ATLAS Experiment and GCE

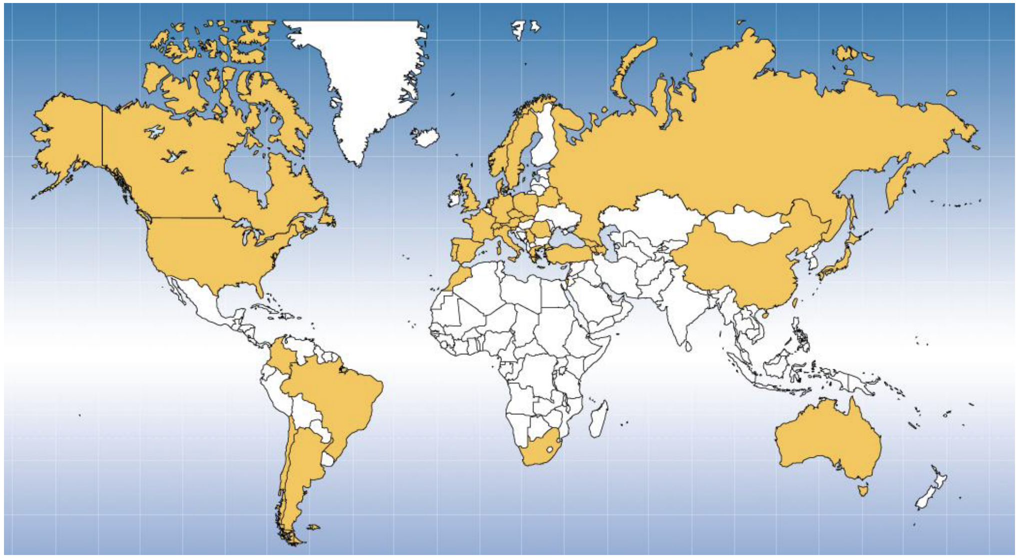
Google IO Conference

San Francisco, CA

May 15-17, 2013

Sergey Panitkin (BNL) and Andrew Hanushevsky (SLAC), for the ATLAS Collaboration

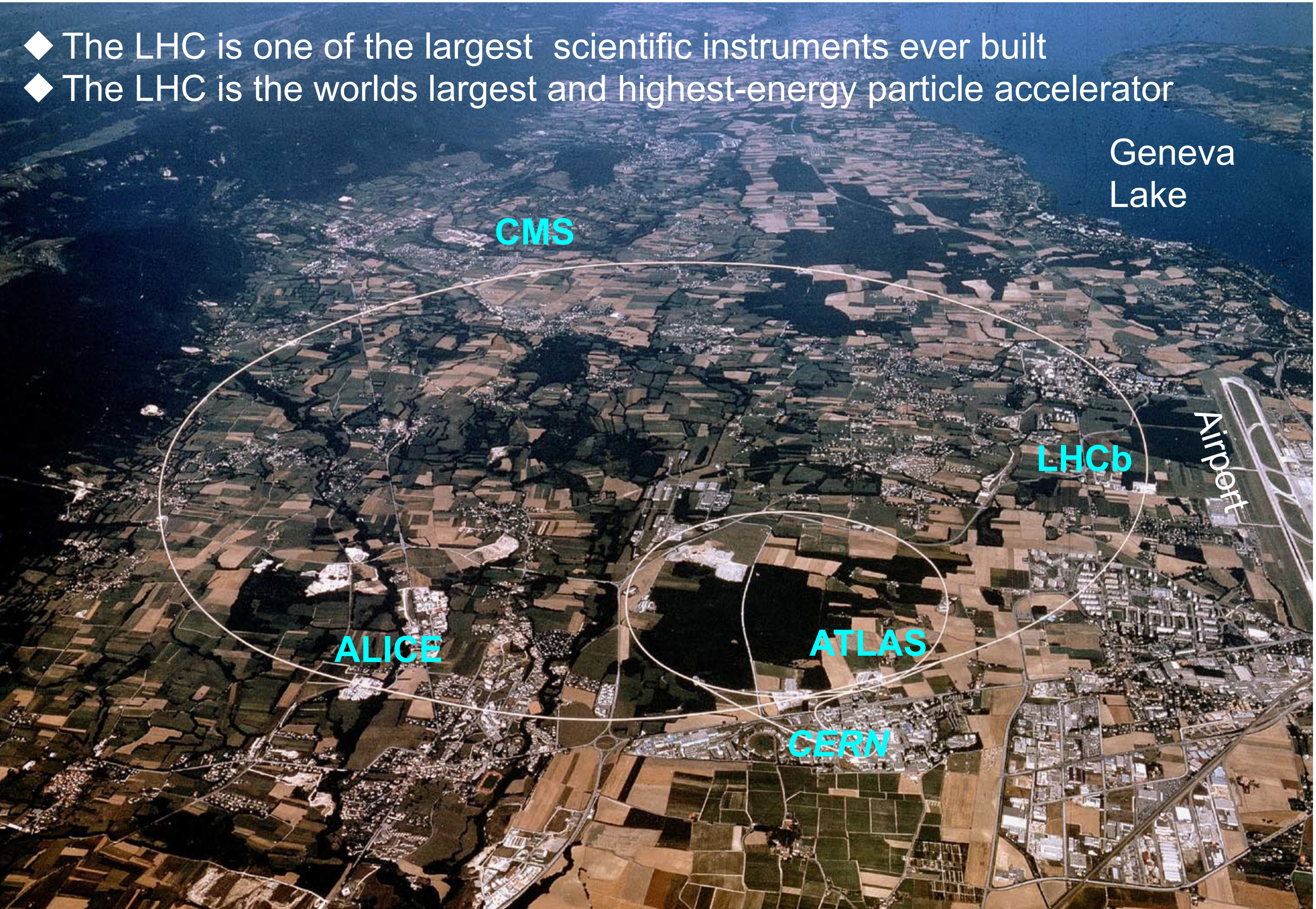
ATLAS Experiment



- The ATLAS is one of the six particle detectors at Large Hadron Collider (LHC) at CERN, one of the two general purpose detectors
- The ATLAS Experiment at LHC is an international collaboration of about 3000 scientists and engineers from 38 countries. They come from 174 Universities and Labs. There are about 1200 graduate students working in ATLAS

Large Hadron Collider. View from above

- The LHC is one of the largest scientific instruments ever built
- The LHC is the worlds largest and highest-energy particle accelerator

