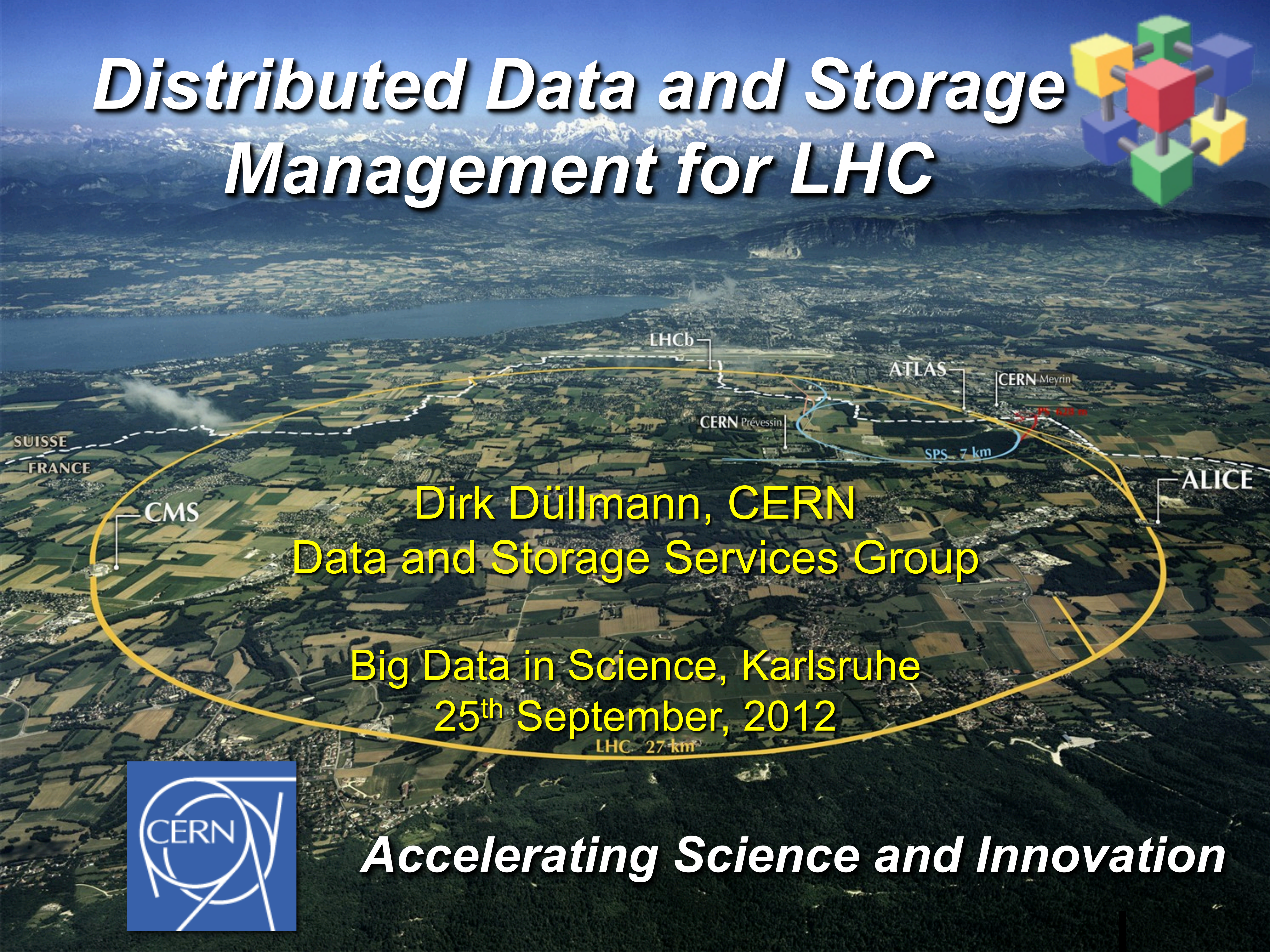
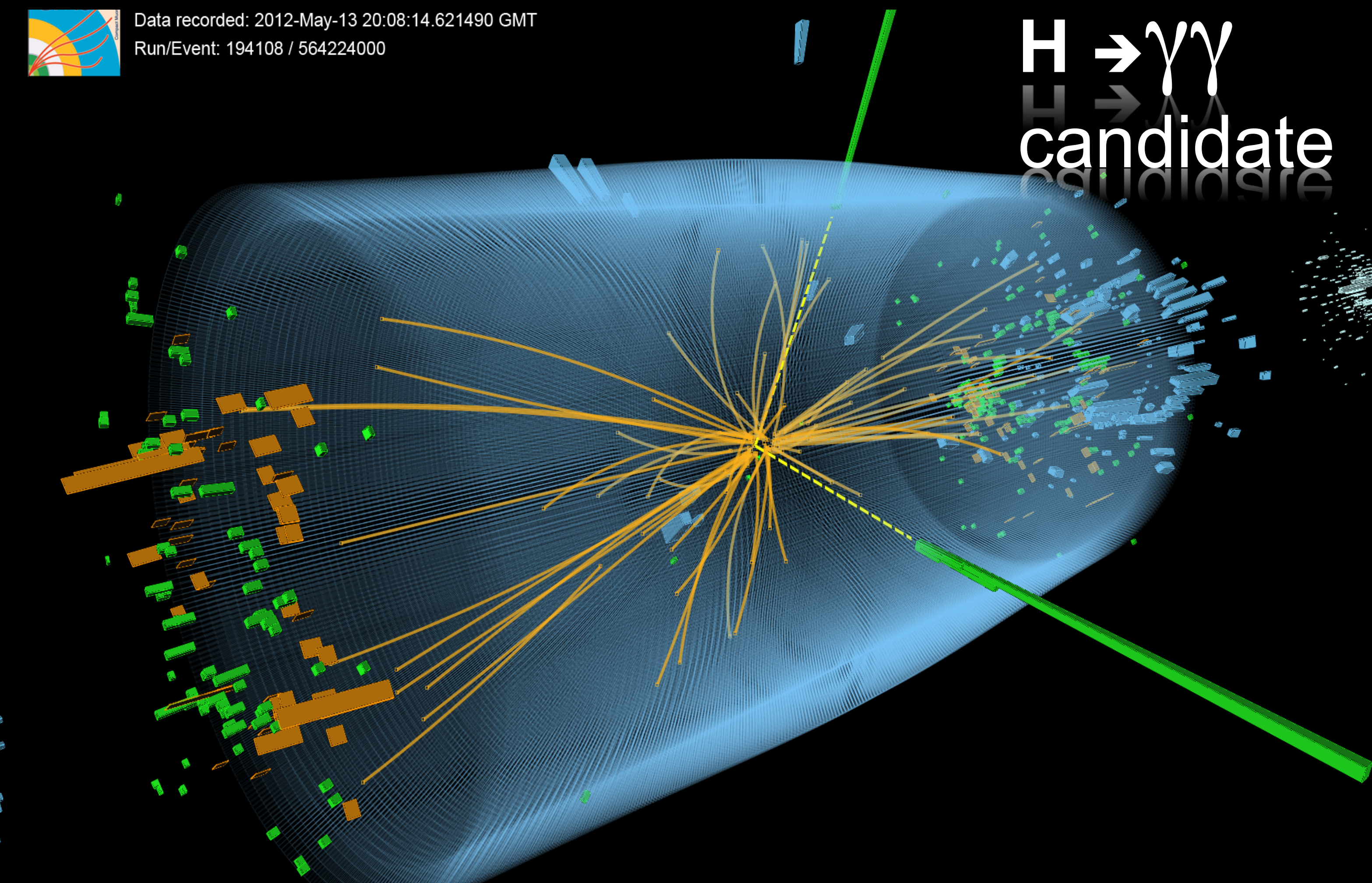
Distributed Data and Storage Management for LHC

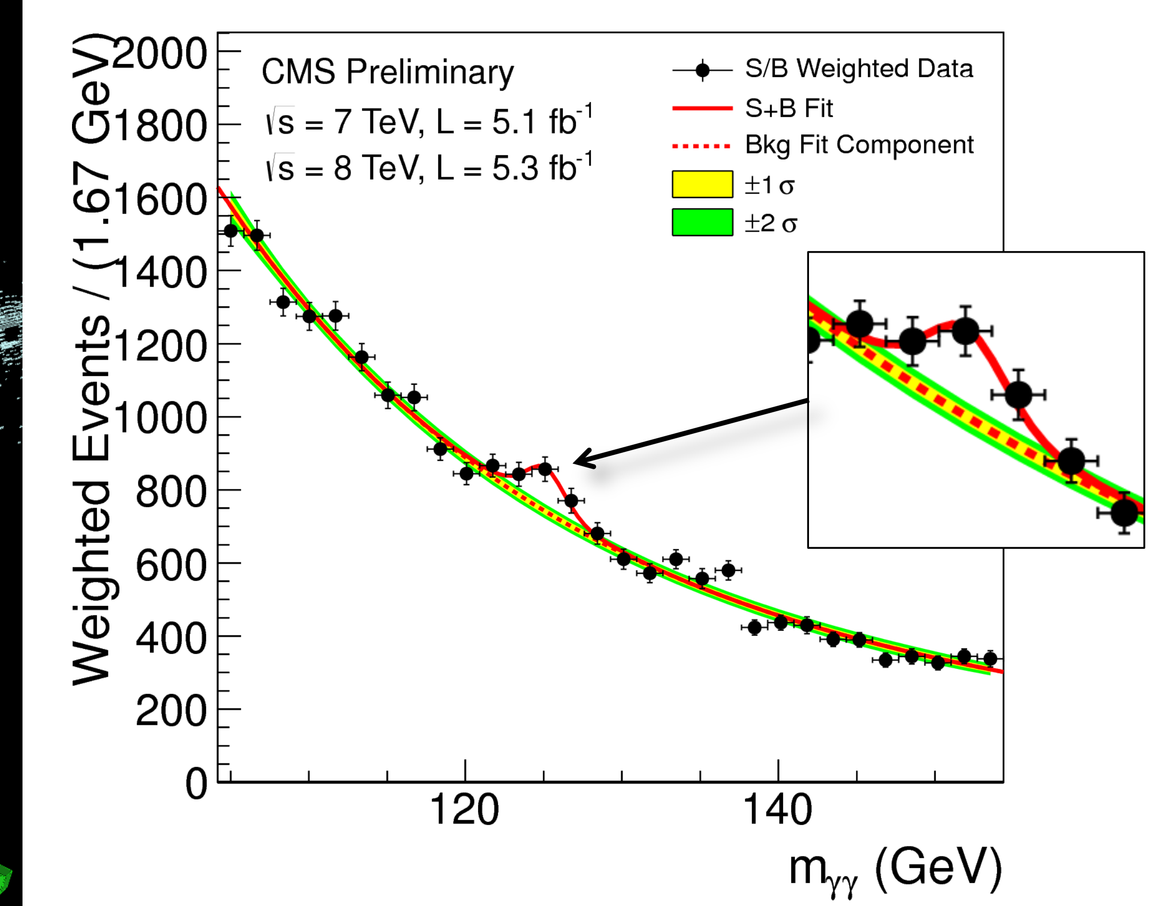


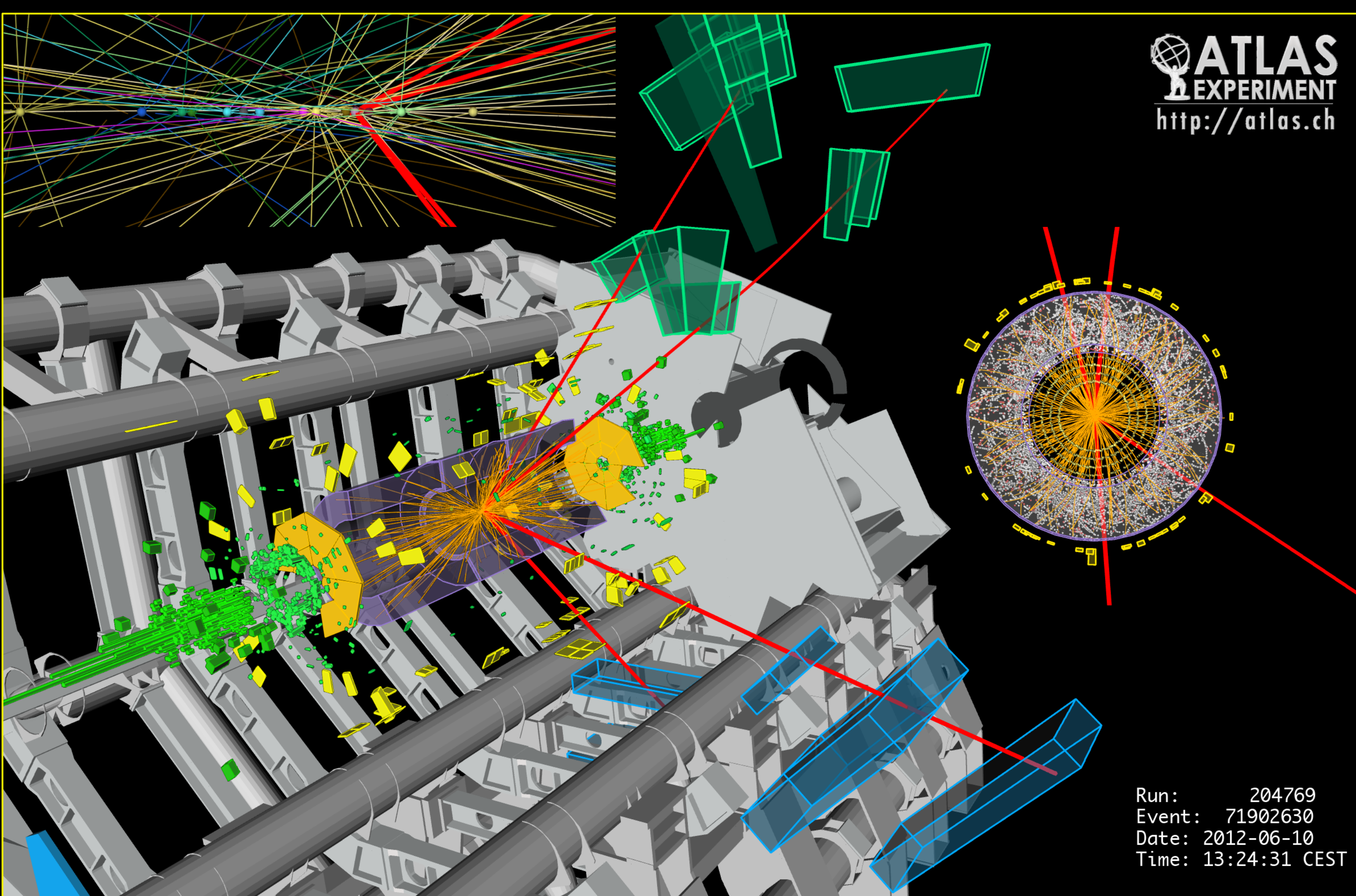
Dirk Düllmann, CERN
Data and Storage Services Group

