Grid Computing in Particle Physics

Norbert Neumeister
Department of Physics
Purdue University

NWICG Workshop, University of Notre Dame
August 16-17, 2005

Outline

- Introduction
  - What is Particle Physics
  - The Large Hadron Collider
  - The Compact Muon Solenoid detector
- The Grid
  - Motivation and Requirements
  - Grid computing
- LHC Computing and Analysis Challenges
  - Distributed analysis
  - Tier-2 center at Purdue
- Summary

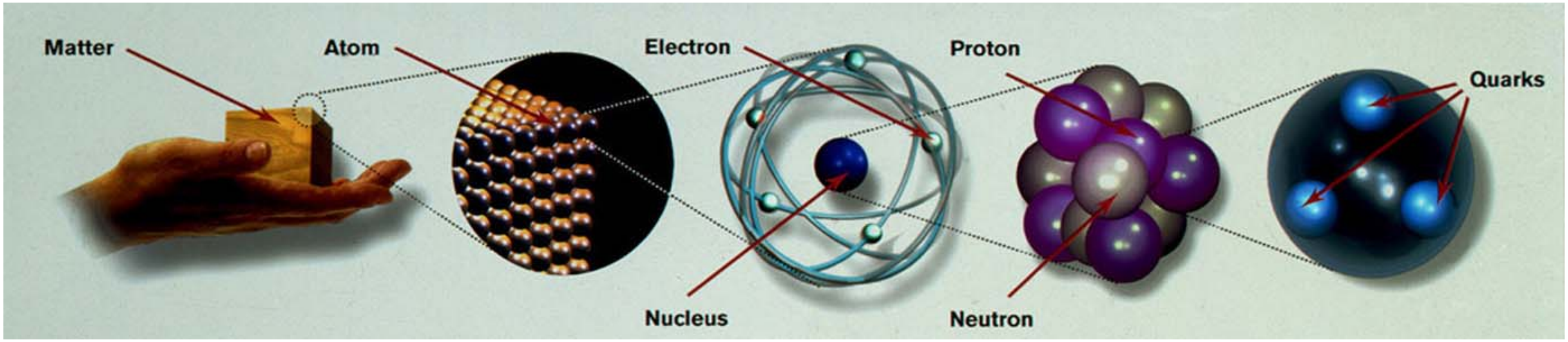
Particle Physics

Aim to answer the two following questions:

- **What are the elementary constituents of matter?**
- **What are the fundamental forces that control their behavior at the most basic level?**

Tools:

Particle Accelerators, Particle Detectors, Computers



atom 10^{-10} m

nucleus 10^{-14} m

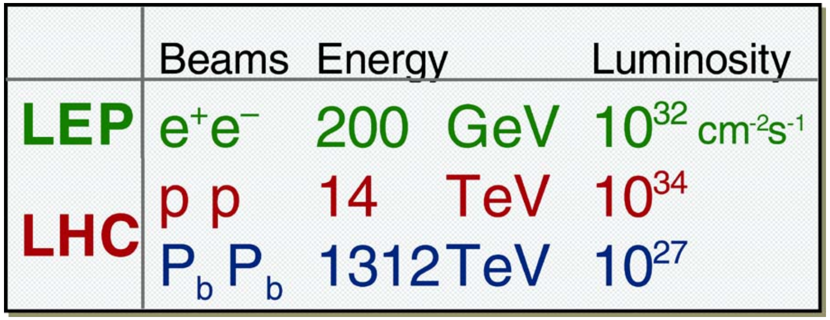
nucleon 10^{-15} m

quark 10^{-18} m

The Large Hadron Collider

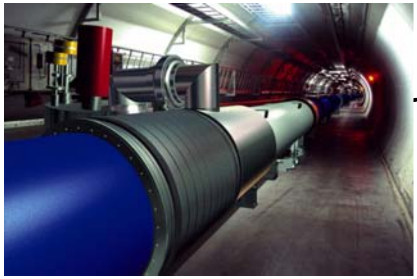
Energy frontier, high Luminosity proton proton collider at CERN, Geneva, Switzerland

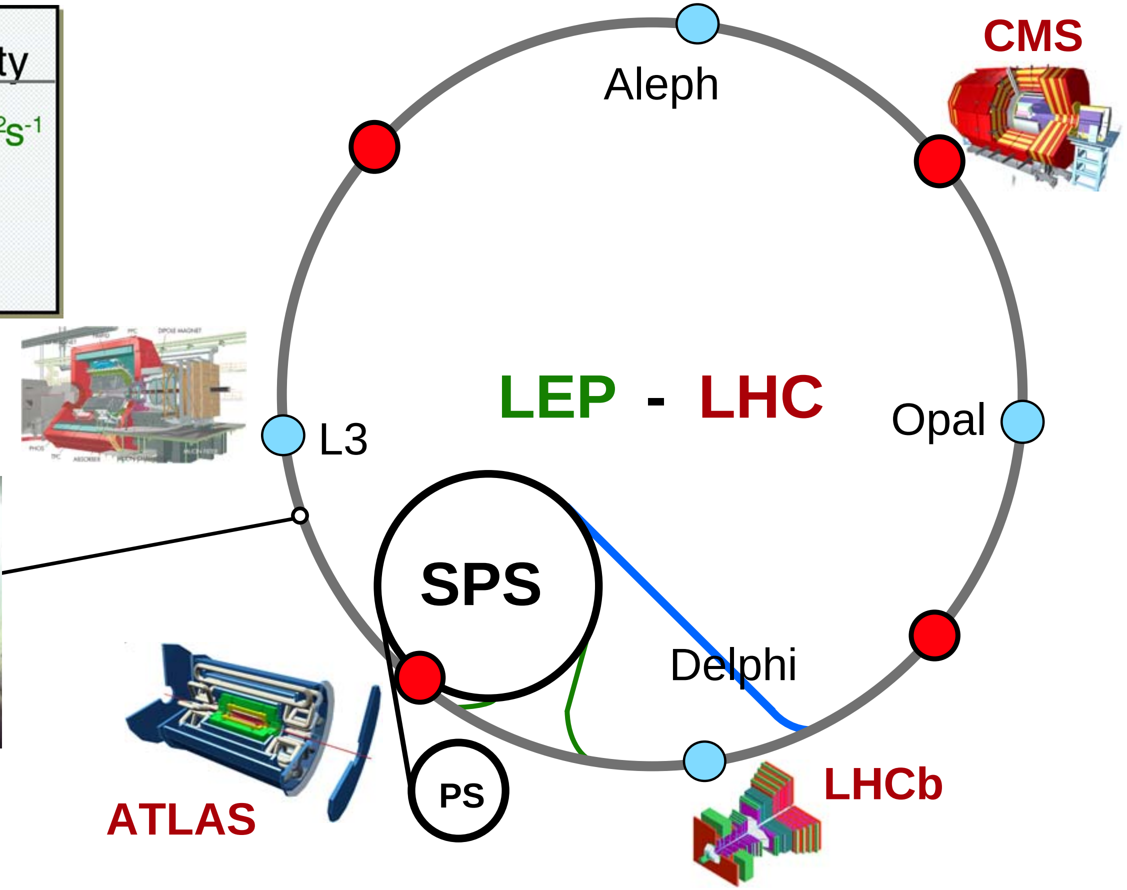
Experiments at the LHC



	Beams	Energy		Luminosity
LEP	e^+e^-	200	GeV	10^{32} cm^{-2}s^{-1}
LHC	p p	14	TeV	10^{34}
	$P_b P_b$	1312	TeV	10^{27}

Two super-conducting magnet rings in the LEP tunnel





Experiments:

