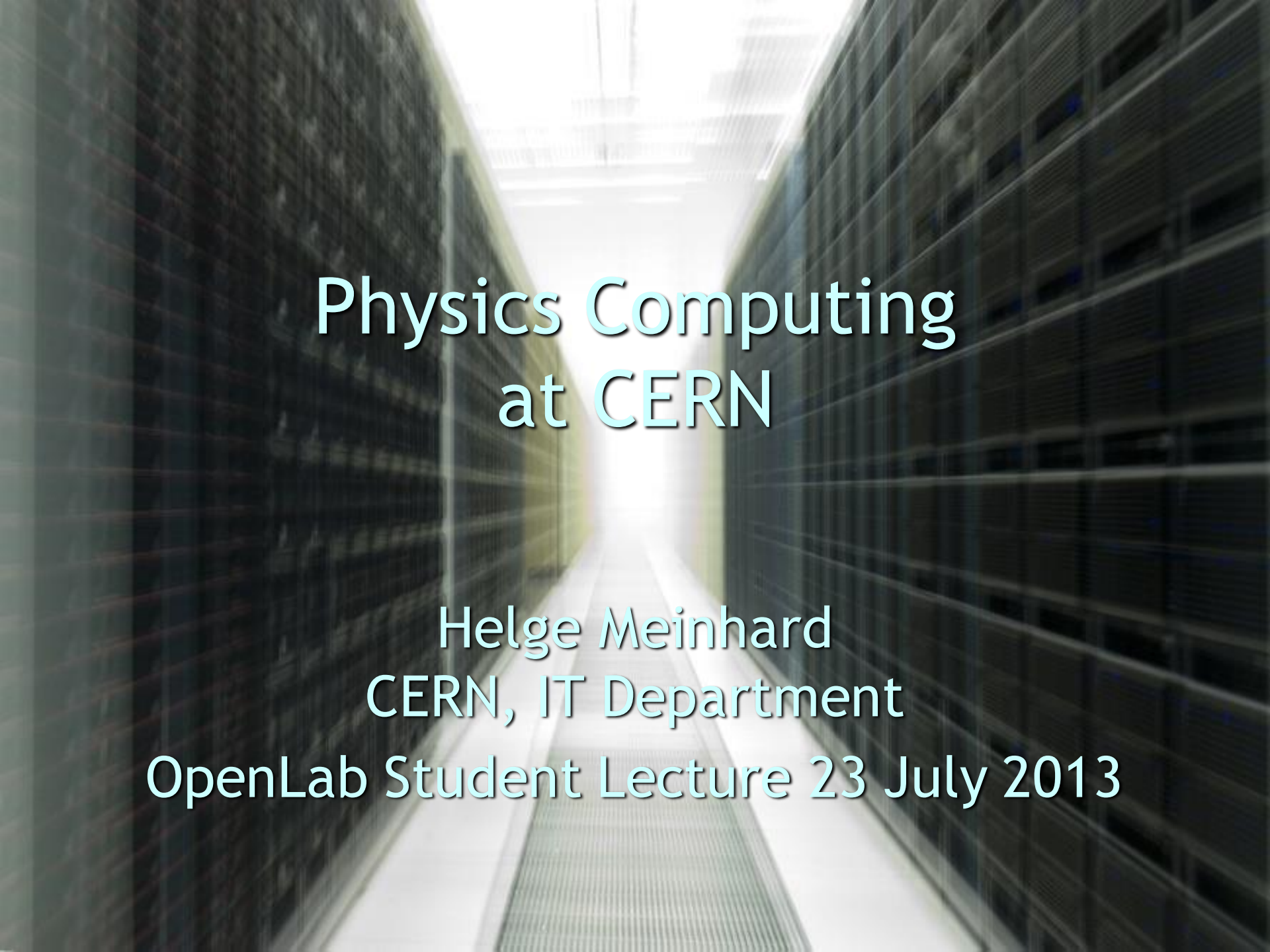


Physics Computing at CERN

Helge Meinhard
CERN, IT Department

OpenLab Student Lecture 23 July 2013

Outline

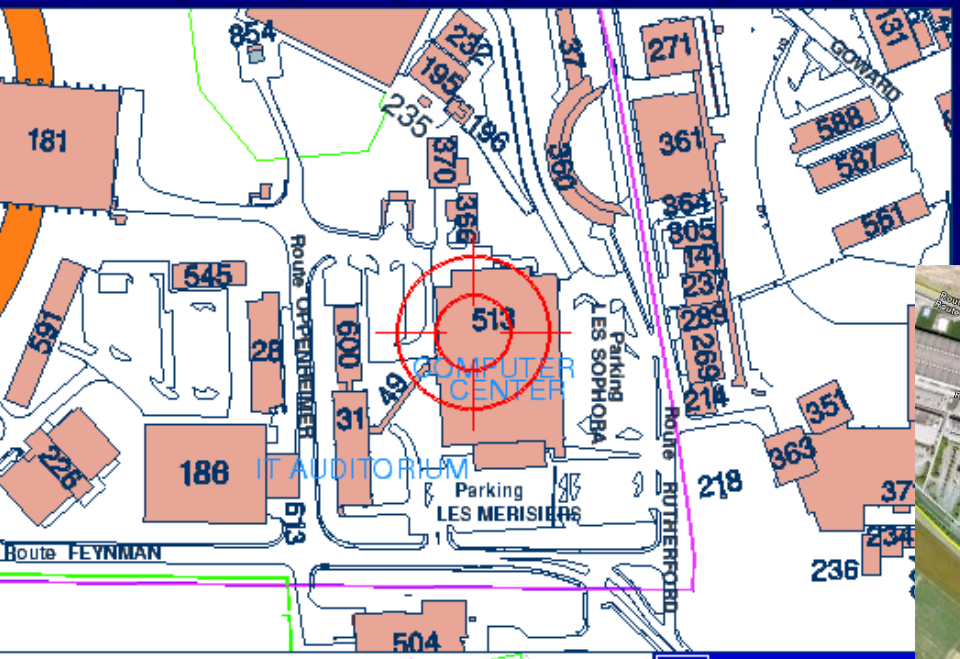
- CERN's computing facilities and hardware
- Service categories, tasks
- Infrastructure, networking, databases
- HEP analysis: techniques and data flows
- Network, plus, batch, storage
- Between HW and services: Agile Infrastructure
- References

