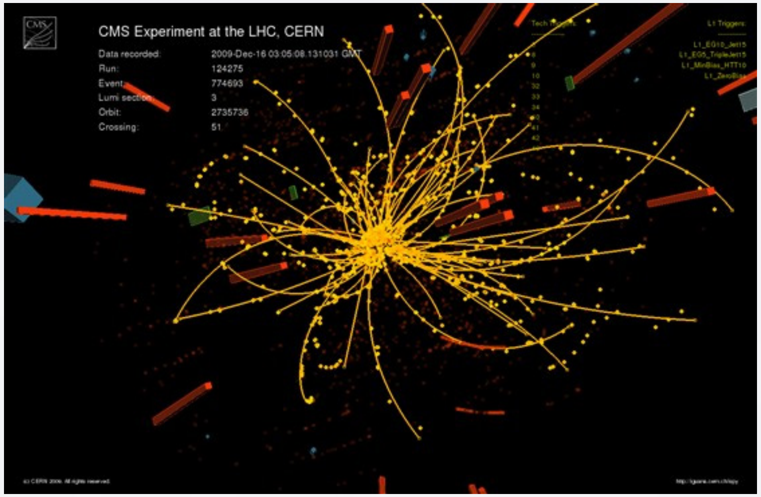
The Extreme Scale: Data Analytics at LHC (the Large Hadron Collider)

A view from 10'000 m



ISC 2012

Sverre Jarp, CERN openlab CTO

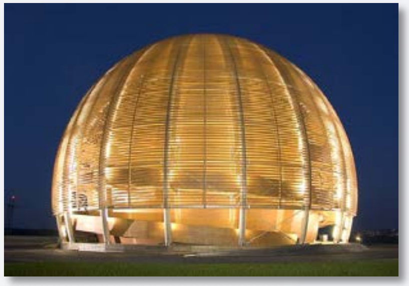
Hamburg, 20 June 2012

Analytics ?

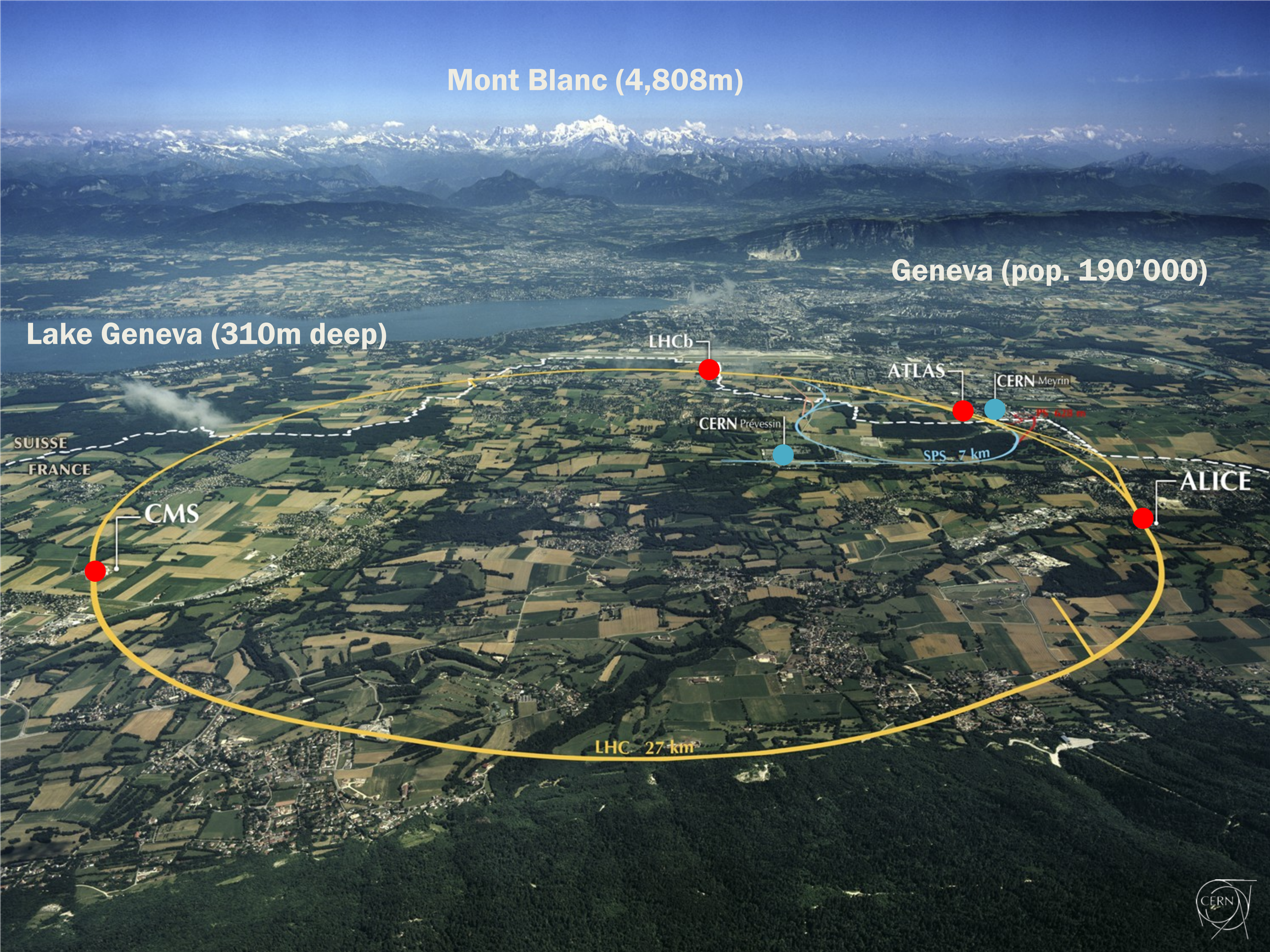
- **According to Wikipedia:**
  - **“Analytics is the discovery and communication of meaningful patterns in data. It relies on the simultaneous application of Statistics, Computer Programming and Operations Research to approach problems in business and industry. Analytics often favours Data Visualization to communicate insight”**

About CERN

- **CERN is the European Organization for Nuclear Research in Geneva**
  - Particle accelerators and other infrastructure for high energy physics (HEP) research
  - Worldwide community
    - 20 members states (+ 3 incoming members)
    - Observers: Turkey, Russia, Japan, USA, India
    - About 2300 staff
    - >10’000 users (about 5’000 on-site)
    - Budget (2011) ~1000 MCHF

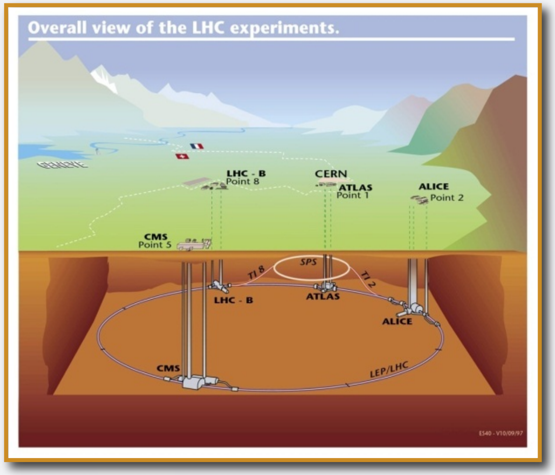


- **Birthplace of the World Wide Web**

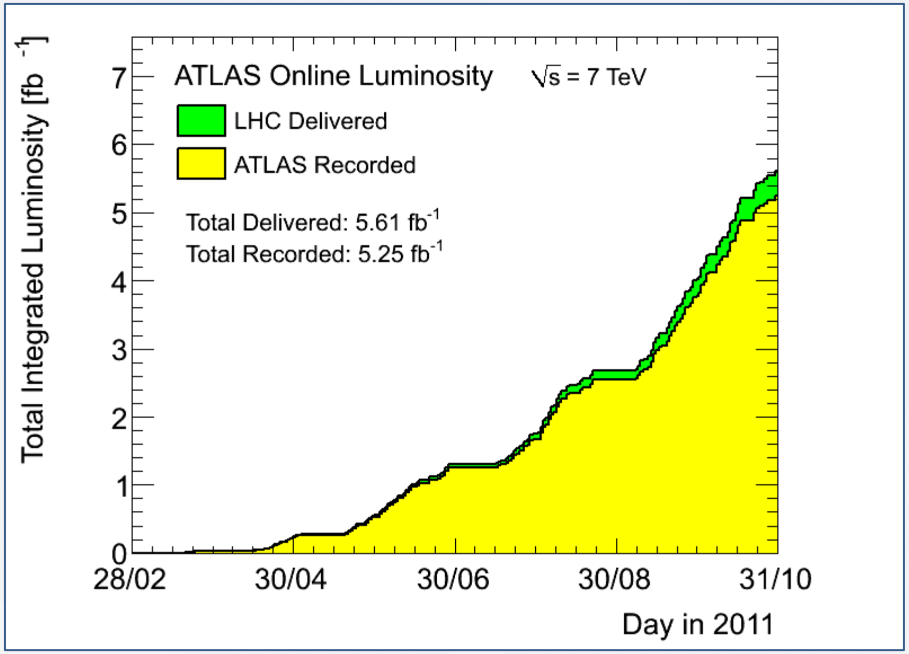

Mont Blanc (4,808m)
Geneva (pop. 190'000)
Lake Geneva (310m deep)
LHCb
ATLAS
CERN Meyrin
PS 628 m
CERN Prévessin
SPS 7 km
SUISSE
FRANCE
ALICE
CMS
LHC 27 km
CERN

Large Hadron Collider (LHC)

