Status and Plans of FNAL Tier1

Ian Fisk

March 22, 2004

FNAL Tier1 Introduction

Fermilab is a Tier1 Facility for CMS

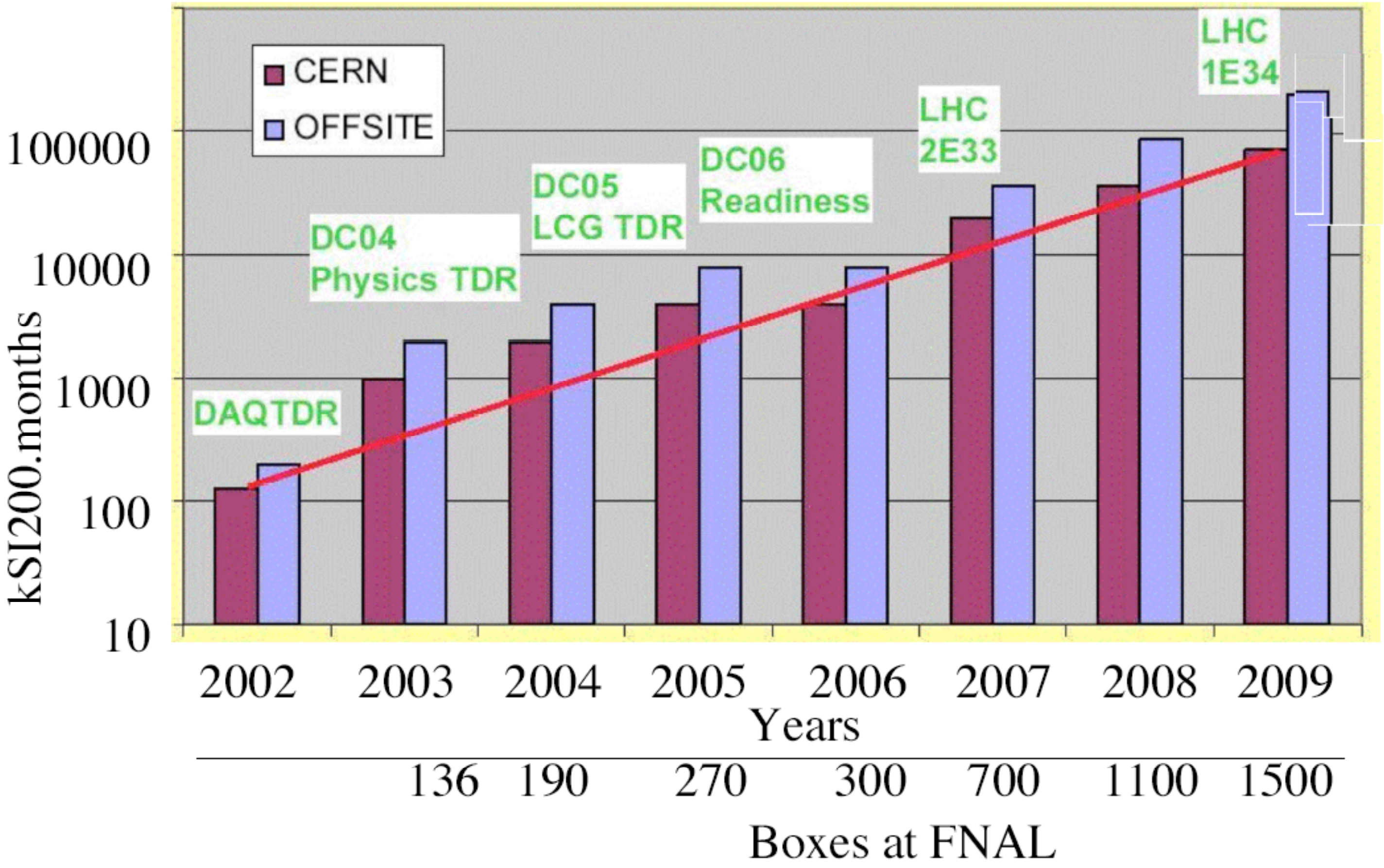
- Funding in the US is handled differently from other places in that the FNAL Tier1 is uniquely funded for CMS

Fermilab is also the host laboratory for US-CMS, which manages and supports the US-CMS Tier-2 Facilities and works closely with iVDGL

- Currently there are 3 prototype Tier-2 centers
  - Caltech
  - University of California, San Diego
  - University of Florida
  - We anticipate 5 production facilities eventually
- International Virtual Data Grid Laboratory was funded to build facilities in the US and facilitate grid development
  - iVDGL primary support for the Tier2 prototypes

Expected Ramp of the Facility

We are using the following CMS estimates for required computing at a function of time



FNAL Tier1 Represents about 10% of the total

- ➡ Roughly on schedule for 2003 and 2004

Tier-1 Hardware Status

136 Worker Nodes (Dual 1 U Xeon Servers and Dual 1U Athlon)

- ➡ 240 CPUs for Production (174 kSI2000)
- ➡ 32 CPUs for Analysis (26 kSI2000)
  - All systems purchased in 2003 are connected over gigabit

