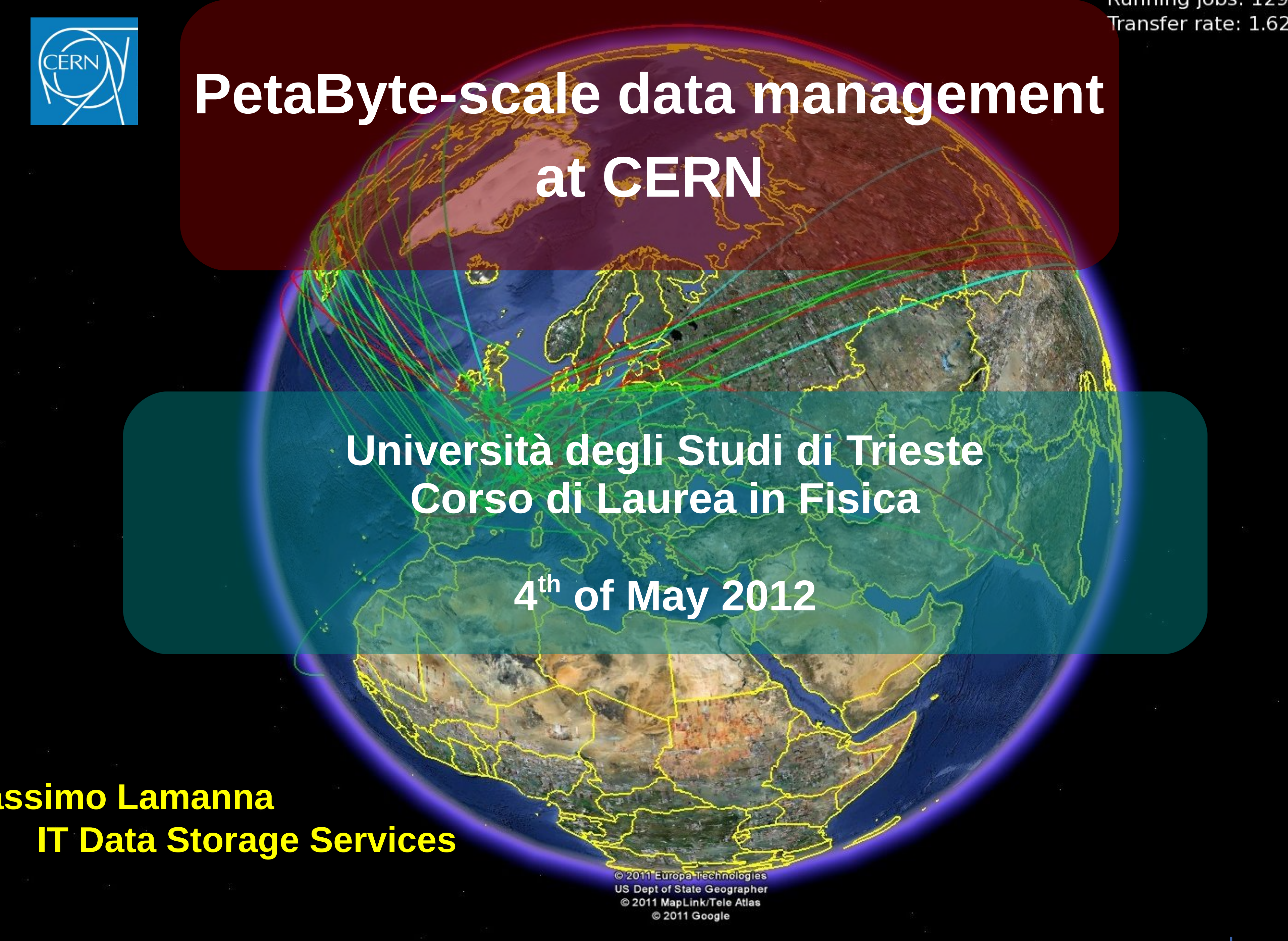
PetaByte-scale data management at CERN

Università degli Studi di Trieste
Corso di Laurea in Fisica

4th of May 2012

assimo Lamanna
IT Data Storage Services



Service Level Status overview

Computer Centre By Numbers

19 Feb 2012 Sun 11:06:32

Service information

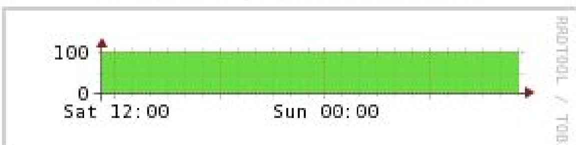
full name:	**Computer Centre By Numbers**
short name:	CCBYNUM
group:	IT-CF-FPP
site:	CERN
email:	imre.szebenyi@cern.ch
manager:	**Imre Szebenyi**

Service availability (more)

availability:	
percentage:	100%
status:	available
last update:	11:01:02, 19 Feb 2012 (6 minutes ago)
expires after:	1440 minutes

rss feed with status changes

availability in the last 24 hours (more):

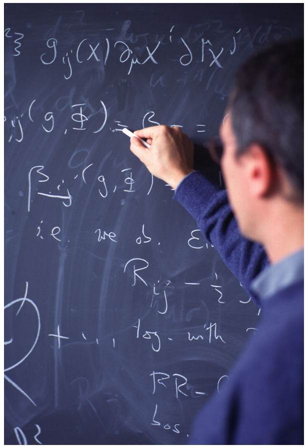


Additional service information (more)

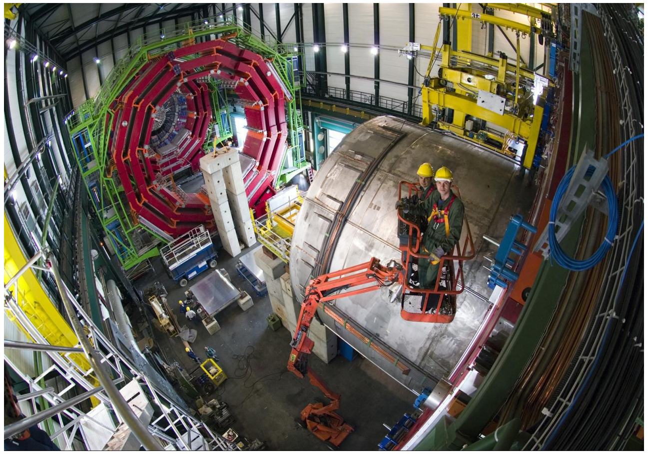
Number of processors:	14,972
Number of cores:	64,623
Memory capacity (TiB):	165
Memory modules:	55,729
Raw HDD capacity (TiB):	62,660
Number of HDD's:	62,023
Number of systems:	7,975
Number of RAID controllers:	3,607
Number of enclosures:	1,554
SPEC CPU2006:	503,637
Number of racks:	1,070
Number of virtual machines:	3,908
Number of Fibre channel ports:	742
Number of 1G ports:	16,773
Number of 10G ports:	622
Current power consumption (kW):	2,186
Current power consumption (kVA):	2,305

- 24x7 operator and system admin support
- Management and Automation framework for large scale Linux clusters
- Hardware installation & retirement
- ~7,000 hardware movements/year; ~1000 disk failures/year

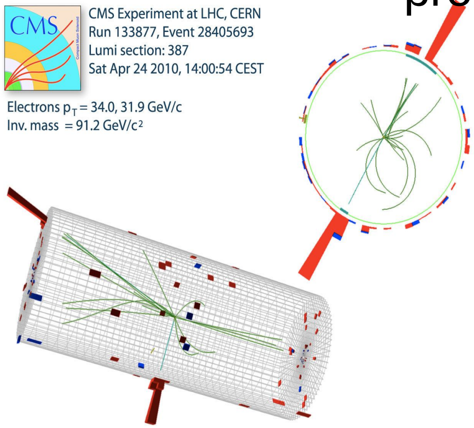
What are all these computers/disks/networks… for?



(Detector) simulation

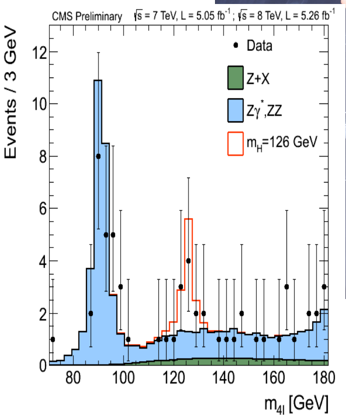


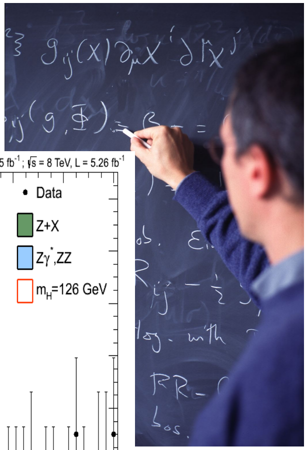
Data acquisition and processing



Data distribution

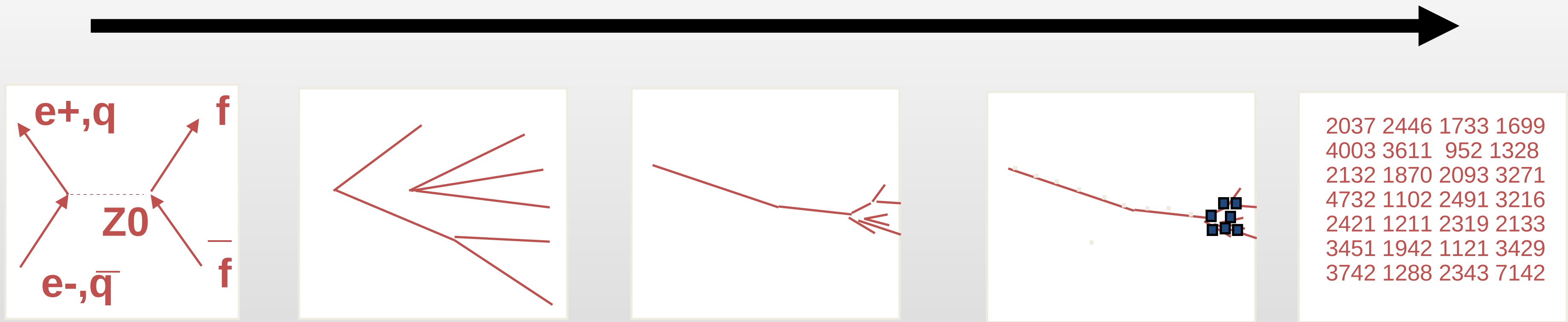
Data analysis





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

From Physics to Raw Data



Basic physics

Fragmentation, Decay

Interaction with detector material
Multiple scattering, interactions

Detector response
Noise, pile-up, cross-talk, inefficiency, ambiguity, resolution, response function, alignment, temperature

Raw data (Bytes)

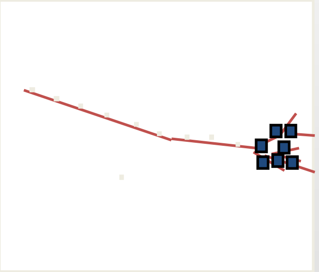
Read-out addresses, ADC, TDC values, Bit patterns

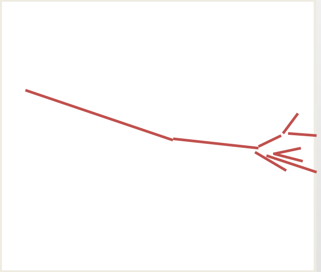
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

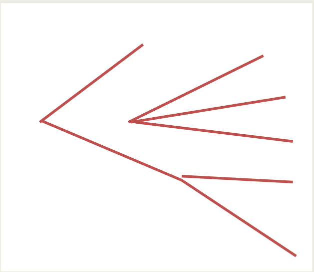
From Raw Data to Physics

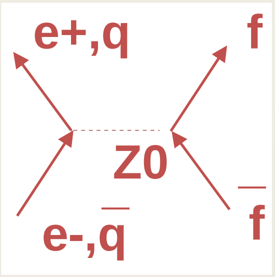
CERN IT Department

2037	2446	1733	1699
4003	3611	952	1328
2132	1870	2093	3271
4732	1102	2491	3216
2421	1211	2319	2133
3451	1942	1121	3429
3742	1288	2343	7142