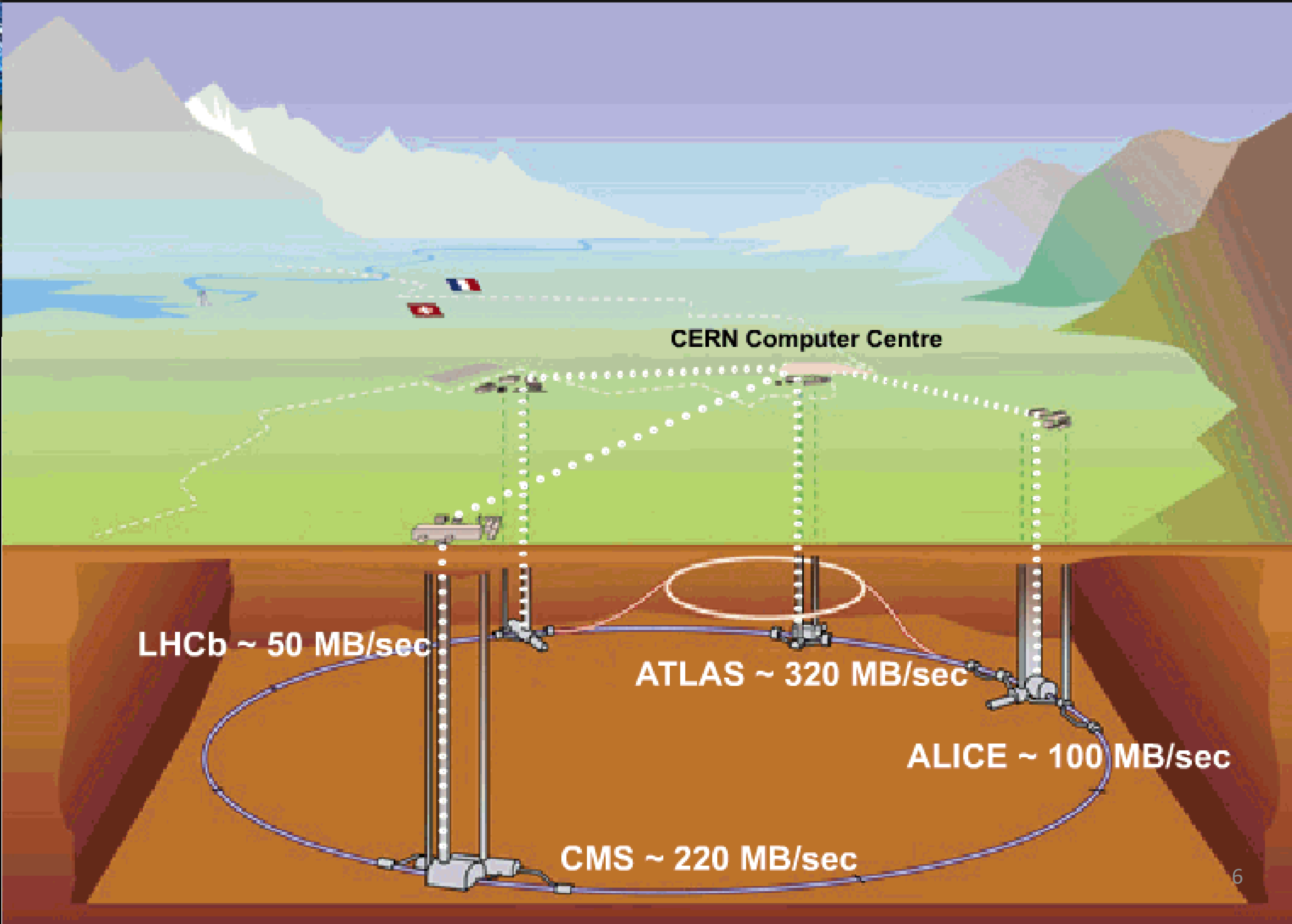
- **The biggest machine ever built**
  - 27 km, 100 meters below ground
- **Activities started in 2009**
  - Highest energy in an accelerator
  - Large data sample of recorded collisions (events) available for high energy physics (HEP) measurements



- ➢ 10^7 collisions per second
- ➢ Fortunately most collisions are uninteresting !

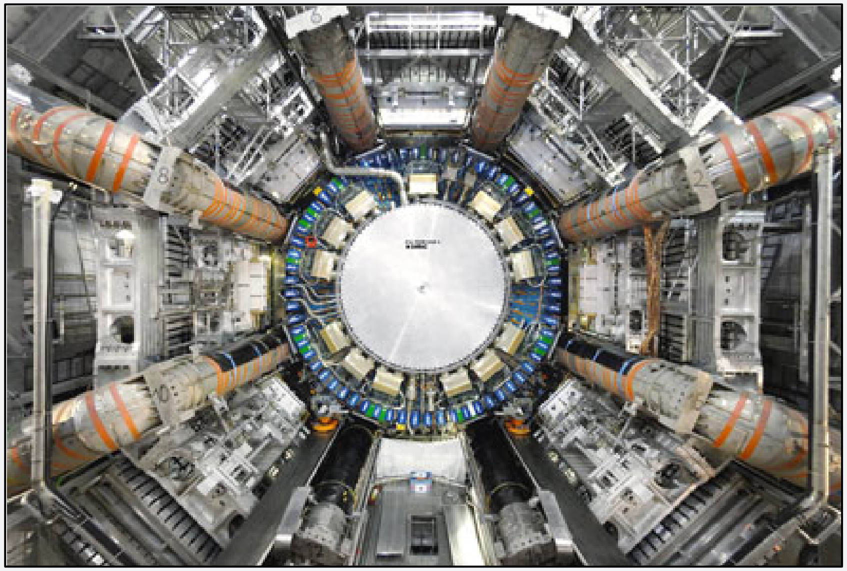


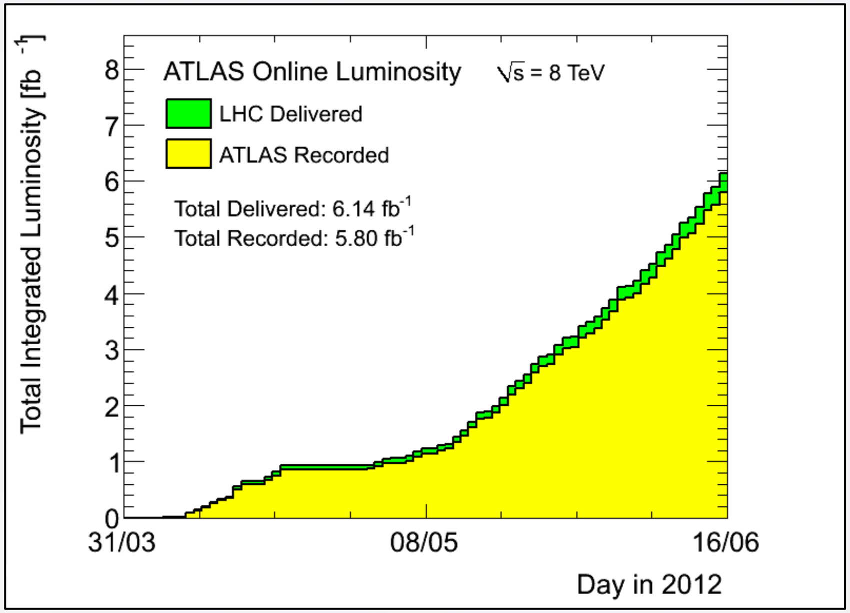
Data acquisition and storage for LHC @ CERN



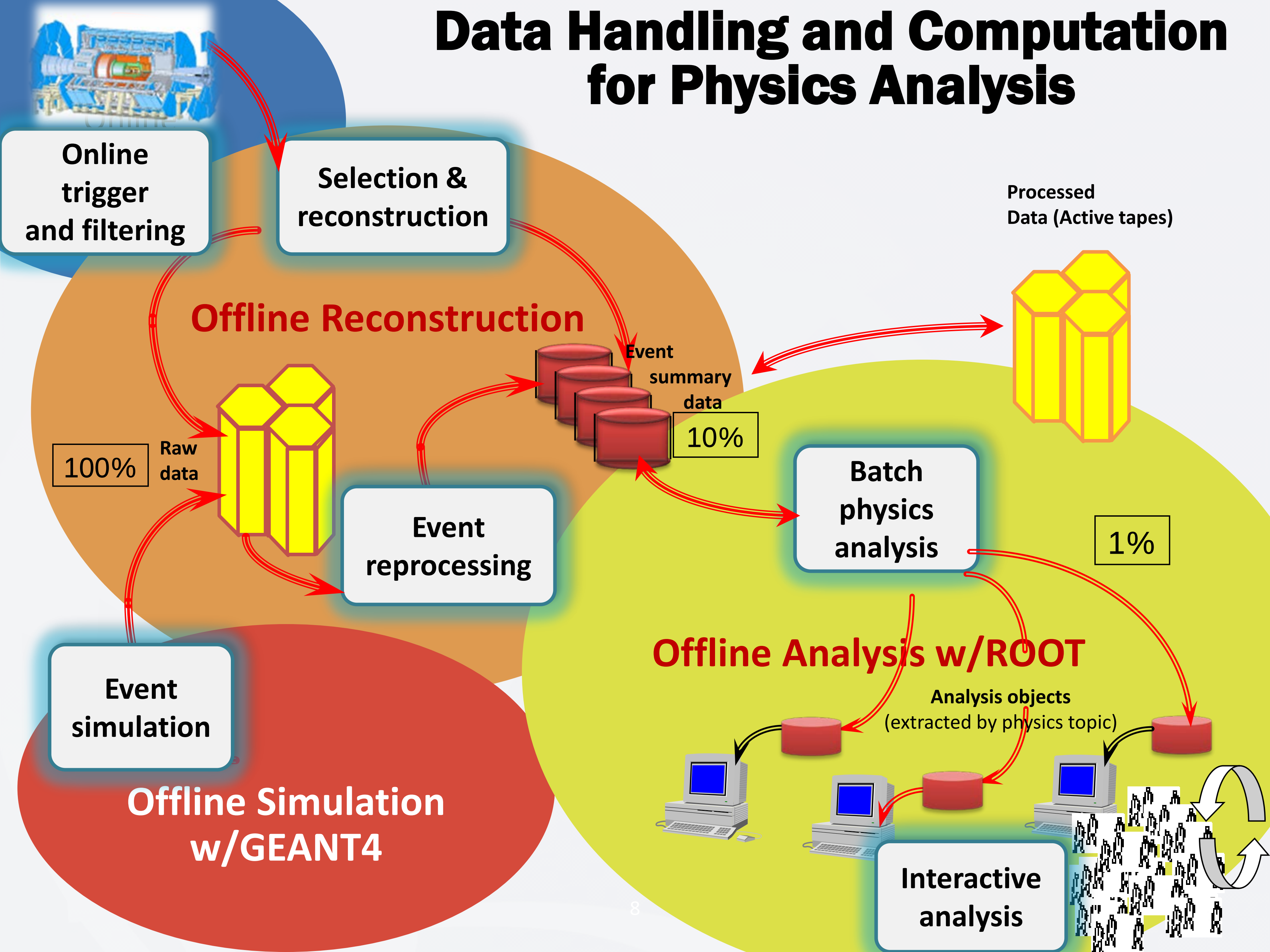
LHC is more productive than ever

- **From our home page (13 June 2012):**
  - It has already delivered more collisions than in the whole of 2011
  - Last year, ATLAS and CMS each recorded a total of around 5.6 inverse femtobarns of data. This measure of accelerator performance is equivalent to about 560 trillion proton-proton collisions. The accelerator today passed last year's totals and is well on its way its goal of delivering 1500 trillion proton-proton collisions in 2012.
  - The LHC is now operating at 1380 proton bunches per beam, the maximum value set for this year, with around 1.5 × 10^{11} protons in each bunch. The accelerator has also far exceeded the best instantaneous collision rate achieved last year: the maximum peak luminosity in 2011 was 3.6 × 10^{33} collisions per square centimetre per second; the LHC has now reached 6.8 × 10^{33} cm-2 s-1.
  - The higher collision energy of 4 TeV per beam this year (compared to 3.5 TeV per beam in 2011) and the resulting higher number of collisions are expected to enhance the machine's discovery potential considerably, opening up new possibilities in the searches for new and heavier particles.





