CERN IT Department

CERN Lustre Evaluation

Arne Wiebalck

Sun HPC Workshop, Open Storage Track
Regensburg, Germany
8th Sep 2009

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Agenda

A Quick Guide to CERN

Storage Use Cases

Methodology & Initial Findings

Current Thoughts & Wish List

Conclusion

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Agenda

A Quick Guide to CERN

Storage Use Cases

Methodology & Initial Findings

Wish List

Conclusion

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

A Quick Quide to CERN

CERN: European Organization for Nuclear Research

Located at the Swiss/French border near Geneva

Twenty Member States:

Austria	Belgium	Bulgaria	Czech Rep.
Denmark	Finland	France	Germany
Greece	Hungary	Italy	Netherlands
Norway	Poland	Portugal	Slovak Rep.
Spain	Sweden	Switzerland	UK

Plus eight Observer States:

European Commission, India, Israel, Japan, Russian Federation, Turkey, UNESCO and USA

Budget (2008): 1154 MCHF (~715 MEUR)

Personnel: 2600 Staff, 700 Fellows/Associates

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Fundamental Laws of Nature

Why do particles have mass?

Newton could not explain it - and neither can we…

What is 96% of the Universe made of?

We only know 4% of it!

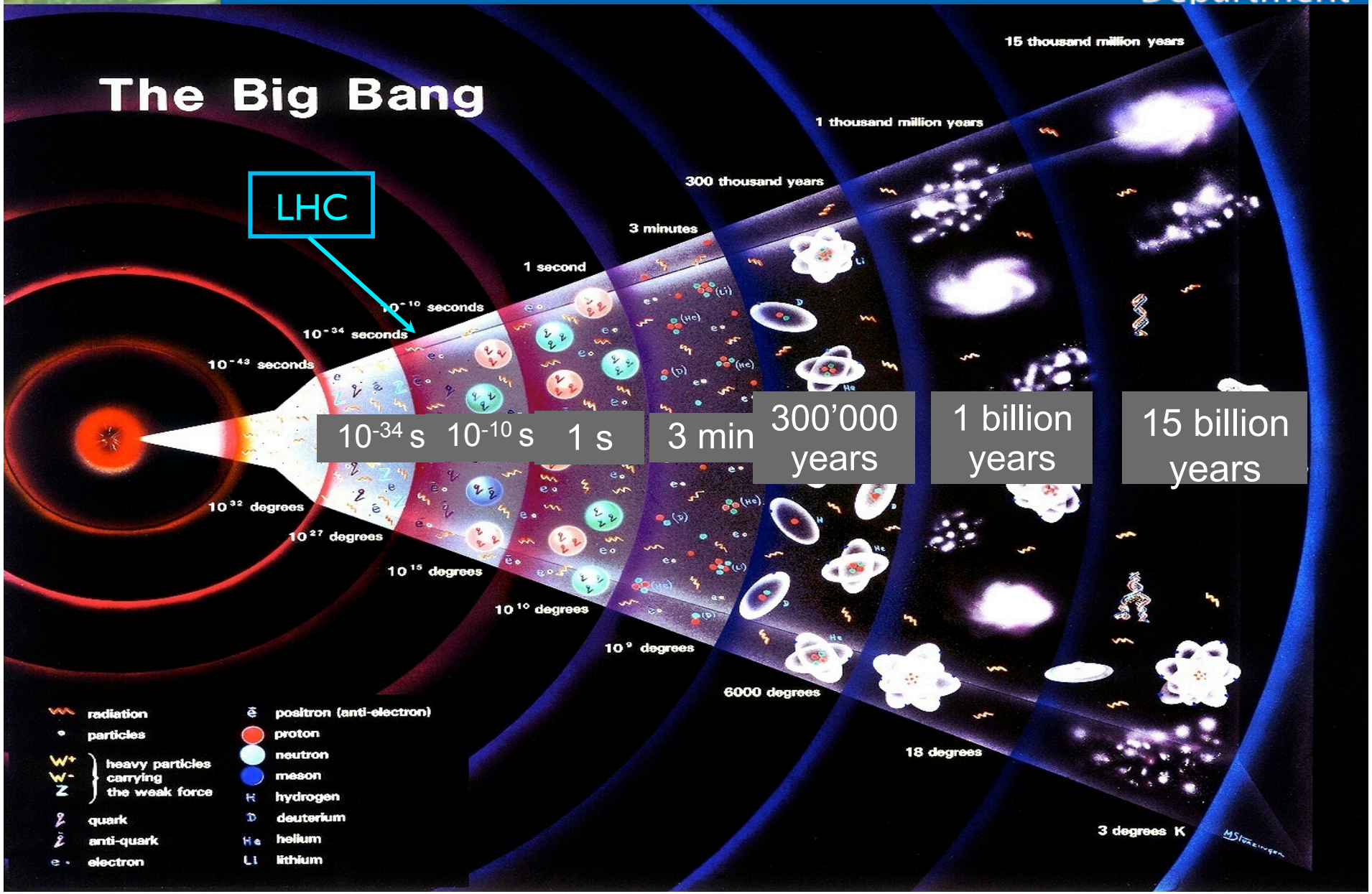
Why is there no antimatter left in the Universe?

