

Advanced Computing Strategy
Presentation to URA Visiting
Committee
May 8, 2006

Vicky White
Head, Computing Division

“Computing” everywhere

- We do a great deal of “computing” (including DAQ and general IT services and scientific computing)
 - and all of our programs and the lab rely on it
 - we have 4 Petabytes of scientific data in our storage systems
 - we have > 3000 dual-processor CPUs in our Farms and analysis clusters
 - we have more than 10,000 addresses on the Fermilab network
- The success of our current and future programs relies not only on
 - An adequate amount of computation, storage, networking, databases and reliable IT services
- But on a continuous and evolving Strategy for Advanced Computing

Advanced Computing Strategy

1. Provide the facility, networking, information management and computer security infrastructure to build on and evolve into the ILC era.
2. Collaborate in worldwide efforts and distributed computing solutions such as Grids
3. Build innovative and specialized computing solutions where we need to for today's scientific programs and for LHC turn-on
4. Do R&D on computing solutions for future experiments and the ILC

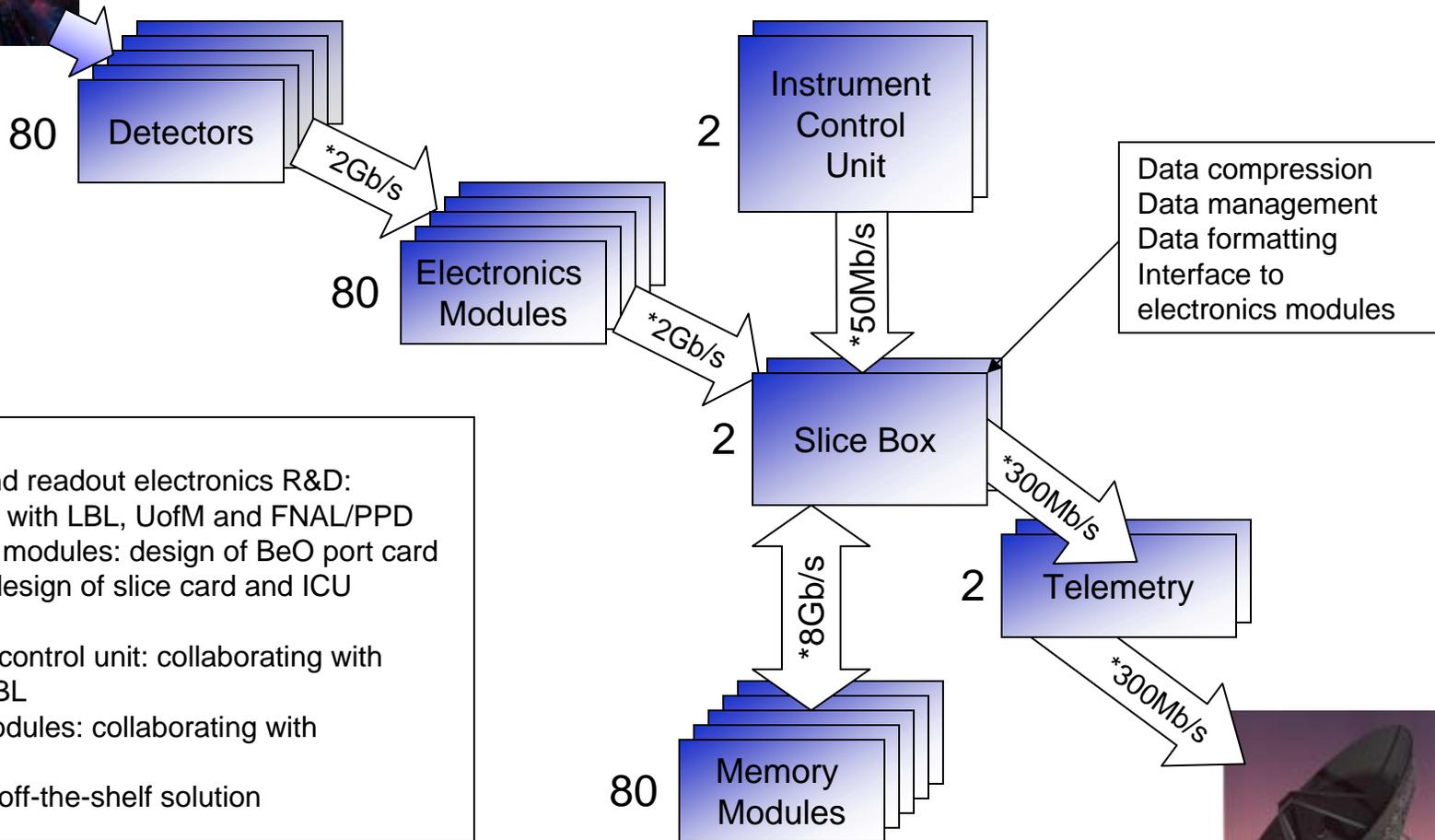
Highlights from past year (in reverse order)

4. Do R&D on computing solutions for future experiments and the ILC
 - SNAP and NOVA DAQ R&D
 - ILC Accelerator modeling -> separate talk
 - ILC Accelerator control systems
3. Build innovative and specialized computing solutions where we need to for today's scientific programs and for LHC turn-on
 - Lattice QCD Facilities
 - LambdaStation and Wide Area Networking R&D
 - FermiGrid and Petabyte Storage and disk cache systems

Highlights from past year (in reverse order)

2. Collaborate in worldwide efforts and distributed computing solutions such as Grids
 - Open Science Grid, CMS, WLCG
 - Patriot and PhyStat
 - Storage Resource Manager and Grid Security issues Build innovative and specialized computing solutions where we need to for today's scientific programs and for LHC turn-on
1. Provide the facility, networking, information management and computer security infrastructure to build on and evolve into the ILC era.
 - New Computing Centers (space, power, cooling)
 - Wide Area Networking
 - Computer Security Infrastructure for the lab and for Open Science
 - Information systems

SNAP DAQ – Overview



* Peak data rates displayed. Data comes in bursts with a 10% duty cycle.



Some DAQ Challenges

- Each exposure lasts 300s followed by 30s of readout
 - Due to noise, no readout can happen during exposures
 - The image from each detector has to be collected, compressed, framed, and stored into the memory bank in 30s
 - Our current design can perform these tasks in 8s!