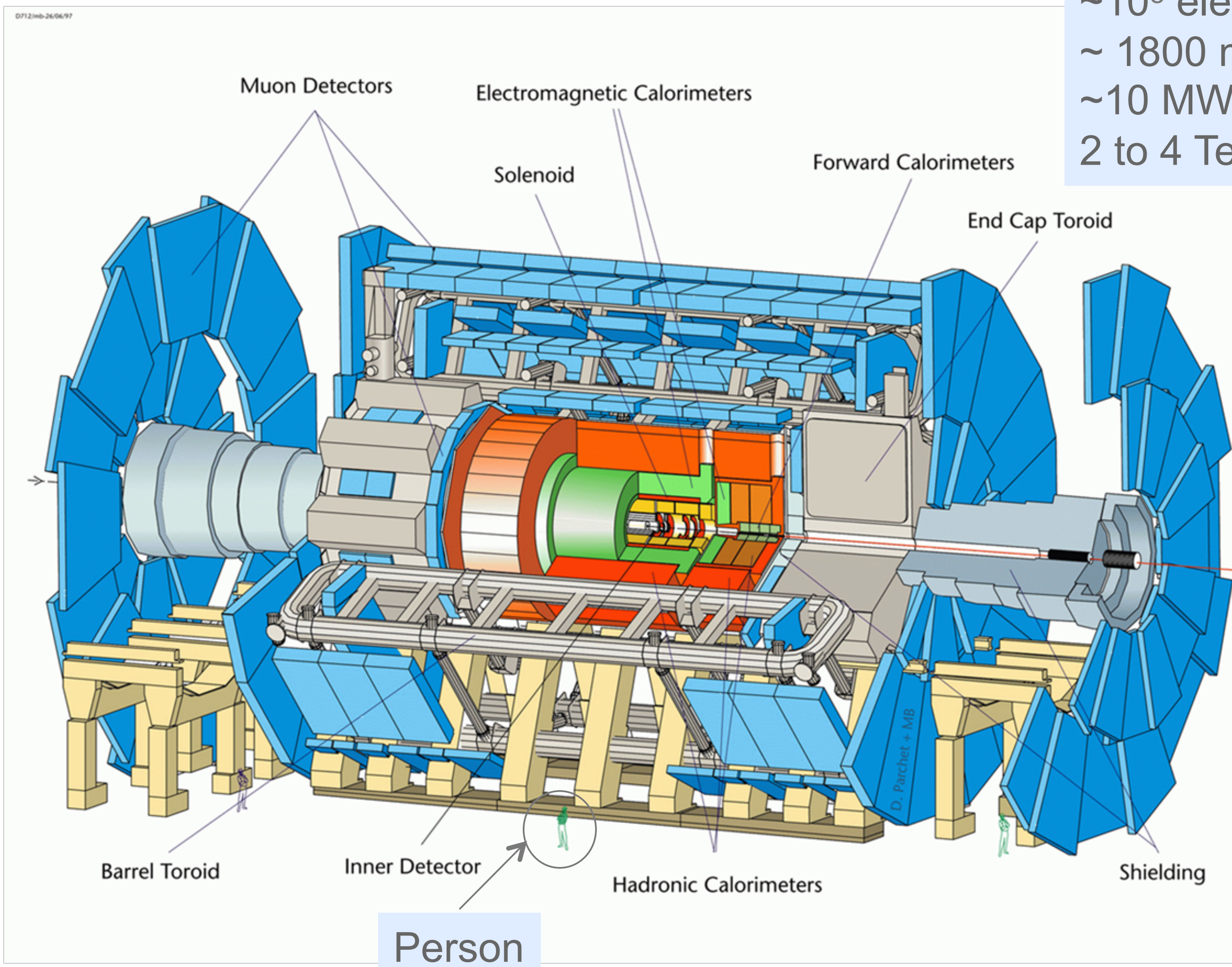


Large Hadron Collider. View from below

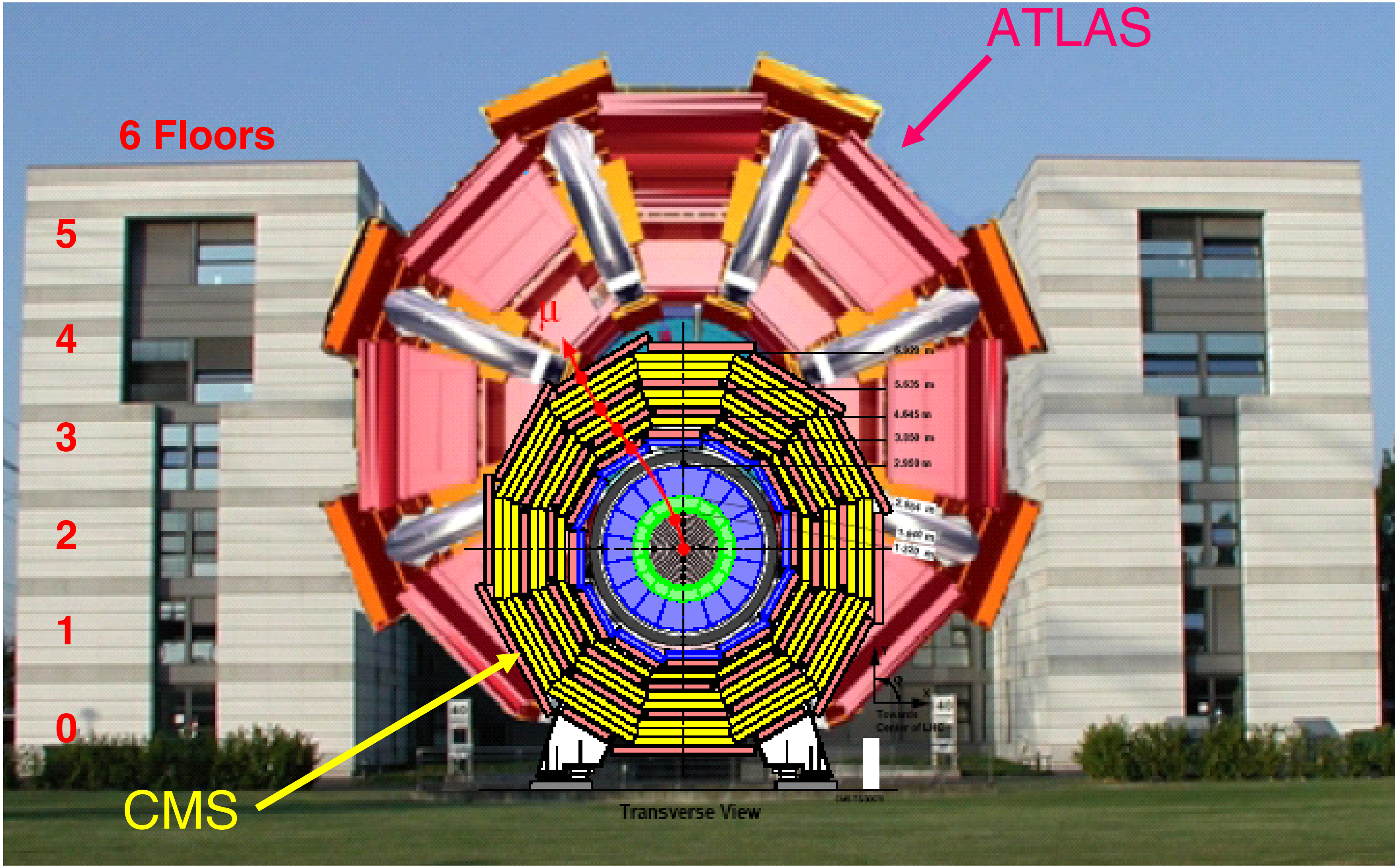
- The LHC tunnel is 26,659 meters long, with ~9600 superconducting magnets needed to accelerate and collide protons and heavy ions at the highest energy ever achieved in laboratory - 14 TeV
- Largest cryogenic installation in the world

ATLAS Detector

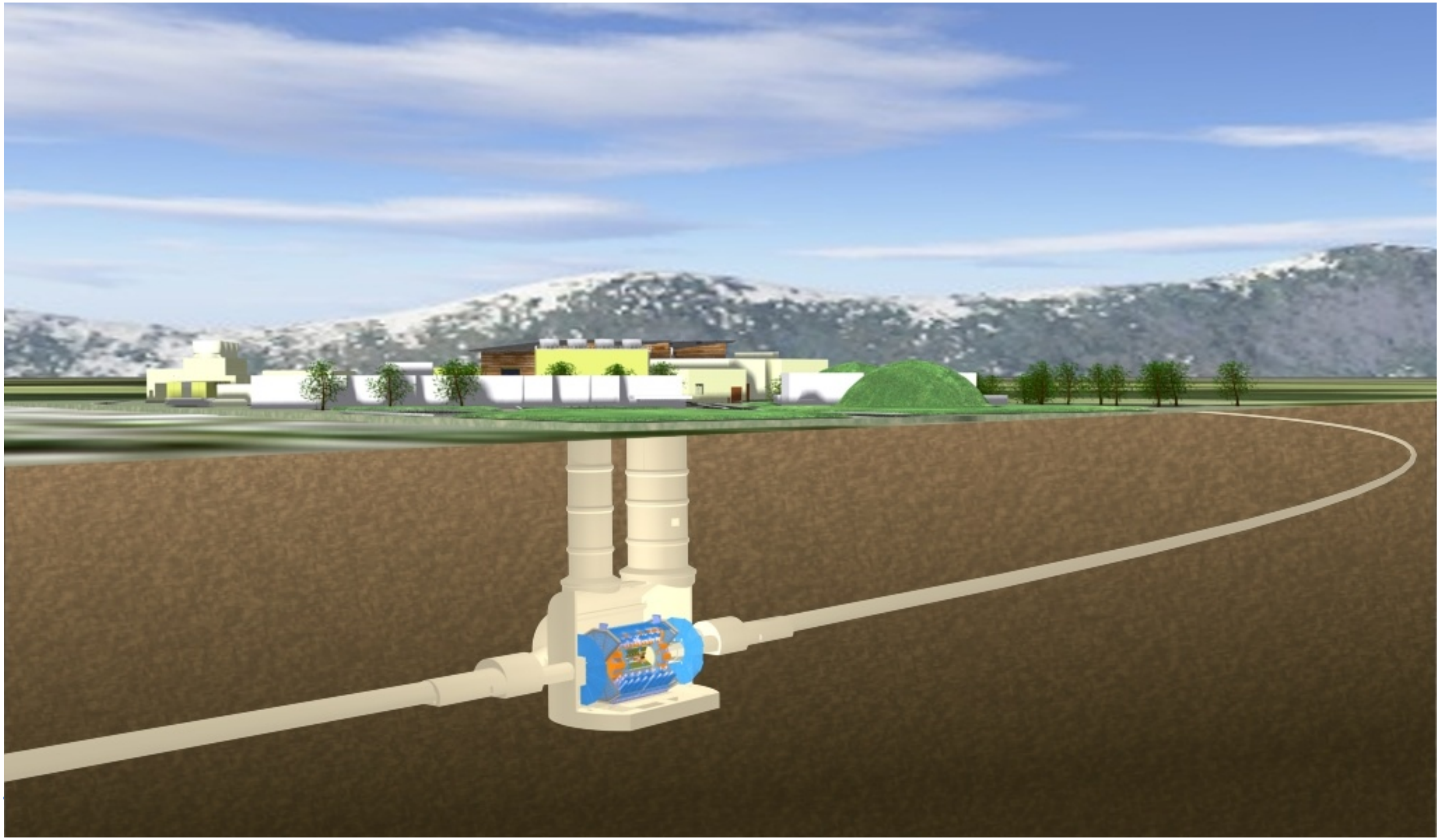
Length : ~46 m (150 ft)
Radius : ~12 m (40 ft)
Weight : ~7000 tons
~10^8 electronic channels
~ 1800 miles of cables
~10 MW of electric power
2 to 4 Tesla mag. field