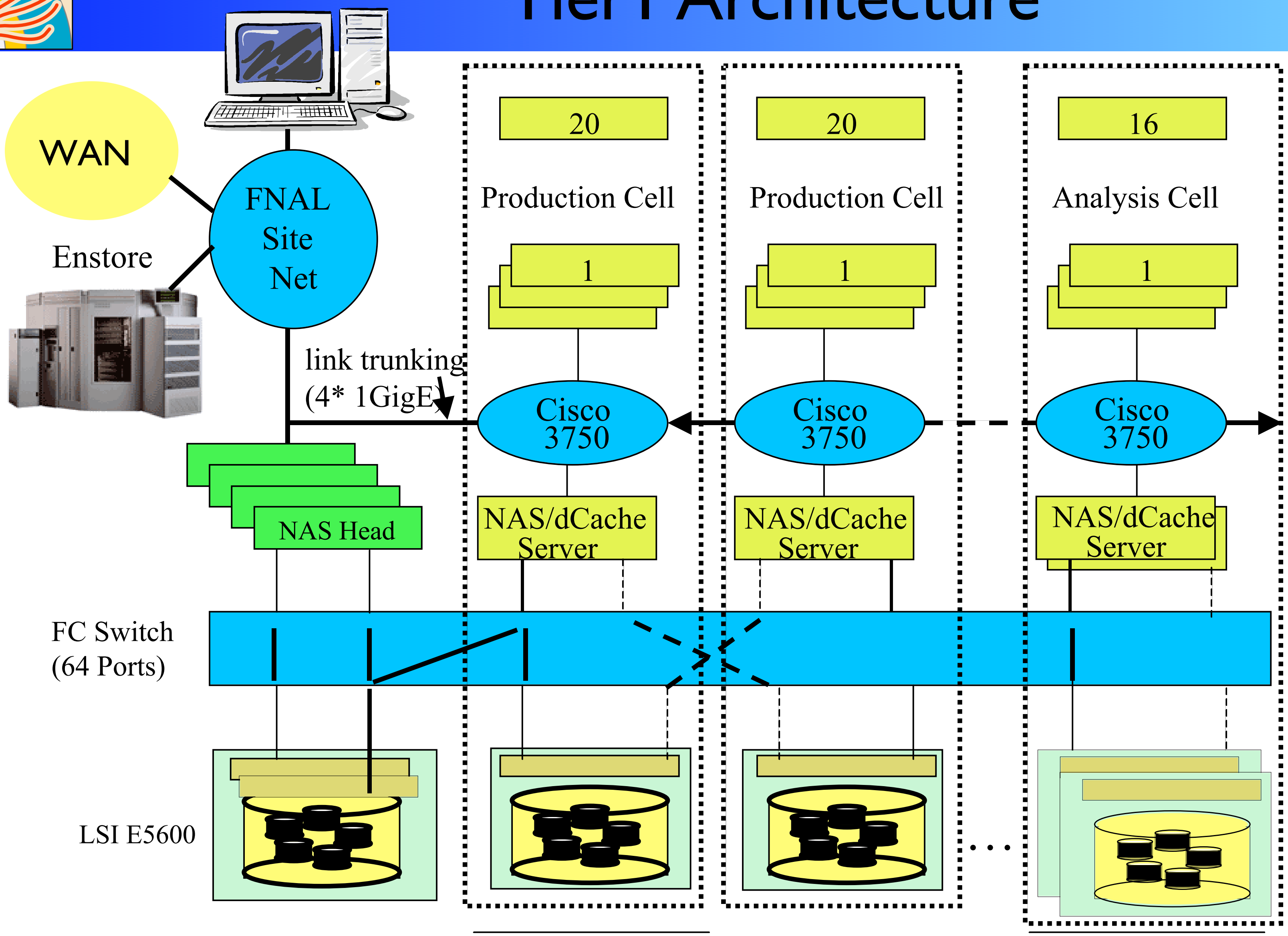
37 TB of Disk Storage

- ➡ 24TB in Production for Mass Storage Disk Cache
  - In 2003 we switched to SATA disks in external enclosures connected over fiber channel
    - Only marginally more expensive than 3ware based systems, and much easier to administrate.
- ➡ 5TB of User Analysis Space
  - Highly available, high performance, backed-up space
- ➡ 8TB Production Space

70TB of Mass Storage Space

- ➡ Limited by tape purchases and not silo space

Tier1 Architecture



System designed for high throughput data access and good connectivity

Tier-1 Upgrades for 2004

We plan to add 5 cells (100 Systems, 200 CPUs) to our existing setup in 2004

- 4 of these would be used for event production
- 1 for analysis

Raising the CPU total for processing to 400kSI2000

- Approximately where we had hoped to be in 2004

Increasing the mass storage cache by 20TB

Increase the space in the Fiber Channel RAID by 3TB

- Used primarily for user analysis space
- Also Production applications that require high performance disk storage