- Data is sent to earth once a day ~2h, but confirmation from the ground comes next day. Hence, memory must store at least 2 days worth of data
 - Compression is paramount to reduce memory size → \$ and weight
 - Even with a 2.5:1 compression ratio, memory bank has 2.6Tb of FLASH and 21Gb of DRAM

NOvA DAQ Software

Challenge 1: *Buffer 3.5 GB/s, in 10 Byte chunks, for multiple seconds, efficiently search retroactively for all data in a specified 30 Osec window, and correlate and format selected data for analysis.*

R&D Work: *Investigate a parallel buffering system receiving data from a parallel data paths. All data for a selected time window of order msec routed to one buffer and given coarse time stamp. Store data in large predefined buffers with array of reference pointers and coarse timestamps. Search array of coarse timestamps first and continue with fine grain.*

Challenge 2: *Configure, control and monitor a system with 216 buffer nodes, 324 concentrator embedded processor boards, and 20088 front end boards in a 24x7 environment.*

R&D Work: *Investigate a publish and subscribe messaging system for control, configuration and error handling using subsystem master servers to interface run control with the subsystems.*

ILC R&D in the Computing Division

- Computing Division (CD) ILC effort is mainly focused on the accelerator
 - Acceleration Simulations
 - Controls (and Instrumentation)
- Accelerator Simulations (See talk by P. Spentzouris)
 - Using software developed by the SciDAC collaboration, Run II simulation and ILC Collaboration efforts;
 - Studies are done in collaboration with AD/TD;
 - Current areas of focus for CD:
 - Main Linac studies
 - Damping Ring
 - Crab Cavity (part of Beam Delivery System)

ILC Controls R&D

- CD is collaborating on design of the ILC control system
 - Global effort with DESY, KEK, US institutions
 - Prototyping and testing at ILC test facilities
- R&D planning in close collaboration with other US groups (FNAL/ANL/SLAC/LBNL/ORNL/UPenn)
 - Timing and RF distribution
 - High availability and reliable hardware and software
 - Frameworks for controls software
 - Low Level RF simulations and controls
 - Feedback systems (in the future)



Remote Operations

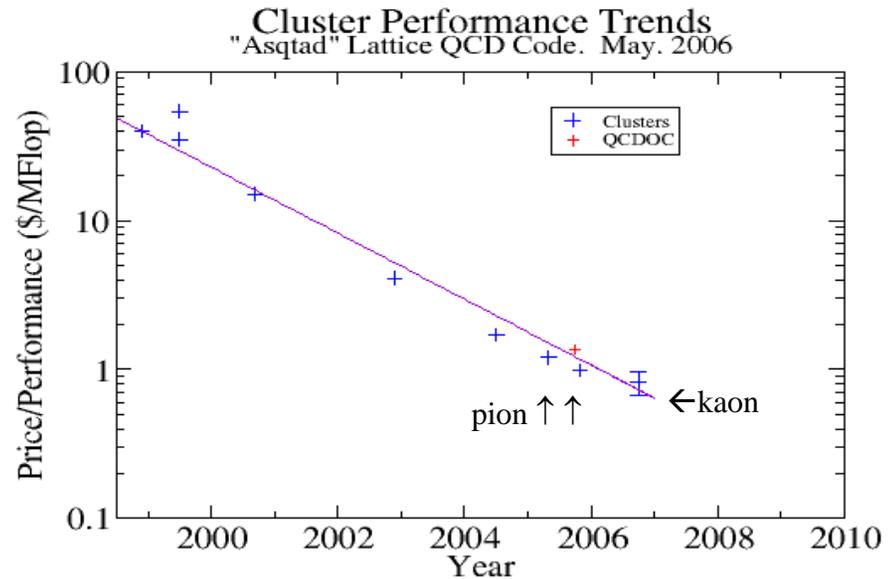
- Remote operations; tools to facilitate global collaboration
 - ILC test facilities will have worldwide participation
 - Collaborate with LHC@FNAL project
 - Forum to investigate/test new collaboration tools



Lattice QCD Computing

- USQCD – Distributed Lattice QCD computing facility at BNL, Jlab and Fermilab
- Fermilab is providing overall project management and strong technical and scientific leadership
- Yearly FY05-FY09 - \$2.0M from HEP and \$0.5M from NP as an OMB300 investment
- \$2.5M/yr in SciDAC funding (hopefully continuing)
- Host lab(s) support

Custom chip QCDOC at BNL
Commodity clusters at Jlab and Fermilab



Clusters are the most effective lattice QCD computers today

Fermilab Lattice QCD clusters

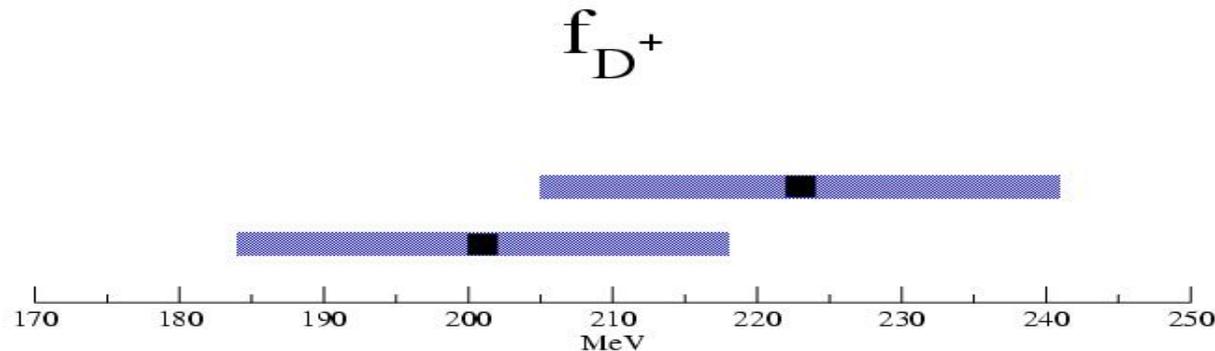
- Run tightly coupled message passing (MPI) parallel programs.
- High bandwidth low latency infiniband networks.
- “Pion”: 520 nodes, Pentium 640. Production since May 05. Rated 2.2TF/s (TOP500).
- “Kaon”: 500+ nodes dual core Opteron. Delivery in Aug 06. Production in Fall of 06. Expect 2.2x more powerful.

These are not computers you can buy on e-Bay !
State of the art innovative integration of commodity components, along with specialized communications and QCD software

Some spectacular LQCD results

- B_c Mass prediction last year and
- Decay constant f_{D^+} prediction

"It became clear that both groups [CLEO-c and Fermilab Lattice & MILC Collaborations] could have substantial results just in time for the Lepton-Photon Symposium in Uppsala at the end of June. Since both communities felt that it was very important for the LQCD result to be a **real prediction**, they agreed to embargo both of their results until the conference... The **two results agree well within the errors of about 8%** for each." CERN Courier **45**, 6 (2005).