Raw data

Convert to physics quantities

Detector response apply calibration, alignment,

Interaction with detector material Pattern, recognition, Particle identification

Fragmentation, Decay Physics analysis

Basic physics

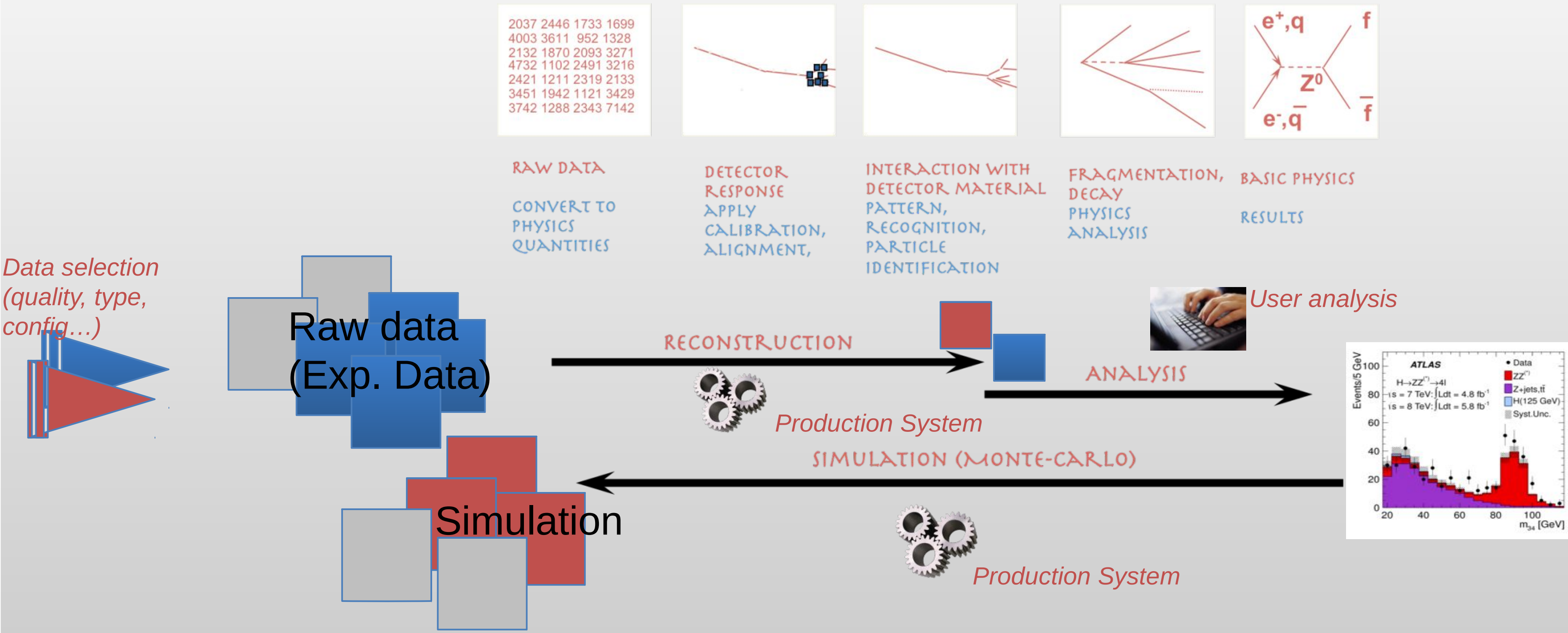
Results

Reconstruction

Analysis

Simulation (Monte-Carlo)

Analysis flow (user view)

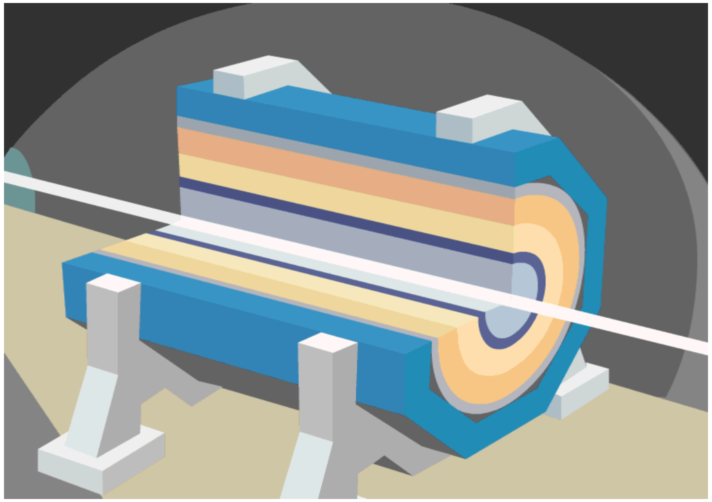


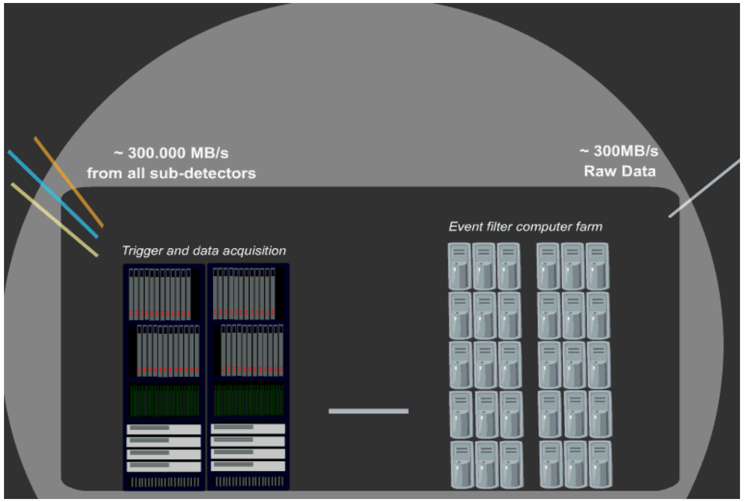
But how this is done *in practice*? Of course we need CPUs, disks, networks etc..
The problem is orchestrating hardware resources, software and humans :)

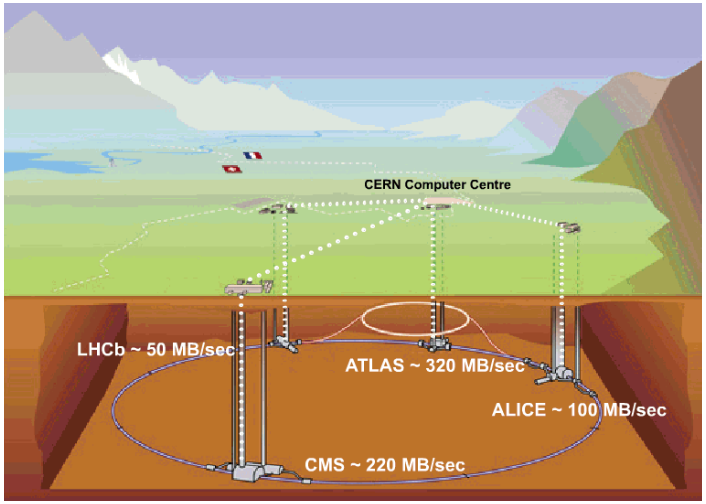
“All” data are stored in files (aggregated as “datasets” = collections of files). Only a small fraction of data in real DBs (e.g calibrations). This is one characteristics of HEP computing.

