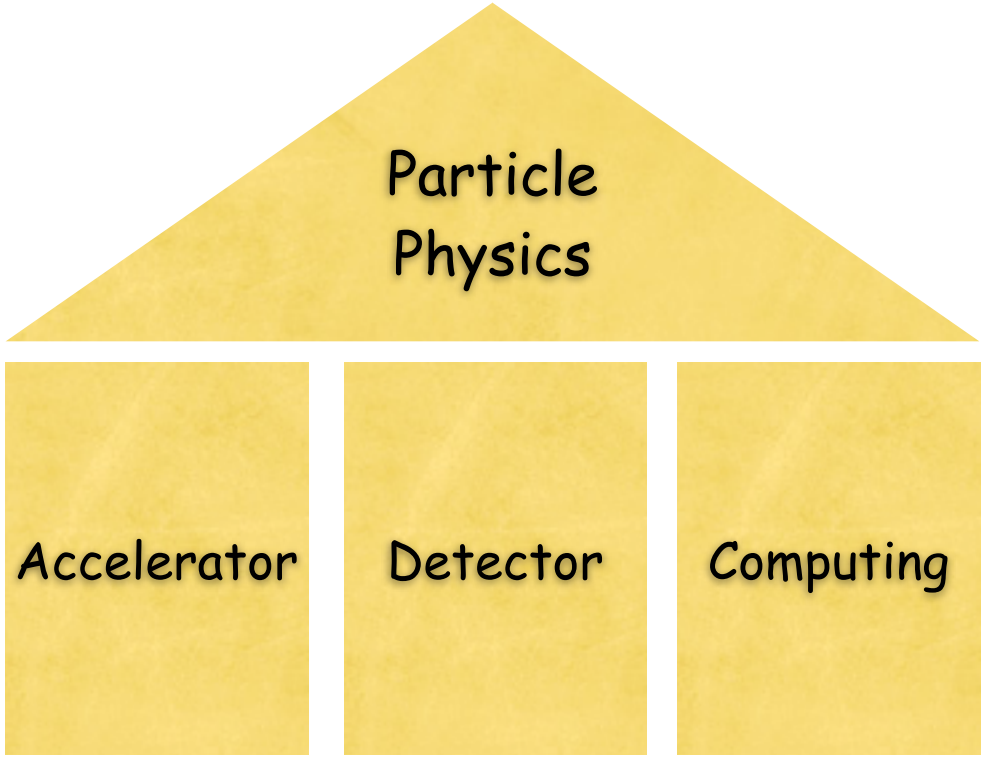
Advanced Computing

Oliver Gutsche

FRA Visiting Committee

April 20/21 2007

Computing Division: Advanced Computing

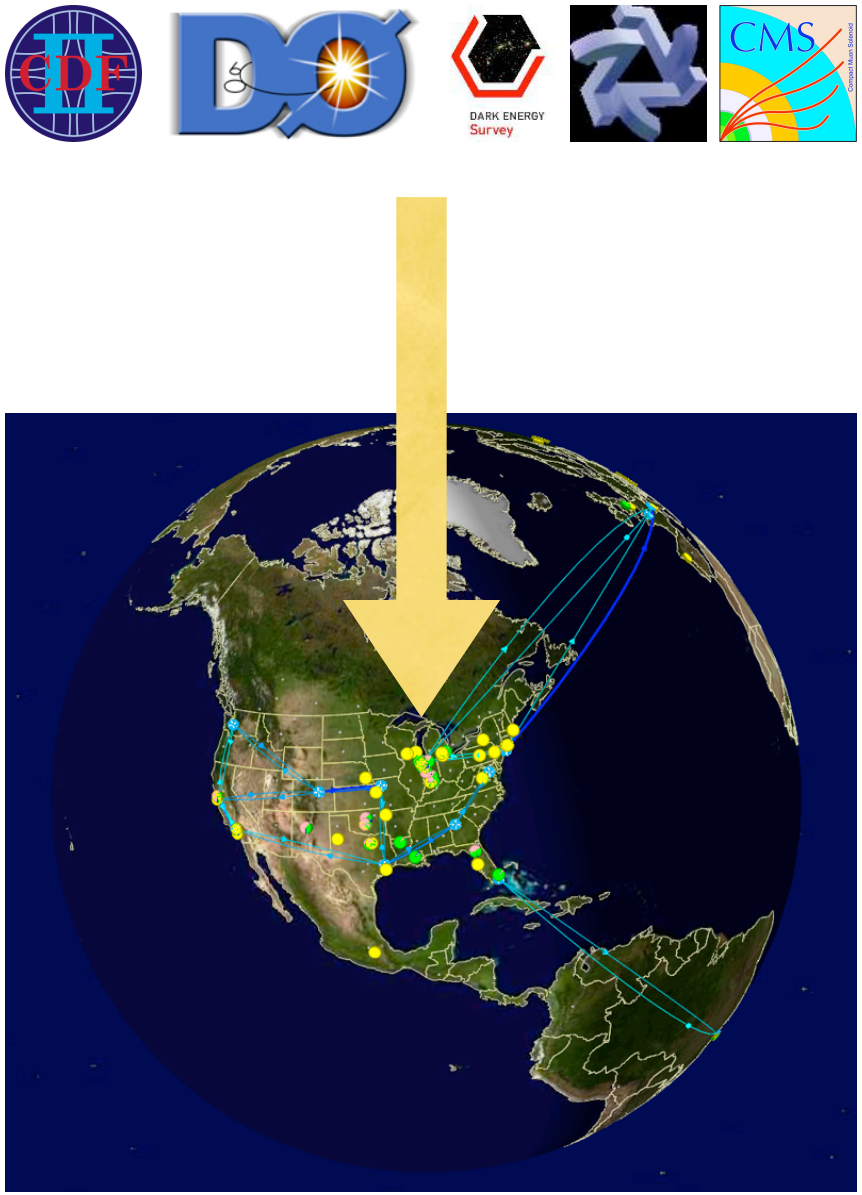


Physicist view of **ingredients** to do Particle Physics

(apart from buildings, roads, ...)

- Particle Physics at Fermilab relies on the Computing Division to:
  - Play a full part in the mission of the laboratory by providing adequate computing for all **ingredients**
- To ensure reaching Fermilab's goals now and in the future, the Computing Division follows its continuous and evolving **Strategy for Advanced Computing** to
  - Develop, innovate and support forefront computing solutions and services

Challenges



- Expected current and future challenges:
  - Significant increase in scale
  - Globalization / Interoperation / Decentralization
  - Special Applications