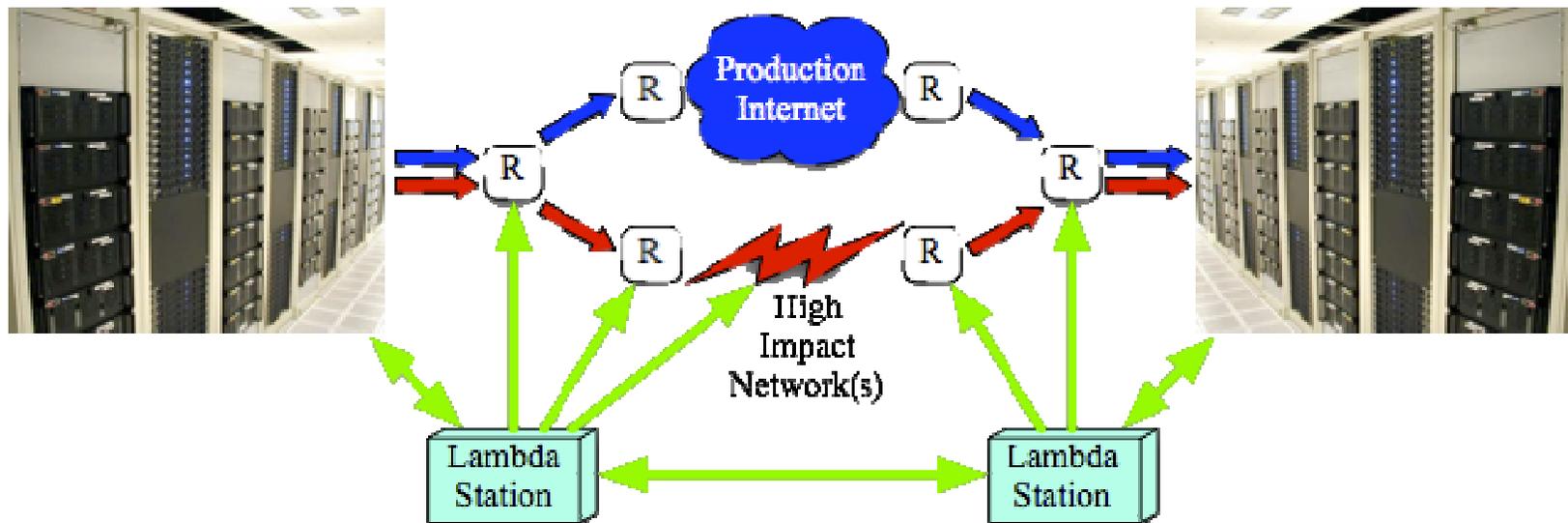
C. Aubin, *et al.*, Phys. Rev. Lett. **95** (2005) 122002

Advanced Networking

- Driven by Run-II and LHC needs
 - Also laying foundations for ILC era
- ESnet Chicago Metropolitan Area (MAN) upgrade will provide 4×10 Gb/s production capacity.
- FNAL will keep/regain $2 \times 10 + 2 \times 1$ Gb/s R&D capacity (now carrying some overflow production traffic).
- Fermi Light Paths: Lab connectivity infrastructure at StarLight will be extended to “Layer 0” – photonic switching to global lambdas.

Lambda Station

- DOE- Advanced Scientific Computing Research -funded project.
- Provides Web Services (SOAP) interface for applications to request site and Wide Area Networking (WAN) routing for special handling of high-volume data.
- Client calls integrated with real Storage systems and disk cache systems – testing with select USCMS Tier-2 sites.





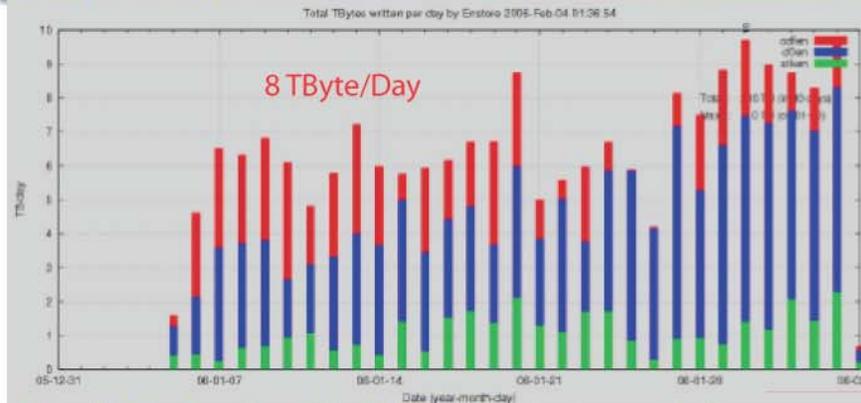
Tapes Are Really A Good Thing!



Tape Writing at FNAL

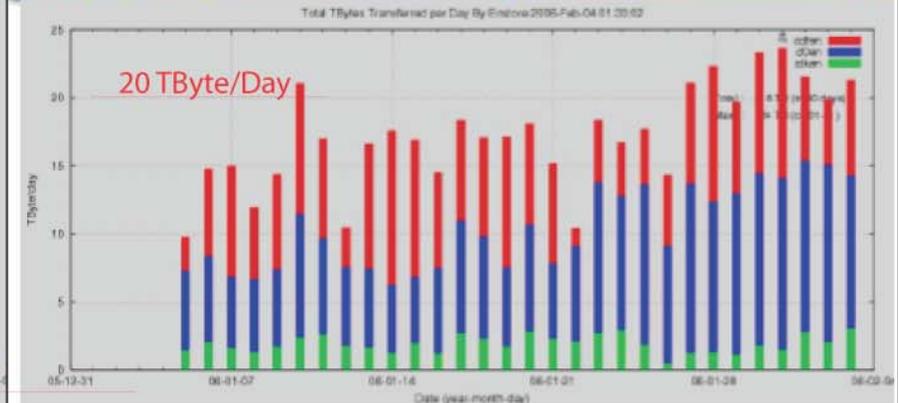


Tape Reading at FNAL



Fermilab is already writing 9 TB of data to tape each day.

- 3.6 PB/year is an average of 10 TB/day
- Experience exists to do this now



Delivering more than 15 TB/day of data from tape to users is common at Fermilab.

Jon Bakken

DOE/NSF-2006 Review

February 7, 2006

Jon Bakken

DOE/NSF-2006 Review

February 7, 2006

46

◆ CMS requires "library style" data storage at Tier-1 centers!

★ a lot of sequential access to data, in particular to MC data!