Data Handling and Computation for Physics Analysis



Online trigger and filtering

Selection & reconstruction

Offline Reconstruction

Processed Data (Active tapes)

100%

Raw data

Event summary data

10%

Event reprocessing

Batch physics analysis

1%

Offline Analysis w/ROOT

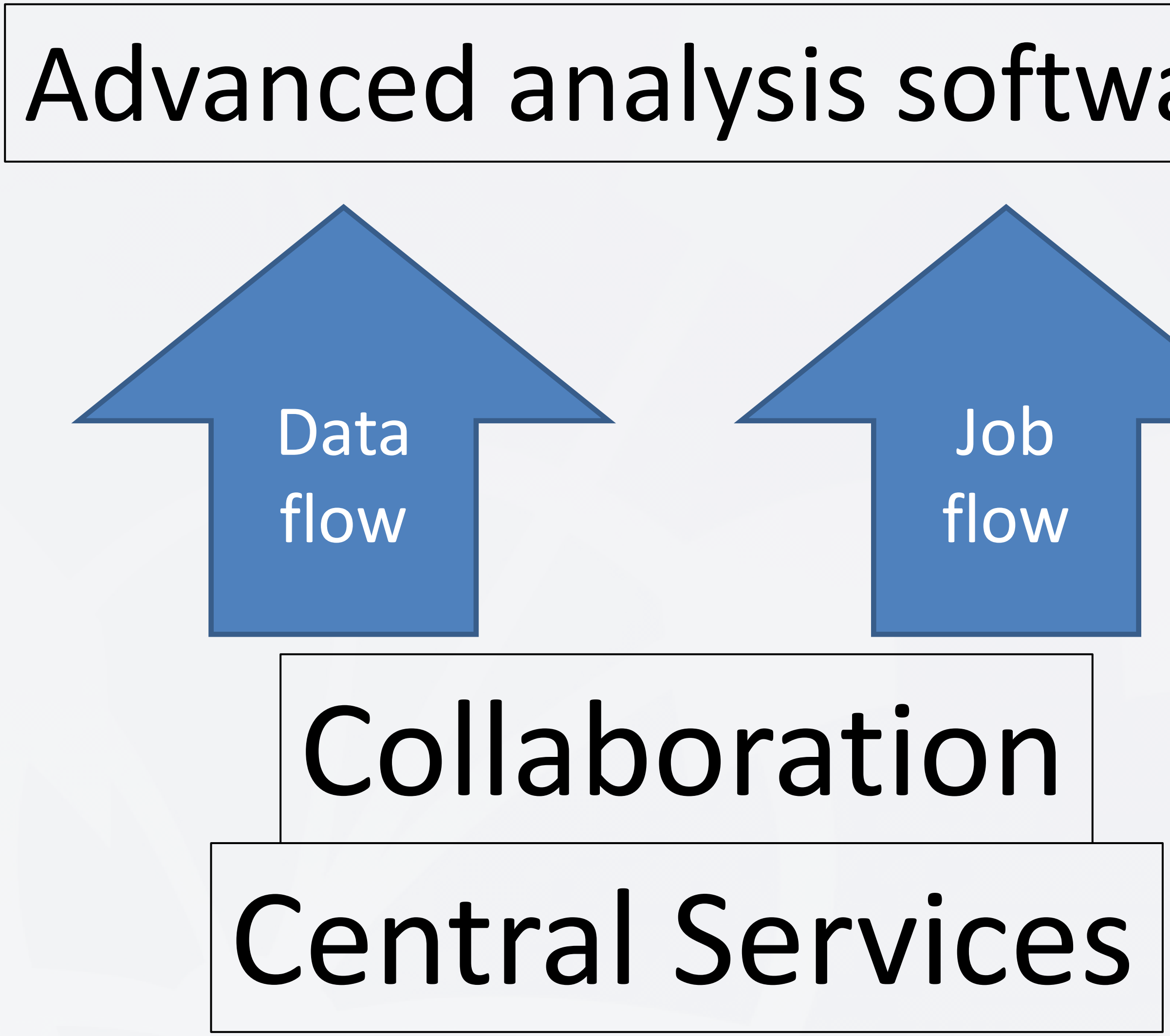
Analysis objects (extracted by physics topic)