➡ The Advanced Computing Strategy invests both in
  - Technology
  - Know-How

  to meet all today's and tomorrow's challenges

Outline

Advanced Computing

Scale

- Facilities
- Networking
- Data handling

Globalization

- GRID
  - FermiGrid
  - OSG
  - Security

Special Applications

- Lattice QCD
- Accelerator modeling
- Computational Cosmology

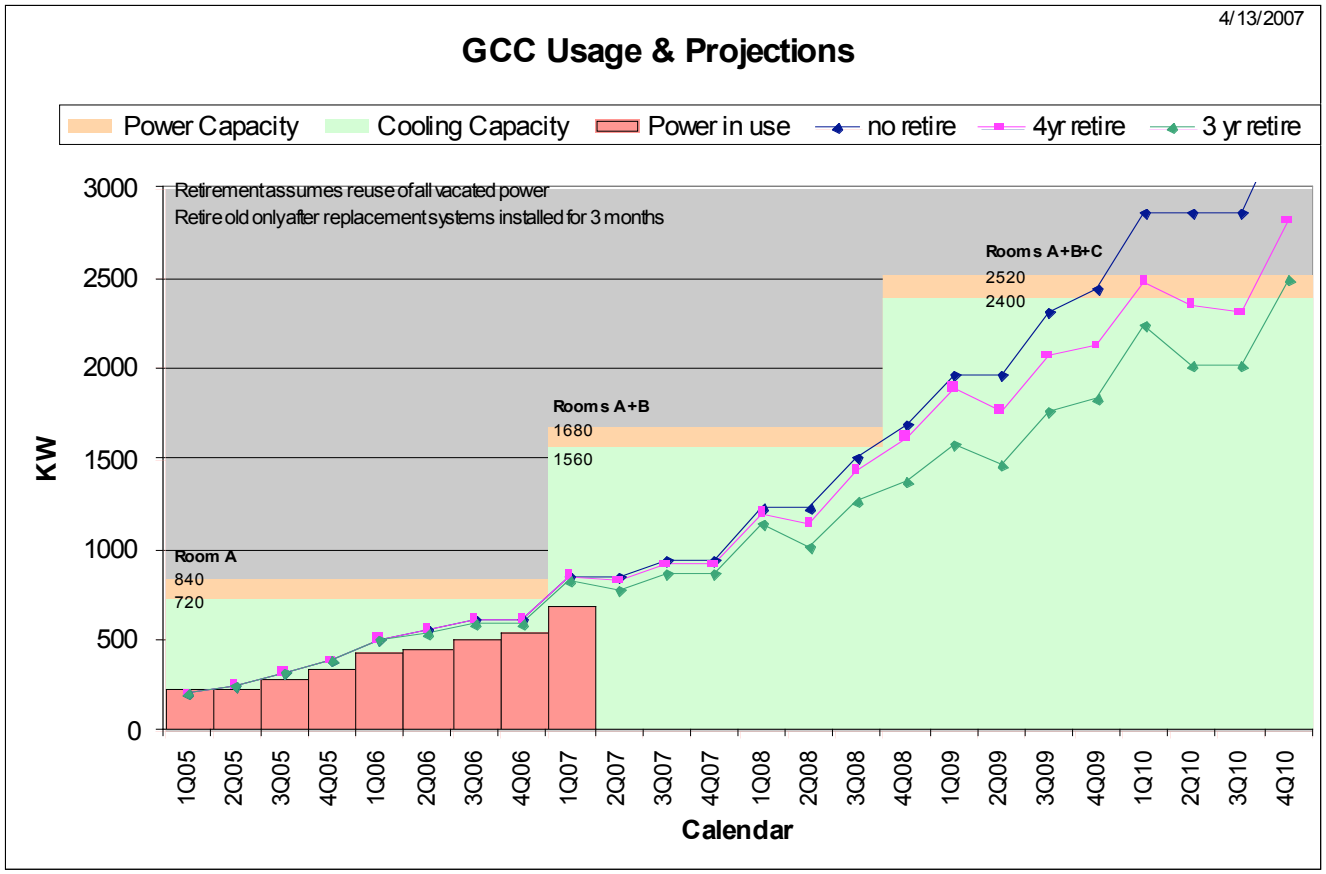
Facilities - Current Status

- Computing Division operates computing hardware and provides and manages needed computing infrastructure, i.e., space, power & cooling
- Computer rooms are located in 3 different buildings (FCC, LCC and GCC)
- Mainly 4 types of hardware:
  - Computing Box (Multi-CPU, Multi-Core)
  - Disk Server
  - Tape robot with tape drive
  - Network equipment

computing boxes	6300
disk	> 1 PetaByte
tapes	5.5 PetaByte in 39,250 tapes (available 62,000)
tape robots	9
power	4.5 MegaWatts
cooling	4.5 MegaWatts

Facilities - Challenges

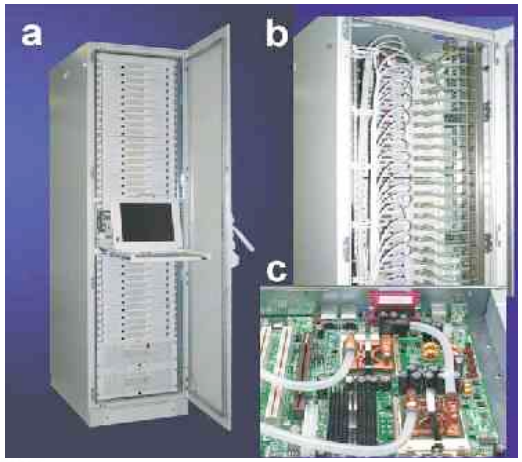
- Rapidly increasing power and cooling requirements for a growing facility
  - More computers are purchased (~1,000/yr)
  - Power required per new computer increases
  - More computers per sq. ft. of floor space

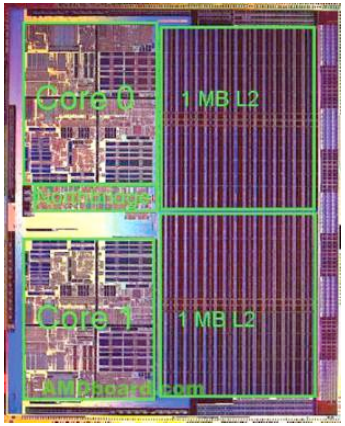


- ➡ Computing rooms have to be carefully planned and equipped with power and cooling, planning has to be reviewed frequently
- ➡ Equipment has to be thoroughly managed (becomes more and more important)
- ➡ Fermilab long-term planning ensures sufficient capacity of the facilities

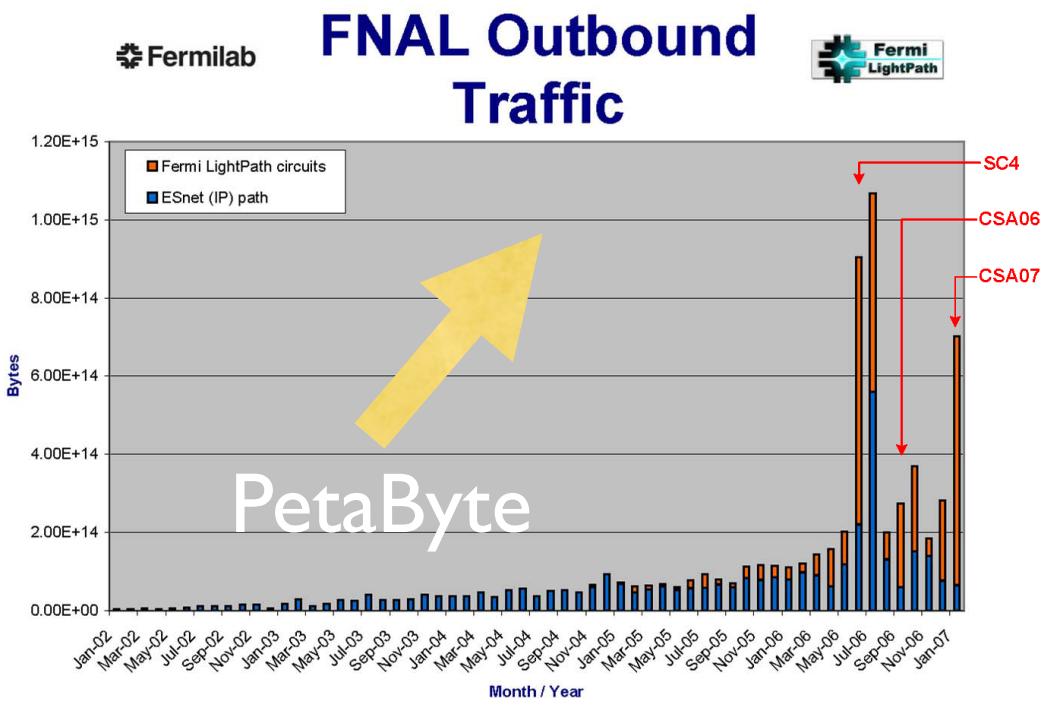
Facilities - Future developments

- To improve sufficient provisioning of computing power, electricity and cooling, following new developments are under discussion:
  - Water cooled racks (instead of air cooled racks)
  - Blade server designs
    - Vertical arrangement of server units in rack
    - Common power supply instead of individual power supplies per unit
    - higher density, lower power consumption
  - Multi-Core Processors due to smaller chip manufacturing processes
    - Same computing power at reduced power consumption

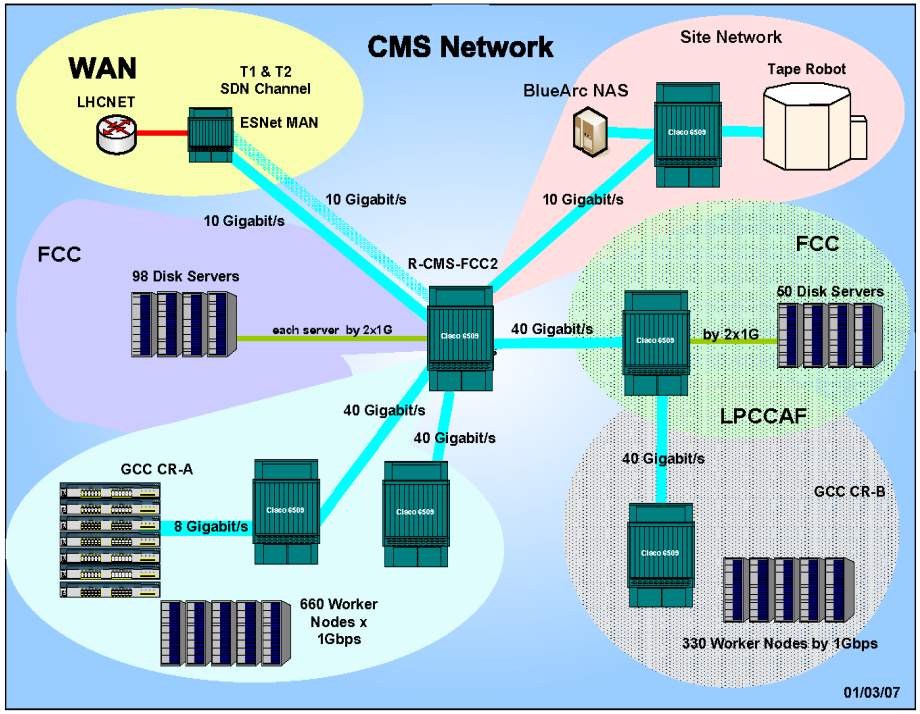




Networking - Current Status



- Wide Area Network traffic dominated by CMS data movements
- Traffic reaches over 1 PetaByte / month during challenges (dedicated simulations of expected CMS default operation conditions)