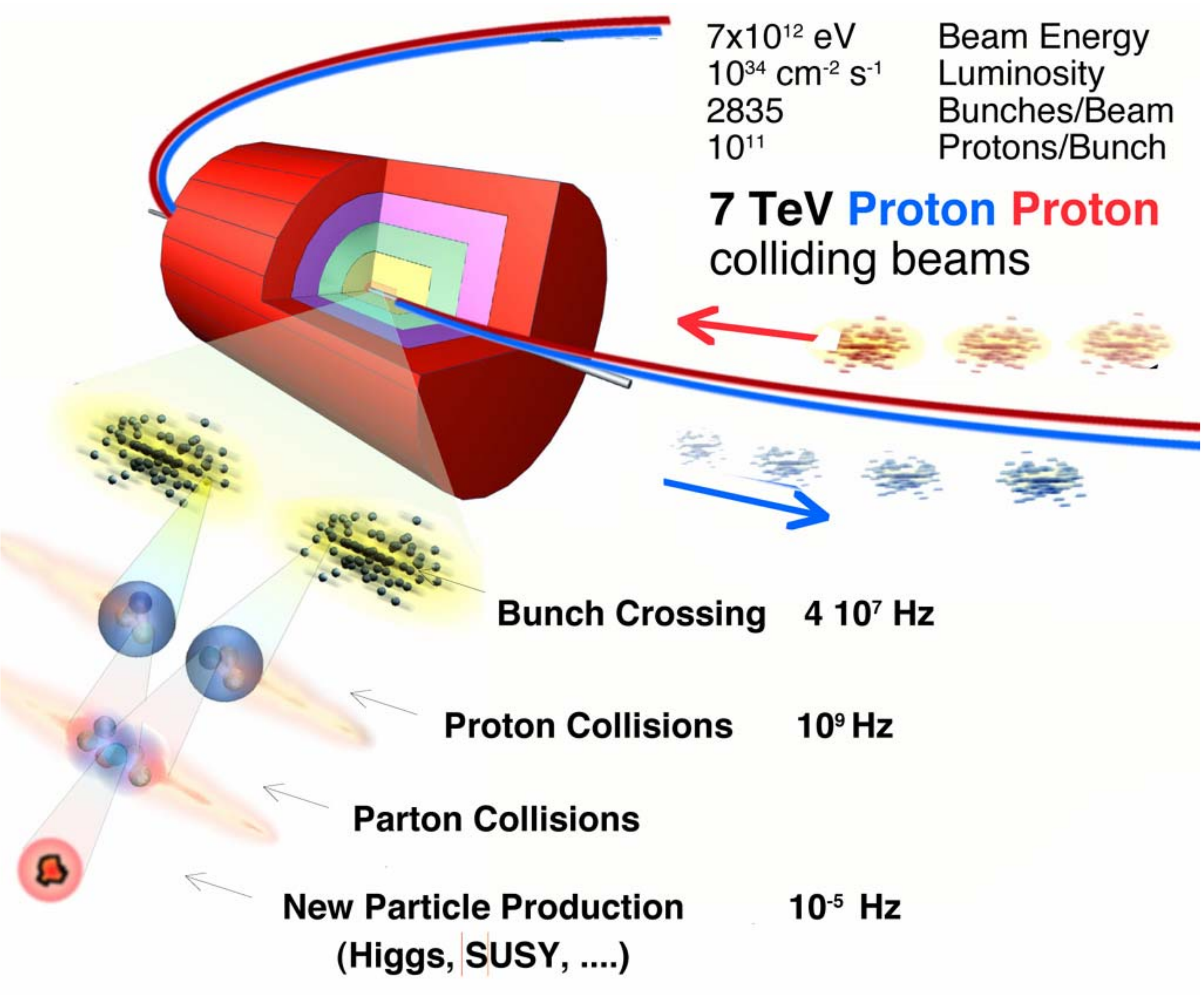
ATLAS A Toroidal LHC ApparatuS (Study of Proton-Proton collisions)

CMS Compact Muon Solenoid (Study of Proton-Proton collisions)

ALICE A Large Ion Collider Experiment (Study of Ion-Ion collisions)

LHCb (Study of CP violation in B-meson decays at the LHC)

Large Hadron Collider



$7\mathrm{x}10^{12}$ eV	Beam Energy
10^{34} cm^{-2} s^{-1}	Luminosity
2835	Bunches/Beam
10^{11}	Protons/Bunch

7 TeV Proton Proton colliding beams

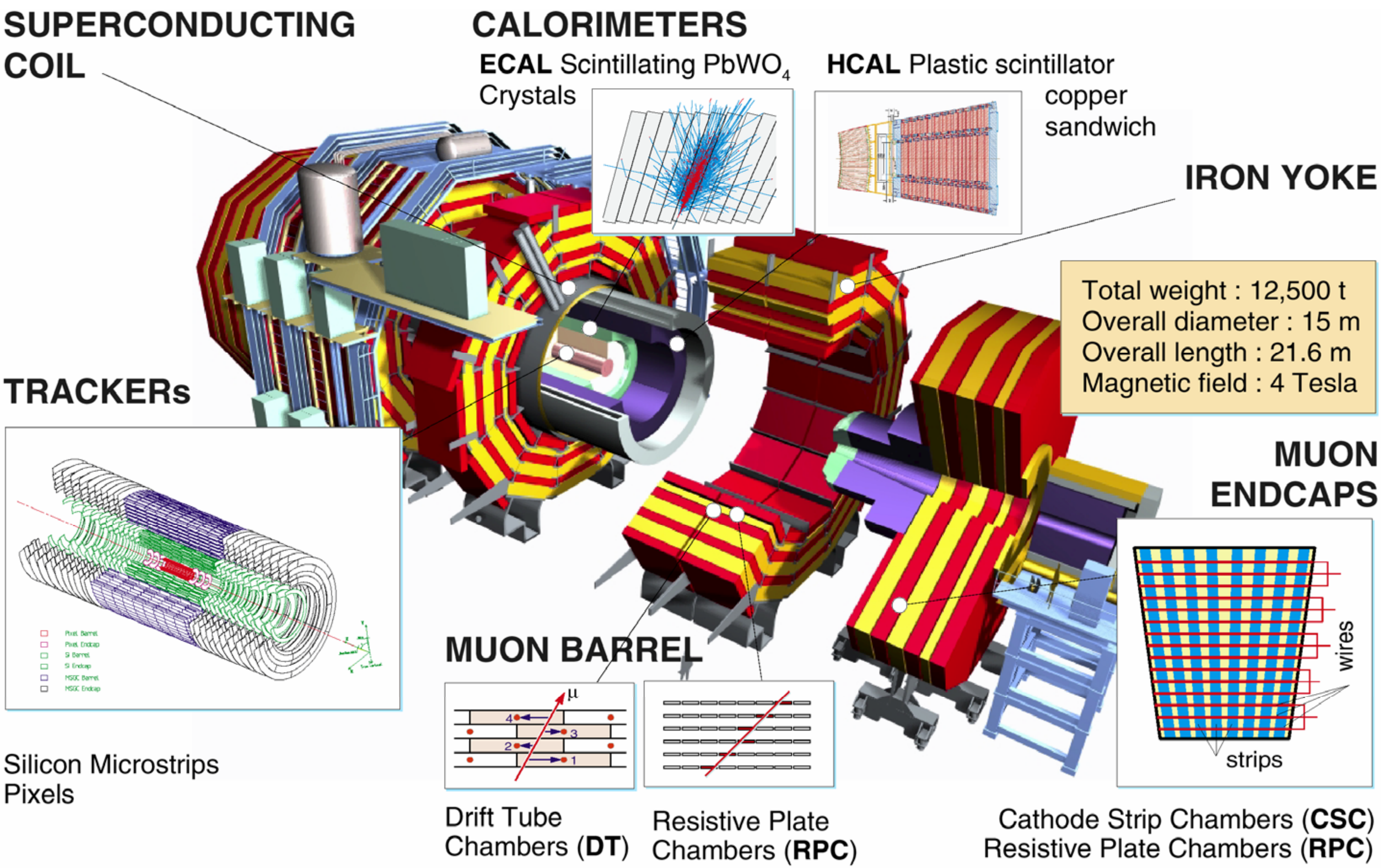
Bunch Crossing	**4 10^{7} Hz**
Proton Collisions	**10^{9} Hz**
Parton Collisions	
New Particle Production (Higgs, SUSY, ...)	**10^{-5} Hz**

Selection of 1 event in 10,000,000,000,000

LHC Challenges

- **Detector design**
  - Must have fast response:
    - otherwise will integrate over many bunch crossings (response time: 20-40 ns)
  - Must be highly granular:
    - minimize probability that pile-up particles be in the same detectors element as interesting object
  - Must be radiation resistant:
    - high flux of particles from pp collisions → high radiation environment
- **Event selection**
  - Out of a billion interactions/sec select one hundred for further analysis
    - need to reject most interactions
    - high rejection power while preserving sensitivity for rare processes
  - On-line event selection ultimately determines the physics output of the LHC experiments
- **Computing/Analysis**
  - Store data at $\approx$100 Hz; **Petabytes of data per year**
  - Large scale distributed computing

Compact Muon Solenoid Detector



SUPERCONDUCTING COIL

CALORIMETERS

ECAL Scintillating $PbWO_4$ Crystals

HCAL Plastic scintillator copper sandwich

IRON YOKE

Total weight : 12,500 t
Overall diameter : 15 m
Overall length : 21.6 m
Magnetic field : 4 Tesla