Increasing the number of tape drives and number of tapes

- 4 9940B Drives bringing the total to 8
- A total of 100TB of disk space

Future Fermilab Facility Upgrades

Networking

- ➨ Fermilab is physically close to Starlight in Chicago (60km)
  - From there the DOE supported link from Starlight to CERN provides ~10GB.
- ➨ The current Fermilab link is 622Mbit/s
  - Primarily network traffic is from Tevatron detectors off-site
- ➨ DOE has a long term strategy for a Metro ring with high performance and availability
- ➨ For a research network and improve access, Fermilab is arranging a fiber connection to StarLight
  - Contracts are in place
  - We hope to see light in the fiber before then end of the year.
  - It should provide a good short term and long term network solution for US-CMS

Fermilab Infrastructure Upgrades

Fermilab is having the same infrastructure issues in terms of power and cooling that all other large centers are experiencing

- ➨ Computers are small, they use a lot of power, and they generate a lot of heat.

The primary Fermilab computing building is essentially at capacity

- ➨ Places stress on VOs as we argue for available resources

Fermilab is in the process of reconditioning an experimental hall as a computing facility

- ➨ First occupancy expect by the end of summer
  - Experiment facilities have good power
  - Cooling infrastructure is being upgraded

Power crunch leads to examining usage

- ➨ Looked at expensive blade systems
- ➨ Looked at alternative processors (Opteron)

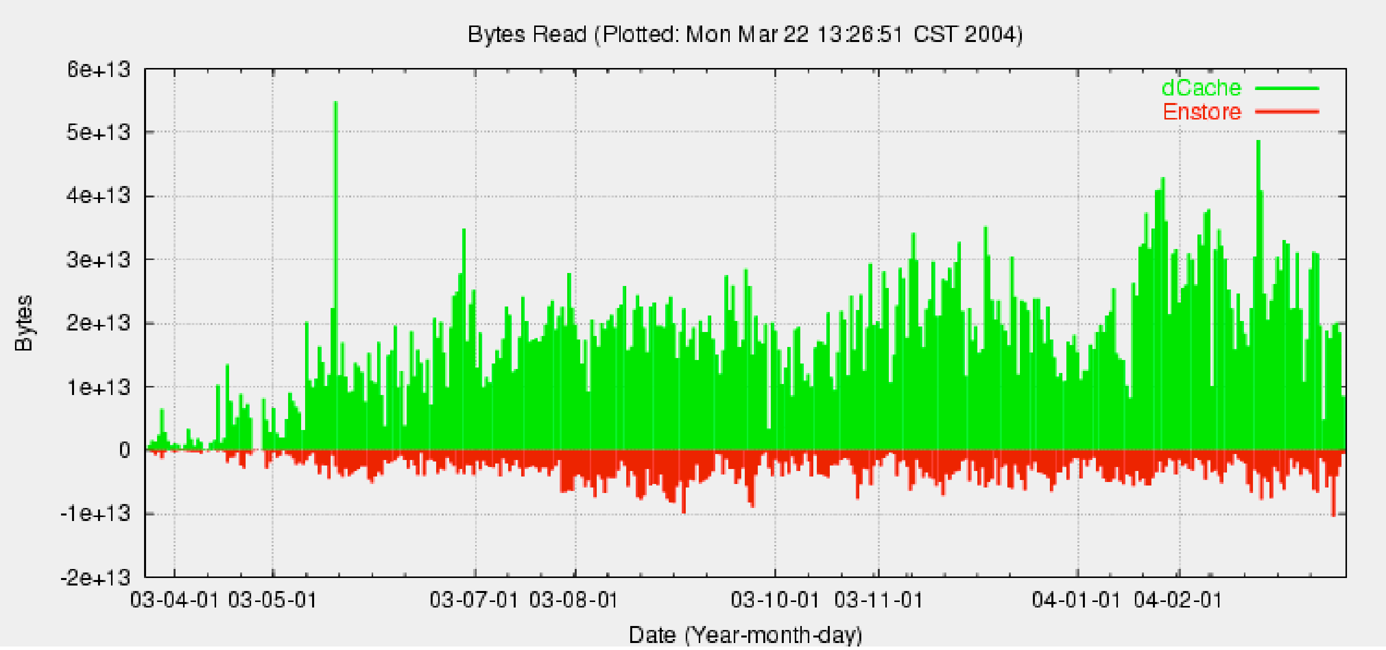
Storage Model

US-CMS Expects to use the Enstore Storage System with a dCache Disk Caching System

- dCache is a co-developed project between DESY and FNAL.