Nature should be symmetrical

What was matter like during the first second of the universe's life, right after the "Big Bang"?

A journey towards the beginning of time

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it


The Big Bang
LHC
15 thousand million years
1 thousand million years
300 thousand years
3 minutes
1 second
10^{-10} seconds
10^{-34} seconds
10^{-43} seconds
10^{-34} s
10^{-10} s
1 s
3 min
300'000 years
1 billion years
15 billion years
10^{32} degrees
10^{27} degrees
10^{15} degrees
10^{10} degrees
10^{9} degrees
6000 degrees
18 degrees
3 degrees K
radiation
particles
heavy particles carrying the weak force
quark
anti-quark
electron
positron (anti-electron)
proton
neutron
meson
hydrogen
deuterium
helium
lithium

CERN's Tools

The world's most powerful accelerator: LHC

- A 27 km long tunnel filled with high-tech instruments
- Equipped with thousands of superconducting magnets
- Accelerates particles to energies never before obtained
- Produces particle collisions creating microscopic "big bangs"

Very large sophisticated detectors

- Four experiments each the size of a cathedral
- Hundred million measurement channels each
- Data acquisition systems treating Petabytes per second

Top level computing to distribute and analyse the data

- A Computing Grid linking ~200 computer centres around the globe
- Sufficient computing power and storage to handle 15 Petabytes per year, making them available to thousands of physicists for analysis

