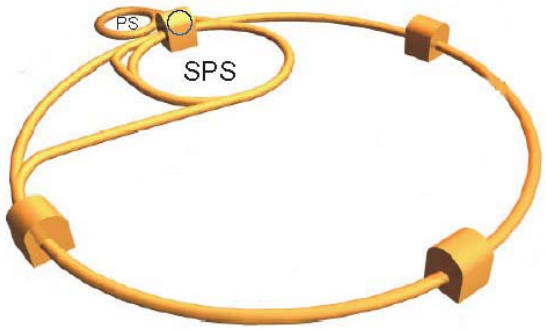
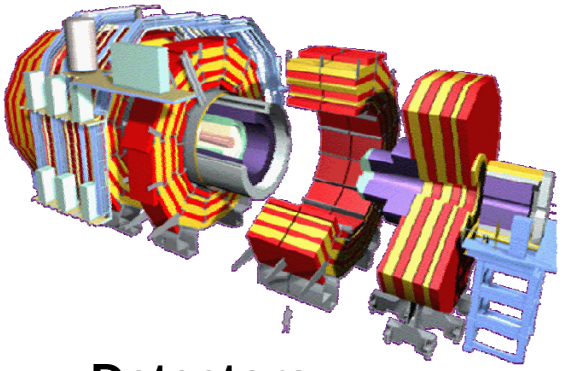
# Computing Strategien am CERN
# oder
# Kleinste Teilchen und riesige Datenmengen

Dr. Bernd Panzer-Steindel
Dr. Andreas Hirstius
September 2007

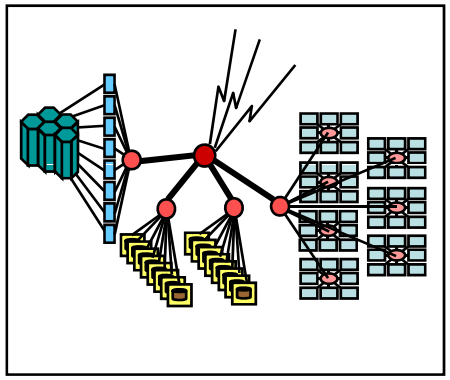



Particle Accelerators
LHC Large Hadron Collider

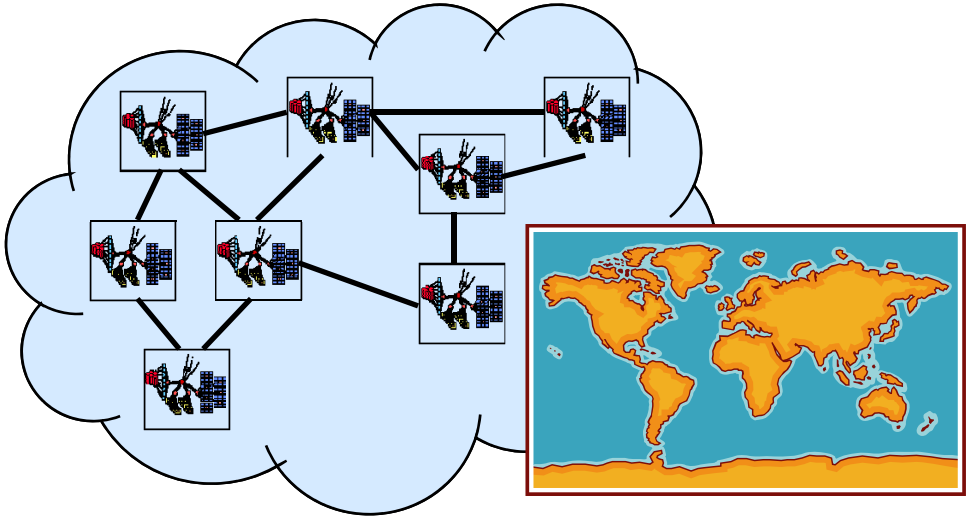


Detectors

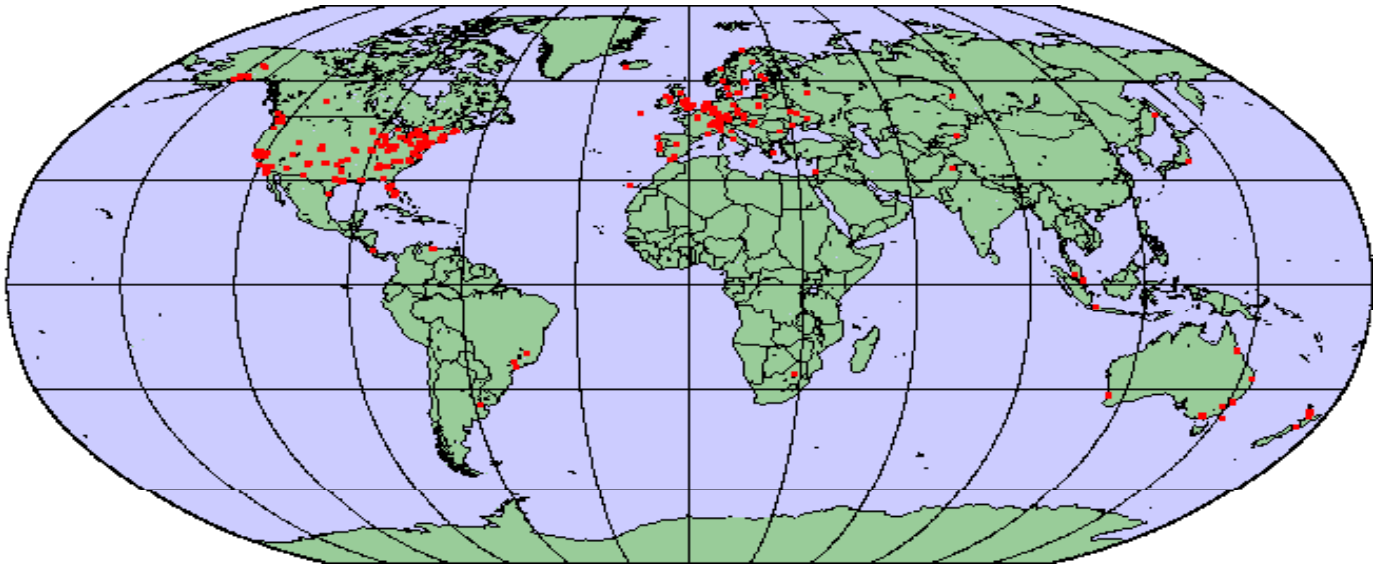
# The Tools of High Energy Physics



Local Computing Center



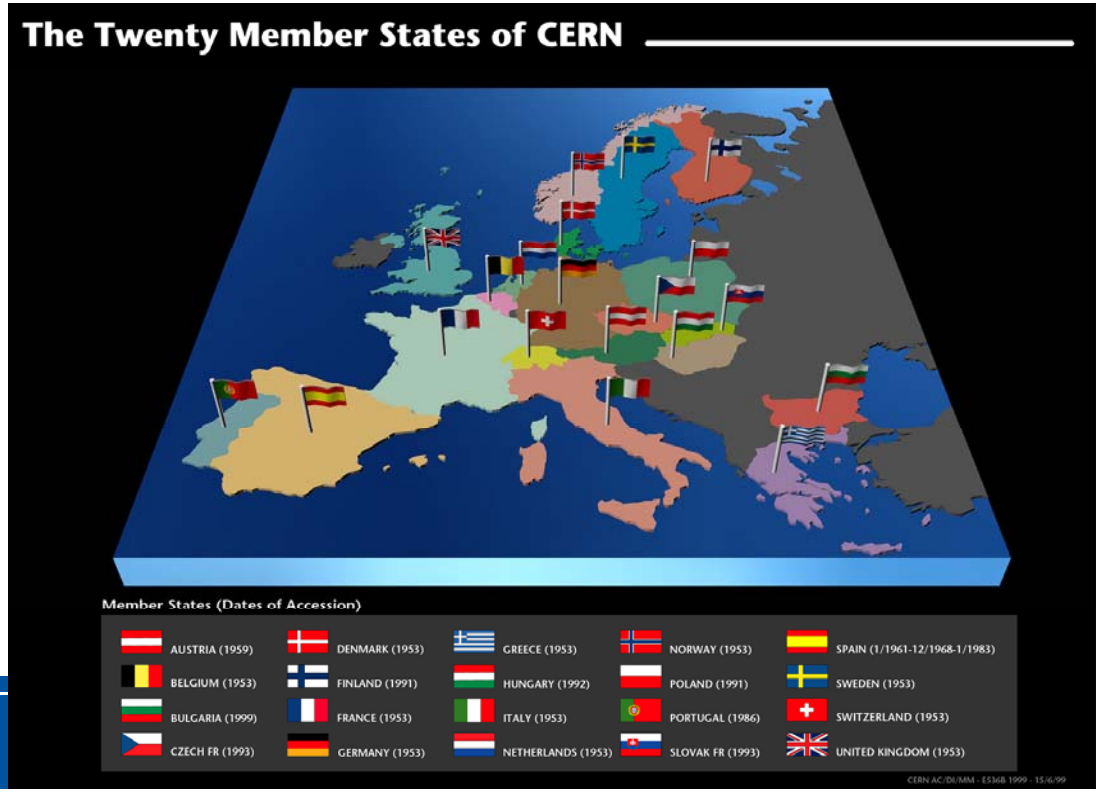
GRID, world-wide collaboration of Computer Centers



***Europe:***
***296 institutes***
***4716 users***

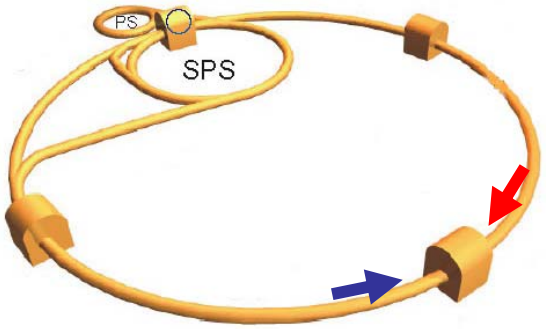
***Elsewhere:***
***224 institutes***
***2059 users***

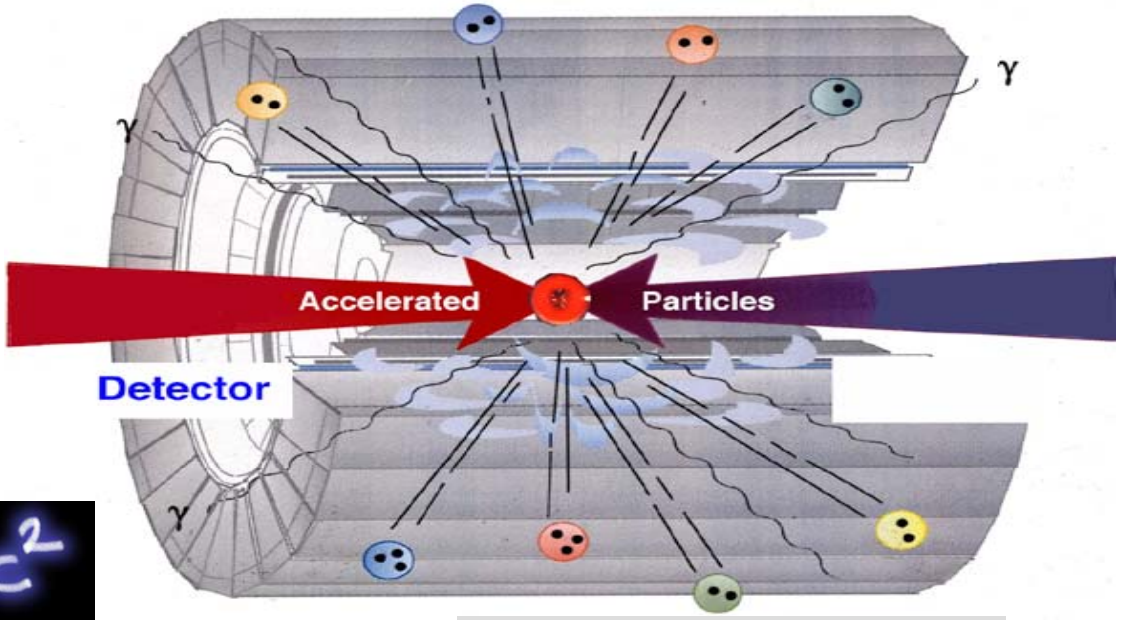
**CERN has some 6,800 visiting scientists from more than 500 institutes and 80 countries from around the world**