TRACKERs

Silicon Microstrips
Pixels

MUON BARREL

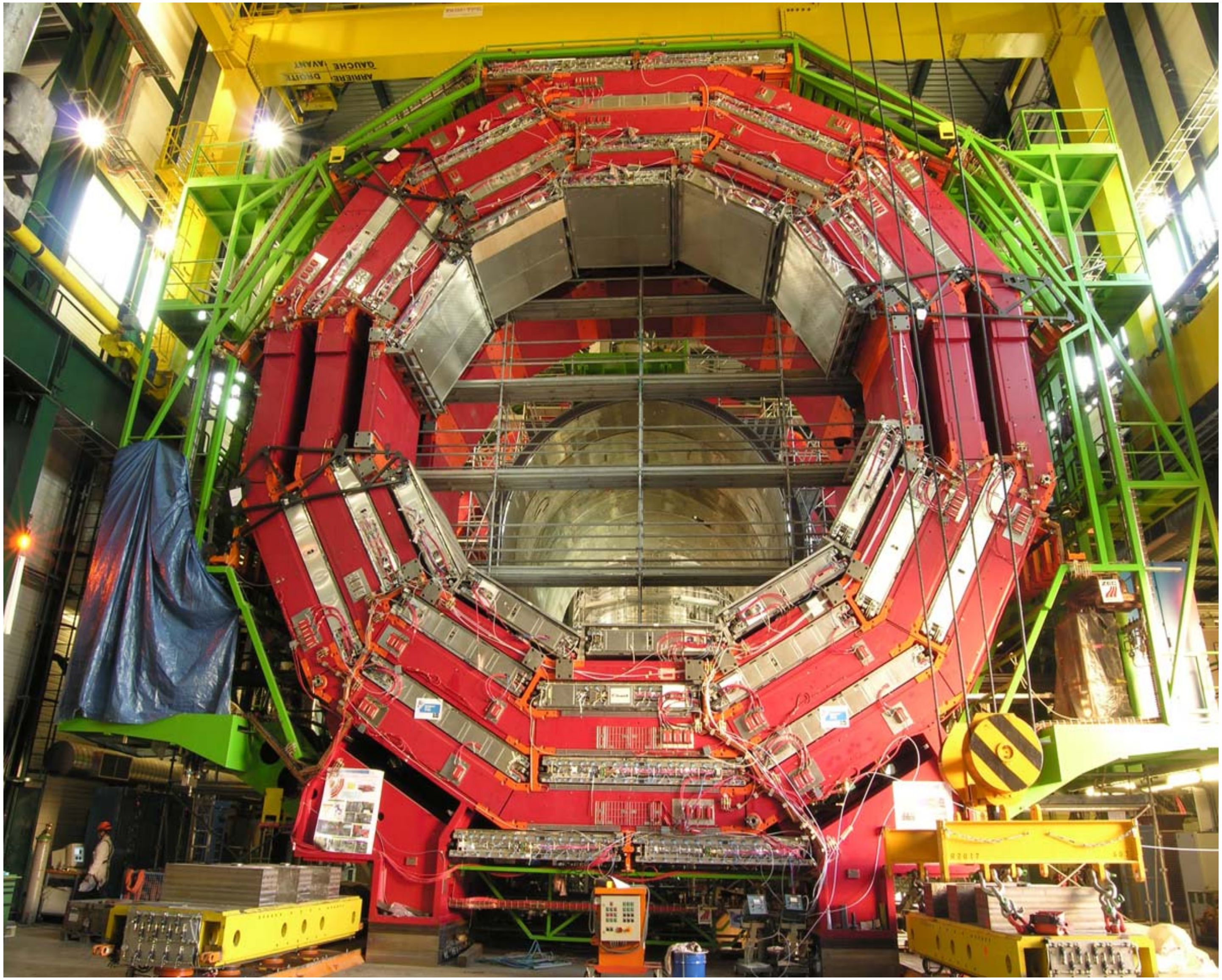
Drift Tube Chambers (**DT**)

Resistive Plate Chambers (**RPC**)

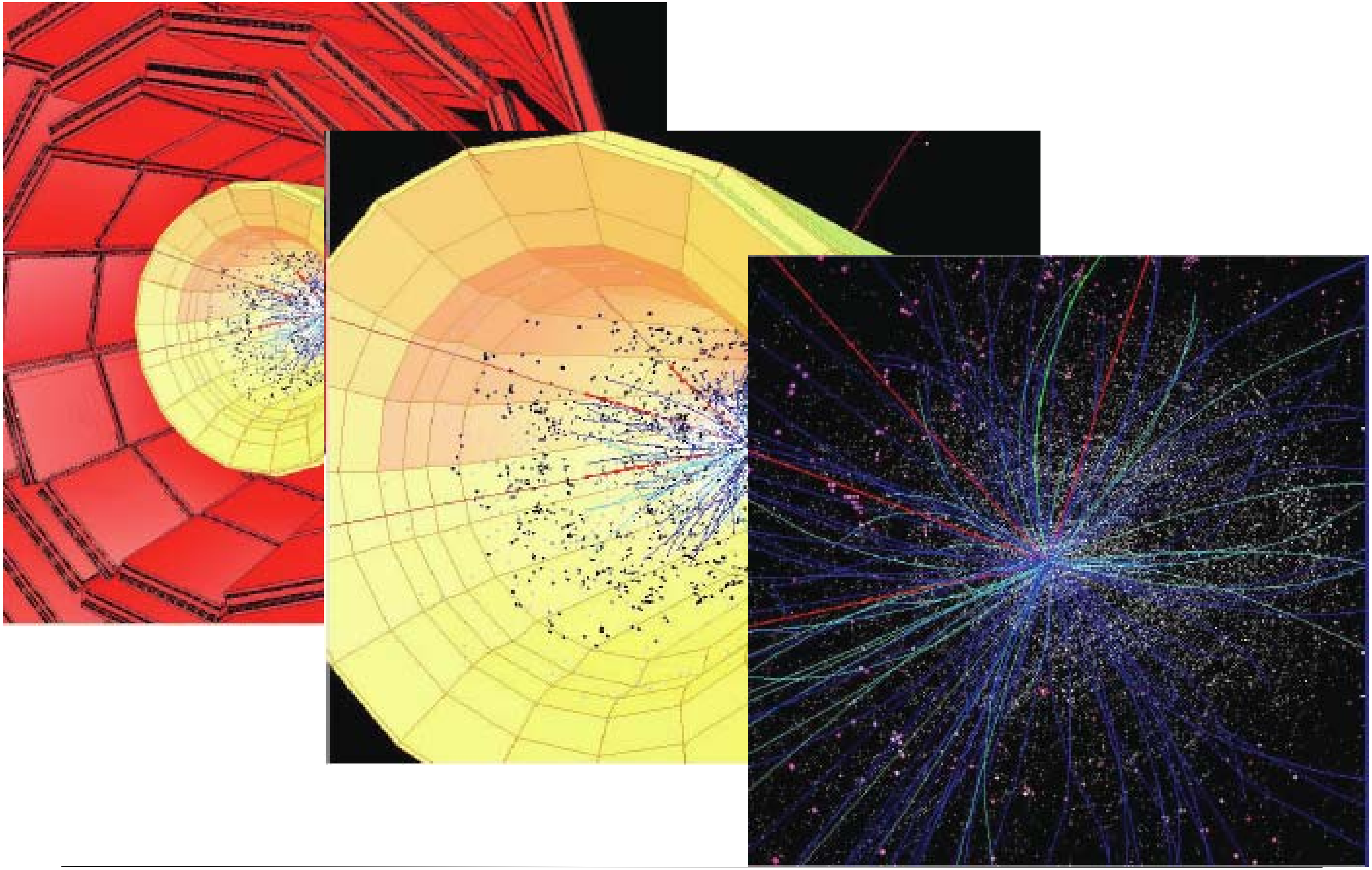
MUON ENDCAPS

Cathode Strip Chambers (**CSC**)
Resistive Plate Chambers (**RPC**)