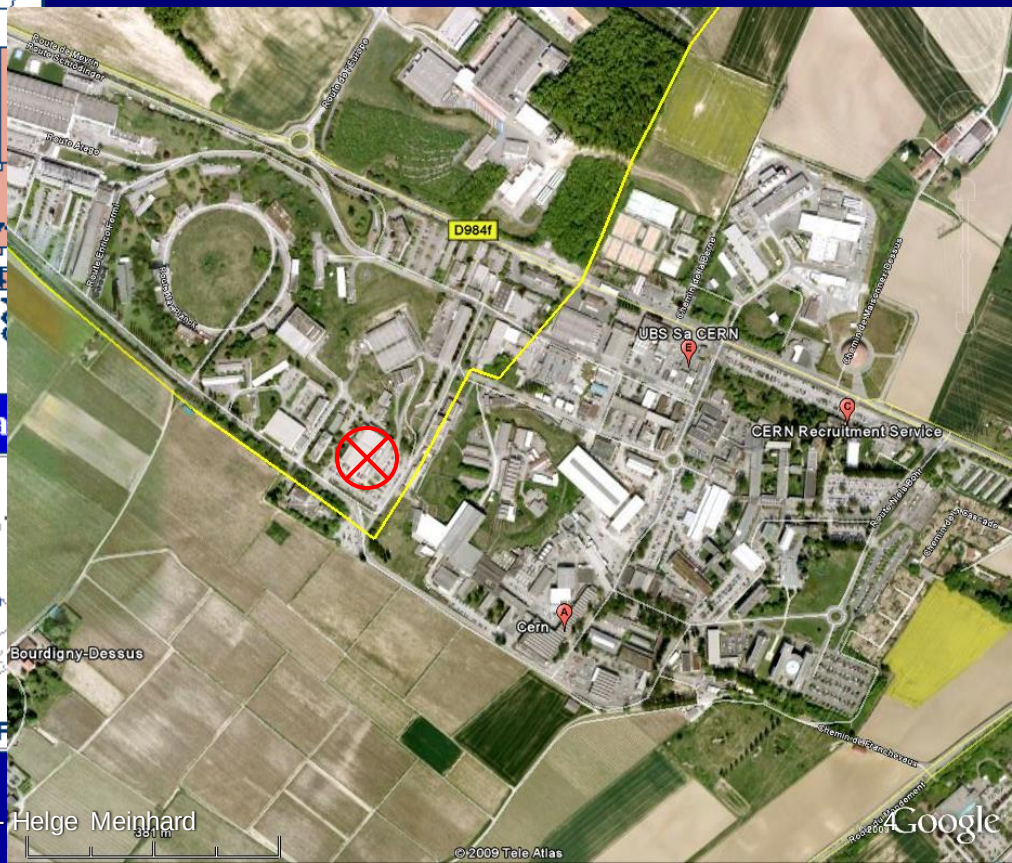
Outline

- **CERN's computing facilities and hardware**
- Service categories, tasks
- Infrastructure, networking, databases
- HEP analysis: techniques and data flows
- Network, plus, batch, storage
- Between HW and services: Agile Infrastructure
- References

Location (1)

Building 513 (opposite of restaurant no. 2)





Building 513

Large building with 2700 m^2 surface for computing equipment, capacity for 3.5 MW electricity and 3.5MW air and water cooling

Chillers

Transformers

Building 513 – Latest Upgrade

- Scope of the upgrade:
  - Increase critical UPS power to 600 kW (with new critical UPS room) and overall power to 3.5 MW (from 2.9 MW)
  - New dedicated room for critical equipment, new electrical rooms and critical ventilation systems in 'Barn'
  - Dedicated cooling infrastructure for critical equipment (decoupled from physics)
    - New building for cooling system
  - Critical equipment which cannot be moved to new rooms to have new dedicated cooling
    - Networking area and telecoms rooms
  - Restore N+1 redundancy for all UPS systems
  - New critical room operational since January 2013
  - Critical services gradually being migrated into this room and expected to be completed early 2014
  - Last finishing touches to be completed soon

New HVAC building

Other facilities (1)

- Building 613: Small machine room for tape libraries (about 200 m from building 513)
- Hosting centre about 15 km from CERN: 35 m^2, about 100 kW, critical equipment

Other facilities – Wigner (1)

- Additional resources for future needs
  - 8’000 servers now, 15’000 estimated by end 2017
- Studies in 2008 into building a new computer centre on the CERN Prevessin site
  - Too costly
- In 2011, tender run across CERN member states for remote hosting
  - 16 bids
- In March 2012, Wigner Institute in Budapest, Hungary selected; contract signed in May 2012

Other facilities – Wigner (2)

