FERMILAB-CONF-09-072-CD

A Multi-Purpose Computing Center : FNAL Proceedings of Data Preservation and Long-Term Analysis Workshop DESY 26-28 January 2009

Stephen Wolbers
Fermilab*

Abstract
The Fermilab Computing Center is described with a special emphasis given to the scientific computing systems and the data storage and archiving systems.

1 Introduction

The scope and focus of this paper is the Fermilab scientific computing facility. It does not cover, or does not cover very well, related issues such as data Grids, cloud computing and storage, commercial storage, data integrity, authorization, access rates, and novel storage technologies. These are all important considerations in discussing data centers and should be kept in mind when one explores issues related to computing centers and long-term data storage.

2 Description of Fermilab Computing Center

The Fermilab computing facility consists of three buildings that house 9 computing rooms, multiple uninterruptible power supplies (UPS), and one large generator. The major characteristics of the three buildings are the following:

2.1 Feynman Computing Center (FCC)

The Feynman Computing Center currently comprises the following:

- 3 computer rooms
- 19000 ft^2 (1765 m^2) computer room floor space
- UPS backup for all computing systems
- 1.5 MWatt backup diesel generator
- A measured uptime of 99.99% (planned & unplanned outages included)

FCC is the primary location for high-availability services and therefore the computers and networks required to provide those services. This includes part of the robotic storage system, compute servers, disk servers, web servers, email, backup systems, business systems, and networks (WAN and LAN). A picture of some of the systems that can be found in FCC is shown in Figure 1.

*Operated by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the United States Department of Energy

Fig. 1: Computing Equipment in FCC (left) and GCC (right)

2.2 Grid Computing Center (GCC)

A second major computing building for scientific computing is the Grid Computing Center. It has the following characteristics:

- 4 computer rooms
- 10388 ft^2 (965 m^2) computer room floor space
- ≈3 MW UPS
- Ability to connect generators to provide uninterrupted power to the computer rooms during scheduled power outages
- A measured uptime of 99.78% (planned & unplanned), 99.88% (unplanned only)

GCC is used to house large compute farms, a large fraction of robotic storage, part of the lattice QCD and other high performance computing systems, disk systems, and the necessary networking.

2.3 Lattice Computing Center (LCC) and the future of Fermilab scientific computing facilities

2.3.1 LCC

The Lattice Computing Center has the following characteristics:

- 2 computer rooms
- 2800 ft^2 (260 m^2) floor space
- No UPS or generator
- A measured uptime of 98.53% (planned & unplanned), 98.58 (unplanned only)

LCC is used to house the first generations of the portion of the US Lattice QCD project cluster located at Fermilab. The space is adequate for that purpose but is not backed by UPS or by generator. In addition, LCC serves as spare space, power and cooling for equipment if new computing room space is not available elsewhere.

2.3.2 Future facilities and upgrades

Upgrades to computing space, power and cooling are extremely important to the ability of Fermilab to support the science program. The upgrades being planned include a new high availability computing room in FCC, upgrades and additions to GCC, UPS and generator backup for GCC and other upgrades and improvements as they become necessary. These changes are driven by the lab's scientific program and are limited by the ability to plan and fund the upgrades. There is a clear and pressing need for facility upgrades and there is constant and dedicated effort to provide these facilities. Green computing [1] considerations are used as part of the planning process for purchases and for upgrades.

2.4 Data Storage and Services

Many services are provided by Fermilab Computing. Business computing systems for the laboratory represent one type of service. They are professionally managed, include off-site storage and backup, disaster recovery plans, and other characteristics of this type of computing. Core scientific and laboratory services include email, web services, file servers, CVS and other code repositories, Windows servers and other similar computing systems. The data and files associated with these facilities are backed up and stored in a TiBS [2] backup system with a one year retention. The total size of this backup is approximately 1 PetaByte.