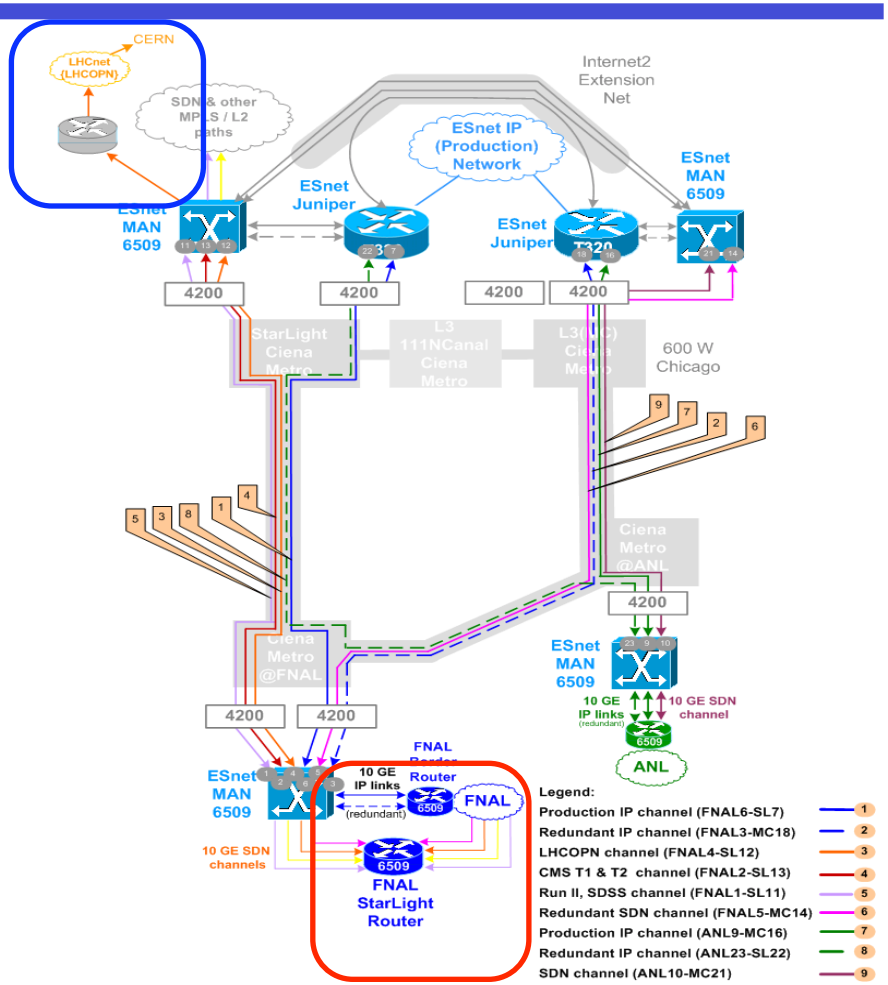


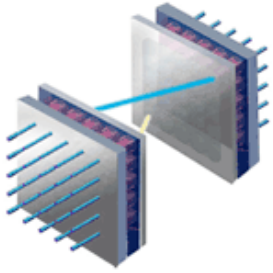
- Local Area Network (on site) provides numerous 10 Gbit/s connections (10 Gbit/s ~ ≥1GB/s) to connect computing facilities

➡ Wide and Local Area Network sufficiently provisioned for all experiments and CMS wide area data movements

Networking - Challenges and Future Developments

- Wide Area and Local Area Connections well provisioned and designed with an upgrade path in mind
- FNAL, ANL and ESnet commission a Metropolitan Area Network to connect Local and Wide Area efficiently with very good upgrade possibilities and increased redundancy
- In addition, 2 10 Gbit/s links are reserved for R&D of optical switches





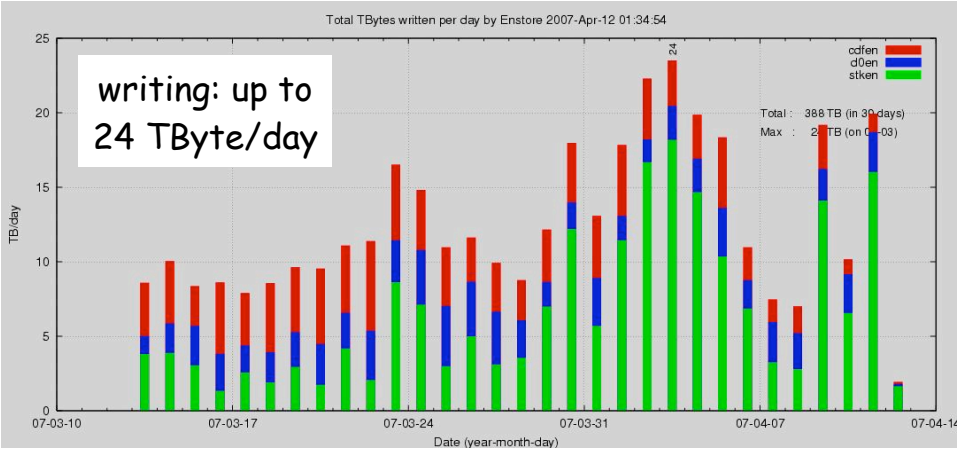
Schematic of optical switching

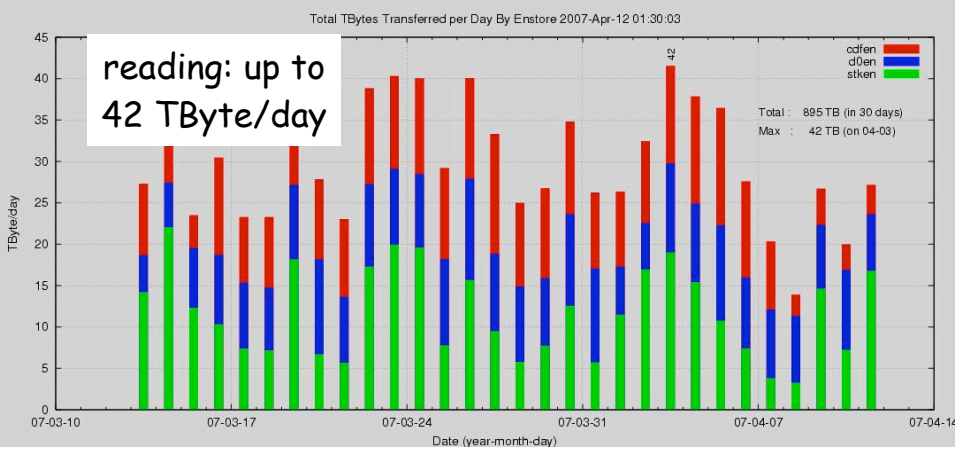
➡ The Advanced Network Strategy's far thinking nature enabled the successful support of CMS data movement

➡ Further R & D in this area will continue to strengthen Fermilab's competence and provide sufficient bandwidth for the future growing demands

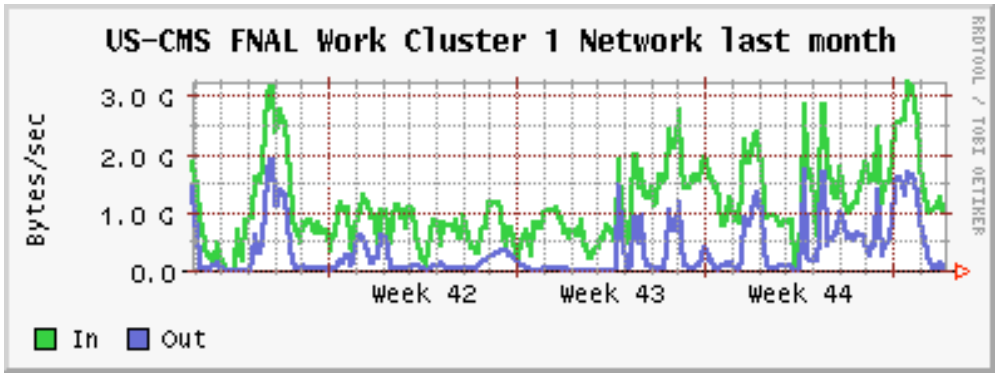
Data Handling - Current Status

- Data handling of experimental data consists of:
  - Storage: "active library-style" archiving on tapes in tape robots
  - Access: disk based system (dCache) to cache sequential/random access patterns to archived data samples
- Tape status: writing up to 24 TeraByte/day, reading more than 42 TeraByte/day



