Location (1)

Building 513 (opposite of restaurant no. 2)
Building 513 (1)

Large building with 2700 m² surface for computing equipment, capacity for 2.9 MW electricity and 2.9 MW air and water cooling
Building 513 (2) - Ongoing Work

**Reasons for the upgrade:**

- **Cooling:**
  - Insufficient cooling for critical UPS room
  - CC not cooled when running on UPS without diesel
  - Insufficient cooling when running on diesel
    - Pumps & Ventilation Units running but no chiller and insufficient stored cold water

- **Power**
  - Insufficient critical power available
  - No redundancy for critical UPS (> 240 kW) - currently running at 340kW
  - No redundancy for physics UPS (> 2.7 MW) - aiming to run at 2.9MW by end of year

- **Other**
  - Limited fire protection in B513
  - Critical areas and physics areas strongly coupled
    - Share the same locations, cooling infrastructure and fire risks
Building 513 (3) - Ongoing Work

- Scope of the upgrade:
  - Dedicated cooling infrastructure for critical equipment (decoupled from physics)
    - New building for cooling system (construction that is underway in front of the building)
  - New dedicated room for critical equipment, new electrical rooms and critical ventilation systems in ‘Barn’
  - Critical equipment which cannot be moved to new rooms to have new dedicated cooling
    - Networking area and telecoms rooms
  - Increase in critical UPS power to 600kW (with new critical UPS room) and overall power to 3.5MW
  - Restore N+1 redundancy for all UPS systems
Building 513 (4) - Ongoing Work
Location (2)

- Building 613: Small machine room for tape libraries (about 200 m from building 513)
- Hosting centre about 15 km from CERN: 35 m², about 100 kW, critical equipment
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 (June 2011)

- **Data Centre Operations (Tier 0)**
  - 24x7 operator support and System Administration services to support 24x7 operation of all IT services.
  - Hardware installation & retirement
    - ~7,000 hardware movements/year; ~1800 disk failures/year
  - Management and Automation framework for large scale Linux clusters

<table>
<thead>
<tr>
<th>Racks</th>
<th>828</th>
</tr>
</thead>
<tbody>
<tr>
<td>Servers</td>
<td>11,728</td>
</tr>
<tr>
<td>Processors</td>
<td>15,694</td>
</tr>
<tr>
<td>Cores</td>
<td>64,238</td>
</tr>
<tr>
<td>HEPSpec06</td>
<td>482,507</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disks</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw disk capacity (TiB)</td>
<td>63,289</td>
</tr>
<tr>
<td>Memory modules</td>
<td>56,014</td>
</tr>
<tr>
<td>Memory capacity (TiB)</td>
<td>158</td>
</tr>
<tr>
<td>RAID controllers</td>
<td>3,749</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tape Drives</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tape Drives</td>
<td>160</td>
</tr>
<tr>
<td>Tape Cartridges</td>
<td>45000</td>
</tr>
<tr>
<td>Tape slots</td>
<td>56000</td>
</tr>
<tr>
<td>Tape Capacity (TiB)</td>
<td>34000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High Speed Routers (640 Mbps → 2.4 Tbps)</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet Switches</td>
<td>350</td>
</tr>
<tr>
<td>10 Gbps ports</td>
<td>2000</td>
</tr>
<tr>
<td>Switching Capacity</td>
<td>4.8 Tbps</td>
</tr>
<tr>
<td>1 Gbps ports</td>
<td>16,939</td>
</tr>
<tr>
<td>10 Gbps ports</td>
<td>558</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IT Power Consumption</th>
<th>2456 KW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Power Consumption</td>
<td>3890 KW</td>
</tr>
</tbody>
</table>
Infrastructure Services

Software environment and productivity tools

User registration and authentication
22’000 registered users

Mail
2 million emails/day, 99% spam
18’000 mail boxes

Web services
10’000 web sites

Tool accessibility
Windows, Office, CadCam, …

Home directories (DFS, AFS)
~400 TB, backup service
~ 2 billion files

PC management
Software and patch installations

Infrastructure needed:
> 400 servers
Network Overview

ATLAS

Central, high speed network backbone

Experiments

LHCb

All CERN buildings
12’000 active users

CMS

Computer centre processing clusters

World Wide Grid centres
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 125 ORACLE database instances on > 400 service nodes, total ~ 100 TB
  - 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
  - ...