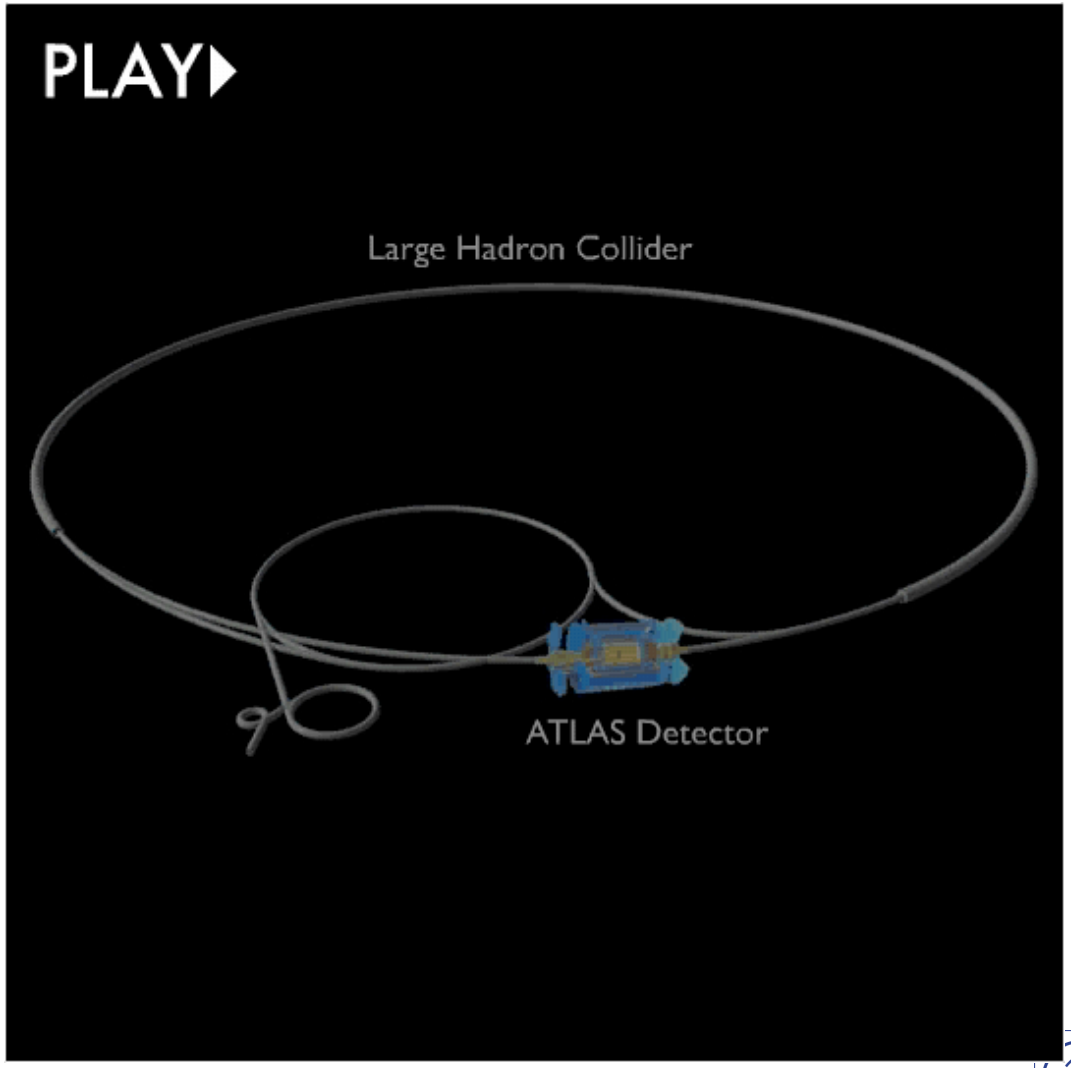
CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Proton Acceleration & Collision

FIO

CERN IT Department

- Protons are accelerated by several machines up to their final energy (7+7 TeV)
- Such collisions take place 40 million times per second, day and night, for about 100 days per year
- Head-on collisions are produced right in the centre of a detector, which records the new particle being produced