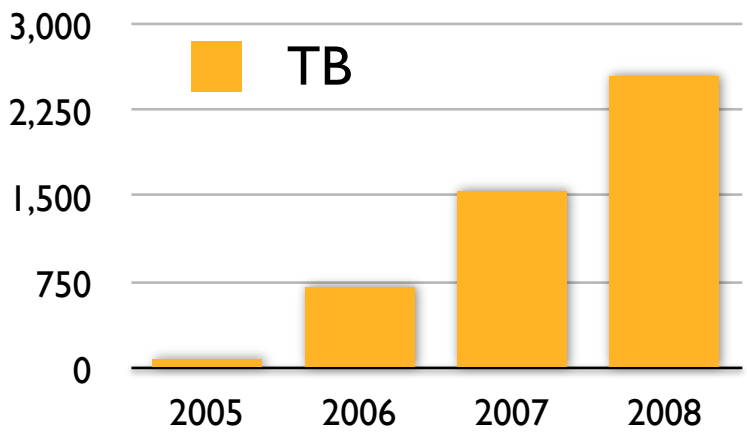


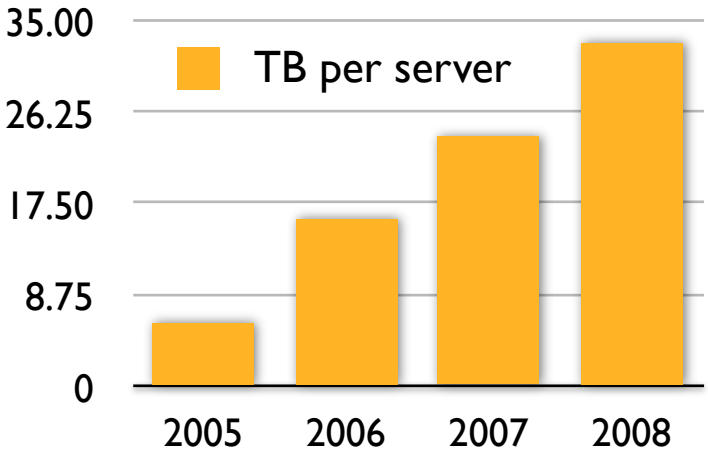
- dCache status, example from CMS:
  - up to 3 GigaBytes/second
  - sustained more than 1 GigaByte/second



Data Handling - Challenges and Future Developments

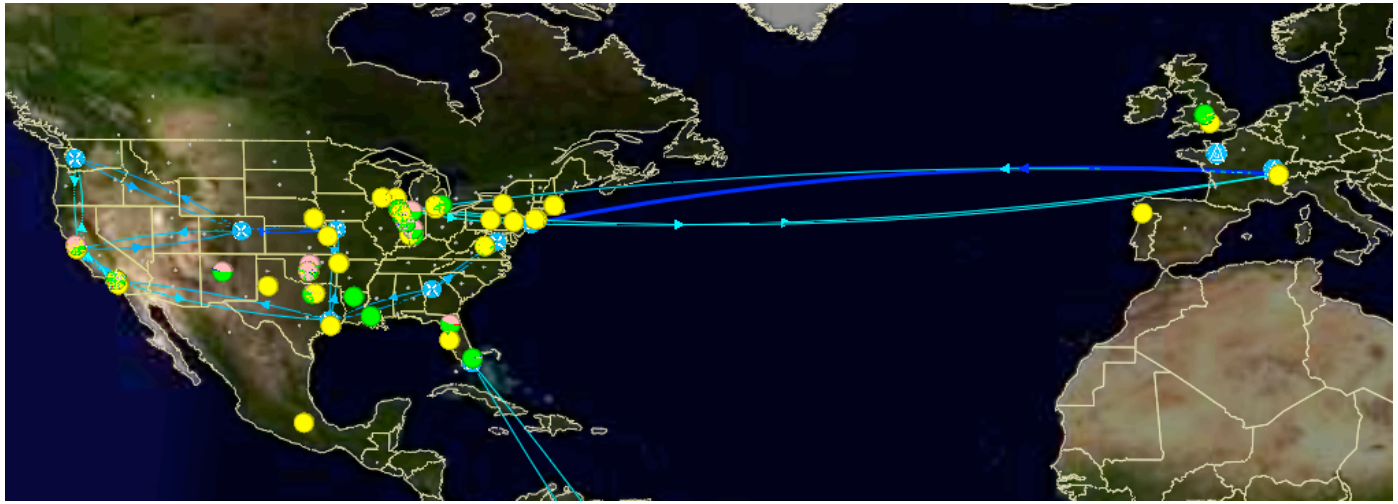
- Tape technology is matured and future developments only related to individual tape size and robot technology
- dCache operation depends on deployment of disk arrays.
  - Current status for CMS at Fermilab: 700 TeraByte on 75 servers
  - Excepted ramp up:





  - New technologies will help to decrease power consumption and space requirements: SATABeast
    - up to 42 disks arranged vertically in 4u unit
    - using 750 GigaByte drives:
      - capacity 31.5 TeraBytes, usable 24 TeraByte
    - expected to increase to 42 TeraBytes with 1 TeraByte drives

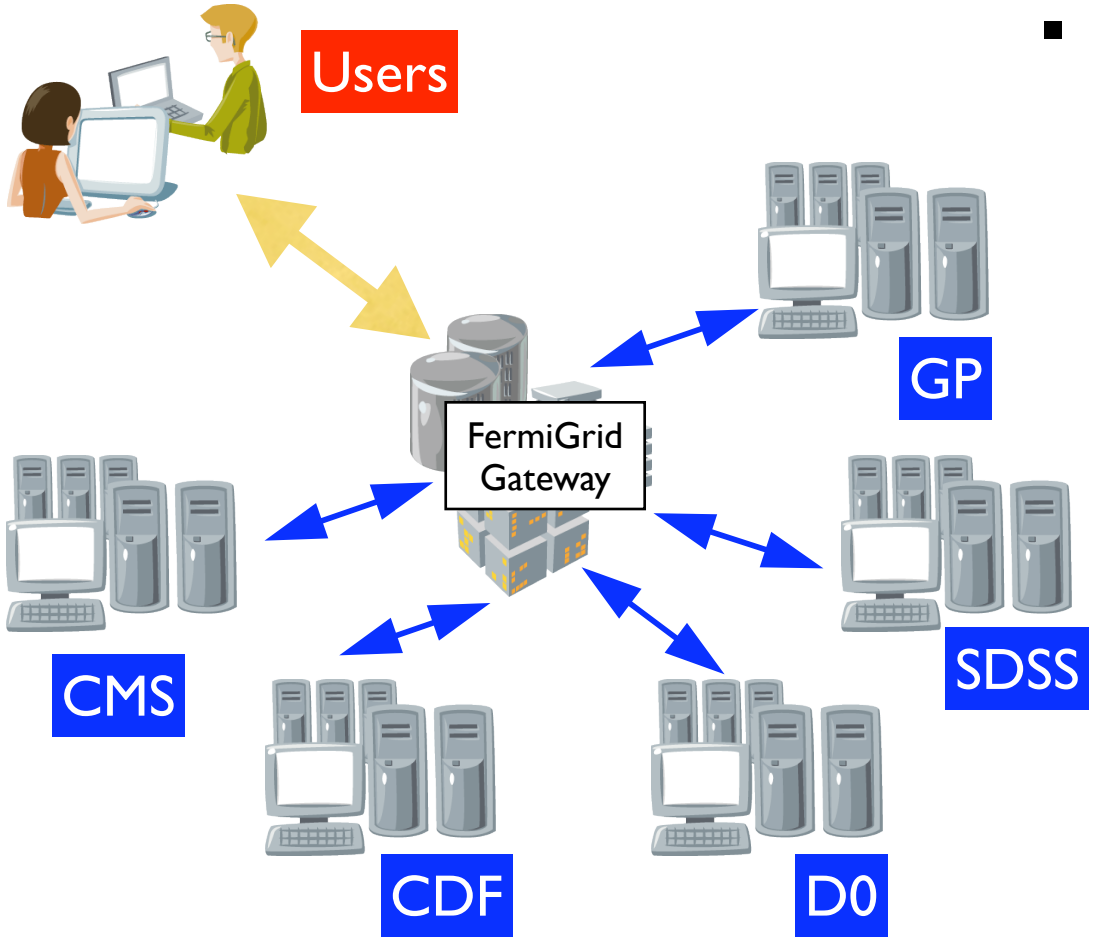
GRID



- Particle Physics Experiments in the LHC era compared to previous experiments:
  - Collaborations consist of significant more collaborators wider distributed over the world
  - Significantly larger computational scales requiring more hardware
- GRID concept provides needed computing for LHC experiments by
  - interconnecting computing centers worldwide (COLLABORATION)
  - providing fairshare access to all resources for all users (SHARING)
- Fermilab plays a prominent role in developing and providing GRID functionalities

