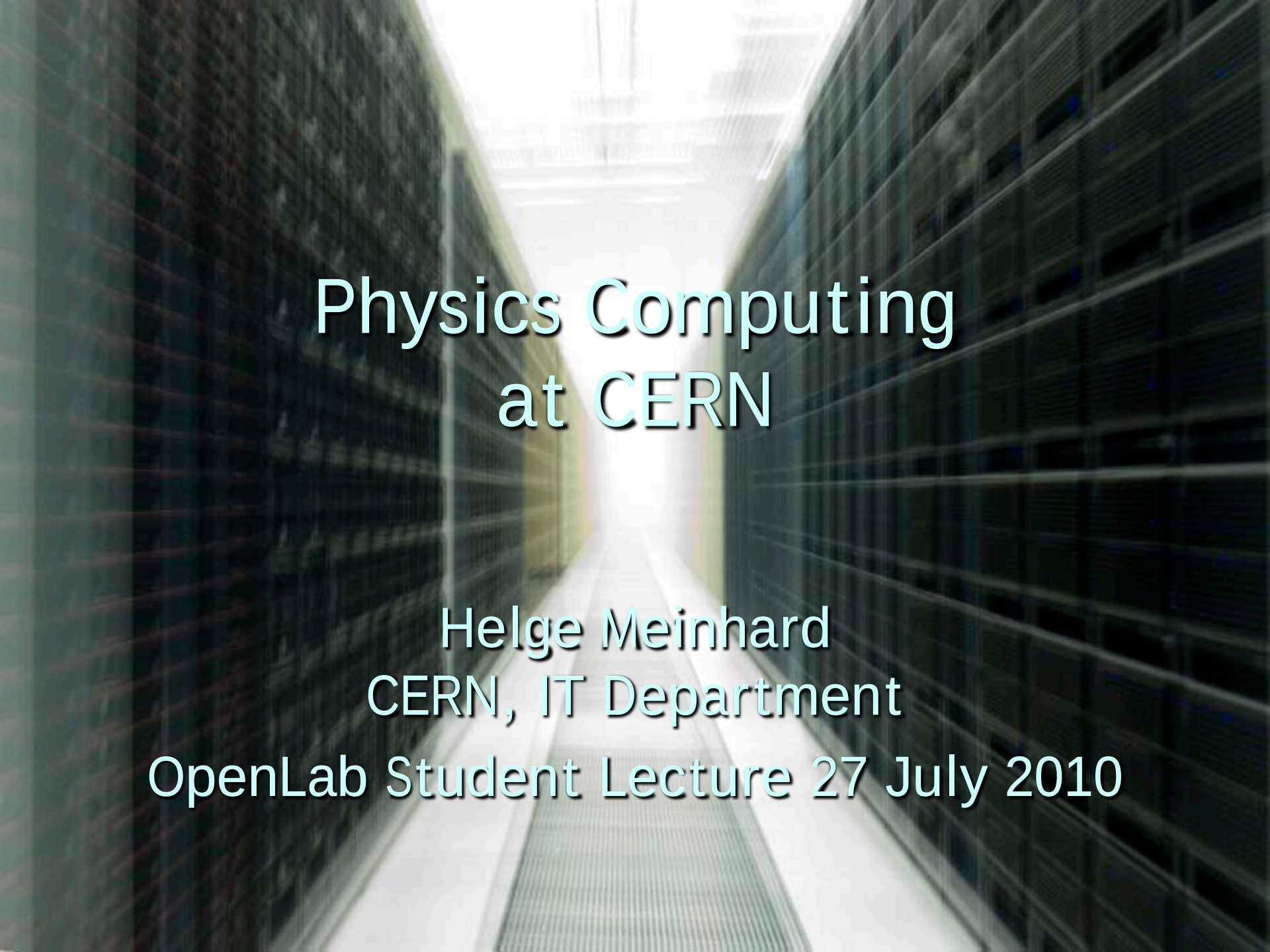


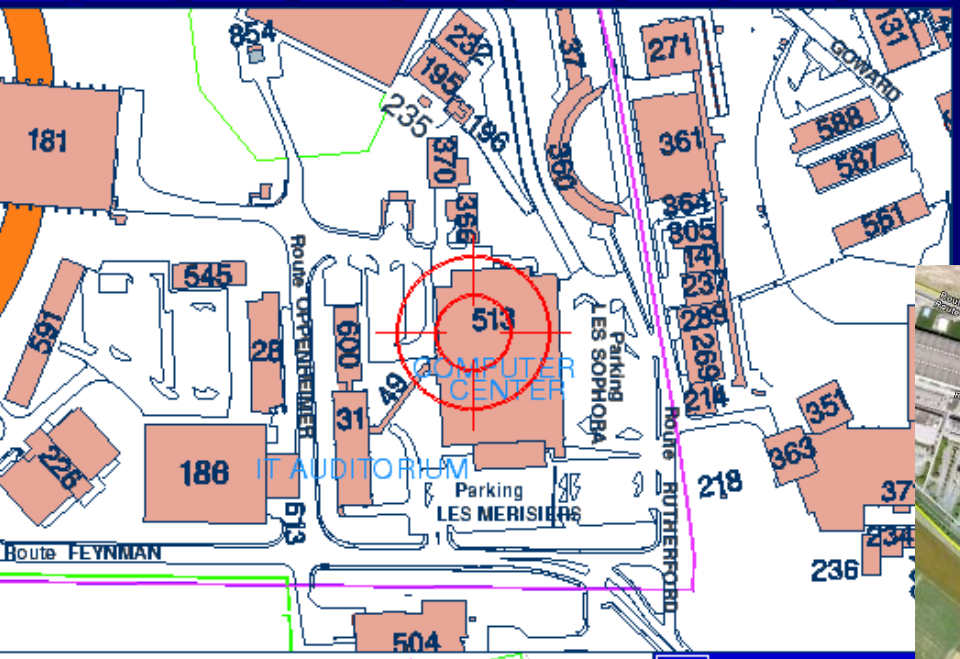
Physics Computing at CERN

Helge Meinhard
CERN, IT Department

OpenLab Student Lecture 27 July 2010

Location

Building 513 (opposite of restaurant no. 2)



Building

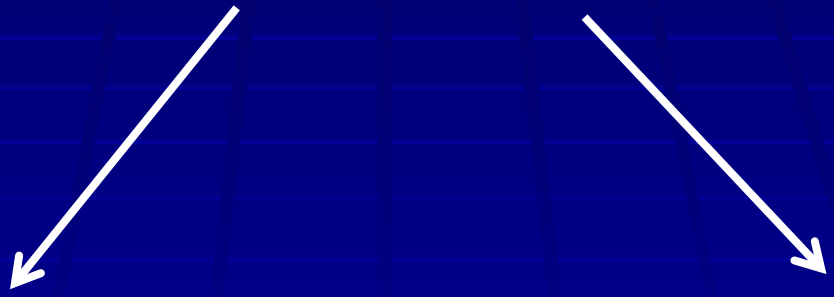
Large building with 2700 m^2 surface for computing equipment, capacity for 2.9 MW electricity and 2.9 MW air and water cooling

Chillers

Transformers

Computing Service Categories

Two coarse grain computing categories



Computing infrastructure
and
administrative computing

Physics data flow
and
data processing

Task overview

- *Communication tools:* mail, Web, Twiki, GSM, …
- *Productivity tools:* office software, software development, compiler, visualization tools, engineering software, …
- *Computing capacity:* CPU processing, data repositories, personal storage, software repositories, metadata repositories, …

- Needs underlying infrastructure
  - Network and telecom equipment
  - Processing, storage and database computing equipment
  - Management and monitoring software
  - Maintenance and operations
  - Authentication and security

CERN CC currently (April 2010)

- 7’500 systems, 50’000 processing cores
  - CPU servers, disk servers, infrastructure servers
- 19’800 TB usable on 55’000 disk drives
- 24’000 TB used, 50’000 tape cartridges total (70’000 slots), 160 tape drives
- Tenders in progress or planned (estimates)
  - 1’750 systems, 17’400 processing cores
  - 12’000 TB usable on 17’000 disk drives