# Particle Accelerator

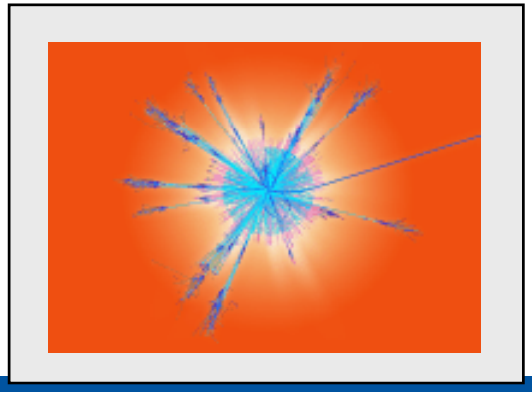
## *The most powerful microscope*



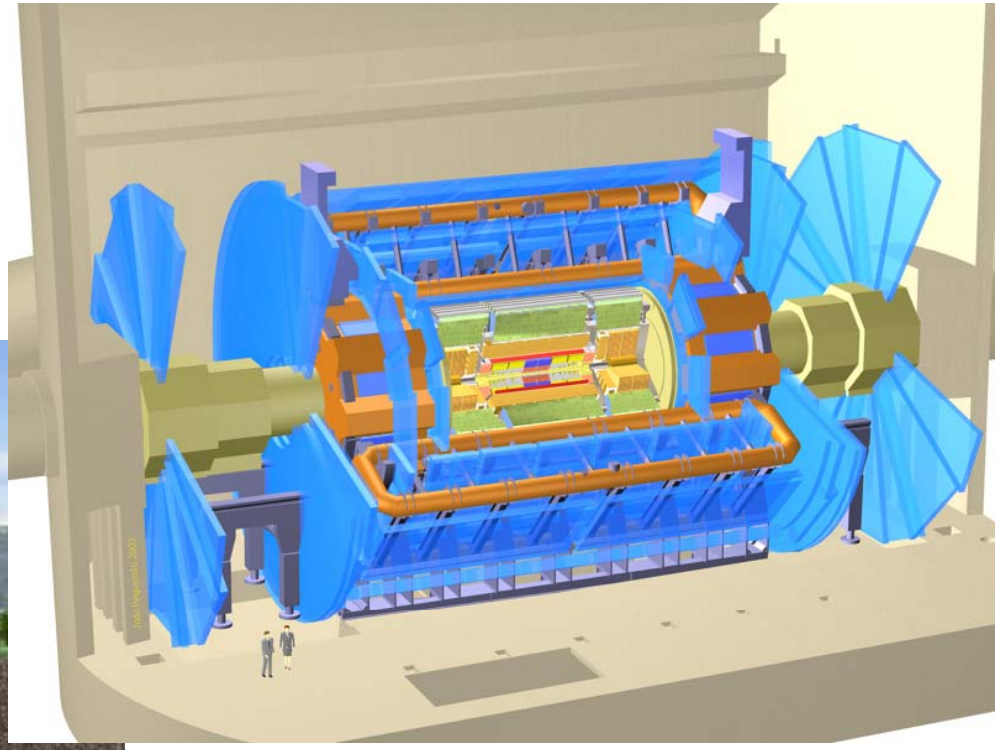


**‘snapshot of nature’**

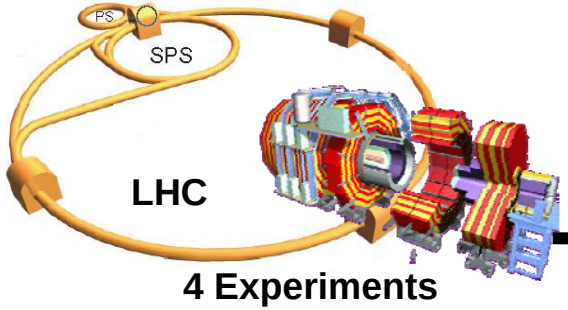


***Creating conditions similar to the Big Bang***

# The ATLAS Experiment

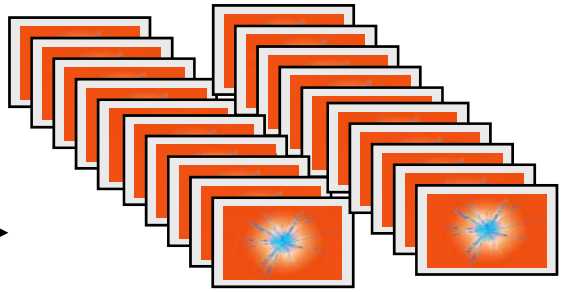


| | |
|---|---|
| *Diameter* | *25 m* |
| *Barrel toroid length* | *26 m* |
| *End-wall chamber span* | *46 m* |
| *Overall weight* | *7000 Tons* |
| *Electronic channels* | *150 million* |




4 Experiments

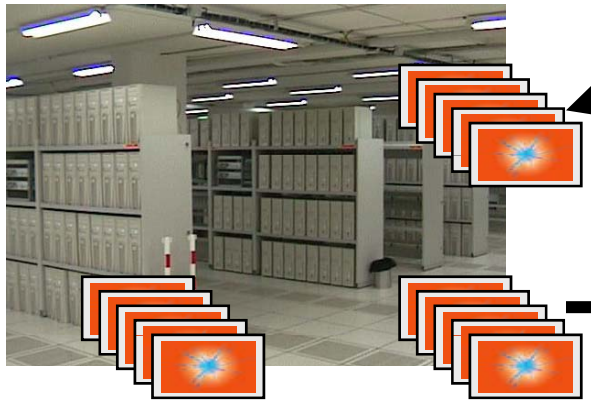
1000 million
'snapshots of nature'
(=events) per second



Filter and first selection

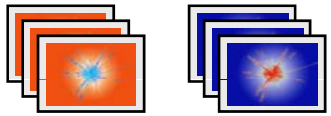
**We are looking for 1 'good' snapshot in 10 000 000 000 000 'photos'**

800 selected 'snapshots'
per second (= 1 CD, 800/s)
to the CERN computer center



# The Dataflow

Create sub-samples



World-Wide Analysis

100000/s

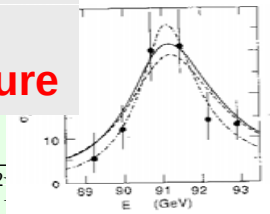
1000/s

1600/s

Store on disk and tape

Export copies

**Physics**
**Explanation of nature**



$$\sigma_{f\bar{f}} \approx \sigma^0_{f\bar{f}} \times \frac{s\Gamma_Z^2}{(s-m_Z^2)^2+s^2}$$

$$\sigma^0_{f\bar{f}} = \frac{12\pi}{m_Z^2}\frac{\Gamma_{ee}\Gamma_{ff}}{\Gamma_Z^2} \quad \text{and} \quad \Gamma_{ff} = \frac{G_F m_Z^3}{6\pi\sqrt{2}} \times (v_f^2 + a_f^2) \times N_{col}$$

1 Byte = one letter (A,B,C,..,Z) or one number (0,1,2,..,9)

100 Byte = one SMS

1.000 Byte = 1 Kilobyte = one E-Mail

1.000.000 Byte = 1 Megabyte = one book or 10 minutes phone-call
one photograph or one LHC 'snapshot'

40.000.000 Byte = 40 Megabyte = 10 MP3 songs

700.000.000 Byte = 700 Megabyte = 1 CD-ROM (music or PC-game)

1.000.000.000 Byte = 1 Gigabyte = 1 second data flow of the LHC experiments

5.000.000.000 Byte = 5 Gigabyte = one DVD

1.000.000.000.000 Byte = 1 Terabyte = library with 100.000 books

200.000.000.000.000 Byte = 200 Terabyte = 10 billion Web-pages or the size of the
American Library of Congress

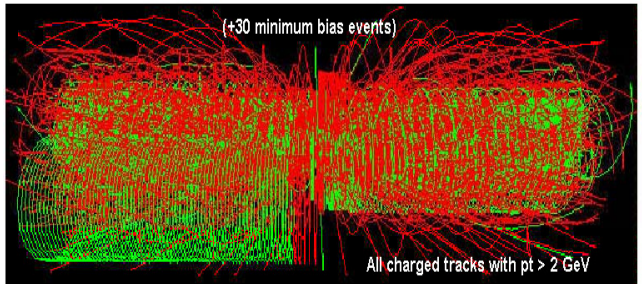
1.600.000.000.000.000 Byte = 1,6 Petabyte = world-wide produced information on paper
(newspaper, books,…)

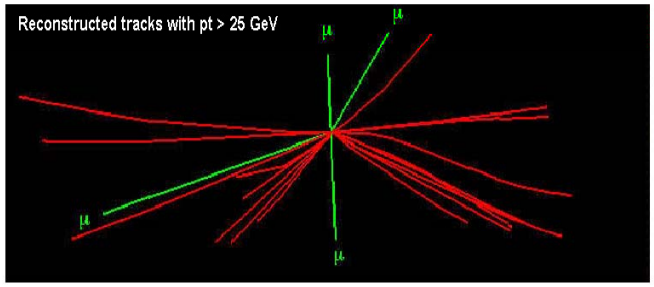
10.000.000.000.000.000 Byte = 10 Petabyte = yearly produced amount of LHC data

5.000.000.000.000.000.000 Byte = 5 Exabyte = yearly world-wide information from Radio,
TV , satellite-pictures, books, newspaper, etc.

17.000.000.000.000.000.000 Byte = 17 Exabyte = yearly information in telephone calls

# Tasks




simplify the ‘snapshots’ and extract the physics

all ‘snapshots’ are independent of each other
→simplifies the computing , ‘embarrassingly parallel’

there is lot’s of ‘noise’ to be subtracted
→need to understand precisely the environment (detectors, accelerator, etc.)
is the measured effect really the underlying basic physics or just
an artifact from the measurement itself ?!