We do not expect to have to keep all real and simulated data on disk

- Below is a plot of the CDF data analysis through dCache who have 15% of there total data set of disk cache
  - Total data in tape is currently 0.75pB



Farm Configuration and Management

At the beginning of 2003, the US-CMS began deploying the ROCKS cluster management software at the Tier-1to help improve the configuration control over the hardware facilities

- ➡ ROCKS was developed by the National Partnership for Academic Computing Infrastructure (NPACI)
  - ● Allows system configurations to be specified, stored, and reproduced.
  - ● Appliances can be specified for all facilities components
  - ● Entire cluster can be upgraded in about 20 minutes ensuring that all the installed packages are consistent

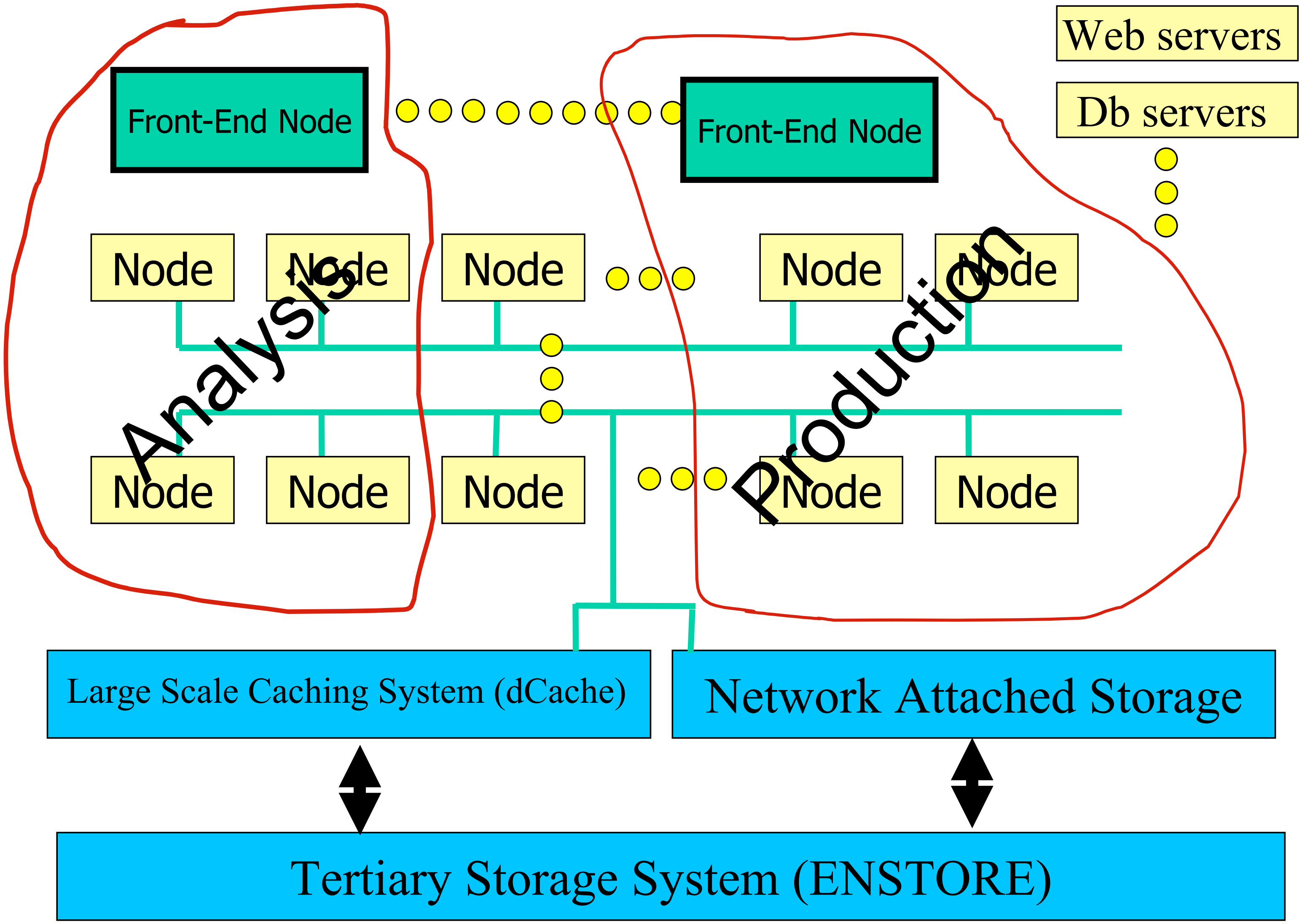
The system works quite nicely and was customizable to support the Fermilab environment

- ➡ Improved and extended over the course of the year

Installed at all the Tier-2 installations

- ➡ Allows cross development and improves support

Dynamic Partitioning of Resources



Plans for LAN-WAN Networking

Currently Fermilab connects all systems to the outside network.

- ➡ This is not the case at all the Tier-2 centers
  - Security and implementation model can be site specific

We see that functionality of a firewall is attractive

- ➡ Traditional firewalls are not compatible with high performance advanced networks
  - We see that development is needed
- ➡ US-CMS envisions enabling fine grained authorization to control network usage, but to do so on the system or router level.

Interfaces and services need have enough functionality and flexibility to function in a variety of configurations

- ➡ There are often good reasons why sites are configured the way they are

Development of Facility Interfaces

The Tier-1 Center at Fermilab has an active program development on facility related interfaces for distributed computing.