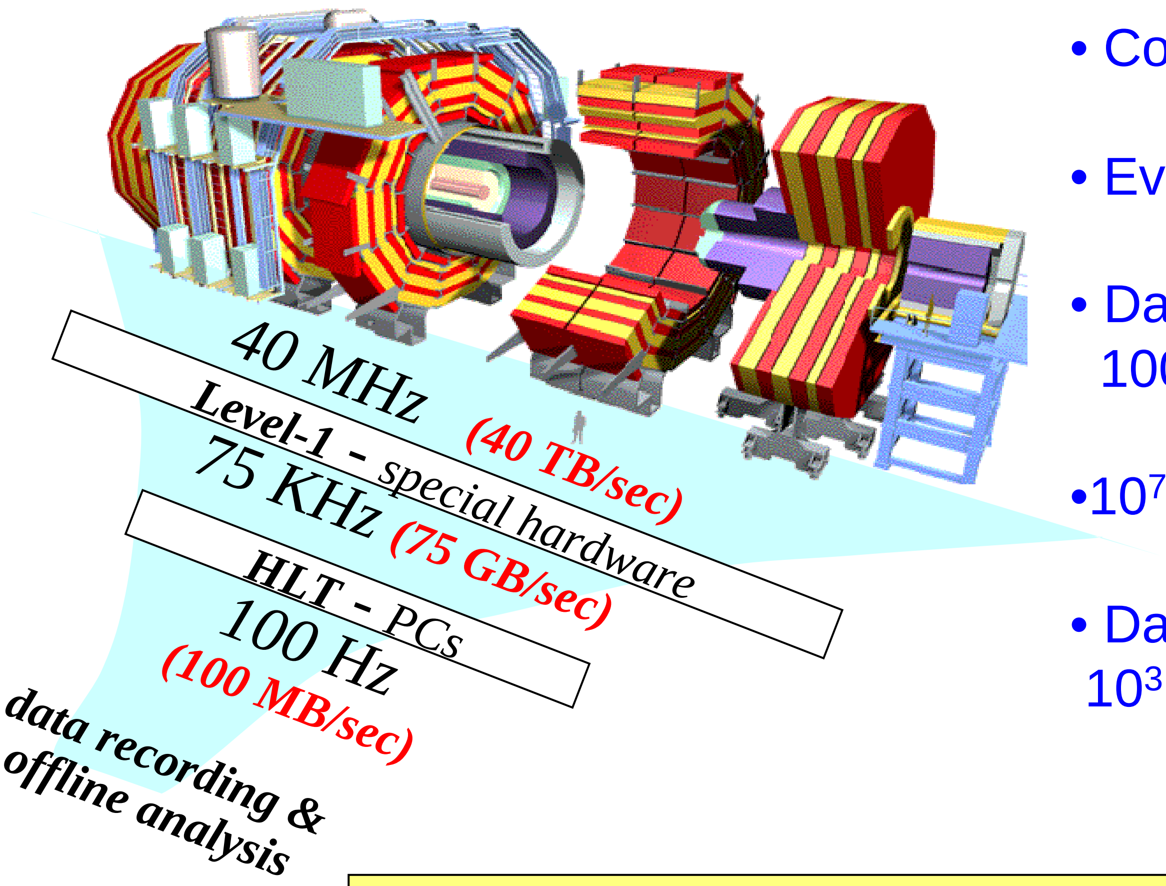
CMS Under Construction



Event Reconstruction



Data Recording



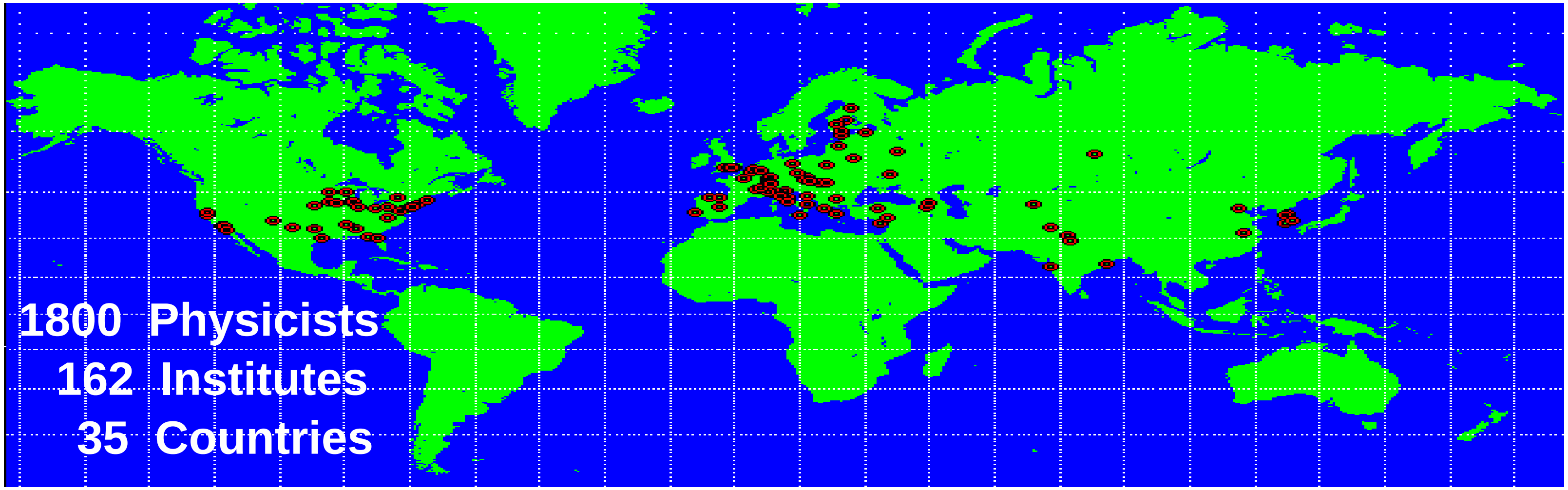
- Collision rate: 40 MHz
- Event size: ≈ 1 MByte
- Data size: 1 MByte/event
  100 events/s → 100 MByte/s
- 10^7 s data taking per year (30%)
- Data size: 1 PetaByte =
  10^3 TByte = 10^6 GByte per year

~ PetaBytes / year
~10^9 events/year
~10^3 batch and interactive users

LHC Computing

Different from Previous Experiment Generations

- Geographical dispersion: of people and resources
- Complexity: the detector and the LHC environment
- Scale: Petabytes per year of data



Major challenges associated with

- Coordinated Use of Distributed computing resources
- Remote software development and physics analysis
- Communication and collaboration at a distance

NEED FOR A NEW TYPE OF COMPUTING!

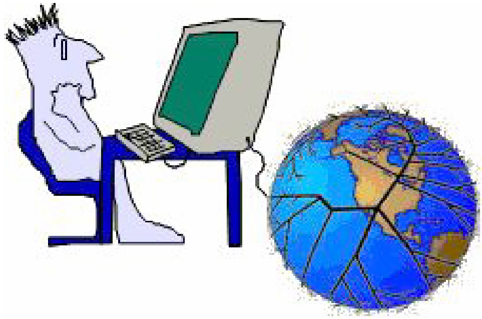
Grid: The Motivation

- **Problem:**
  - Need to store and analyze a massive amount of data
  - Need for computing power rises faster than Moore's Law
- **Possible solutions:**
  - a) Find someone to pay for additional computers
  - b) Use existing, distributed resources

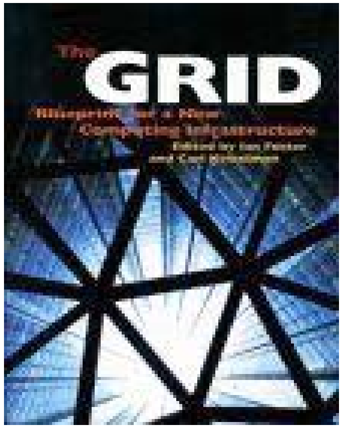
Solution b) needs *standardized framework* for distributed computing

- Problem applies to many areas, beyond the boundaries of particle physics
- Using existing, distributed resources is a good idea anyway
- Single countries can't afford multi-billion dollar investments into particle physics anymore
- Might still be willing to donate a portion of overall expenses
- But: want to spend money locally!
- More willing to invest in local computer centres than into resources elsewhere ...
- Need to bundle existing or newly established computing resources to form a coherent, standardized ensemble capable of running very large scale jobs (p.physics, weather simulation, biology, ...)

Grid: The Vision*



- Computing Power from a plug in the wall
- The world is your computer
- Analogy: Power Grid
- Seamless exchange of computing power
- Logical extension of the World Wide Web?
- "The Web on Steroids"
- A hype (which has good and bad sides ...)
- Takes distributed computing to a new level


The GRID
Blueprint for a New Computing Infrastructure
Edited by Ian Foster and Carl Kesselman

Warning:

Different people mean different things when talking about “The Grid”

*a la Foster & Kesselman

Distributed Computing

- "When the network is as fast as the computer's internal links, the machine disintegrates across the net into a set of special purpose appliances" (Gilder Technology Report, June 2000)

- "Grid technologies and infrastructures can be defined as supporting the sharing and coordinated use of diverse resources in dynamic, distributed 'virtual organisations' " (OGSA white paper)

Limitations:

- High-speed network connection (the faster, the more data can be exchanged between participating nodes).
- "Speed" means :

  a) High Bandwidth - Scales!