Event simulation

Offline Simulation w/GEANT4

Interactive analysis

Overview



Advanced analysis software

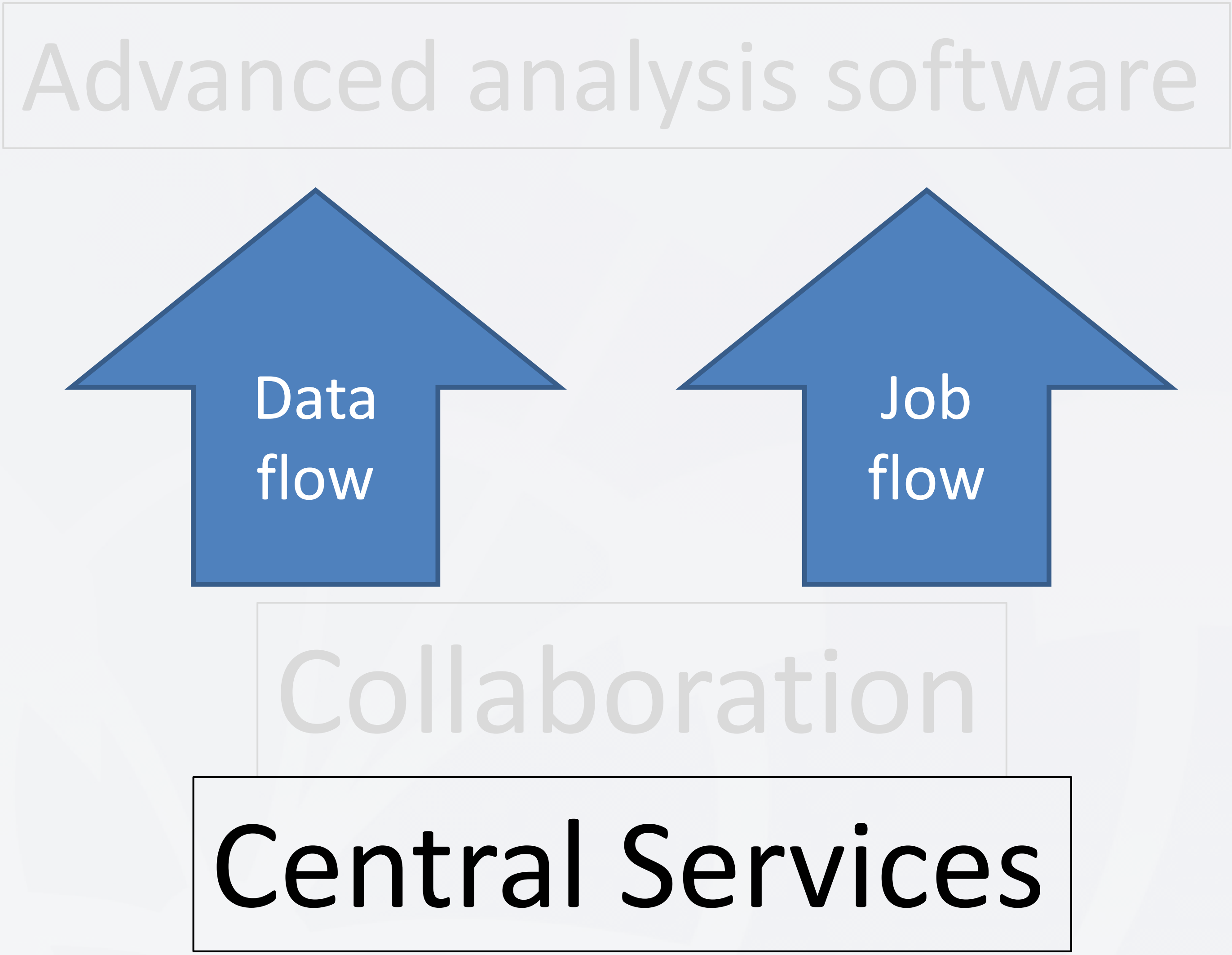
Data flow

Job flow

Collaboration

Central Services

Overview



Advanced analysis software

Data flow

Job flow

Collaboration

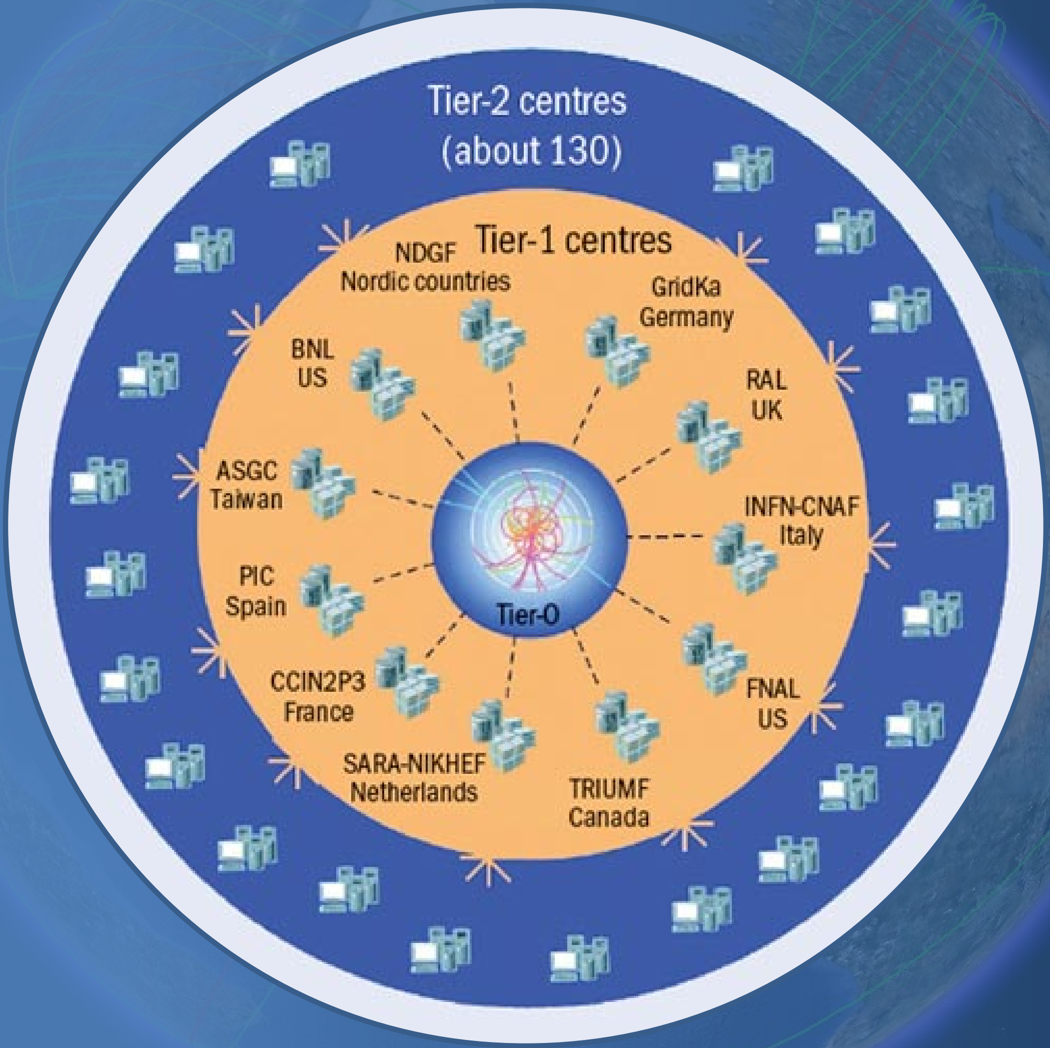
Central Services

The Worldwide LHC Computing Grid

Tier-0 (CERN): data recording, reconstruction and distribution

Tier-1: permanent storage, re-processing, analysis

Tier-2: Simulation, end-user analysis



nearly 160 sites

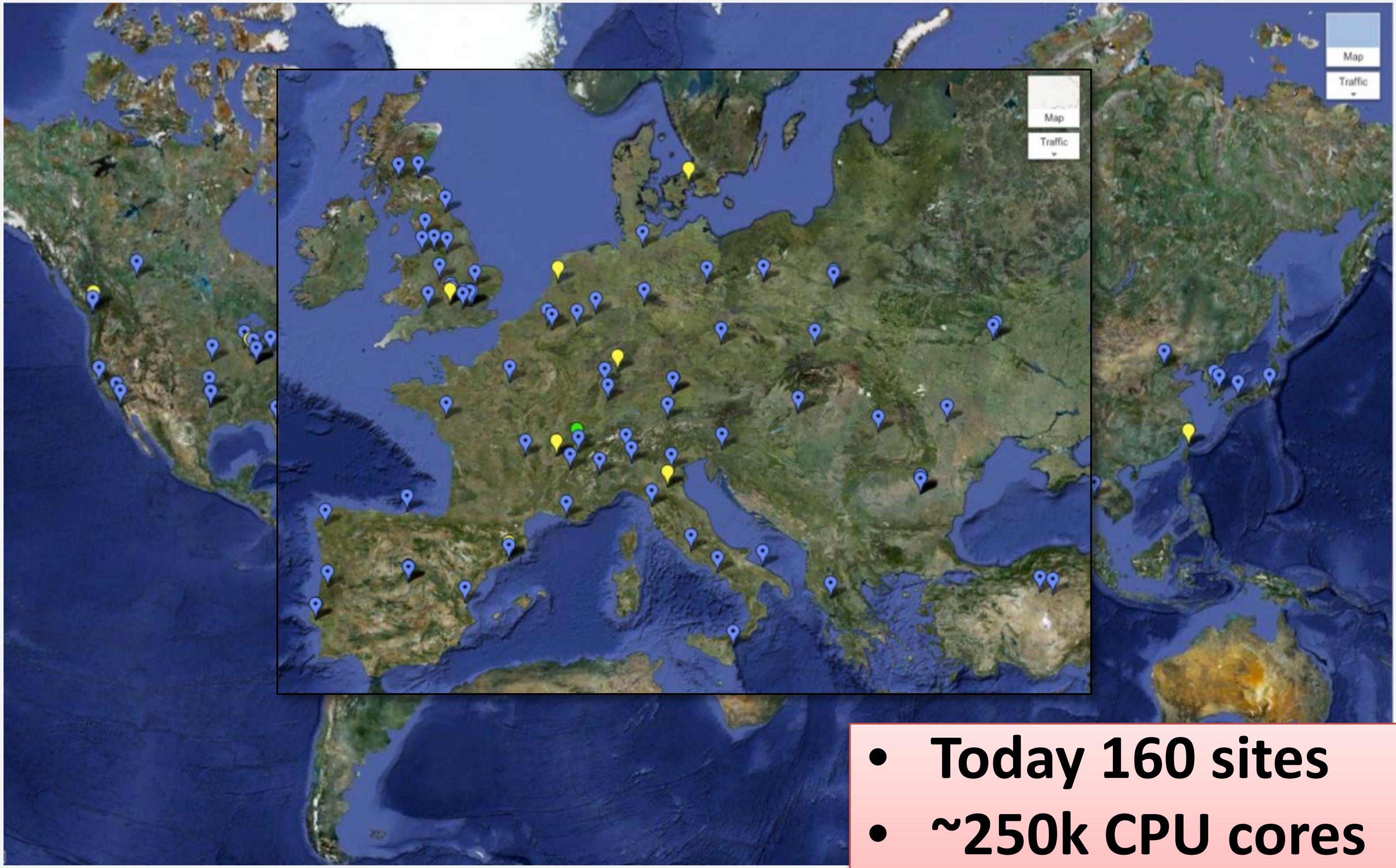
~250'000 cores

173 PB of storage

> 1 million jobs/day

10 Gb links

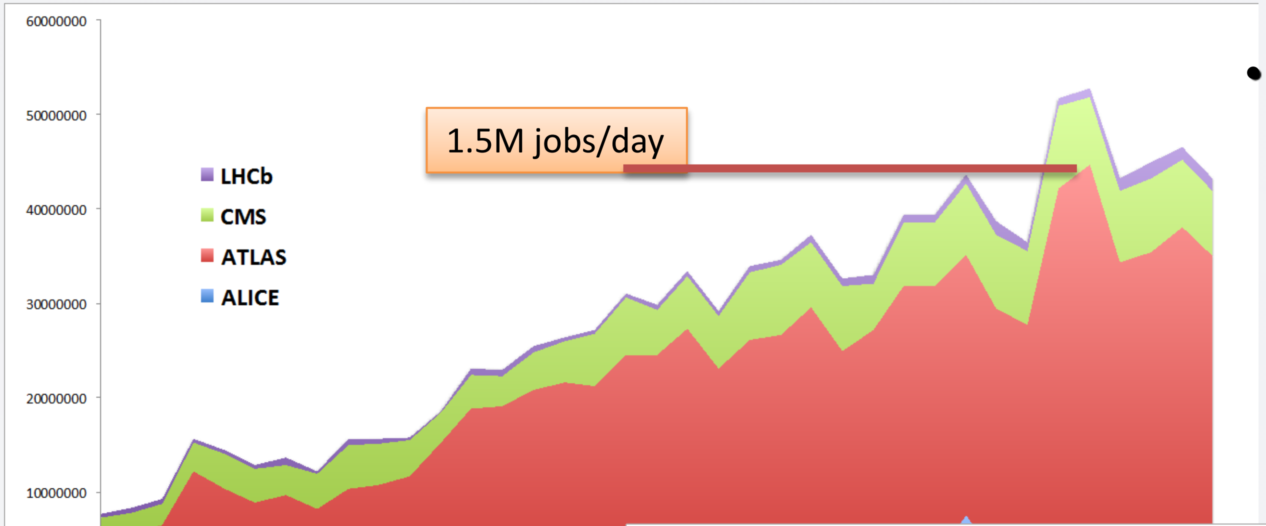
Larger picture: WLCG Grid Sites



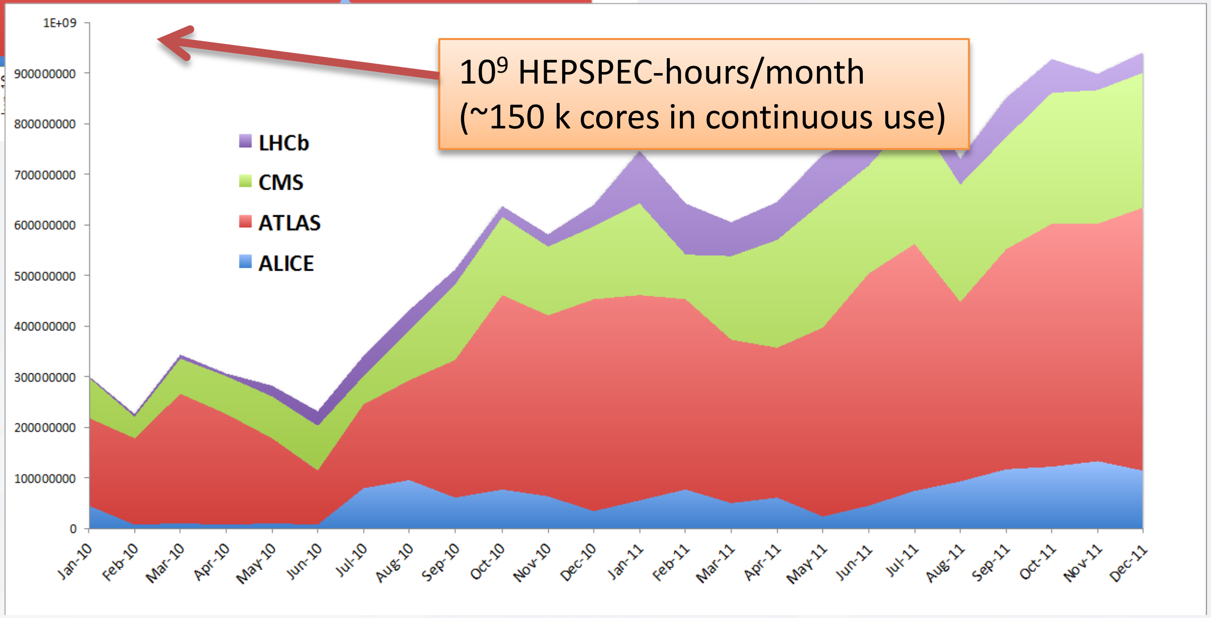
● Tier 0 ● Tier 1 ● Tier 2

- Today 160 sites
- ~250k CPU cores
- >150 PB disk

WLCG usage: continues to grow



1.5M jobs/day



10^9 HEPSPEC-hours/month
(~150 k cores in continuous use)