CMS-Computing: Tier structure:
20% T0 at CERN, 40% at T1s and 40% at T2s
Fermilab is the largest CMS T1

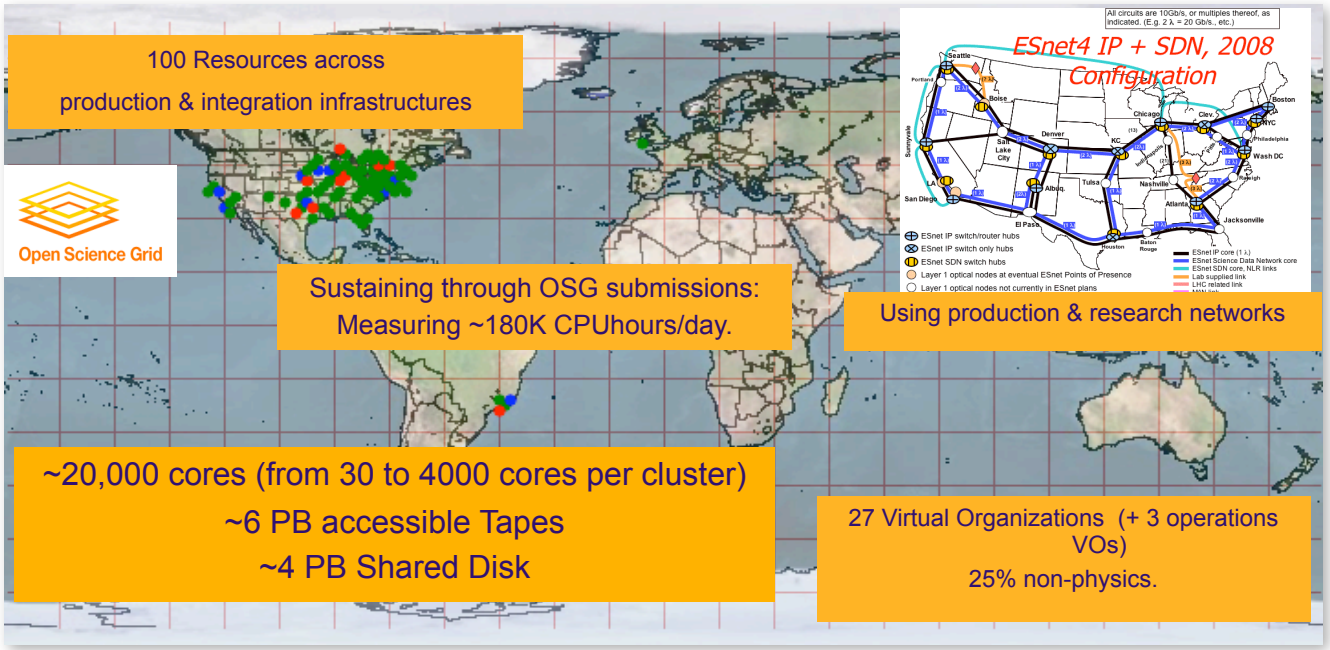
FermiGrid



- FermiGrid is a Meta-Facility forming the Fermi Campus GRID
  - Provides central access point for all Fermilab computing resources from the experiments
  - Enables resource sharing between stakeholders: D0 is using CMS resources opportunistically through the FermiGrid Gateway
  - Portal from the Open Science Grid to Fermilab Compute and Storage Services
- Future developments will continue work in developing campus GRID tools and authentication solutions and also concentrate on reliability and develop failover solutions for GRID services

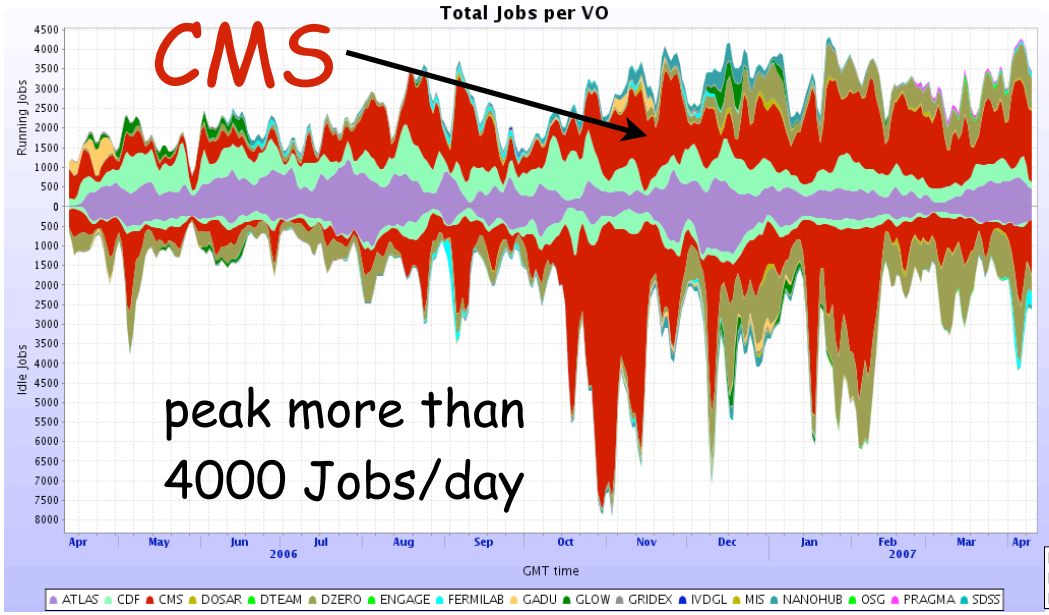
OSG - Current Status

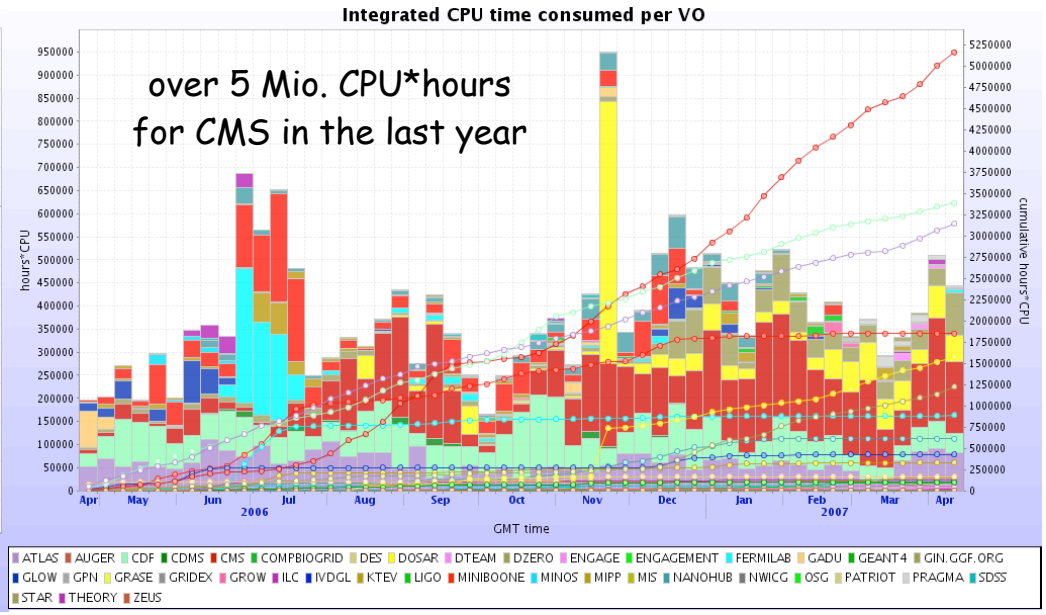
- Open Science Grid (OSG) is the common GRID infrastructure of the U.S.
- SciDAC-2 funded project, goals:
  - Support data storage, distribution & computation for High Energy, Nuclear & Astro Physics collaborations, in particular delivering to the needs of LHC and LIGO science.
  - Engage and benefit other Research & Science of all scales through progressively supporting their applications.