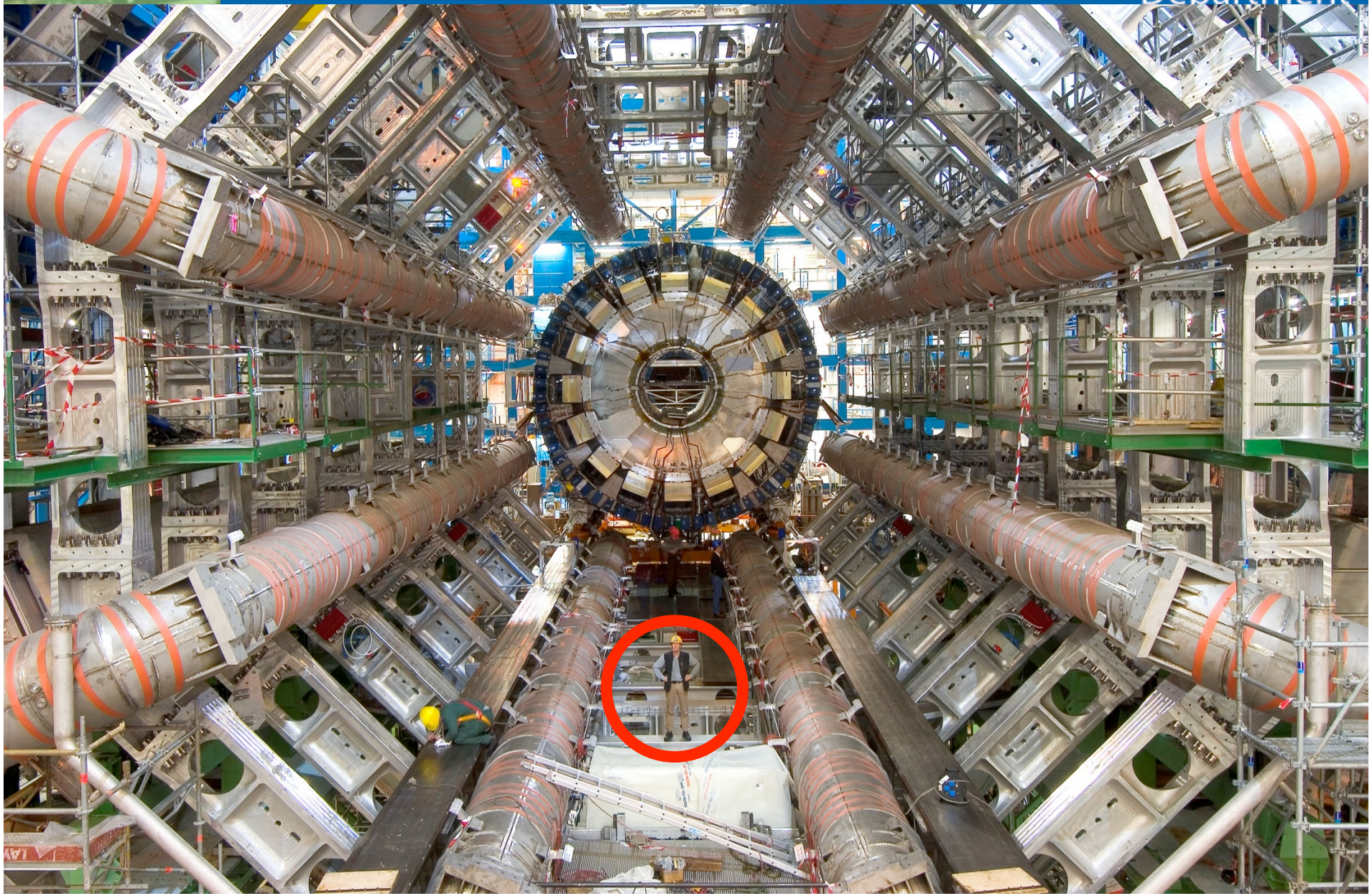
CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