Infrastructure Services

Software environment and productivity tools

User registration and authentication
22'000 registered users

Mail
2 million emails/day, 99% spam
18000 mail boxes

Web services
> 8000 web sites

Tool accessibility
Windows, Office, CadCam, …

Home directories (DFS, AFS)
150 TB, backup service
1 billion files

PC management
Software and patch installations

Infrastructure needed :
> 400 servers

Network Overview

Experiments

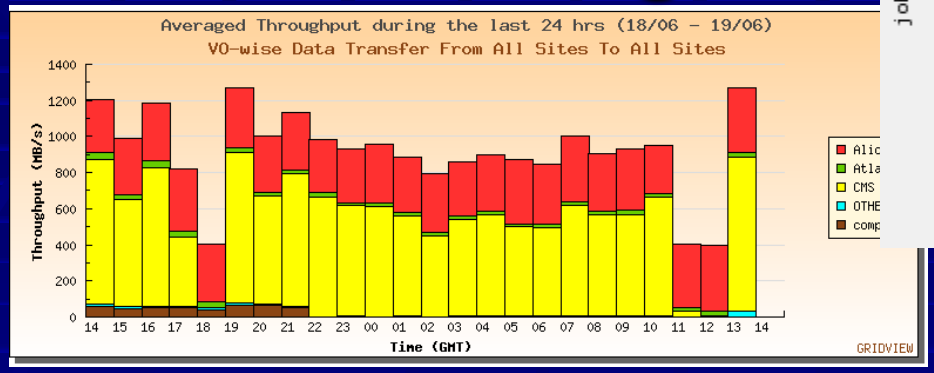
All CERN buildings
12000 users

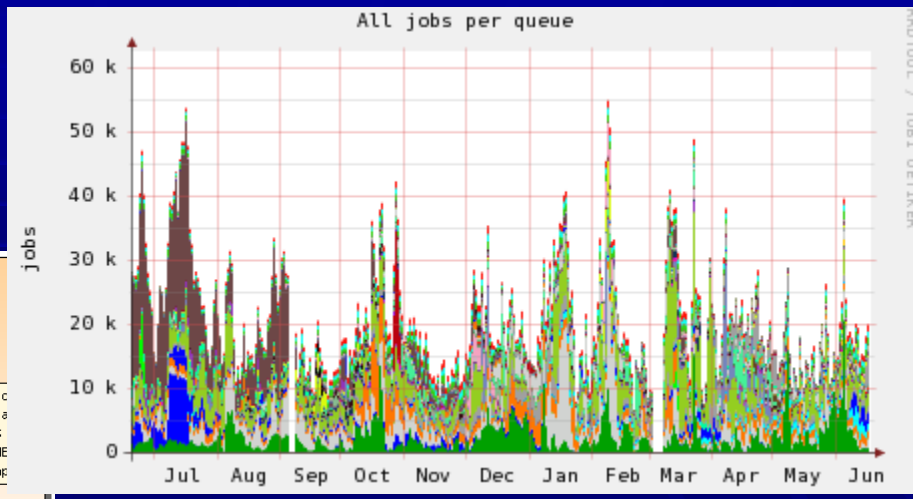
Central, high speed network backbone

World Wide Grid centres

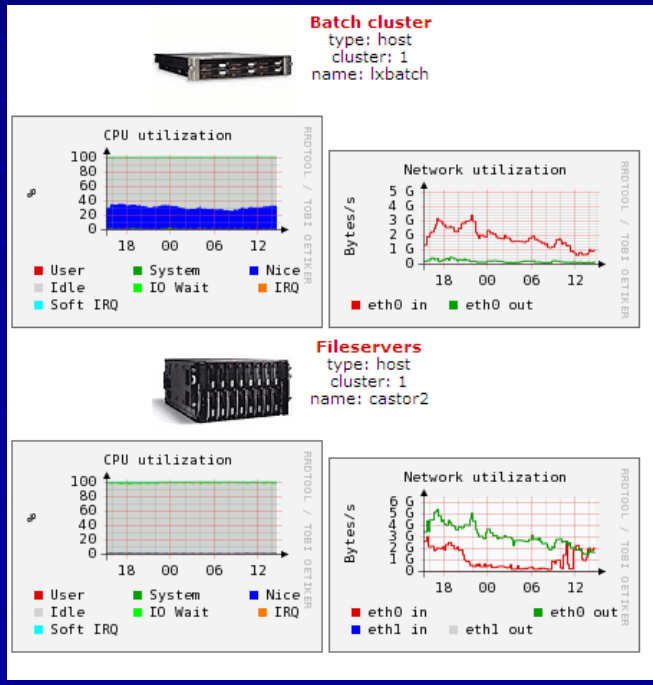
Computer centre processing clusters

Monitoring





- Large scale monitoring
  - Surveillance of all nodes in the computer centre
  - Hundreds of parameters in various time intervals, from minutes to hours, per node and service
  - Data base storage and Interactive visualisation



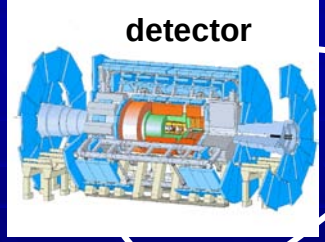
Bookkeeping: Database Services

- More than 200 ORACLE data base instances on > 300 service nodes
  - Bookkeeping of physics events for the experiments
  - Meta data for the physics events (e.g. detector conditions)
  - Management of data processing
  - Highly compressed and filtered event data
  - …
- LHC machine parameters
- Human resource information
- Financial bookkeeping
- Material bookkeeping and material flow control
- LHC and detector construction details
- …

HEP analyses

- Statistical quantities over many collisions
  - Histograms
  - One event doesn't prove anything
- Comparison of statistics from real data with expectations from simulations
  - Simulations based on known models
  - Statistically significant deviations show that the known models are not sufficient
- Need more simulated data than real data
  - In order to cover various models
  - In order to be dominated by statistical error of real data, not simulation

Data Handling and Computation for Physics Analyses



event filter (selection & reconstruction)

reconstruction

raw data

event reprocessing

event summary data

processed data

batch physics analysis

analysis

analysis objects (extracted by physics topic)

interactive physics analysis

event simulation

simulation

Data Flow – online

Detector

150 million electronics channels

1 PBytes/s

Level 1 Filter and Selection

Fast response electronics, FPGA, embedded processors, very close to the detector

150 GBytes/s

High Level Filter and Selection

O(1000) servers for processing, Gbit Ethernet Network

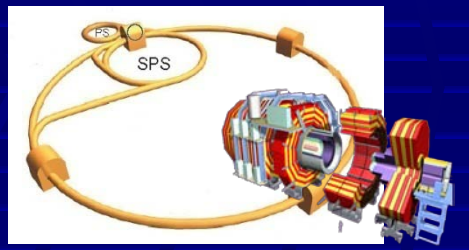
0.6 GBytes/s

N x 10 Gbit links to the computer centre

CERN computer centre

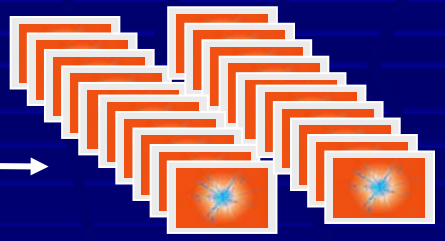
Limits: Essentially the budget and the downstream data flow pressure