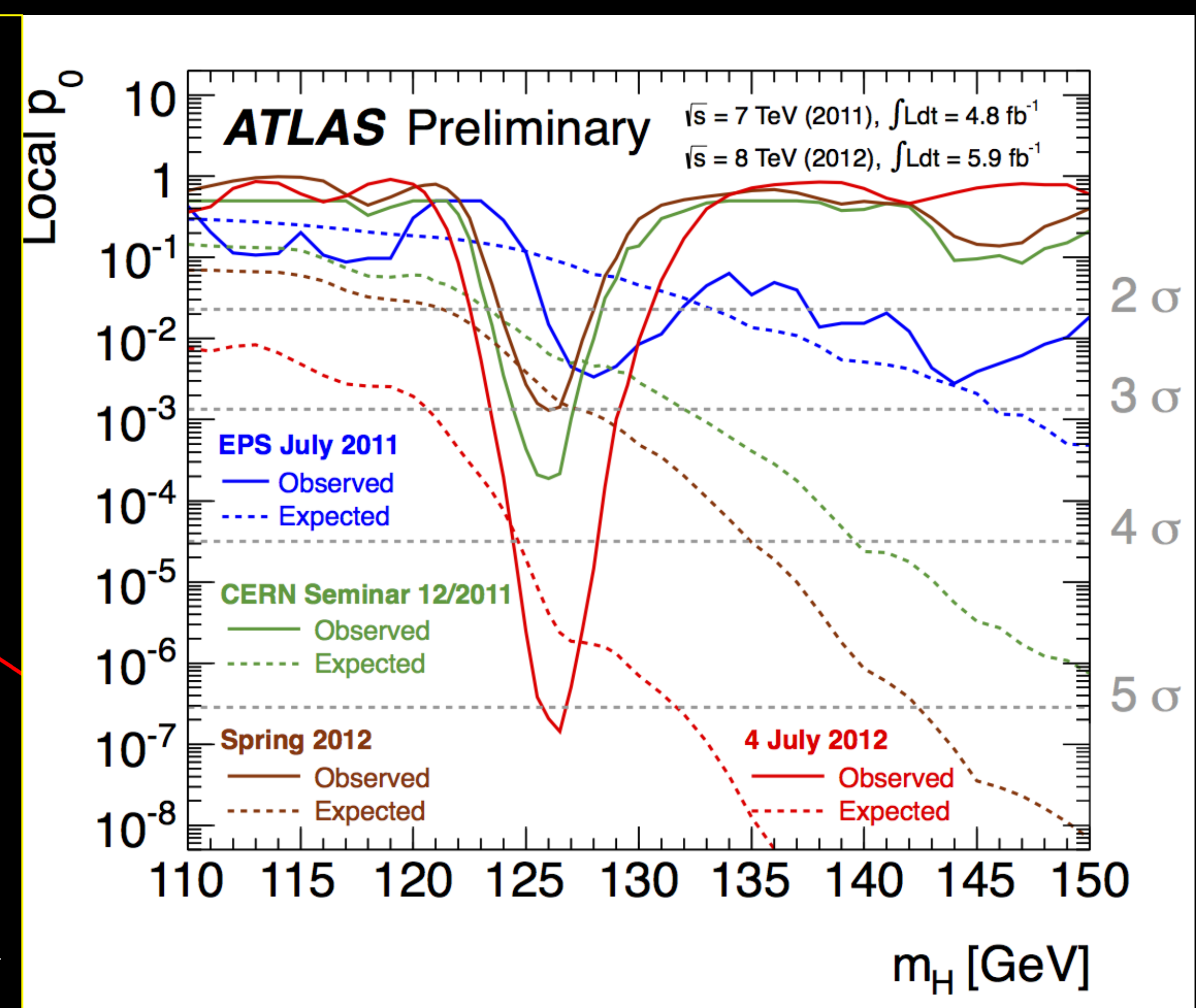
Big Data in Science, Karlsruhe
25th September, 2012

Accelerating Science and Innovation


Data recorded: 2012-May-13 20:08:14.621490 GMT
Run/Event: 194108 / 564224000
H → γγ
candidate


CMS Preliminary
√s = 7 TeV, L = 5.1 fb⁻¹
√s = 8 TeV, L = 5.3 fb⁻¹
S/B Weighted Data
S+B Fit
Bkg Fit Component
±1 σ
±2 σ
Weighted Events / (1.67 GeV)
m_γγ (GeV)


ATLAS EXPERIMENT
http://atlas.ch
Run: 204769
Event: 71902630
Date: 2012-06-10
Time: 13:24:31 CEST


ATLAS Preliminary
√s = 7 TeV (2011), ∫Ldt = 4.8 fb⁻¹
√s = 8 TeV (2012), ∫Ldt = 5.9 fb⁻¹
Local p₀
EPS July 2011
Observed
Expected
CERN Seminar 12/2011
Observed
Expected
Spring 2012
Observed
Expected
4 July 2012
Observed
Expected
2 σ
3 σ
4 σ
5 σ
m_H [GeV]

Global Effort → Global Success

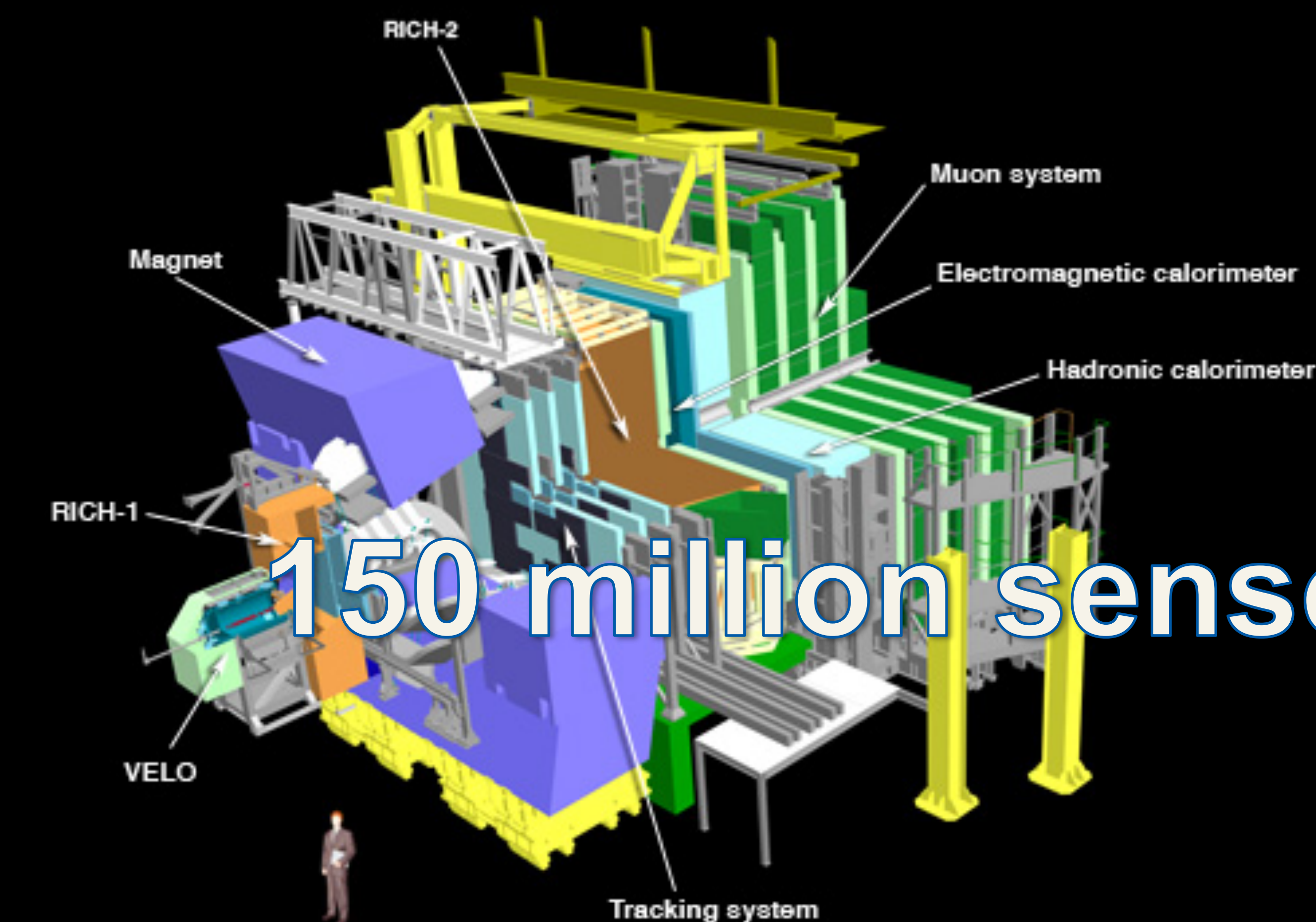
Results today only possible due to extraordinary performance of accelerators – experiments – Grid computing

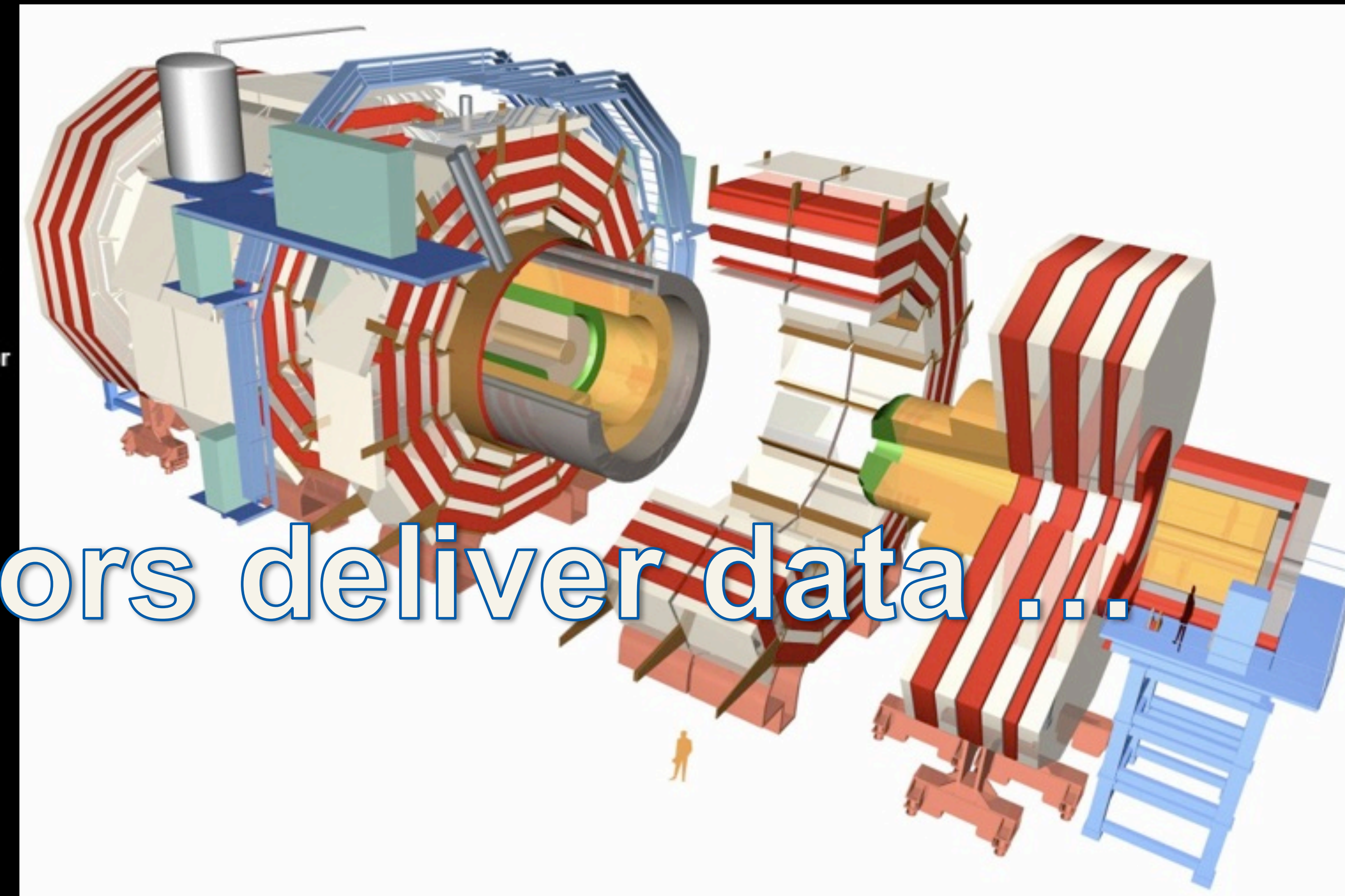
Observation of a new particle consistent with a Higgs Boson (but which one…?)