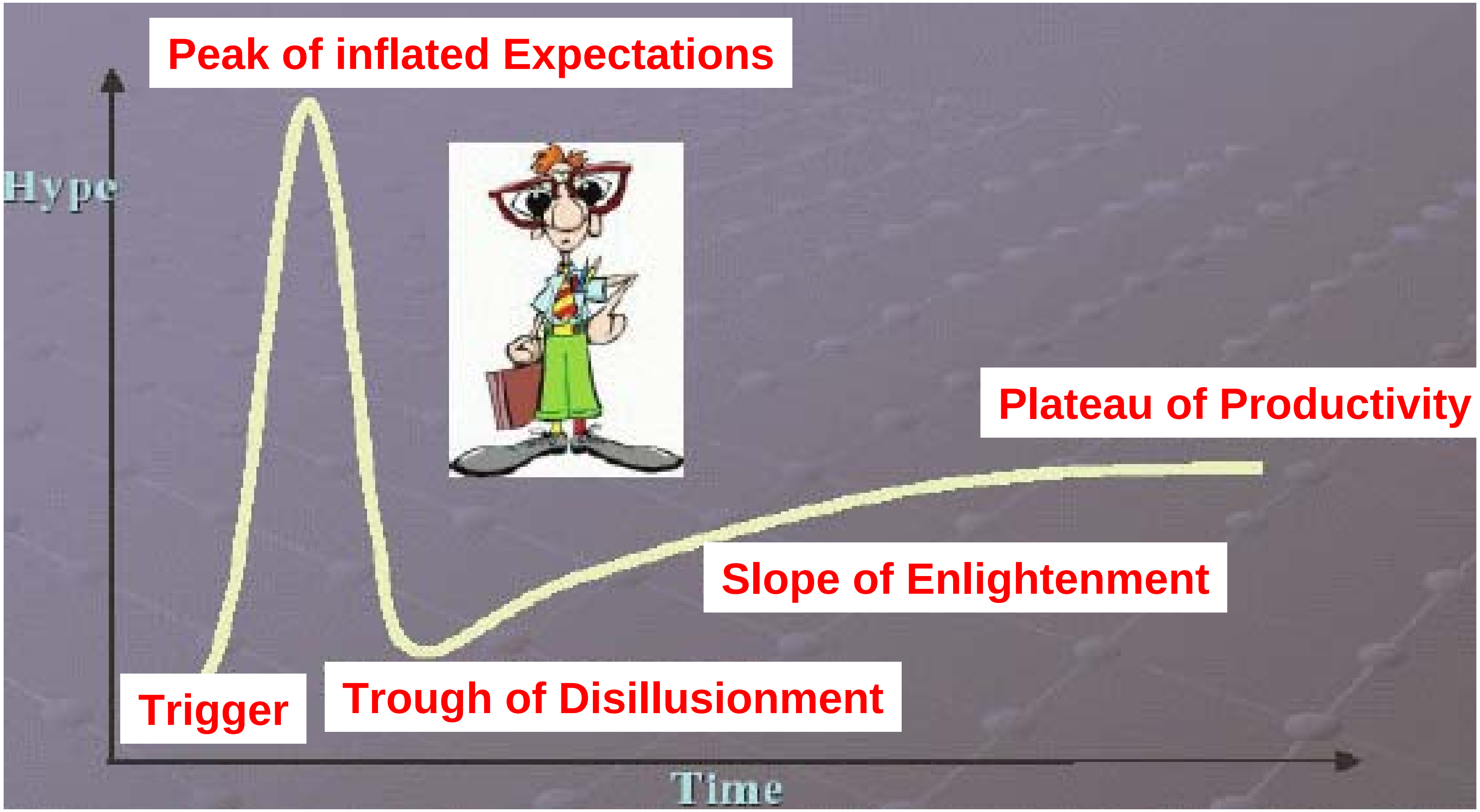
  b) Low Latency - Doesn't scale! (Latency limits possible application types)

Requirements:

- High Bandwidth / Low Latency connections are needed!
- Common middleware framework needed; Standards
- One current focus of research and development:

  from "local GRIDs" to the "World Wide GRID": GEANT, TeraGrid, etc.

The Grid



Distributed Computing is partially Social Engineering

It pays to THINK

Software Framework

- The Globus Toolkit (www.globus.org)
- Base technology for the majority of Grid projects
- Effort led by Foster and Kesselman (Argonne Lab and USC)
- Layer between Operating System and GRID application: "Protocol layer”
- Takes care of many aspects of GRID computing including the communication between the participating program fragments, authentication, security, ...

- Basic functionality: submit job → like batch submission
- Does not contain a Resource Broker
- The Globus Toolkit includes among other components:
  - GSI (Grid Security Infrastructure): Authentication
  - GASS (Global Access to Secondary Storage)
  - Uses RSL (Resource Specification Language) to specify (min. size of memory, OS, etc.)
- **When GRID computing becomes more mature, more services will migrate from the middleware into the Operating System. Needed for seamless integration!**

The Grid Services Concept

- Middleware (Software tools and API) that allows to better manage large and distributed resources, thus reducing the complexity
  - Work scheduling, data management, application monitoring, fabric management, storage management
- Basic Middleware Framework: *GLOBUS*
- Standard services that
  - Provide uniform, high-level access to a wide range of resources (including networks)
  - Address interdomain issues: security, policy
  - Permit application-level management and monitoring of end-to-end performance
  - Perform resource discovery
  - Manage authorization and prioritization
- Broadly deployed (like Internet Protocols)

Open Science Grid

Open Science Grid

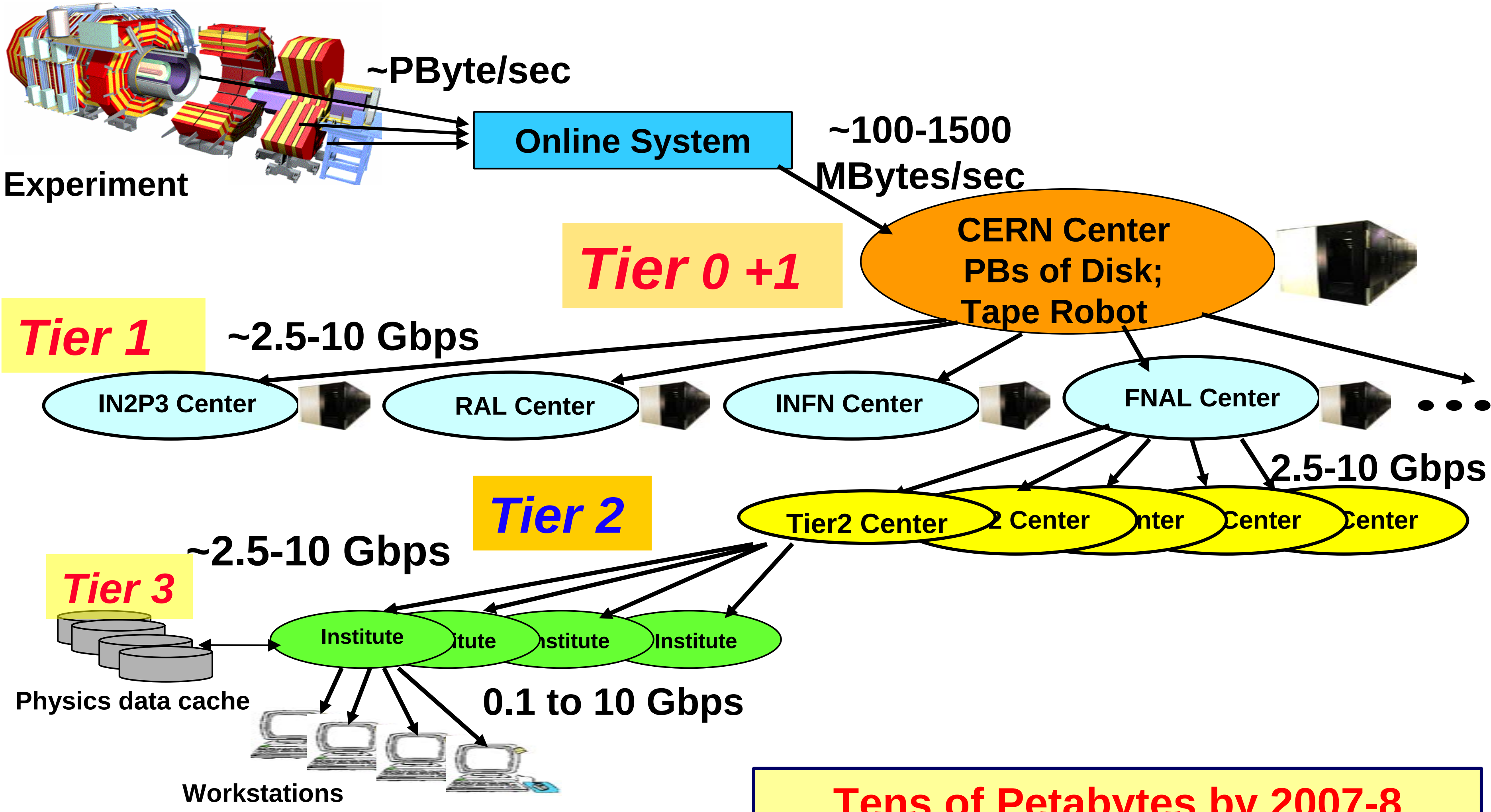
- The Open Science Grid (OSG) is the Grid under the governance of the Open Science Grid Consortium operated as a sustained and production infrastructure for the benefit of its members and users.
- Will build a National Production Grid Infrastructure in the United States, open to any Science.
- The Open Science Grid interacts with Grids external to the Consortium through federation and partnerships:
  - Build on and evolve existing infrastructures: Grid3+, shared laboratory & experiment university facilities, Campus Grids, all in close collaboration with Grid technology providers.
  - US LHC sponsors OSG contributions to the Worldwide LHC Computing Grid
- **US LHC has a key role in contributing to & leadership in the roadmap:** presenting resources to the common Grid infrastructure; developing distributed data and processing services to a common environment; running applications on the shared infrastructure; education and outreach