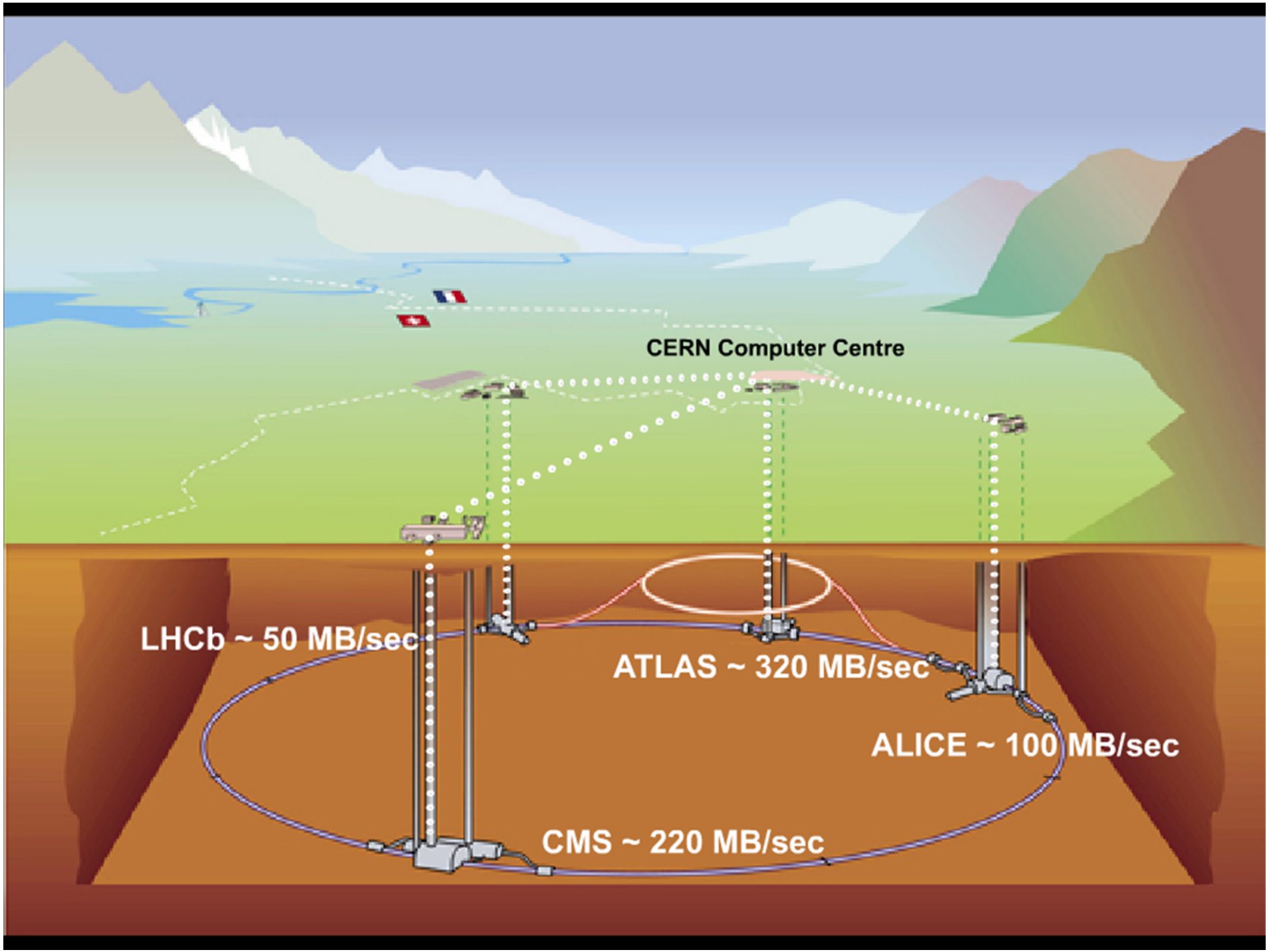
The Large Hadron Collider (LHC)