Historic Milestone but only the beginning

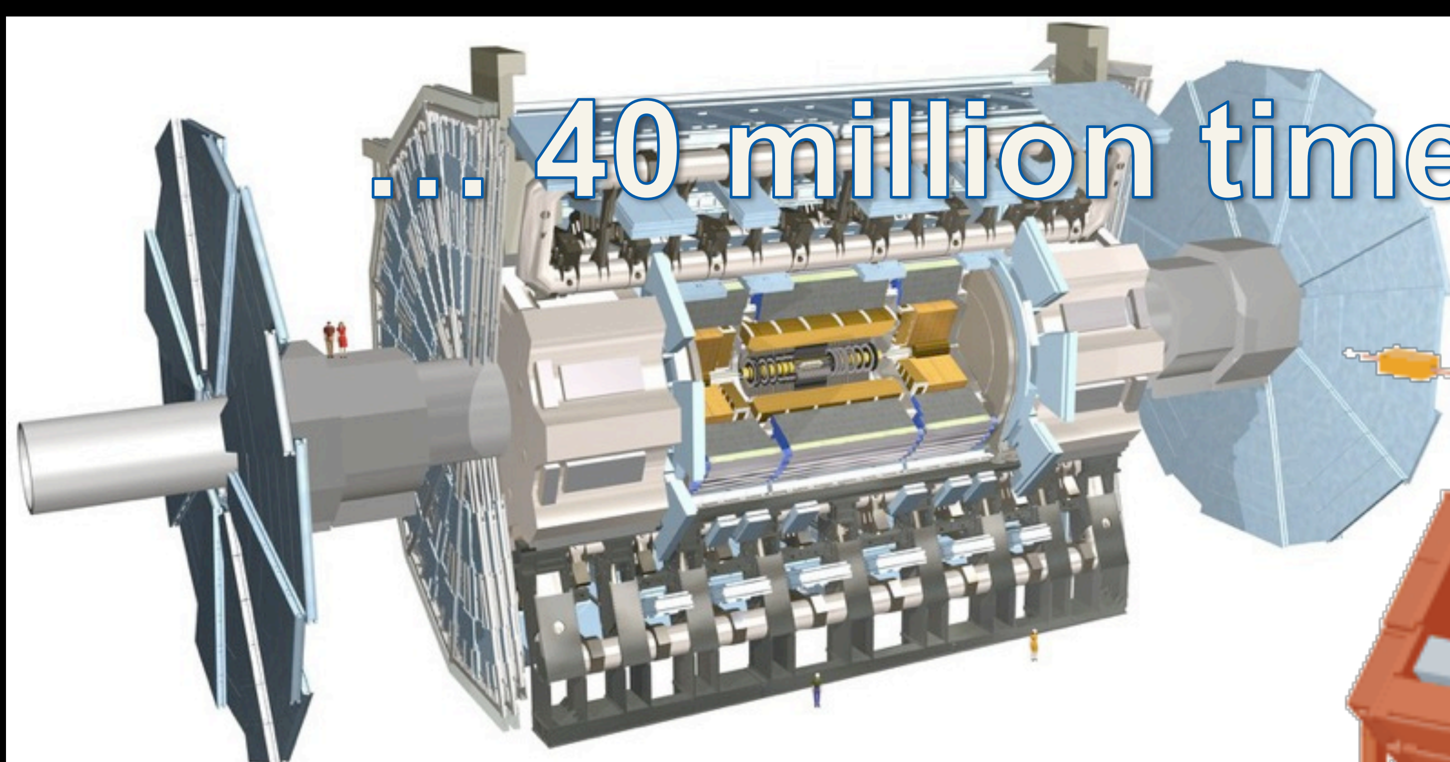
Global Implications for the future

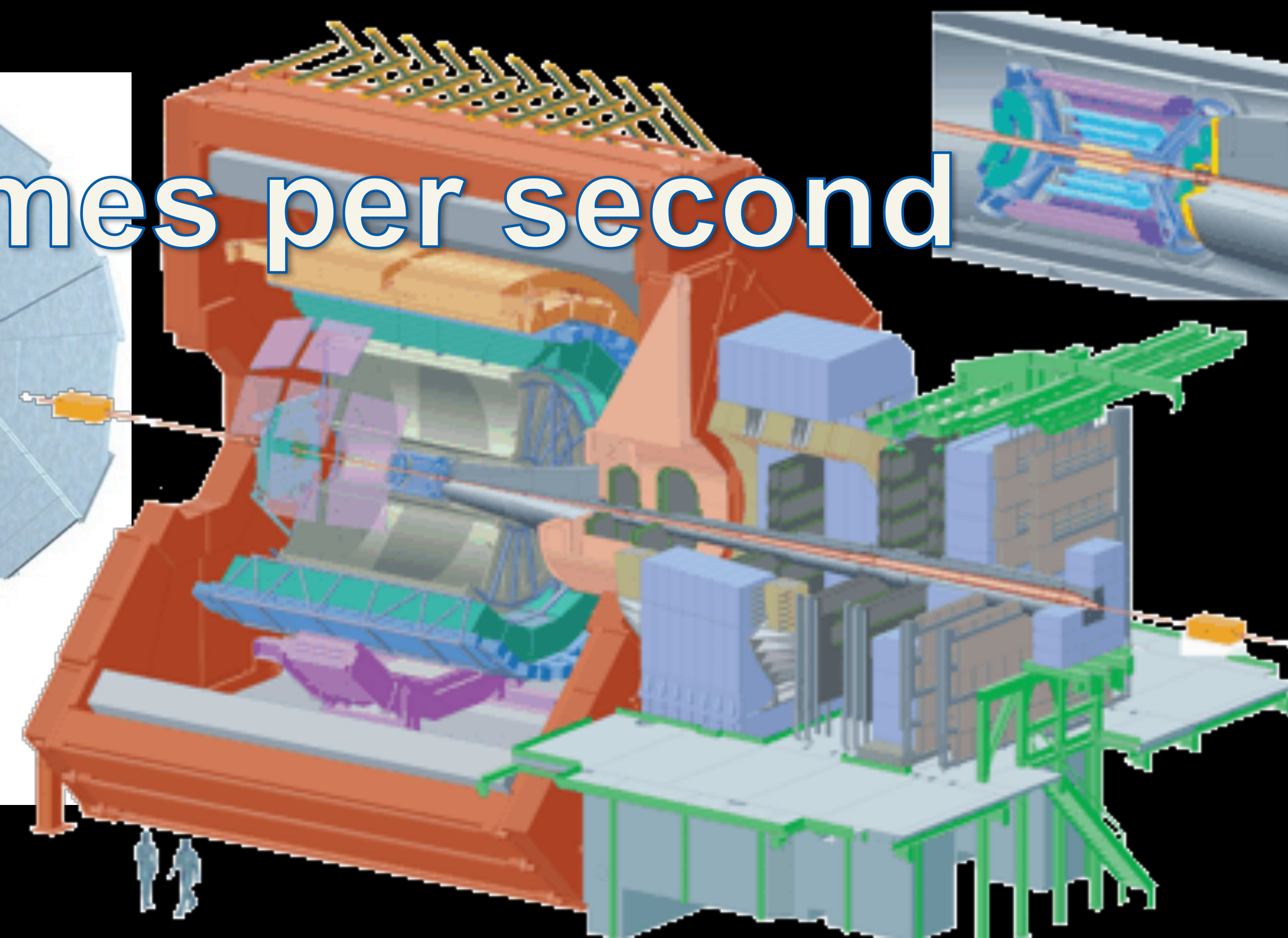
R-D Heuer





150 million sensors deliver data …

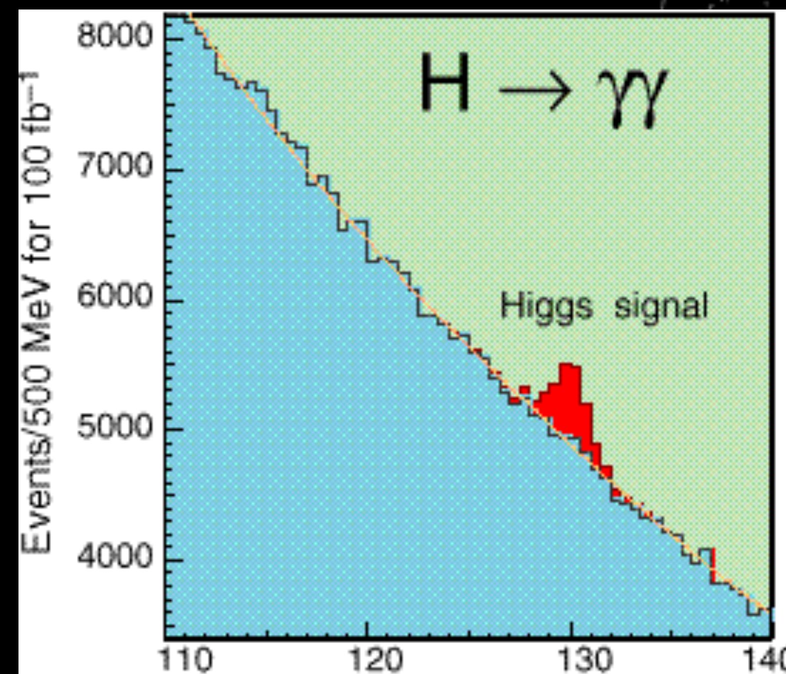


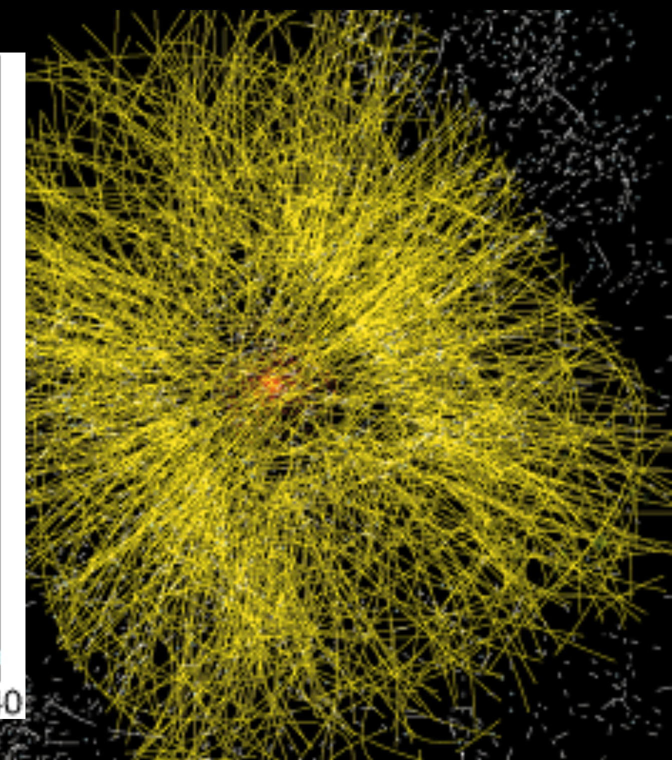


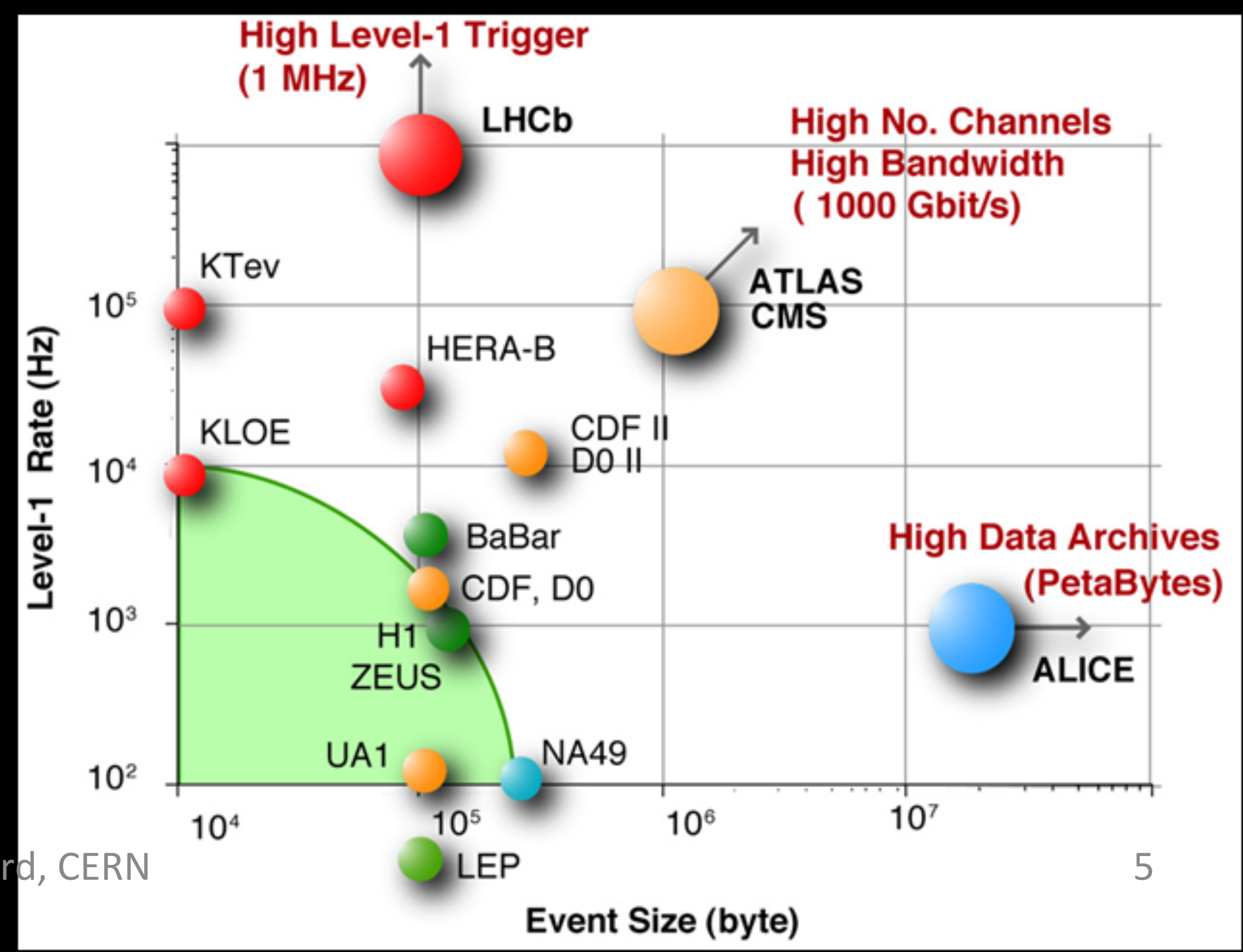
… 40 million times per second

The LHC Computing Challenge

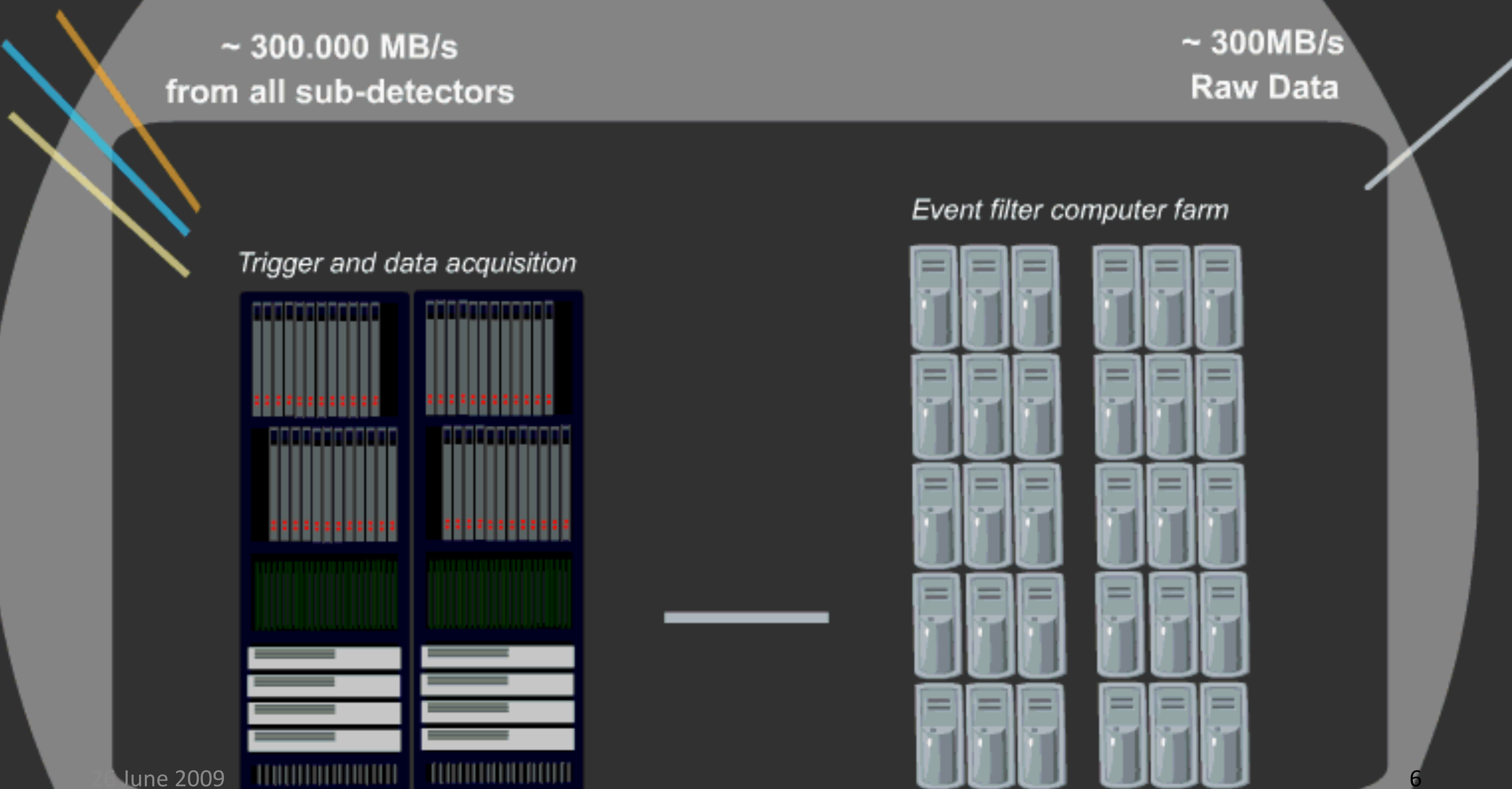
- Signal/Noise: 10^{-13} (10^{-9} offline)
- Data volume
  - High rate * large number of channels * 4 experiments
  - ➔ **15 PetaBytes of new data each year** → 22 PB in 2011
- Compute power
  - Event complexity * Nb. events * thousands users
  - ➔ **200 k CPUs** → 300 k CPU
  - ➔ **45 PB of disk storage** → 170 PB
- Worldwide analysis & funding
  - Computing funding locally in major regions & countries
  - Efficient analysis
  - ➔ **GRID technology**



