The EuroStore Project − Results and deployment @ DESY

M. Gasthuber[1], P. Fuhrmann[1], D. Roweth[2]

[1] Deutsches Elektronen Synchrotron (DESY/Hamburg), Germany
[2] Quadrics Limited, Bristol UK

Abstract

The EuroStore project has demonstrated the feasibility to built a highly scalable and easy to run storage system covering the core requirements of tomorrows applications in HEP. The resulting prototype will be used as a base for a newly proposed follow−on project. The new project will approach to a full product with new features covering large−scale storage demands for this decade. This paper will outline the usage of the EuroStore results within the DESY production storage environment.

Keywords: Mass Storage, EuroStore, PNFS, Disk−Cache, HSM, Java, Filesystems

1. The EuroStore project

The EuroStore project starts in March 1998 as an EC funded ESPRIT project[1] for the duration of 2 years with the goal to develop a prototype of a highly scalable and high performance mass storage system for the industry and research community. The project is formed from industry and research partners[2] with basic distinction between development and end−user partners. The development

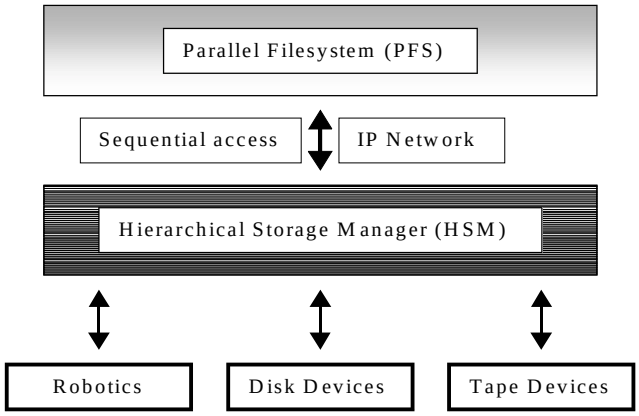

Figure 1: basic design

activities are spitted between QSW and DESY. The overall design approach are shown in Figure 1. The project have selected the Solaris platform for the prototype reference and development base. For

1 ESPRIT Project 26317
2 The partners are: QSW, DESY, CERN, HNMS, AMC, TERA, HCSA

more detailed information about the EuroStore project, its goals, the user requirements and the design see[1].

1.1 PFS

The parallel filesystem PFS builds the access layer for the applications serving a standard POSIX compliant filesystem (shared and random access). PFS can be constructed out of several client filesystem (it is a UNIX System V layered filesystem) of various types (UFS, NFS). PFS itself has two major components − the map filesystem, storing the mapping and other information about the data file, and the data filesystem storing the actual user data. The data filesystem type is one of the existing filesystem like UFS or NFS. PFS can be configured in various ways to allow extreme high performance and/or secure configurations. PFS exists prior to the EuroStore project and was extended during the project with the following features:

- HSM interface with file migration support (including Space Management)
- MPI File I/O support (for parallel applications)
- mmap implementation
- concurrency control with global lock manager instance
- backup and restore in a HSM aware fashion
- RMS (Resource Management System) integration

1.2 HSM

The HSM is the component managing the complete tertiary storage (today mostly tapes). It serves as a 'black hole' for data with a very simple interface to the client − the parallel filesystem PFS. The interface act as a put/get interface dealing with complete files and a flat namespace for stored objects (the bitfile id). As usual the control and data path are separated and today both are using the TCP/IP protocol for communication. The HSM was developed from scratch and implemented in Java which opens up new possibilities for the implementation. The HSM is built out of sub−services implemented as separate process and/or thread, to construct the resulting, network based, HSM service. This allows very good scalability combined with very good performance characteristics. The communication between all these separated sub−services is based on a message passing environment[3] sending and receiving Java objects in a secure (ssh protocol) way. Built into the HSM are interfaces allowing easy migration to other/new storage hardware (drives and robotics). The administration tasks are either based on the standard ssh application (command line) or on a secure (signed and through https) Java Applet as the graphical alternative.

1.3 Results

The prototype system installed at CERN are running on a farm of 4 Sun E450 machines connected through QSW's Elan network (proprietary) and fast plus gigabit ethernets. The Elan network supports a high data transfer rate combined with a very low latency supporting IP and is thus the preferred data path for the HSM − PFS data exchange. The machines have ~150 GB disks connected to all of them running the PFS filesystem. The tertiary storage is built from STK Powderhorn libraries housing STK 9840 (Eagle) tape drives. I. Augustin[2] present more details of the system setup and characteristics combined with the initial assessment results. The initial impression shows that:

- PFS has a high potential to be extremely performant
- The Java implementation of the HSM causes no performance penalties
- The system shows the expected scaling options
- Data Management can be done with less complex/big systems

2 The Model

The long term abstract model of the Data Management architecture at DESY is built out of three major components, which are:

- The nameservice − mapping of human readable pathnames to system specific informations, like bitfile−id etc.

3 Called 'Cell Communication' which exists prior to the EuroStore project

- The disk cache − serving shared/random access storage to the clients and cache data from the tertiary storage.
- The HSM − the tertiary storage manager serving the perfect 'black hole' for data. The data exchange to and from the HSM is sequential.