# Die Charakteristika von Physik Computing - HEP

- Die Ereignisse sind statistisch unabhängig
  - trivial (lies: einfach) zu parallelisieren
- Weitaus größter Teil der Daten ist read-only
  - Es gibt “Versionen” anstatt “Updates”
- Meta-Daten in Databanken, Physik Daten in “einfachen” Dateien
- Die Rechenleistung wird in SPECint gemessen (anstatt SPECfp)
  - Aber Fließkommaleistung ist sehr wichtig
- Sehr große Gesamtanforderungen:
  - Datenmenge, Input/Output Anforderungen (Netzwerk, Disk, Tape), Anforderungen an die Rechenleistung
- “Chaotische” Auslastung –
  - Forschungsumgebung - Physikergebnisse werden von Gruppen von Physikern durch iterative Analyse gewonnen

→ Unvorhersehbar → praktisch unbegrenzte Anforderungen

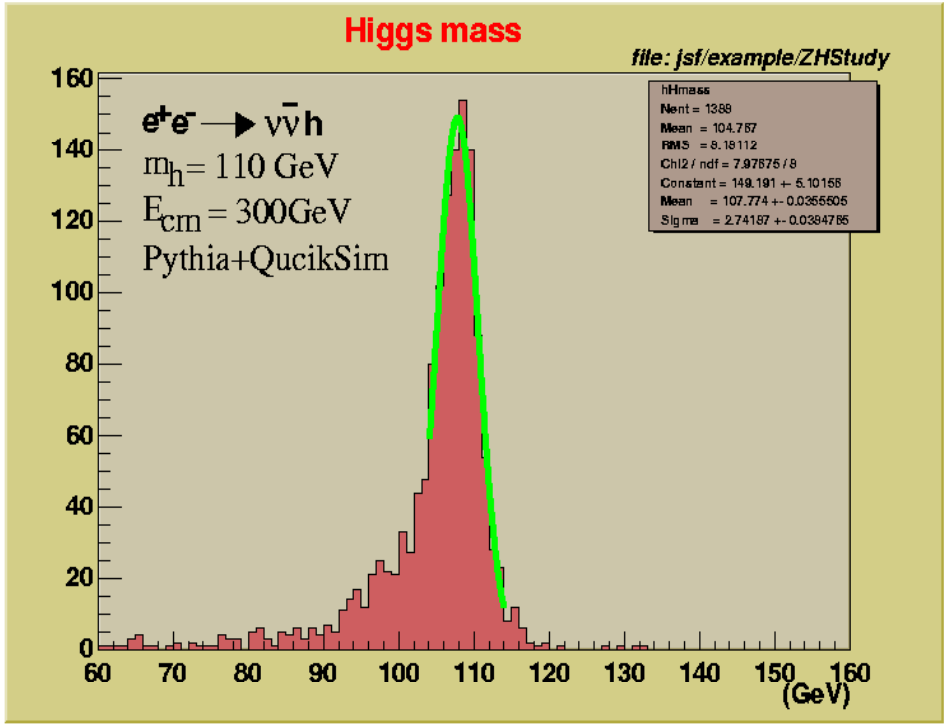
# Die Physik Software

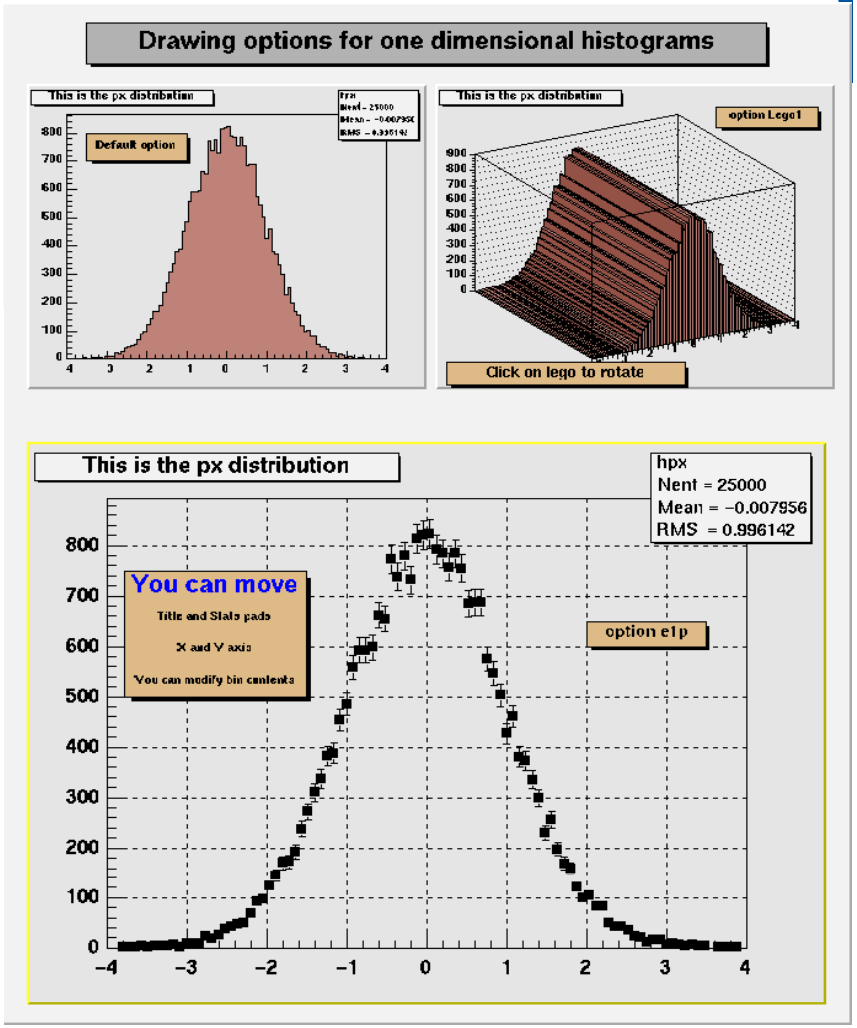
- Die Programme die von den Physikern benutzt werden lassen sich in zwei Kategorien einteilen
  - die **Datenanalyse**
  - die (Monte-Carlo) **Simulation**.

- Es wurden sogenannte “Frameworks” für die verschiedenen Gebiete entwickelt
  - ***GEANT4*** für die (Detektor-) *Simulation*
  - ***ROOT*** für die *Datenanalyse*
  - ***CLHEP*** eine C++ Klassenbibliothek

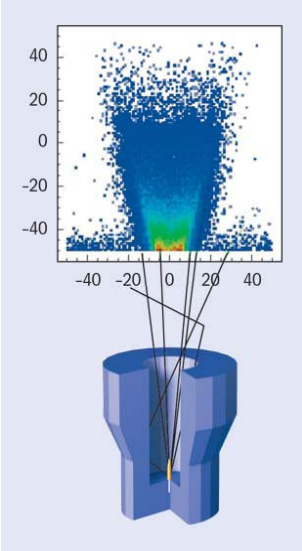
- Die Experimente bauen auf diesen Frameworks auf und entwickeln die experiment-spezifischen Software Packete.
  - Mehrere Millionen Zeilen Quellcode
  - Insgesamt hunderte “Entwickler”: Profis, Diplomanten, Doktoranten, usw.

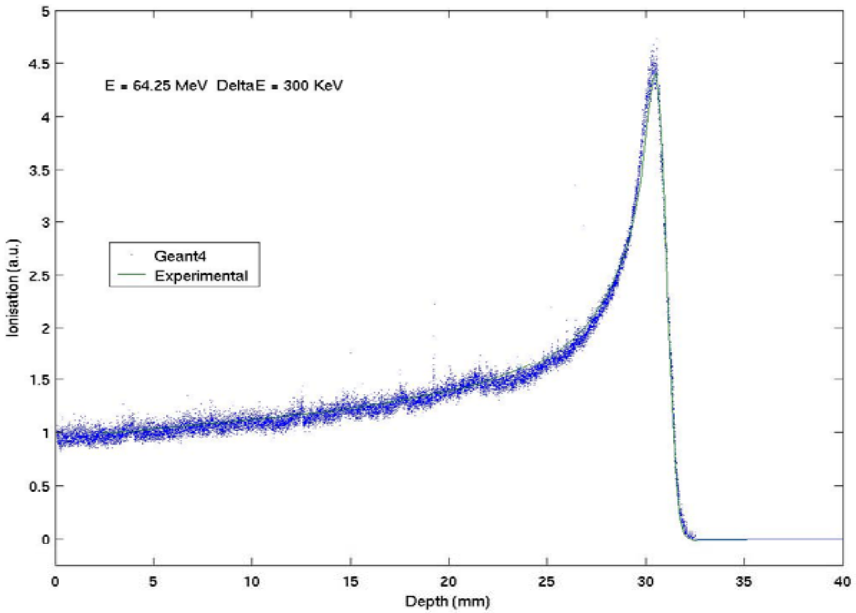
# ROOT – Die Datenanalyse



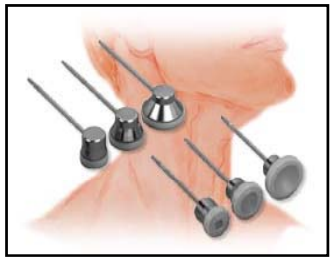



# GEANT4 – Anwendungen in der Medizin



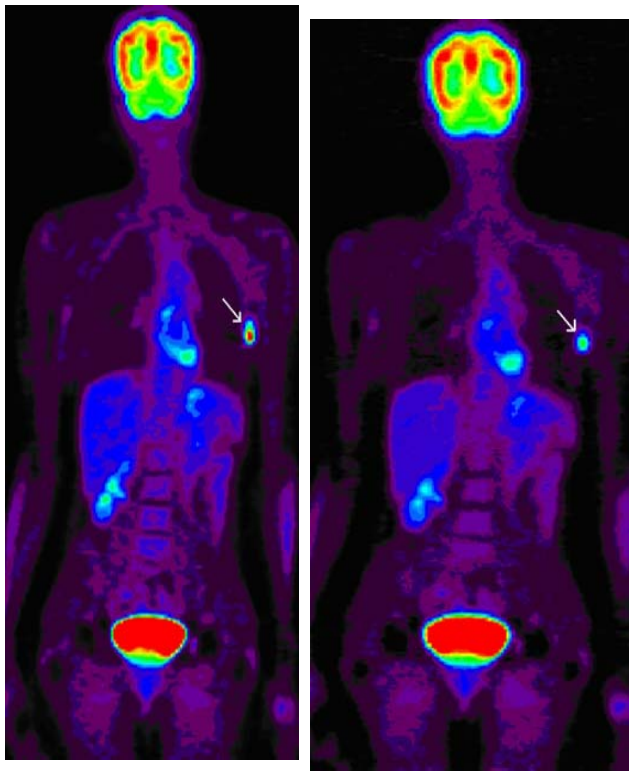




Hadrontherapie: Vergleich experimenteller Daten mit der GEANT4 Simulation



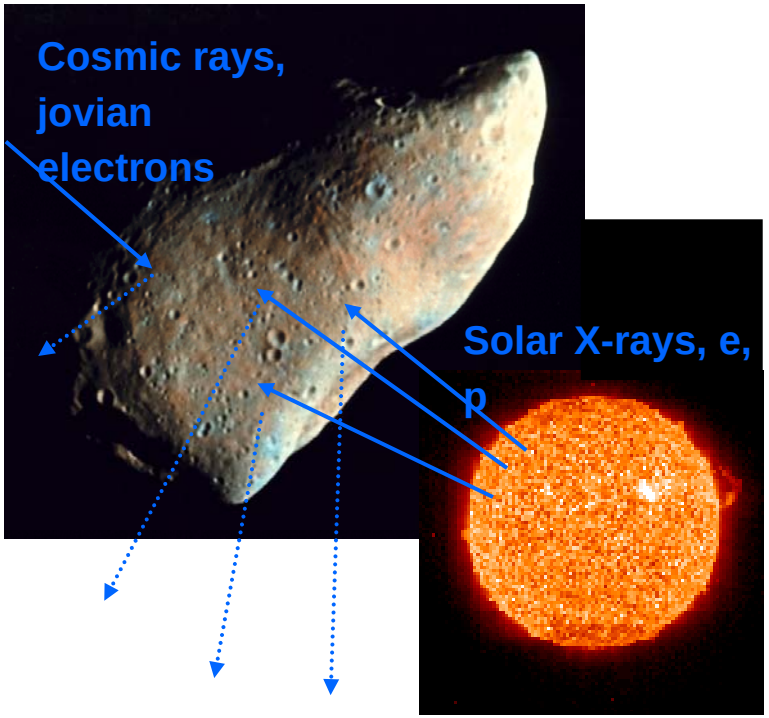
Brachytherapie Vorrichtung und Simulation der resultierenden Verteilung der Dosis