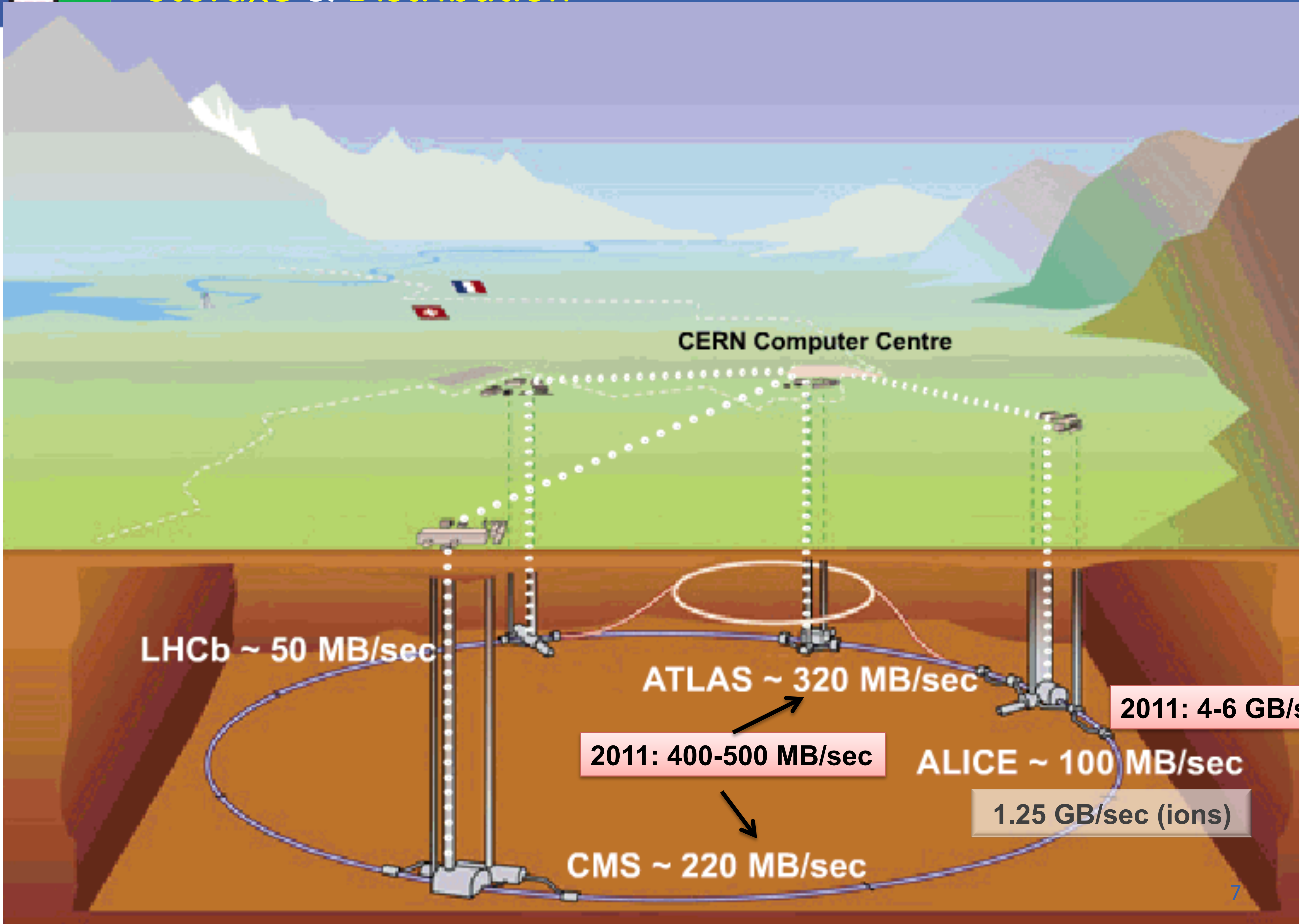




The Data Acquisition

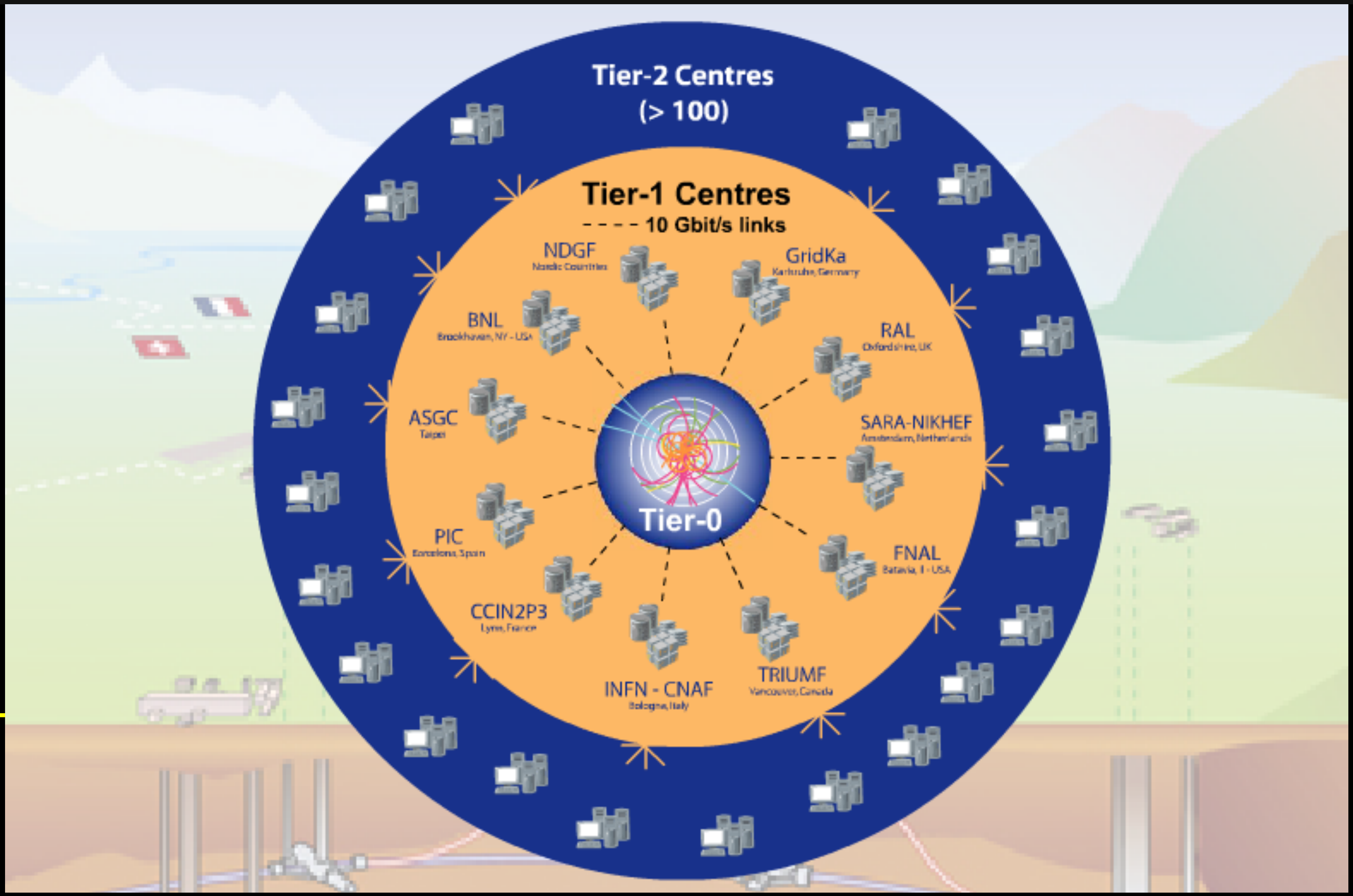


Tier 0 at CERN: Acquisition, First pass reconstruction, Storage & Distribution



WLCG – what and why?

- A distributed computing infrastructure to provide the production and analysis environments for the LHC experiments
- Managed and operated by a worldwide collaboration between the experiments and the participating computer centres
- The resources are distributed – for funding and sociological reasons
- Our task was to make use of the resources available to us – no matter where they are located



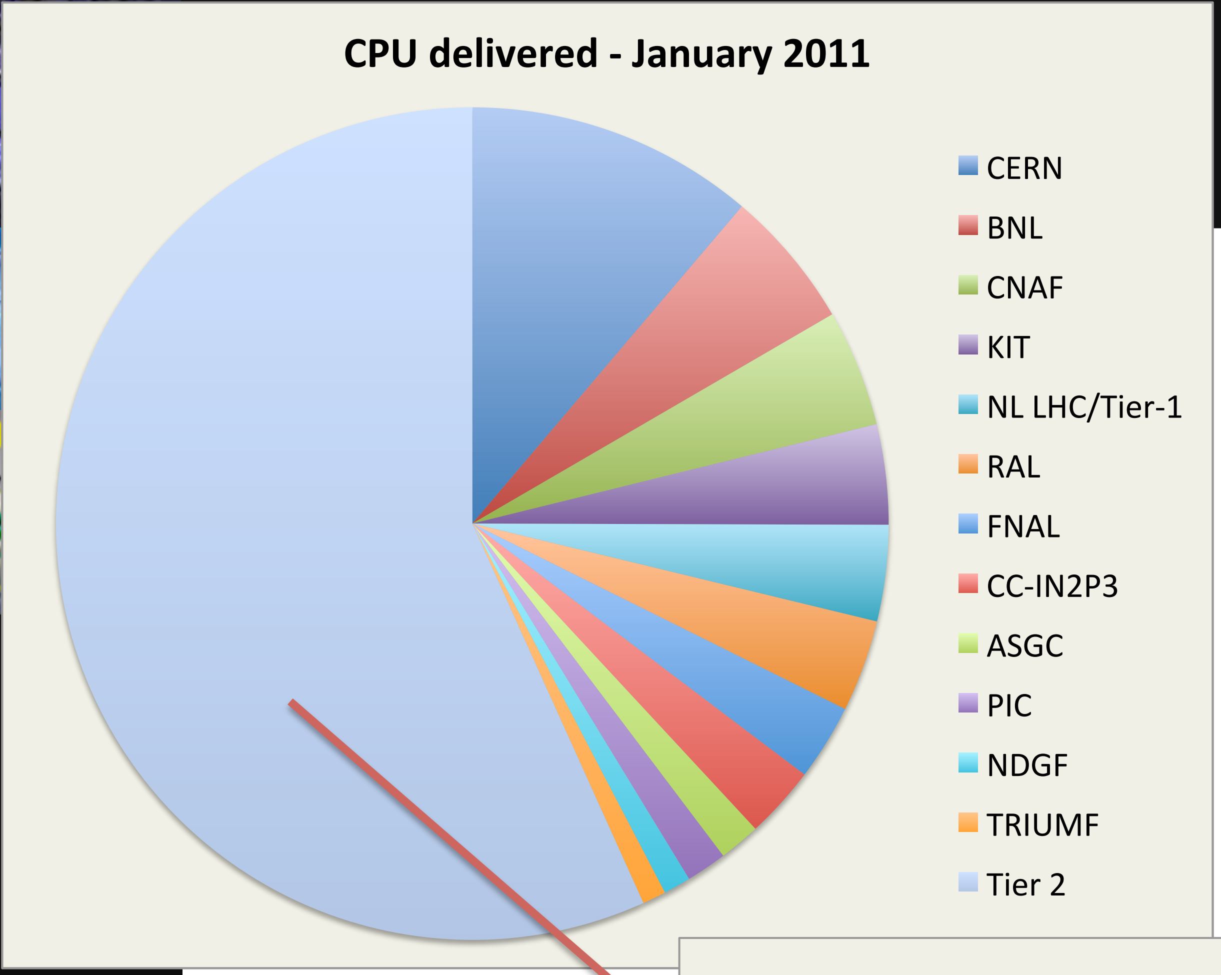
Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

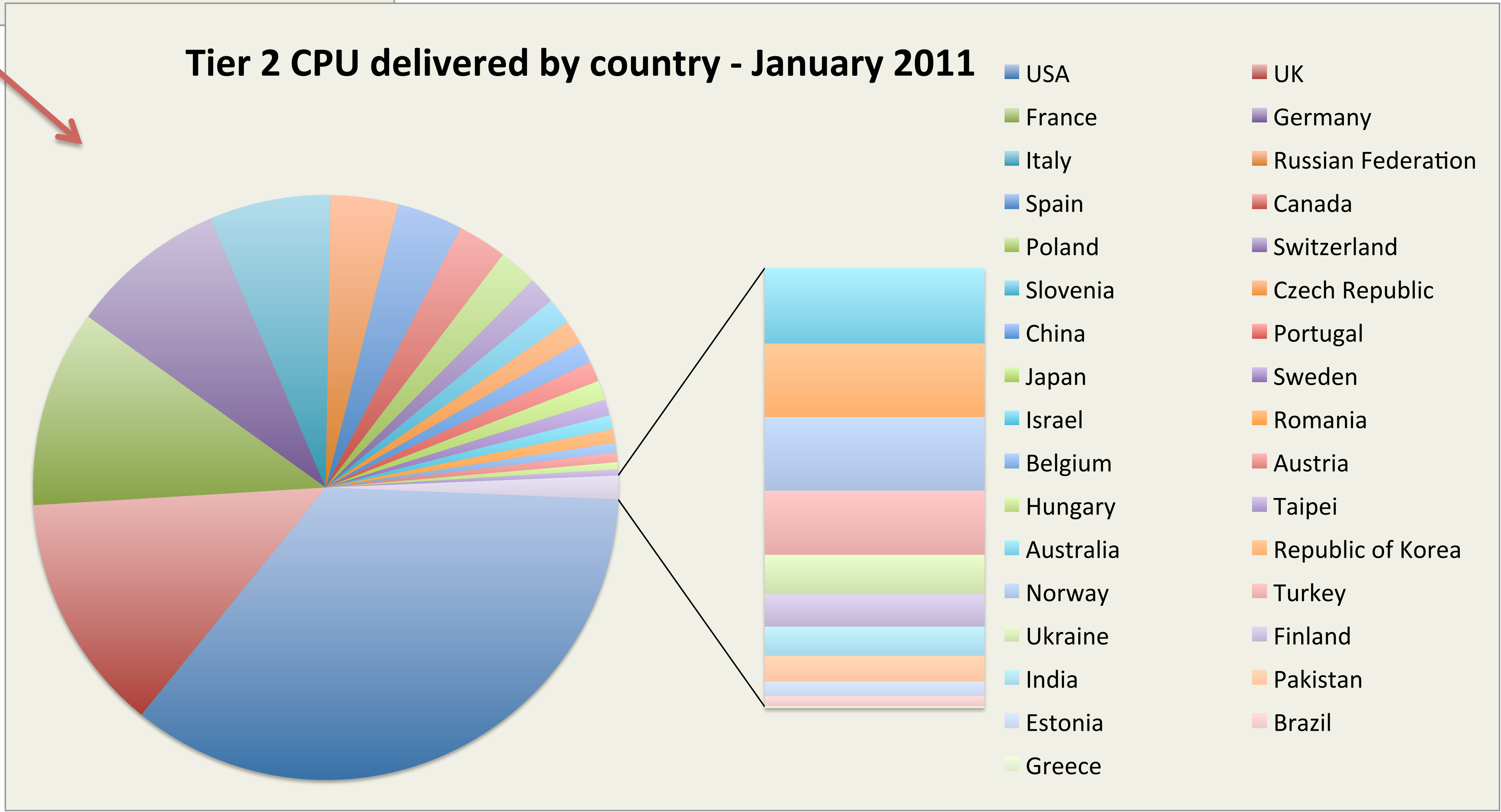
Tier-1 (11 centres):
- Permanent storage
- Re-processing
- Analysis

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

CPU – around the Tiers

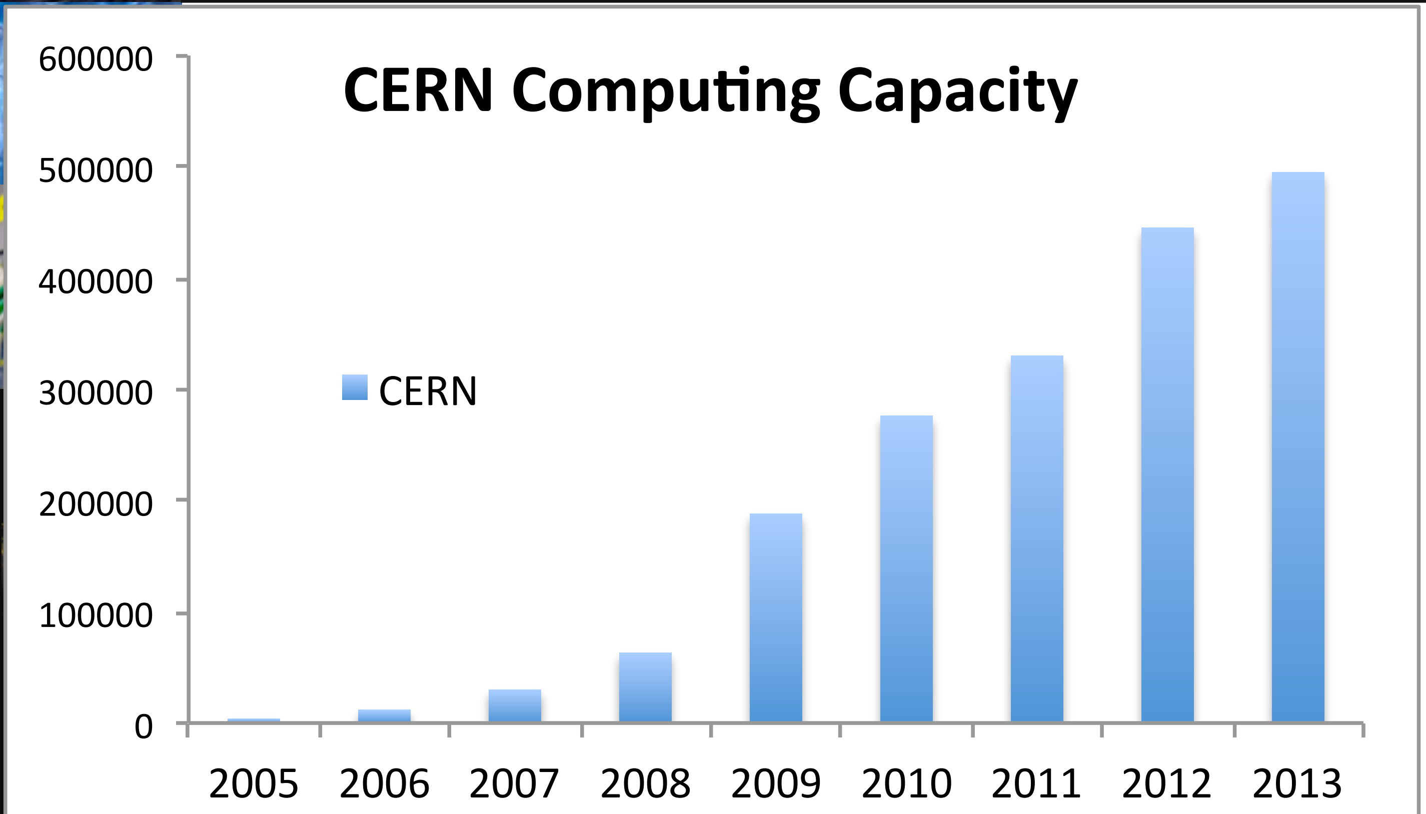
- The grid really works
- All sites, large and small can contribute
  - And their contributions are needed!