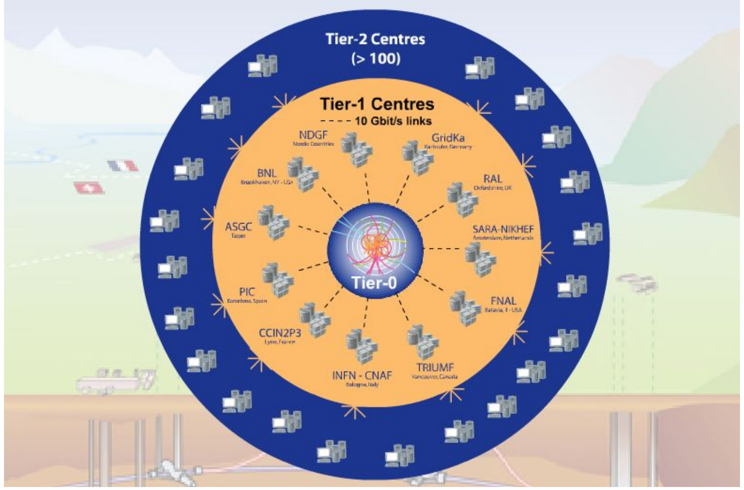
The role of the CERN Computer Centre

CERN IT Department

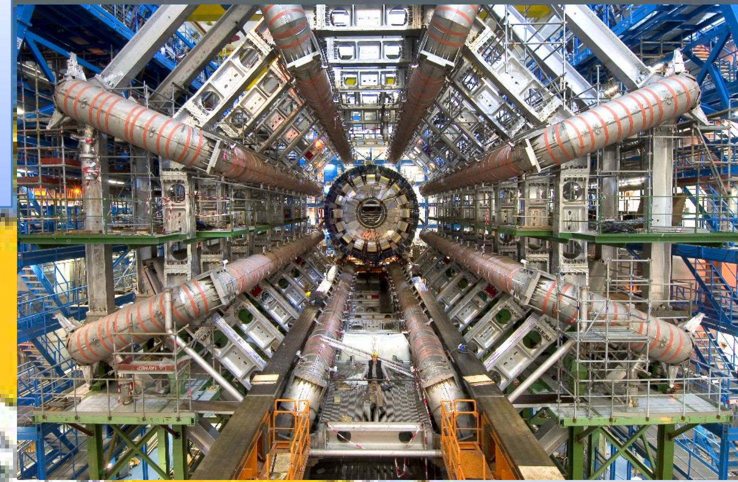






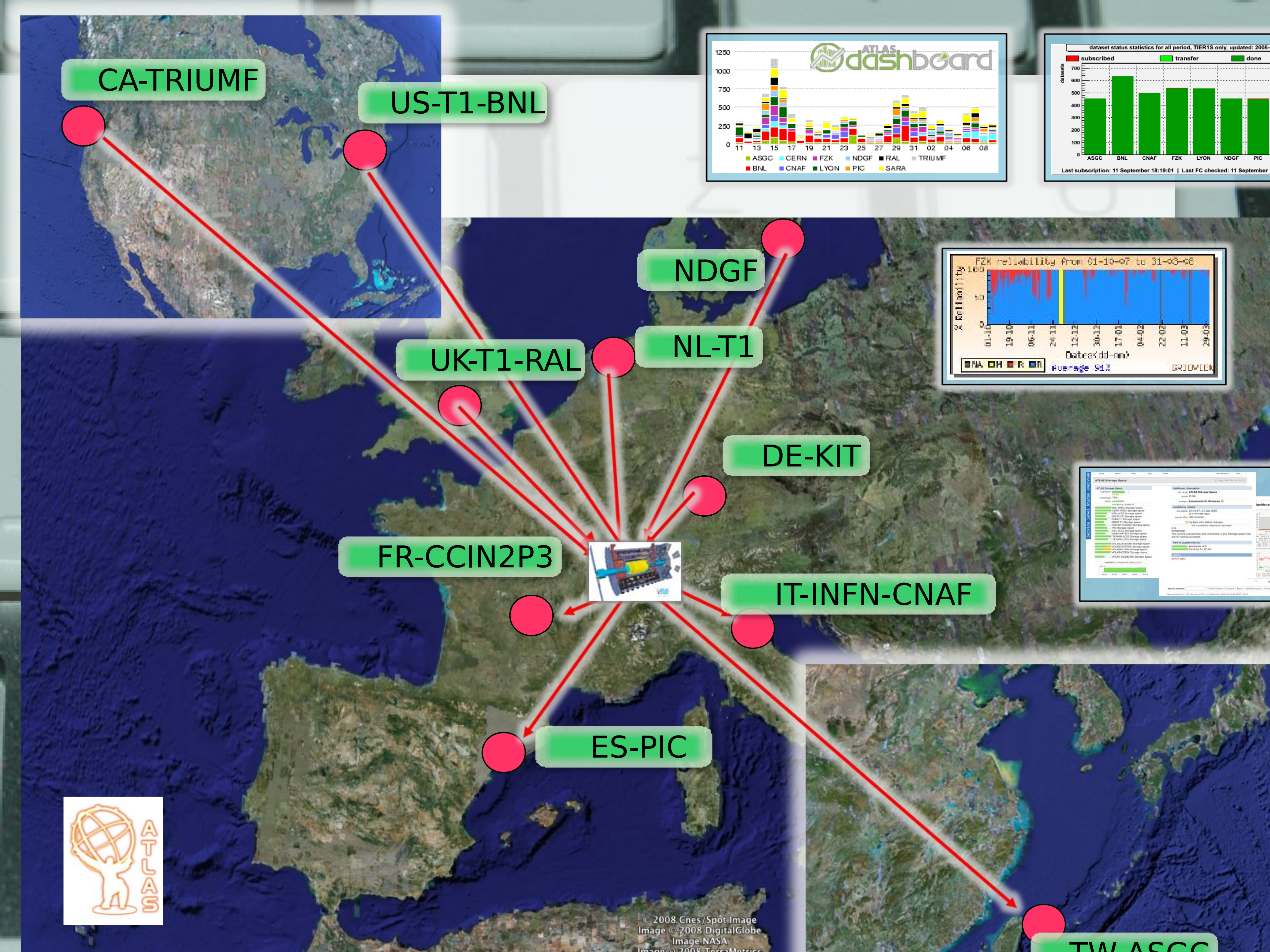


ATLAS: 7,000 tons, 150 million sensors generating data 40 millions times per second i.e. a petabyte/s (1 million GB/s)



ATLAS EXPERIMENT
http://atlas.ch

ATLAS is around than 3,000 collaborators
From 169 universities from 37 countries
~1000 students!!!


CA-TRIUMF
US-T1-BNL
NDGF
NL-T1
UK-T1-RAL
DE-KIT
FR-CCIN2P3
IT-INFN-CNAF
ES-PIC
TW-ASGC
ATLAS dashboard
dataset status statistics for all period, TIER1S only, updated: 2008-
subscribed
transfer
done
ASGC
BNL
CNAF
FZK
LYON
NDGF
PIC
Last subscription: 11 September 18:19:01 | Last FC checked: 11 September
ASGC
CERN
FZK
NDGF
RAL
TRIUMF
BNL
CNAF
LYON
PIC
SARA
ATLAS
© 2008 Cnes/Spot Image
Image © 2008 DigitalGlobe
Image NASA

One of CERN main responsibilities

(LHC scientific programme – LHC data management)

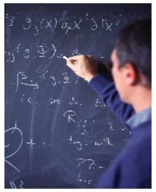
CERN IT Department

Tier 0 function

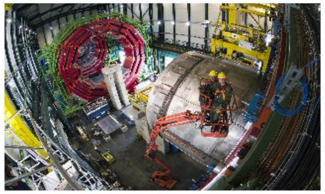
- Experimental data to tape (RAW)
- Data distribution (Tier0 → Tier1s)
- Data reconstruction and redistribution
  - RAW → ESD/DST,AOD..
- Archival data (e.g. simulation data generated in other centres)

Analysis centre

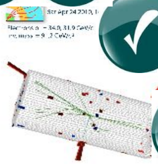
- Each LHC physicist has access to CERN!
- CERN-based analysis groups



(Detector) simulation

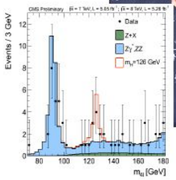


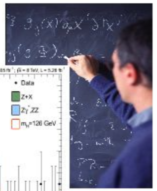
Data acquisition and processing



Data distribution

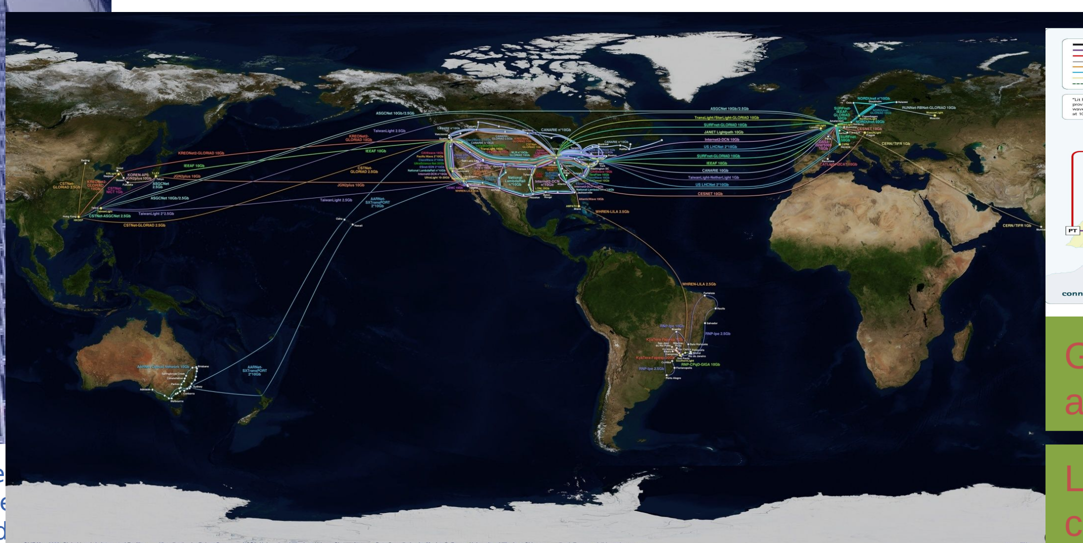
Data analysis

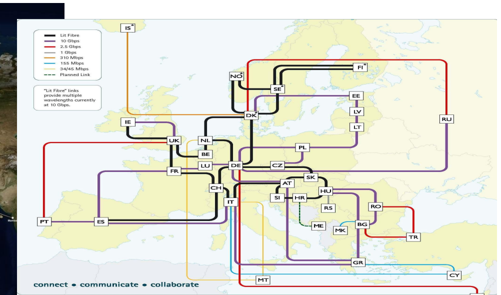




CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN Computer Centre: Storage, Distribution and Processing (Reconstruction and Analysis)





Géant: the pan-european Research and Education Network

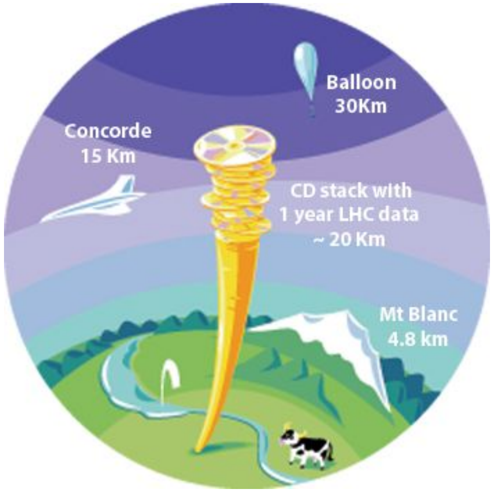
LHCOPN: dedicated links with major computer centres worldwide

CERN IT De
CH-1211 Ge
Switzerland
www.cern.ch/it

The HEP Data Challenge

CERN IT Department

- LHC will run for 20 years
- Experiments ***are*** producing about **15 Million Gigabytes** of data each year (about 20 million CDs!)
- LHC data analysis requires a computing power equivalent to **~100,000 of today's fastest PC processors**
- Requires many cooperating computer centres, as CERN can *only* provide ~20% of the capacity



A challenge for physics…
… and a challenge for technology research and industry as well

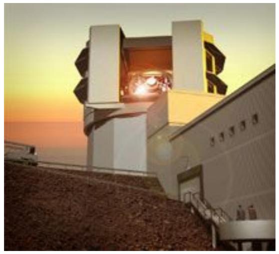
Big data non HEP specific (… any longer)

HEP (LCG):

- CASTOR
- EOS
- Other LCG systems
  - dCache (most popular among Tier1s) - DESY
  - DPM (most popular among Tier2s) - CERN
  - STORM – INFN
  - Xroot - SLAC
  - BestMan - Berkeley

Non HEP:

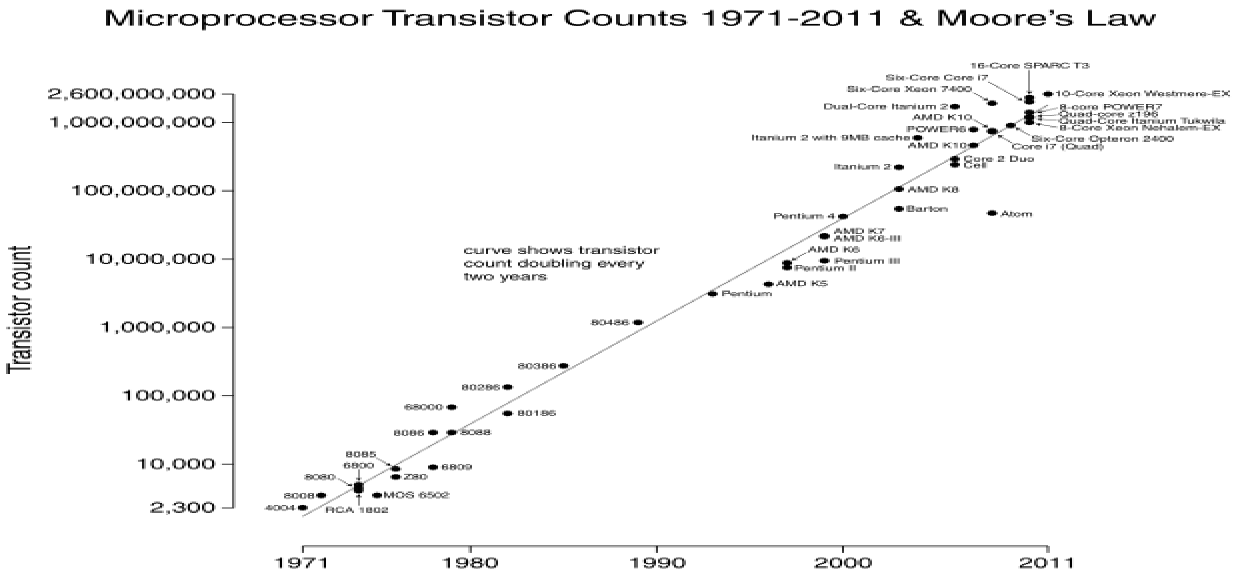
- **Google, Facebook, etc…**
  - e.g. HADOOP