Currently the active development projects are

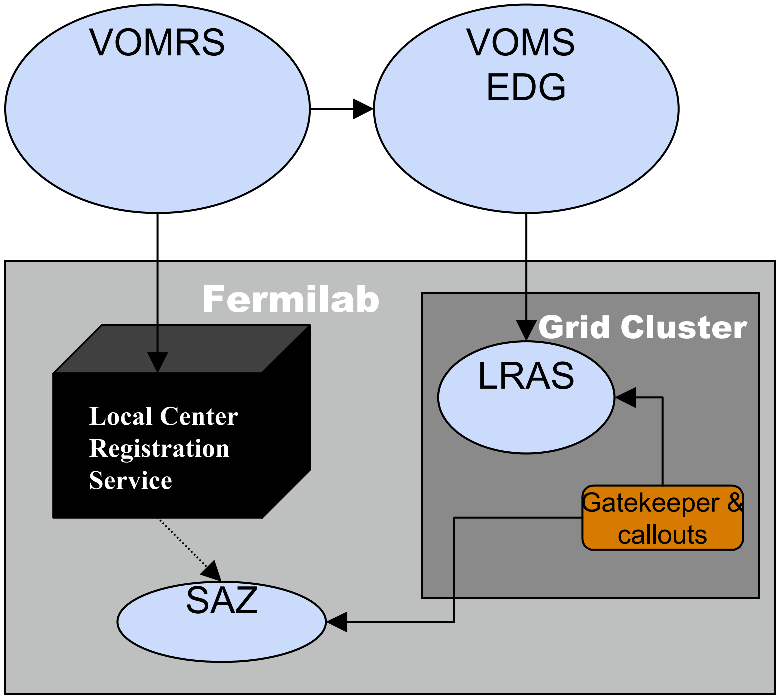
- ➨ User registration, authentication and authorization
  - Enabling fine grained authorization
  - Satisfying site security
- ➨ Grid Interface to Mass Storage
  - Allowing managed access to storage resources
- ➨ Enabling LCG interfaces to access US resources
  - Satisfying site policy
  - Enabling LCG connected users to take advantage of US grid resources

US-CMS VOX Project

US-CMS Started a project to improve user registration

The goals of the project were to

- understand and model the registration workflow
- provide VO registration mechanism
- negotiate and monitor member authorization to grid resources
- To facilitate the remote participation of physicists in effective and timely analysis of data



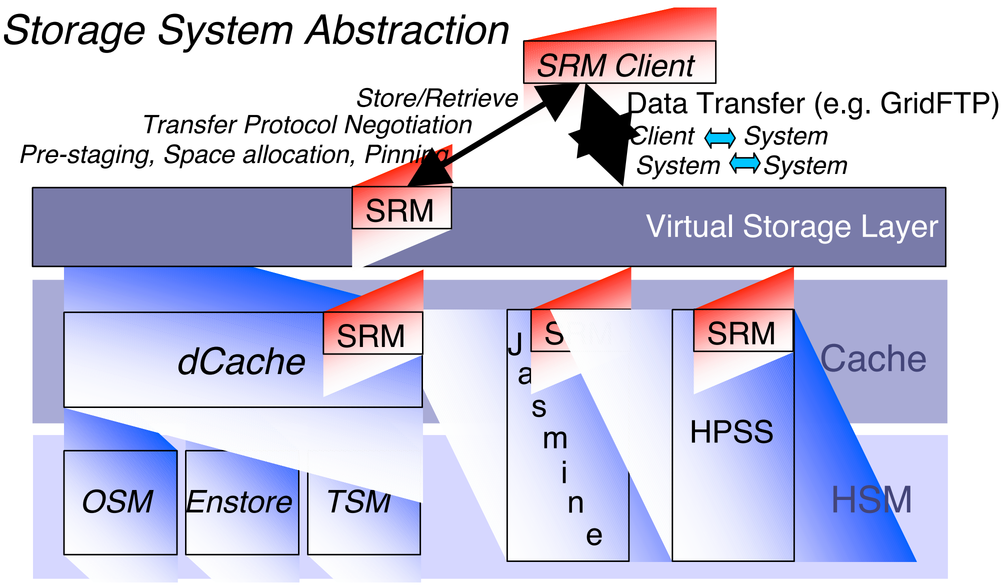
Common Storage Element Development

While compute resource scheduling is fairly advanced, mechanisms for storage resource management is lacking. Issues include

- Shared storage resource allocation & scheduling
- Staging management - Files are typically archived on a mass storage system (MSS)
- Wide area networks – minimize transfers
- File replication and caching

The Storage Element Joint Development is based on SRM.

- ➡ SRM provides a uniform interface to diverse and distributed physical storage devices (MSS, Disks, Data Caching services, etc.)