Data Flow – offline



4 detectors

1000 million events/s



Filter and first selection

1…2 GB/s

Create sub-samples

World-wide analysis

Physics
Explanation of nature

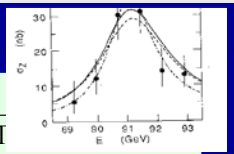
10 GB/s

3 GB/s

Store on disk and tape

Export copies

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



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

SI Prefixes

Prefix	Symbol	1000^m	10^n	Decimal	Short scale	Long scale	Since[1]
yotta	Y	1000^8	10^{24}	1 000 000 000 000 000 000 000 000	Septillion	Quadrillion	1991
zetta	Z	1000^7	10^{21}	1 000 000 000 000 000 000 000	Sextillion	Trilliard	1991
exa	E	1000^6	10^{18}	1 000 000 000 000 000 000	Quintillion	Trillion	1975
peta	P	1000^5	10^{15}	1 000 000 000 000 000	Quadrillion	Billiard	1975
tera	T	1000^4	10^{12}	1 000 000 000 000	Trillion	Billion	1960
giga	G	1000^3	10^9	1 000 000 000	Billion	Milliard	1960
mega	M	1000^2	10^6	1 000 000	Million		1960
kilo	k	1000^1	10^3	1 000	Thousand		1795
hecto	h	$1000^{2/3}$	10^2	100	Hundred		1795
deca	da	$1000^{1/3}$	10^1	10	Ten		1795
		1000^0	10^0	1	One		
deci	d	$1000^{-1/3}$	10^{-1}	0.1	Tenth		1795
centi	c	$1000^{-2/3}$	10^{-2}	0.01	Hundredth		1795
milli	m	1000^{-1}	10^{-3}	0.001	Thousandth		1795
micro	µ	1000^{-2}	10^{-6}	0.000 001	Millionth		1960[2]
nano	n	1000^{-3}	10^{-9}	0.000 000 001	Billionth	Milliardth	1960
pico	p	1000^{-4}	10^{-12}	0.000 000 000 001	Trillionth	Billionth	1960
femto	f	1000^{-5}	10^{-15}	0.000 000 000 000 001	Quadrillionth	Billiardth	1964
atto	a	1000^{-6}	10^{-18}	0.000 000 000 000 000 001	Quintillionth	Trillionth	1964
zepto	z	1000^{-7}	10^{-21}	0.000 000 000 000 000 000 001	Sextillionth	Trilliardth	1991
yocto	y	1000^{-8}	10^{-24}	0.000 000 000 000 000 000 000 001	Septillionth	Quadrillionth	1991

1. The metric system was introduced in 1795 with six prefixes. The other dates relate to recognition by a resolution of the CGPM.
2. The 1948 recognition of the micron by the CGPM was abrogated in 1967.

Source: wikipedia.org

Data Volumes at CERN

- Each year: 15 Petabytes
  - Tower of CDs: which height?
- Stored cumulatively over LHC running
- Only real data and derivatives
  - Simulated data not included
    - Total of simulated data even larger

- *Library of Congress: 200 TB*
- *E-mail (w/o spam): 30 PB*
  30 trillion mails at 1 kB each
- *Photos: 1 EB*
  500 billion photos at 2 MB each
  - *50 PB on Facebook*
- *Web: 1 EB*
- *Telephone calls: 50 EB*

… growing exponentially…

Physical and Logical Connectivity

Complexity / scale

Components

Hardware

Software

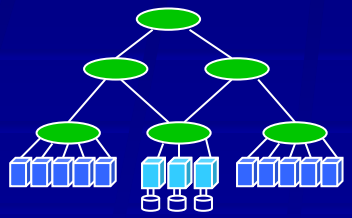
CPU, disk, memory, mainbord

CPU, disk server

Operating system, device drivers

Network, interconnects

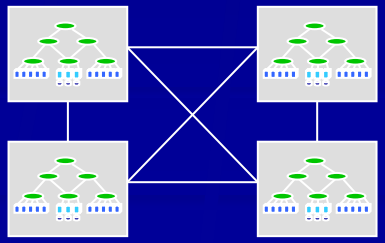
Cluster, local fabric

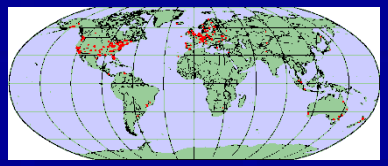


Resource management software

Wide area network

World-wide cluster

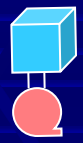




Grid and cloud management software

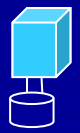
Computing Building Blocks

Commodity market components:
not cheap, but cost effective!
Simple components, but many of them



Tape server =
CPU server
+ fibre channel connection
+ tape drive

CPU server or worker node:
dual CPU, quad core,
16 or 24 GB memory



Disk server =
CPU server
+ RAID controller
+ 24 SATA disks

Market trends more important than technology trends

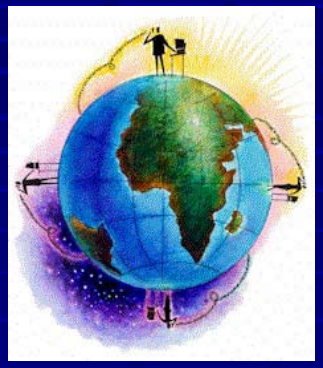
Always watch TCO: Total Cost of Ownership

Hardware Management

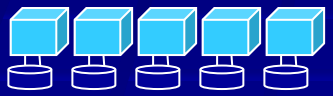
- About 7000 servers installed in centre
- Assume 3 years lifetime for the equipment
  - Key factors: power efficiency, performance, reliability
- Demands by experiments require investments of ~ 15 MCHF/year for new PC hardware and infrastructure
- Infrastructure and operation setup needed for
  - ~2000 nodes installed per year
  - ~2000 nodes removed per year
  - Installation in racks, cabling, automatic installation, Linux software environment