FIO

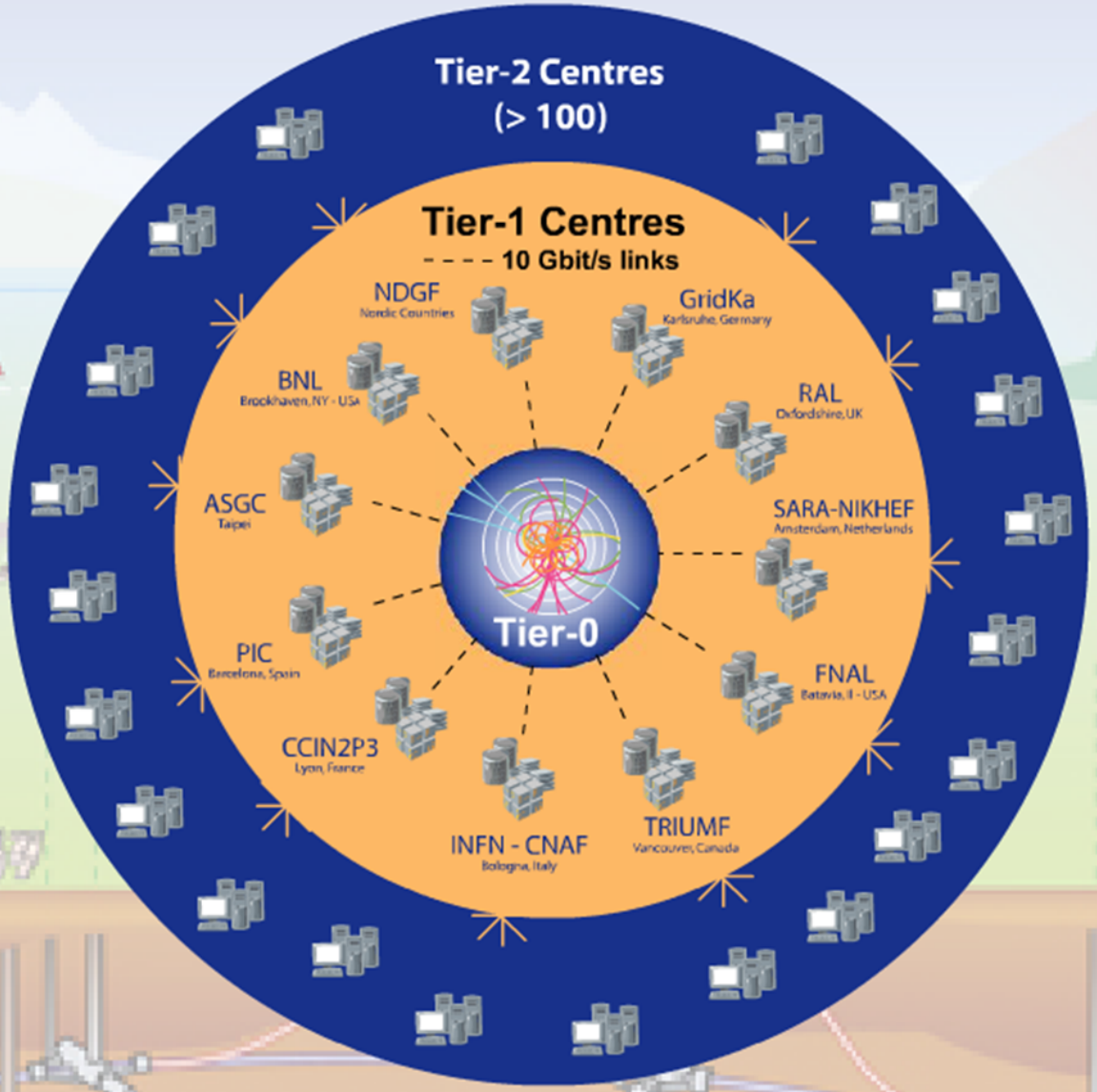
The ATLAS Experiment

CERN IT Department




CERN Computer Centre
LHCb ~ 50 MB/sec
ATLAS ~ 320 MB/sec
ALICE ~ 100 MB/sec
CMS ~ 220 MB/sec

The LHC Computing Grid



Agenda

A Quick Guide to CERN

Storage Use Cases

Initial Findings

Wish List

Conclusion

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Storage Use Cases

- **HSM System**
  - CERN Advanced STORage Manager (CASTOR)
  - 18.3 PB, 120 million files, 1'200 servers
- **Analysis Space**
  - Analysis of the experiments' data
  - XRootD plus CASTOR
- **Project Space**
  - >150 projects
  - Experiments' code (build infrastructure)
  - CVS/SVN, Indico, Twiki, …
- **User home directories**
  - 20'000 users on AFS
  - 50'000 volumes, 25TB, 1.5 billion acc/day, 50 servers
  - 400 million files

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Why a Lustre Evaluation?