◆ Tape libraries = cheap storage w/ very performant access

★ Fermilab Tier-1 estimated costs to deploy the 2008 resources:

★ for disk: **\$1400/TB** (2.0 PB), for tape: **\$400/TB** (4.7 PB)

★ initial tape library costs ~\$700k, incremental media costs \$200/TB

LATBauerdick/Fermilab

ISGC 2006 — CMS Computing

May 3, 2006

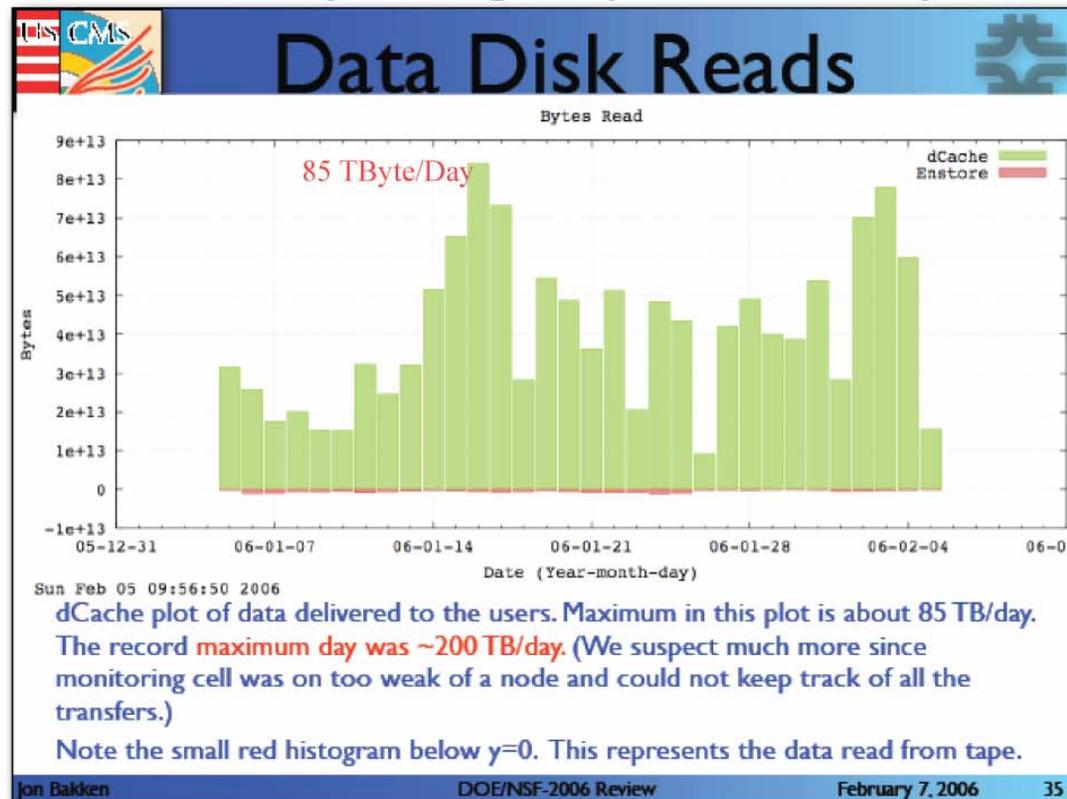
9



High Throughputs With Tapes + Disks



- ◆ Combination of high-throughput tape libraries and disk caches
- ★ amazing performance is feasible, see Suen Hou's talk w/ CDF numbers!
- ★ to achieve this does require to gain quite a bit of experience



➔ but PBs of disks w/o local tape potential commissioning&maintenance nightmare!

FermiGrid –our strategy

Grid is about collaboration and sharing

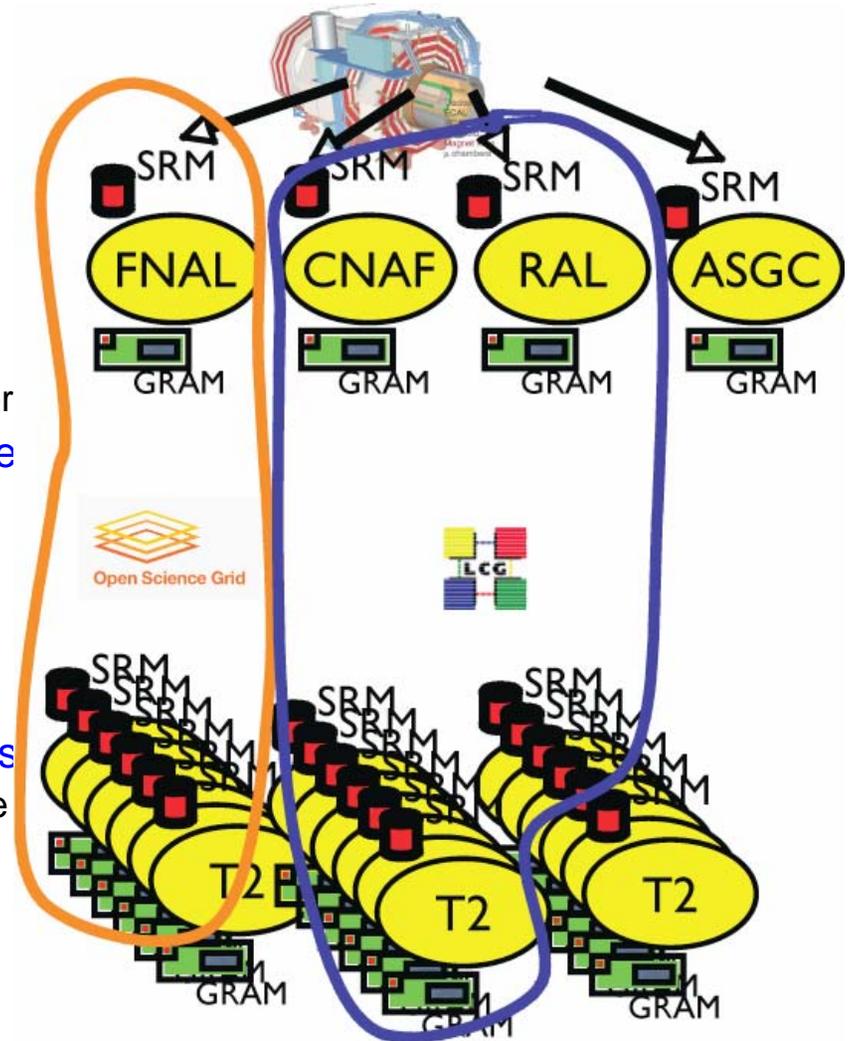
FermiGrid is a meta-facility with 4 main elements:

1. Interoperability and ability to share resources between all the different scientific program stakeholders at Fermilab (e.g. Dzero used CMS resources in its data reprocessing earlier this year)
2. Common Fermilab Site Grid Services
 - Virtual Organization hosting (VOMS, VOMRS),
 - Site-wide Globus GRAM gateway,
 - Site AuthoriZation, MyProxy, GUMS.
3. Interfaces to the Open Science Grid
4. Grid interfaces to Storage systems.