Scientific computing facilities and services represent the main topic of this talk. The archived data volume as of late January 2009 is approximately 14 PetaByte, managed by the Enstore mass storage system. Databases are used by all of the systems and the data associated with the databases are sometimes managed and stored as part of the core scientific and laboratory services and in other cases are managed and stored as part of the scientific computing facilities.

2.5 User Communities

The user community of the Fermilab computing facilities is quite large and includes many groups of users indirectly affiliated with the laboratory. Nevertheless, it is useful to list the primary user groups as this is quite revealing and shows the breadth of the services and systems required. The major internal users include the Computing Division and the rest of the laboratory, all of whom rely on the central services such as networking, email, web services and business services to conduct their work.

The experiments that Fermilab either hosts or collaborates with are significant user communities. They include the Run 2 experiments CDF and D0; CMS, including the Tier 1 computing center and the LHC Physics Center; the neutrino program experiments; other HEP experiments engaged in R&D, construction and data analysis; and the astrophysics experiments Sloan Digital Sky Survey (SDSS), Dark Energy Survey (DES), Joint Dark Energy Mission (JDEM) and Cold Dark Matter Search (CDMS). Other scientific efforts also make use of the facilities. Some of them require significant computing resources. They include the US Lattice QCD project, accelerator simulation, cosmological computing simulations, ILC accelerator and detector simulations, and Project X proton accelerator proposed for Fermilab. Other user communities include newly proposed experiments such as Mu2e as well as groups engaged in large simulations and calculations.

3 Scientific Data Management

3.1 Components and architecture

Scientific data includes much more than just the "raw" data written out to mass storage from the experiments. In general experiments process raw data and the output of that processing is saved. This in some cases is larger than the original raw data. Subsets of data at any stage of processing chain are kept. Further processing and selections are performed, data is compressed (usually by keeping less information from each event) and the compressed data sets are also kept. The most common format in HEP for the final analysis datasets is an NTuple.

In addition to the event data from the detectors there is a great deal of information associated with the raw data that is required for proper analysis. This includes metadata, databases, calibrations and descriptions of the detector, beam and environmental conditions. Software used to read data, process it, and analyze it is also part of the required information.

Simulations are essential to analyzing experiments. The number of simulated events is quite large and the amount of information kept from each event can be substantial. The combination of these two factors leads to a need for large volumes of simulated data that has to be properly stored and managed. Of course the software that produces, reads, writes and analyzes the simulated data also has to be preserved and maintained.

The implementation of the basic architecture of HEP experiments includes the application code, the interfaces and middleware that allow and support applications, data movement, job queuing, etc. It also include grid and data handling systems, disk or other cache storage, mass storage, calibrations and databases, and all of the systems used to tie all of this together. Good examples can be found by looking at CDF, D0, Babar, CMS, or ATLAS offline production and analysis systems - all of which are fairly complicated systems.

All of the components of these systems must function properly for an experiment to properly collect, process and analyze data. When one contemplates long-term analysis there has to be a consideration of what subset or modifications to this large and elaborate system must be made.

3.2 Scientific Data Storage

At Fermilab the techniques and technology used to archive data has fundamentally changed over the past 20 years.

3.2.1 pre Collider Run 2

Prior to the availability of cost-effective robotic tape systems that could manage large data sets (for the time), the strategy pursued at Fermilab was to write maximum volumes of data to storage for use in offline analysis. This required minimal cost per byte and increased data storage capacity at the expense of lower reliability and increased complexity. But the physics needs of the experiments demanded large data samples. The only tape technologies that could provide the necessary storage at reasonable cost were commodity tapedrives and media, primarily 8mm and DLT. Commodity technology was difficult to manage and maintain and some fraction of the data was not readable. The consequences of the data loss was fairly small, given that most results were statistics limited and in many or most cases new generations of experiments allowed for

improved results.