Functional Units

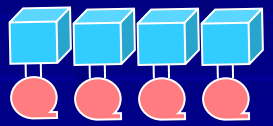
Detectors



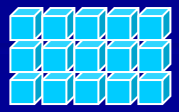
Data import
and export



Disk storage



Tape storage
‘Active’ archive and backup



Event processing capacity
CPU server

Meta-data storage
Data bases

Software “Glue”

- Basic hardware and software management
  - Installation, configuration, monitoring (Quattor, Lemon, ELFms)
  - *Which version of Linux? How to upgrade? What is going on? Load? Failures?*
- Management of processor computing resources
  - Batch scheduler (LSF of Platform Computing Inc.)
  - *Where are free processors? How to set priorities between users? Sharing of resources? How are results flowing back?*
- Storage management (disk and tape)
  - CERN developed HSM called Castor
  - *Where are the files? How to access them? How much space is available? What is on disk, what on tape?*

Job Data and Control Flow (1)

Here is my program and I want to analyse the ATLAS data from the special run on June 16th 14:45h or all data with detector signature X

Processing nodes (CPU servers)

'Batch' system to decide where is free computing time

Management software

Data management system where is the data and how to transfer to the program

Disk storage

Database system
Translate the user request into physical location and provide meta-data (e.g. calibration data) to the program

Job Data and Control Flow (2)

CERN installation

Users

Repositories
- code
- metadata
- .....

Interactive
lxplus

Data processing
Lxbatch

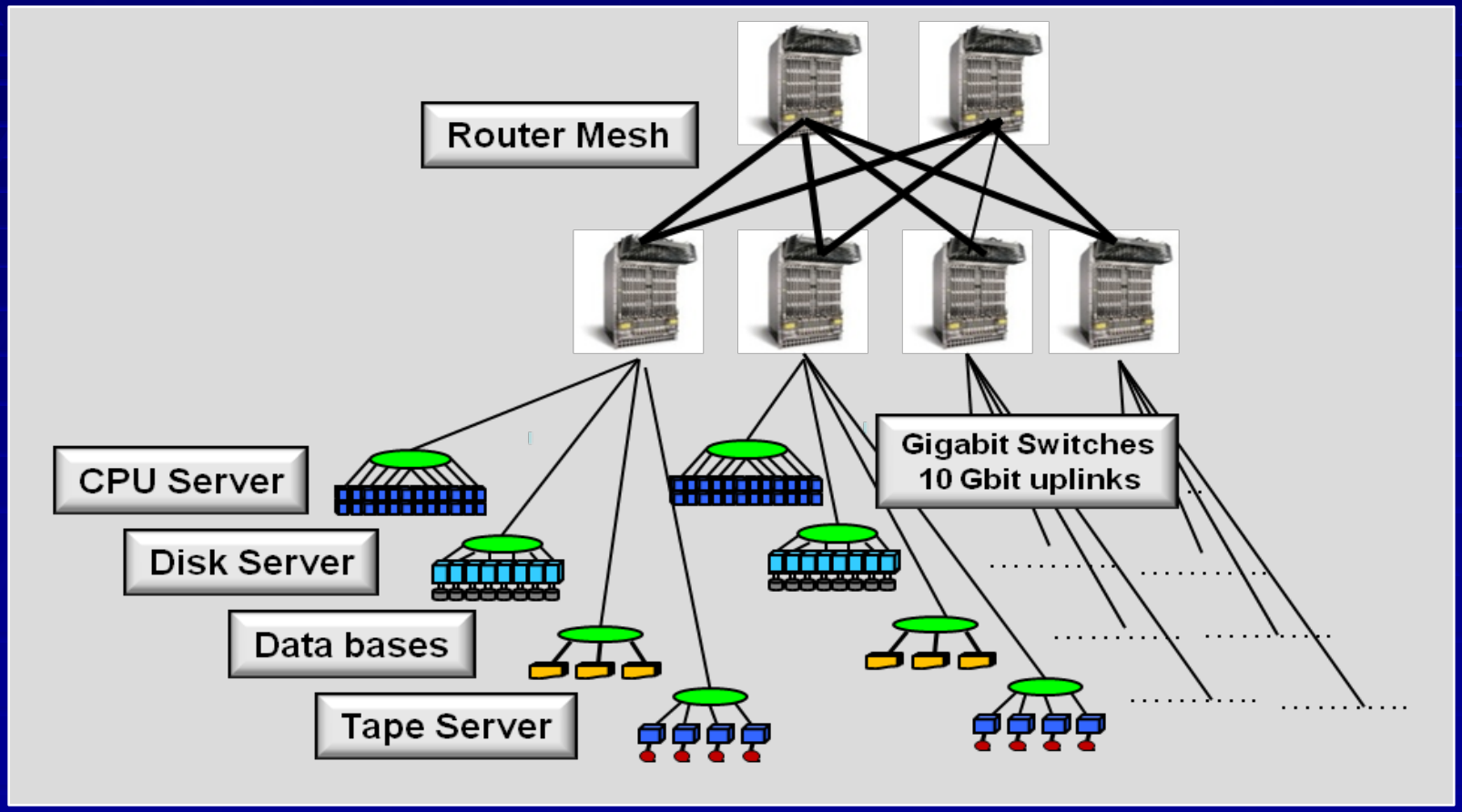
Bookkeeping
Data base

Disk storage

Tape storage

CASTOR
Hierarchical Mass Storage Management System (HSM)