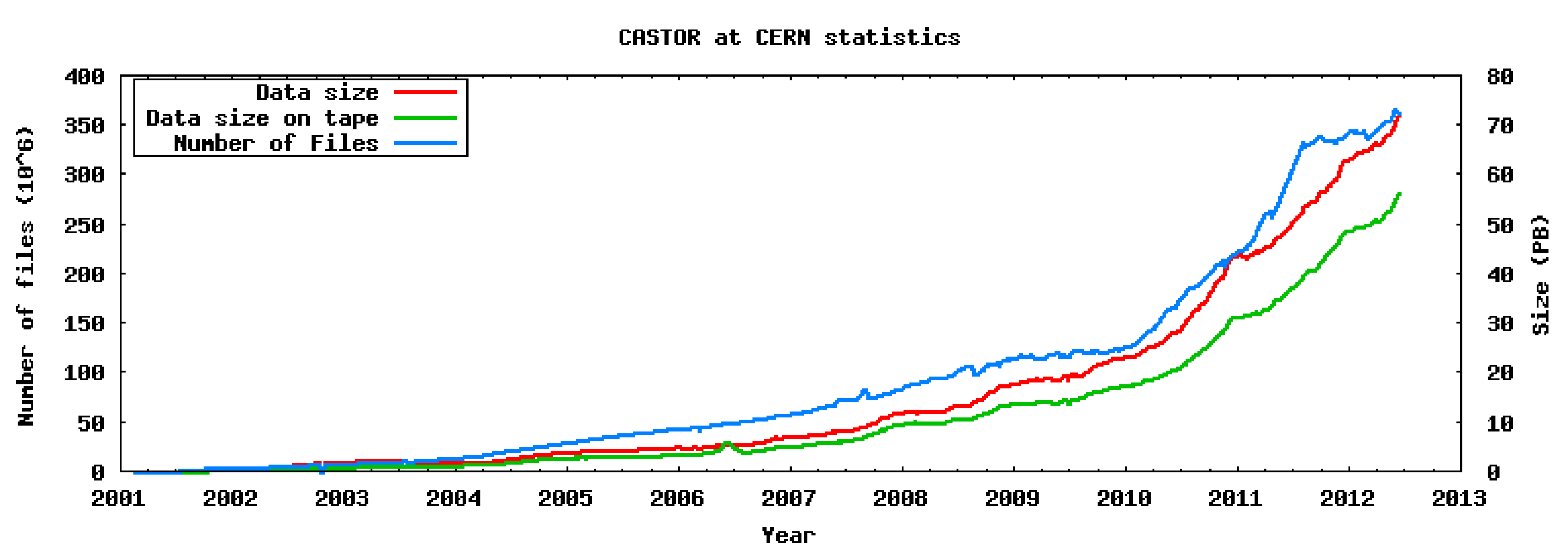
- **WLCG usage pattern:**
  - Continuous
  - Ever increasing load
  - Some spikes
  - Mirrored by contributing national Grids

- **Grid middleware handles both job flow and data flow**

Tier-0: Central Data Management

- **Hierarchical Storage Management: CASTOR**
  - Rich set of features:
    - Tape pools, disk pools, service classes, instances, file classes, file replication, scheduled transfers (etc.)
  - DB-centric architecture
- **Disk-only storage system: EOS**
  - Easy-to-use, stand-alone, disk-only for user and group data with in-memory namespace
    - Low latency (few ms for read/write open)
    - Focusing on end-user analysis with chaotic access
    - Adopting ideas from other modern file systems (Hadoop, Lustre, etc.)
    - Running on low-cost hardware (JBOD and sw RAID)

CASTOR statistics

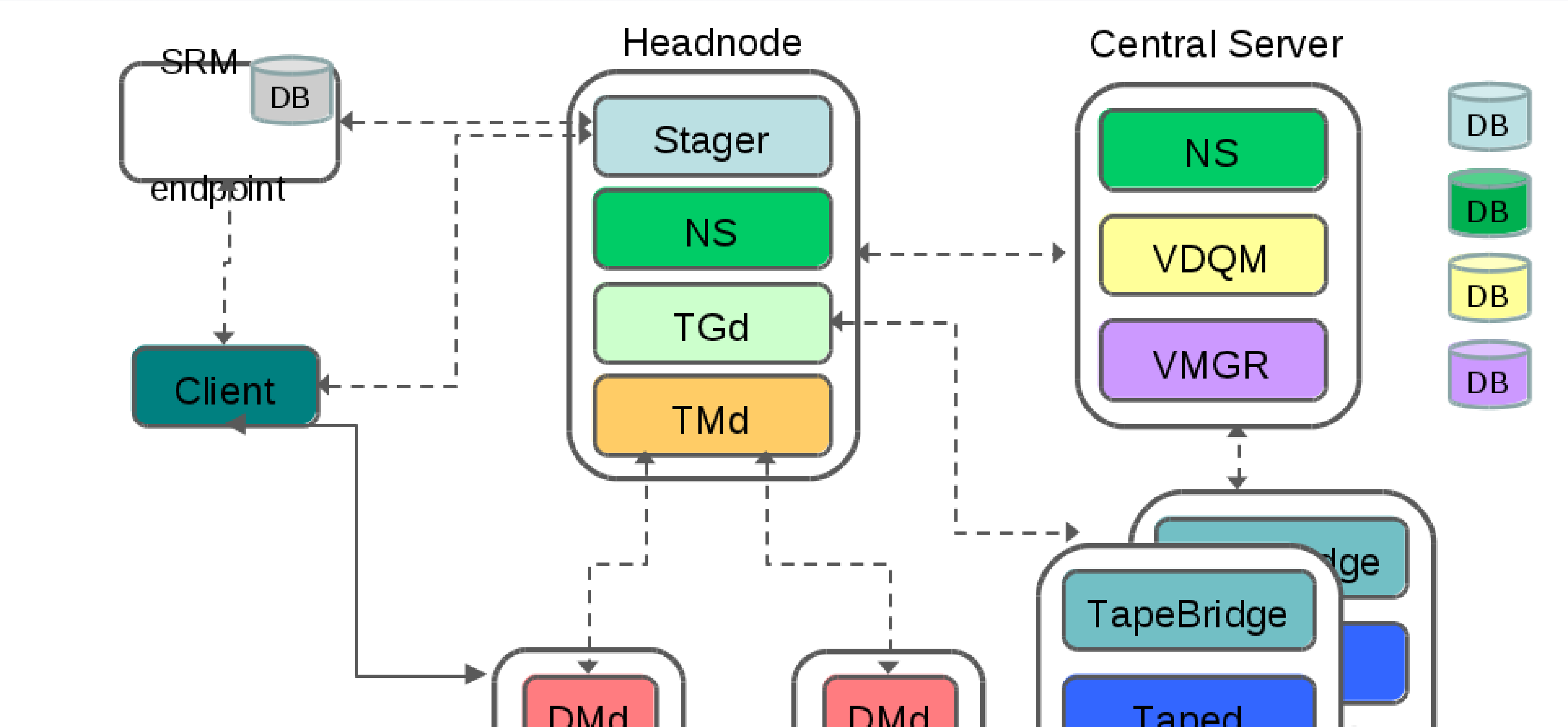


Current status:

66 petabytes across 362 million files

Credit: J.Iven

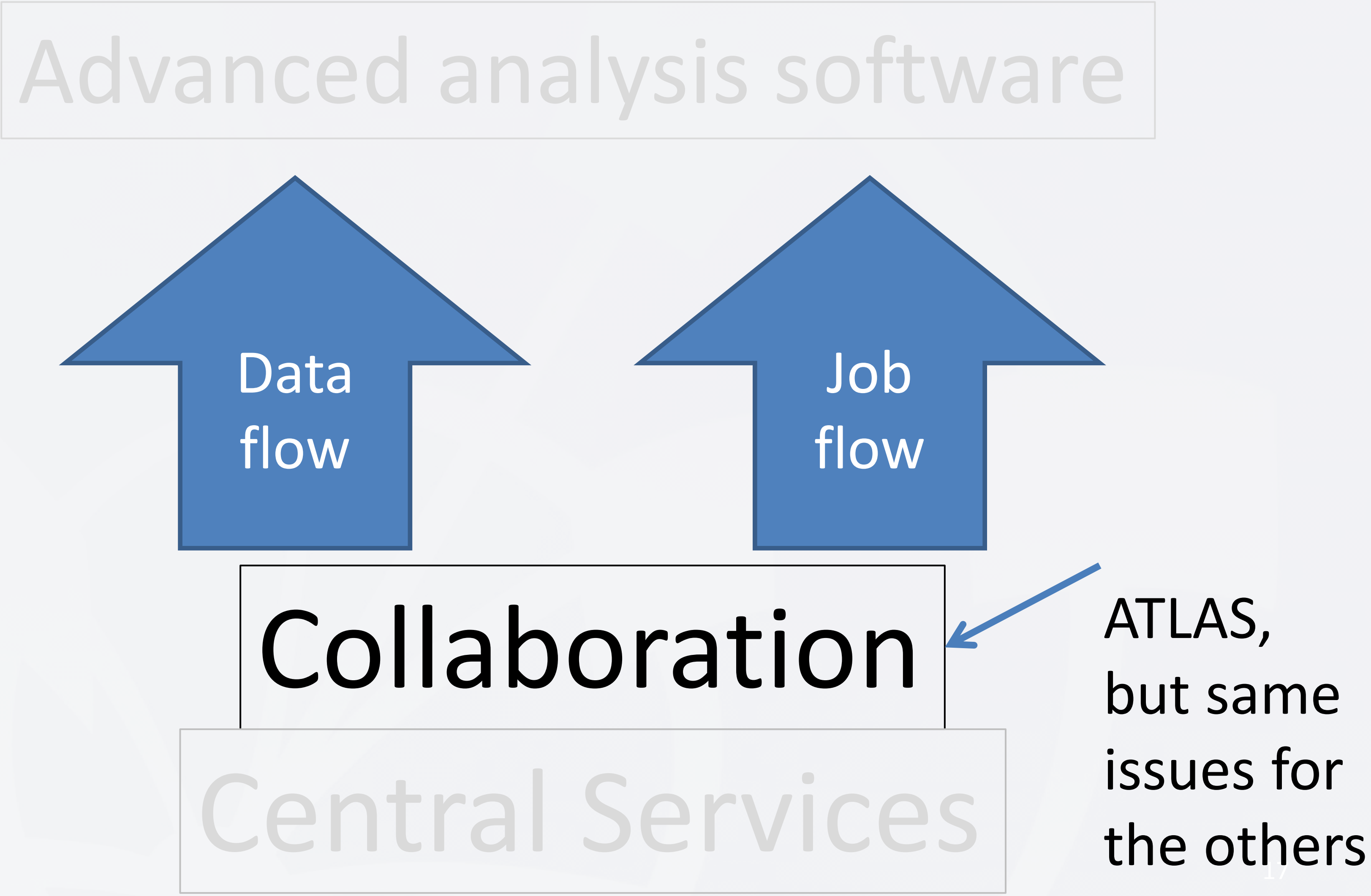
CASTOR architecture



NS: NameServer
TMd: TransferManagerd
DMd:DiskManagerd
TGd: TapeGatewayd
VMGR:tape catalogue
VDQM: drive scheduler