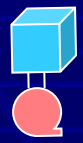
- Timescales Data Centre
  - Construction started in May 2012
  - First room available January 2013
  - Inauguration with Hungarian Prime Minister June 18th 2013
  - Construction finished at end of June 2013
- Timescales Services
  - First deliveries (network and servers) 1Q13
  - 2x100Gbps links operation in February 2013
    - Round trip latency ~ 25ms
  - Servers installed, tested and ready for use May 2013
  - Expected to be put into production end July 2013
  - Large ramp up foreseen during 2014/2015
- This will be a “hands-off” facility for CERN
  - Wigner manage the infrastructure and hands-on work
  - We do everything else remotely

Other facilities – Wigner (3)

Other facilities – Wigner (4)

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, six or eight cores,
3...4 GB memory per core

Disk server =
CPU server
+ RAID or disk controller
+ ~ 16...36 internal or external SATA disks

Market trends more important than technology trends

Always watch TCO: Total Cost of Ownership

CERN CC currently

- 24x7 operator support and System Administration services to support 24x7 operation of all IT services
- Hardware installation, retirement and repair
  - Per year: ~4'000 hardware movements; ~2'700 hardware interventions/repairs

Metrics	Number
Number of 10GB NICs	2'782
Number of 1GB NICs	18'051
Number of cores	88'014
Number of disks	79'167
Number of memory modules	63'028
Number of processors	17'196
Number of "boxes"	10'035
Total disk space (TiB)	123'975.41
Total memory capacity (TiB)	296.12

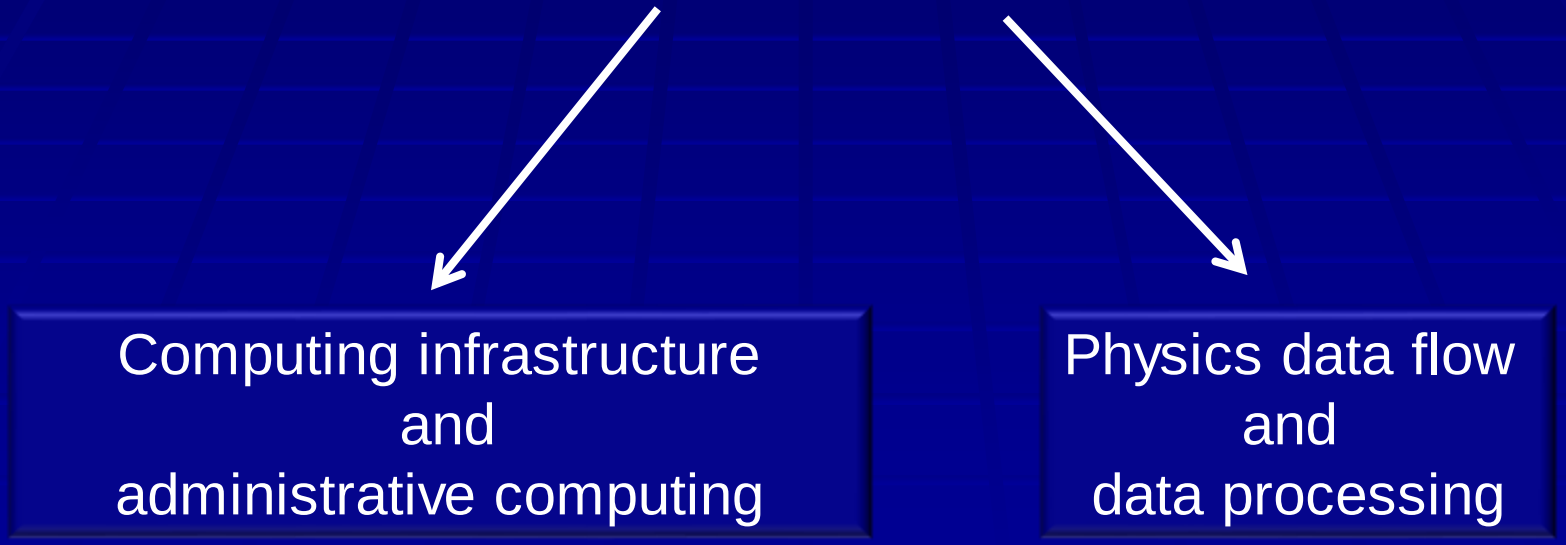
Preliminary – probably incomplete – to be checked

Outline

- CERN's computing facilities and hardware
- **Service categories, tasks**
- Infrastructure, networking, databases
- HEP analysis: techniques and data flows
- Network, plus, batch, storage
- Between HW and services: Agile Infrastructure
- References

Computing Service Categories

Two coarse grain computing categories



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
  - Computing equipment for processing, storage and databases
  - Management and monitoring software
  - Maintenance and operations
  - Authentication and security

Outline

- CERN's computing facilities and hardware
- Service categories, tasks
- **Infrastructure, networking, databases**
- HEP analysis: techniques and data flows
- Network, plus, batch, storage
- Between HW and services: Agile Infrastructure
- References