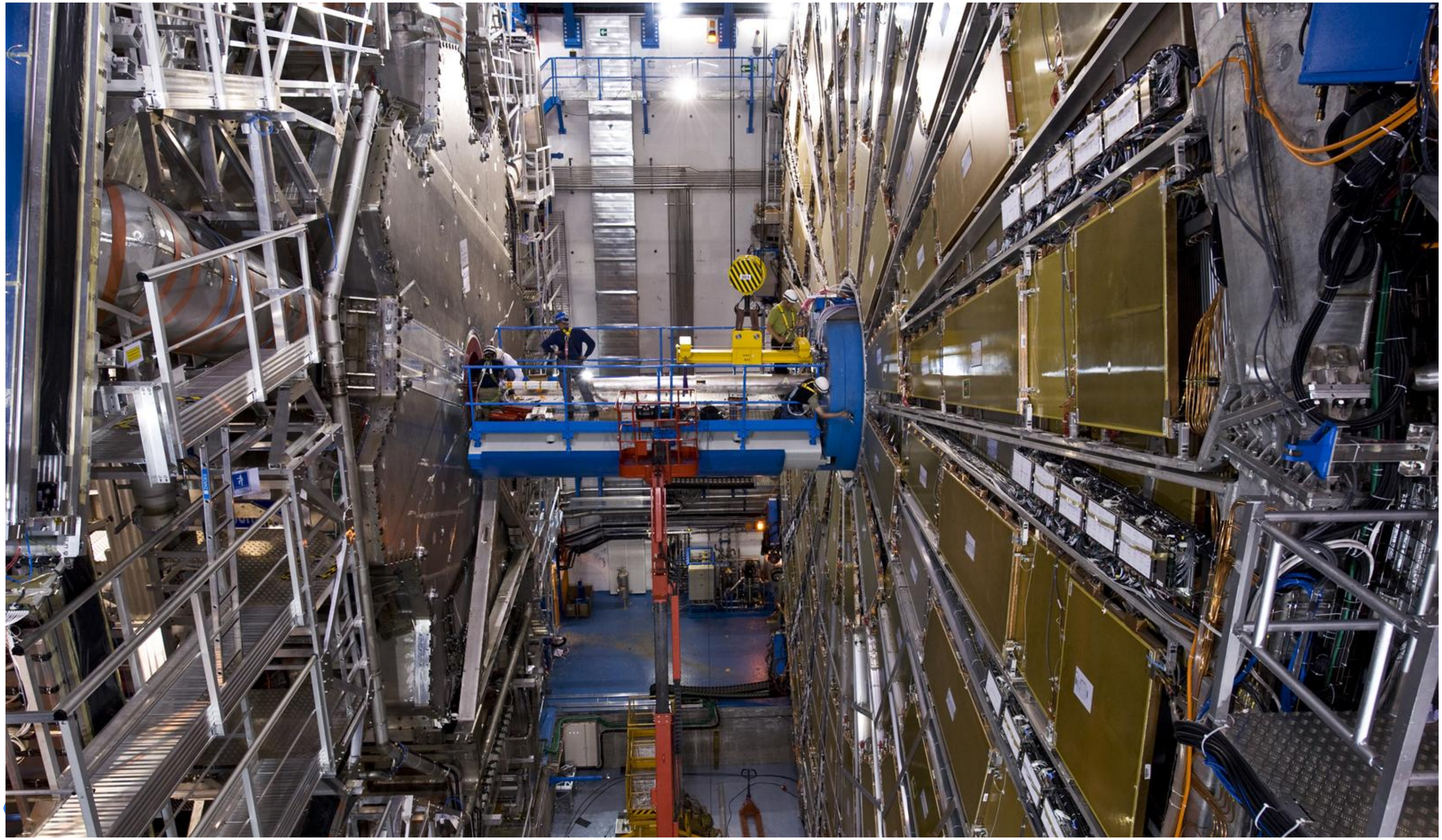
Mr Big and Mr Heavy



ATLAS Underground



ATLAS Detector



ATLAS Physics Goals

- Explore high energy frontier of particle physics
- Search for new physics
  - Higgs boson and its properties
  - Physics beyond Standard Model – SUSY, Dark Matter, extra dimensions, Dark Energy, etc
- Precision measurements of Standard Model parameters

