Mass Storage Systems at Fermilab: An early experience with the High Performance Storage System.

Ken Fidler, Krzysztof Genser, Steve Kalisz, Jeff Mack, Alexander Moibenko, David Sachs

Fermi National Accelerator Laboratory, Batavia, Illinois, USA.

Fermilab is integrating and testing the High Performance Storage System (HPSS) as a foundation for providing large scale data storage and management services. Additional software tools called Fermilab Mass Storage System (FMSS) were developed to provide functionality beyond that of the native HPSS API. Initial experience with HPSS and FMSS is described.

With datasets increasing in size to many tens of terabytes today along with expected increases to hundreds of terabytes for Run II of the Tevatron Collider beginning in 1999, Fermilab has begun to deploy new generations of software and hardware to enable efficient storage, access and management of these datasets. Because of its robust design and scalable architecture, Fermilab has chosen the High Performance Storage System (HPSS) as a foundation for providing these large scale data storage and management services.

HPSS is a major software development project and collaboration. The primary HPSS development team consists of IBM Global Government Industry and four U.S. Department of Energy National Laboratories: Los Alamos, Lawrence Livermore, Oak Ridge and Sandia. Other collaboration partners who have provided significant contributions to the HPSS development effort include NASA Langley Research Center and Cornell University. At present, there are five other members in the collaboration designated as early deployment sites. These early deployment sites are working closely with the HPSS development and support teams to install, integrate and test HPSS functionality in a variety of large and complex mass storage environments. Fermilab is one of these early deployment sites.

HPSS architecture[1], fundamentally using distributed strategies, is based on the IEEE Mass Storage Model, Version 5[2] and allows separate network data and control paths. To provide a reliable and robust distributed system infrastructure, HPSS uses DCE Remote Procedure Call technology[3] and Transarc's Encina[4] as a client-server transaction manager. HPSS is scalable in several dimensions, including distribution and multiprocessing of servers, data transfer rates, storage capacity, number of files and their sizes. When additional capacity is required, its expandable topology facilitates adding new physical resources to the system such as processors, disks, tapes, robotic tape libraries and networks. HPSS provides multiple storage hierarchies which are useful to separate users with different storage needs. Its multiple storage classes allow for logical grouping of storage media with similar I/O characteristics.

HPSS is written using ANSI standard C following POSIX standards and importantly, requires no special modifications to operating system kernels. The major HPSS components are: Name Server (NS), Bitfile Server (BFS), Repack, Storage Server (SS), Migration, Purge Server (MPS), Physical Volume Library (PVL), Physical Volume Repository (PVR), Movers (MVR) and Storage System Manager (SSM).

The Name Server translates a human readable file name to an HPSS object identifier. Objects in the HPSS are files, directories and links. The Bitfile Server provides the abstraction of logical bitfiles to its clients. A logical bitfile is an uninterpreted bit string of a length up to 2**64 bytes. Bitfiles are mapped into storage segments which are handled by the Storage Server. The Storage Server maps