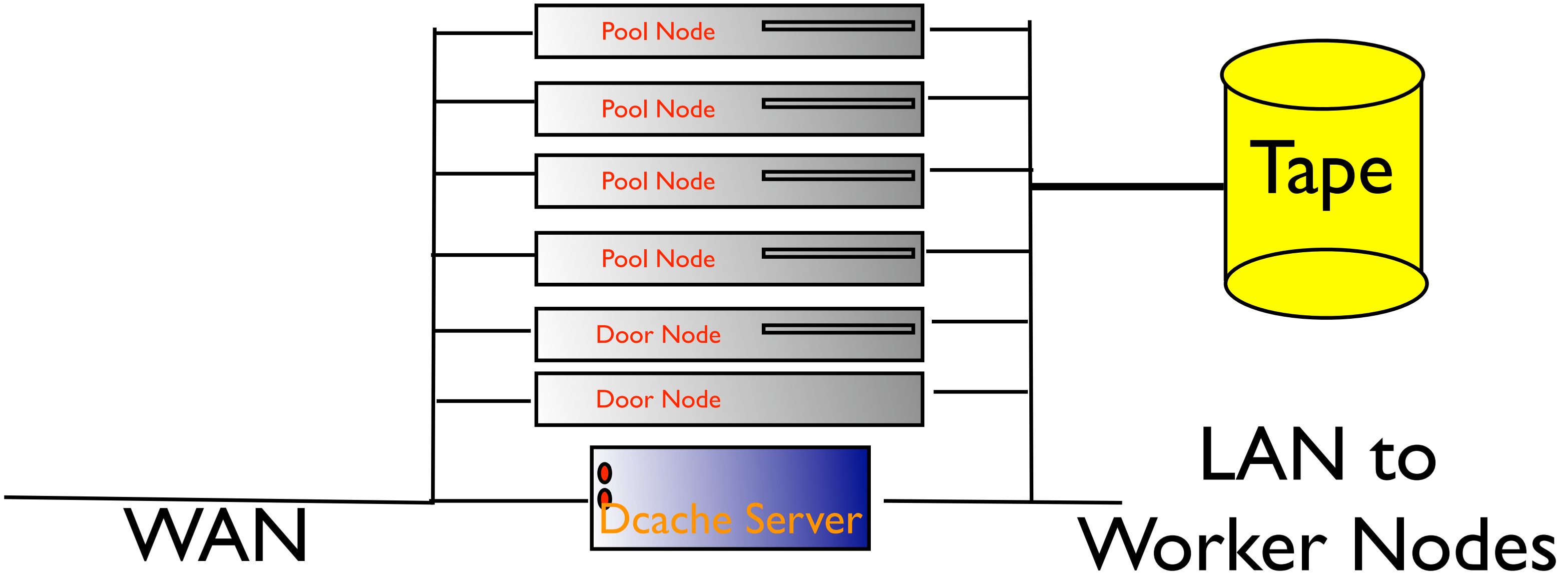
SRM + dCache Standard Installation

While the functionality can be spread across several physical systems, the dCache Server

- ➡ Handles the name space, the database, SRM, establishing transactions with the Pool nodes

The Pool nodes are the physical storage

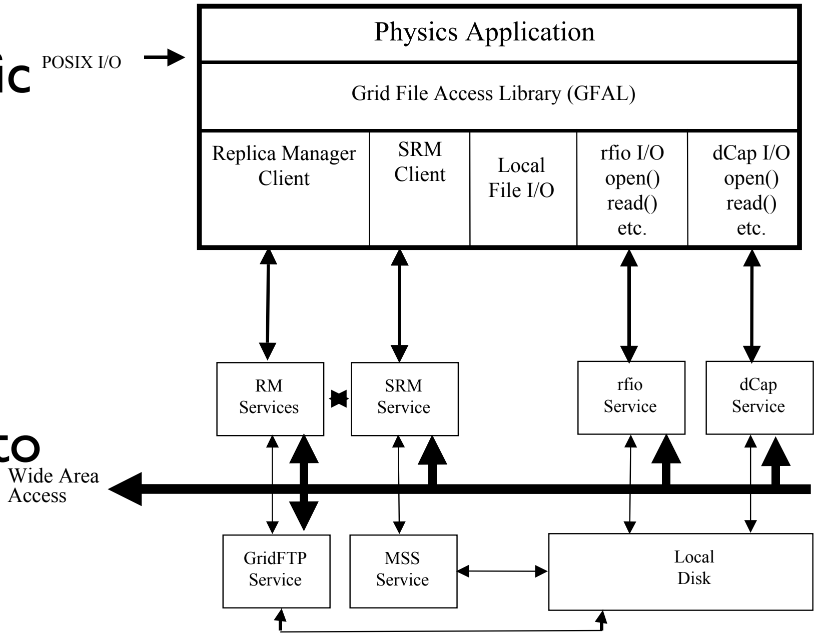
- ➡ Pool nodes can either be dedicated (RAID arrays or other physical storage devices) or they can also serve as worker nodes
  - Efficient way to take advantage of space on the worker nodes



Application View

Application interacts with a library

- Hides the storage system specific access method and provides higher level services
  - Registering files in replication services
  - Transparently arranging access to remote files



Data Streaming Project

US-LHC has formed a project with CERN IT to work on improving data streaming from CERN to Tier1 centers

- LHC experiments are interested in archiving a copy of the raw data across the Tier1 centers
- Interesting to evaluate extending the physics reach of the detector by have dedicated analysis streams that cannot be reconstructed at CERN due to limited resources, but might be reconstructed at the Tier1s

We are currently in procurement phase of the project.