21 July 2011
Physics computing - Helge Meinhard
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

- Simulation
- Reconstruction
- Analysis

Event filter (selection & reconstruction)

Raw data → Event reprocessing

Event summary data → Batch physics analysis

Analysis objects (extracted by physics topic)

Interactive physics analysis

Processed data
Data Flow - online

Detector
150 million electronics channels

Level 1 Filter and Selection
1 PBytes/s

High Level Filter and Selection
150 GBytes/s

CERN computer centre

Limits:
Essentially the budget and the downstream data flow pressure

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

O(1000) servers for processing, Gbit Ethernet Network

N x 10 Gbit links to the computer centre

21 July 2011
Data Flow - offline

4 detectors

1000 million events/s

Filter and first selection

1...2 GB/s

Create sub-samples

World-wide analysis

Store on disk and tape

Export copies

Physics
Explanation of nature

\[ \sigma \approx \sigma^0 \frac{s \Gamma_Z^2}{(s-m_Z^2)^2 + s^2 \Gamma_Z^4} \]

\[ \sigma^0 = \frac{12\pi}{m_Z} \frac{\Gamma_Z \Gamma_{\text{eff}}}{\Gamma_{\text{eff}}^2} \]

\[ \Gamma_{\text{eff}} = \frac{G_F m_Z^5}{16\pi^3} \approx (v_s^2 + a_s^2) N_{\text{col}} \]
## SI Prefixes