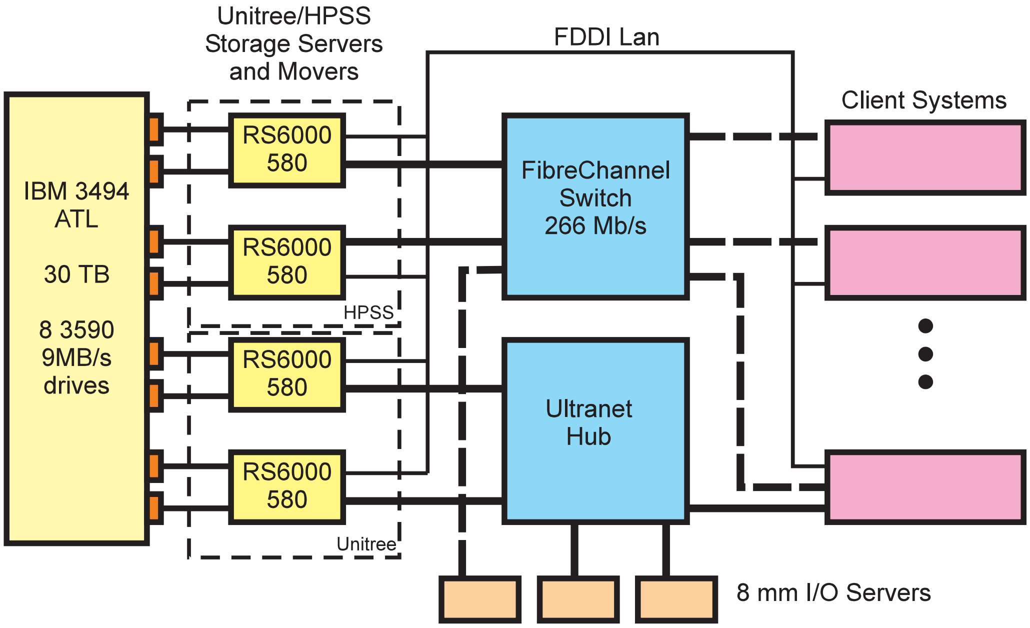

Fig. 1. Fermilab Central Mass Storage Configuration

storage segments into virtual volumes and subsequently into physical volumes. It also schedules the mounting and dismounting of removable volumes through the Physical Volume Library and provides defragmentation of physical volumes including tape cartridges. The Physical Volume Library maintains a mapping of physical volumes to cartridges and a mapping of cartridges to Physical Volume Repository. The PVL also controls drive allocation. The Physical Volume Repository manages all HPSS cartridges. Movers are responsible for transferring data from a source to a destination and perform a set of device control operations. The Migration-Purge Server provides a migration of bitfiles to lower storage class levels of storage hierarchies according to site-defined migration and purge policies. The Storage System Manager monitors and controls all HPSS resources according to particular management policies set by each site. SSM is built upon the SAMMI product, a GUI layout and runtime system from Kinesix. Important for sites migrating from NSL-Unitree to HPSS such as Fermilab, HPSS also includes tools to translate the metadata of its predecessor, NSL-Unitree. These tools could be enhanced to convert other hierarchical storage management (HSM) systems like Convex UniTree and DMF.

Fermilab's mass storage system configuration consists of two hierarchical storage management systems: NSL-Unitree and HPSS (fig.1).

Both share a single IBM 3494 robotic tape library with a capacity of 30 terabytes. The HPSS portion consists of one general HPSS server node containing the Name Server, Bitfile Server and PVL/PVR. Two Mover nodes are presently installed and interconnected with quarter speed (266 Mb/s) Fibre Channel. Each mover is connected to one IBM 3590 tape drive in the robotic tape library. Soon, two more tape drives and one mover node will be added. When NSL-Unitree — HPSS migration is completed in the next few months, the HPSS system will have four mover nodes and eight tape drives.

Five HEP groups currently use a total of 10.5 terabytes in 45000 files in the NSL-Unitree system. The total amount of data stored in HPSS is 3.9 terabytes in 222000 files and is used by six experiments. All new requests for mass storage resources are allocated only HPSS space. Two major groups have been migrated from NSL-Unitree to HPSS so far. No significant problems were encountered during the migration. The remaining Unitree data is expected to be converted to HPSS within a few months, most likely in one final step.

Being a member of the HPSS Collaboration, Fermilab has the opportunity to be on various committees to discuss deployment plans, future enhancements, as well as specific design issues. One of the current topics of particular importance to Fermilab is the definition of File Family — a concept to build functionality which will enable physical grouping of related files on sequential media to allow for even more efficient data access and minimization of the tape mounts.

As an early deployment site Fermilab has been assigned a member of the IBM HPSS staff as the direct contact. This individual has access to the Fermilab system and closely monitors and assists the various planning and administrative activities being performed.

Because two HSM systems are currently used at Fermilab, a semi-transparent and uniform user interface was developed which hides the HPSS or NSL-Unitree specific calls. This interface, the Fermilab Mass Storage System (FMSS), is a set of tools to store, retrieve and manage user data files maintained by the underlying HSM.