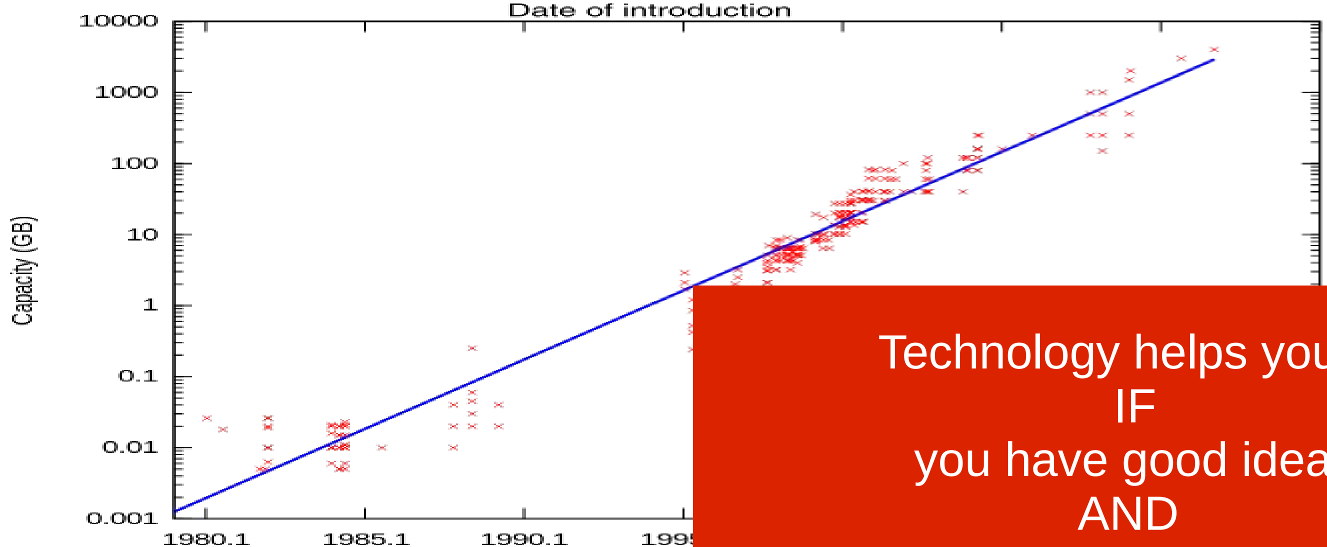


- **Astronomy**:
  - LSST: xroot for DB access (large-scale distributed queries)
- **Everyone:**
  - Cloud Computing
    - OpenStack swift
  - More file system
    - AFS, NFS 4.1, CEPH

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

Moore's law



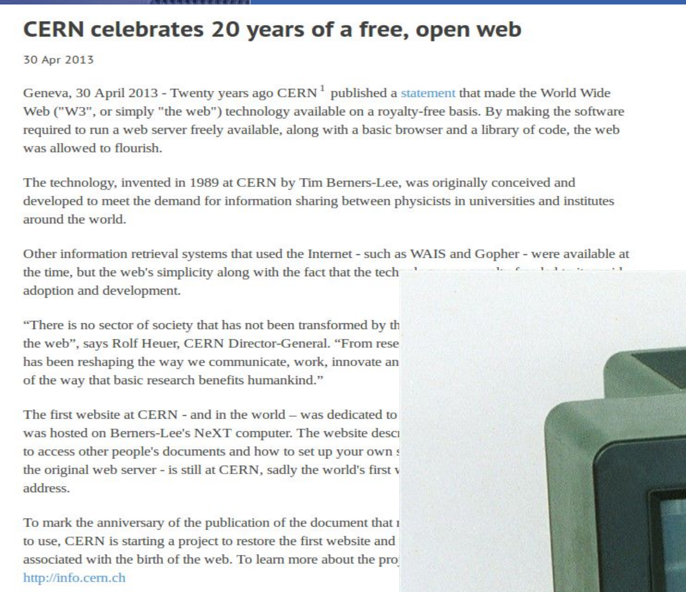


Technology helps you...
IF
you have good ideas
AND
you can master the global complexity

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

“Historical” example…

1990s:
The web was invented at CERN!
The machine used by Tim Berners-Lee in 1990 to develop and run the first WWW server, multi-media browser and web editor.


CERN celebrates 20 years of a free, open web

30 Apr 2013

Geneva, 30 April 2013 - Twenty years ago CERN [1] published a statement that made the World Wide Web ("W3", or simply "the web") technology available on a royalty-free basis. By making the software required to run a web server freely available, along with a basic browser and a library of code, the web was allowed to flourish.

The technology, invented in 1989 at CERN by Tim Berners-Lee, was originally conceived and developed to meet the demand for information sharing between physicists in universities and institutes around the world.

Other information retrieval systems that used the Internet - such as WAIS and Gopher - were available at the time, but the web's simplicity along with the fact that the tech
adoption and development.

“There is no sector of society that has not been transformed by th
the web”, says Rolf Heuer, CERN Director-General. “From rese
has been reshaping the way we communicate, work, innovate an
of the way that basic research benefits humankind.”

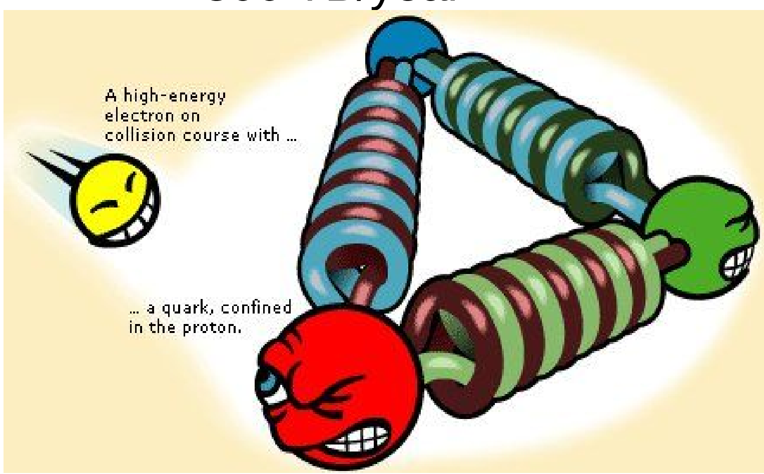
The first website at CERN - and in the world – was dedicated to
was hosted on Berners-Lee's NeXT computer. The website desc
to access other people's documents and how to set up your own
the original web server - is still at CERN, sadly the world's first w
address.

To mark the anniversary of the publication of the document that
to use, CERN is starting a project to restore the first website and
associated with the birth of the web. To learn more about the pro
http://info.cern.ch

COMPASS proposal (1996)

CERN IT Department

- **(Data) challenge**
  - Semi inclusive → reconstruct “the rest” of the deep inelastic event
    - 35 MB/s
    - 300 TB/year



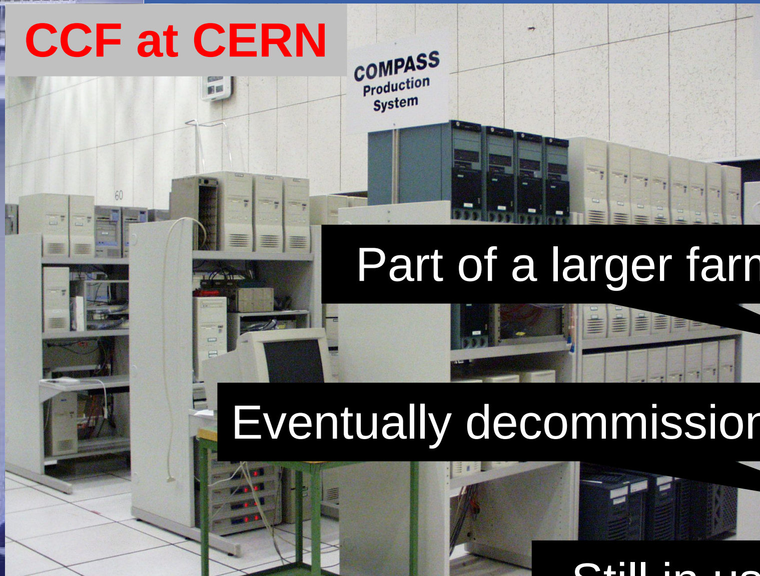
More than a prototype!!!

- 3 times less data 14 years before LHC (LHC experiment) (7 Moore's cycles hence ~2^7)
- 40x factor "against" us ($2^7/3$)

- **Use a Linux PC farm**
  - Instead of the “usual super computers”
- **C++**
  - Instead of good ol' Fortran IV
- All data in a **data base** (Objectivity/DB)
  - Instead of “plain” files
  - Object store (noSQL)
- First experiment using **CASTOR**
  - Instead of writing tapes “by ourself”

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

CCF at CERN



Part of a larger farm

Eventually decommissioned

Still in use

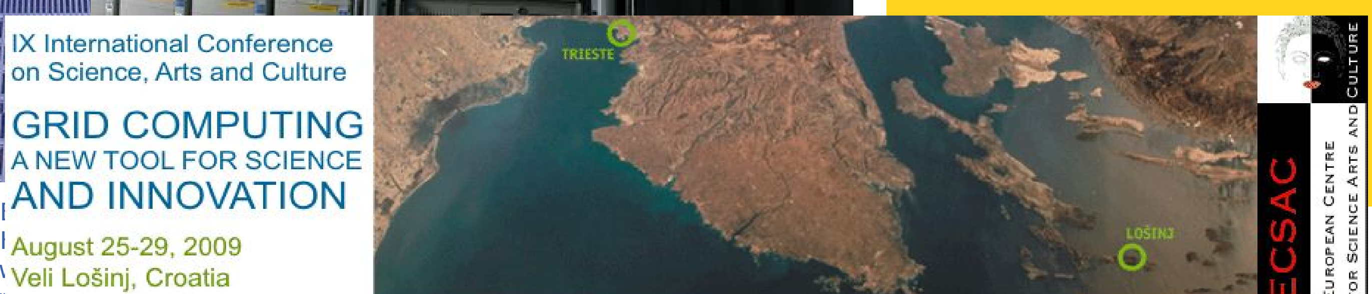
ACID in Trieste (INFN)

Still in use

Close collaboration with CERN IT

COMPASS was the first/among the first to:

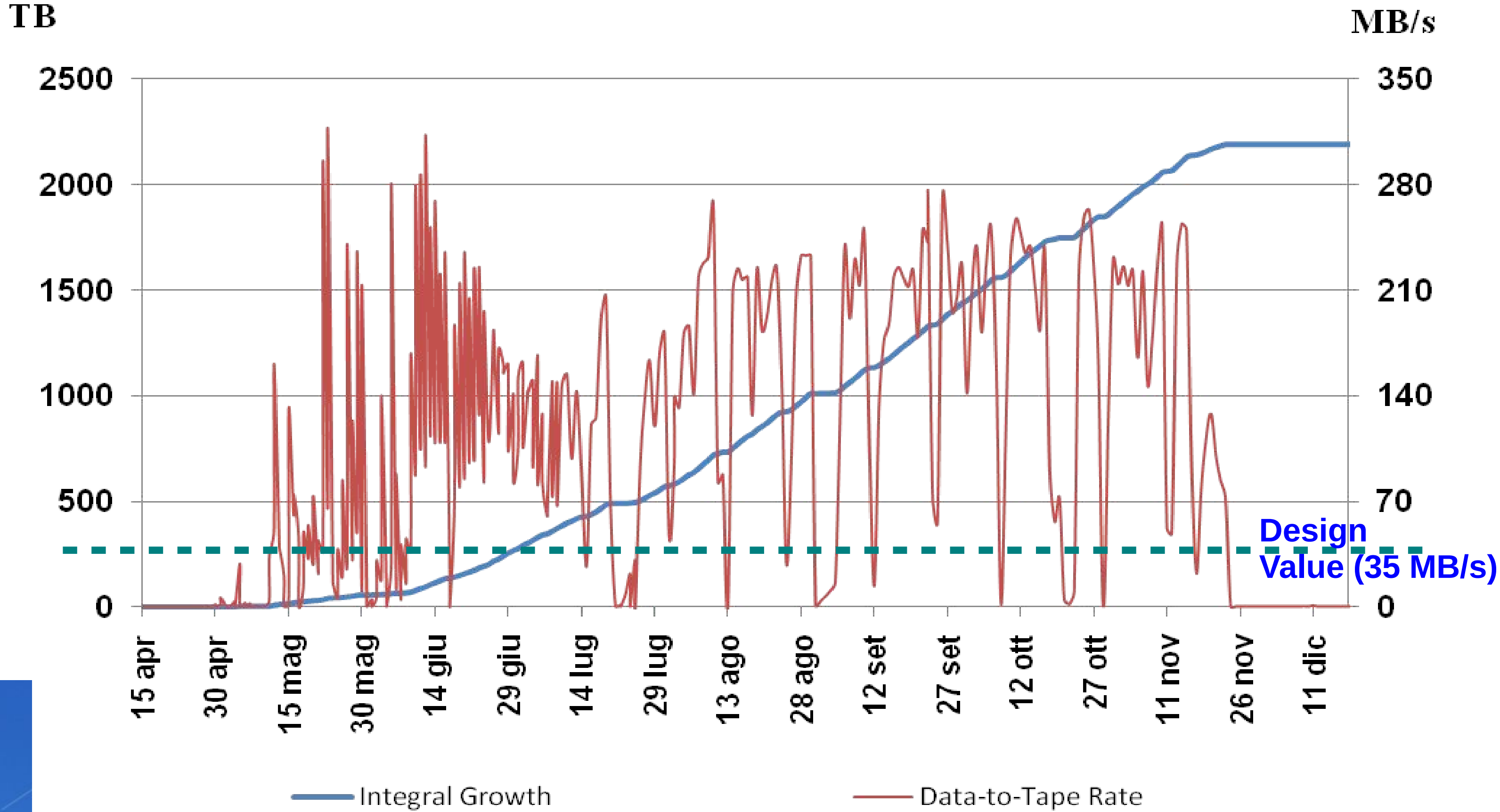
- Migration to Linux PC farms (conclusions of CHEP2000)
  - CCF at CERN
  - ACID in Trieste
- Use of Objectivity/DB at the 10^2 TB scale
- First production user of CASTOR
- Deliver and operate CORAL (C++ framework) for data processing
  - And the algorithmic parts, notably the RICH reconstruction