FIO

CERN IT Department

Lustre:

- Performance
- Scalability
- HEPiX FSWG

CERN:

- Home directories & projects on AFS
- CASTOR as HSM
- XRootD plus CASTOR for analysis

Lustre: An opportunity for consolidation?

Does Lustre meet the requirements?

Can Lustre be operated?

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Agenda

A Quick Guide to CERN

Storage Use Cases

Methodology & Initial Findings

Wish List

Conclusion

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Project Plan

- ✓ **Assemble a list of points to look at**
  - ✓ Understand requirements of the use cases
  - ✓ Manageability
- ✓ **Gather information**
  - ✓ Lustre training
  - ✓ Attend LUG
  - ➢ Exchange experiences with other sites
- ➢ **Get hands-on experience**
  - ➢ Set up test instances, preprod instance
  - ➢ Familiarize with certain functionality
  - ➢ Integrate with CERN‘s tools for fabric management
- • **Document the findings**

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Use Case Requirements

FIO

CERN IT Department

HSM	Analysis Space	Home Directories/ Analysis Space
• Generic interface, CASTOR/TSM • Scalable • Support for random access	• Low latency access (disk-based) • O(1000) open/sec • Several GB/s aggregate bandwidth • ~10% Backup • Mountable plus XrootD access • ACLs & quota	• Strong authentication • Backup • Wide-area access • Small files • Availability • ACLs & Quota

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Areas of Interest

- **Strong Authentication**
- **Backup**
- **Fault-tolerance**
- **Special performance & Optimization**
- **HSM interface**
- **Life Cycle Management (LCM) & Tools**

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Strong Authentication

- v1.8: no Kerberos
- v2.0: early Adaptation, v2.x full Kerberos (late 2010 / early 2011)
- „robust“ and „usable“
- Implementation not (yet) complete
- PSC early adopter, WAN setup

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Backup

- **Meta data (MDT)**
  - ✓ LVM snapshots plus tar of EA
  - ✓ Orphanfiles, Ghostfiles
- **User data (OST)**
  - × Crawling the name space
  - ✓ With v1.8.x: e2scan
  - ? With v2.0.x: changelogs
- **Used in pre-prod instance (to TSM)**

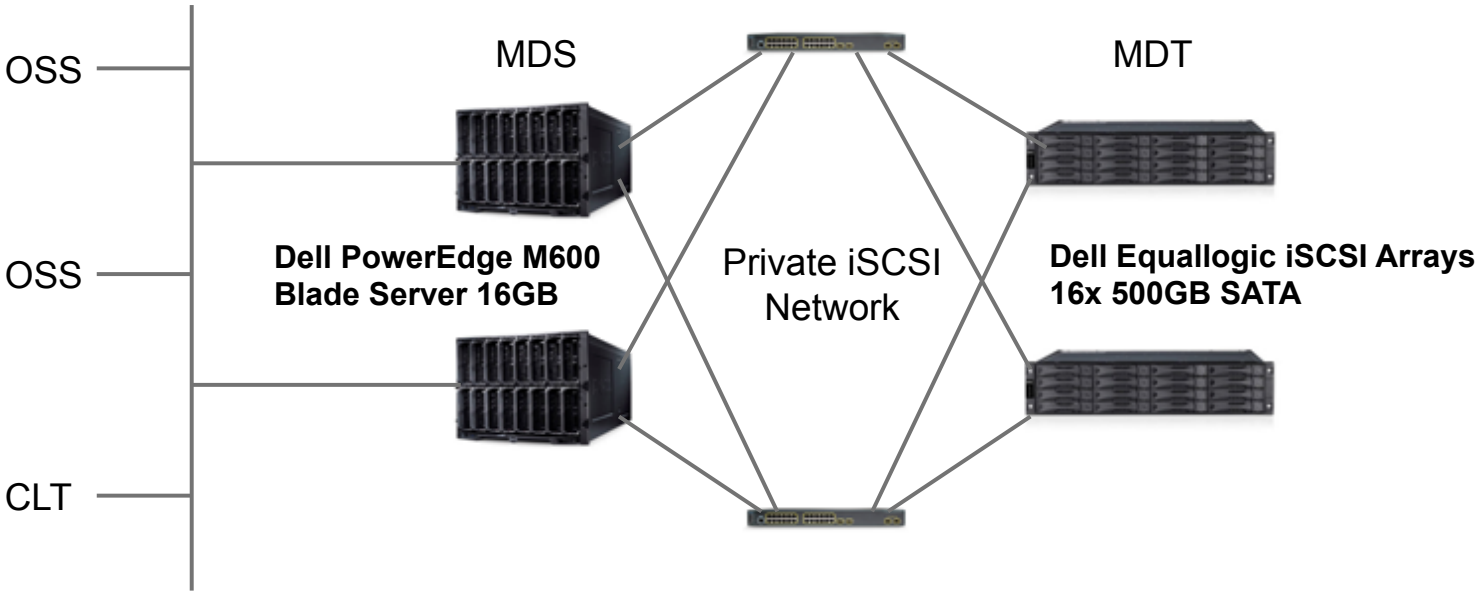
CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Fault-tolerance