## Simulation verschiedenster Radiotherapiemethoden

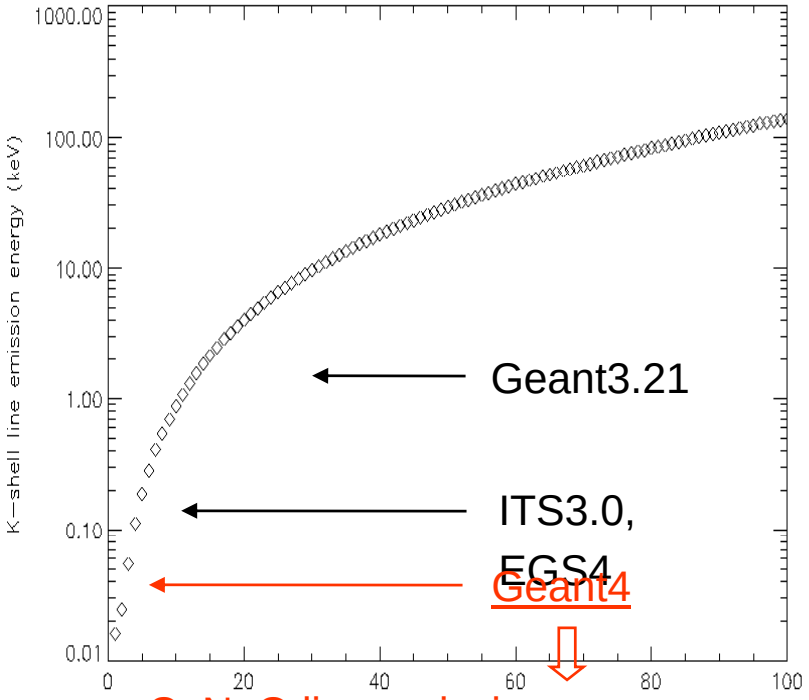


Positronen Emissions Tomography (PET) Bilder
links: “echte” Aufnahme rechts: die Simulation

# X-Ray Surveys of Asteroids and Moons




*Courtesy SOHO EIT*

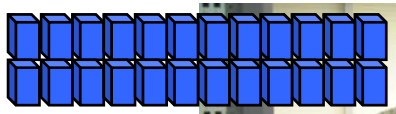



Induced X-ray line emission:
indicator of target composition
(~100 μm surface layer)

C, N, O line emissions included

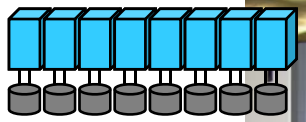
# The Data Flow

Physics Analysis Programs

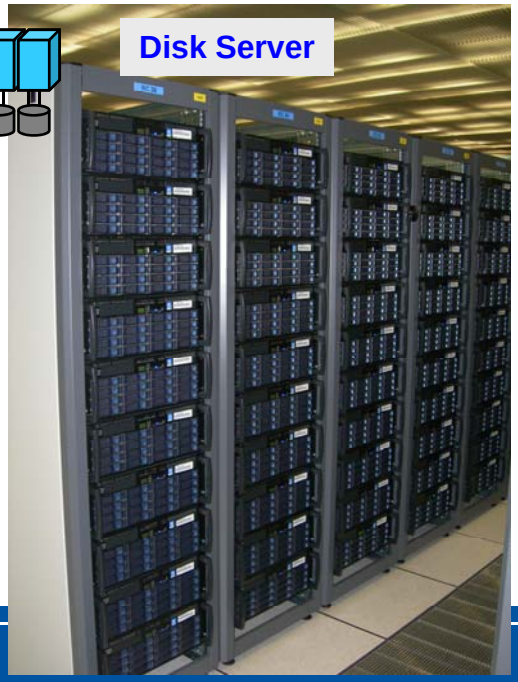


CPU Server

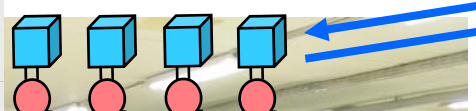




Disk Server

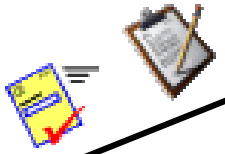


Tape Server and Tape Library



# Storage

*WINDOWS, ORACLE, FC disks, SAN*



**Administrative Data**

**2 million electronic documents**
**70000 electronic signatures per month**
**80000 emails per day**
**250000 orders per year (650 million CHF)**

**backup**
**per hour and per day**

**Users**

**accessibility → 24*7*52 = always**
**tape storage → forever**

*LINUX, SATA disks, Gigabit*

**continuous storage**

**Physics Data**

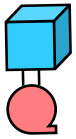
**10000 TB and**
**20 million files per year**

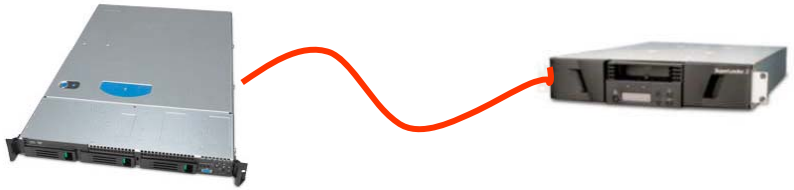
# Hardware Building Blocks

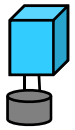
**commodity market components**
**cost effective**
**simple components, but many of them**

**CPU server**

dual CPU, dual core,
8 GB memory

**Tape server**

=
CPU server + fibre channel connection
+ tape drive

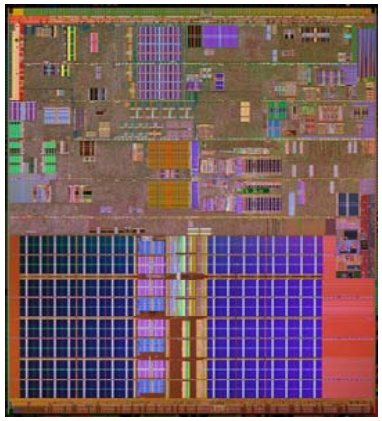


**Disk server**

=
CPU server + RAID controler +24 SATA disks

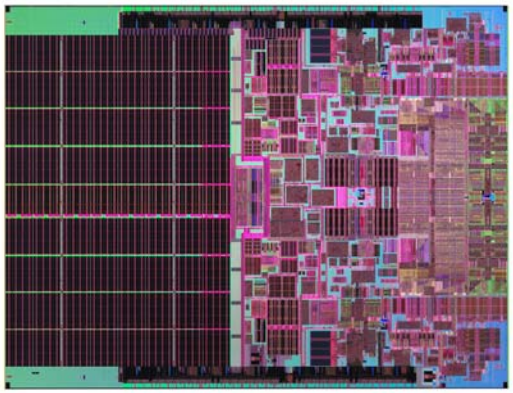
**market trends more important**
**than technology trends**

# Technologie

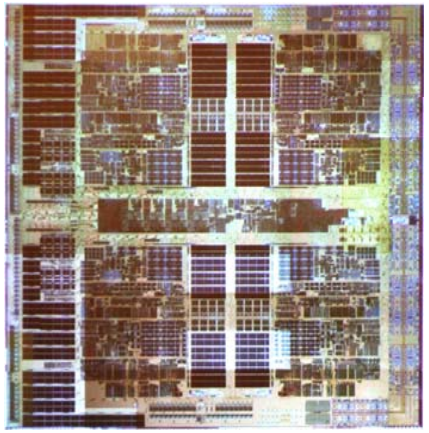
- In den letzten 18Monaten vollzog sich ein Paradigmenwechsel bei den Prozessorherstellern
  - ➔ Leistungssteigerungen werden nicht mehr (nur) durch höhere Frequenzen erreicht!
  - ➔ Auf einem “Prozessor” befinden jetzt sich mehrere unabhängige “cores”
    - ➔ Jeder “core” ist ein voll funktionstüchtiger “Prozessor” ...
  - ➔ Die software muss in der Lage sein mit mehreren cores umzugehen!!!
  - ➔ Im Augenblick bis zu 4 cores in einem Prozessor
  - ➔ In Zukunft 8, 16, 32 oder noch mehr denkbar (Intel hat 80 cores als Forschungsprojekt demonstriert !)



Single core



Dual core



Quad-core

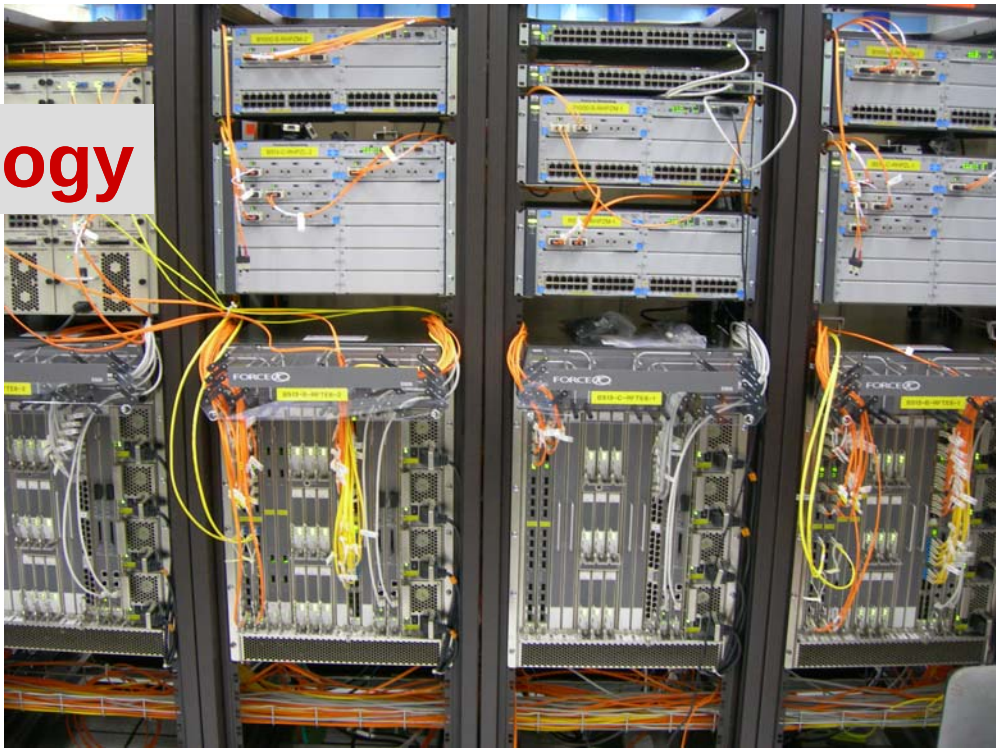
# Software ‘glue’