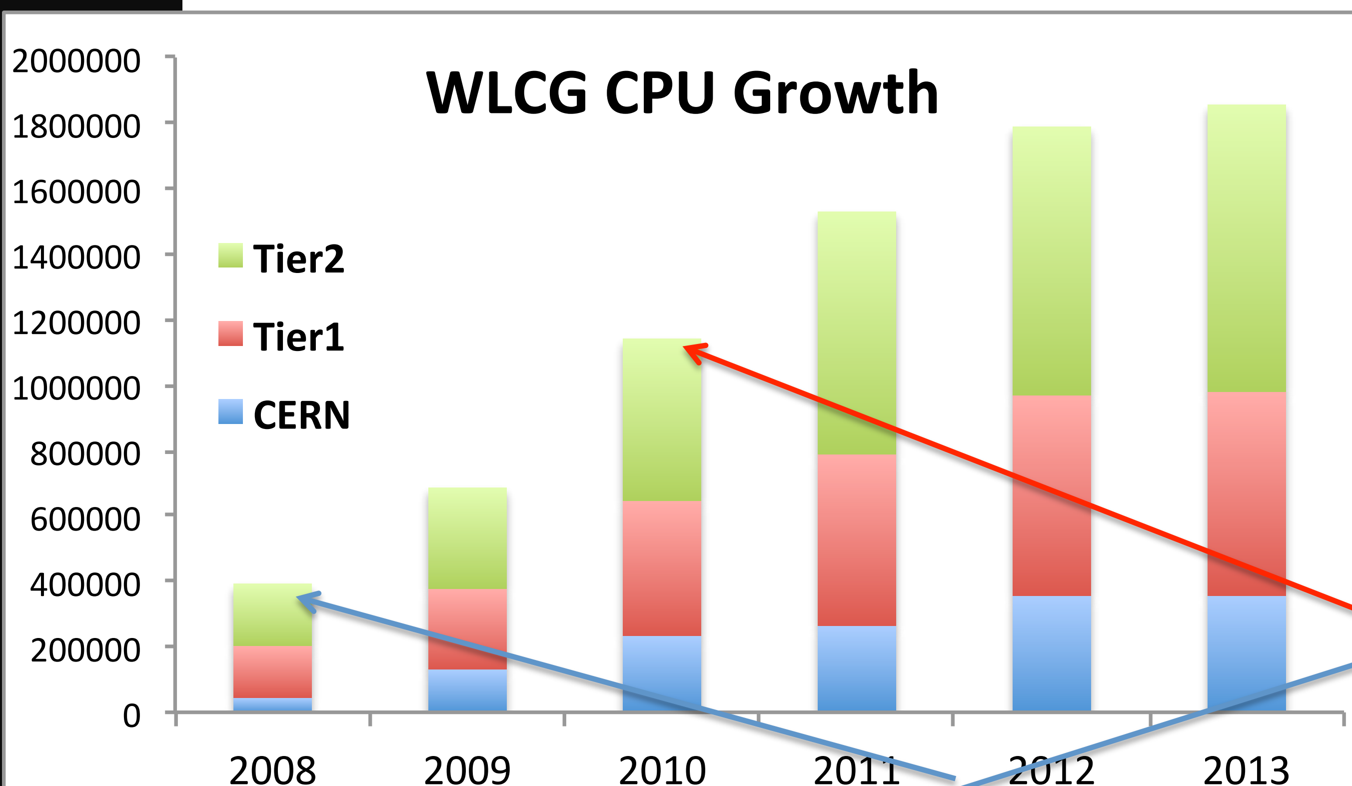


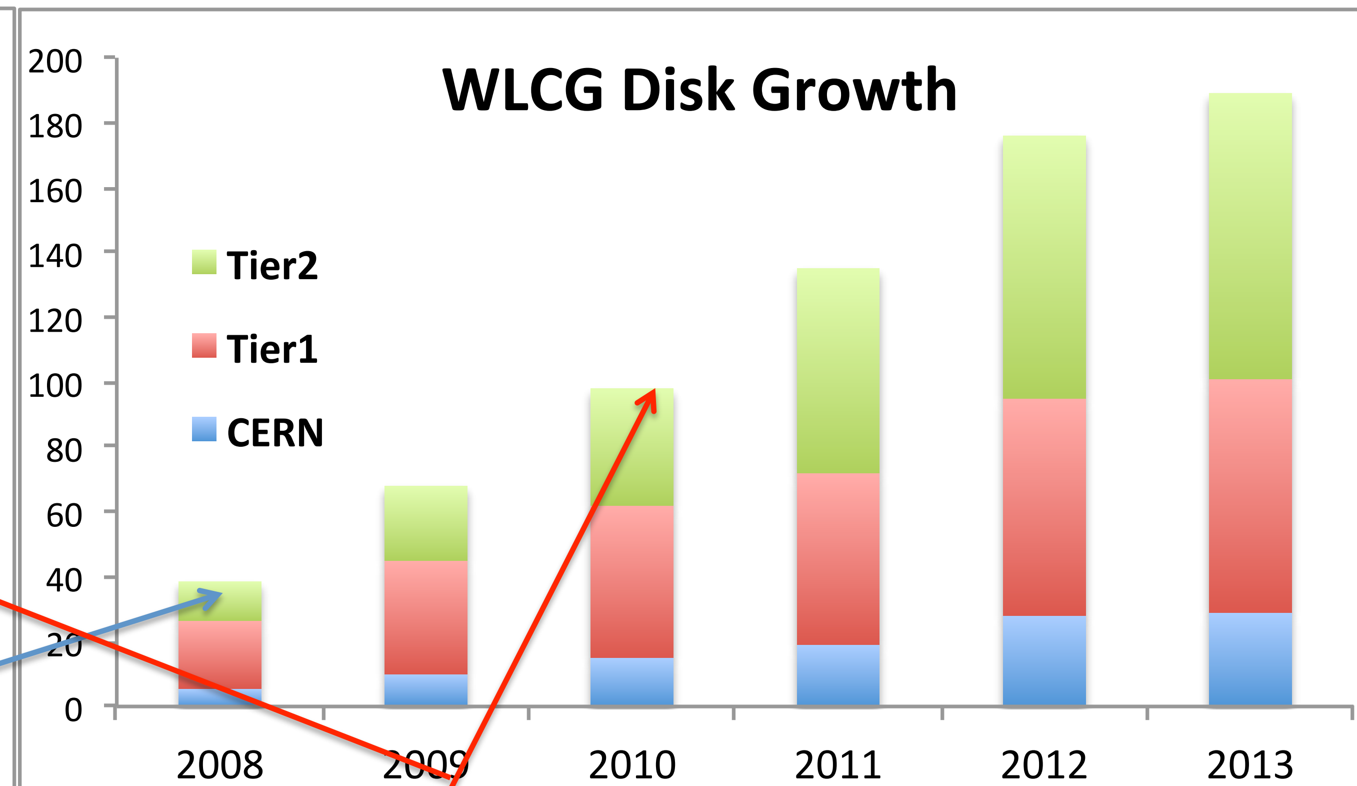
WLCG
Worldwide LHC Computing Grid

Evolution of capacity: CERN & WLCG



2013/14: modest increases to process "parked data"
2015 → budget limited ?
- experiments will push trigger rates
- flat budgets give ~20%/year growth