Grid Computing

•CMS is assembling a distributed computing model where the experiment is supported by interoperable grids and the layers are connected by grid service interfaces

- Majority of resources are located off the experiment site
 - 40% at Tier-1 centers and 40% at Tier-2 center
- Infrastructure supported by the Open Science Grid and the LHC Computing Grid
- Jobs are dispatched through consistent interfaces through gatekeepers to the batch farms
 - Currently Globus GRAM V2
- Data is transferred between storage elements
 - Storage interface is published with the Storage Resource Manager
- The facilities are described with a consistent information system
 - Based on the GLUE schema

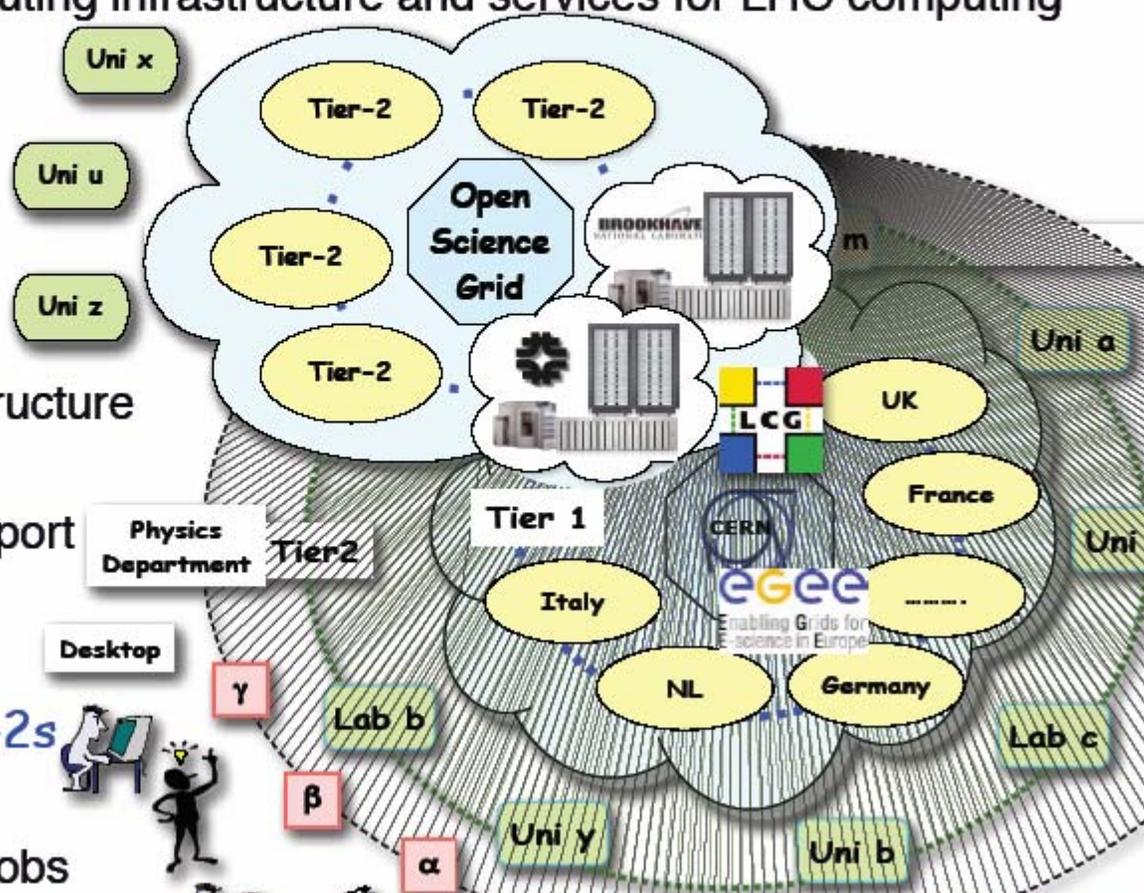




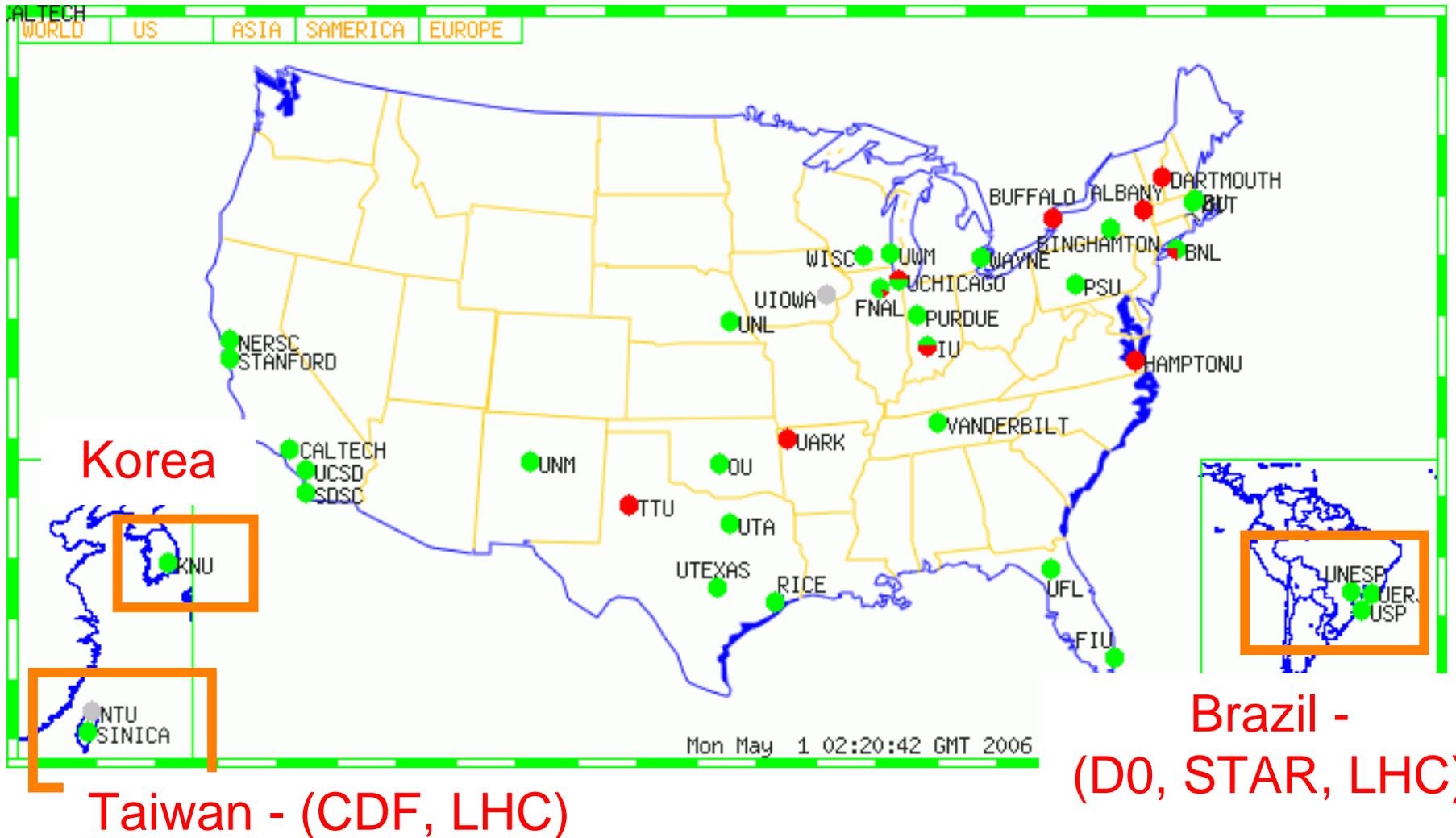
Worldwide LHC Computing Grid WLCG



- ◆ Oct 2005 founded collaboration of regional centers: WLCG
 - ★ to provide the computing infrastructure and services for LHC computing
- ◆ WLCG-MoU
 - ★ resource pledges
- ◆ WLCG Service
 - ★ storage services
 - ★ “batch slots”
 - ★ data transfer infrastructure
 - ★ middleware
 - ★ Grid operations, support
 - ★ Service Challenges
 - ★ ...
- ◆ 7 Tier-1s, ~14 Tier-2s
 - ★ ready to accept CMS datasets and jobs