- Fermilab is in a leadership position in OSG
  - Fermilab provides the Executive Director of the OSG
  - Large commitment of Fermilab's resources by access via FermiGrid

OSG - Challenges and Future Developments

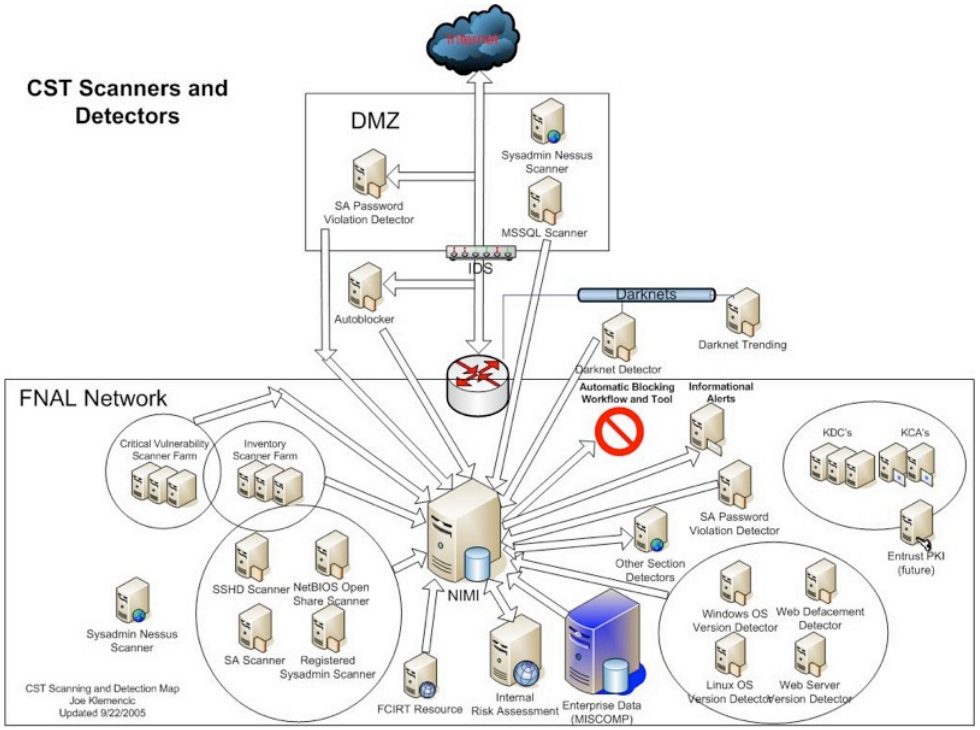




- Last year's OSG usage shows significant contributions of CMS and FNAL resources
- Future developments will concentrate on:
  - Efficiency and Reliability
  - Interoperability
  - Accounting
- Further future developments from collaboration of Fermilab Computing Division, Argonne and University of Chicago in many areas:
  - Accelerator physics
  - Peta-Scale Computing
  - Advanced Networks
  - National Shared Cyberinfrastructure

Security

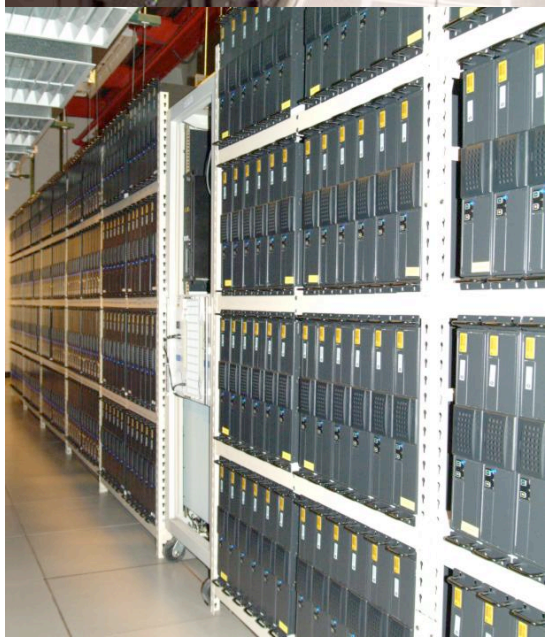
- Fermilab strives to provide secure operation of all its computing resources and prevent compromise without putting undue burdens on experimental progress:
  - Enable high performance offsite transfers without performance-degrading firewalls
  - Provide advanced infrastructure for
    - Aggressive scanning and testing of on-site systems to assess that good security is practiced
    - deployment of operation system patches so that systems exposed to internet are safe



- GRID efforts open a new dimension for security related issues
  - Fermilab is actively engaged in handling security in the collaboration with non-DOE institutions (e.g. US and foreign universities, etc.) and within worldwide GRIDs
  - Fermilab provides the OSG security officer do provide secure GRID computing

Lattice QCD - Current Status

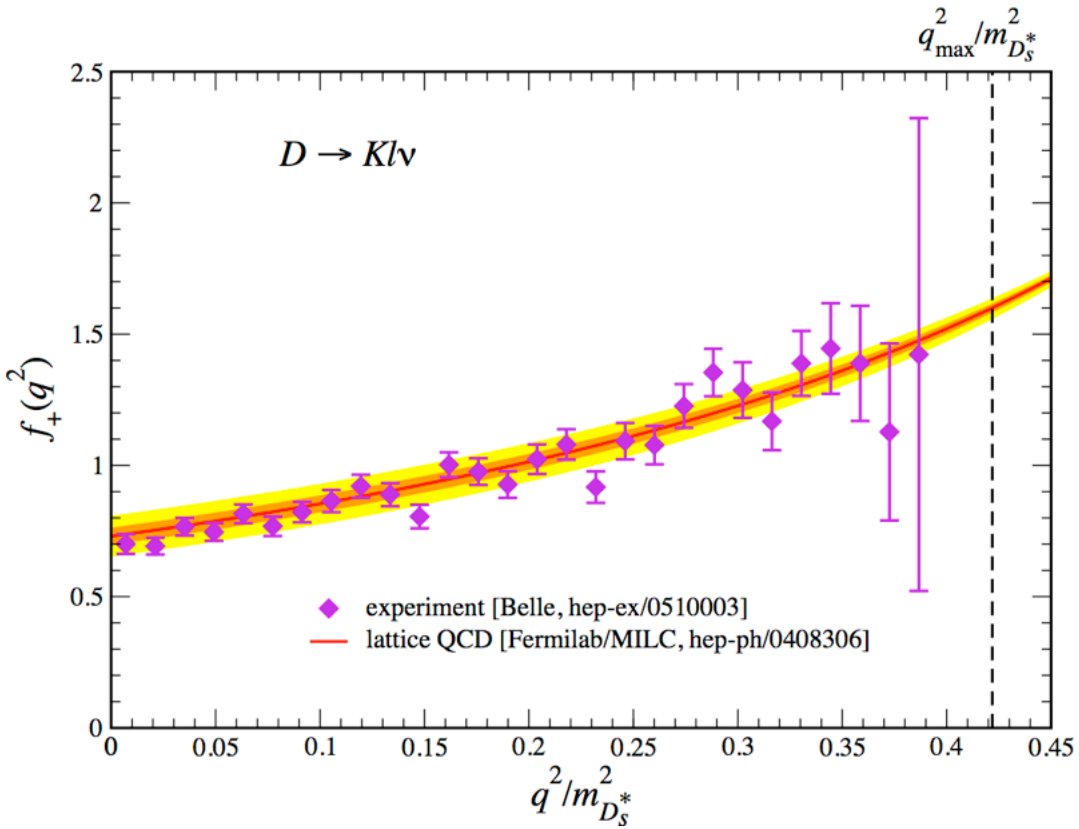
- Lattice QCD requires computers consisting of hundreds of processors working together via high performance network fabrics
  - Compared to standard Particle Physics applications, the individual jobs running in parallel have to communicate with each other with very low latency requiring specialized hardware setups
- Fermilab operates three such systems:
  - "QCD" (2004) – 128 processors coupled with a Myrinet 2000 network, sustaining 150 GFlop/sec
  - "Pion" (2005) – 520 processors coupled with an Infiniband fabric, sustaining 850 GFlop/sec
  - "Kaon" (2006) – 2400 processor cores coupled with an Infiniband fabric, sustaining 2.56 TFlop/sec



Lattice QCD - Example

- Recent results:
  - D meson decay constants
  - Mass of the B_c

    (One of 11 top physics results of the year (AIP))
  - D meson semileptonic decay amplitudes (see histogram)
- Nearing completion
  - B meson decay constants
  - B meson semileptonic decay amplitudes
  - Charmonium and bottomonium spectra
  - BBbar mixing



Lattice QCD - Future Developments

- New computers:
  - For the DOE 4-year USQCD project, Fermilab is scheduled build:
    - a 4.2 TFlop/sec system in late 2008
    - a 3.0 TFlop/sec system in late 2009
- Software projects (funded by SCiDAC-2):
  - new and improved libraries for LQCD computations
  - multicore optimizations
  - automated workflows
  - reliability and fault tolerance
  - visualizations