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Symbol</th>
<th>$1000^m$</th>
<th>$10^n$</th>
<th>Decimal</th>
<th>Short scale</th>
<th>Long scale</th>
<th>Since[11]</th>
</tr>
</thead>
<tbody>
<tr>
<td>yotta</td>
<td>Y</td>
<td>$1000^8$</td>
<td>$10^{24}$</td>
<td>1,000,000,000,000,000,000,000,000,000</td>
<td>Septillion</td>
<td>Quadrillion</td>
<td>1991</td>
</tr>
<tr>
<td>zetta</td>
<td>Z</td>
<td>$1000^7$</td>
<td>$10^{21}$</td>
<td>1,000,000,000,000,000,000,000</td>
<td>Sextillion</td>
<td>Trilliard</td>
<td>1991</td>
</tr>
<tr>
<td>exa</td>
<td>E</td>
<td>$1000^6$</td>
<td>$10^{18}$</td>
<td>1,000,000,000,000,000,000</td>
<td>Quintillion</td>
<td>Trillion</td>
<td>1975</td>
</tr>
<tr>
<td>peta</td>
<td>P</td>
<td>$1000^5$</td>
<td>$10^{15}$</td>
<td>1,000,000,000,000,000</td>
<td>Quadrillion</td>
<td>Billiard</td>
<td>1975</td>
</tr>
<tr>
<td>tera</td>
<td>T</td>
<td>$1000^4$</td>
<td>$10^{12}$</td>
<td>1,000,000,000,000</td>
<td>Trillion</td>
<td>Billion</td>
<td>1960</td>
</tr>
<tr>
<td>giga</td>
<td>G</td>
<td>$1000^3$</td>
<td>$10^9$</td>
<td>1,000,000,000</td>
<td>Billion</td>
<td>Milliard</td>
<td>1960</td>
</tr>
<tr>
<td>mega</td>
<td>M</td>
<td>$1000^2$</td>
<td>$10^6$</td>
<td>1,000,000</td>
<td>Million</td>
<td></td>
<td>1960</td>
</tr>
<tr>
<td>kilo</td>
<td>k</td>
<td>$1000^1$</td>
<td>$10^3$</td>
<td>1,000</td>
<td>Thousand</td>
<td></td>
<td>1795</td>
</tr>
<tr>
<td>hecto</td>
<td>h</td>
<td>$1000^{2/3}$</td>
<td>$10^2$</td>
<td>100</td>
<td>Hundred</td>
<td></td>
<td>1795</td>
</tr>
<tr>
<td>deca</td>
<td>da</td>
<td>$1000^{1/3}$</td>
<td>$10^1$</td>
<td>10</td>
<td>Ten</td>
<td></td>
<td>1795</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1000^0$</td>
<td>$10^0$</td>
<td>1</td>
<td>One</td>
<td></td>
<td></td>
</tr>
<tr>
<td>deci</td>
<td>d</td>
<td>$1000^{-1/3}$</td>
<td>$10^{-1}$</td>
<td>0.1</td>
<td>Tenth</td>
<td></td>
<td>1795</td>
</tr>
<tr>
<td>centi</td>
<td>c</td>
<td>$1000^{-2/3}$</td>
<td>$10^{-2}$</td>
<td>0.01</td>
<td>Hundredth</td>
<td></td>
<td>1795</td>
</tr>
<tr>
<td>milli</td>
<td>m</td>
<td>$1000^{-1}$</td>
<td>$10^{-3}$</td>
<td>0.001</td>
<td>Thousandth</td>
<td></td>
<td>1795</td>
</tr>
<tr>
<td>micro</td>
<td>μ</td>
<td>$1000^{-2}$</td>
<td>$10^{-6}$</td>
<td>0.000001</td>
<td>Millionth</td>
<td></td>
<td>1960[2]</td>
</tr>
<tr>
<td>nano</td>
<td>n</td>
<td>$1000^{-3}$</td>
<td>$10^{-9}$</td>
<td>0.000000001</td>
<td>Billionth</td>
<td>Milliardth</td>
<td>1960</td>
</tr>
<tr>
<td>pico</td>
<td>p</td>
<td>$1000^{-4}$</td>
<td>$10^{-12}$</td>
<td>0.00000000001</td>
<td>Trillionth</td>
<td>Billionth</td>
<td>1960</td>
</tr>
<tr>
<td>femto</td>
<td>f</td>
<td>$1000^{-5}$</td>
<td>$10^{-15}$</td>
<td>0.0000000000001</td>
<td>Quadrillionth</td>
<td>Billiardth</td>
<td>1964</td>
</tr>
<tr>
<td>atto</td>
<td>a</td>
<td>$1000^{-6}$</td>
<td>$10^{-18}$</td>
<td>0.000000000000001</td>
<td>Quintillionth</td>
<td>Trillionth</td>
<td>1964</td>
</tr>
<tr>
<td>zepto</td>
<td>z</td>
<td>$1000^{-7}$</td>
<td>$10^{-21}$</td>
<td>0.000000000000000001</td>
<td>Sextillionth</td>
<td>Trilliardth</td>
<td>1991</td>
</tr>
<tr>
<td>yocto</td>
<td>y</td>
<td>$1000^{-8}$</td>
<td>$10^{-24}$</td>
<td>0.000000000000000001</td>
<td>Septillionth</td>
<td>Quadrillionth</td>
<td>1991</td>
</tr>
</tbody>
</table>

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

- Compare with (numbers from mid 2010):
  - 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

Components

- CPU, disk server
- Cluster, local fabric
- World-wide cluster

Hardware

- CPU, disk, memory, mainbord
- Network, interconnects
- Wide area network

Software

- Operating system, device drivers
- Resource management software
- Grid and cloud management software

Complexity / scale
Computing Building Blocks

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

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

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

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

Market trends more important
than technology trends

Always watch TCO:
Total Cost of Ownership
Hardware Management

- Almost 12'000 servers installed in centre
- Assume 3...4 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
  - ~3’500 nodes installed per year
  - ~3’500 nodes removed per year
  - Installation in racks, cabling, automatic installation, Linux software environment
Functional Units