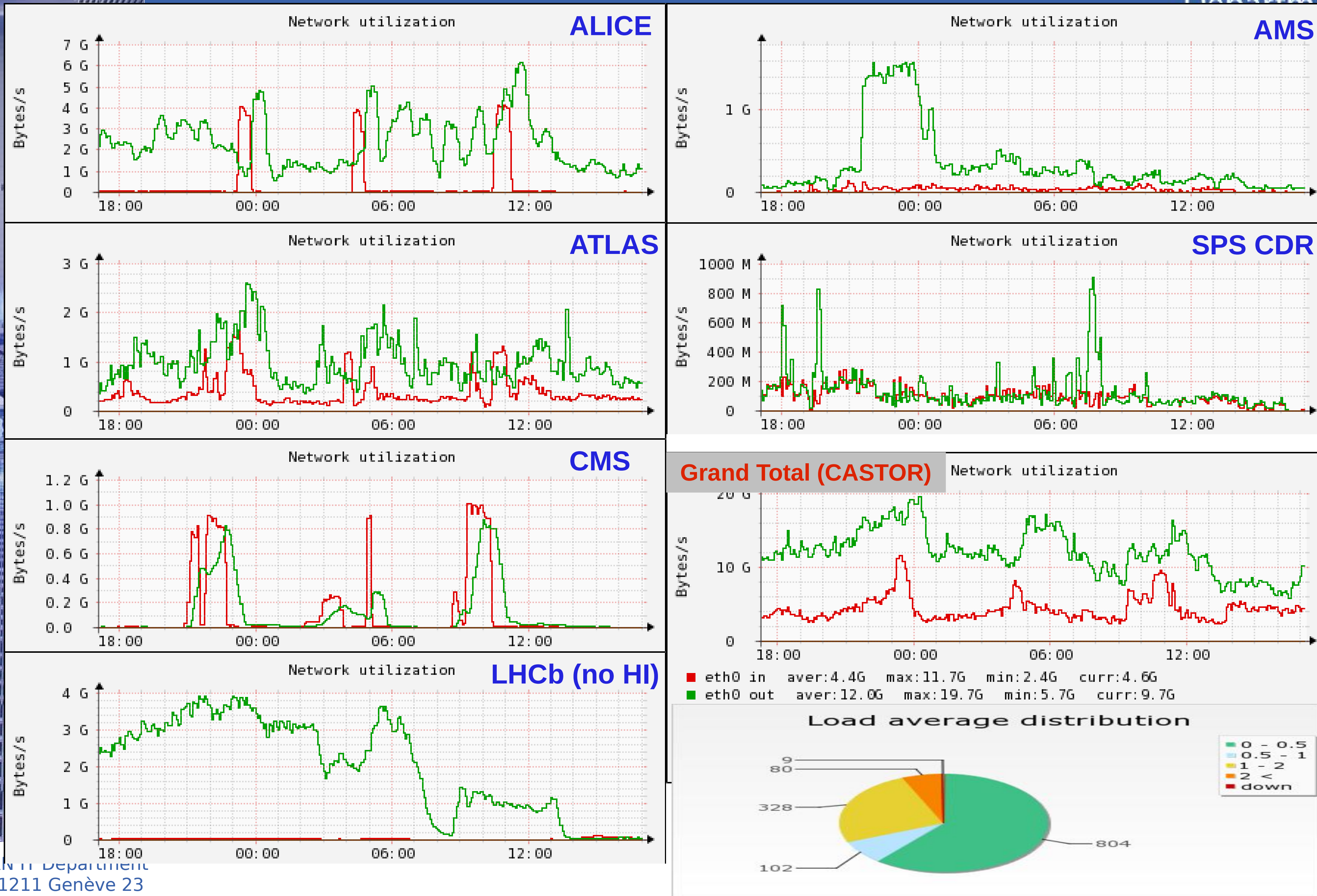


CERN IT Department

COMPASS

Run 2010 - Collected data to CASTOR




Network utilization
ALICE
Bytes/s
7 G
6 G
5 G
4 G
3 G
2 G
1 G
0
18:00
00:00
06:00
12:00
Network utilization
AMS
Bytes/s
1 G
0
Network utilization
ATLAS
3 G
2 G
1 G
Network utilization
SPS CDR
1000 M
800 M
600 M
400 M
200 M
Network utilization
CMS
1.2 G
1.0 G
0.8 G
0.6 G
0.4 G
0.2 G
0.0 G
Grand Total (CASTOR)
Network utilization
10 G
eth0 in aver:4.4G max:11.7G min:2.4G curr:4.6G
eth0 out aver:12.0G max:19.7G min:5.7G curr:9.7G
Network utilization
LHCb (no HI)
4 G
3 G
2 G
1 G
Load average distribution
0 - 0.5
0.5 - 1
1 - 2
2 <
down
9
80
328
102
804

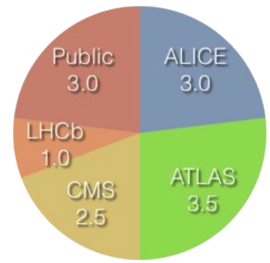
Data management today

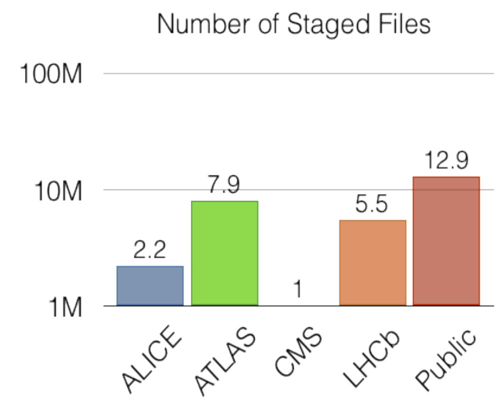
CERN IT Department

- CASTOR for Tier0
  - "RAW" data, tape access
- EOS for analysis
  - Analysis data, disk only
- Optimised for different use cases (T0 and analysis)

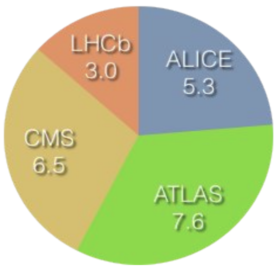
CASTOR
CERN Advanced STORage manager

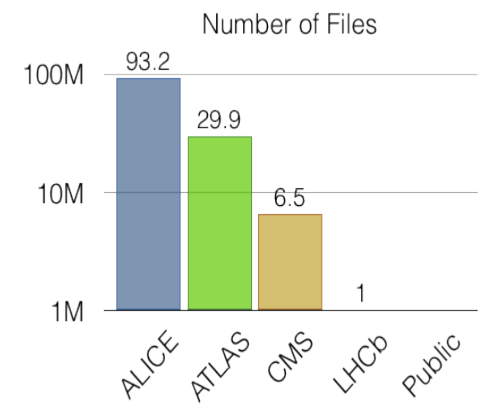
Installed (usable) disk capacity





Installed (usable) capacity





Total installed disk capacity :
13.0 PB + 22.4PB
Over 100 PB in the CERN store (disk+tape)
NB: 1 week of LHC (2012) = 1 PB on tape!

DSS/FDO weekly report (03 May 2013)

ACRON service

Instance	Jobs	Efficiency (*)	Unique user/host
@cern	859.6 K	94.8 %	689.0

AFS service

Instance	Capacity	Files	Δ	Size	Δ
@cern	501.7 TB	1.7 G	20.3 M	198.8 TB	3.2 TB

FILER service

Instance	Capacity	Files	Δ	Size	Δ
@cern	192.0 TB	118.4 M	0.7 M	23.7 TB	0.1 TB

CASTOR service

Instance	Files	Δ	Size	Δ	OnTape	Δ
@cern	311.2 M	-0.2 M	86.3 PB	132.0 TB	74.6 PB	121.0 TB

EOS service

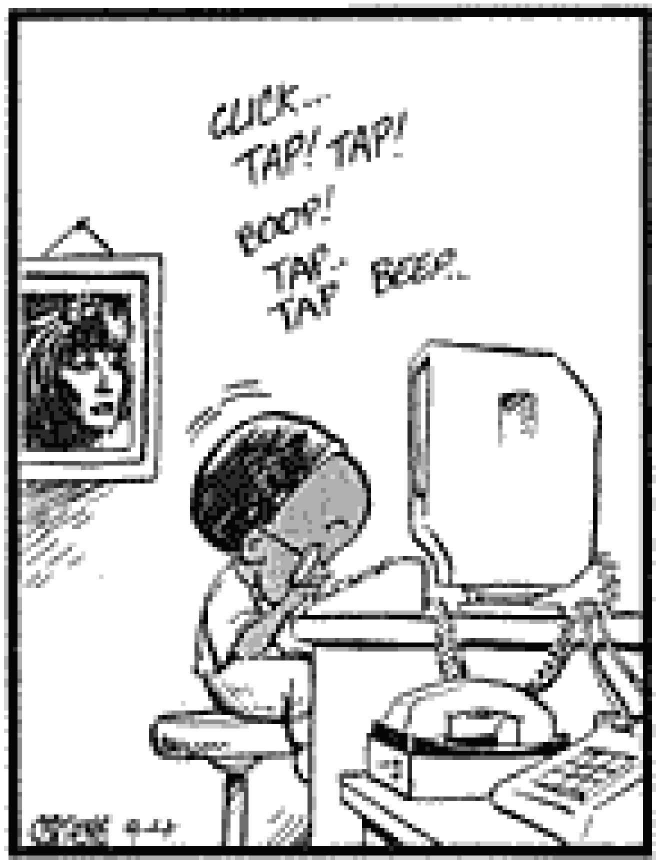
Instance	Files	Δ	Size	Δ
@cern	134.5 M	-1.7 M	15.7 PB	462.1 TB
alice	96.8 M	-2.4 M	4.1 PB	-86.4 TB
atlas	29.9 M	0.5 M	4.2 PB	132.8 TB
cms	6.8 M	0.2 M	4.8 PB	400.0 TB
lhcb	1.0 M	0.1 M	2.5 PB	15.8 TB