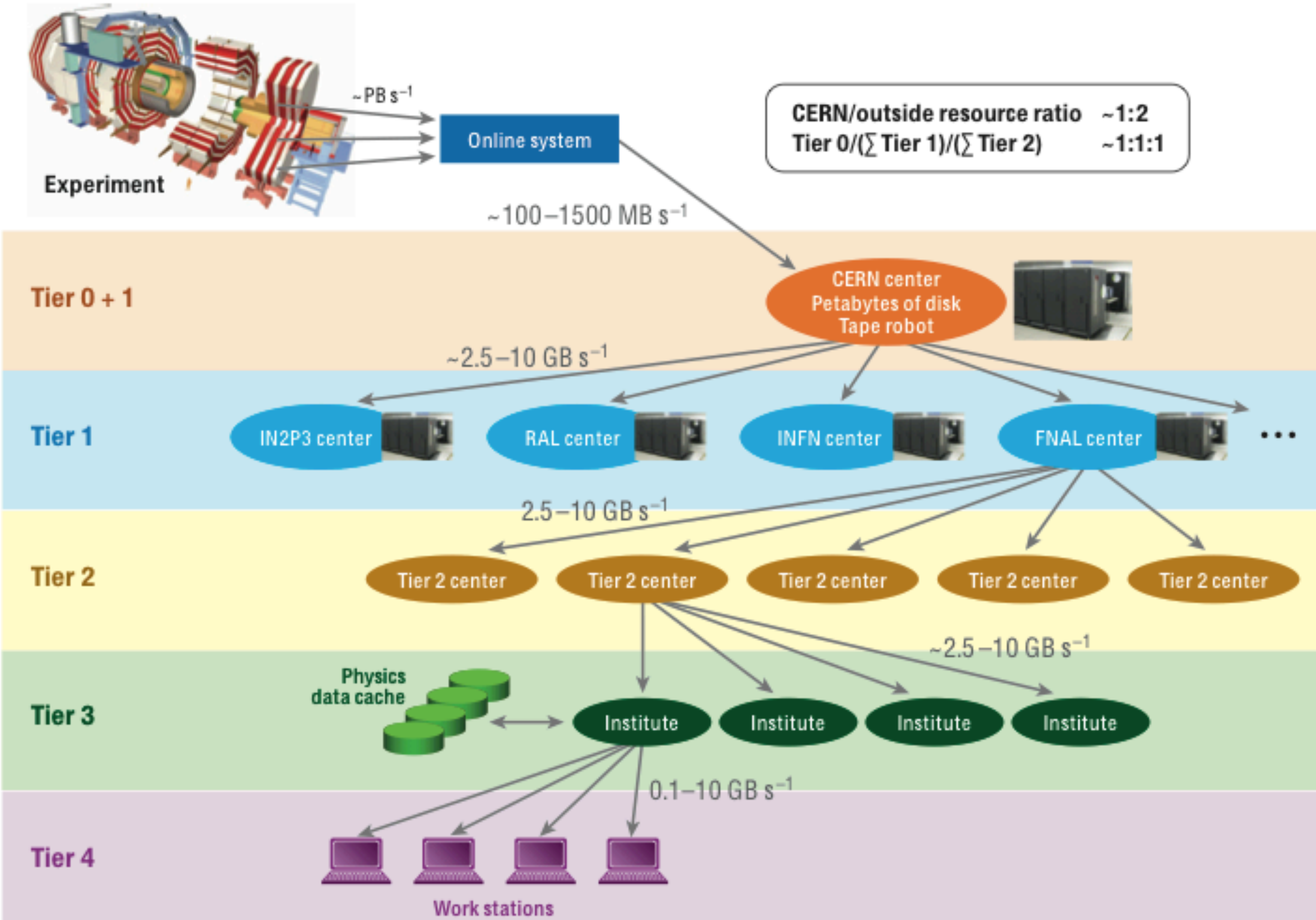




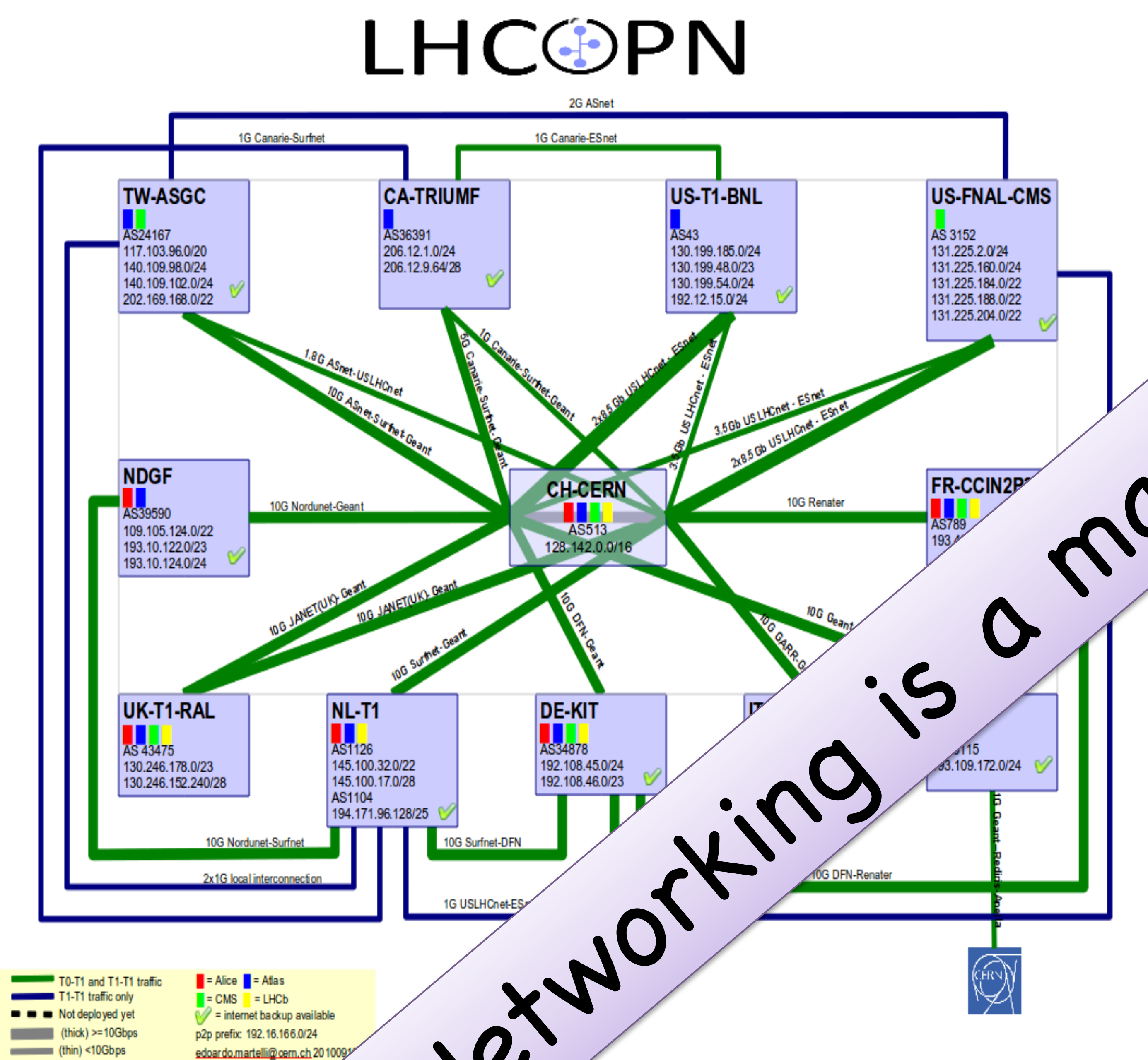
What we thought was needed at LHC start

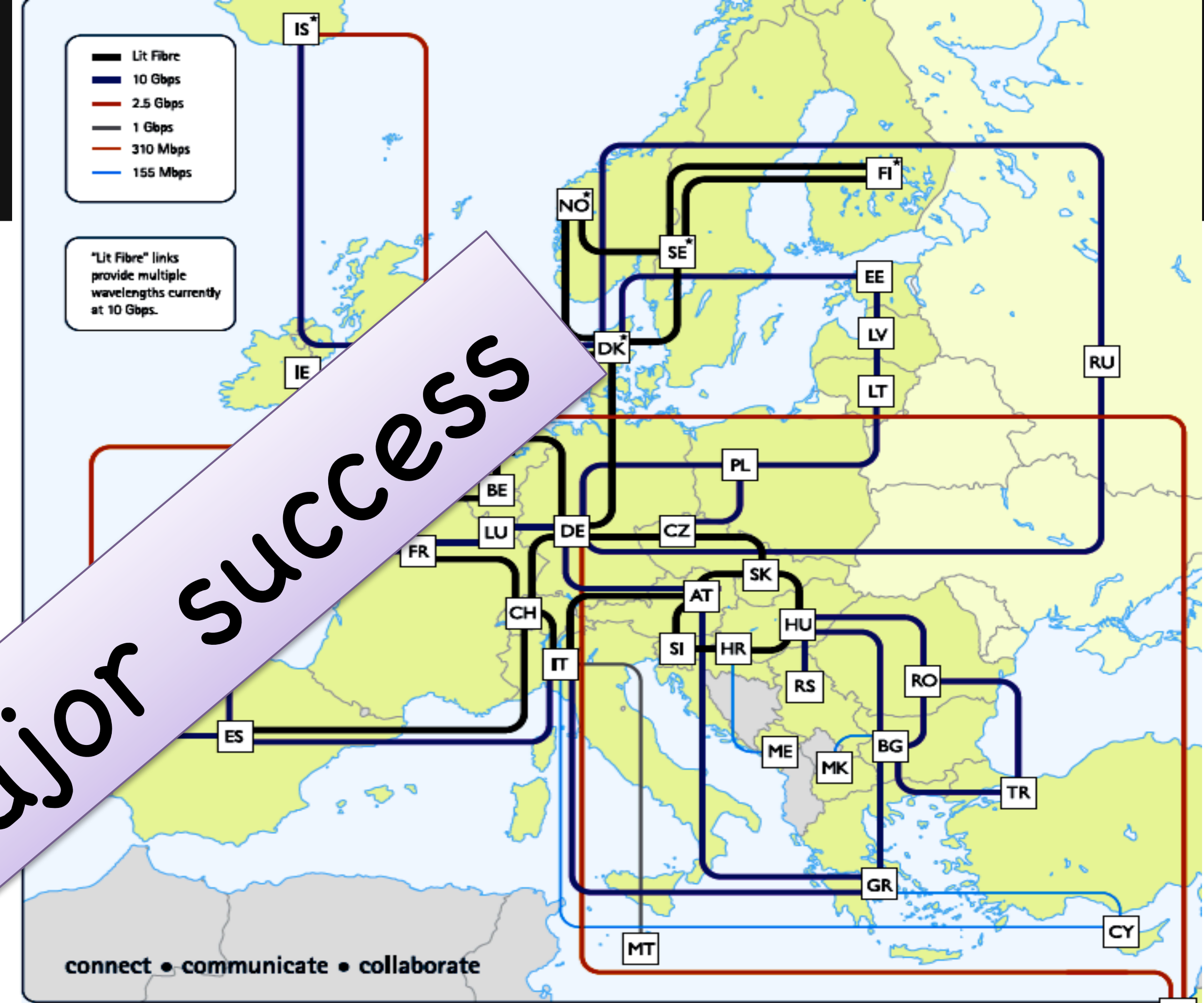
What we actually used at LHC start!

Original Computing model

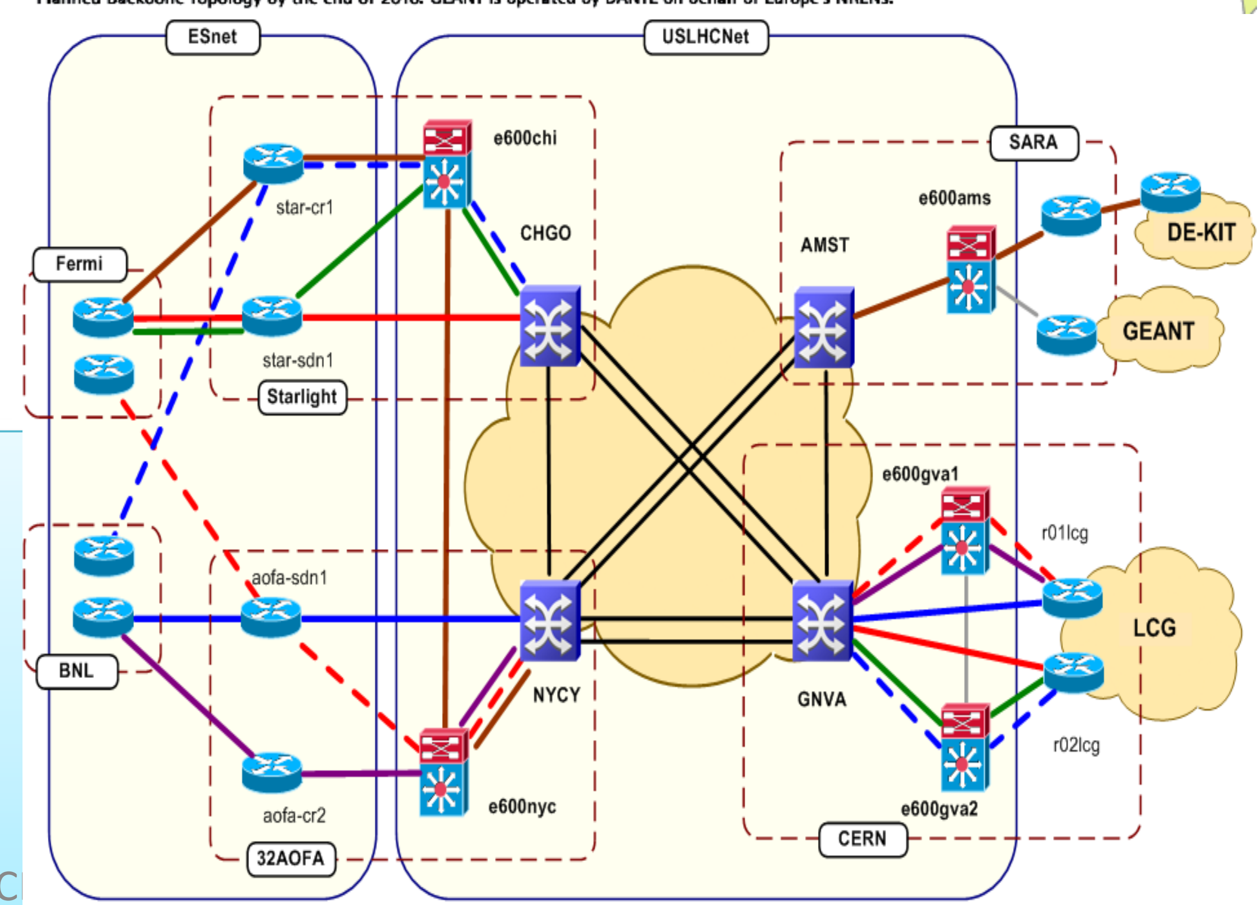


LHC Networking





Planned Backbone Topology by the end of 2010. GÉANT is operated by DANTE on behalf of Europe's NRENs.

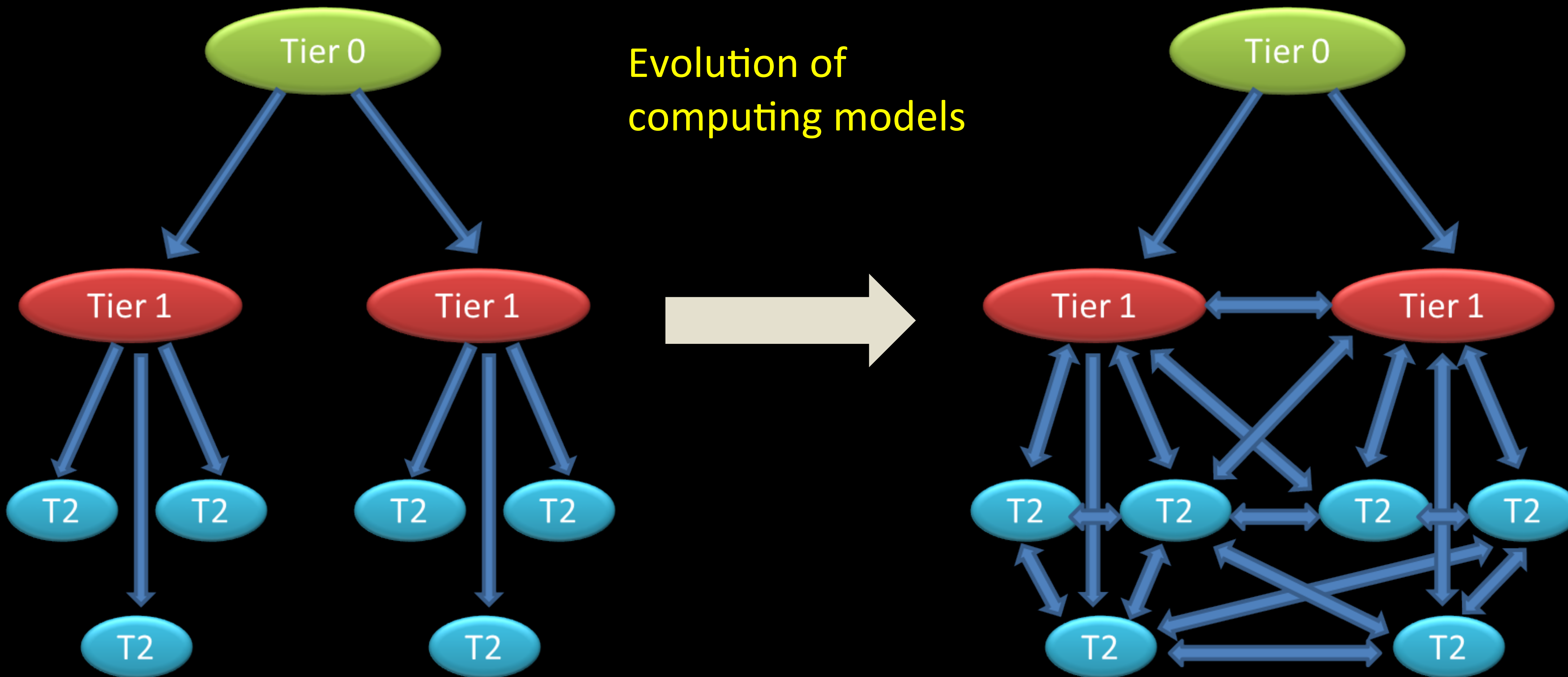


Networking is a major success

- Relies on
  - OPN, GEANT, US-LHCNet
  - NRENs & other national & international providers