Infrastructure Services

Software environment and productivity tools

User registration and authentication
35'600 registered users, 48'300 accounts

Mail
230'000 emails/day, 64% spam
31'000 mail boxes

Web services
12'500 web sites

Tool accessibility
Windows, Office, CadCam, …

Home directories (DFS, AFS)
~550 TB, backup service
~ 3 billion files

PC management
Software and patch installations

Infrastructure needed :
> 700 servers

Network Overview

Experiments

All CERN buildings
12’000 active users

Central, high speed network backbone

Wigner

World Wide Grid centres

Computer centre processing clusters

Bookkeeping: Database Services

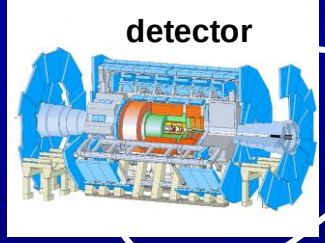
- Numerous ORACLE data base instances, total > 370 TB data (not counting backups)
  - Bookkeeping of physics events for the experiments
  - Meta data for the physics events (e.g. detector conditions)
  - Management of data processing
  - Highly compressed and filtered event data
  - …
- LHC machine parameters, monitoring data
- Human resource information
- Financial bookkeeping
- Material bookkeeping and material flow control
- LHC and detector construction details
- …

Outline

- CERN's computing facilities and hardware
- Service categories, tasks
- Infrastructure, networking, databases
- **HEP analysis: techniques and data flows**
- Network, plus, batch, storage
- Between HW and services: Agile Infrastructure
- References

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 summary data

processed data

event reprocessing

batch physics analysis

analysis

analysis objects (extracted by physics topic)

event simulation

simulation

interactive physics analysis

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

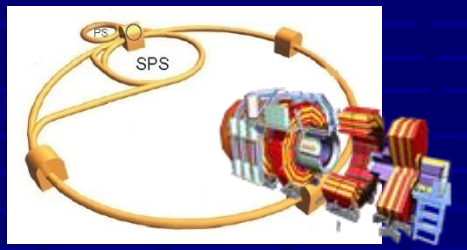
N x 10 Gbit links to the computer centre

CERN computer centre

Constraints:

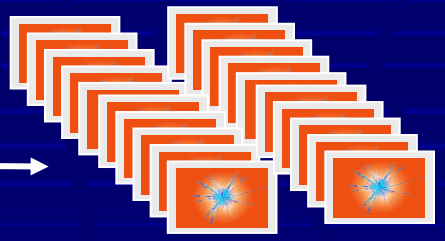
- *Budget*
- *Physics objectives*
- *Downstream data flow pressure*

Data Flow – offline



4 detectors

1000 million events/s



Filter and first selection

1…4 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. 23 July 2013 recognition of the micron by the CGPM was approved in 1967

Source: wikipedia.org

Data Volumes at CERN

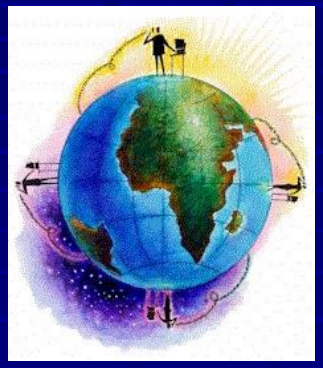
- Original estimate: 15 Petabytes / year
  - Tower of CDs: which height?
  - Stored cumulatively over LHC running
  - Only real data and derivatives
    - Simulated data not included
      - Total of simulated data even larger
- Written in 2011: 22 PB
- Written in 2012: 30 PB

- *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*

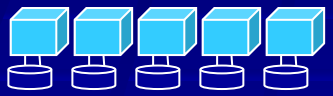
… growing exponentially…

Functional Units

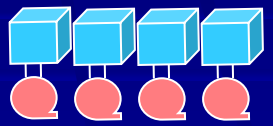
Detectors



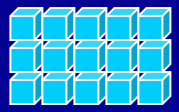
Data import
and export



Disk storage



Tape storage
'Active' archive and backup



Event processing capacity
CPU server

Meta-data storage
Data bases

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

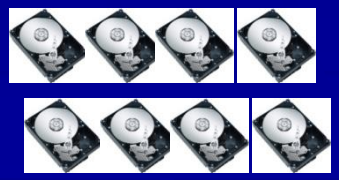
Processing nodes (CPU servers)

'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

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

Users

Repositories
- code
- metadata
- .....