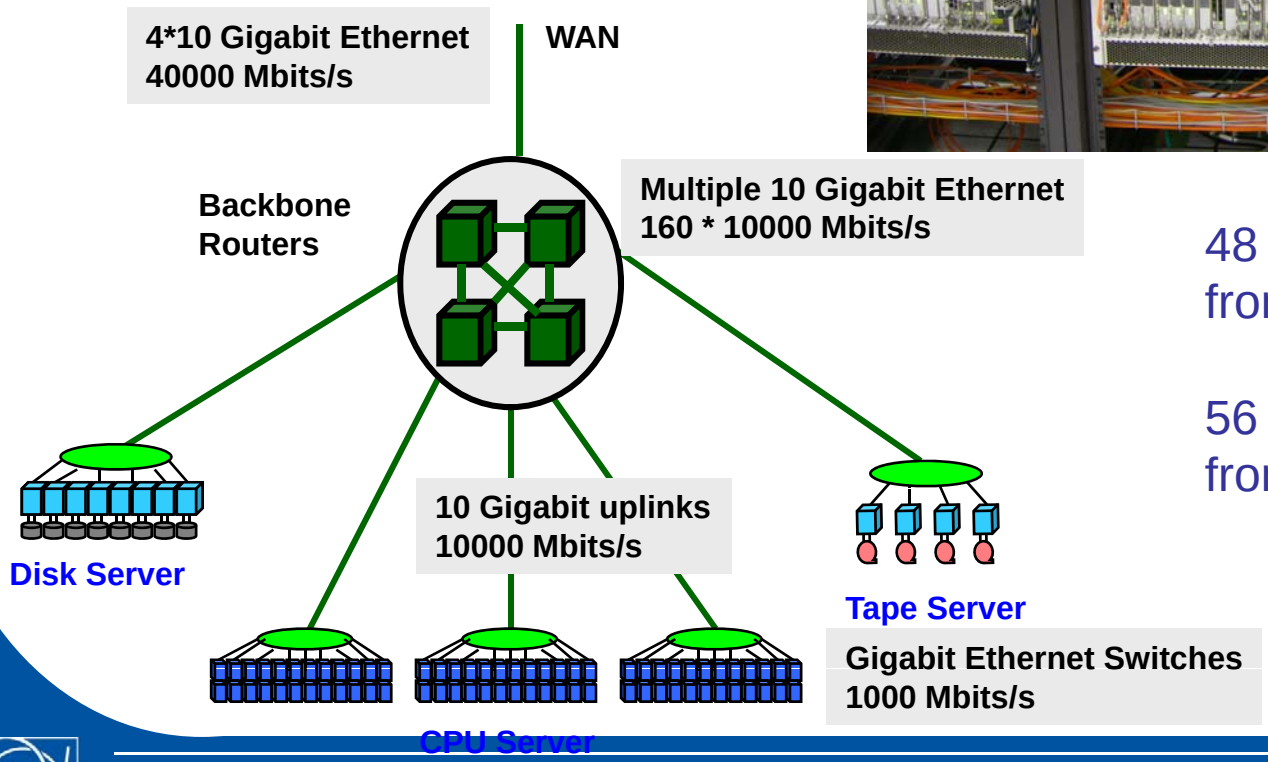
- **management of the basic hardware and software : installation, configuration and monitoring system**
  Which version of Linux ? How to upgrade the software ?
  What is going on in the farm ? Load ? Failures ?

- **management of the processor computing resources : Batch system (LSF from Platform Computing)**
  Where are free processors ? How to set priorities between different users ? sharing of the resources ? How are the results coming back ?

- **management of the storage (disk and tape) : CASTOR (CERN developed Hierarchical Storage Management system)**
  Where are the files ? How can one access them ?
  How much space is available ?

# Schematic network topology

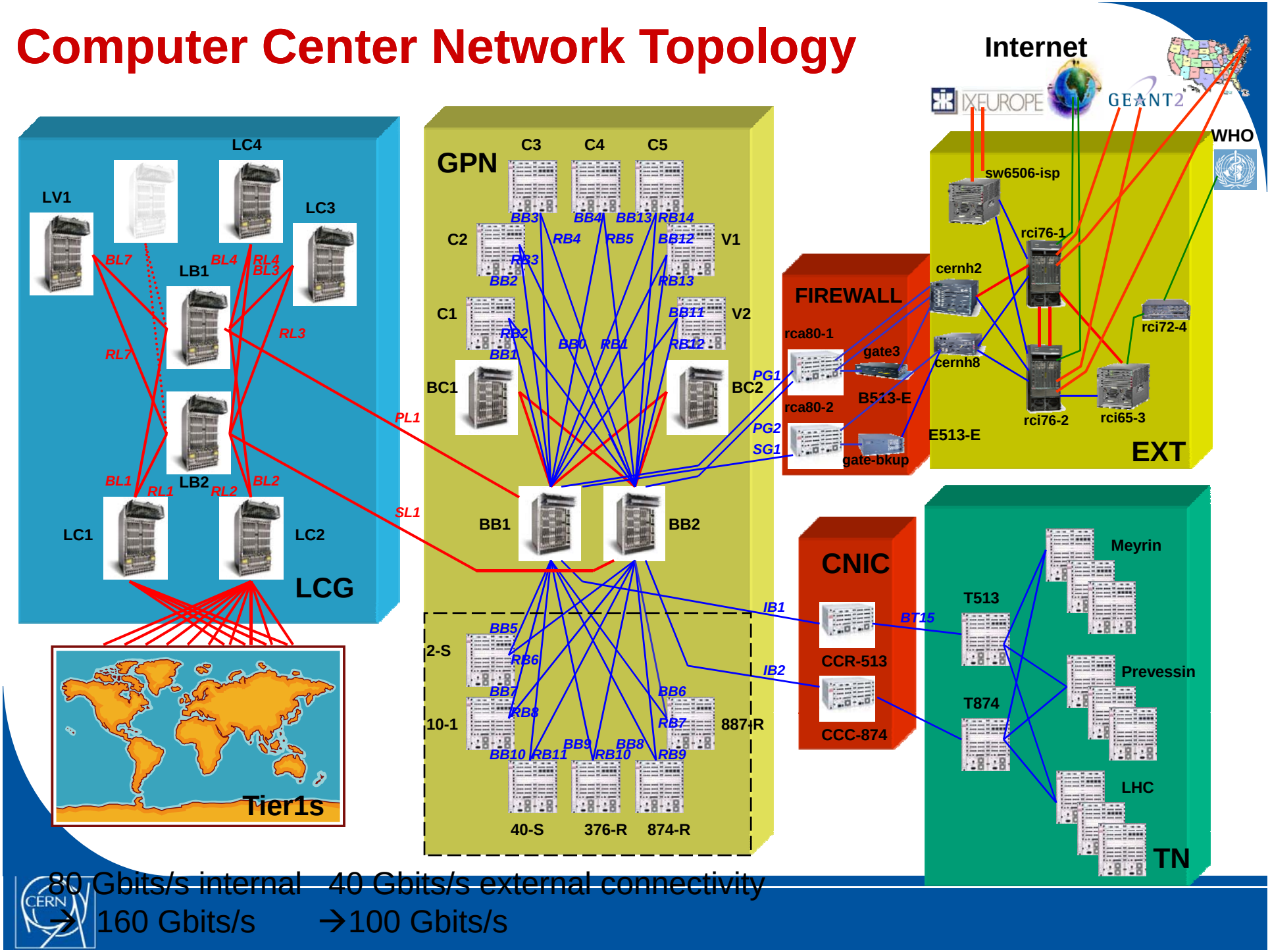





48 port Gigabit switches from HP

56 port 10 Gigabit routers from Force10

# Computer Center Network Topology




80 Gbits/s internal 40 Gbits/s external connectivity

→160 Gbits/s →100 Gbits/s

# Security

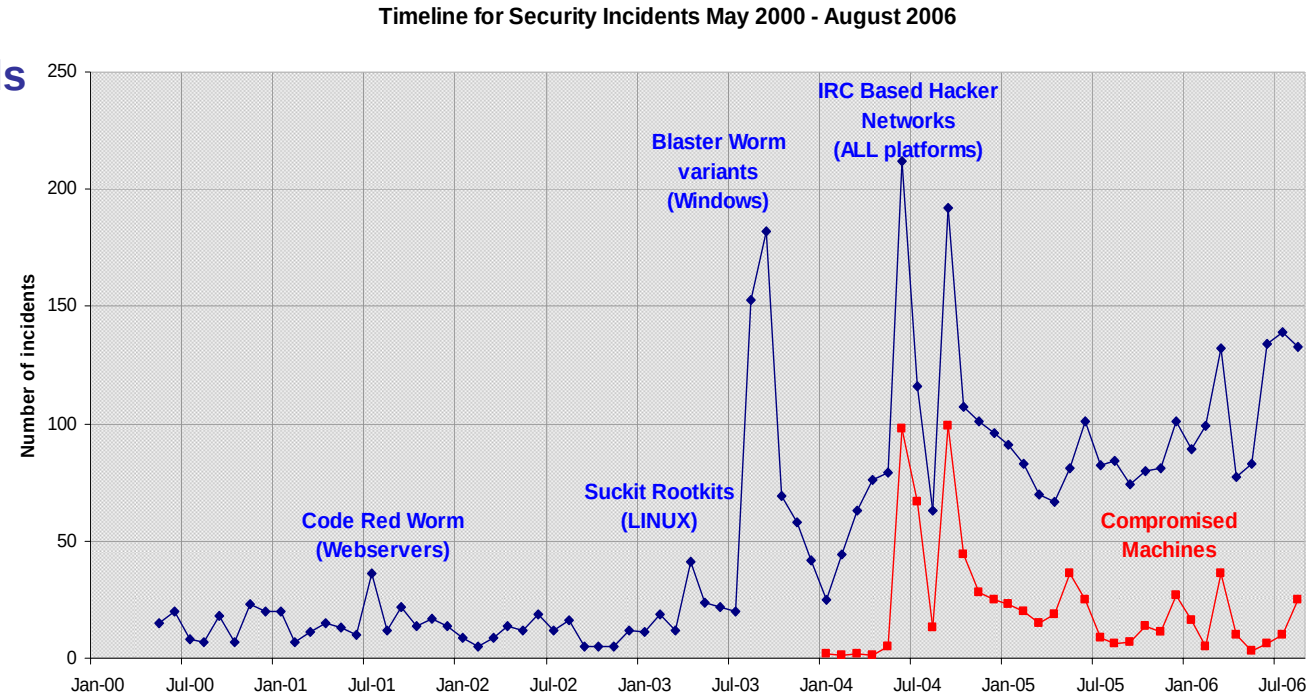
5 person security team
plus security effort for all services