Open Science Grid: More than a US Grid



OSG 1 day last week

Routed from Local
UWisconsin Campus
Grid

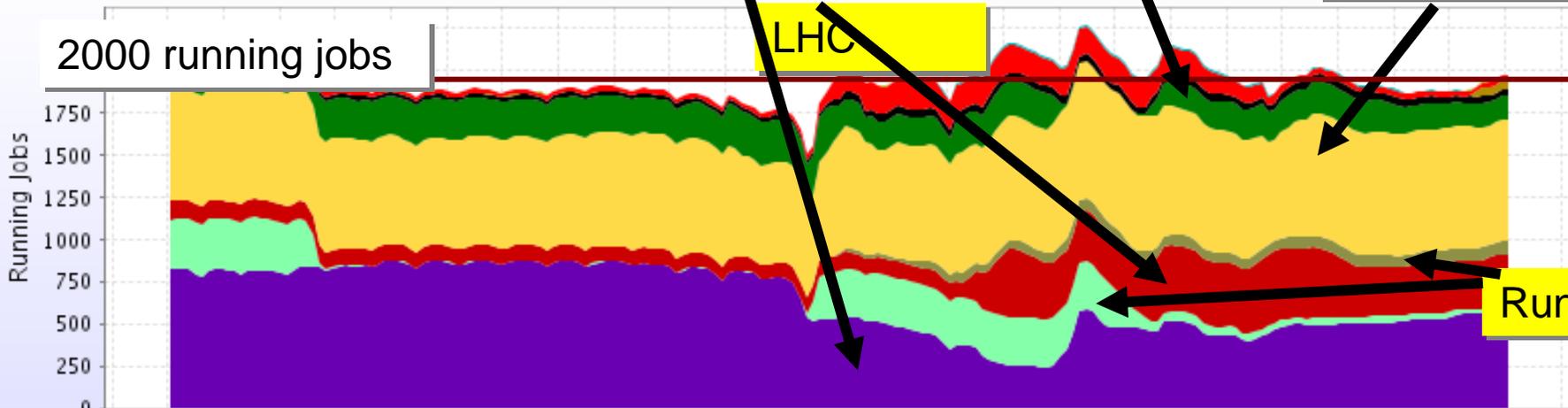
Bioinformatics

Total Jobs per VO

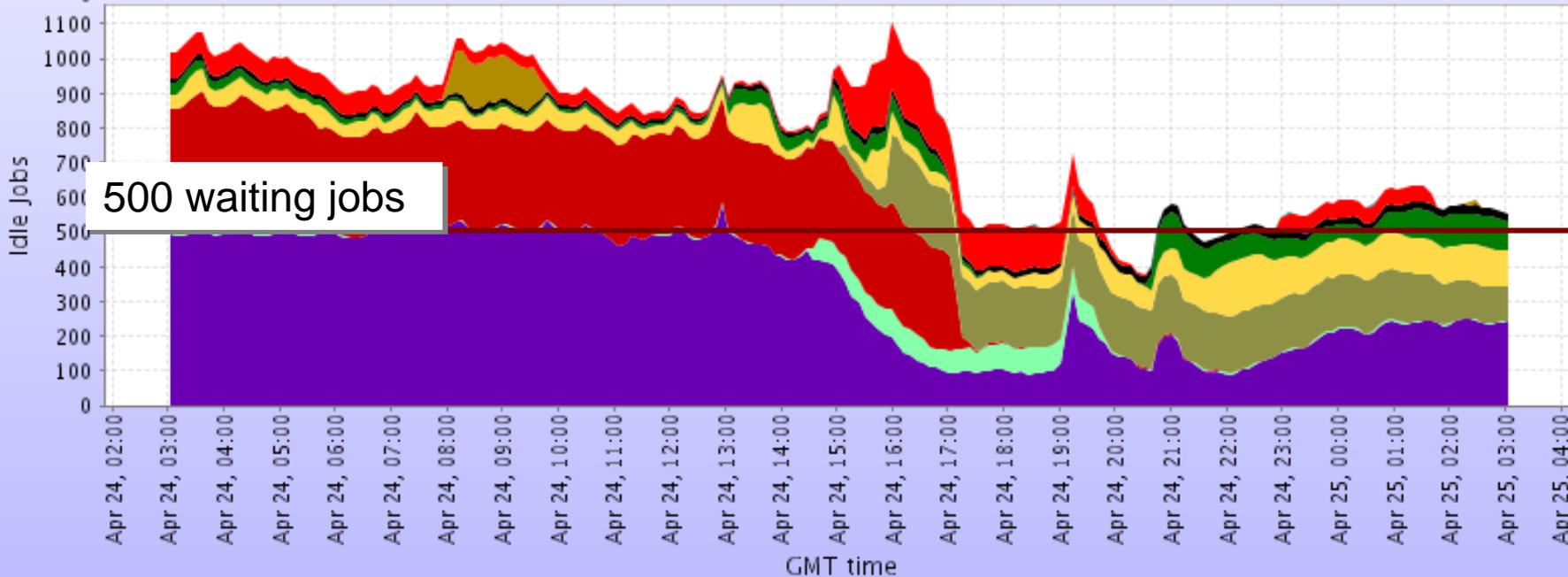
LHC

Run II

2000 running jobs



500 waiting jobs



- ATLAS
- AUGER
- CDF
- CMS
- DZERO
- GADU
- GLOW
- GRIDEX
- IVDGL
- KTEV
- MINIBOONE
- MIS
- SDSS

PHYSTAT – phystat.org repository

- A broadly accessible collection of
 - Tools and utilities
 - Modules and Libraries
 - Code fragments and technical documentation

Pertaining to statistics used in physics
- Idea emerged as an adjunct of the PHYSTAT Conferences on Statistical problems in Particle Physics, Astrophysics, and Cosmology
- Hoped for Scientific outcomes
 - Broader use of established code techniques and packages, in lieu of individual development of the same functionality
 - Presents a forum for consensus across the discipline on the optimal (or at least a “standard”) way to apply new statistics notions

Patriot – Monte Carlo repository

Patriot

FBSNG on the web ~ 200 worker & 2 I/O nodes

Form: FBSNG
Time: Wed Sep 20 11:30:24 2004
Report: List of queues

| Name | Status | Health | System Type | Slaves | Pkts | Waiting | Ready | Running | Total |
|---------------|----------|--------|-------------|--------|------|---------|-------|---------|-------|
| Active queues | ACQIE | OK | Acq_Worner | 1000 | 0 | 0 | 0 | 1 | 1 |
| Queue | Queue | OK | Acq_Worner | 236 | 0 | 0 | 0 | 68 | 68 |
| Process Types | HTCondor | OK | HTCondor | 1000 | 20 | 0 | 0 | 64 | 79 |
| Storage | Storage | OK | Storage | 1.50 | 1000 | 0 | 0 | 1 | 1 |

Putting Tools Together

Multi-Terabyte Mass Storage of final results

enstore
Product Description
Enstore provides distributed access to and management of data stored on tape. It provides a generic interface so experimenters can efficiently use mass storage systems as easily as if they were native file systems.

Standardize Structure for Datasets

STDHEP & MCFIO

```
PARAMETER (MCHIEP=4000)
COMMON /RSPF77/RSRHEP, AKAP, LSTRHP (MCHIEP), LDRHP (MCHIEP),
& JMRHEP (2, MCHIEP), JDRHEP (2, MCHIEP), PHRP (5, MCHIEP), VHRP (4, MCHIEP)
DOUBLE PRECISION PHRP, VHRP
```

Disk storage for results of intermediate steps

Dfarm - Disk Farm System

Abstract
Disk Farm allows using disk space distributed among nodes of a big computing farm by organizing physical disk partitions into a single remote space structure similar to LINUX file system. Disk Farm users access data stored in Disk Farm through a subset of UNIX file system primitive operations such as "mkdir" directory, "list files", "get file", "put file", etc.
Disk Farm helps several negative effects of node unreliability by allowing the user to create replicas of data files on multiple farm nodes.