Interactive
lxplus

Data processing
lxbatch

Bookkeeping
Data base

Disk storage

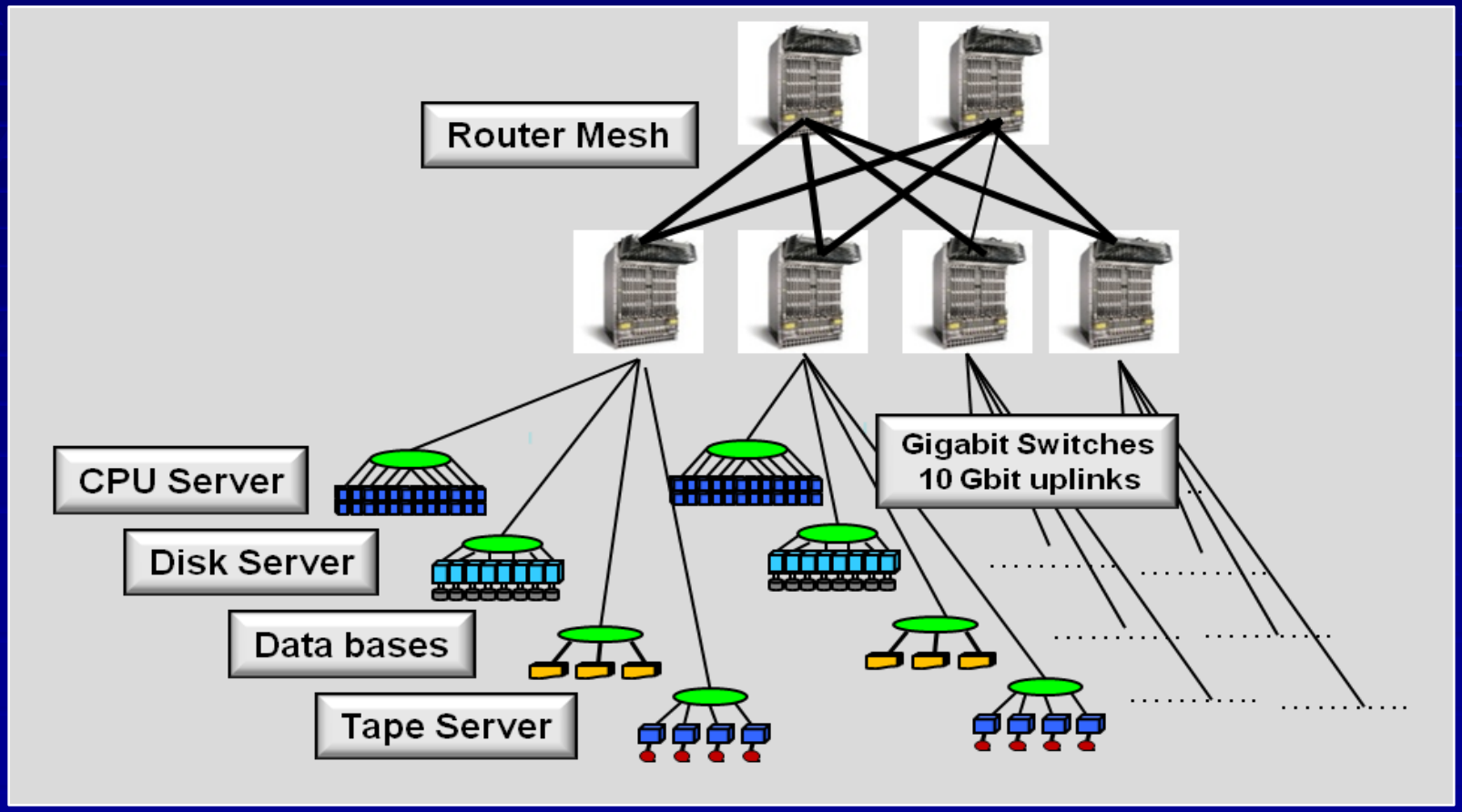
Tape storage

CASTOR
Hierarchical Mass Storage Management System (HSM)

Outline

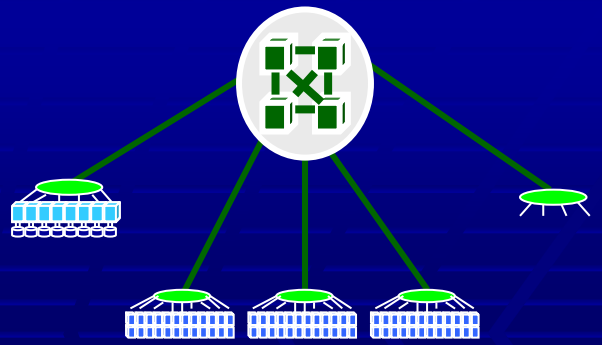
- CERN's computing facilities and hardware
- Service categories, tasks
- Infrastructure, networking, databases
- HEP analysis: techniques and data flows
- **Network, plus, batch, storage**
- Between HW and services: Agile Infrastructure
- References

CERN Farm Network



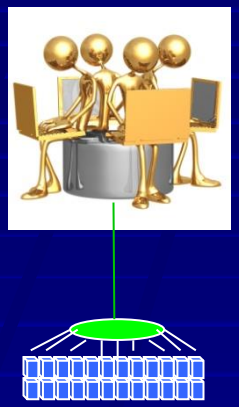
- Switches in the distribution layer close to servers
- (Possibly multiple) 10 Gbit uplinks
- 1 Gbit or 10 Gbit to server

CERN Overall Network

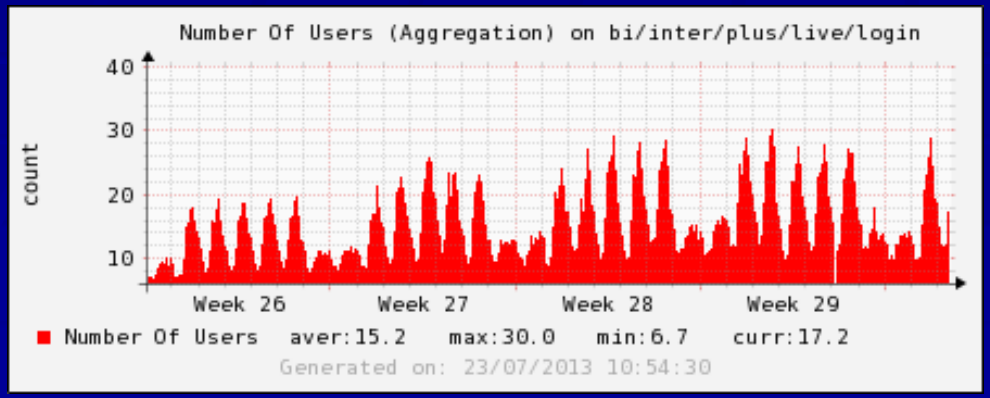


- Hierarchical network topology based on Ethernet: core, general purpose, LCG, technical, experiments, external
- 180+ very high performance routers
- > 6’000+ subnets
- 3’600+ switches (increasing)
- 75’000 active user devices (exploding)
- 80’000 sockets – 5’000 km of UTP cable
- 5’000 km of fibers (CERN owned)
- 200 Gbps of WAN connectivity