The number of tape cartridges written was very large, over 500,000. The Computing Division (CD) was responsible for tape management, storage and access of all of this data. Experiments were responsible for producing any necessary duplicate tapes, derived datasets, and other detailed data management issues. CD provided many of the necessary software tools. Both the experiments and CD were responsible for the methods and software used to read and write the tapes, the file structure on tapes, long-term storage, and migration of data where appropriate. This data management strategy and implementation can be best described as ad-hoc but effective at the time.

3.2.2 Run 2 era

When planning for collider Run 2 it was clear that robotic tape systems and the media and tapedrives associated with them had become economical for data volumes anticipated in Run 2. Although hierarchical robotic mass storage existed at Fermilab earlier it was limited to small datasets, primarily for final physics analysis. The data storage plans for Run 2 computing included the robotics and associated tapedrives and tapes. All of the other experiments at Fermilab also were included in the planning and implementation of central robotic tape storage. Mass storage at Fermilab has grown dramatically and now stores and manages about 14.5 Petabytes of data (January, 2009).

Fermilab's commitment to Run 2 data storage includes long-term rapid access for production and analysis, low data loss, data management, data access and mass storage software. The timescale for the data storage and management includes the complete data-taking period (2001-2010+) and some period of time beyond the data-taking period for continued analysis of the data. This period has not been defined yet.

3.2.3 Sloan Digital Sky Survey

A slightly different and interesting situation exists for the data and information of the Sloan Digital Sky Survey (SDSS). SDSS is covered by a formal agreement with Astrophysical Research Consortium (ARC). The data recorded was copied and maintained in multiple locations. In addition to the raw and processed data used by the experimental collaborators a series of data releases available to anyone has been made over a period of 7 years. The final data release occurred last year [3]. Multiple mirrored copies are available. These are made available under a 5-year agreement with ARC. Fermilab and at least two other sites are hosts for the data. It will be interesting to see how the longer-term responsibility and funding for this data proceeds.

3.2.4 The LHC era

Though labeled as the LHC era in reality this period began a few years ago and includes all scientific computing associated with Fermilab. Fermilab is a Tier 1 and LHC Physics Center. There are certain responsibilities associated with these roles in terms of data access and archiving, many of them are spelled out in MOU's and other documents. And of course Run 2, neutrino, lattice, astrophysics, cosmology and many other scientific communities all continue to use and require mass storage and the related facilities.

All scientific data and file backups are managed robotically. A system that automatically migrates data to higher density media is in place. Theoretically, access to data can be maintained indefinitely. This simplifies the access to older data but does add some cost and operational complexity to the storage systems. A data policy is being developed to address issues such as risk analysis and data loss, geographical placement of data, and possible uniform frameworks for scientific data management.

3.3 Computing Center Responsibilities for Data Management

Computing centers are responsible for long-term custodianship of many types of information that reside on or pass through their systems. These responsibilities include and are implemented by the following:

- Agreements with experiments and other customers.
- Disaster planning and recovery.
- Risk analysis and agreements to implement plans to reduce risks where possible.
- Adherence to agency requirements and regulations.
- Implementation of techniques to minimize data loss and attempt recovery when loss occurs.
- Data duplication if necessary.
- Data migration to higher density media.

In terms of long-term access and other associated issues such as data formats, file structure and methods to access the data are part of the data center's role.

4 Open Issues

There are many open issues related to data preservation and long-term analysis. These include the lifetime of agreements and data preservation, software and access method support over long time periods, data format, metadata, file size, hardware and software migration, experiment requirements, laboratory requirements, HEP and public access issues, required effort, expertise and budget, new technologies, Grid and cloud computing, LHC experiments, and issues related to international laboratories.

Other issues that should be explored include who has access to data and to which part or parts of the data (and the necessary software), data access methods and data delivery rates, data integrity, data security, documentation (this includes user instruction), agreements and policies, and custodianship, including ideas for institutions and consortia such as libraries, SPIRES or INSPIRE, or Google to take that role.

References

[1] See http://en.wikipedia.org/wiki/Green_computing.

[2] See http://www.teradactyl.com/.

[3] See http://www.sdss.org/dr7/.