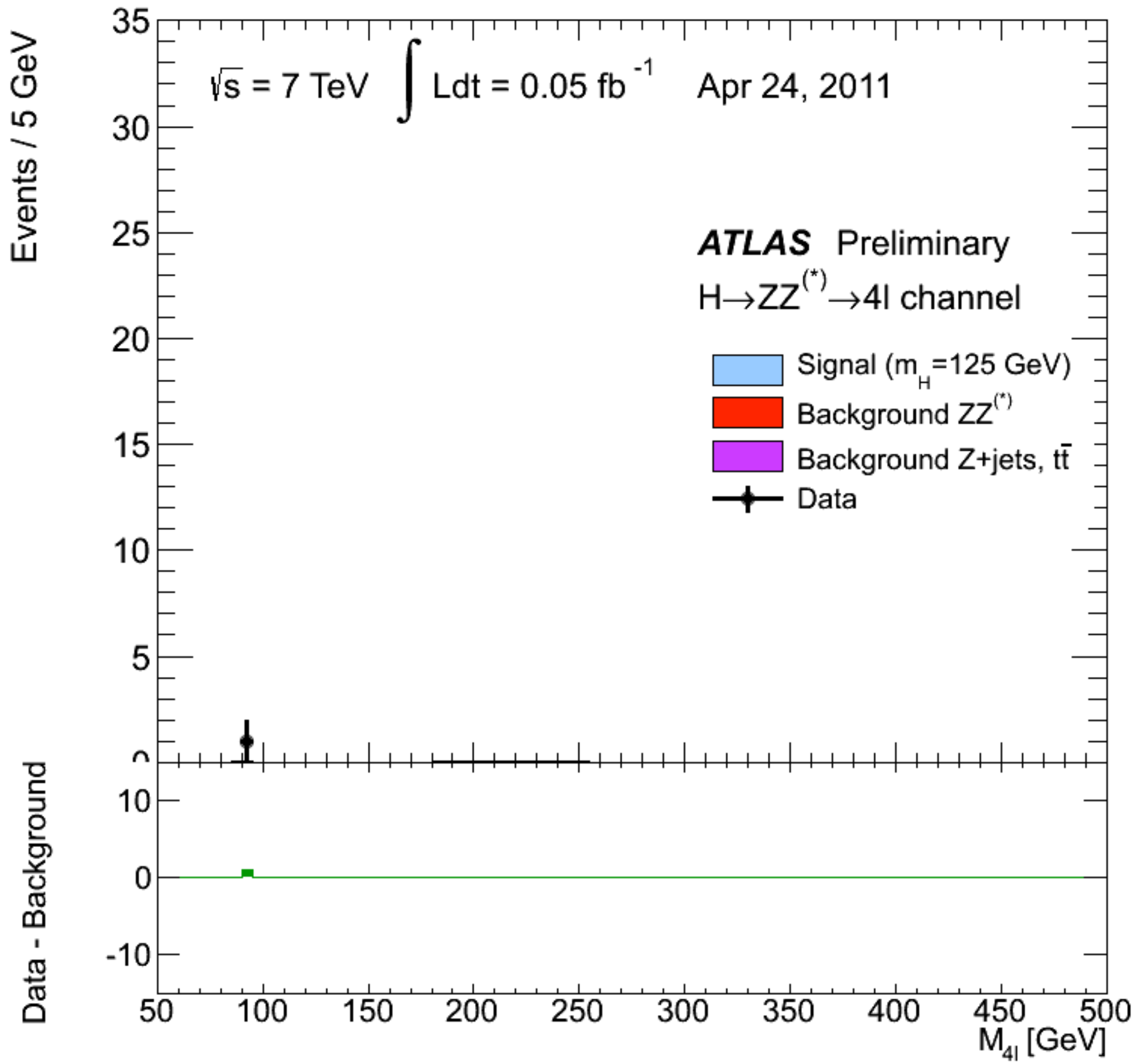
Higgs Boson Discovery

Two seminars from CMS and ATLAS were given on July 4, 2012 at CERN.

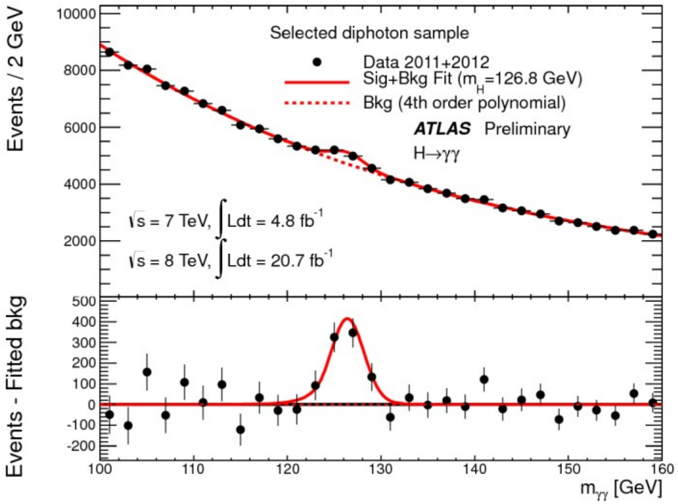
>10,000 print stories
>1,034 television spots (worldwide)

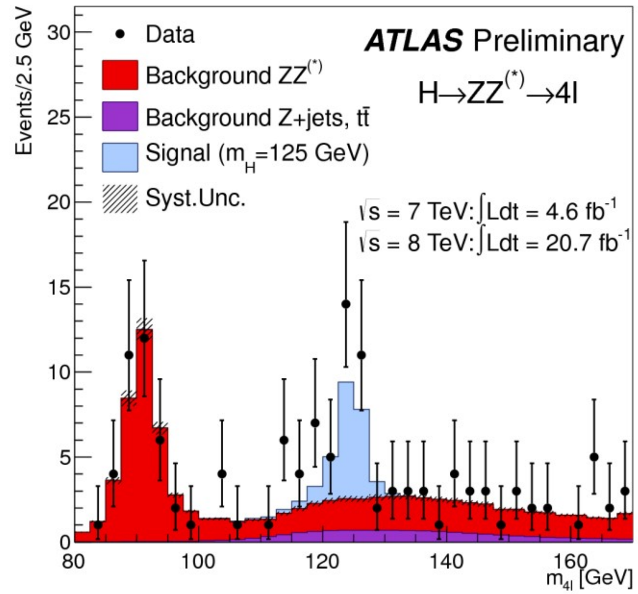
Two scientific papers were submitted on July 31, 2012

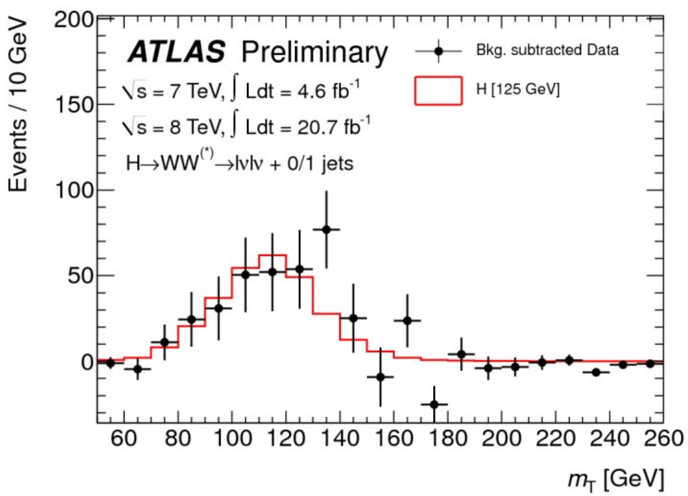
Higgs Boson Discovery



Evidence for a new Particle





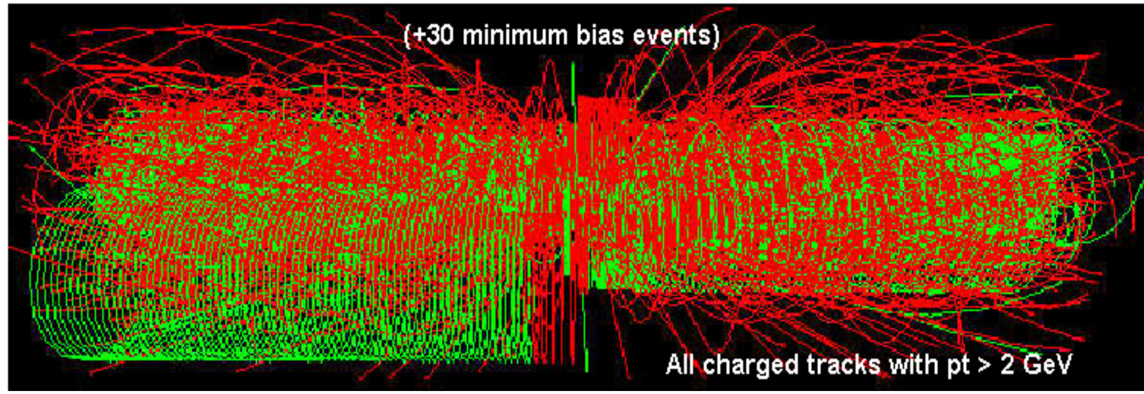


- γγ, ZZ and WW channels updated with full data

- Clear evidence for a new particle seen in these channels

ATLAS Data Challenge

- 800,000,000 proton-proton interactions per second
- 0.0002 Higgs per second
- ~150,000,000 electronic channels
- >10 PBytes of data per year

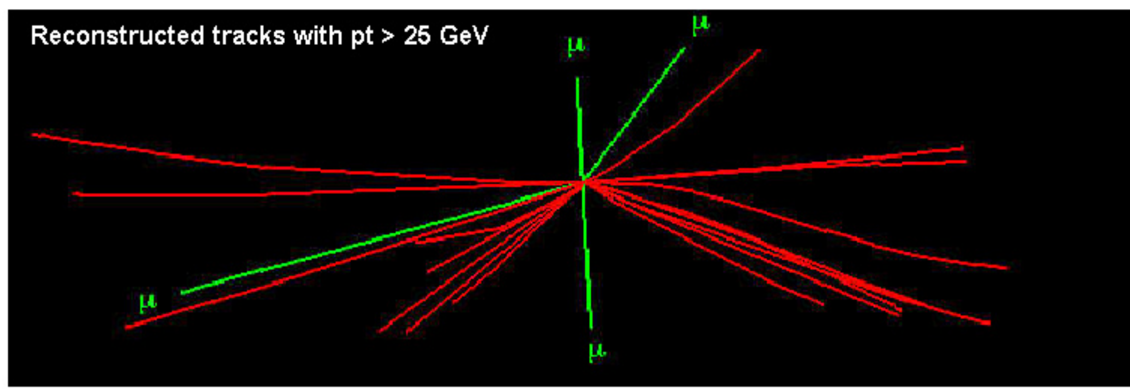


Selectivity: 1 in 10^{13}

Like looking for 1 person in a thousand world populations

Or for a needle in 20 million haystacks!