(credit: L.Mascetti)

Overview



Advanced analysis software

Data flow

Job flow

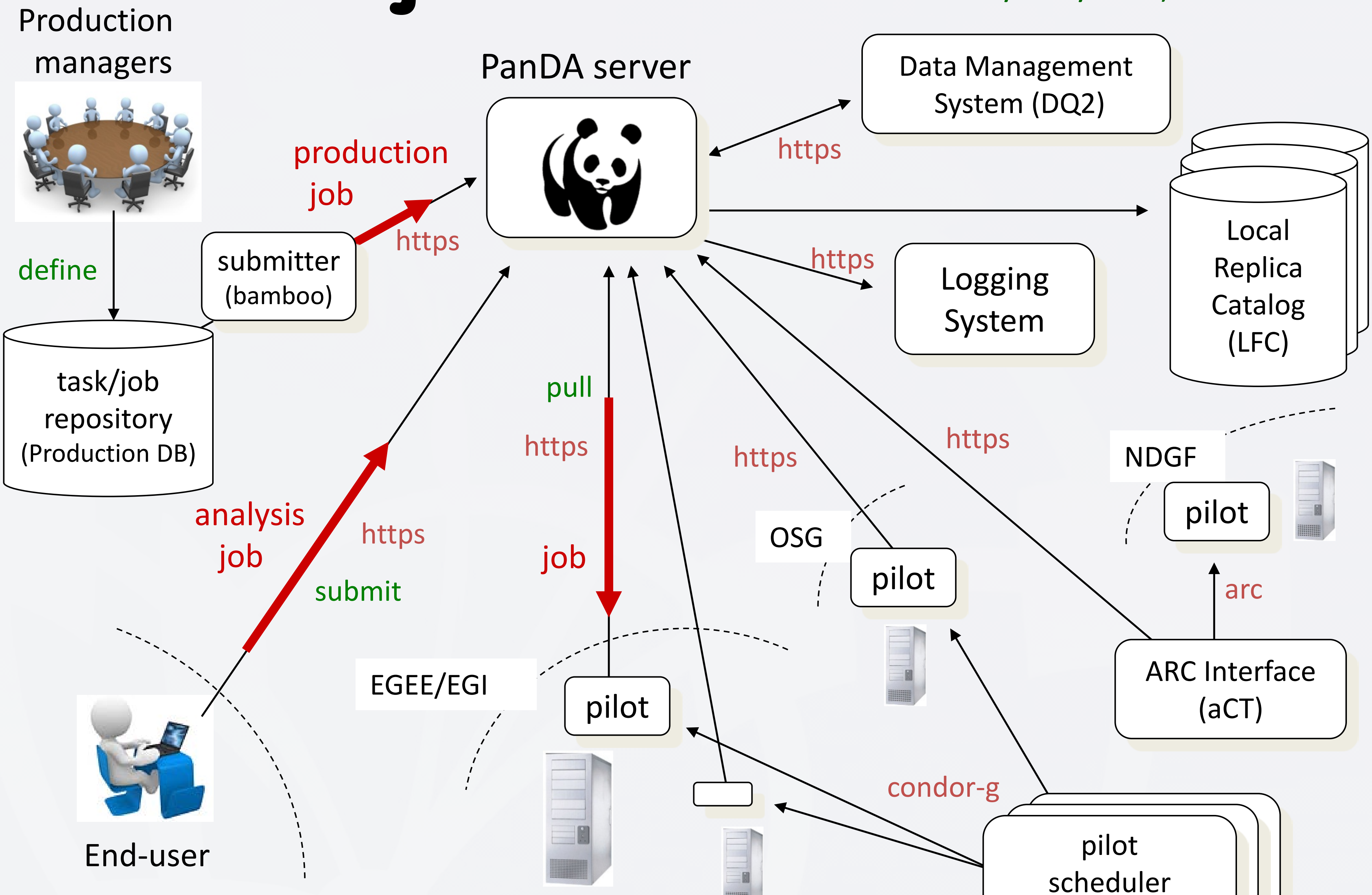
Collaboration

Central Services

ATLAS, but same issues for the others

ATLAS job control

(Production and Distributed Analysis System)



Production managers

define

task/job repository (Production DB)

submitter (bamboo)

production job

https

PanDA server

https

Data Management System (DQ2)

Local Replica Catalog (LFC)

https

Logging System

pull

https

job

https

https

NDGF

pilot

arc

ARC Interface (aCT)

OSG

pilot

analysis job

https

submit

End-user

EGEE/EGI

pilot

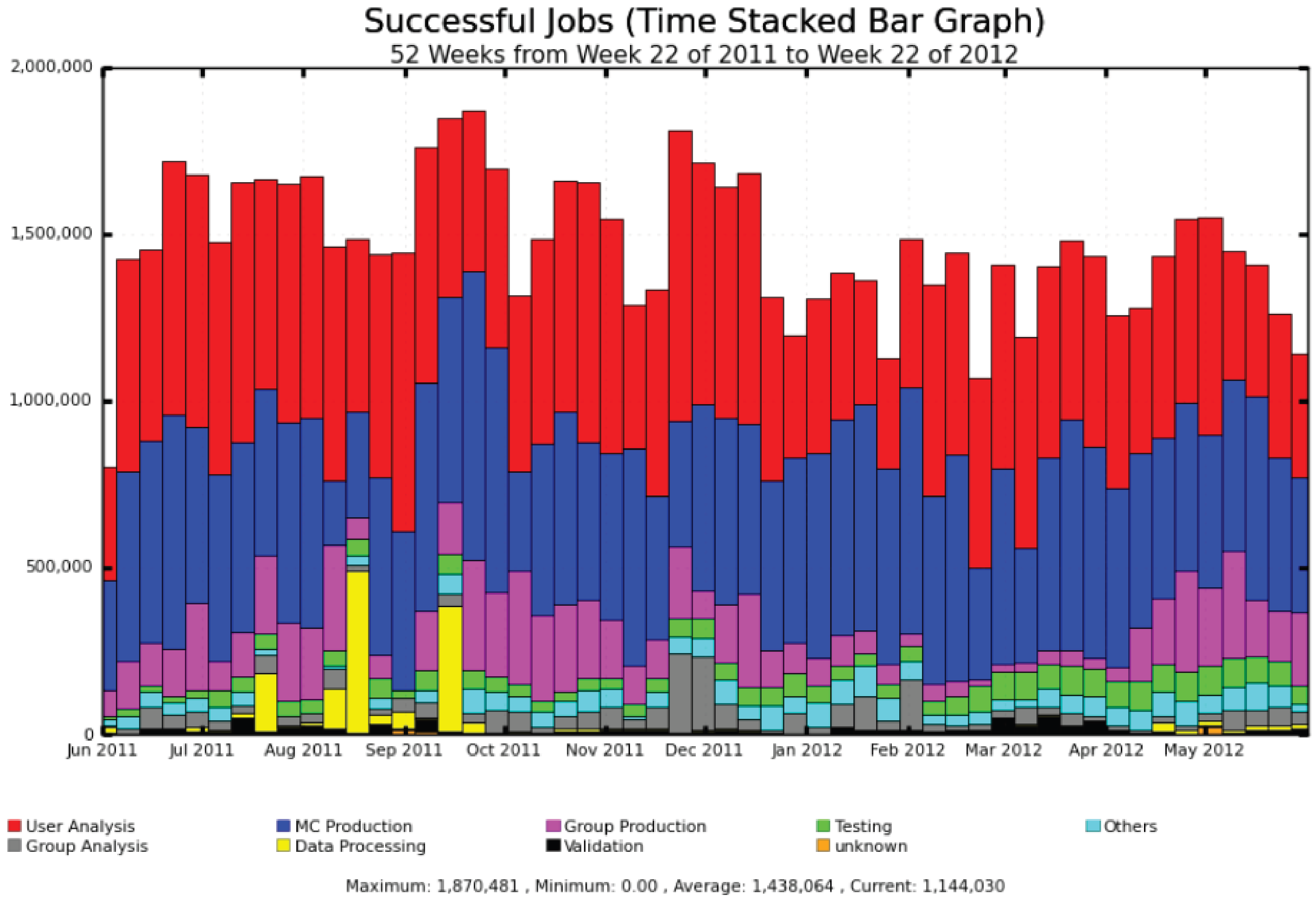
condor-g

pilot scheduler (autopyfactory)