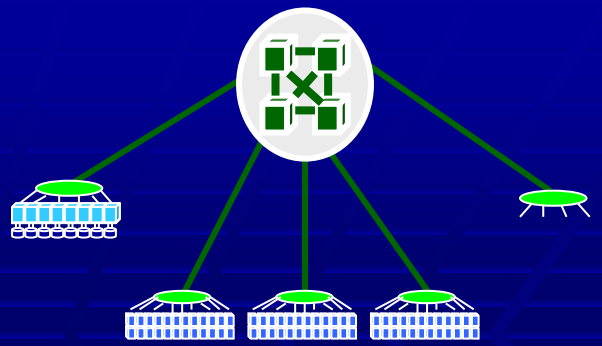
CERN Farm Network



Switches in the distribution layer close to servers, 10 Gbit uplinks, majority 1 Gbit to server, slowly moving to 10 Gbit server connectivity

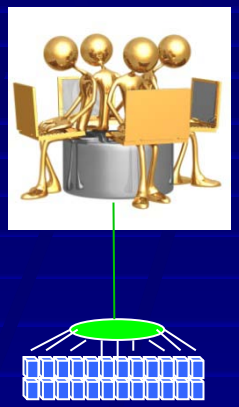
Expect 100 Gbytes/s internal traffic (15 Gbytes/s peak today)

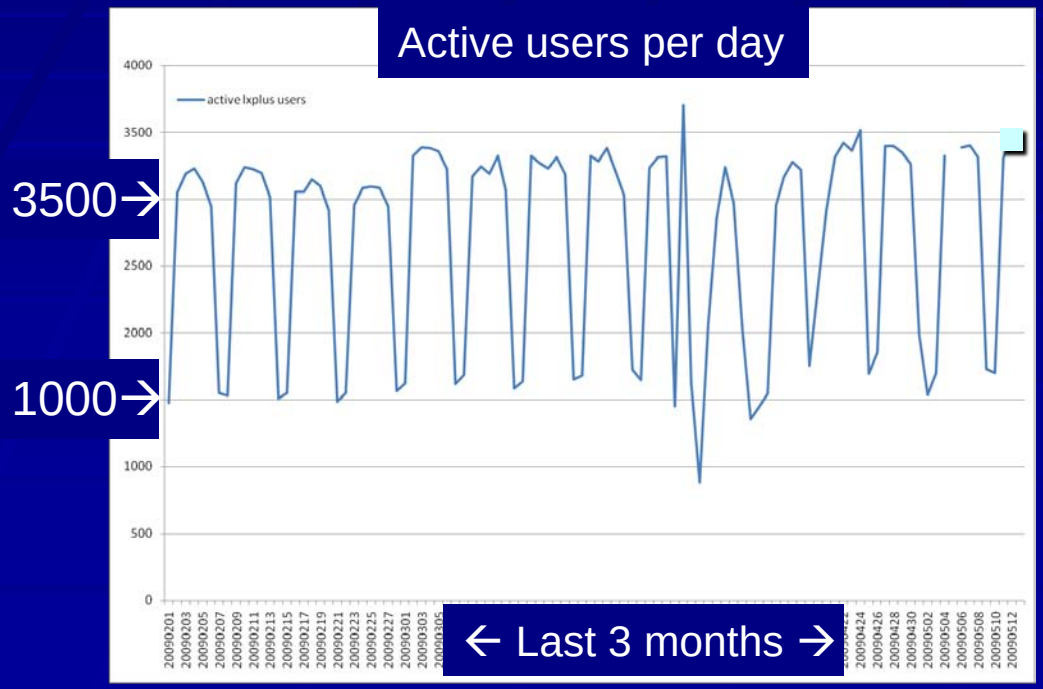
CERN Overall Network



- Hierarchical network topology based on Ethernet
- 150+ very high performance routers
- 3’700+ subnets
- 2200+ switches (increasing)
- 50’000 active user devices (exploding)
- 80’000 sockets – 5’000 km of UTP cable
- 5’000 km of fibers (CERN owned)
- 140 Gbps of WAN connectivity

Interactive Login Service: lxplus





- Interactive compute facility
- 70 CPU servers running Linux (RedHat variant)
- Access via ssh from desktops and notebooks under Windows, Linux, MacOS X

Used for compilation of programs, short program execution tests, some interactive analysis of data, submission of longer tasks (jobs) into the lxbatch facility, internet access, program development etc.