- **Lustre comes with support for failover**
  - active/active for OSSs
  - active/passive for MDS
  - Used at other sites
- **Technology**
  - Shared storage: FC, DRBD, iSCSI
  - Heartbeat

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

MDS Test Setup



- **Fully redundant against component failure**
  - iSCSI for shared storage
  - Linux device mapper + md for mirroring
- **Performance?**
- **OSS: disk server (DAS/iSCSI targets), iSCSI arrays**

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Special performance

- **Small files**
  - No client-side caching, 1MB transfer unit, double lookup
  - „How bad is it?“
  - Recent improvements (OSS read caches)?
- **Understand tuning options**

HSM

- **CEA (J.-C. Lafoucrière, A. Degrémont)**
  - interface design
  - Implementations
    - HPSS, Sun SAM-QFS, Enstore (FNAL)?
  - HSM Mailing list quite active
  - Beta version available later this year

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Life Cycle Management

- **Adding capacity (add an OSS)**
  - Possible, but quiesce clients (?!)
  - Allocation policy, coordinated filling
- **Removing capacity (remove an OSS)**
  - **no user-transparent data migration**
- **MDT (re-)sizing**
- **Managing Quota**
- **Develop procedures / tools**
  - Lustre upgrades
  - System upgrades (kernel)

Agenda

A Quick Guide to CERN

Storage Use Cases

Methodology & Initial Findings

Wish List

Conclusion

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Wish List

- **Complete the support for strong authentication**
- **More control over the system**
  - Client-server coupling (recovery)
  - Too powerful users (striping, pools)
- **Stronger Support for Life Cycle Management**
  - User transparent data migration
- **File replication**
  - Availability, different striping policies
  - Easier maintenance (high level vs. redundant storage)
  - Usable for migration?
- **Outsourcing of privileges**

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Current Thoughts

- **New versions bring some helpful features**
  - v1.8: VBR, adaptive timeouts, OSS readcache
  - v2.0: Kerberos, change logs, Clustered MDS
- **Increased emphasis on Quality Assurance**
- **Milestones-based pre-releases**
  - the „moving targets“ problem
- **Responsiveness of Lustre team**
- **Interest in „non-performance“ features by other sites**

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Summary

- **CERN is looking into Lustre as a candidate for Storage Consolidation**
- **Investigation on manageability is missing**
- **Input is welcome, share your experiences!**

arne.wiebalck@cern.ch

Some links

- CERN: http://www.cern.ch
- CASTOR: http://www.cern.ch/castor
- XrootD: http://xrootd.slac.stanford.edu/
- AFS: http://www.openafs.org

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it