- Detectors
- Disk storage
- Event processing capacity
- CPU server
- Meta-data storage
- Data bases
- Tape storage
  - ‘Active’ archive and backup
- Data import and export

21 July 2011

Physics computing - Helge Meinhard
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?
Here is my program and I want to analyse the ATLAS data(130,209),(286,454) from the special run on June 16th 14:45h or all data with detector signature X

‘Batch’ system to decide where is free computing time

Management software

Processing nodes (CPU servers)

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

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

Disk storage
Job Data and Control Flow (2)

CERN installation

Repositories
- code
- metadata
- …….

Data processing
lxbatch

Interactive
lxplus

Users

Bookkeeping
Data base

Disk storage

CASTOR
Hierarchical Mass Storage Management System (HSM)

Tape storage
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
- 2,200+ switches (increasing)
- 50,000 active user devices (exploding)
- 80,000 sockets - 5,000 km of UTP cable
- 5,000 km of fibers (CERN owned)
- 180 Gbps of WAN connectivity
Interactive Login Service: lxplus

- Interactive compute facility
- > 80 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.

Interactive users per server

![Graph showing number of users per week]

- Number of Users: average 58.3, max 77.5, min 37.4, current 71.1
Processing Facility: lxbatch

- Today about 3’700 processing nodes with 33’000 processing cores
- Jobs are submitted from lxplus, or channeled through GRID interfaces world-wide
- About 200’000 user jobs per day recently
- 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)

- Large disk cache in front of a long term storage tape system
- 1’800 disk servers with 21 PB usable capacity
- Redundant disk configuration, 2...3 disk failures per day
  - part of the operational procedures
- Logistics again: need to store all data forever on tape
  - > 20 PB storage added per year, plus a complete copy every 4 years (“repack”, change of technology)
- CASTOR data management system, developed at CERN, manages the user IO requests
- Expect a demand growth rate of ~30% per year
260 million files; 36 PB of data on tape already today

Experiments Production Data and Experiments User Data in CASTOR

Total: ~ 47 PB
On tape: ~ 36 PB

In: avg 2.6 GB/s, peak 6.4 GB/s
Out: avg 7 GB/s, peak 13.5 GB/s (binning: day)
Storage

Administrative data
- 3.3 million electronic documents
- 280,000 electronic docs per year
- 55,000 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

Physics data
- 36,000 TB and
- 50 million files per year

backup per hour and per day

continuous storage

Users

21 July 2011
Miscellaneous Services

- **TWiki: Collaborative Web space**
  - About 250 Twikis, between just a few and more than 8’000 Twiki items each

- **Version control services**
  - SVN with SVNWEB/TRAC
  - CVS (repository based on AFS)
    - Legacy service, to be phased out in 2013

- ...
CERN Computer Centre

ORACLE Database servers

- ~400 servers
- ~300 TB raw

CPU servers
- 33,000 processor cores

Disk servers
- 1800 NAS servers, 21,000 TB, 40,000 disks

Tape servers and tape libraries
- 160 tape drives, 53,000 tapes
- 71,000 TB capacity, 53,000 TB used

Network router
- 160 Gbits/s

2.9 MW electricity and cooling
- 2700 m²

~400 servers
~300 TB raw
THE REST OF THE WORLD...
World-wide computing

- CERN’s resources by far not sufficient
- World-wide collaboration between computer centres
  - WLCG: World-wide LHC Computing Grid
- Web, Grids, clouds, WLCG, EGEE, EGI, EMI, ...: See Markus Schulz’ lecture on August 2nd
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 with very high efficiency, 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


Computing and Physics

http://event.twgrid.org/chep2010/

In case of further questions don’t hesitate to contact me:
Helge.Meinhard@cern.ch