Worldwide LHC Computing

Regional Center Concept:

- Common, fair-shared access for Physicists everywhere
- Maximize total funding resources while meeting the total computing and data handling needs
- Balance between proximity of datasets to appropriate resources, and to the users
  - **Tier-n Model**
- Efficient use of network bandwidth: higher throughput
  - **Local > regional > national > international**
- Utilizing all intellectual resources, in several time zones
  - **CERN, national labs, universities, remote sites**
  - **Involving physicists and students at their home institutions**
- Greater flexibility to pursue different physics interests, priorities, and resource allocation strategies by region

CMS Global Data Grid



Experiment

~PByte/sec

Online System

~100-1500 MBytes/sec

Tier 0 +1

CERN Center
PBs of Disk;
Tape Robot

Tier 1

~2.5-10 Gbps

IN2P3 Center

RAL Center

INFN Center

FNAL Center

• • •

2.5-10 Gbps

Tier 2

Tier2 Center

2 Center

nter

Center

Center

~2.5-10 Gbps

Tier 3

Institute

itute

nstitute

Institute

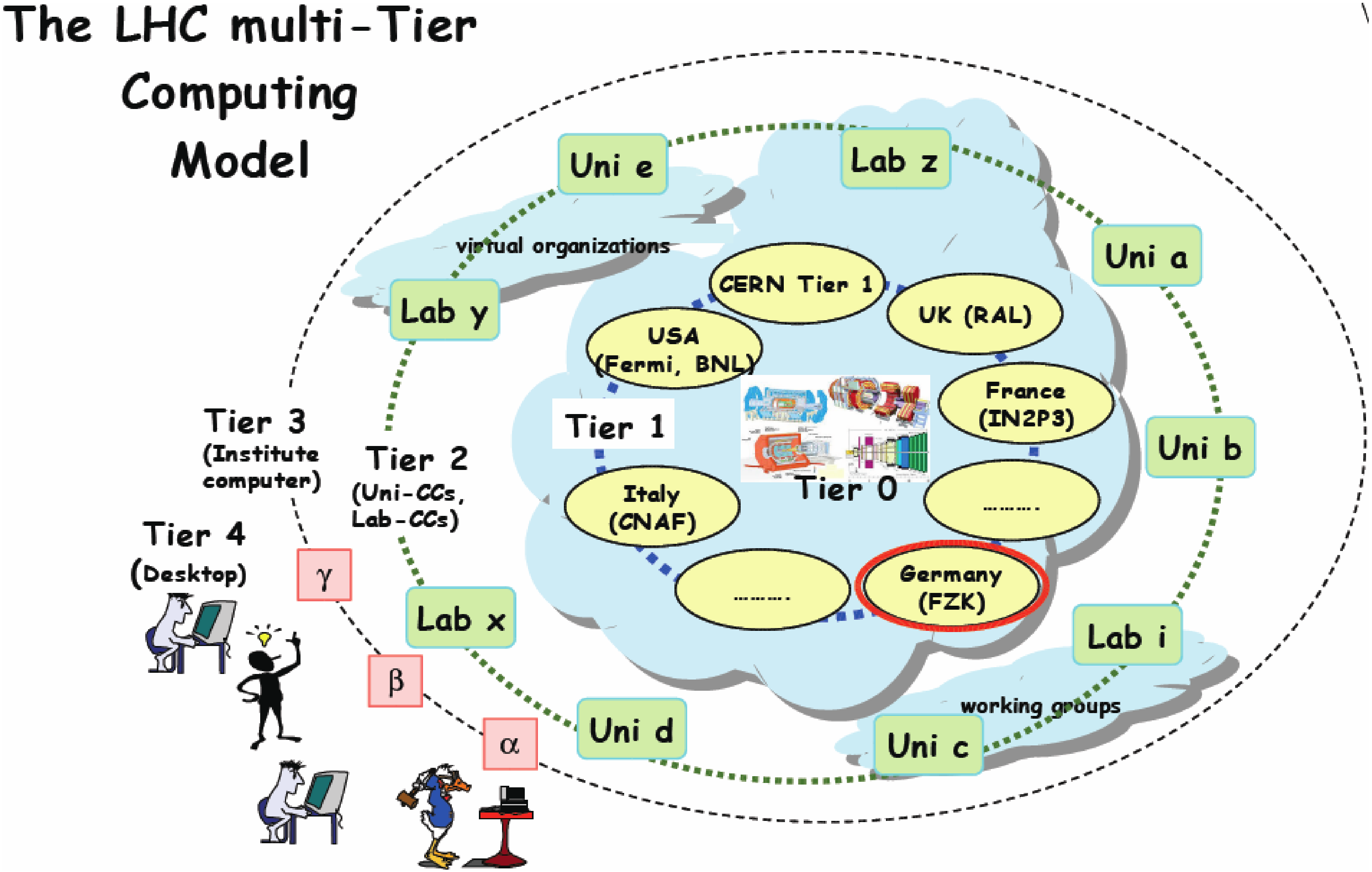
Physics data cache

0.1 to 10 Gbps

Workstations

Tens of Petabytes by 2007-8
An Exabyte ~5-7 Years later

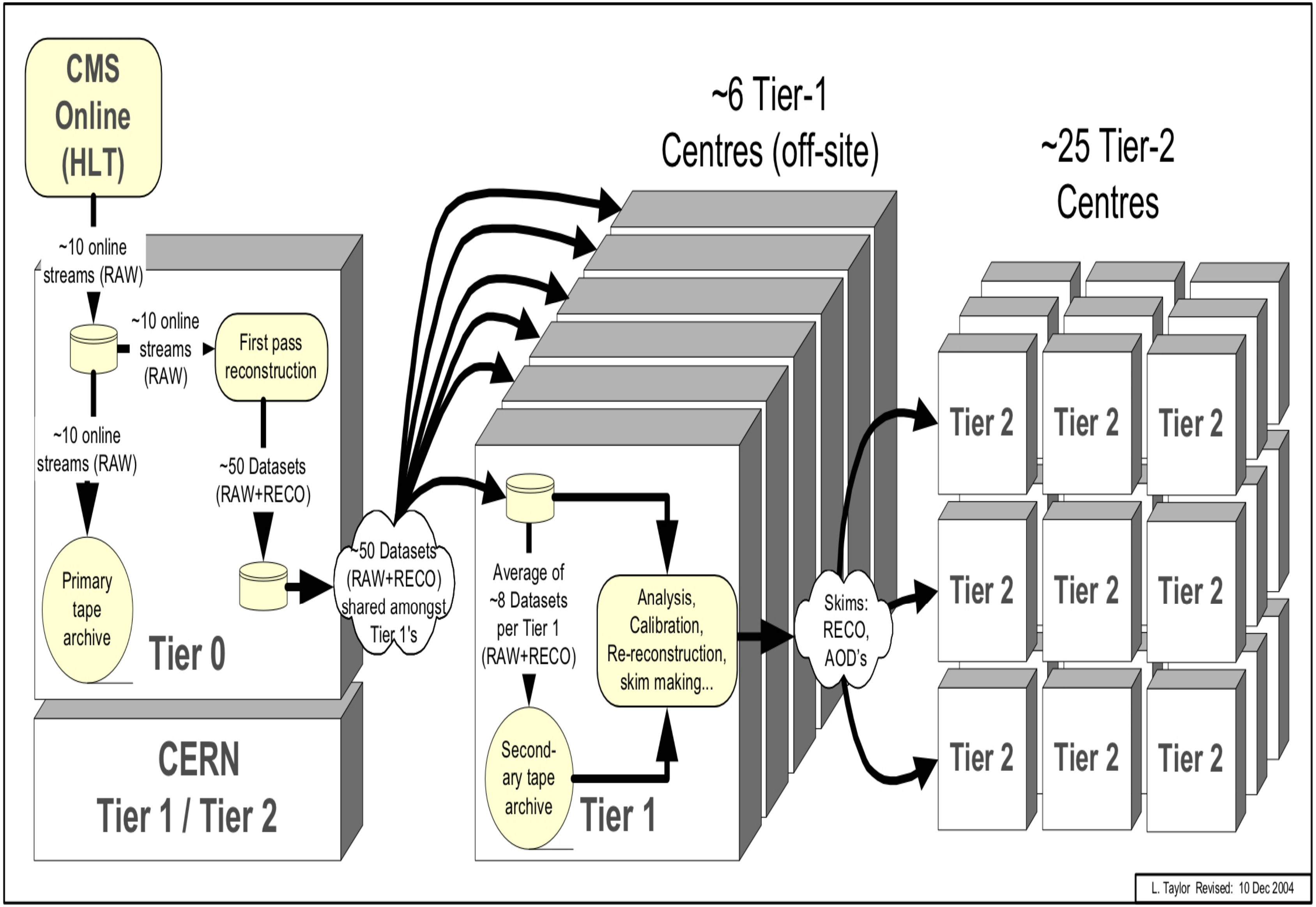
LHC Computing Model



Processing

- The Tier-0 Centre (1)
  - Calibration, alignment, reconstruction in-line with data acquisition. This facility, possibly extended by part of the online filter system, may be used for re-processing during the 100 days when the accelerator is not running.
- Tier-1 Centres (~6)
  - Re-processing of the current year's data, and selective re-processing in subsequent years of the cumulated data.
  - ESD-AOD selection and some general end-user analysis.
- Tier-2 Centres (~25)
  - Simulation, reconstruction, calibration of simulated data.
  - General end-user analysis for local communities or physics groups.