TOP 500 Supercomputer

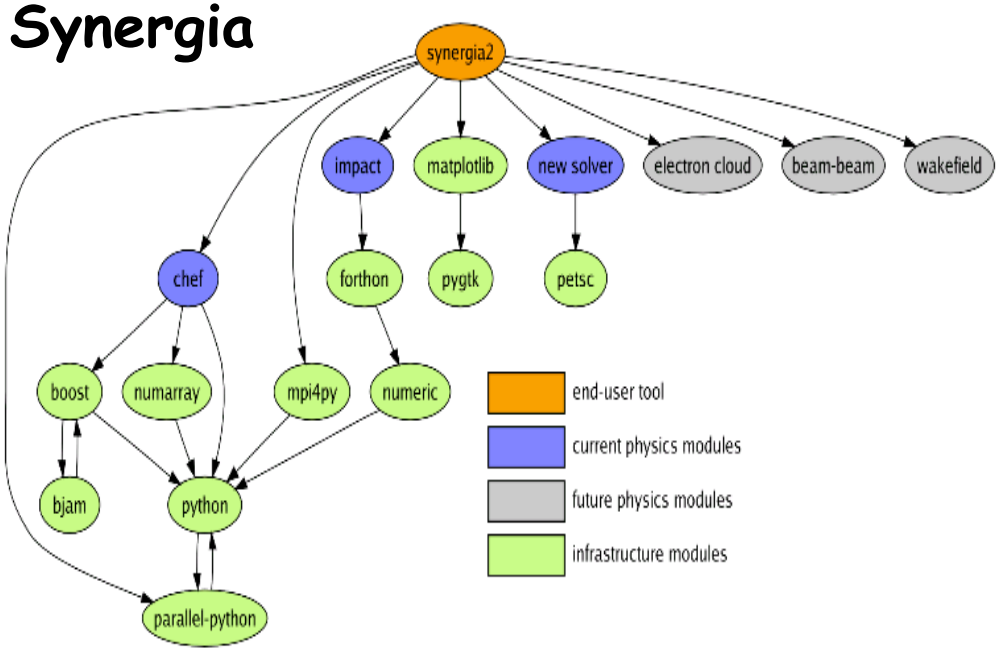
Rank	Site	Computer	TFlops
1	DOE/NNSA/LLNL	eServer Blue Gene Solution	280.6
2	NNSA/Sandia National Laboratories	Sandia/ Cray Red Storm, Opteron 2.4 GHz dual core	101.4
3	IBM Thomas J. Watson Research Center	eServer Blue Gene Solution	91.29
4	DOE/NNSA/LLNL	eServer pSeries p5 575 1.9 GHz	75.76
5	Barcelona Supercomputing Center	BladeCenter JS21 Cluster, PPC 970, 2.3 GHz, Myrinet	62.63
6	NNSA/Sandia National Laboratories	PowerEdge 1850, 3.6 GHz, Infiniband	53
7	Commissariat a l'Energie Atomique (CEA)	NovaScale 5160, Itanium2 1.6 GHz, Quadrics	52.84
8	NASA/Ames Research Center/NAS	SGI Altix 1.5 GHz, Voltaire Infiniband	51.87
9	GSIC Center, Tokyo Institute of Technology	Sun Fire x4600 Cluster, Opteron 2.4/2.6 GHz and ClearSpeed Accelerator, Infiniband	47.38
10	Oak Ridge National Laboratory	Cray XT3, 2.6 GHz dual Core	43.48
⋮			
90	Lawrence Livermore National Laboratory	ASCI White, SP Power3 375 MHz	7.304
91	NERSC/LBNL	SP Power3 375 MHz 16 way	7.304
92	NCSA	TeraGrid, Itanium2 1.3/1.5 GHZ, Myrinet	7.215
93	US Army Research Laboratory (ARL)	eServer Opteron 2.2 GHz. Myrinet	7.185
94	Swiss Scientific Computing Center (CSCS)	Cray XT3, 2.6 GHz	7.182
95	Fermi National Accelerator Laboratory	Opteron 2.0 GHz, Infiniband	6.894
96	University of Sherbrooke	PowerEdge SC1425 3.6 GHz-Infiniband	6.888
97	Nagoya University	PRIMEPOWER HPC2500 (2.08 GHz)	6.86
98	Los Alamos National Laboratory	Opteron 2.6 GHz, Infiniband	6.677
99	Joint Supercomputer Center	MVS-15000BM, eServer BladeCenter JS20 (PowerPC970 2.2 GHz), Myrinet	6.64553
100	Bio Tech	BladeCenter LS20, Opteron 2.2 GHz Dual core, GigEthernet	6.61893

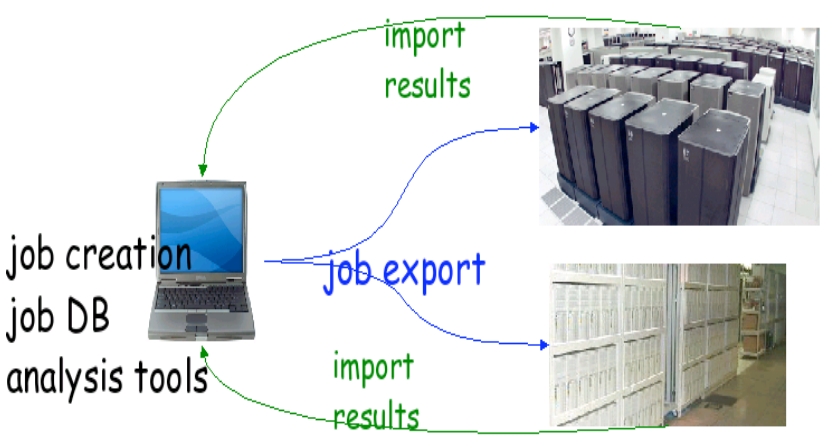
Kaon

Accelerator Modeling - Current Status

- Introduction to accelerator modeling
  - provide self-consistent modeling of both current and future accelerators
  - main focus is to develop tools necessary to model collective beam effects, but also to improve single-particle-optics packages
- Benefits from Computing Division's experience in running specialized parallel clusters from Lattice QCD
  (both in expertise and hardware)

Accelerator modeling framework:
Synergia

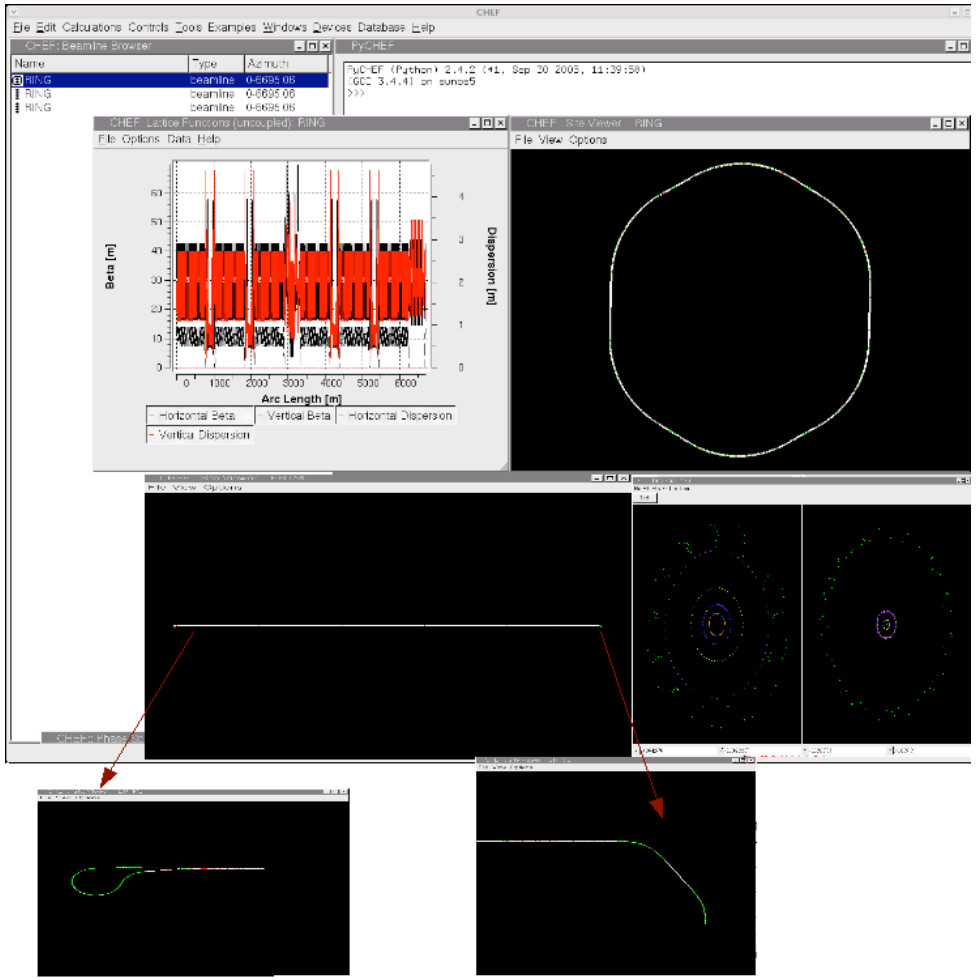


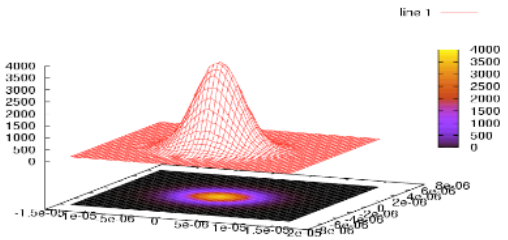


- Since '01 member of a multi-institutional collaboration funded by SciDAC to develop & apply parallel community codes for design & optimization.
- SciDAC-2 proposal submitted Jan '07, with Fermilab as the lead institution