regular security patches, weekly, monthly plus emergency at any time

several levels of firewalls

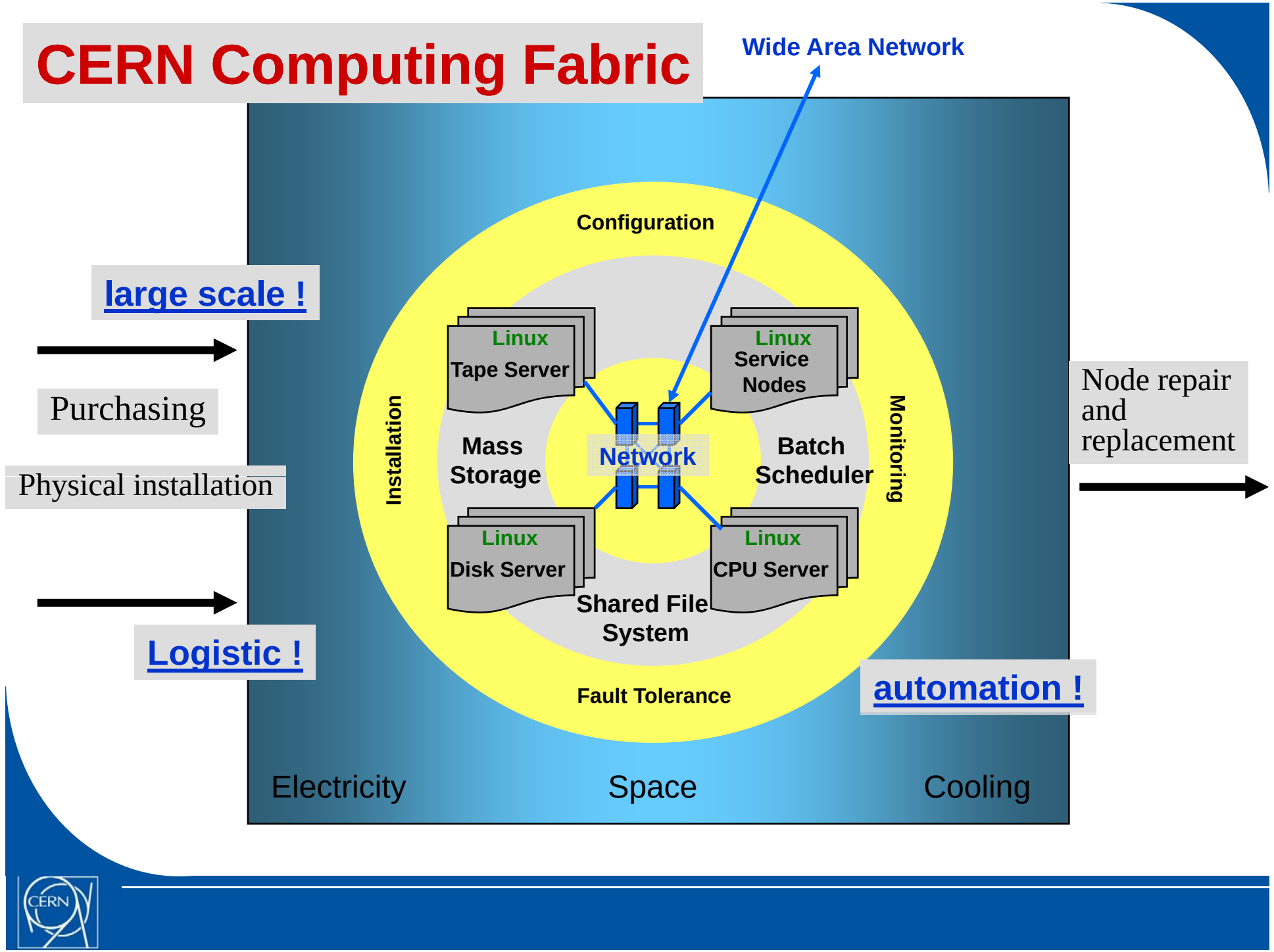
detailed automatic monitoring

encryption of sensitive data

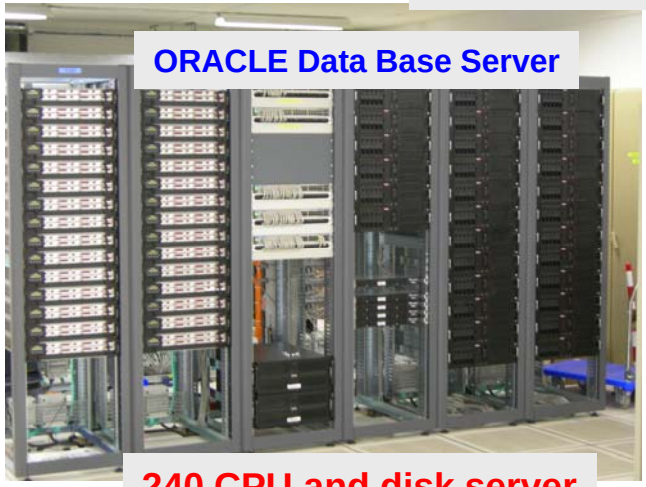



**Focus : protection of sensitive data (administration)**
**hijacking of large clusters**

# CERN Computing Fabric



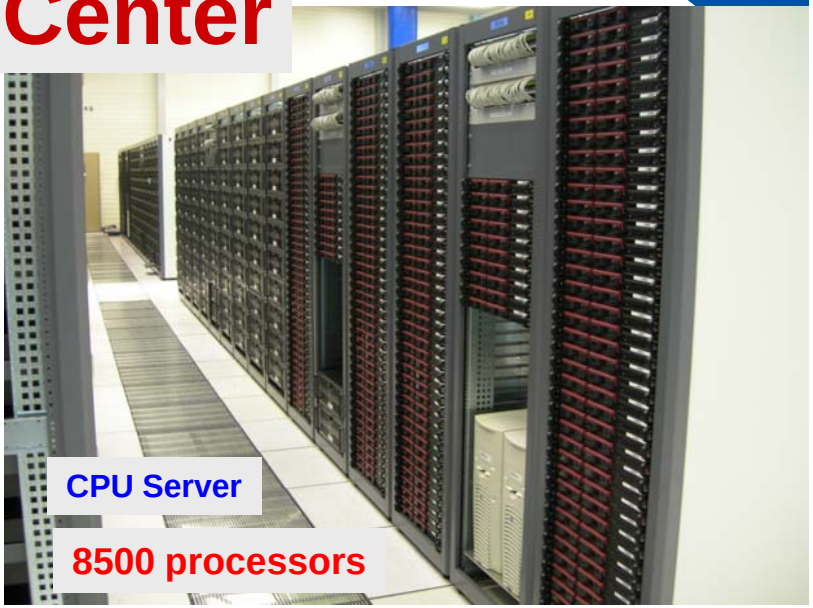
# CERN Computer Center



ORACLE Data Base Server

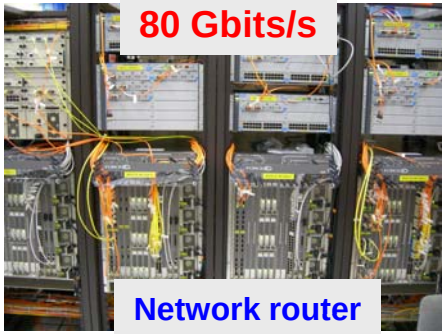
240 CPU and disk server

2.5 MW Electricity and Cooling



CPU Server

8500 processors

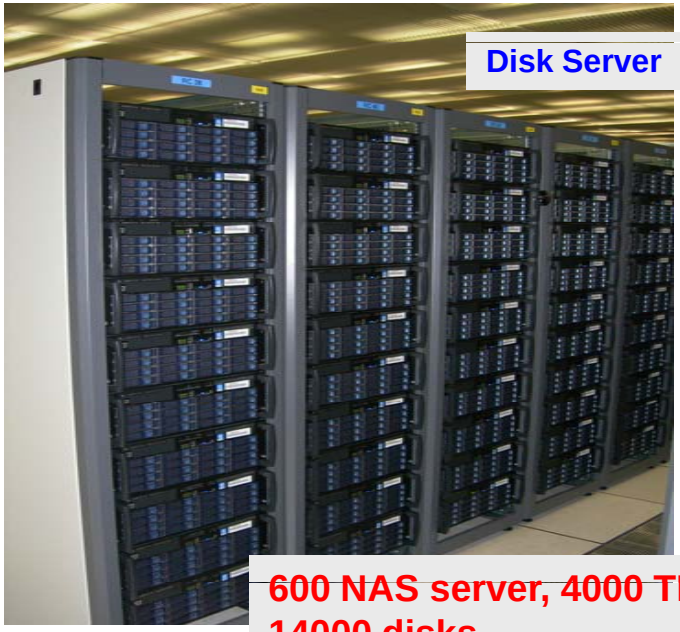


80 Gbits/s

Network router

Tape Server and Tape Library

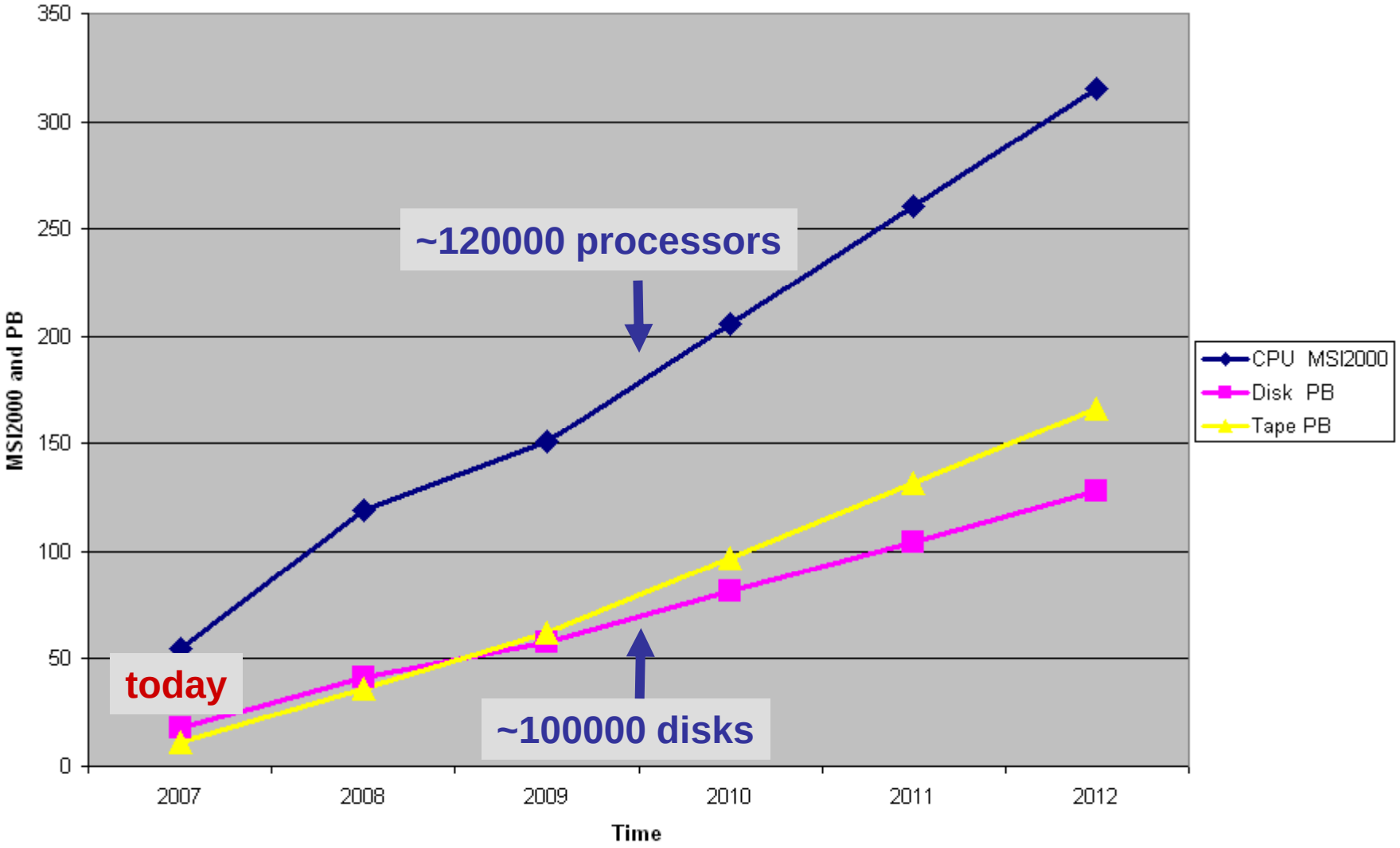
100 tape drives , 20000 tapes
10 PB capacity