Figure 2 shows the major players and the interaction paths to the clients. DESY today has two

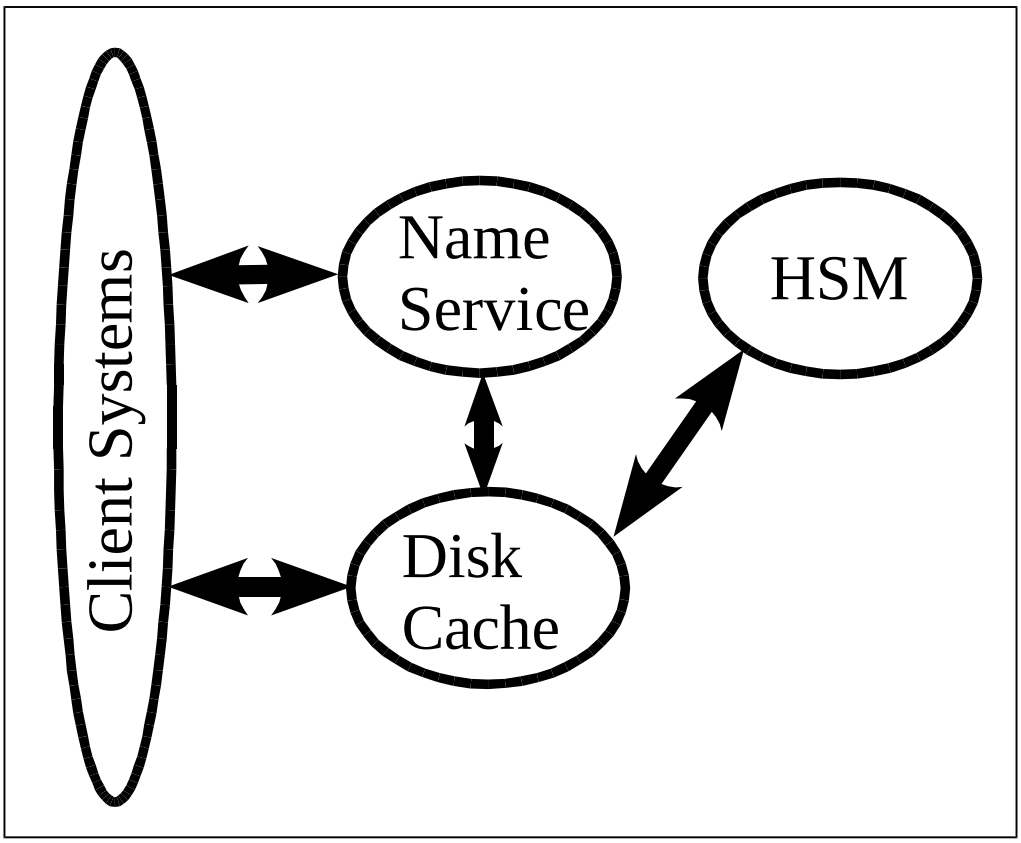

Figure 2: major players in the model

components of the above model in production (Nameservice and HSM) whereas the HSM functionality is currently served by the old OSM system. The nameservice component we use PNFS, a versatile nameservice mapping pathnames into 7 distinct levels of opaque data (i.e. bitfile id). The client access protocol is NFS v2 to allow nearly everybody to participate from the nameservice. For further information regarding PNFS visit http://watphrakeo.desy.de/pnfs or look at [3]. The Disk Cache is currently under development process and the prototype has been completed in December 99. The development is a joint collaboration between FNAL and DESY. For more information regarding the Disk Cache project and its goals, see [4].

3 Implementation

The implementation plan for the short term planning is to use PNFS as the nameservice, the newly created Disk Cache (plan is to have a pre−production version by Q2 2000), and the EuroStore HSM as a replacement for the old OSM system, serving the perfect *black hole* for all data. The initial plan was to use the EuroStore HSM as the base, but recently a new collaboration has submitted a follow−on proposal (EuroStore II) with the goal to extend the list of features and to built a *real* product[4]. If the proposal will be approved much of the activities necessary to run the EuroStore HSM prototype HSM in production will be embedded in the EuroStore II efforts. The timeline for the replacement of the old OSM system by the EuroStore HSM is now tentatively the end of the year 2000, depending on the approval and the real project start of EuroStore II.

4 Full Documentation, fully tested, more platforms (Linux), and more supported storage hardware

References

1 M. Gasthuber, P. Fuhrmann, D. Rowet, "EuroStore − Design and First Results", 16’th IEEE Symposium on Mass Storage Systems, 7’th NASA Goddard Space Flight Center Conference on Mass Storage, San Diego, USA, March 1999

2 I. Augustin, "Test Results of the EuroStore Mass Storage System", CHEP2000, Padua, Winter 2000.

3 P. Fuhrmann, "A Perfectly Normal Namespace for the DESY OSM", CHEP’97, Berlin, Spring 1997

4 P. Fuhrmann, "A Distributed Rate−Adapting Buffer Cache for Mass Storage Systems", CHEP2000, Padua, Winter 2000