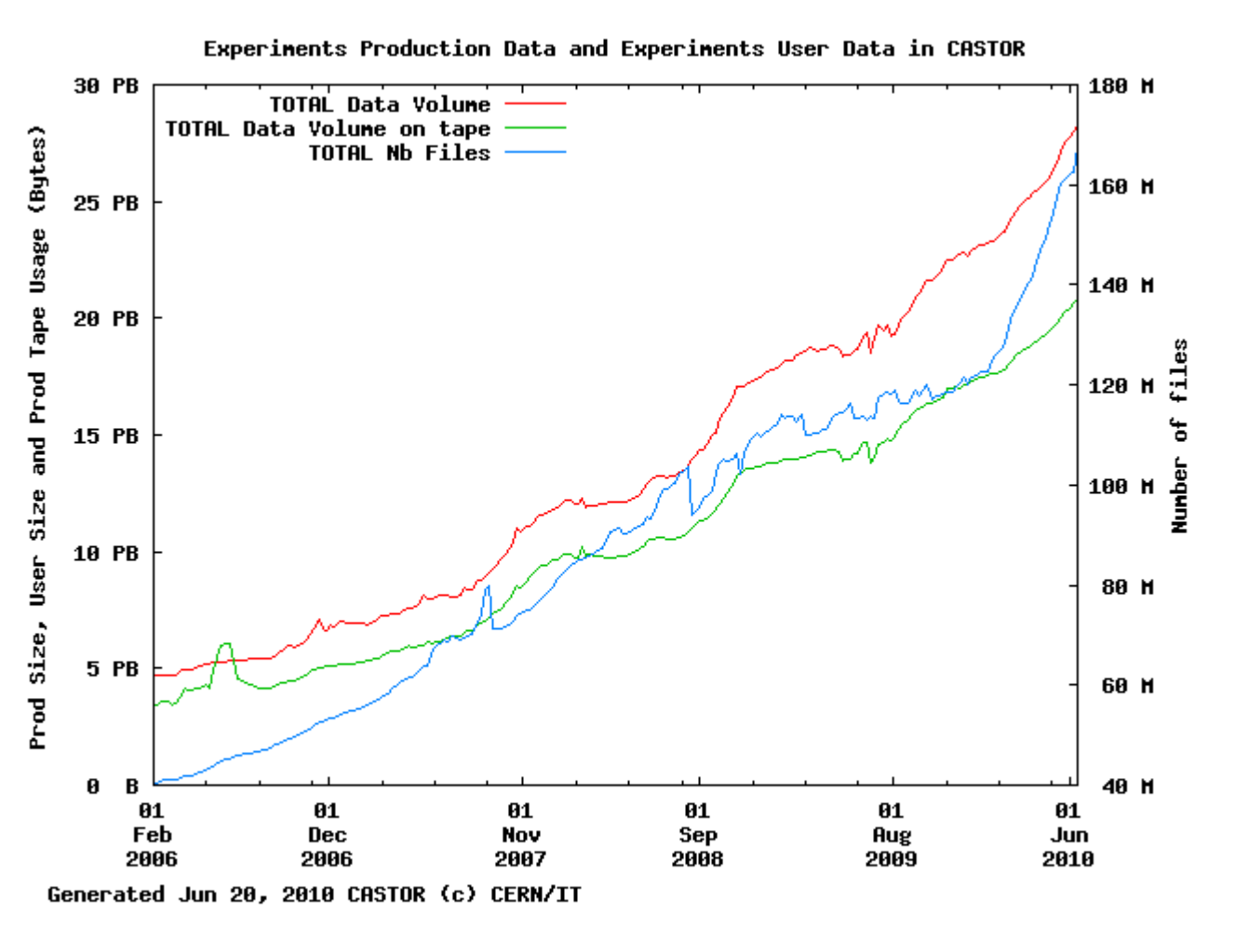
Processing Facility: lxbatch

- Today about 4’000 nodes with 31’000 processing cores
- Jobs are submitted from lxplus, or channeled through GRID interfaces world-wide
- About 150000 user jobs per day
- Reading and writing > 1 PB per day
- Uses LSF as a management tool to schedule the various jobs from a large number of users
- Expect a demand growth rate of ~30% per year

Data Storage (1)

-

Data Storage (2)