Disk Server

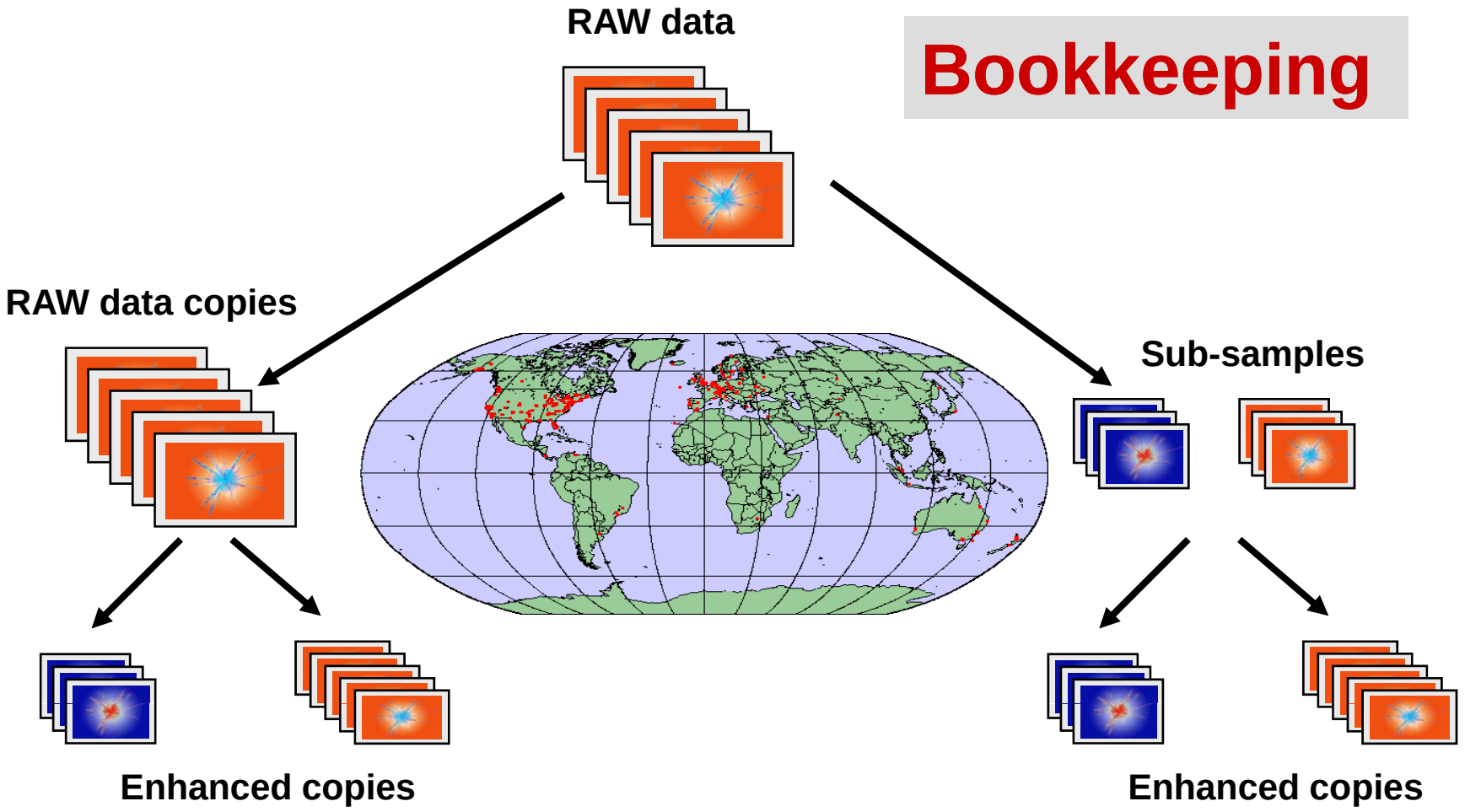
600 NAS server, 4000 TB
14000 disks

# Resources for Computing




**CERN can only contribute ~15% of these resource**
**→ need a world-wide collaboration**

# Bookkeeping




**10000 million ‘snapshots’ created per year**
**10 000 000 000 000 000 Bytes = 10 Petabytes**

**Distributed  world-wide to over 500 institutes**

**Each and every ‘snapshot’ is catalogued and needs to be traced**

# Physical and logical coupling

Complexity

Components

Hardware

Software

10

CPU+disk+memory +motherbord

Operating system WINDOWS LINUX

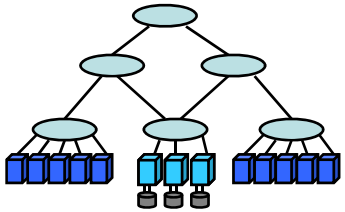
PC

30000

Local Area Network

Resource Management software

Cluster

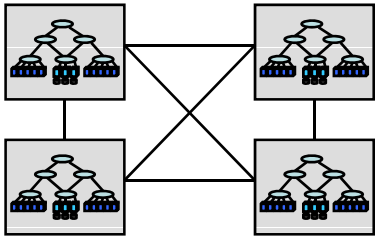


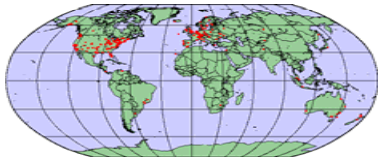
1000000

Wide area network

Grid middleware

World Wide Cluster





# Solution: the Grid

- Use the Grid to unite computing resources of particle physics institutes around the world

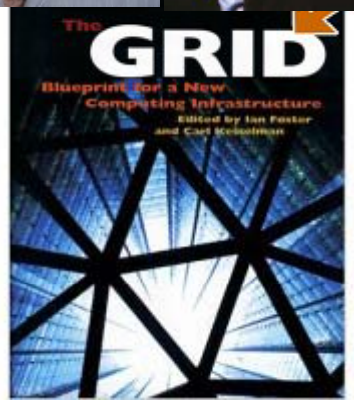
**Tim Berners-Lee invented the World Wide Web at CERN in 1989**

The **World Wide Web** provides seamless access to information that is stored in many millions of different geographical locations

The **Grid** is an infrastructure that provides seamless access to computing power and data storage capacity distributed over the globe

# Grid history

- Name “Grid” chosen by analogy with electric power grid (Foster and Kesselman 1997)
- Vision: plug-in computer for processing power just like plugging in toaster for electricity.
- Concept has been around for decades (distributed computing, metacomputing)
- Key difference with the Grid is to realize the vision on a global scale.




I want to analyze the LHC measurements

Am I allowed to work in this center ?

?

Where are the data ?

How do I access them ?

Where is a free computer ?

How do I get the results back ?

?

?

?

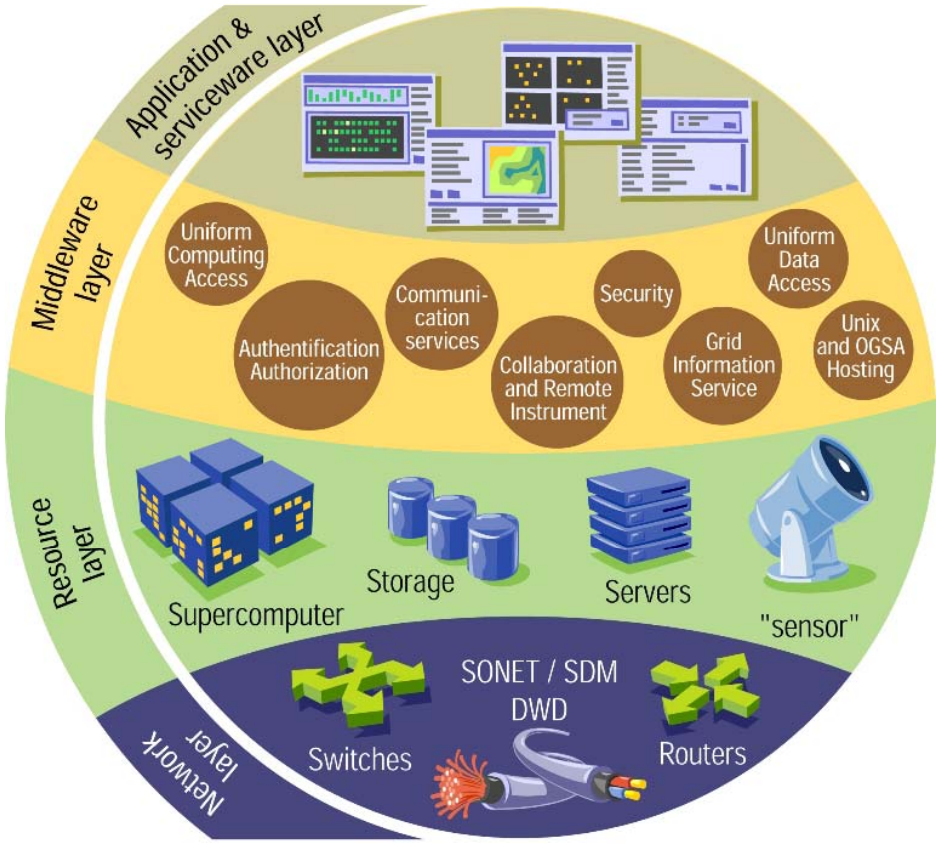
?

There are many different centers !

Each one with different hardware and software !

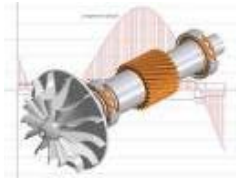
# How does the Grid work?

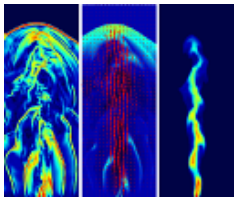
- It relies on advanced software, called **middleware**.
- Middleware automatically finds the **data** the scientist needs, and the **computing power** to analyse it.
- Middleware balances the load on different resources. It also handles **security, accounting, monitoring** and much more.



# Das EGEE Projekt

- **E**nabling **G**rid for **E**-scienc**E** - **EGEE**
  - 1 April 2004 – 31 März 2006
  - 71 Partner in 27 Ländern, verbunden in regionalen Grids
- **EGEE-II**
  - 1 April 2006 – 31 März 2008
  - 91 Partner in 32 Ländern
  - 13 Verbünde
- **EGEE-III (in Verhandlung)**
- Ziele
  - Large-scale, production-quality Infrastruktur für e-Science
  - Neue Benutzer und Ressourcen aus Industrie und Forschung anziehen
  - Verbesserung und Pflege der “gLite” Grid middleware

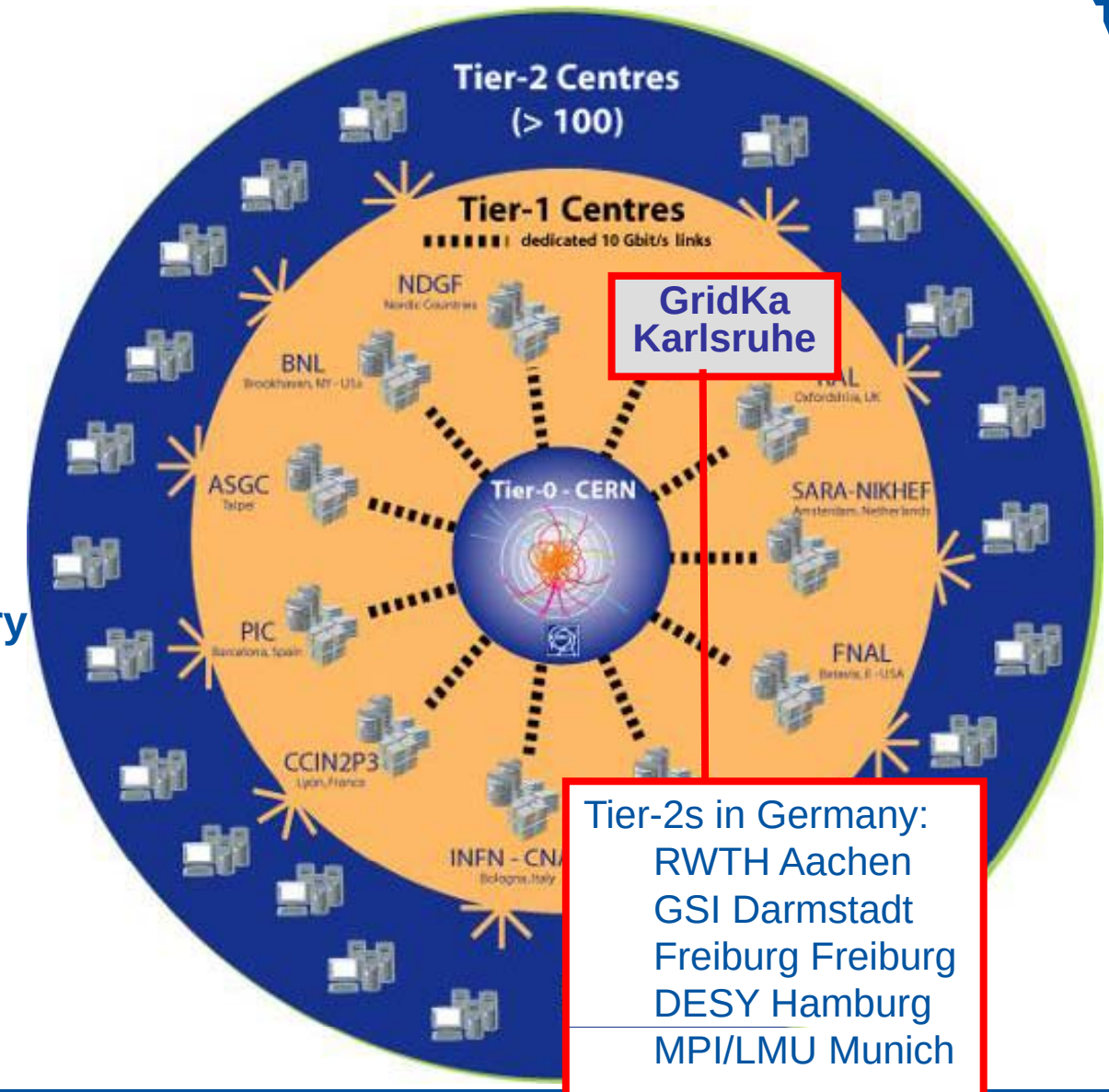





# LHC Computing Grid project (LCG)

- Initiated in 2002
- Collaboration with various EU projects

- More than 150 computing centres
- 12 large centres for primary data management: CERN (Tier-0) and eleven Tier-1s
- 38 federations of smaller Tier-2 centres
- 40 countries involved