Interactive Login Service: lxplus



Interactive users



- Interactive compute facility
- 45 virtual CPU servers running RHEL 6 (default target)
- 45 CPU servers running Linux (RHEL 5 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, …

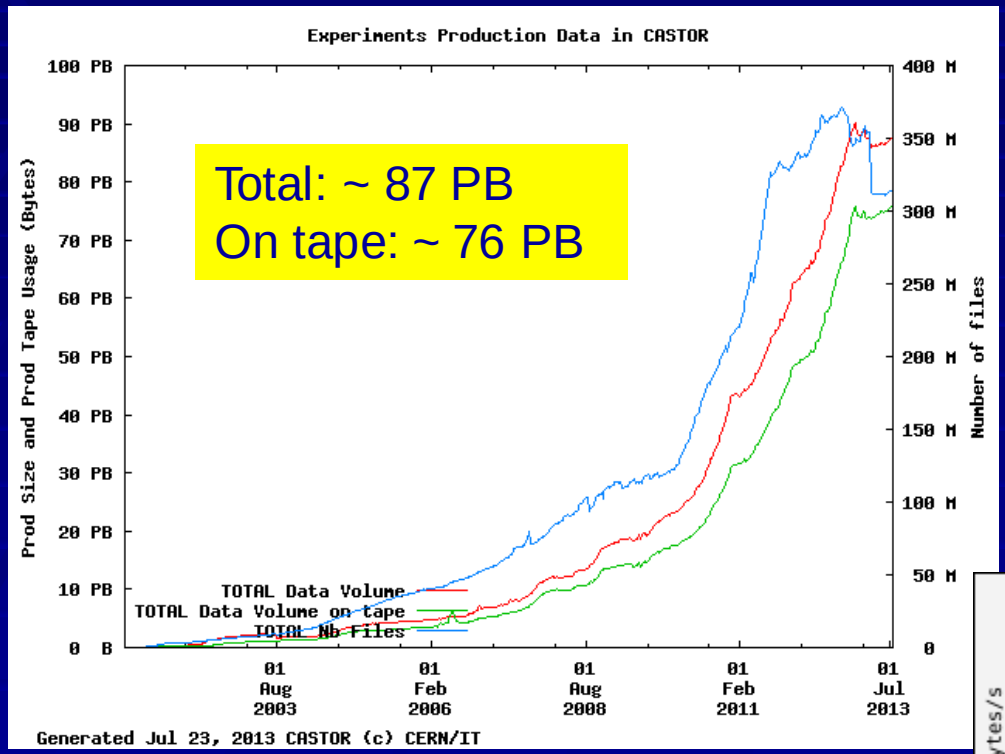
Processing Facility: lxbatch

- Today about 3’650 processing nodes
  - 3’350 physical nodes, SLC5, 48’000 job slots
  - 300 virtual nodes, SLC6, 8’000 job slots
- Jobs are submitted from lxplus, or channeled through GRID interfaces world-wide
- About 300’000 user jobs per day recently
- Reading and writing up to 2 PB per day
- Uses IBM/Platform Load Sharing Facility (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)

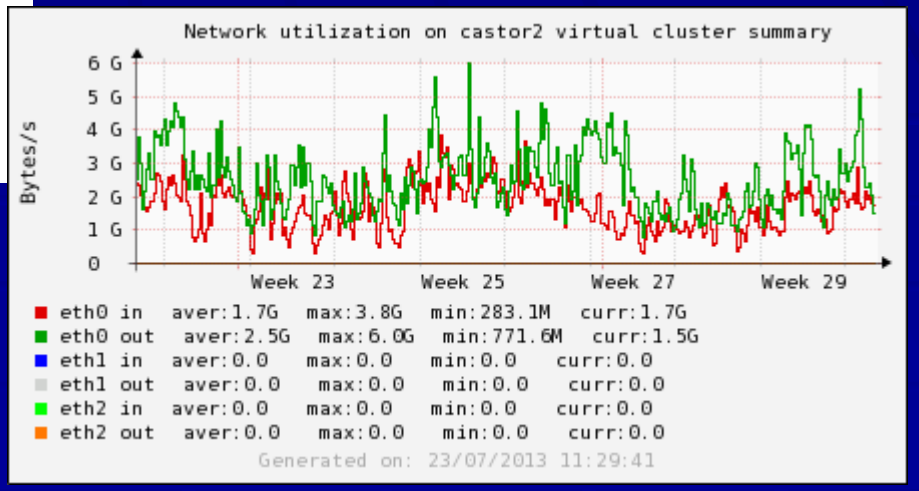
- Large disk cache in front of a long term storage tape system: CASTOR data management system, developed at CERN, manages the user IO requests
  - 535 disk servers with 17 PB usable capacity
  - About 65 PB on tape
  - Redundant disk configuration, 2…3 disk failures per day
    - part of the operational procedures
- Logistics again: need to store all data forever on tape
  - > 25 PB storage added per year, plus a complete copy every 4 years (“repack”, change of technology)
- Disk-only use case (analysis): EOS
- Expect a demand growth rate of ~30% per year