enstore

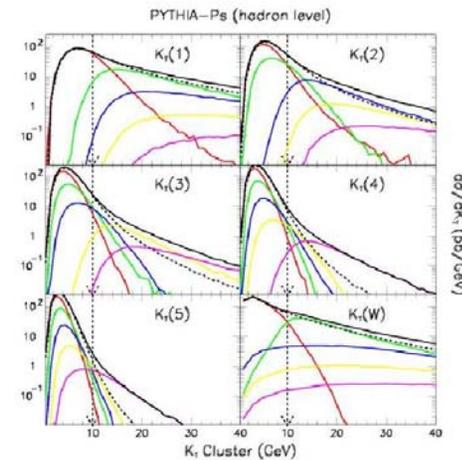
SAM

New on the Grid

Steve Mrenna – pioneering an approach to discovery that leans on computing tools

Clever Matching of Tree Graphs and Parton Showers

Make Better Predictions



Advanced Computing Strategy

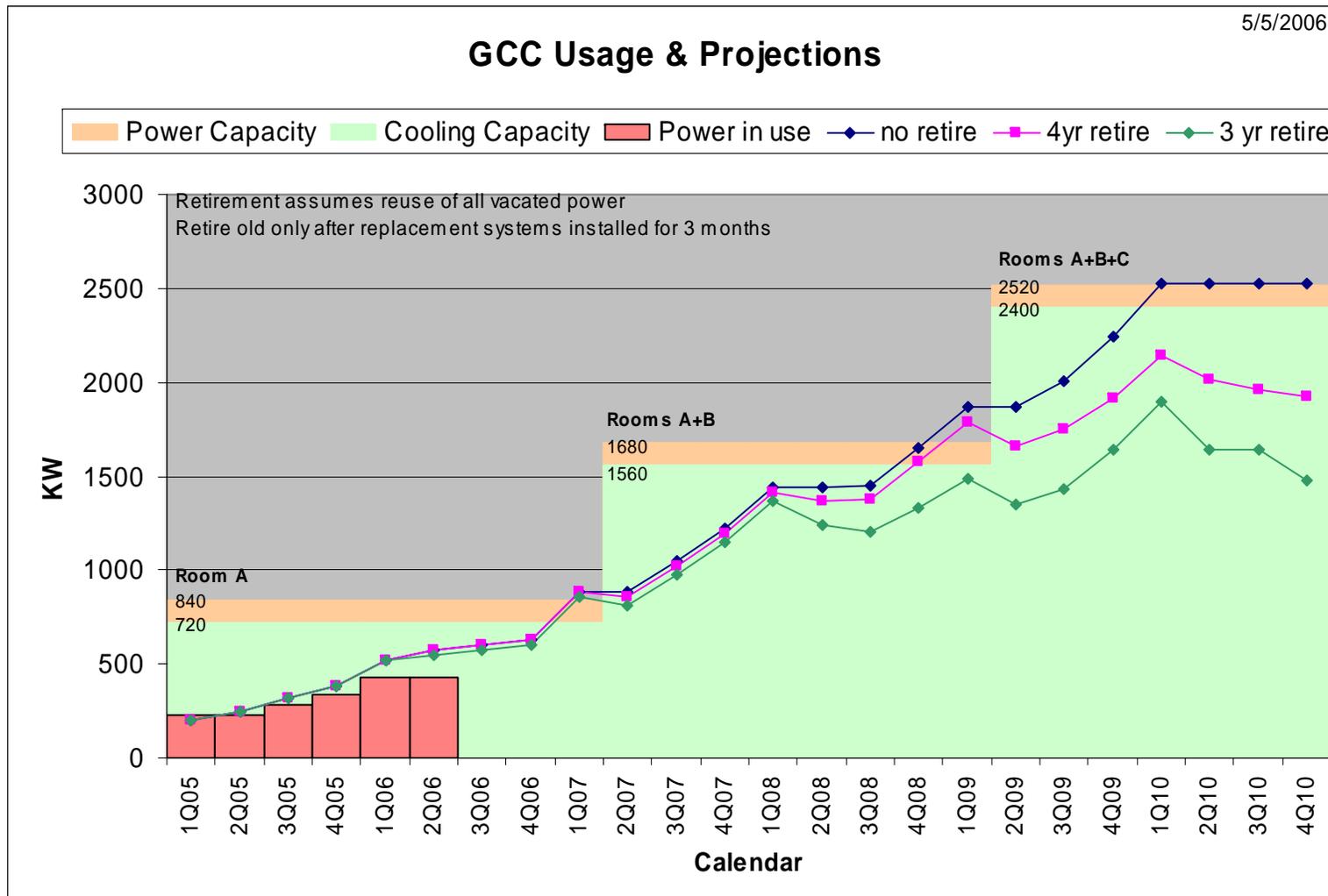
➔ Now a few extra slides on

1. Provide the facility, networking, information management and computer security infrastructure to build on and evolve into the ILC era.
2. Build innovative and specialized computing solutions where we need to for today's scientific programs and for LHC turn-on
3. Collaborate in worldwide efforts and distributed computing solutions such as Grids
4. Do R&D on computing solutions for future experiments and the ILC

Fermilab Grid Computing Center

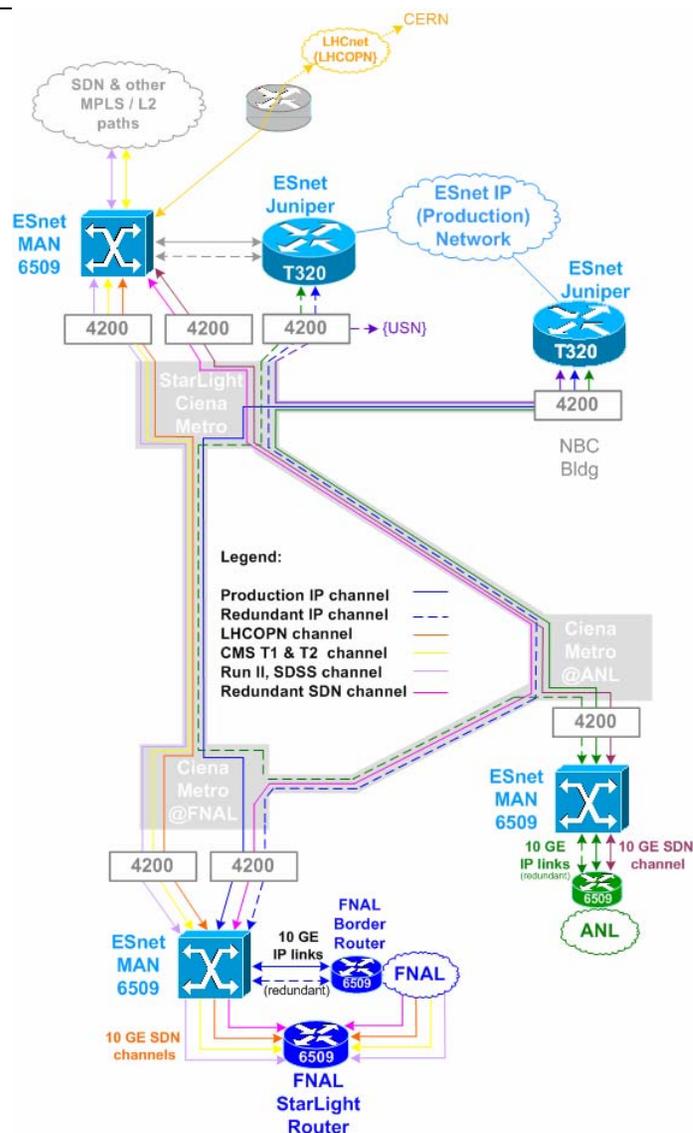


Grid Computing Center Growth



Planned WAN Upgrades

- ESnet Metropolitan Area Network (MAN) will upgrade FNAL connectivity to 6 x 10Gb/s
 - 10 Gb/s production (ESnet) network path, with 10 Gb/s redundant/failover path
 - Four 10 Gb/s channels for specifically designated high impact traffic:
 - One dedicated for CMS T0/T1 traffic
 - Second dedicated to other CMS traffic
 - Third used for Run-II, SDSS, etc
 - Fourth will serve redundancy role
 - MAN expected to be operational by July
- Work group LAN connections to WAN paths to be upgraded to 10GE by July
- Existing two (pre-MAN...) 10 Gb/s channels available for research projects



Information Management Architecture

- Goal: unite the Information management infrastructure (that Computing Division uses to do its business and provide services to the lab) under a common architecture.
- Major emphasis is to make the Division's business more efficient by:
 - automating workflow of existing (often manual) processes;
 - providing uniform & coherent access to enterprise business data;
 - interfacing seamlessly to Laboratory systems;
- Architecture is being applied to:
 - Computing Division HelpDesk
 - Open Science Grid & Fermigrid Support Centers
 - Computer security processes
 - Operations automation
 - Document management
 - Project planning & management
 - Budget planning & execution
 - ... and others
- Advanced features:
 - Enterprise-quality “core” (Oracle, Remedy, PostGRES);
 - API hooks for local user tools (Access, FileMaker, Crystal, etc.);
 - Common workflow based on open-source standards (Plone, AlphaFlow);

Computer Security and Open Science

- Surprisingly this is “Advanced Computing”
 - Keeping us safe
 - Having a good auditable NIST-compliant computer security program (we got outstanding ratings and compliments from recent Site Assist Visit by a team from Office of Science, DOE office of independent oversight and DOE CIO)
 - Making sure we can continue to work collaboratively and openly on Science
 - Evolving the policies and technologies for global Grid computing
 - Fermilab people are leading in these efforts