Computing model evolution

Evolution of computing models

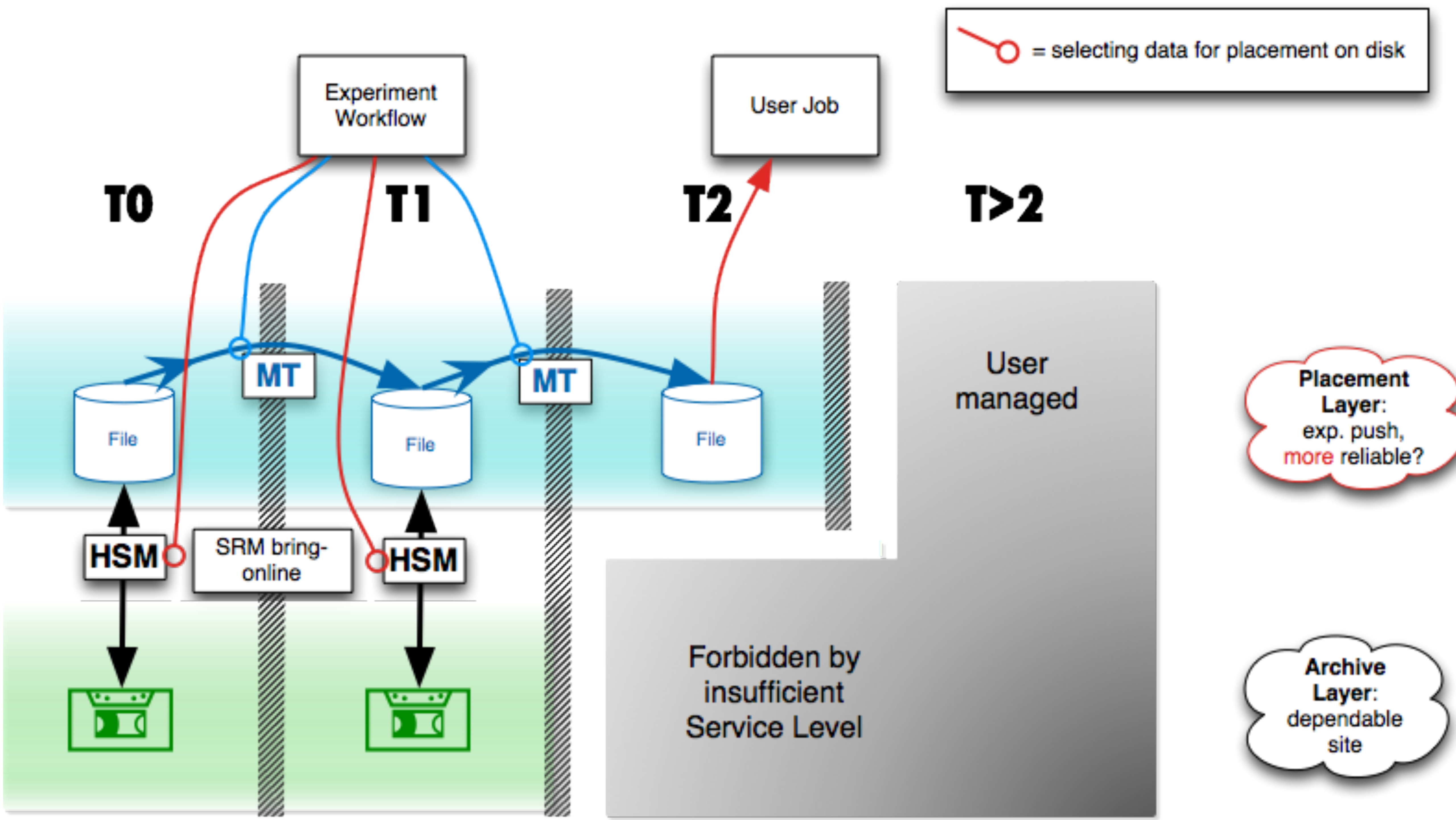


Hierarchy

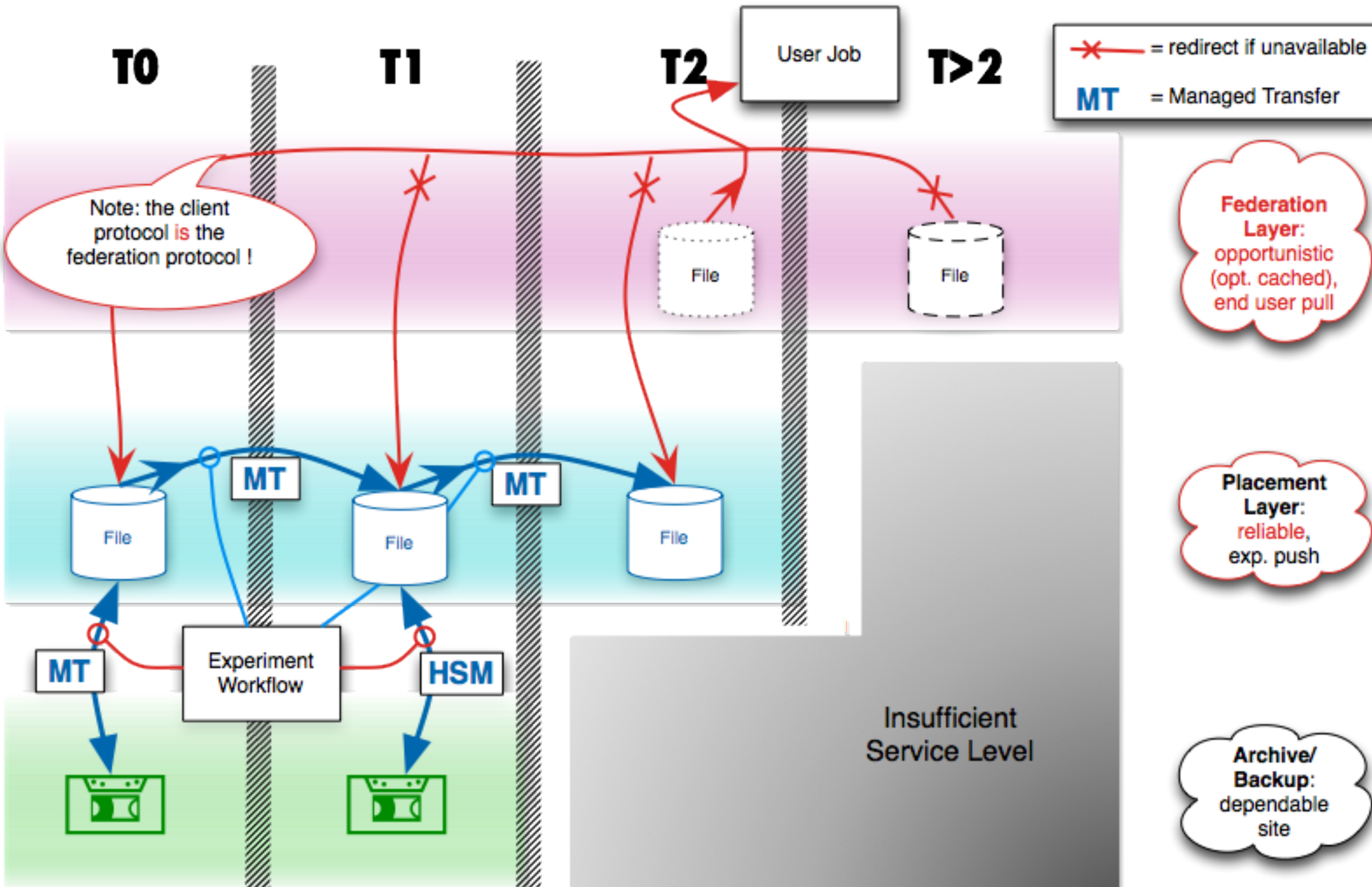
Mesh

WLCG Worldwide LHC Computing Grid

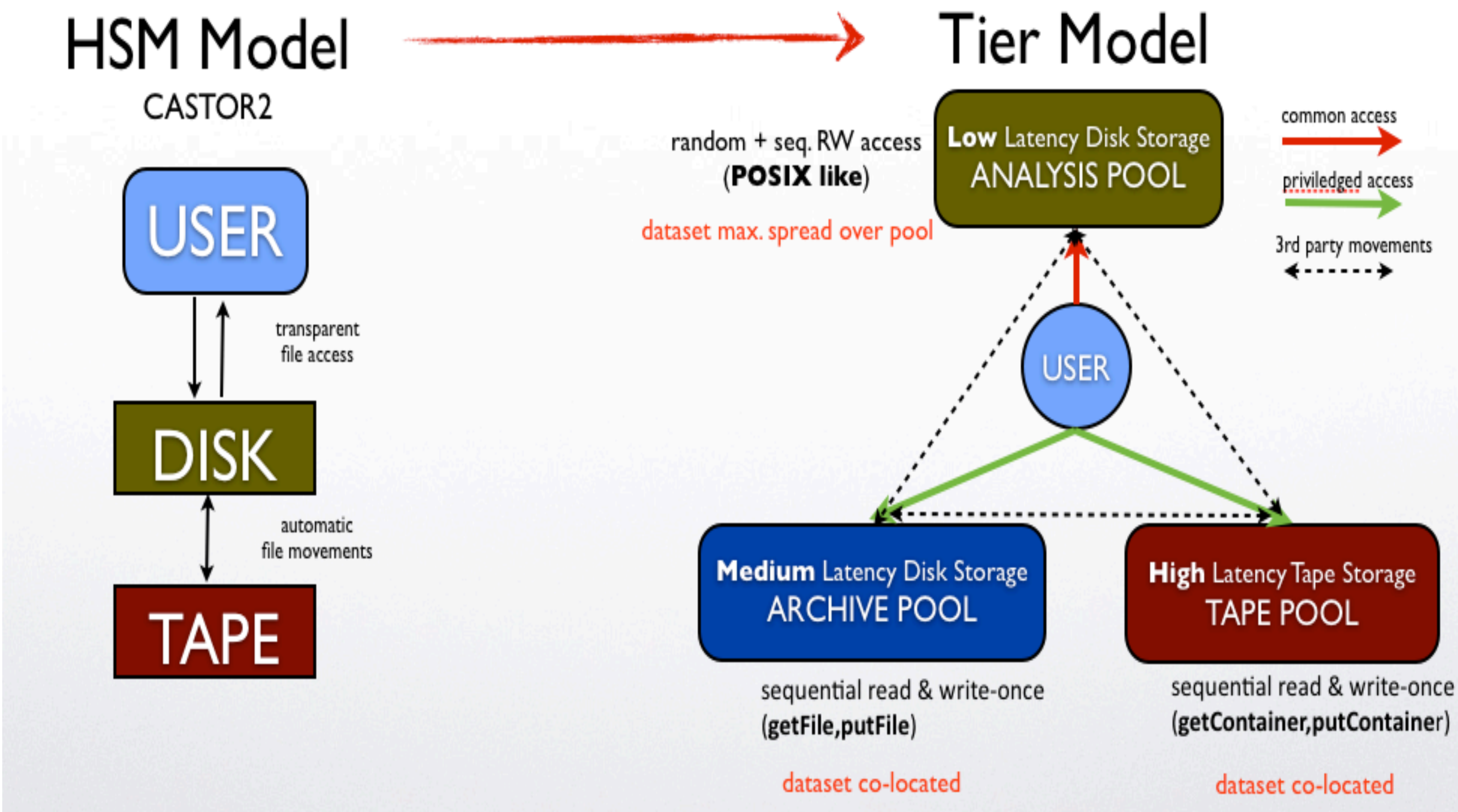
Data Placement



Placement and Federation



CERN : Model change from HSM to more loosely coupled Data Tiers



EOS Design Targets

- Project start: April 2010
- Initial focus: user analysis at CERN
  - many individual users with “chaotic” work patterns
  - many small output files, large shared read-only input files
    - often only partial file access
    - many file seeks over “uninteresting” input events or branches
- Using xroot as client server framework
  - with an in-memory name space (no DB)
  - availability via file-level replication (configurable)
    - reduce operational effort at large volume scale

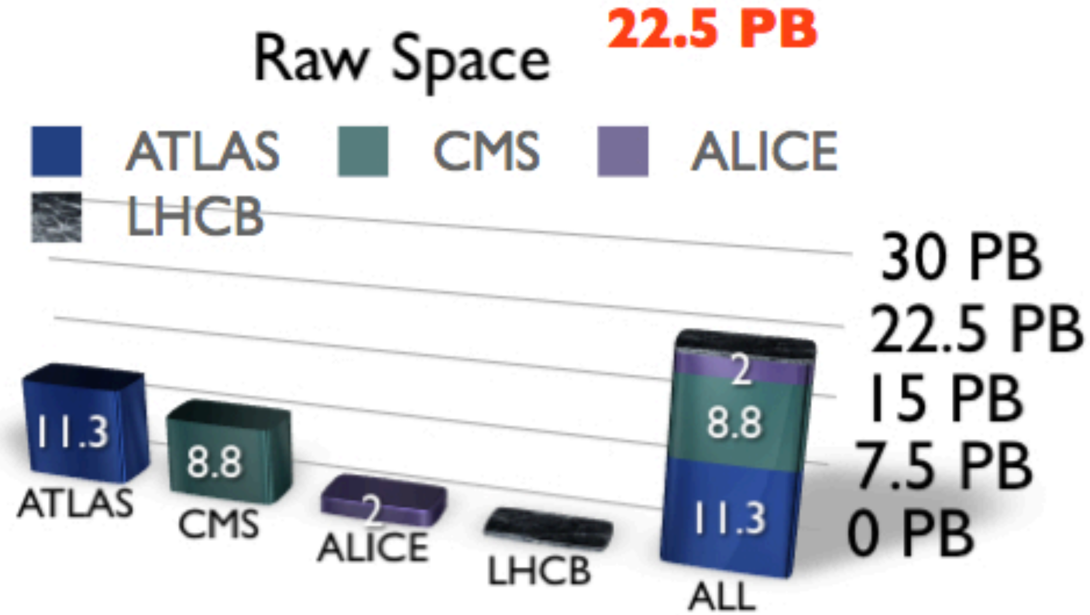
Pessimistic calculation assuming 1 MB file size

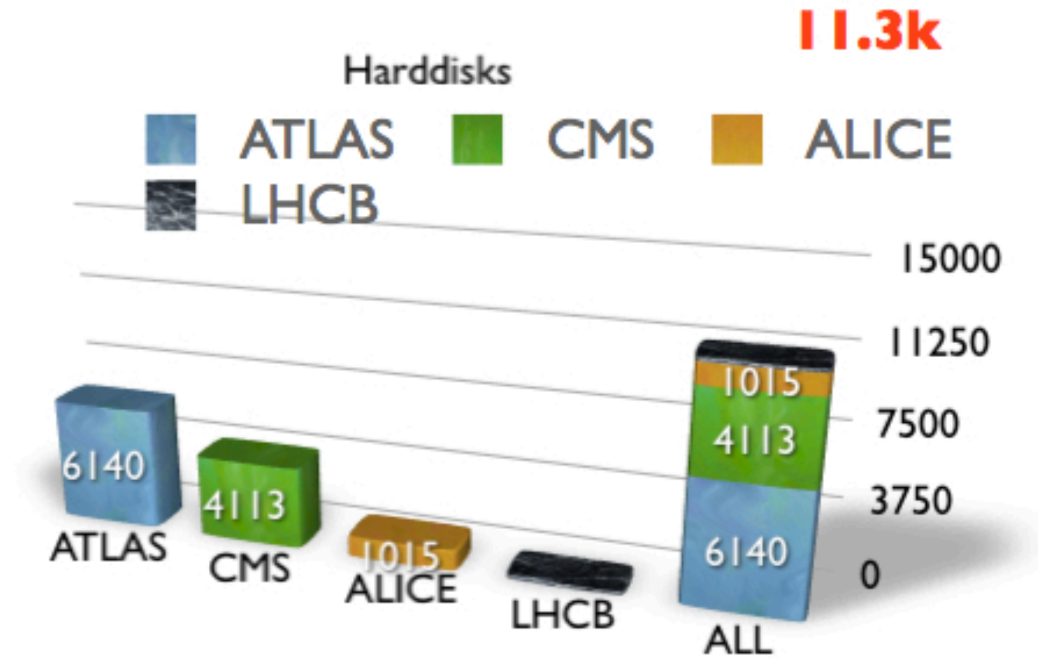
	Access Latency [s]	Files	File Container	Volume [bytes]
Analysis Pool	$10^{-3} – 10^{-2}$	$10^{9} – 10^{10}$ Billions	10^{6} - 10^{7} Millions	$10^{15} – 10^{16}$ Petabytes
Archive Pool	$10^{-2} – 10^{1}$	$10^{11} – 10^{12}$	10^{8} -10^{9} 100 Million+	$10^{17} – 10^{18}$ Exabytes
Tape Pool	$10^{1} – 10^{3}$	10^{11} - 10^{12}	10^{8} -10^{9} 100 Million+	$10^{17} – 10^{18}$ Exabytes