FMSS is implemented as a client-server system. FMSS client code can run on a variety UNIX systems including AIX, IRIX, and Digital UNIX and uses SUN RPC calls and a slightly modified FTP to communicate with the HSM. All the NSL-Unitree/HPSS specifics, including the DCE RPC calls are implemented in the FMSS server. FMSS provides a reliable communication and fast transfer mechanism in a distributed environment and secures the access to the storage system. A UNIX-like user interface makes FMSS easy to learn and use. The FMSS tools have the following functionality:

- Transparent access to data in one or more HSM systems.
- Automatic validation for securely storing and retrieving data files.
- Bulk data areas for storage of large files and archive areas for storage of smaller files.
- Independent storage allocations and quotas for different groups and subgroups.
- Data export to and from external (non-HSM) magnetic tapes.

When a user runs the FMSS client, the client first authenticates the user with the server. As soon as the access is validated, the user is able to retrieve, create and delete files, change permissions, etc... FTP is used to transfer the data. The respective HSM's API is used to handle file and directory names as well as their attributes. A mechanism exists to selectively grant access to mass storage areas not owned by the specific user.

Users have access to two storage areas, bulk and (in future) archive. The bulk area is a large storage area dedicated to storing experimental data. This area has no duplicate copies created because of the large amount of data. The archive area is intended for saving more critical data and is considerably smaller than the bulk area. To minimize the loss of data due to tape failures, duplicate tape copies of all files will be made for this archive area when implemented.

A FMSS base directory is provided for each group in the bulk area and for each user in the archive area. In the bulk area there may also be specific subdirectories for particular users of a group. The descriptions of all authorized groups and users are stored in the user table of the FMSS server. This table contains FMSS groups, user names, FTP user names and passwords, and additional information specific to the HSM system such as class of service for HPSS datasets.

Each group and user may specify a set of users allowed to access the mass storage system even if those users do not have direct access permission. Permitted host and user names are contained in .access files inside the group and user subdirectories. A special .quota file is used to enforce storage limits on a group basis.

The following set of commands is provided by the FMSS shell API: cp, rm, mv, mkdir, ls, chmod, status, query, activity. In general the command functionality is the same as for its UNIX counterpart. Other commands are specific to the HSM and FMSS.

To provide reliable communication and recovery from possible network and other failures, the FMSS client periodically checks the accessibility of the FMSS server during an established session. If the server is inaccessible, the client will try to reestablish the session after the expiration of a time-out. The FMSS server keeps track of all active clients and if within a specified time it does not receive any requests from the client, the client will be removed from the list of active clients. If one of the resources required by FMSS is unavailable or fails, FMSS will automatically retry the operation until either the request can be completed or the time-out limit is reached, thus preventing unwanted termination of long batch jobs accessing mass storage datasets. A logging facility is used to record significant events in both the FMSS client and server to provide information necessary for effective trouble shooting and usage analysis.

FMSS has been utilized at Fermilab since July 1996 using NSL-Unitree and since December 1996 using HPSS. Both versions have proven to be reliable and easy to use. Fermilab is actively participating in the HPSS collaboration to ensure that features and functionality important in solving large-scale HEP data management problems are incorporated into HPSS.

References

[1] Teaff, D., D. Watson and R.A. Coyne, *The Architecture of the High Performance Storage System, Fourth NASA Goddard Conference on Mass Storage Systems Technologies*, March 1995.

[2] IEEE Storage system Standards Working Group (SSSWG) (Project 1244), *Reference Model for Open Storage Systems Interconnection, Mass Storage Reference Model Version 5*, Sept.1994.

[3] Open Software Foundation, *Distributed Computing Environment Version 1.0 Documentation Set*. Open Software Foundation, Cambridge, Mass. 1992

[4] Dietzen, Scott, Transarc Corporation, *Distributed Transaction Processing with Encina and the OSF/DCE*, Sept. 1992.