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

Data durability on a disk farm

11 RAID1 pairs

```
% df -H
Filesystem            Size   Used  Avail Use% Mounted on
…
/dev/sdb              1.9T   2.2G   1.9T   1% /srv/castor/01
/dev/sdc              2.0T   2.2G   2.0T   1% /srv/castor/02
/dev/sdd              2.0T   2.2G   2.0T   1% /srv/castor/03
/dev/sde              2.0T   2.2G   2.0T   1% /srv/castor/04
/dev/sdf              2.0T   2.2G   2.0T   1% /srv/castor/05
/dev/sdg              2.0T   2.2G   2.0T   1% /srv/castor/06
/dev/sdh              2.0T   2.2G   2.0T   1% /srv/castor/07
/dev/sdi              2.0T   2.2G   2.0T   1% /srv/castor/08
/dev/sdj              2.0T   2.2G   2.0T   1% /srv/castor/09
/dev/sdk              2.0T   2.2G   2.0T   1% /srv/castor/10
/dev/sdl              2.0T   2.2G   2.0T   1% /srv/castor/11
```



2 TB * 11 filesystem (* 2 copies)
~ 44 TB raw storage in one box

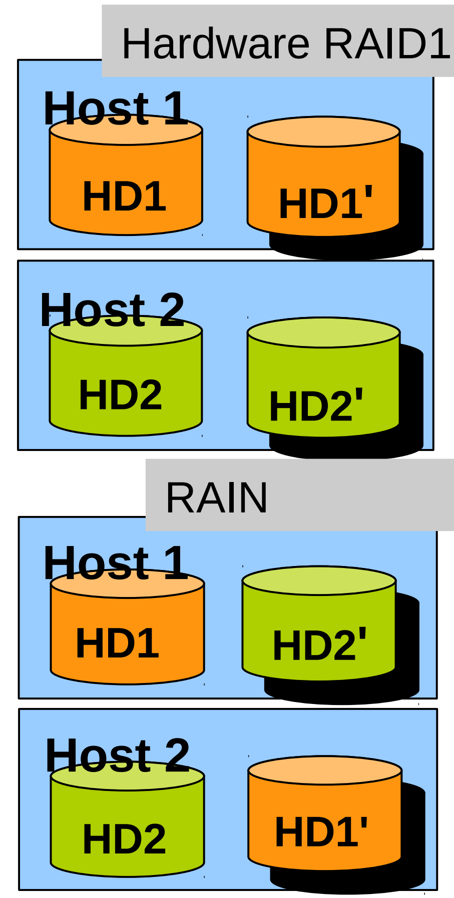
1000 boxes → 2000 disks (JBOD)

+ compact (~1 PB in 4 boxes)
– potentially “too little spindles” (analysis)
– potentially “too small network I/F” (analysis)
– data placement problems?

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

Different setups (mirror)

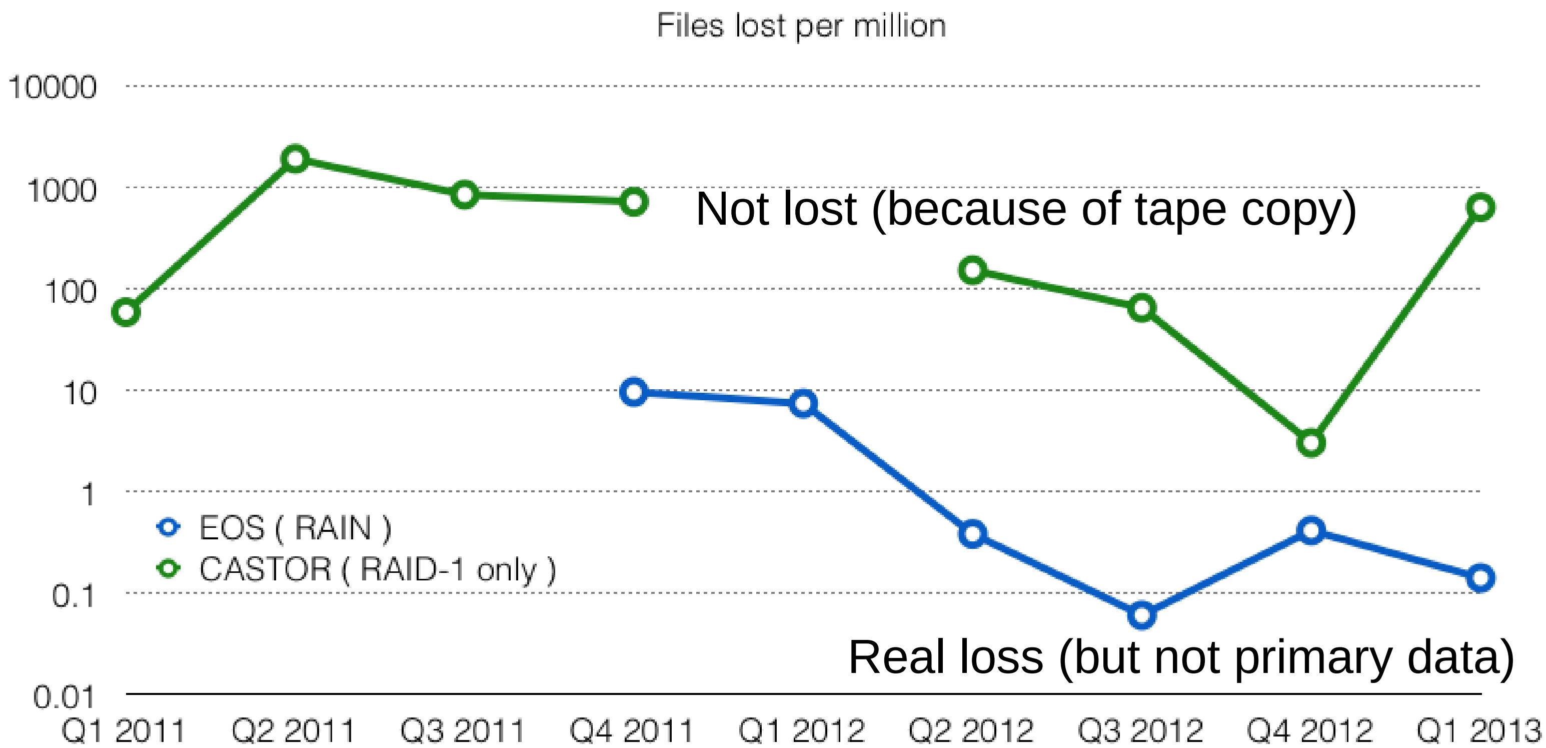
- Durability (data on HD1)
  - ≈ 1 – p(HD1) * p(HD1')
    - First approximation dominated by the number of replicas
    - If anything, in set up 1 there is a correlation (“positive”) which accounts for correlated failures in HD1 and HD1'
- Availability (data on HD1)
  - Setup 1
    - 1– p(host1)
  - Setup 2
    - ≈ 1 – p(host1) * p(host2)



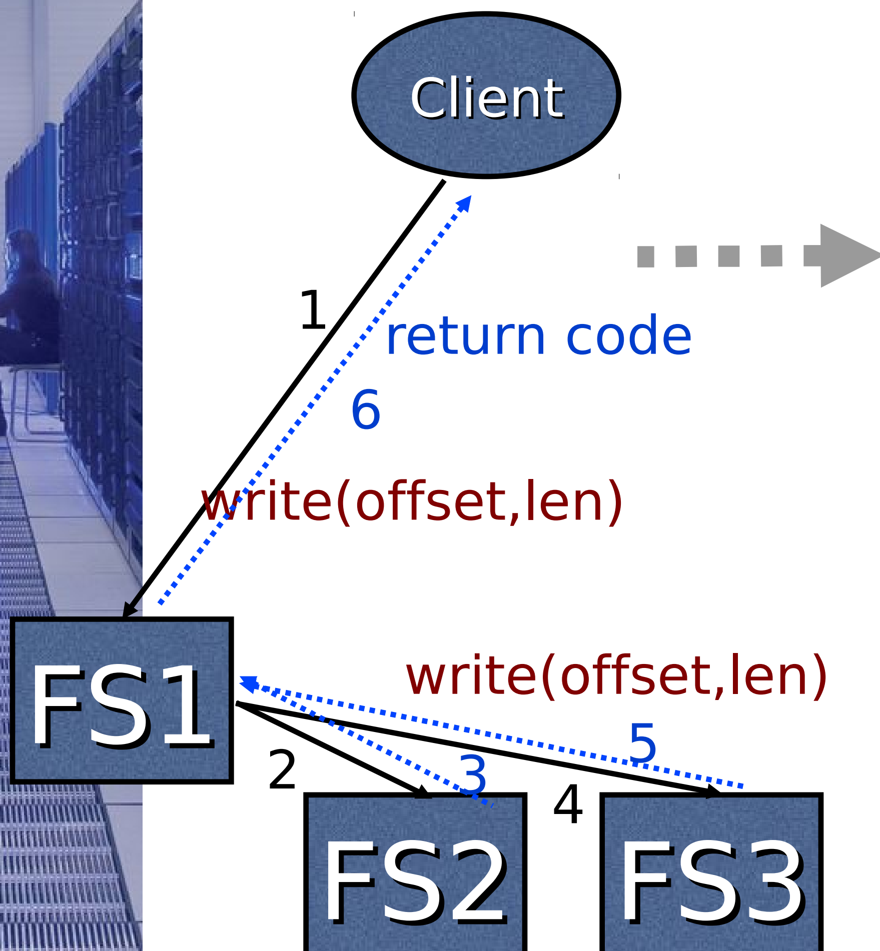
As today:
CASTOR: mainly Hardware RAID (RAID1)
EOS: tunable number of full copies (n>1)

Reliability... Good

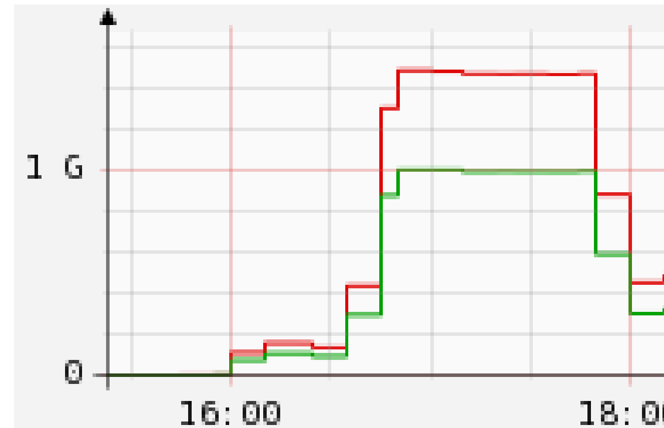
- File loss is not nice but unavoidable with a certain probability
  - RAID-1 does not protect against controller or machine problem, filesystem corruption and finger trouble
  - typically important files can be recovered from offsite
- In case of backup (CASTOR) the tape reliability is helping the disk one



EOS replica mechanics



Network IO for file creations with 3 replicas:

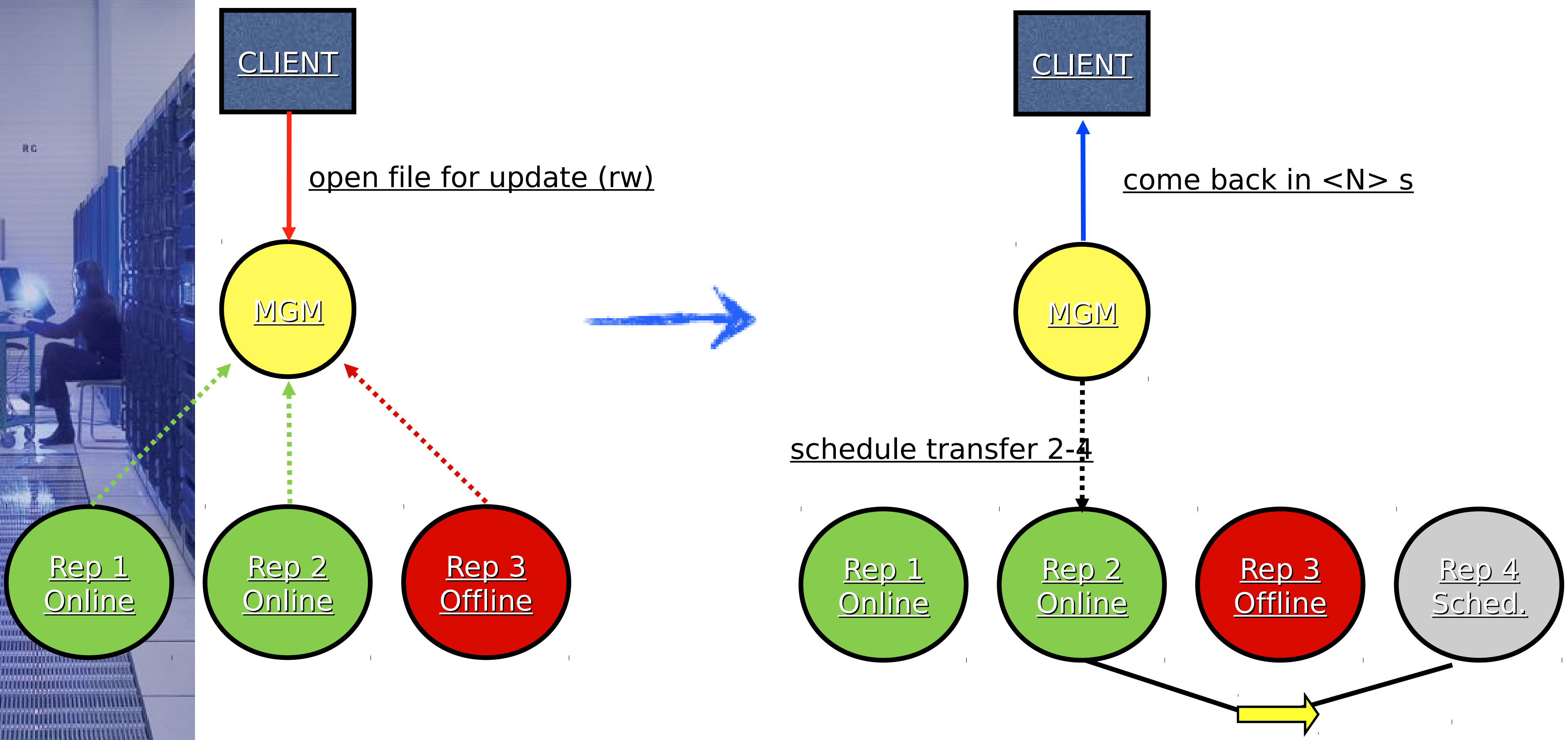


500 MB/s injection result in
- 1 GB/s output on eth0 of all disk servers
  (write out of FS1 0.5 GB/s twice)
- 1.5 GB/s input on eth0 of all disk servers
  (0.5 GB/s x 3 copies)

Plain (no replica)
Replica (here 3 replicas)
More sophisticated redundant storage (RAID5, RAID6, LDPC)

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

EOS replica healing



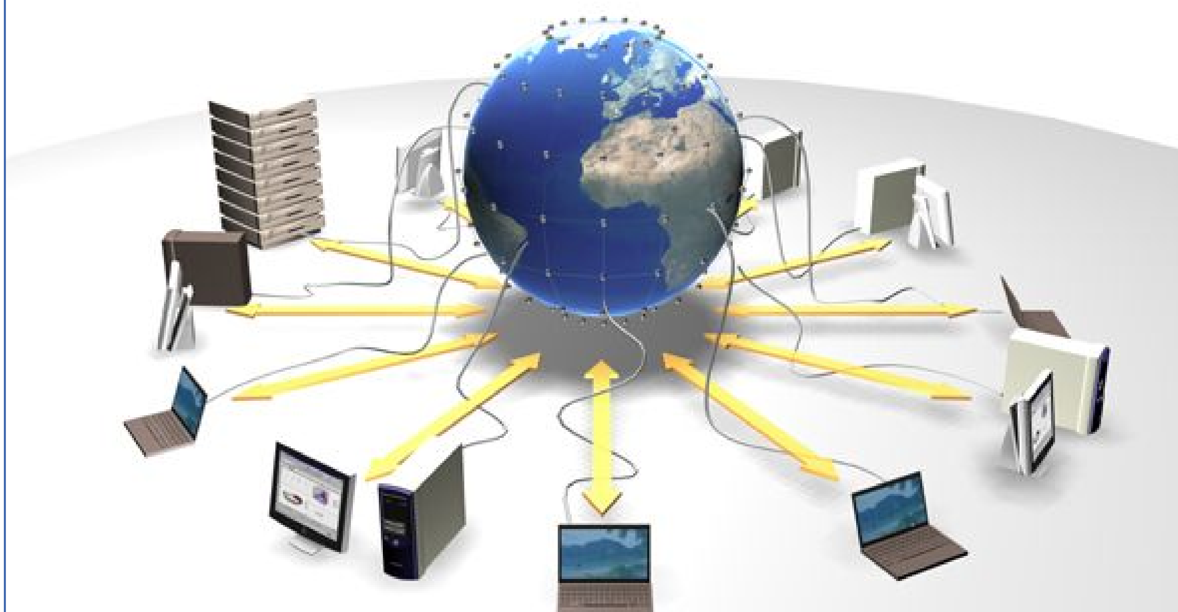
Client **RW** reopen of an existing file triggers
- creation of a new replica
-dropping of offline replica

This is at the heart of our model of operations (asynchronous operations)

Bringing all together...

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

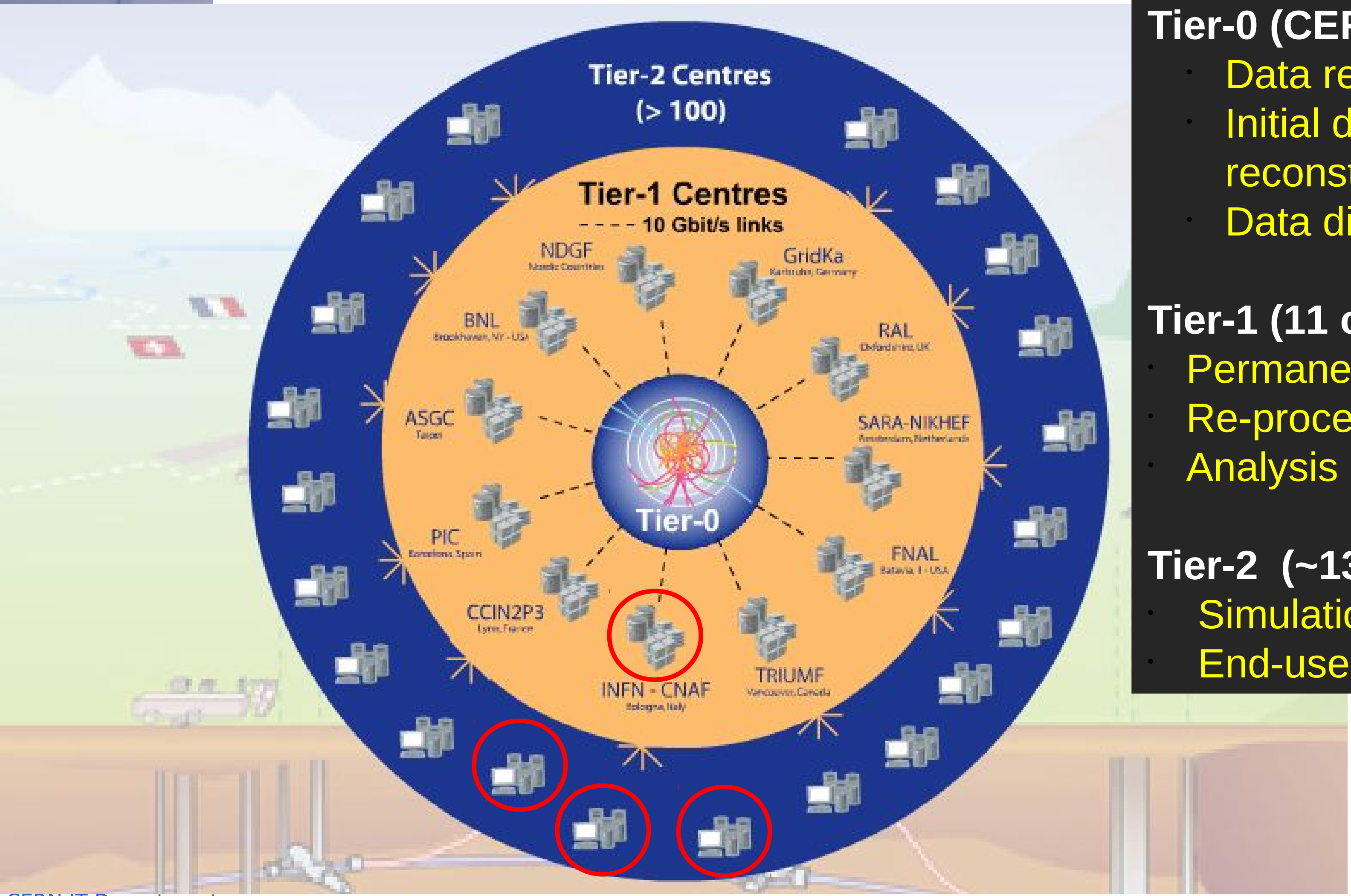
The **Grid** is an infrastructure that provides seamless access to computing power and data storage capacity distributed over the globe



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

WLCG Tiers Organization

CERN IT Department



Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (~130 centres):
- Simulation
- End-user analysis

How does it work?

ATLAS

Not substantially different for the other HEP experiments

Heavily simplified...

What do we want to achieve

The user wants to specify a subset of the data and run applications on it (chain of programs reading intermediate outputs)

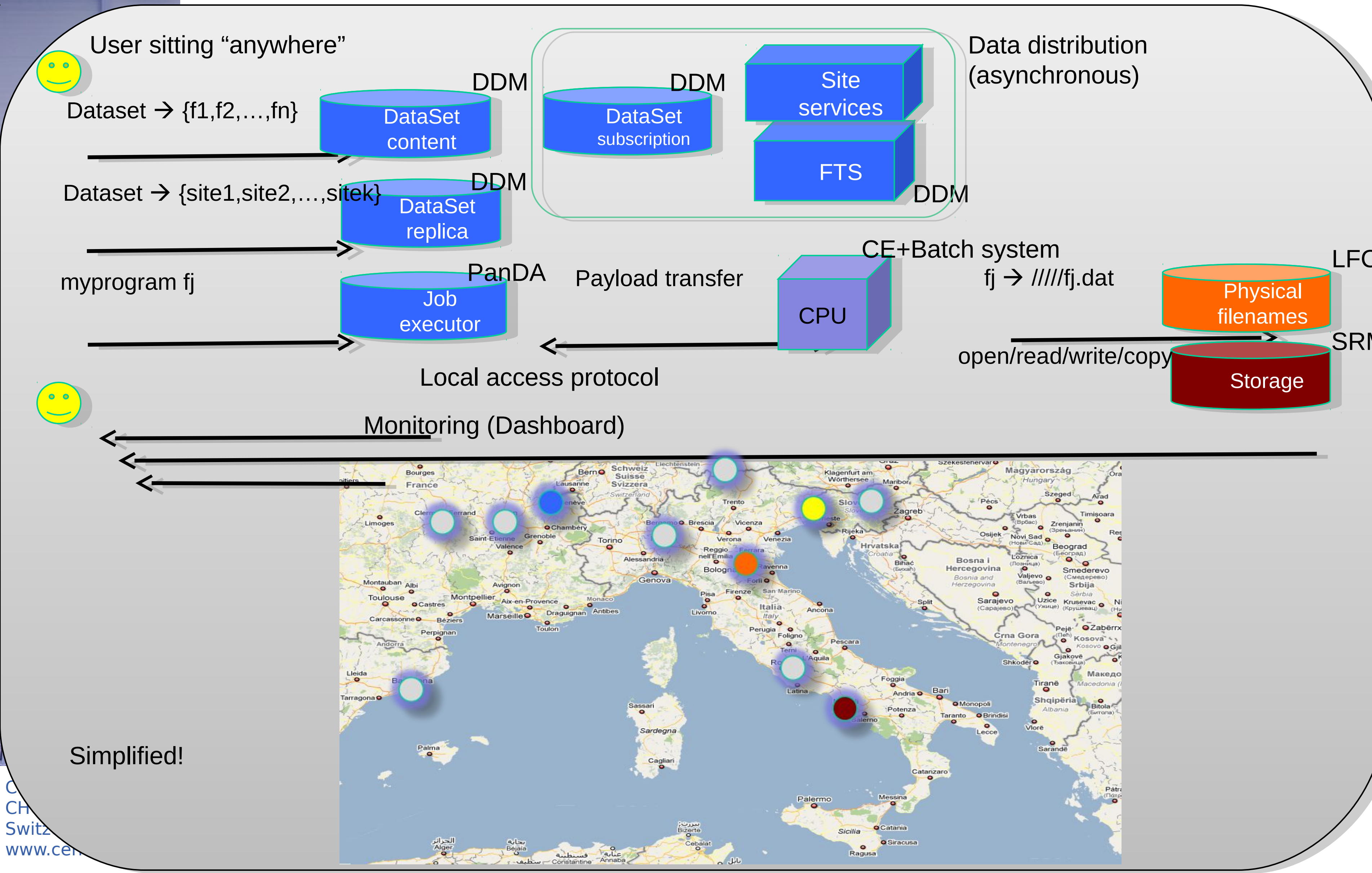
Only at the end of the chain data sizes and computational complexity this can be (possibly) done on a laptop