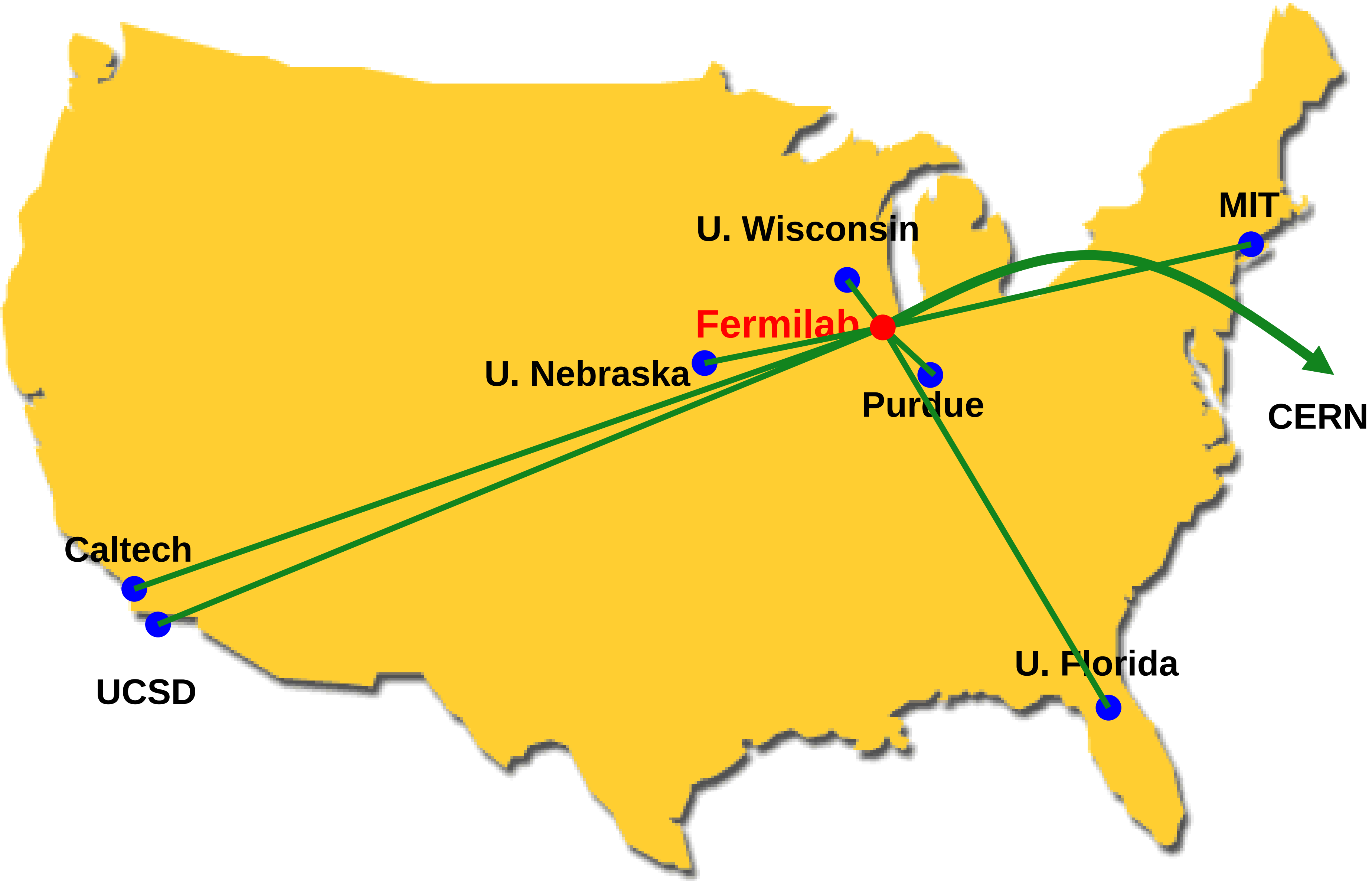
CMS Data Flow



Analysis Computing

- PetaBytes of data to be explored and analyzed by 1000s of globally dispersed scientists, in hundred of teams.
  - Analysis is distributed worldwide
- Tough problem: Many clients, unpredictable behavior, short timescales (conferences), large peaks of load.
- In addition, analysis computing normally competes to some extent with production activities
  - Tape drives, Network, Staging systems
- Technical Issues
  - Balance proximity to large computational and data facilities, vs. proximity to more local resources and users
  - Special issues: Co-scheduling of data and CPU
- Software: Complex C++ code; under constant development
- Processing: A single processor (node) analyzes one event at a time

US CMS Computing



Tier-2 Centers

- Tier-2s are at the core of empowering US universities
- Providing computing services is required, but not sufficient
  - Local excellence and proven ability to provide services
  - Take responsibility in a multi-lateral partnership with Tier-1, local physics communities and other sciences, **Open Science Grid**, LCG, middleware and infrastructure projects
- Functions:
  - Maintain on disk a copy of AODs
  - Provide disk storage for samples of user data
    - Well managed disk storage - grid enabled!
  - Simulation and calibration – Tier-2s are assumed to provide the full simulation capacity for CMS
  - Generated data is sent to a Tier-1 for permanent storage
  - Service for end-user analysis
    - Batch + interactive analysis
- Some Tier-2 centres will have large parallel analysis clusters
  - **It is expected that clusters of Tier-2 centres will be configured for use by specific physics groups.**

Tier-2 Centers

- Resources (2007):
  - CPUs: ~1000k SI2000
  - Disk space: ~250 TB
  - Network: 2.5 - 10 Gigabit/sec
  - Tape storage does not need to be provided
- Regional Center
  - Serve a community of CMS users
  - Member of a Distributed Computing Fabric (for analysis)
    - Be part of CMS infrastructure with agreed upon level of services - both Quantity and Quality
    - High throughput production computing is well understood
    - High throughput distributed analysis support is the challenge
- Resource Management
  - Tier-2 must manage data and CPU resources as part of the distributed computing fabric under both computing and physics analysis hierarchies

Tier-2 at Purdue

- In January 2005 Purdue was selected by US-CMS to build and operate a Tier-2 center.
  - HEP group in partnership with ITaP
- Purdue was chosen as one of ten proposals from the ~40 US-CMS university groups.
- NSF funded for 5 years

- **Future Options:**
  - Can be expanded into a regional Purdue - Notre Dame center
  - Synergy with CMS group at Notre Dame
  - Attach local “super” Tier-3 center

Tier-2 at Purdue

- Strategic partnership between Physics and ITaP
  - Take advantage of existing infrastructure (computing and networking) and operations expertise
- Collaboration with leading TeraGrid team
  - Provide link between Open Science Grid and TeraGrid
- Complement ongoing CMS efforts
  - Focus on Software and Computing
- Provide analysis facilities and data for specific physics topics
- PIs:
  - Norbert Neumeister (Physics)
  - Sebastien Goasguen (ITaP)
- Personnel:
  - Mike Shuey 50% project manager
  - Chris Baumbauer 100% Grid software deployment
  - 50% storage engineer to be recruited
  - Preston Smith 25% + 100% cost shared post doc

Physics at a Tier-2

- Disk Resident Analysis files
  - Selected DST
  - More compact formats
  - Special analysis datasets…
- Up-to-the-minute Physics Analysis Software
- Computing resources which can be rescheduled daily
  - Analysis Group
  - Analysis sub-groups
- Analysis support for Tier-3
  - Lightweight interactive analysis tool works with T2 data and remote display.
- Guaranteed Service Quality and Quantity
  - Manage dedicated production tasks versus opportunistic (dynamic) analyses services
  - Different qualities of service for different tasks

Tier-2 Challenges

- Operating the simulation batch queue is an understood problem
  - Scale of the centers is fairly modest
- Operating the storage is going to require expertise
  - 200TB of storage (even in 2007) is a substantial operations load
  - Infrastructure to turn 200TB of disk into 200TB of useful storage is not trivial to operate
- Making efficient use of networking
  - Lots of development to leverage, but succeeding at this is hard work
- Supporting Analysis
  - This is a challenge requiring technical and operational development
  - Environment work
  - Policy work
  - Support work