Data Storage (2) – CASTOR



Total: ~ 87 PB
On tape: ~ 76 PB

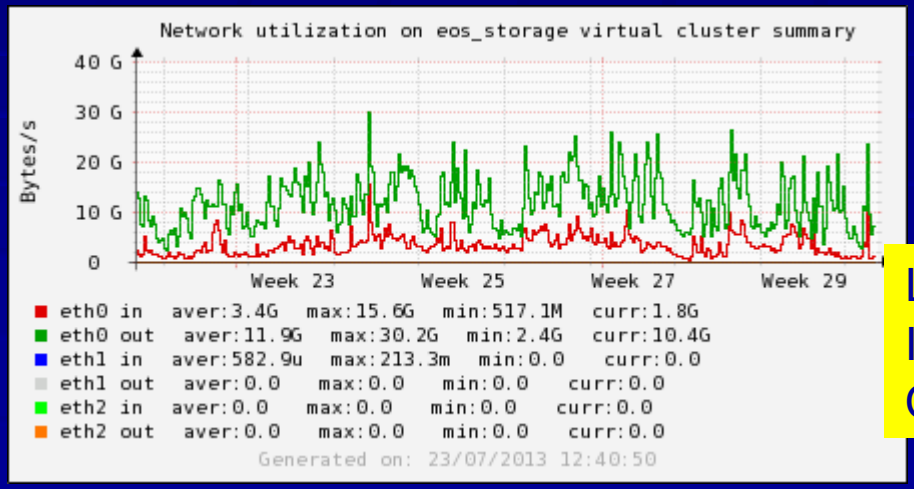
Last 2 months:
In: avg 1.7 GB/s, peak 4.2 GB/s
Out: avg 2.5 GB/s, peak 6.0 GB/s



314 million files; 75 PB of data on tape already today

Data Storage (3) - EOS

- Disk-only storage for analysis use case
  - Requirements very different from CASTOR
- 914 servers, 17 PB usable, 144 M files



Last 2 months:
In: avg 3.4 GB/s, peak 16.6 GB/s
Out: avg 11.9 GB/s, peak 30.6 GB/s

Other Storage for Physics

- Databases: metadata, conditions data, …
- AFS, DFS for user files
- CVMFS for experiment software releases, copies of conditions data, …
  - Not a file system, but an http-based distribution mechanism for read-only data
- …

Miscellaneous Services (in IT-PES)

- Numerous services for Grid computing
- TWiki: Collaborative Web space
  - More than 250 Twikis, between just a few and more than 8'000 Twiki items each (total 142'000)
- Version control services
  - SVN with SVNWEB/TRAC (2'041 active projects, 1'230 GB)
  - Git (257 active projects, 7.5 GB)
- Issue tracking service
  - Atlassian JIRA, Greenhopper, Fisheye, Crucible, Bamboo, …
  - 138 projects, 15'540 issues, 2'942 users
- BOINC: Framework for volunteer computing

World-wide Computing for LHC

- 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 Fabrizio Furano's lecture on July 30th

Outline

- CERN's computing facilities and hardware
- Service categories, tasks
- Infrastructure, networking, databases
- HEP analysis: techniques and data flows
- Network, plus, batch, storage
- **Between HW and services: Agile Infrastructure**
- References

Between Hardware and Services

- Until recently: dedicated hardware, OS and software set up according to service
  - CERN-proprietary tools (written 2001…2003): ELFms, Quattor, LEMON, …
- Challenges:
  - Very remote data centre extension
  - IT staff numbers remain fixed but more computing capacity is needed
  - Tools are high maintenance and becoming increasingly brittle
  - Inefficiencies exist but root cause cannot be easily identified and/or fixed

CERN-IT: Agile Infrastructure

- Reviewed areas of
  - Configuration management
  - Monitoring
  - Infrastructure layer
- Guiding principles
  - We are no longer a special case for computing
  - Adopt a tool chain model using existing open source tools
  - If we have special requirements, challenge them again and again
  - If useful, make generic and contribute back to the community