165 million files; 28 PB of data on tape already today

Storage

Users

Administrative data

1 million electronic documents

5’5000 electronic signatures per month

60’000 emails per day

250’000 orders per year

> 1’000 million user files

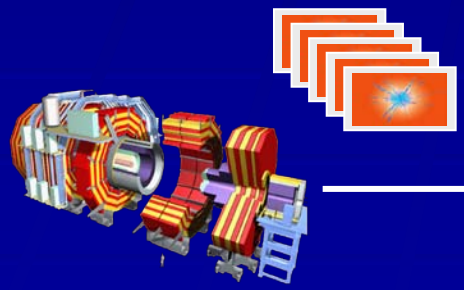
*accessibility → 24*7*52 = always*

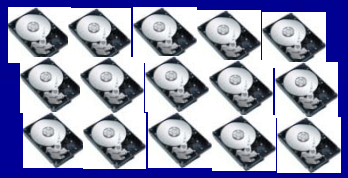
tape storage → forever

backup

per hour and per day

continuous storage





Physics data

21’000 TB and

50 million files per year

Miscellaneous Services

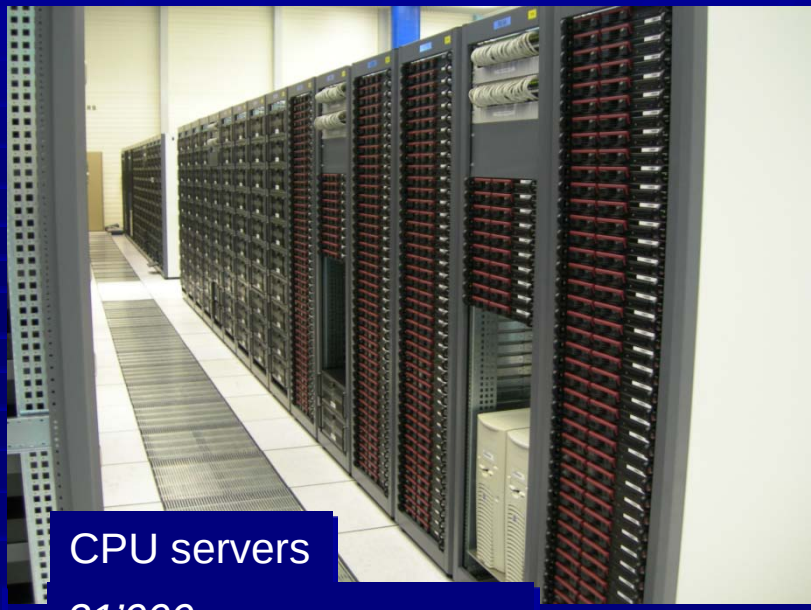
- TWiki: Collaborative Web space
  - About 200 Twikis, between just a few and more than 6’000 Twiki items each
- Version control services
  - CVS (repository based on AFS)
  - LCGCVS (repository on local disks)
  - SVN with SVNWEB/TRAC
- …

CERN Computer Centre

ORACLE Data base servers

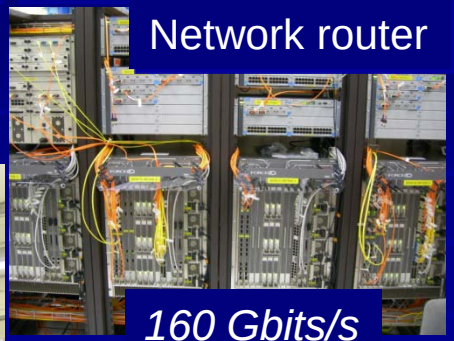
240 CPU and disk server
200 TB

2.9 MW electricity
and cooling
2700 m^2



CPU servers

31'000 processor cores

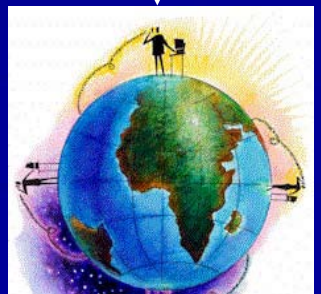


Network router

160 Gbits/s

Tape servers and tape libraries

160 tape drives, 50000 tapes
40000 TB capacity



Disk servers

1800 NAS servers,
21'000 TB, 30'000 disks


BNL
New York
ASGC/Taipei
CCIN2P3/Lyon
TRIUMF
Vancouver
NIKHEF/SARA
Amsterdam
FNAL
Chicago
RAL
Rutherford
THE REST OF THE WORLD...
CNAF
Bologna
PIC
Barcelona
NDGF
Nordic countries
TIER2s
CERN
FZK

World-wide computing

- CERN's resources by far not sufficient
- World-wide collaboration between computer centres
  - WLCG: World-wide LHC Computing Grid
- Web, Grids, EGEE, EMI, clouds, WLCG, …: See Markus Schulz' lecture next week

Future (1)

- Is IT growth sustainable?
  - Demands continue to rise exponentially
  - Even if Moore's law continues to apply, data centres will need to grow in number and size
  - IT already consuming 2% of world's energy – where do we go?
  - How to handle growing demands within a given data centre?
    - Demands evolve very rapidly, technologies less so, infrastructure even at a slower pace – how to best match these three?

Future (2)

- IT: Ecosystem of
  - Hardware
  - OS software and tools
  - Applications
- Evolving at different paces: hardware fastest, applications slowest
  - How to make sure at any given time that they match reasonably well?

Future (3)

- Example: single-core to multi-core to many-core
  - Most HEP applications currently single-threaded
  - Consider server with two quad-core CPUs as eight independent execution units
    - Model does not scale much further
  - Need to adapt applications to many-core machines
    - Large, long effort

Conclusions

- The Large Hadron Collider (LHC) and its experiments is a very data (and compute) intensive project
- Implemented using right blend of new technologies and commodity approaches
- Scaling computing to the requirements of LHC is hard work
- IT power consumption/efficiency is a primordial concern
- We are steadily taking collision data at 2 * 3.5 TeV, and have the capacity in place for dealing with this
- We are on track for further ramp-ups of the computing capacity for future requirements


Thank you

More Information (1)

IT department

http://it-div.web.cern.ch/it-div/

http://it-div.web.cern.ch/it-div/need-help/

Monitoring

http://sls.cern.ch/sls/index.php

http://lemonweb.cern.ch/lemon-status/

http://gridview.cern.ch/GRIDVIEW/dt_index.php

http://gridportal.hep.ph.ic.ac.uk/rtm/

Lxplus

http://plus.web.cern.ch/plus/

Lxbatch

http://batch.web.cern.ch/batch/

CASTOR

http://castor.web.cern.ch/castor/

More Information (2)

Windows, Web, Mail

https://winservices.web.cern.ch/winservices/

Grid

http://lcg.web.cern.ch/LCG/public/default.htm

http://www.eu-egee.org/

Computing and Physics

http://www.particle.cz/conferences/chep2009/programme.aspx

In case of further questions don't hesitate to contact me:

Helge.Meinhard@cern.ch