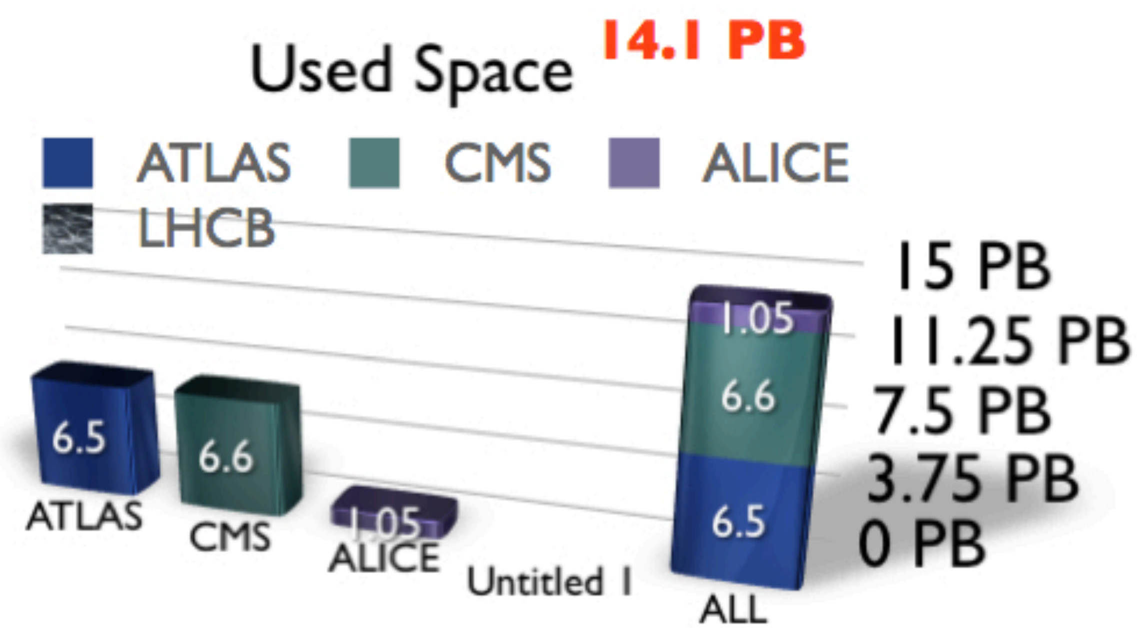
DSS

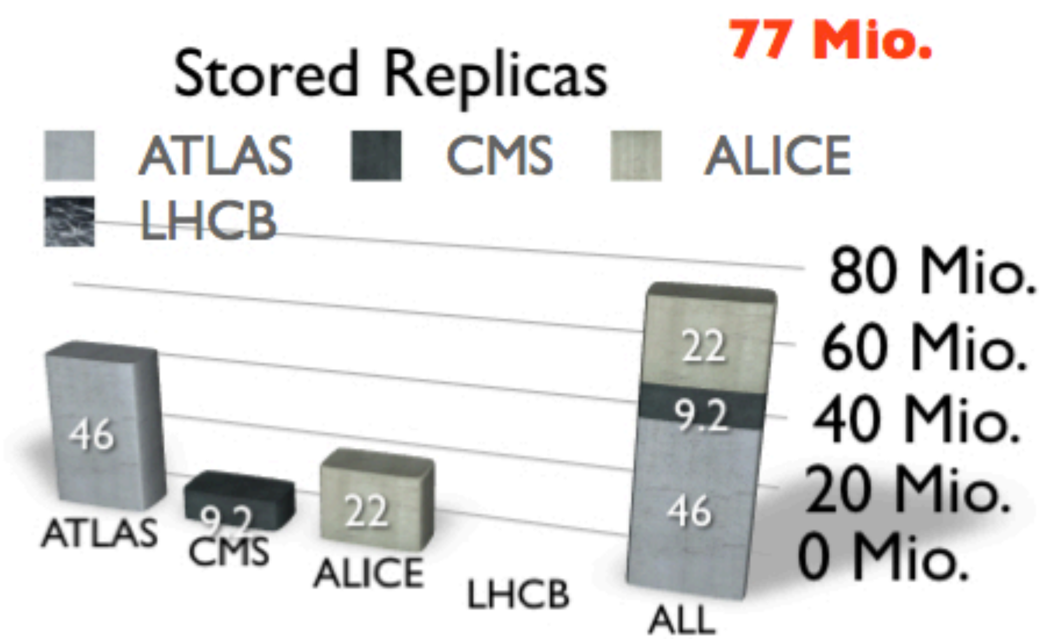
EOS Deployment Status Today

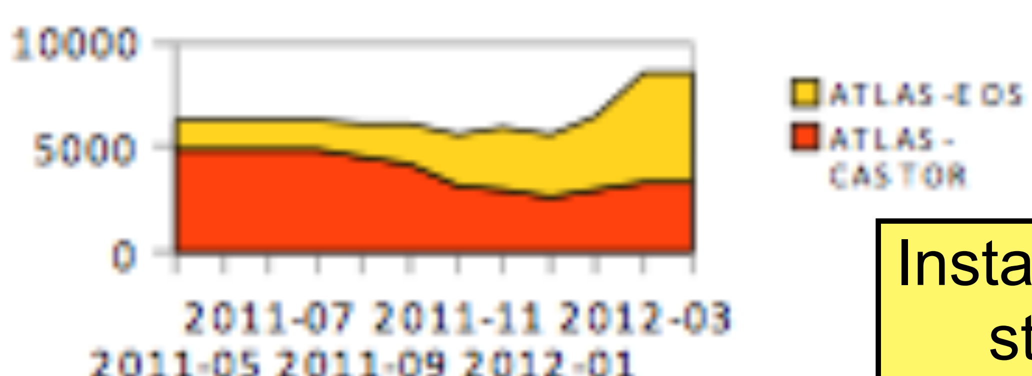
CERN IT Department



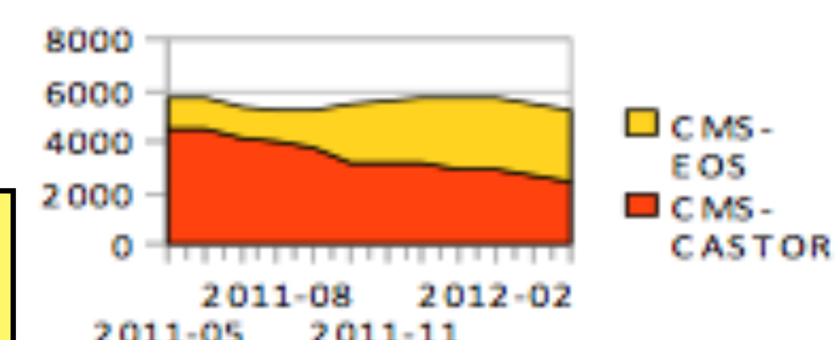








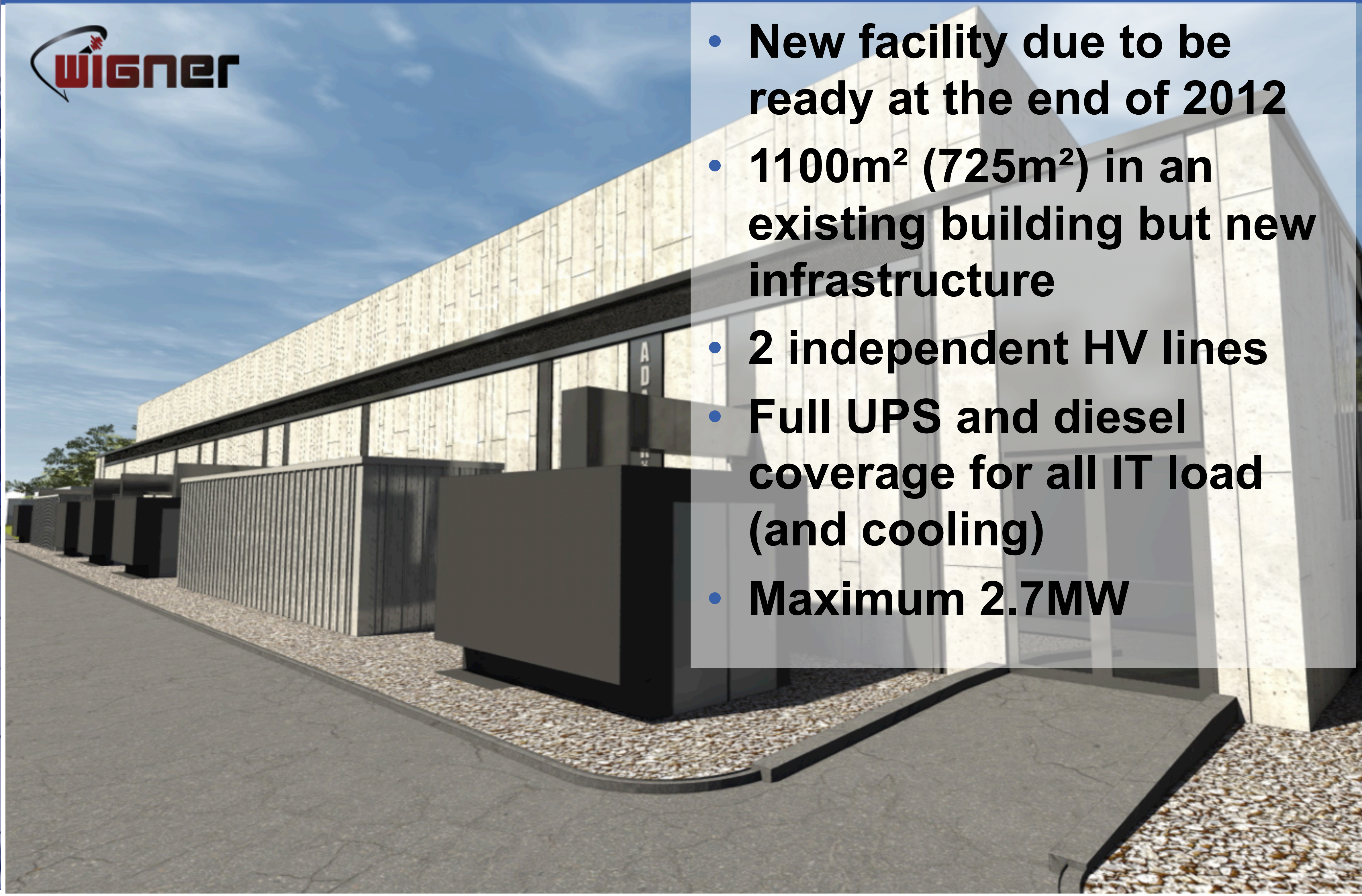
Installation for LHCb
still under test



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

Wigner Data Centre, Budapest

CERN IT Department



- New facility due to be ready at the end of 2012
- 1100m² (725m²) in an existing building but new infrastructure
- 2 independent HV lines
- Full UPS and diesel coverage for all IT load (and cooling)
- Maximum 2.7MW

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

Cloud Storage - Semantics and Protocol

CERN IT Department

- Simple Storage Service (Amazon S3)
  - "just" a storage service
    - in contrast to eg Hadoop, which comes with a distributed computation model exploiting data locality
  - uses a language independent REST API
    - http(s) for transport

- Provide additional scalability by
  - focussing on a defined subset of posix functionality
  - partitioning of namespace into independent buckets

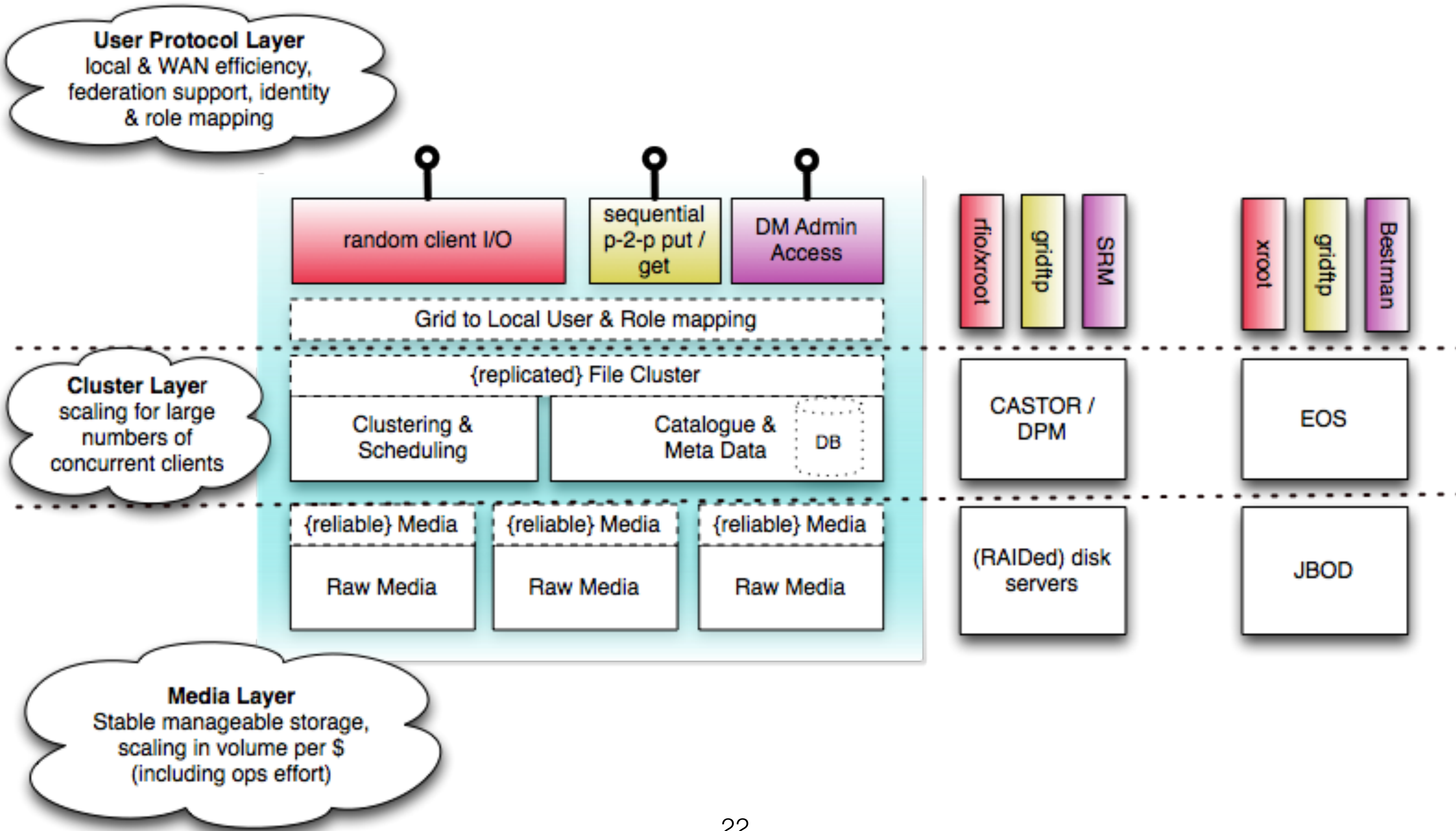
- S3 protocol alone can not provide scalability
  - eg if added on top of a traditional storage system
  - Scalability gains need to be proven for each S3 implementation

Potential Interest for WLCG

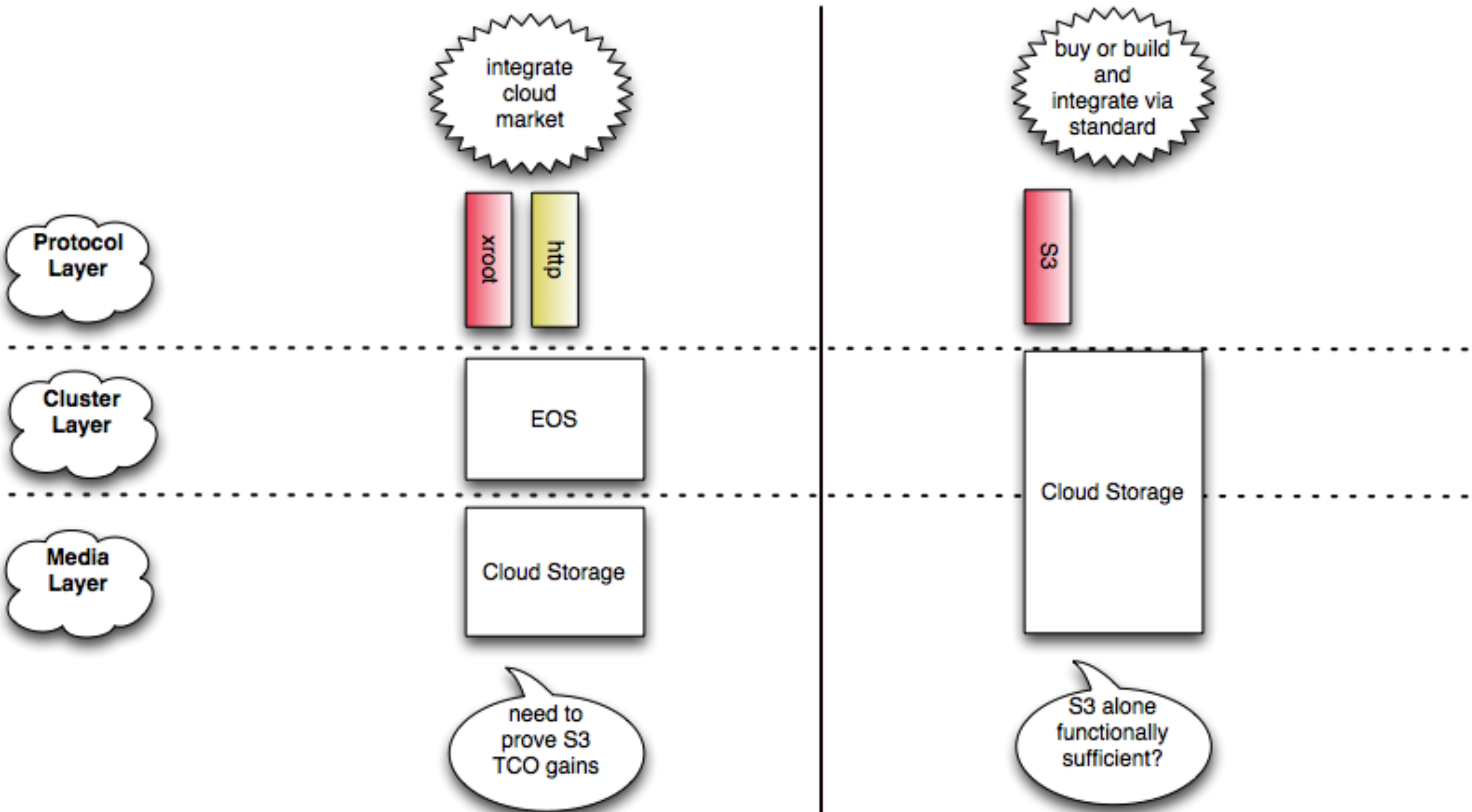
- S3 Protocol could be a standard interface for access, placement or federation of physics data
- Allowing to provide (or buy) storage services without change to user application
  - large sites may provide private clouds storage on acquired hardware
  - smaller sites may buy S3 or rent capacity on demand
- First Steps
  - successful deployment at one site (eg CERN)
  - demonstrate data distribution across sites (S3 implementations) according to experiment computing models

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

Component Layering in current SEs



Potential Future Scenarios



Preliminary Results

- OpenStack/Swift and Huawei reach similar (10-20% less) performance as EOS
  - for full file access for small to moderate number of clients (O(100))
- Analysis type access using the ROOT S3 plugin
  - naive use (no TTreeCache) of both S3 implementations shows significant overhead
  - with enabled cache and vector read this overhead is removed
- S3fs (= fuse mounted S3 storage) almost reaches the same performance for jobs accessing 10-100% of a file
  - assuming that local cache space (/tmp) is available
- Authentication and authorisation
  - not yet mapped from certificates used in WLCG
- Plan to publish a more quantitative comparison at autumn HEPiX

Summary

CERN IT Department

- Distributed Data Management is crucial for obtaining rapid physics results from LHC data
- Initial strategy is being refined to further increase the efficiency of the available resource
- Strategy of decoupling Archive from Disk storage has been implemented at CERN
  - Reducing the total deployment effort and the interference impact for experiment users
- Federated data access is being used or evaluated by several LHC experiments
  - Larger infrastructures have been setup in US/Europe
- Cloud storage evaluation has started at CERN
  - Performance of local S3 based storage looks comparable to current production system
  - Realistic TCO estimation can not yet be done in a small (1PB) test system w/o real users access