FERMILAB-CONF-04-483-CD

THE STATUS OF THE FERMILAB DATA STORAGE SYSTEM.

J. Bakken, E. Berman, Chi-Hao Huang, A. Moibenko, D. Petravick, M. Zalokar
FNAL, Batavia, IL 60510, USA

Abstract

This document describes the Fermilab Data Storage System Enstore, its design concepts, structure, and current status. Enstore provides storage of the data in robotic tape libraries according to requirements of the experiments. High fault tolerance and availability, as well as multilevel priority based request processing allows experiments to effectively store and access data in the Enstore. Amount of data stored in the system currently approaches 2 PBytes. The Enstore system includes 5 robotic tape libraries, more than 100 PC nodes, and 90 tape drives. The distributed structure and modularity of Enstore allows scaling of the system and adding of more storage equipment as the requirements and needs grow. Users access data in Enstore directly using a special command. They can also use ftp, GridFtp, and SRM interfaces to the dCache caching and buffering system [1], which uses Enstore as its lower layer storage.

STRUCTURE OF THE SYSTEM AND ITS MAJOR FEATURES

The Enstore project was started as a preparation for the Fermilab Collider Run II experiments. The estimated storage requirements were 250 Mbytes/s of aggregate throughput sustained for a month of uninterrupted work. Estimated capacity of the storage was several petabytes. The storage system had to provide a centralized primary data store for all kinds of data including online and analysis. This required grouping of data sets according to certain criteria, prioritized request processing, and resource allocation. The storage system had to be very reliable to provide uninterrupted operations over an extended period of time. Failure of one or more hardware elements could not lead to the failure of the whole system. The system should support the addition of new equipment to scale its capacity and rates, as well as new types of equipment. The low cost of the system and its operation was also one of the major requirements. We evaluated storage systems which existed at that time. Most of the them were designed as backup systems. They could not group stored data according to experiment needs. They did not map to the specific needs of high energy physics data processing. They could not sustain a constant data flow at a very high rate of uninterrupted work over a substantial time period. The ratio cost/effectiveness was not satisfactory. Most of the systems were strongly coupled to the hardware, reducing flexibility in the selection of machines and robotic libraries. It was also difficult to get the code modified to meet user requirements.

Fermilab then decided to develop its own data storage system to satisfy requirements of high energy experiments and general storage needs. This system is highly reliable, scalable, flexible mass storage system. It uses inexpensive computers and has the capability to select between a wide range of robotic libraries and tape drives.The Enstore project started in July 1998 and the major components of the system were developed and put into test production in approximately 1 year.

The structure of the Enstore software is presented in Figure 1. It is designed using a client-server architecture and provides a generic interface for users. A configuration server keeps the system configuration information and provides it to the rest of the system. Configuration is described in the configuration file and can be easily modified and downloaded to the configuration server without interruption of the service. A volume clerk maintains the volume database and is responsible for declaration of new volumes, assignments of volumes, user quotas, and volume bookkeeping. A file clerk maintains the file database, assigns unique bit file IDs and keeps all necessary information about files written into Enstore. Multiple distributed library managers provide queuing, optimization, and distribution of user requests to assigned movers. Movers write / read user data to tapes. A mover can be assigned to more than one library manager. A media changer mounts / dismounts tapes in the tape drives at the request of a mover. Events are used in the Enstore system to inform its components about changes in the configuration, completed and ongoing transfers, states of the servers, etc. An event relay transmits these events to its subscribers. Alarm and log servers generate alarms and log messages from Enstore components correspondingly. An accounting server maintains an accounting database containing information about completed and failed transfers, and mounts. A drivestat server maintains a database with information about tape drives and their usage. An Inquisitor monitors the state of the Enstore components. PNFS [2] is a product developed at DESY. It implements a name space and externally looks like a set of Network File Systems.

All Enstore components communicate using IPC based on UDP. Great care has been taken to provide reliable communications under extreme load conditions. The user command, encp, retries in case of an internal Enstore error.

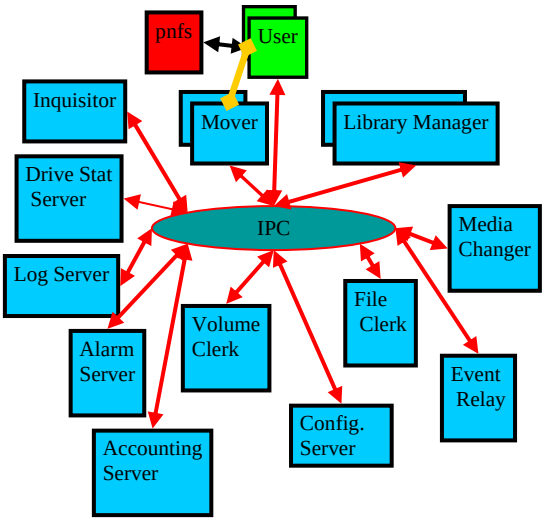

Figure 1: Enstore structure.

The number of user computers is not restricted, and Enstore components can be distributed over unlimited number of nodes, tape libraries and tape drives.

Enstore supports automated and manual storage libraries, which allows for a larger number of tapes than slots in the robotic storage. The user command interface program encp has a syntax similar to the Unix cp command with some additional options, allowing users to specify request processing parameters such as priority, and number of retries, whether to calculate a CRC, and to set a tape dismount time, etc.

Data stored in Enstore are grouped based on the storage group unique to each experiment, and file families inside of the storage group. The storage group is assigned by the storage system administrator while the file family is assigned by the user. Files in the same file family are written to the same set of tapes moderated by a file family width. The file family width controls the amount of simultaneous write transfers for a certain file family.