We are looking for this "signature"



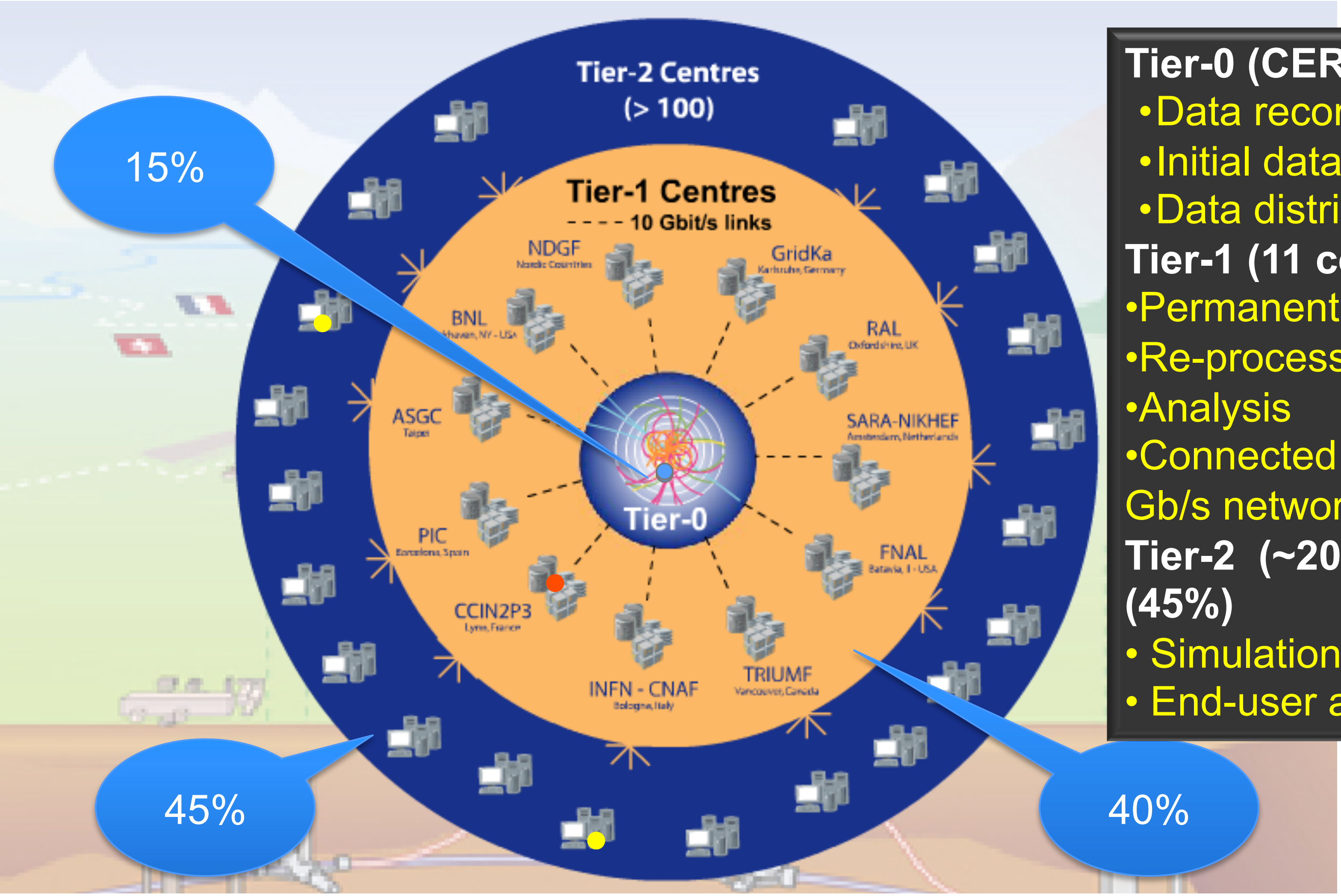
ATLAS – The Big Data Experiment

- ATLAS Detector generates about 1PB of raw data per second – most filtered out in real time by the trigger system
- Interesting events are recorded for further reconstruction and analysis
- As of 2013 ATLAS manages **~140 PB** of data, distributed world-wide to O(100) computing centers
  - Data and simulations
- Expected rate of data influx into ATLAS Grid ~40 PB of data per year in 2014
- Thousands of physicists from ~40 countries analyze the data

ATLAS Computing

- ATLAS uses grid computing paradigm to organize distributed resources
- ATLAS computational resources are managed by PanDA Workload Management System
  - Now successfully manages $O(10^2)$ sites, $O(10^5)$ cores, $O(10^8)$ jobs per year, $O(10^3)$ users

LHC Computing Grid

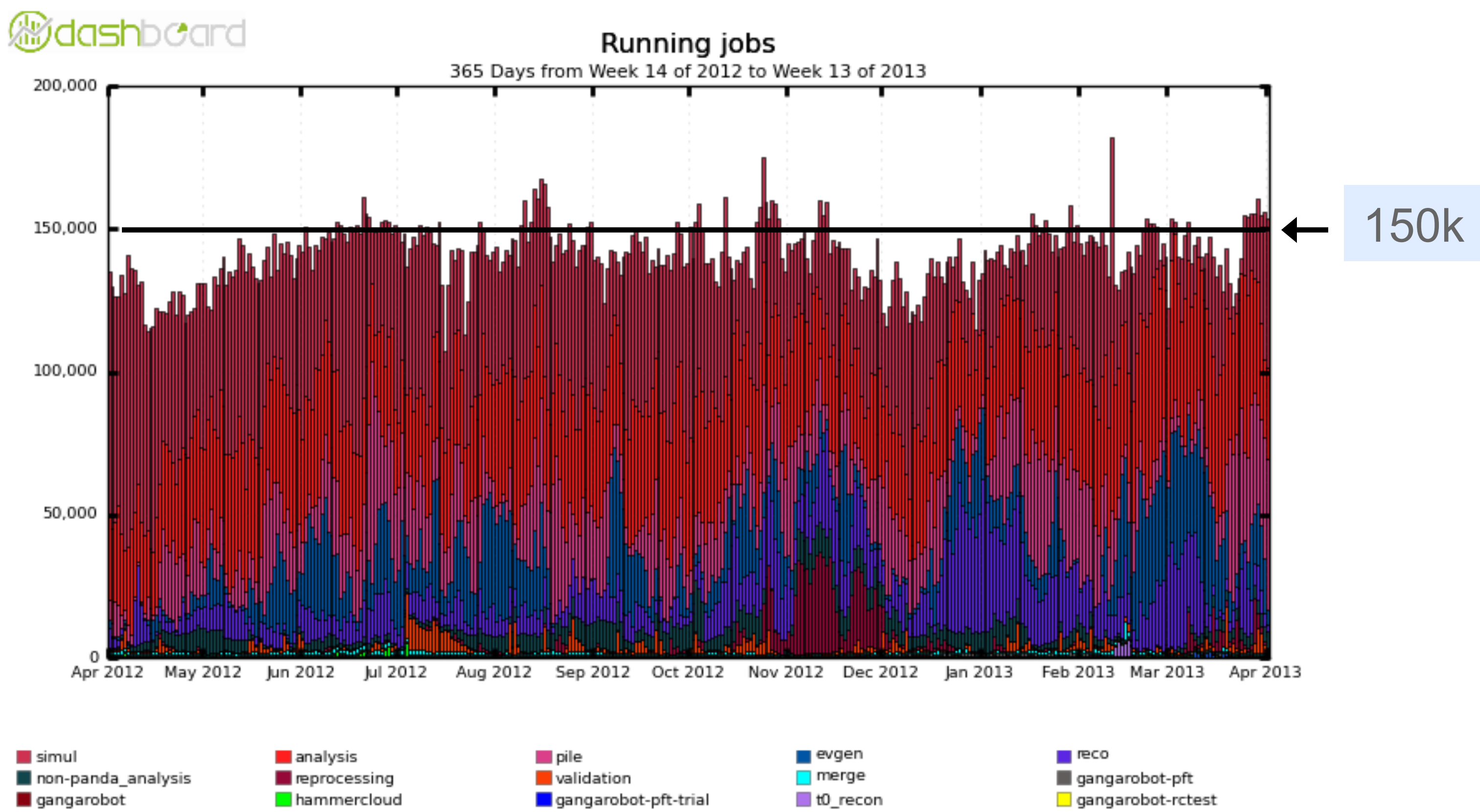


Tier-0 (CERN): (15%)
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres): (40%)
- Permanent storage
- Re-processing
- Analysis
- Connected by direct 10 Gb/s network links