- 1000+ of physicists worldwide after the same data

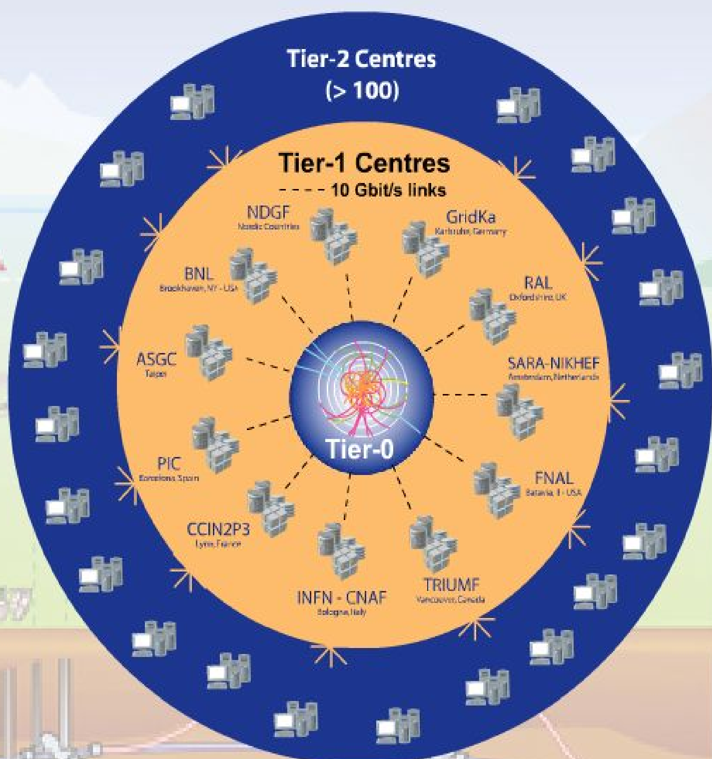
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Behind the scenes…

CERN IT Department



More info:
INFN (Istituto Nazionale Fisica Nucleare):
http://www.infn.it
IGI (Italian Grid Initiative): http://www.italiangrid.org/

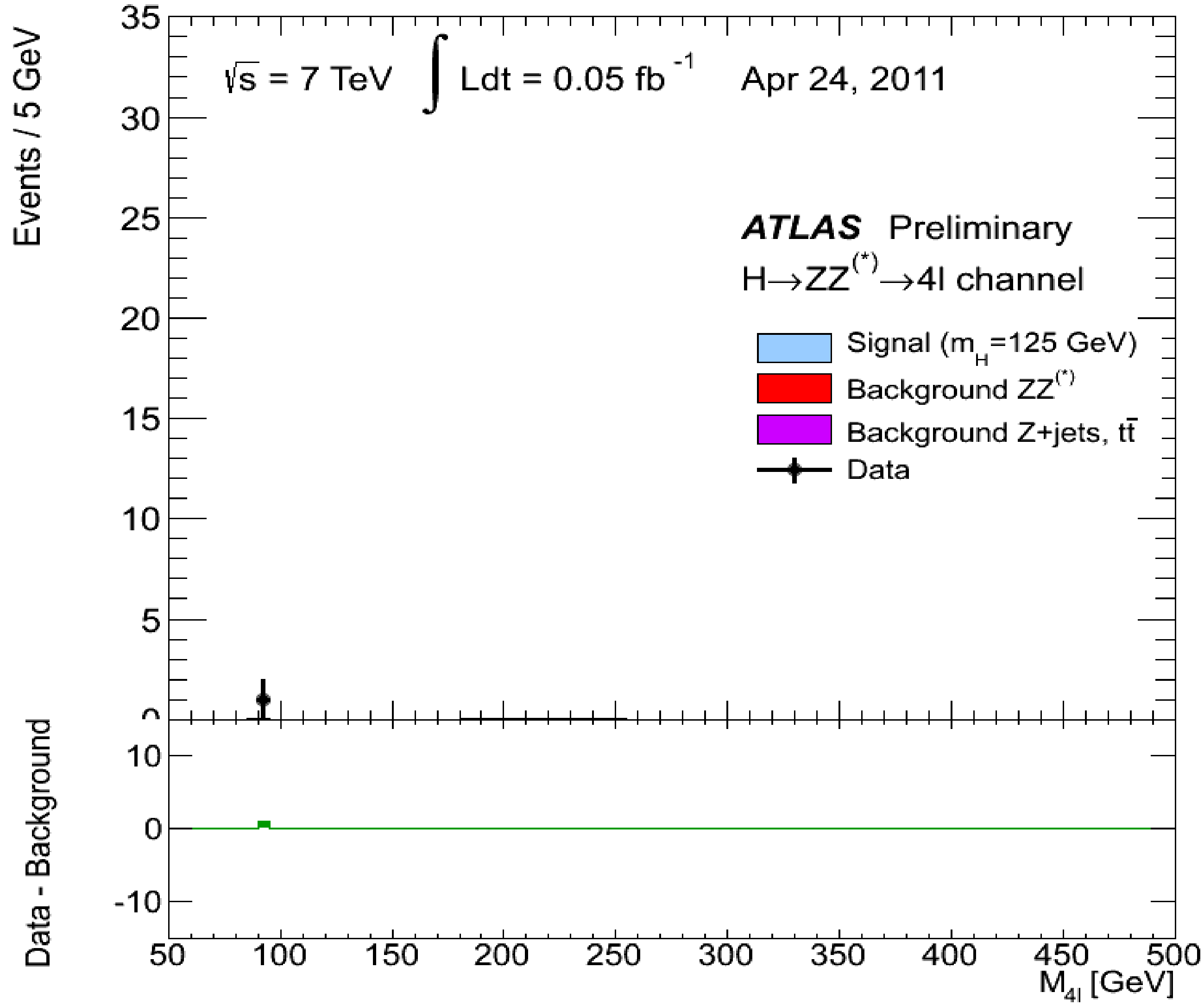


Switzerland
www.cern.ch/it

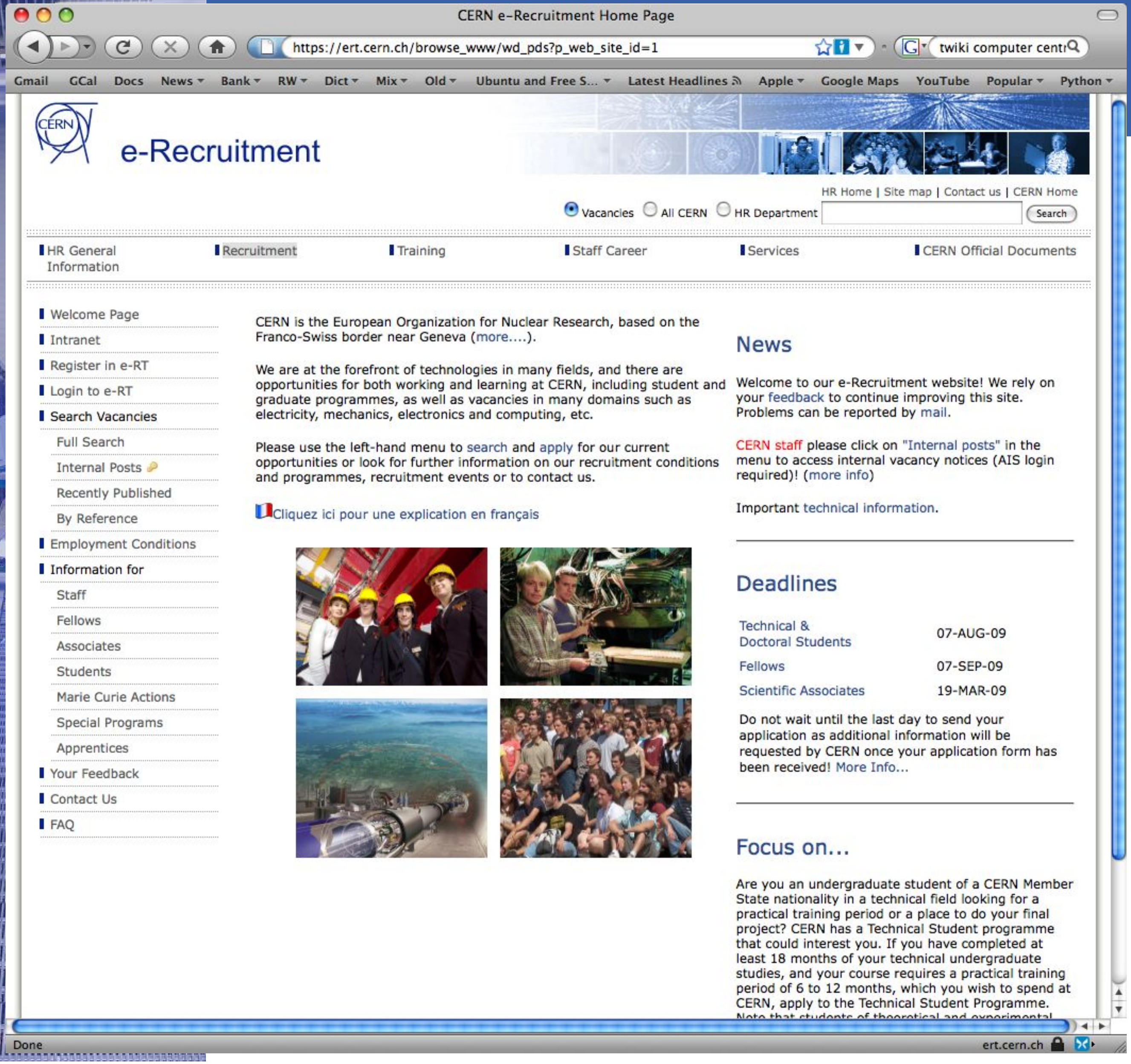
Outlook

- CERN data for LHC:
  - Solid foundation (CASTOR)
  - Successful new product (EOS)
  - CASTOR and EOS coexist (complementary)
- Goals
  - Stability with low operations costs
  - Open to adapt to new ideas (also from non-HEP areas)
- Open directions ahead:
  - Data management at the heart of our activity
    - Critical for the success of physics research
    - An exciting field of study by itself
    - New stuff coming
      - » Is it any good?

Questions?



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



CERN options for students

University level (BS/Master)

- Summer student
- OpenLab summer students

Master thesis

- Technical student (non physicist)

PhD students

- Doctoral students

Young scientists/engineers

- Fellowship
- Other programmes