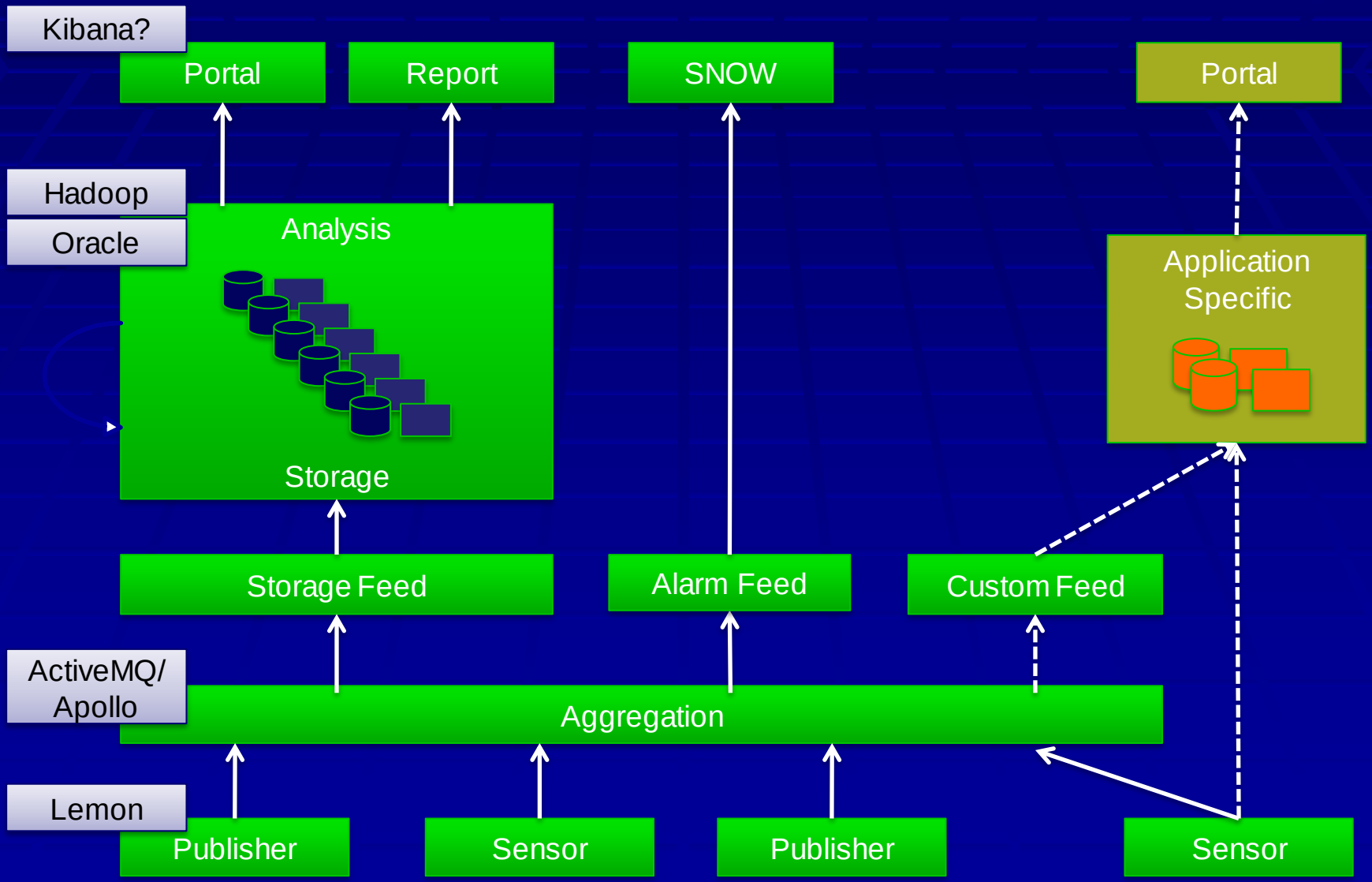
Configuration Management

- Puppet chosen as the core tool
- Puppet and Chef are the clear leaders for ‘core tools’
- Many large enterprises now use Puppet
  - Its declarative approach fits what we’re used to at CERN
  - Large installations: friendly, wide-based community
- The PuppetForge contains many pre-built recipes
  - And accepts improvements to improve portability and function
- Training and support available; expertise is valuable on job market
- Additional tools: Foreman for GUI/dashboard; GIT for version control; Mcollective for remote execution; Hiera for conditional configuration; PuppetDB as configuration data warehouse

Monitoring: Evolution

- Motivation
  - Several independent monitoring activities in IT
  - Based on different tool-chain but sharing same limitations
  - High level services are interdependent
  - Combination of data and complex analysis necessary
    - Quickly answering questions you hadn't though of when data recorded
- Challenge
  - Find a shared architecture and tool-chain components
  - Adopt existing tools and avoid home grown solutions
  - Aggregate monitoring data in a large data store
  - Correlate monitoring data and make it easy to access

Monitoring: Architecture



Move to the clouds

… or Infrastructure as a Service (IaaS)

- Rationale:
  - Improve operational efficiency
    - Machine reception/testing by rapidly deploying batch VMs
    - Hardware interventions with long running programs by live migration
    - Multiple operating system coverage by multi-VMs per hypervisor
  - Improve resource efficiency
    - Exploit idle resources such as service nodes by packing with batch
    - Variable load such as interactive/build machines with suspend/resume
  - Improve responsiveness
    - Self-Service web interfaces rather than tickets to request resources
    - Coffee break response time
  - Support middleware simplification and cloud studies
    - Could we run LHC infrastructure with only an Amazon like interface?
- Previous experience with lxcloud (OpenNebula) and CERN Virtualisation Infrastructure (Microsoft SCVMM)

IaaS: Openstack

- Cloud operating system/orchestrator
- Controls large pools of compute, storage, and networking resources
- Dashboard gives administrators control
- Users to provision resources through a web interface
- Components used: compute, dashboard, image store, object storage, block storage, identity management, network, load balancing/high-availability
- Fully integrated with CERN's Active Directory via LDAP
- Potential long-term impact on Grid middleware

Conclusions

- The Large Hadron Collider (LHC) and its experiments are very data (and compute) intensive projects
- 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
- Computing has worked very well during Run I (at 2 * 4 TeV), instrumental for discovering a Higgs particle
- We are on track for further ramp-ups of the computing capacity for future requirements
  - Additional, remote data centre
  - AI project covering configuration and installation; IaaS; monitoring


Thank you

Outline

- CERN's computing facilities and hardware
- Service categories, tasks
- Infrastructure, networking, databases
- HEP analysis: techniques and data flows
- Network, plus, batch, storage
- Between HW and services: Agile Infrastructure
- **References**

More Information (1)

IT department

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

Monitoring (currently in production)

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/

EOS

http://eos.web.cern.ch/eos/

More Information (2)

Windows, Web, Mail

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

Grid and WLCG

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

http://www.egi.eu/, http://www.eu-emi.eu, http://www.eu-egee.org/

Computing and Physics

http://www.chep2012.org/

Data centre upgrades

http://cern.ch/go/NN98

Agile Infrastructure project

http://cern.ch/go/N8wp

http://cern.ch/go/99Ck

https://indico.cern.ch/conferenceDisplay.py?confId=184791

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

Helge.Meinhard (at) cern.ch