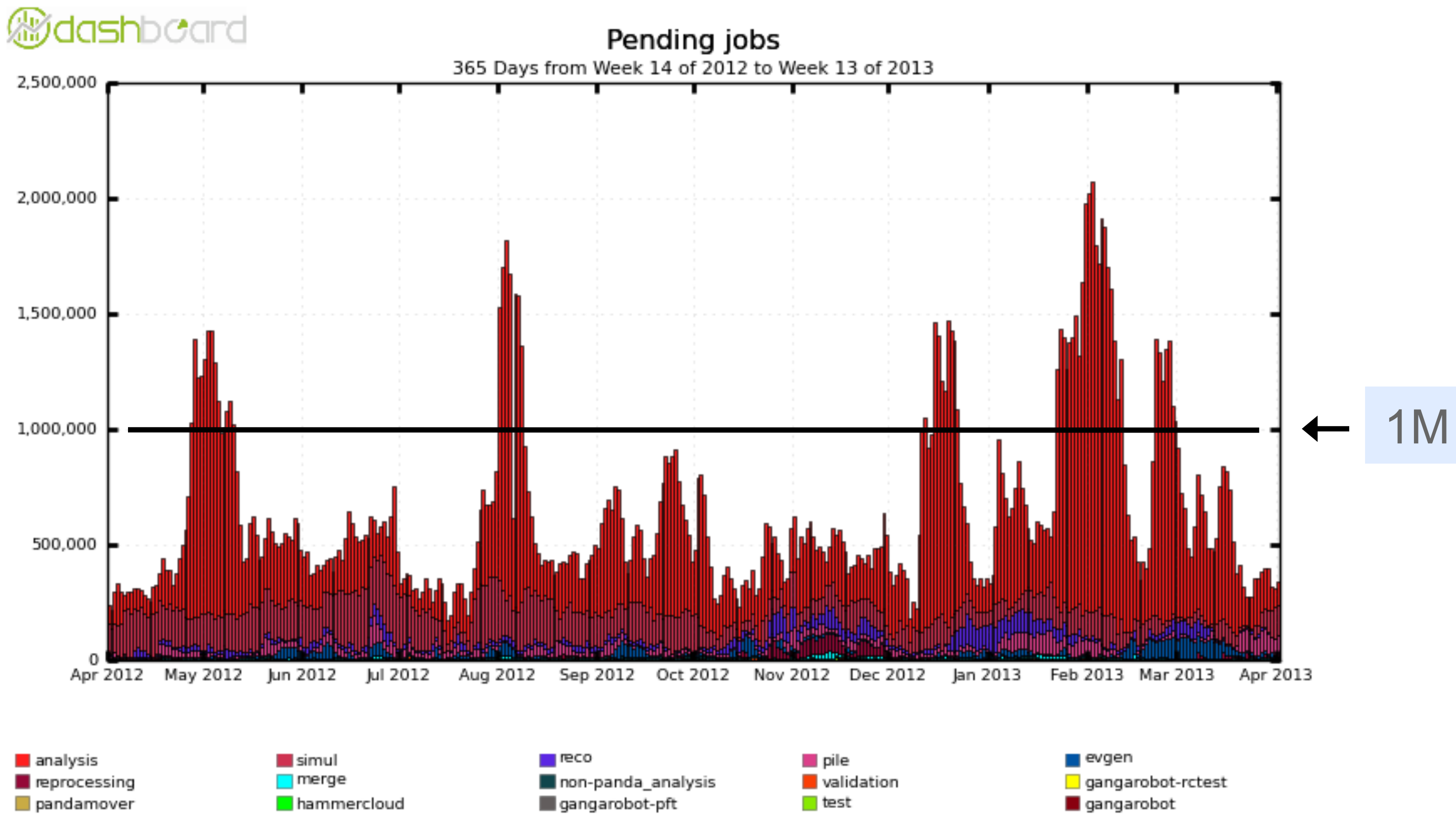
Tier-2 (~200 centres): (45%)
- Simulation
- End-user analysis

ATLAS Distributed Computing



- Includes user and group analysis and Monte-Carlo simulations on ATLAS Grid
- Running on ~100,000 cores worldwide
- Available resources fully used/stressed

ATLAS Distributed Computing II



Spikes in demand for computational resources
Can significantly exceed available ATLAS Grid resources
Lack of resources slows down pace of discovery

Cloud Computing and ATLAS

- A few years ago ATLAS set up cloud computing R&D project to explore virtualization and clouds
- Can we use clouds to cope with spikes in demand for computational resources?
- Experience with variety of cloud platforms
  - Commercial, Academic, Private, Hybrid clouds
- Recent project on Google Compute Engine (GCE)

ATLAS and Google Compute Engine

- We were invited to participate in GCE trial period in August 2012
- Attracted by modern hardware, powerful API, competitive pricing.
- This is Google after all !
- Frustrated that none of the tools that we use supported GCE at that time
- GCE engineers were very helpful
- Google was very gracious in providing more resources than the initial trial quota

ATLAS and Google Compute Engine

- We wanted to try several things on GCE:
  - High performance analysis clusters (PROOF based and other)
  - Cloud storage and data management
  - Use of Xroot for Cloud storage aggregation and interaction with ATLAS Xroot federation
    - **We will talk about Xroot technology in more details later in the talk**
  - PanDA queue for Monte Carlo Simulations