Worker Nodes

Credit: T.Maeno

ATLAS User Analysis



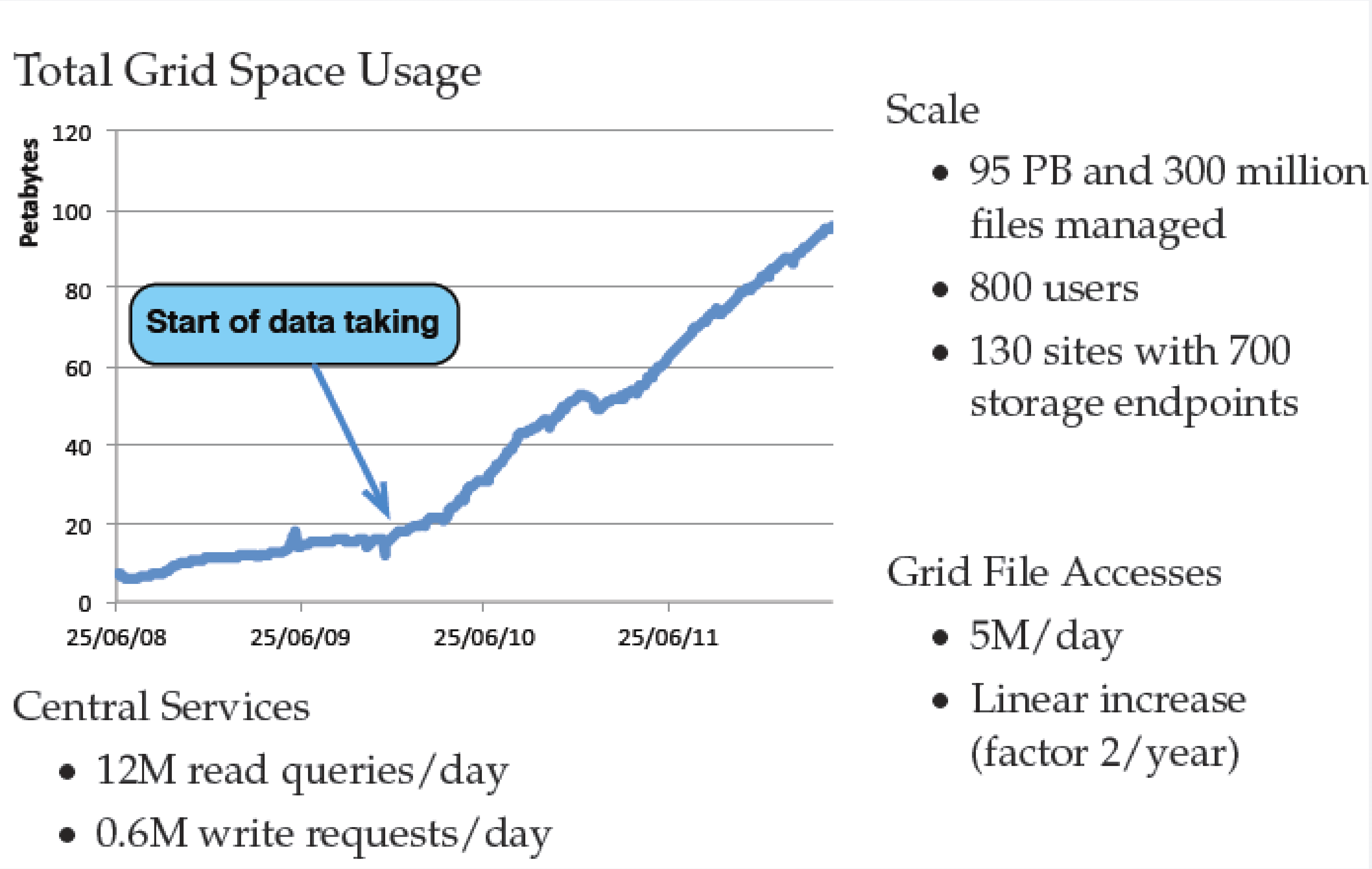
4

ATLAS Distributed Data Management System (DQ2)

- The ATLAS Distributed Data Management project is charged with managing ATLAS data on the grid
- All for the purpose of helping the collaboration store, manage and process LHC data in a heterogeneous distributed environment
-
- Requirements:
- Catalog data
- Transfer data to/from sites
- Delete data from sites
- Ensure data consistency at sites
- Enforce ATLAS computing model requirements

Credit: V.Garonne

ATLAS DQ2 statistics



Total Grid Space Usage

Scale

- 95 PB and 300 million files managed
- 800 users
- 130 sites with 700 storage endpoints

Central Services

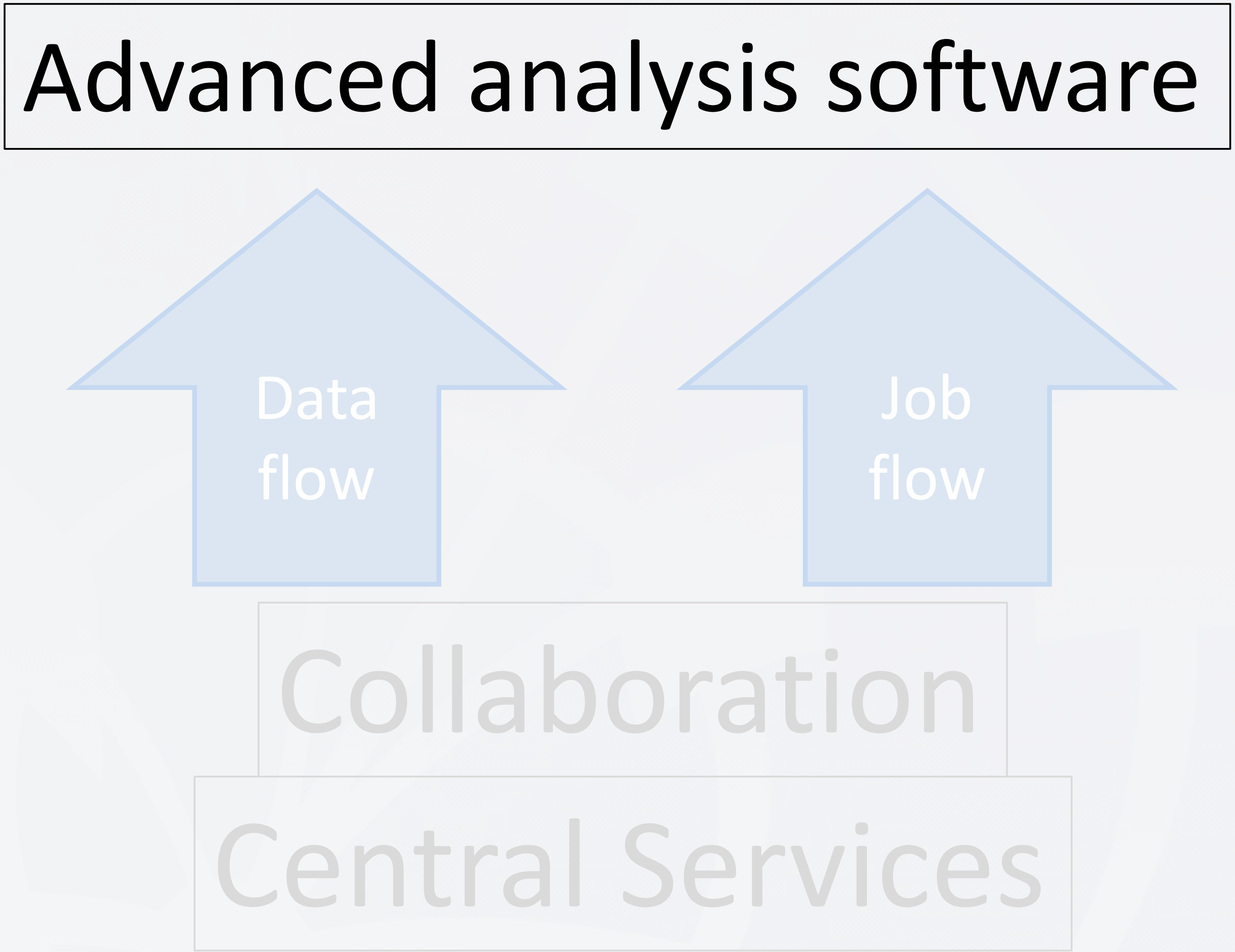
- 12M read queries/day
- 0.6M write requests/day

Grid File Accesses

- 5M/day
- Linear increase (factor 2/year)

Credit: V.Garonne

Overview



Advanced analysis software

Data flow

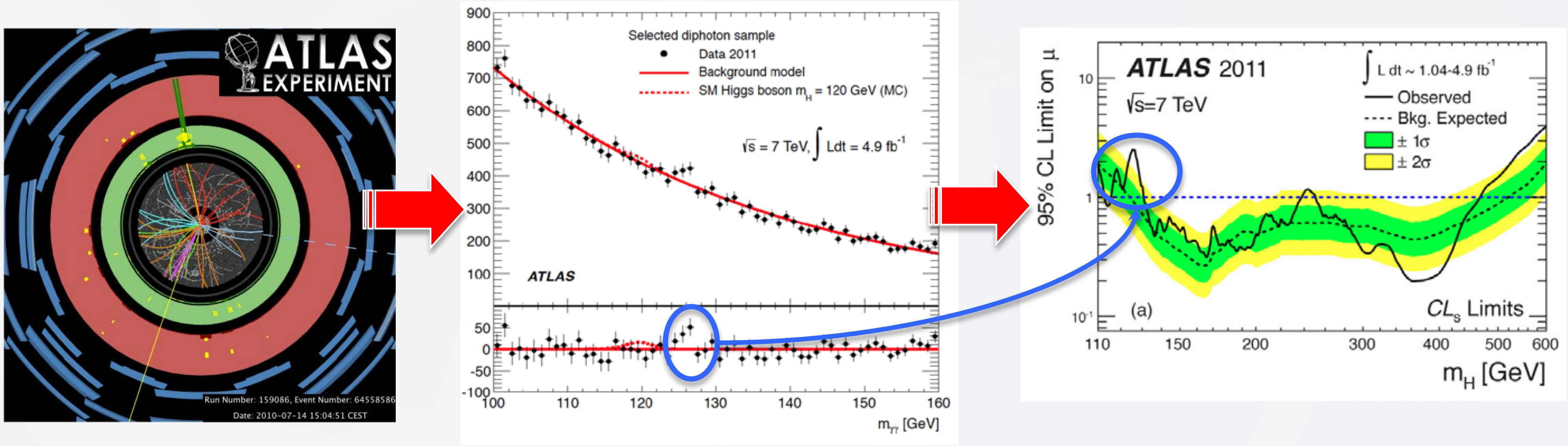
Job flow

Collaboration

Central Services

Data Analysis

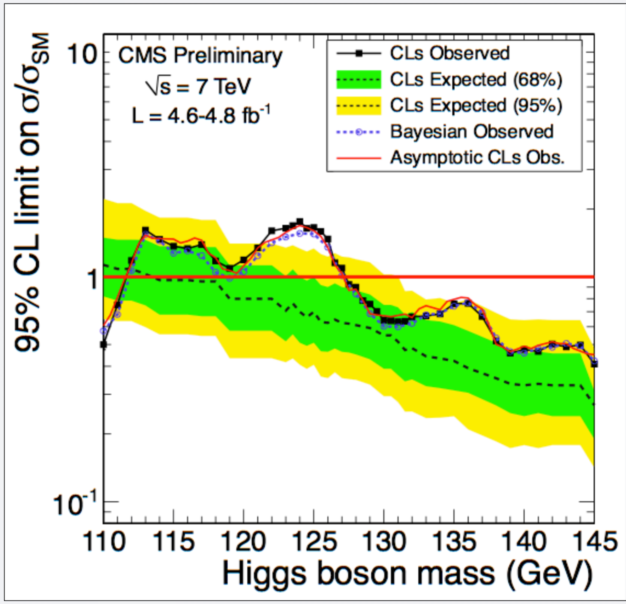
- **Huge quantity of data collected, but most of events are simply reflecting well-know physics processes**
  - New physics effects expected in a tiny fraction of the total events: few tens
- **Crucial to have a good discrimination between interesting events and the rest, i.e. different species**
  - Data analysis techniques play a crucial role in this "fight"