Enstore allows users to specify the priorities for data transfer. There are 2 kinds of priorities: regular and administrative or Data Acquisition (DAQ) Priority. The library manager will dispatch DAQ priority requests ahead of any regular priority requests. In this case the mover may dismount the currently mounted tape and mount the DAQ one. Priorities can be assigned to requests according to configuration parameters in the Enstore configuration file.

Another important feature of the system is its capability to specify the number of tape drives dedicated to an experiment (storage group). Experiments have separate budgets. Some of them may have their own tape drives installed in the general purpose robotic library. These drives are in the common pool but the experiment will preferentially be given access to the number of drives equivalent to its contribution. This amount can be specified in the configuration file and is used while processing the request queue.

Enstore has a powerful monitoring system that monitors a large set of hardware and software components:

- states of the Enstore servers
- amount of Enstore resources such as tape quotas, number of working movers
- user request queues
- plots of data movement, throughput, and tape mounts
- volume information
- generated alarms
- completed transfer
- computer uptime and accessibility
- amount of resources used (memory, CPU, disk space, etc).

The monitored information is available on the web in a set of static and dynamic web pages published on the Enstore web site at http://www-hppc.fnal.gov/enstore/. In more details Enstore monitoring is described in [3].

FNAL ENSTORE SYSEM CONFIGURATION AND STATUS

Currently the Fermilab Enstore Mass Storage System is represented by 3 independent systems:

D0en - for the D0 experiment
CDFen - for the CDF experiment
STKen - for the rest of the Fermilab user community

Altogether they have 1 ADIC and 5 STK robotic tape libraries with 27 STK 9940A, 42 STK 9940B, 4 STK 9840, 9 IBM LTO-1, 4 IBM LTO-2, 4 Mammoth-2, and 8 DLT tape drives. In the near future we plan to replace 9940A tape drives with tape drives supporting the denser data format on the same media. Data transfer rates between clients and mover nodes are equal to the rates of the tape drives. Currently the total amount of data stored in permanent storage is about 2 PB with an average daily transfer rate of more than 10 TB with maximal transfer of 20 TB a day with one day record of 24.9 TB.

To provide parallel access to data, experiments started to deploy DCache [1], a data caching and buffering system. DCache uses disks to store data. The interface between DCache and Enstore was designed to backup data stored in DCache to Enstore and retrieve it from tapes as needed. Several protocols were implemented to access data via DCache including secure and weak ftp (the last is used only for reading data), GridFtp and SRM. Currently The DCache systems contain more than 60 nodes with approximately 160 TB of disk space backed up by Enstore.

To write data on tapes Enstore usually uses a cpio wrapper. But this wrapper has a restriction on the size of a wrapped file of 8GB. To allow bigger file sizes we participated in the design of the CERN wrapper, which theoretically allows file sizes up to 9e16 bytes (86 Exabytes). In practice the file size in this case will be limited by capacity of the tape because Enstore does not

allow spanning files across multiple volumes. Enstore movers were modified to implement this wrapper.

As the amount of data in the robotic library grows and technology changes allow storing of more data on single media, the more important it becomes to support migration of the data. Migration is also needed to make copies of the data when the media ages (lots of tape mounts). In 2003 – 2004 more than 4000 9940A tapes were migrated to 9940B tapes. This project required a lot of attention from developers and administrators. It then was decided to implement an auto migration feature in the framework of the Enstore project. The auto migration is file based (the list of files is supplied for the migration process), and controls the file copying process not interfering with user requests. It makes the migration unnoticeable for users and files are available to users during the migration. The development has been mostly completed and now auto migration is in the testing stage. About 200 9840 tapes have already been migrated using this feature.

Additional Enstore functionality was supporting the ingest of DLT tapes arriving at Fermilab from Apache Point Observatory (Sloan Digital Sky Survey project). Enstore components were modified and a new client program was added to allow creation of metadata in Enstore while data were being read from the tape.

Fermilab is getting ready to become a tier-1 CMS data processing center. For this it needs to create and deploy data storage systems capable of effectively transferring, storing, and processing very large amounts of data. It is planned to install an additional robotic library and include it into the general Enstore system – STKen. Having many independent systems makes administration more difficult. Even more important is the desire to share resources across systems. The federation project addresses this challenge and is currently in the design stage. The implementation of this project should also help solve a problem with distribution of data and resource sharing between different equipment holders. This will help increase the robustness of the overall Fermilab Enstore infrastructure and its fault tolerance.

CONCLUSION

The Enstore data storage system was developed at Fermilab initially as a primary data store for Collider Run II experiments. It has proven to be very efficient and reliable and now is used by more than 30 experiments and groups. Recently, Vanderbilt University successfully deployed Enstore for their needs. At Fermilab, Enstore has been used for several years. It is attractive because it uses inexpensive hardware and can be accommodated to use any robotic storage library and tape drive. The flexible distributed configuration supports the addition of new components (such as mover nodes) without restarting of the system. Administration of the system is quite easy and is continually being improved. Modularity of the system allows addition of or modification of existing functionality. Enstore can also be used as a lower layer in the hierarchical data storage system. This was proven by combining Enstore with DCache. As long as the need for tape stored data exists, Enstore will be able to provide necessary functionality and capacity.

REFERENCES

[1] M. Ernst et al. DCache, a distributed storage data caching system, http://www-dcache.desy.de/chep2001/talk-4-005.pdf

[2] P. Fuhrmann, A Perfectly Normal File System, http://www-pnfs.desy.de/info.html

[3] J. Bakken et al. Monitoring a Petabyte Scale Storage System. Proceedings of CHEP-2004 Conference, Interlaken, September-October 2004.