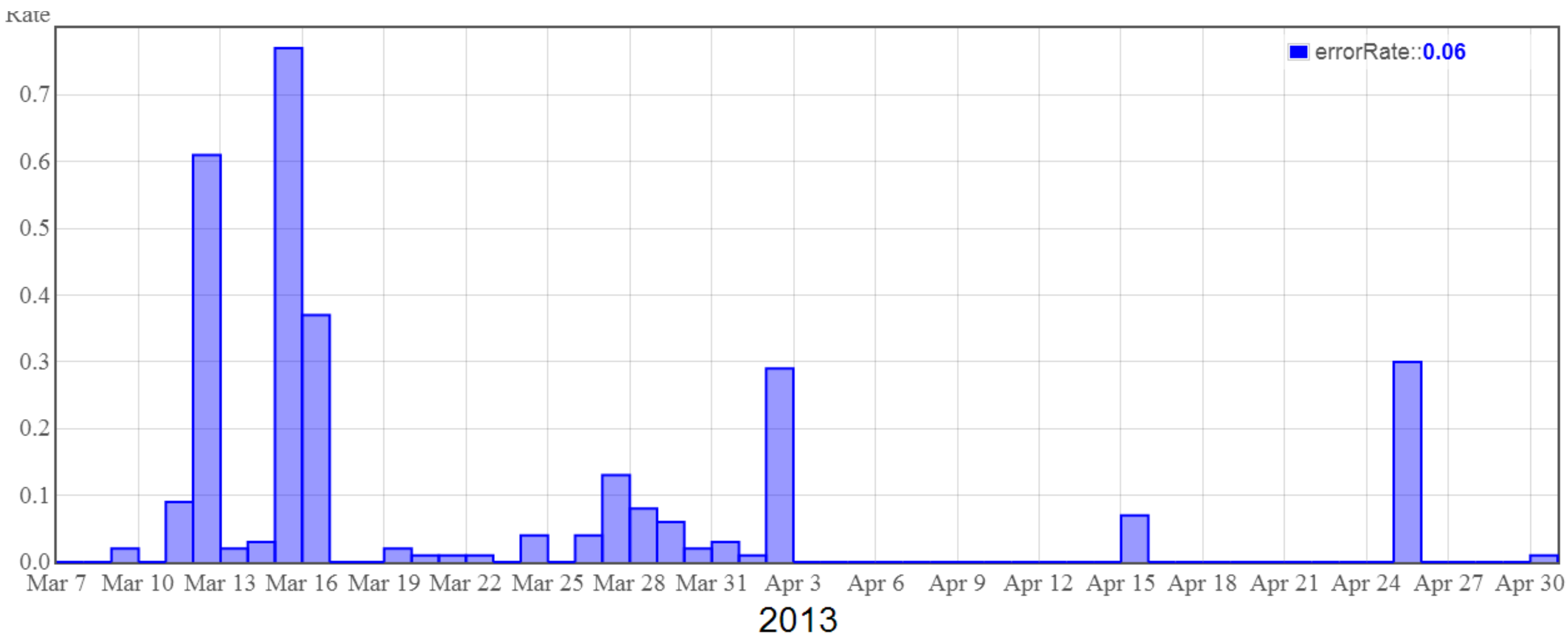
PanDA batch queue on GCE

- Google agreed to allocate additional resources for ATLAS
  - ~5M core-hours,   4000 cores for about 2 month, (original preview allocation was 1k cores)
- Resources were organized as HTCondor based PanDA queue
- Transparent inclusion of cloud resources into ATLAS Grid
- The idea was to test long term stability while running a cloud cluster similar in size to Tier 2 site in ATLAS
- Intended for CPU intensive Monte-Carlo simulation workloads
- Planned as a production type of run. Delivered to ATLAS as a resource and not as an R&D platform.

PanDA batch queue on GCE II

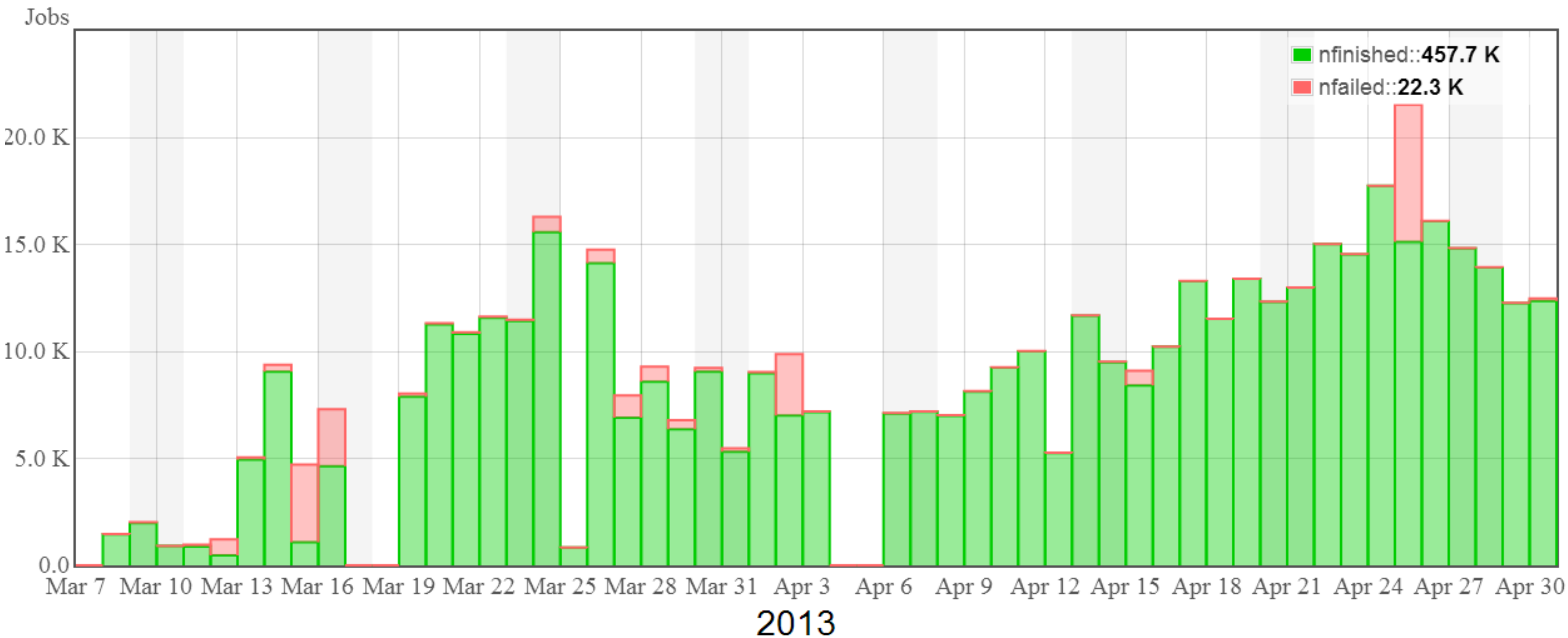
- We ran for about 8 weeks (2 weeks were planned for scaling up)
- Very stable running on the Cloud side. GCE was rock solid.
- Most problems that we had were on the ATLAS side.
- We ran computationally intensive jobs
  - Physics event generators, Fast detector simulation,, Full detector simulation
- Completed 458,000 jobs, generated and processed about 214 M events

PanDA queue on GCE. Failure Rate



- Most of the job failures occurred during start up and scale up phase – as expected
- Most of the failures were on the ATLAS side – file transfer, LFC problems, HTCondor
- No failures were due to GCE problems

Failed and Finished Jobs

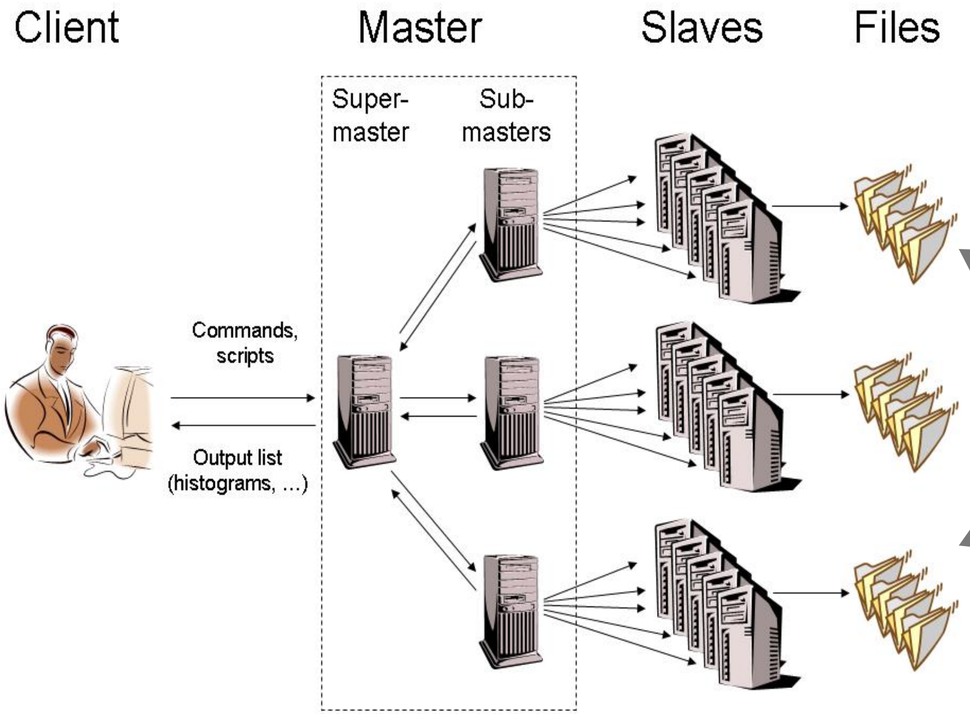


- Most of the job failures occurred during start up and scale up phase – as expected
- Reached throughput of 15k jobs per day