Tier-2s in Germany:
RWTH Aachen
GSI Darmstadt
Freiburg Freiburg
DESY Hamburg
MPI/LMU Munich
Univ. Wuppertal

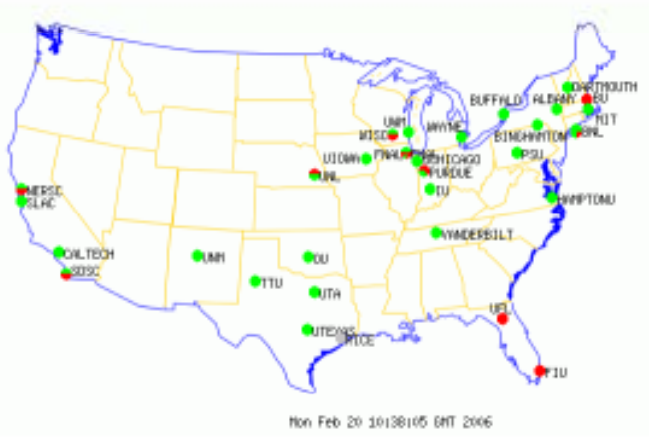
# Grid Projects Collaborating in LHC Computing Grid



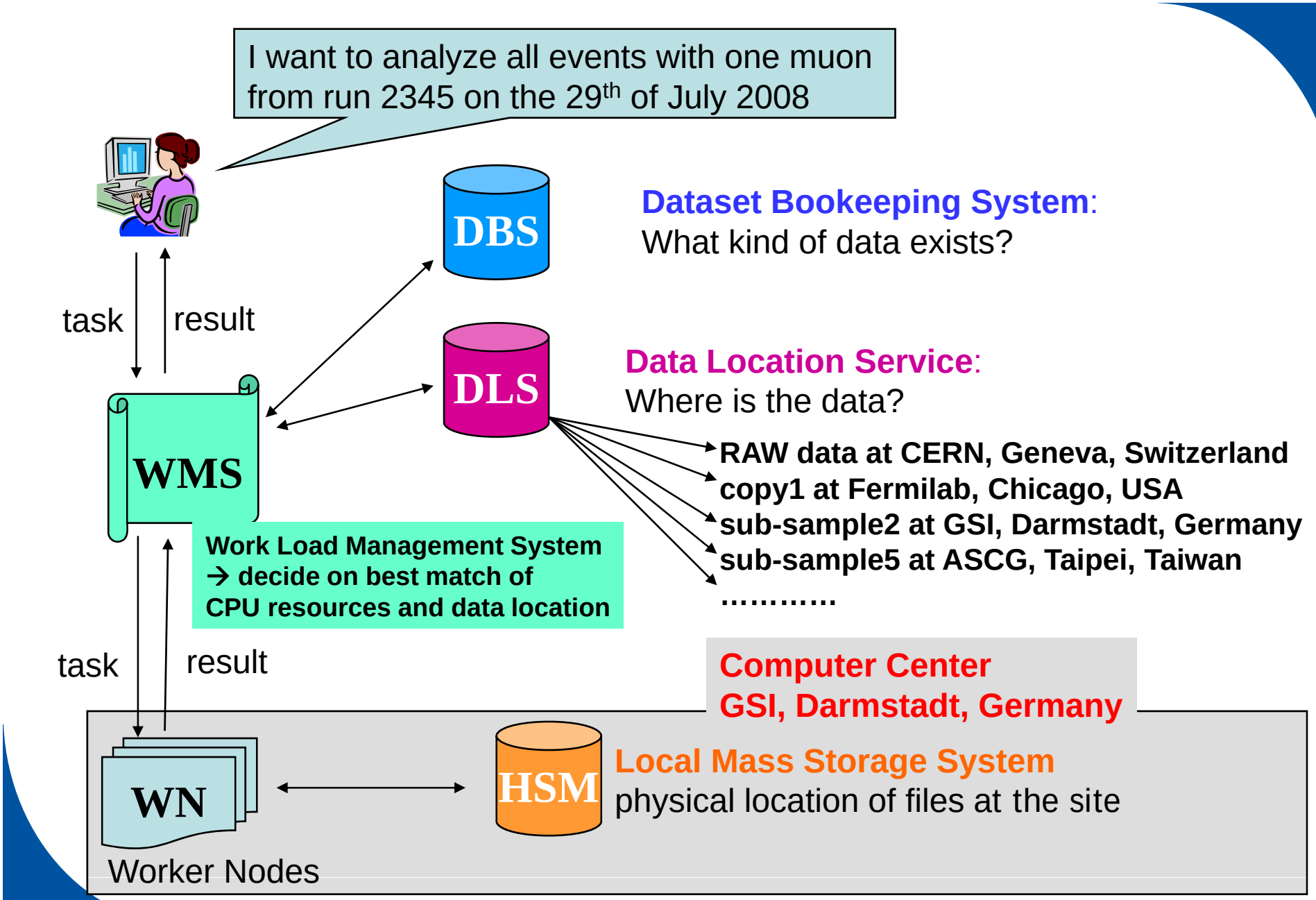





| | | |
|---|---|---|
| **Active sites** | **:** | **> 230** |
| **Countries involved** | **:** | **~ 50** |
| **Available processors** | **:** | **~ 35000** |
| **Available disk space** | **:** | **~ 10 PB** |

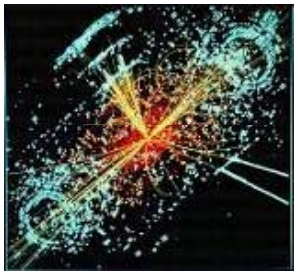


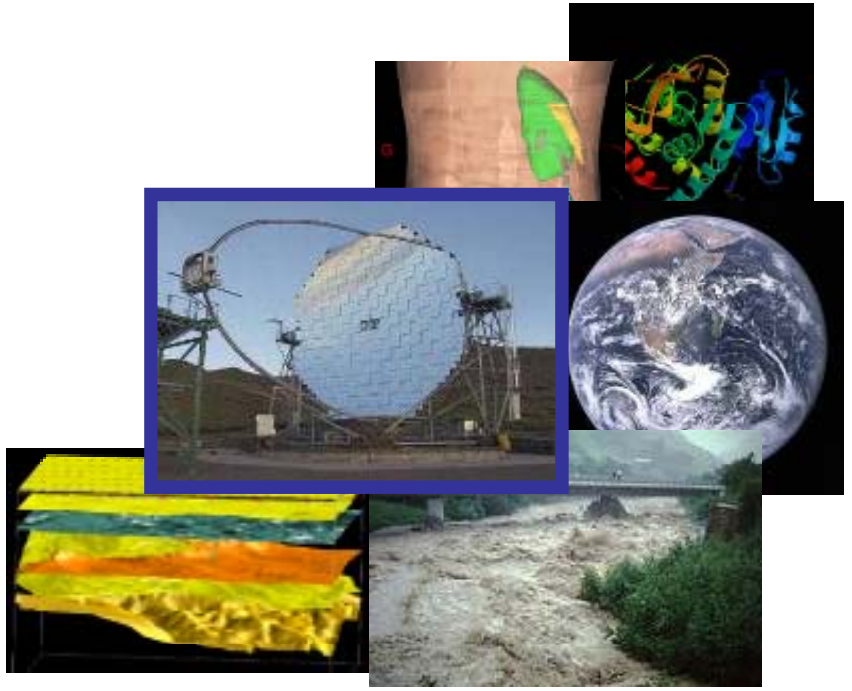

LastBuild:Mon Mar 12 07:16:01 GMT 2007 GstatQuery:2006-12-15


I want to analyze all events with one muon from run 2345 on the 29th of July 2008
DBS
Dataset Bookeeping System:
What kind of data exists?
task
result
DLS
Data Location Service:
Where is the data?
WMS
RAW data at CERN, Geneva, Switzerland
copy1 at Fermilab, Chicago, USA
sub-sample2 at GSI, Darmstadt, Germany
sub-sample5 at ASCG, Taipei, Taiwan
.....
Work Load Management System
→ decide on best match of
CPU resources and data location
task
result
Computer Center
GSI, Darmstadt, Germany
WN
HSM
Local Mass Storage System
physical location of files at the site
Worker Nodes

# Anwendungen in EGEE

- Mehr als 20 Anwendungen aus 7 unterschiedlichen Gebieten
  - High Energy Physics - Hochenergiephysik (Pilot domain)
    - 4 LHC Experimente
    - Andere HEP Experimente (DESY, Fermilab, etc.)
  - Bio-Medizin (Pilot domain)
    - Bioinformatik + bildgebende Verfahren
  - Geo-Wissenschaften
    - Erdbeobachtung
    - Solid Earth Physics
    - Hydrologie
    - Klimaforschung
  - Computational Chemistry
  - Fusions Forschung
  - Astronomie
    - Cosmic microwave background
    - Gamma ray astronomy
  - Geophysik
    - Industrielle Anwendungen

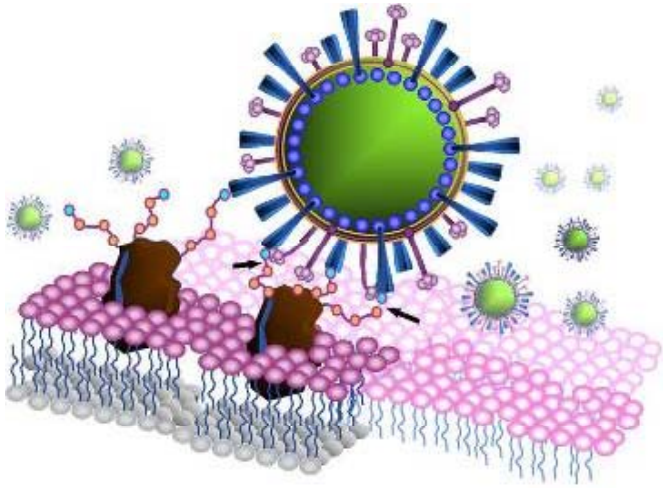




Book of abstracts: http://doc.cern.ch//archive/electronic/egee/tr/egee-tr-2006-005.pdf

# EGEE und die Vogelgrippe

- **EGEE** wurde benutzt um 300,000 Verbindungen auf eine potentielle Wirksamkeit gegen den H5N1 virus zu testen.
- 2000 Computer in 60 Rechen-zentren in Europa, Rußland, Asien und im Mittleren Osten liefen für vier Wochen im April - äquivalent zu 150 Jahren auf einem einzelen Computer.
- Potentielle Wirkstoffe werden getestet und eingestuft...



***Neuraminidase, one of the two major surface proteins of influenza viruses, facilitating the release of virions from infected cells****. Image Courtesy Ying-Ta Wu, AcademiaSinica.*