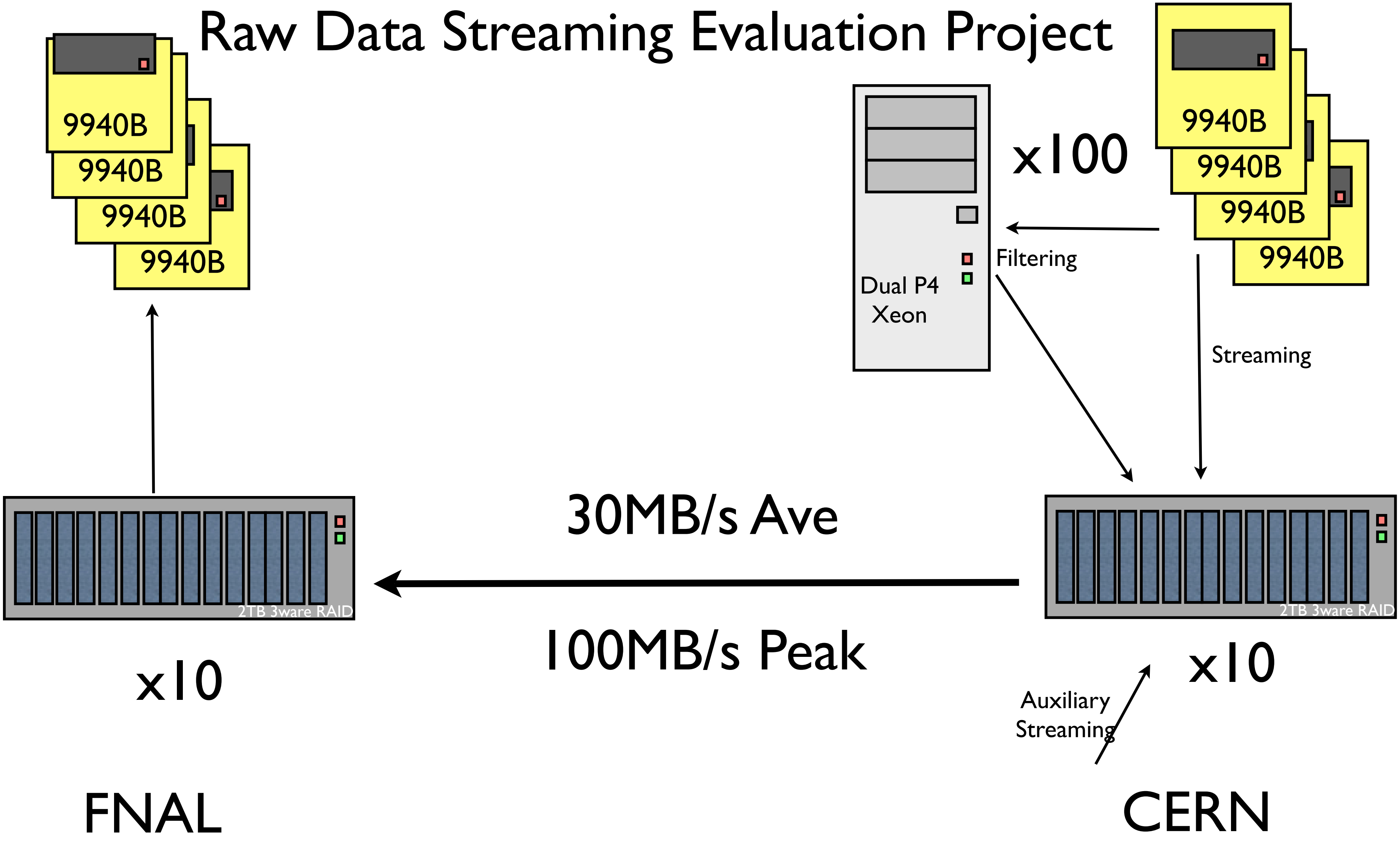
- Equipment at CERN for event selection and deep buffering
- Equipment at FNAL for input buffers and tape throughput

We plan in the spring to demonstrate real time data streaming from CERN

- Start with end-to-end SRM transfers using improved network
- More advanced selection and transfer techniques that exercise the CMS software should follow

Data Streaming Project

Raw Data Streaming Evaluation Project



9940B

9940B

9940B

9940B

Dual P4 Xeon

x100

Filtering

Streaming

30MB/s Ave

100MB/s Peak

2TB 3ware RAID

2TB 3ware RAID

x10

x10

Auxiliary Streaming

FNAL

CERN

LCG Interoperability

There are several methods available to for the Tier-1 facility to interoperate with multiple grid installations

- ➨ Partition the cluster between the projects
  - Currently used at FNAL due to time pressure, but an inflexible and inefficient method
- ➨ Deploy multiple sets of interfaces to the same physical resources
  - Requires decomposing the interfaces to determine how to integrate into the existing structure most efficiently.
    - Often requires development on the interfaces to make them compliant with the site security policies and configuration techniques.
- ➨ Define and deploy compatible interfaces between multiple grid projects
  - A flexible and efficient method from the facility standpoint, but requires cooperation and coordination between grid installations

US-CMS is working on the final two methods. The middle is seen as a viable short term solution. The final is seen as the most desirable.

