.com/eustostore/`