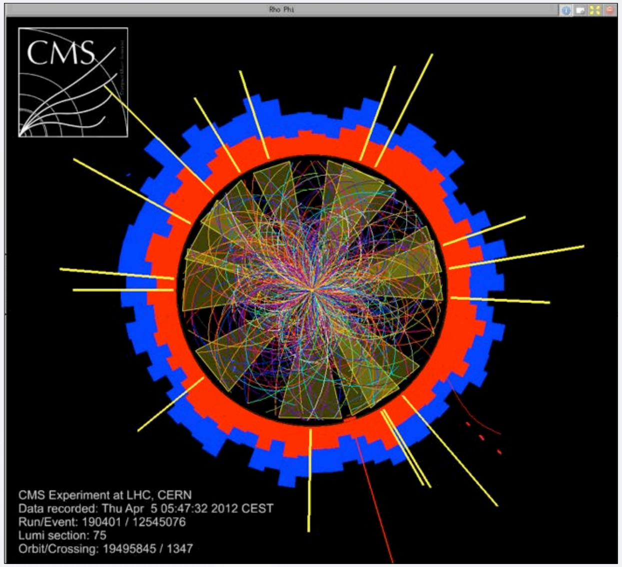


Credit: A.Lazzaro

ROOT Object-Oriented toolkit

- **Data Analysis toolkit**
  - Written in C++ (millions of lines)
  - Open source
  - Integrated interpreter
  - File formats
  - I/O handling, graphics, plotting, math, histogram binning, event display, geometric navigation
  - Powerful fitting (RooFit) and statistical (RooStats) packages on top
  - In use by all HEP experiments

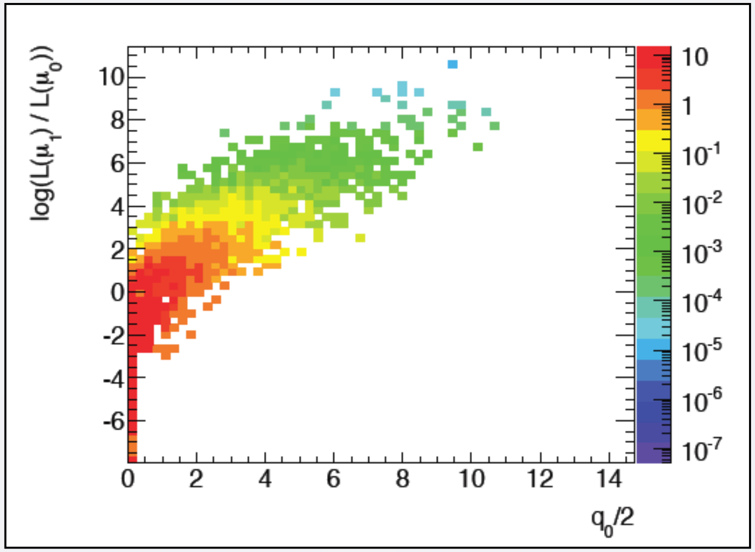


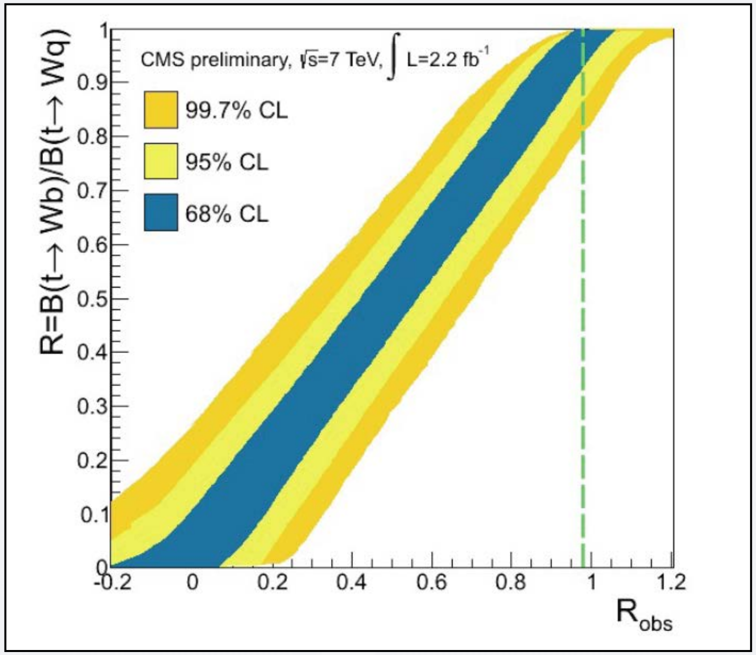

Rho Phi
CMS
CMS Experiment at LHC, CERN
Data recorded: Thu Apr 5 05:47:32 2012 CEST
Run/Event: 190401 / 12545076
Lumi section: 75
Orbit/Crossing: 19495845 / 1347

Credit: F.Rademakers

RooFit/RooStats

- Standard tool for producing physics results at LHC
  - Parameter estimation (fitting)
  - Interval estimation (e.g limit results for new particle searches)
  - Discovery significance (quantifying excess of events)
- Implementation of several statistical methods (Bayesian, Frequentist, Asymptotic)
- New tools for model creation and combinations
  - Histfactory: make RooFit models (RooWorkspace) from input histograms

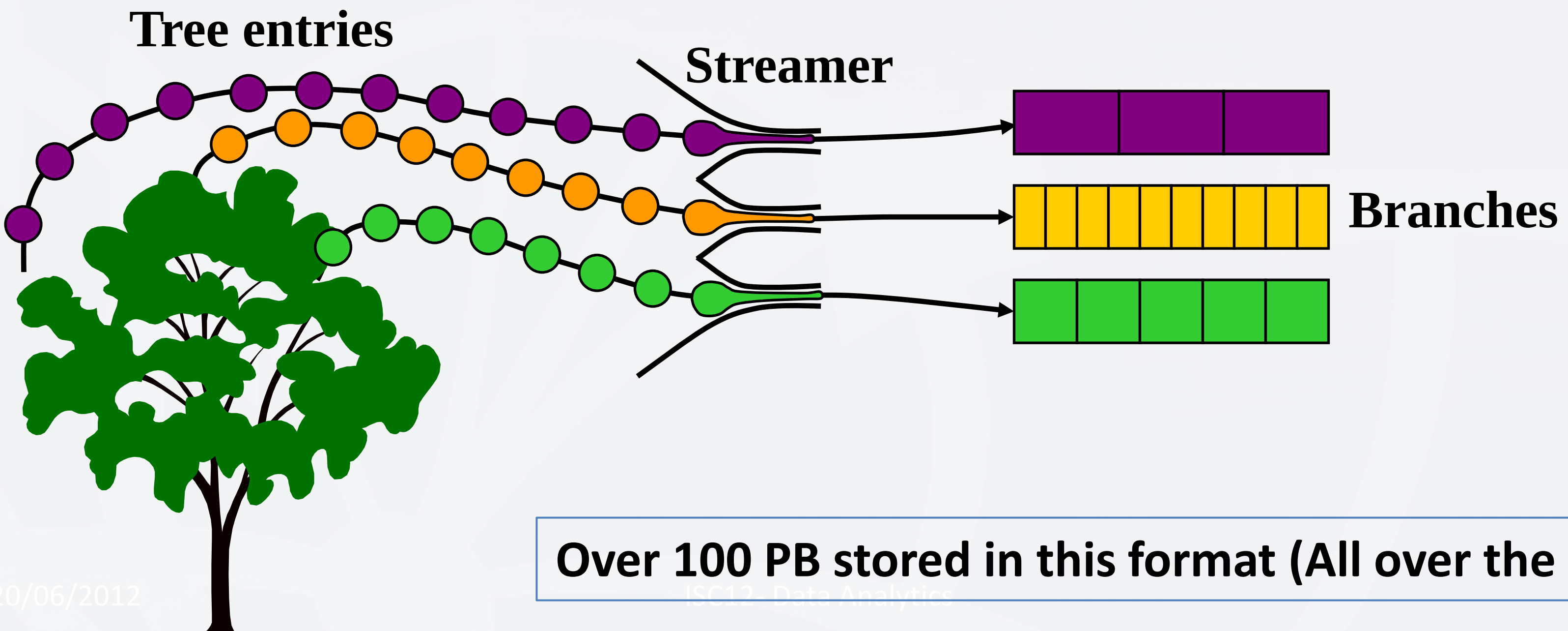




Credit: L.Moneta

ROOT files

- **Default format for all HEP data**
- **Organised as Trees with Branches**
  - Sophisticated formatting for optimal analysis of data
    - Parallelism, prefetching and caching
    - Compression, splitting and merging



Over 100 PB stored in this format (All over the world)

Conclusions

- **Big Data Analytics requires a solid organisational structure at all levels**
- **Must avoid "Big Headaches"**
  - Enormous files sizes and/or enormous file counts
  - Data movement, placement, access pattern, life cycle
  - Replicas, Backup copies, etc.
- **Big Data also implies Big Transactions/Transaction rates**
- **The LHC community started preparing more than a decade before real physics data arrived**
  - Now, the situation is well under control
  - But, data rates will continue to increase (dramatically) for years to come

There is no time to rest!

THANK YOU

Q & A

CERN openlab

References

- http://www.cern.ch/
- http://wlcg.web.cern.ch/
- http://root.cern.ch/
- http://eos.cern.ch/
- http://castor.cern.ch/
- http://panda.cern.ch/
- http://www.atlas.ch/

Backup