Short Term Plans

US-CMS has had to operate our LCG-2 installation under an exemption from site security policies

- ➡ Working to improve the authorization of incoming users

The LCG-2 installation has up to now been small and separated from the rest of the CMS farm

- ➡ Working to allow incoming jobs to access the larger number of batch nodes
  - Adding the required worker node packages to our Rocks configuration
  - Enabling access to the batch queues form the LCG interfaces

US-CMS has a vested interest in allowing the facility to interoperate with multiple grid infrastructures.

- ➡ There are a significant number of non-LCG and non-LHC computing resources that are available to CMS in the US

Example of Grid2003

Grid2003

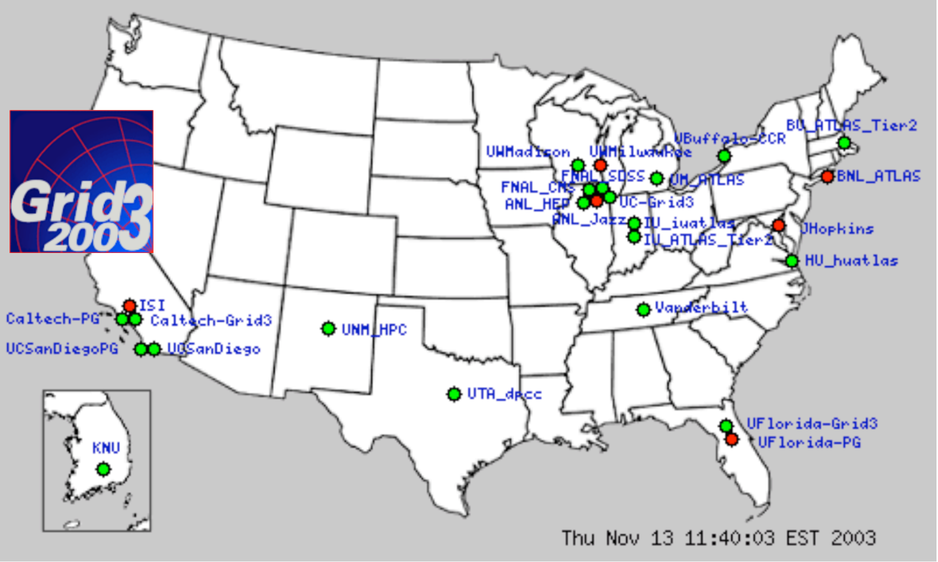
A multi-experiment Grid

-Collaboration of US-ATLAS, US-CMS, LIGO, SDSS,BTeV, and US-Grid Projects

-Aimed for simple installation base on Virtual Data Toolkit: site installation now down to 5 manageable steps

-The model was to deploy fairly simple low level services and aim for a robust environment (processing, services, data transfer services, imformation providers, and VO and authentication management)

Sites in Grid2003



Installation reached 28 sites and 2800 CPUs by SC2003

– Running fairly stably since. US-CMS has performed 13million GEANT4 full detector simulations, and counting, on Grid3 infrastructure since Nov. 2003. This represents about 100 processor years of computing. Participants in Grid2003 have seen a significant contribution of opportunistic computing resources.

Progress Toward Open Science Grid

The US will continue to develop a national grid infrastructure called the Open Science Grid

- It is currently a roadmap for the U.S. to build a national grid infrastructure for science, proposing a program of work to federate the U.S. grid resources into a scalable, engineered and managed grid, the Open Science Grid.
- This new Peta-scale computational service will be built as an open national infrastructure, optimizing shared use of resources for diverse collaborative research. The Open Science Grid will serve as a backbone to merge grid computing efforts of allied experiments in particle and nuclear physics, and can be extended to other scientific communities.
- The road map has been formulated by US-LHC Experiment Projects, Regional Centers, Universities, and Grid Projects.

It is important the US Tier-1 and Tier-2 centers operate efficiently with our LCG and OSG colleagues

Outlook

The FNAL Tier1 Center is on schedule for the start of the experiment

- Hardware ramp on schedule
  - Fermilab is trying to anticipate facility requirements
    - Power, cooling, and networking
- Network upgrades are on schedule

Lots of interesting development to do on interfacing the facility and interoperating

- Authentication and authorization
- Storage Interfaces
  - Data Transfer