Accelerator Modeling - Example

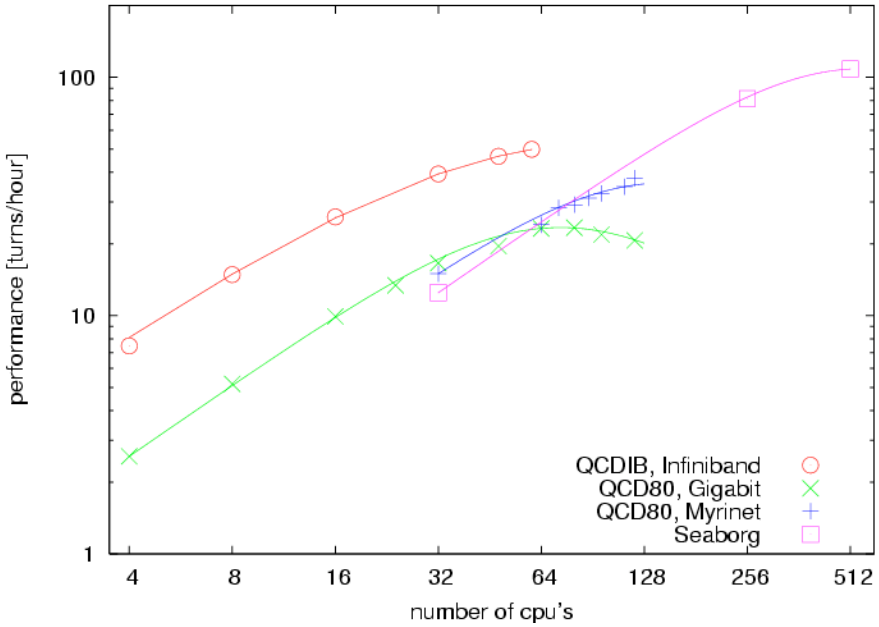
- Current activities cover simulations for Tevatron accelerators and studies for the ILC
- Example:
  - ILC damping ring
    - Study space-charge effects
      - halo creation
      - dynamic aperture
    - using Synergia (3D, self-consistent)
    - Study space-charge in RTML lattice (DR to ML transfer line)

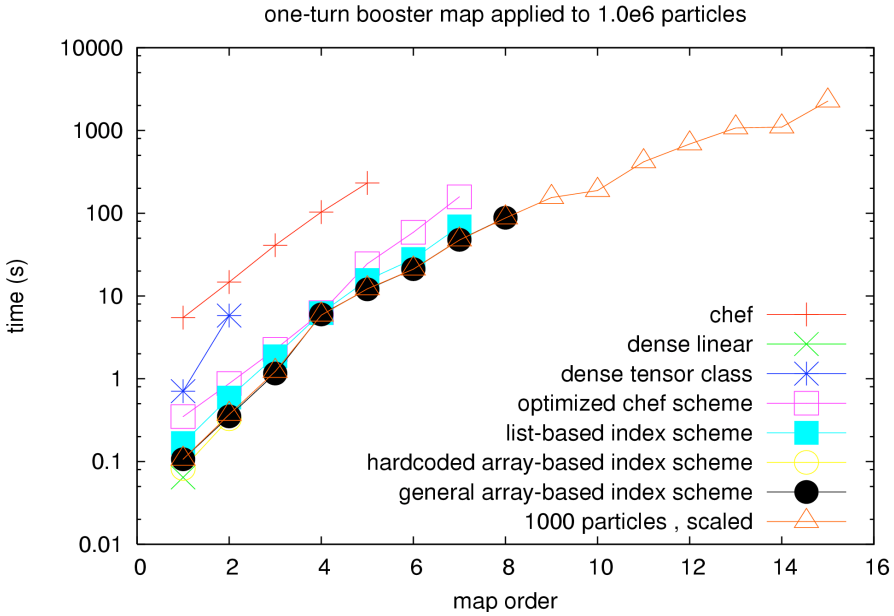




Accelerator Modeling - Future Developments

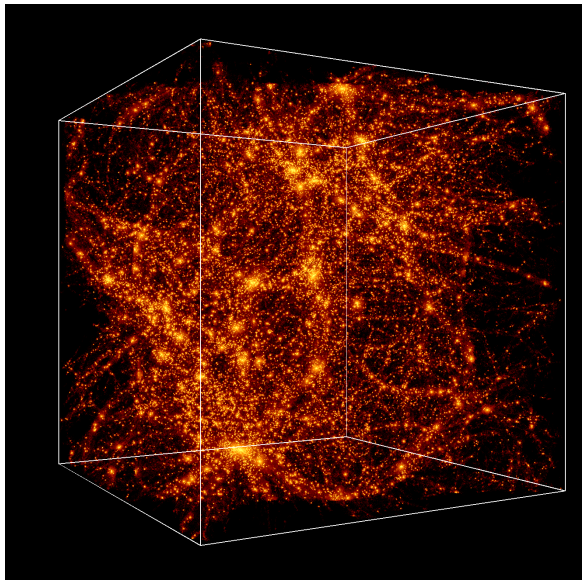
- Currently utilizing different architectures:
  - multi-cpu large-memory node clusters (NERSC SP3)
  - "standard" Linux clusters
    - Recycle Lattice QCD hardware
- Future computing developments:
  - Studies of parallel performance
    - Case-by-case optimization
  - Optimization of particle tracking

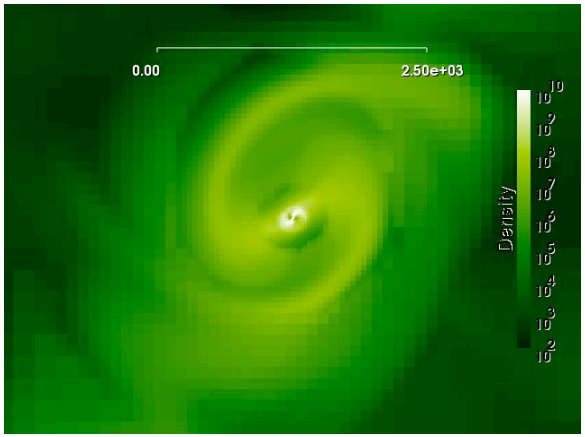




Computational Cosmology

- New project in very early stage
- Proposal to FRA for collaboration with UC on computational cosmology using
  - expertise of the Theoretical Astrophysics Group
  - world-class High Performance Computing (HPC) support of the FNAL Computing Division
- Simulation of large scale structures, galaxy formation, supermassive black holes, etc.
- Modern state-of-the art cosmological simulations require even more inter-communication between processes as Lattice QCD and:
  - ≥ 100,000 CPU-hours (130 CPU-months). Biggest ones take > 1,000,000 CPU-hours.
  - computational platforms with wide (multi-CPU), large-memory nodes.





Summary & Outlook

- The Fermilab Computing Division's continuous and evolving Strategy for Advanced Computing plays a prominent role in reaching the laboratory's goals:
  - It enabled the successful operation of ongoing experiments and provided sufficient capacities for the currently ongoing ramp-up of LHC operations
  - The ongoing R&D will enable the laboratory to do so in the future as well
- The Computing Division will continue to follow and further develop the strategy by:
  - Continuing maintenance and upgrade of existing infrastructure
  - Addition of new infrastructure
  - Significant efforts in Advanced Computing R&D to extend capabilities in traditional and new fields of Particle Physics Computing
- Physicist's summary:
  - Fermilab is worldwide one of the best places to use and work on the latest large scale computing technologies for Particle and Computational Physics