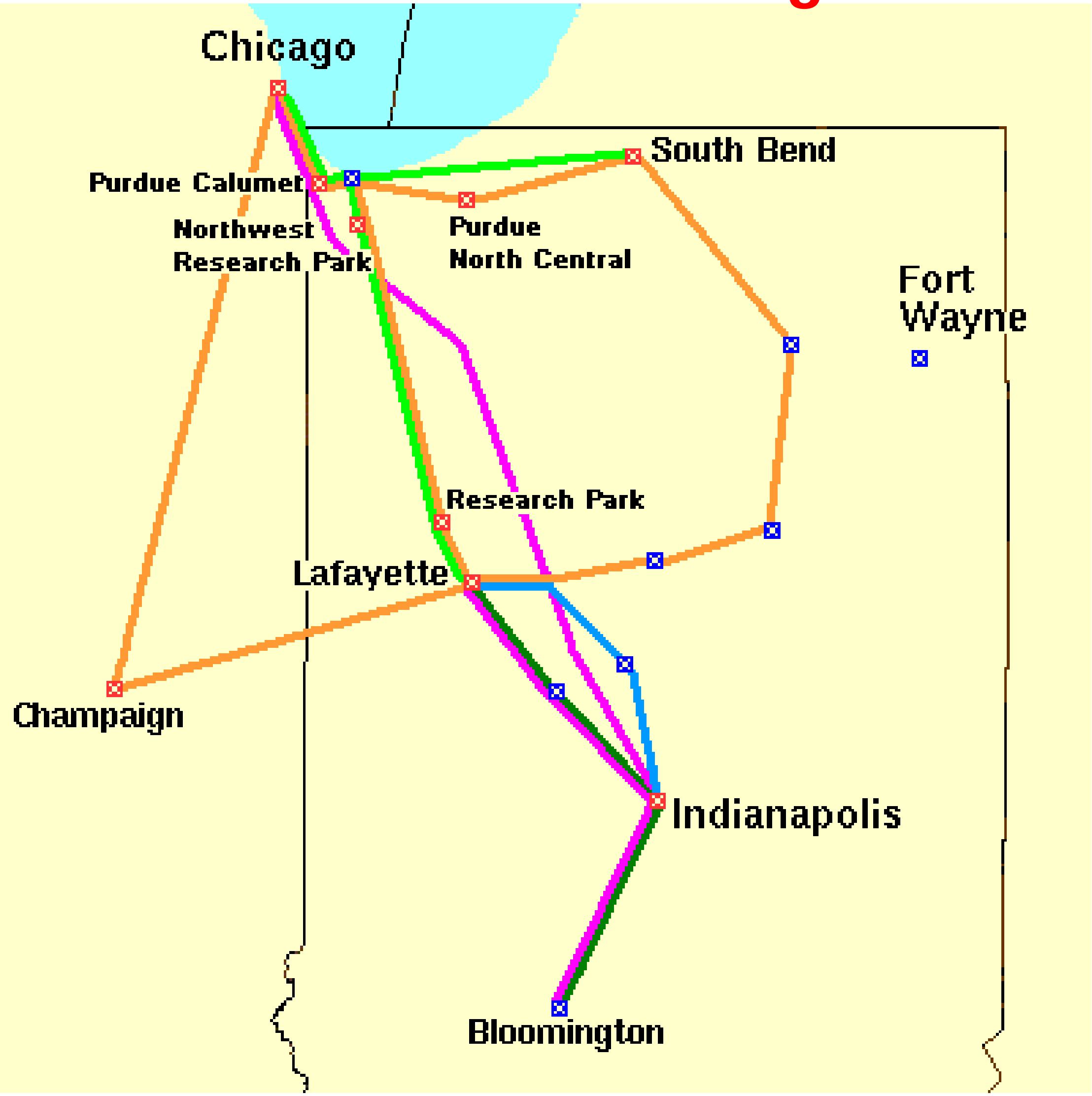
Tier-2 Storage

- **The Tier-2 storage is seen as a true cache**
  - No data that is time consuming to reproduce should be exclusively stored on the Tier-2 disk
  - Flexibility of usage drives the network requirements
- Tier-2 centers should strive for 2.5Gb/s to 10Gb/s by the start of the experiment
- The networking available at the Tier-1 needs to support Tier-2 transfers
- **Data management task is vital**

Network Speed	Time to completely cycle Disk Storage
1 Gb/s	20 days
2.5 Gb/s	8 days
10 Gb/s	2 days

Network

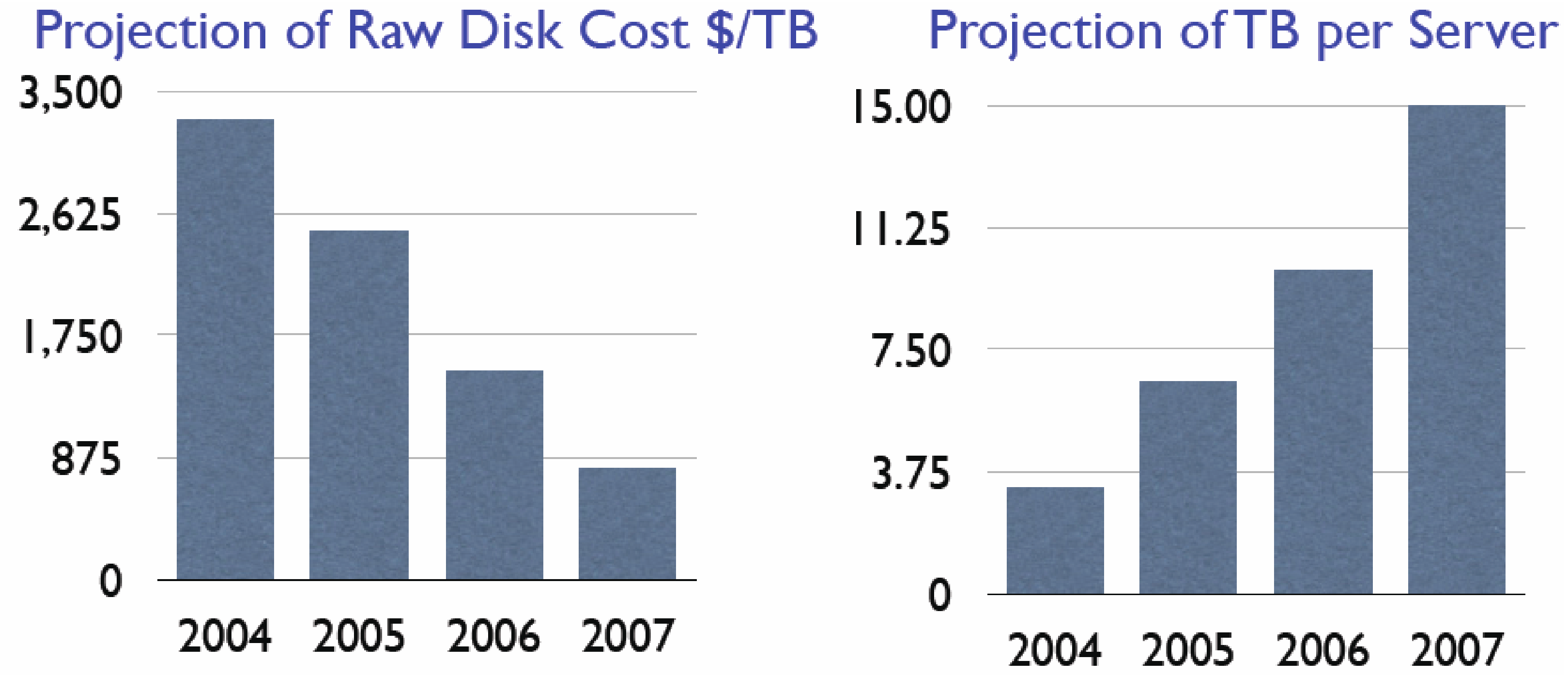
Need connection to StarLight



Storage

Data Space: The default solution is a dCache Storage Element

User Space: IBM TotalStorage



Projection of Raw Disk Cost $/TB

Projection of TB per Server

Data disks have dropping costs per raw TB and the associated servers can serve more TB per unit in 06 and 07. Industry contacts indicate this disk scaling will continue

Summary

- LHC will provide the next big leap in particle physics
  - A number of severe challenges as a result
- Huge amounts of data
  - Event selection and classification
- Need a new type of computing
  - Based on Grid technology
  - Multi-PetaByte scale and resources distributed worldwide
  - Development of a robust PB-scale networked data access and analysis system is mission-critical
- Purdue:
  - Tier-2 center is a step for Purdue to become a major player in the CMS computing project and an essential element to establishing scientific Grid Computing at Purdue.