PROOF/Xroot Clusters on GCE

- ◆PROOF is implementation of MapReduce paradigm based on ROOT framework.
- ◆ROOT framework for data analysis in HENP
  - ◆Developed and supported by ROOT Team at CERN
  - Written in C++
  - Free, Open Source
  - More info at: http://root.cern.ch/ ; http://root.cern.ch/drupal/content/proof
- PROOF allows for efficient aggregation and use of distributed computing resources for data intensive event based analyses
- Uses Xroot for clustering, storage aggregation and data discovery
  - Xroot is well suited for ephemeral storage aggregation into one name space
- PROOF clusters can be federated

PROOF Architecture

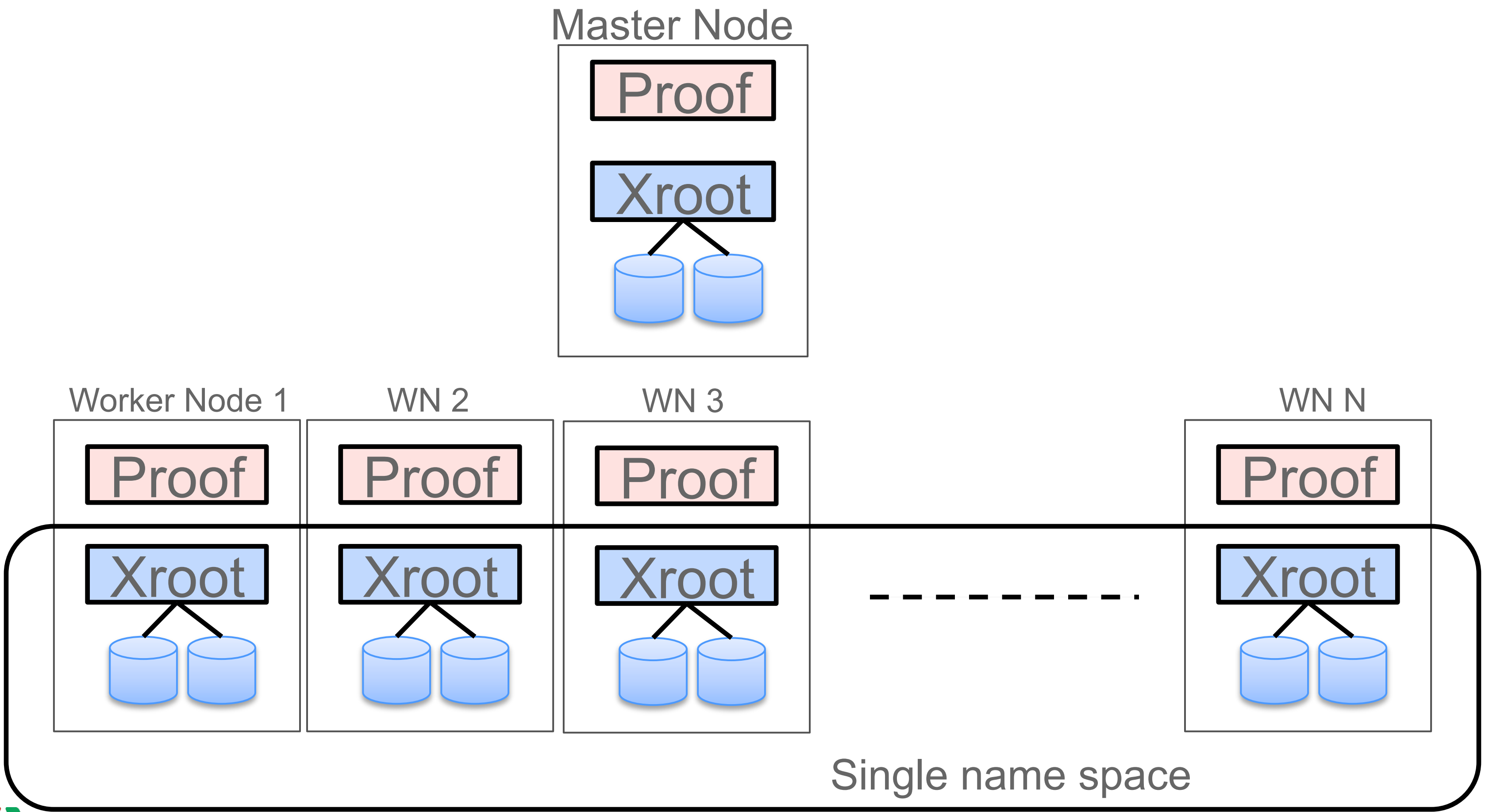


Adapts to wide area *virtual* clusters

Geographically separated domains, heterogeneous machines

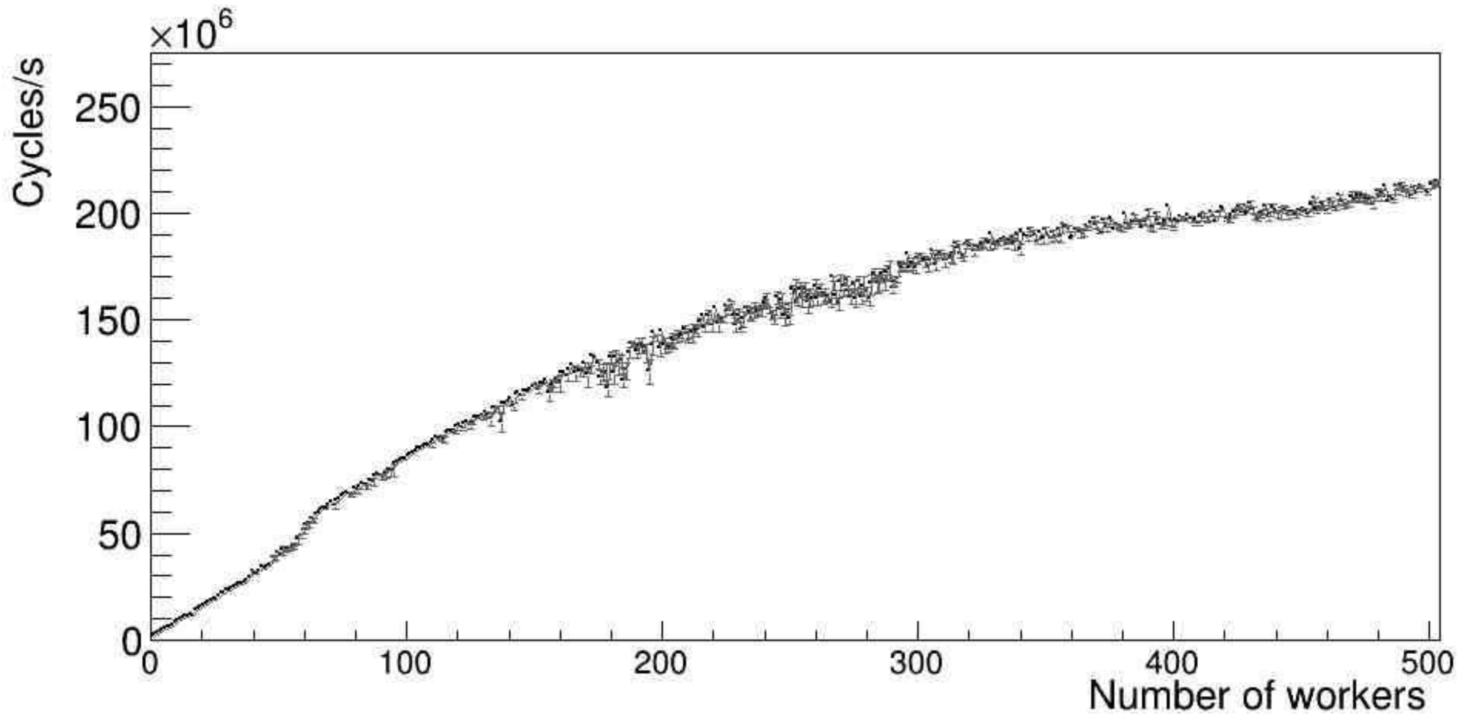
Super master is users' single point of entry. System complexity is hidden from users
Automatic data discovery via Xroot and job matching with local data
Can be optimize for data locality or high bandwidth data server access
Allows interactive analysis on large distributed datasets

PROOF/Xrootd Cluster

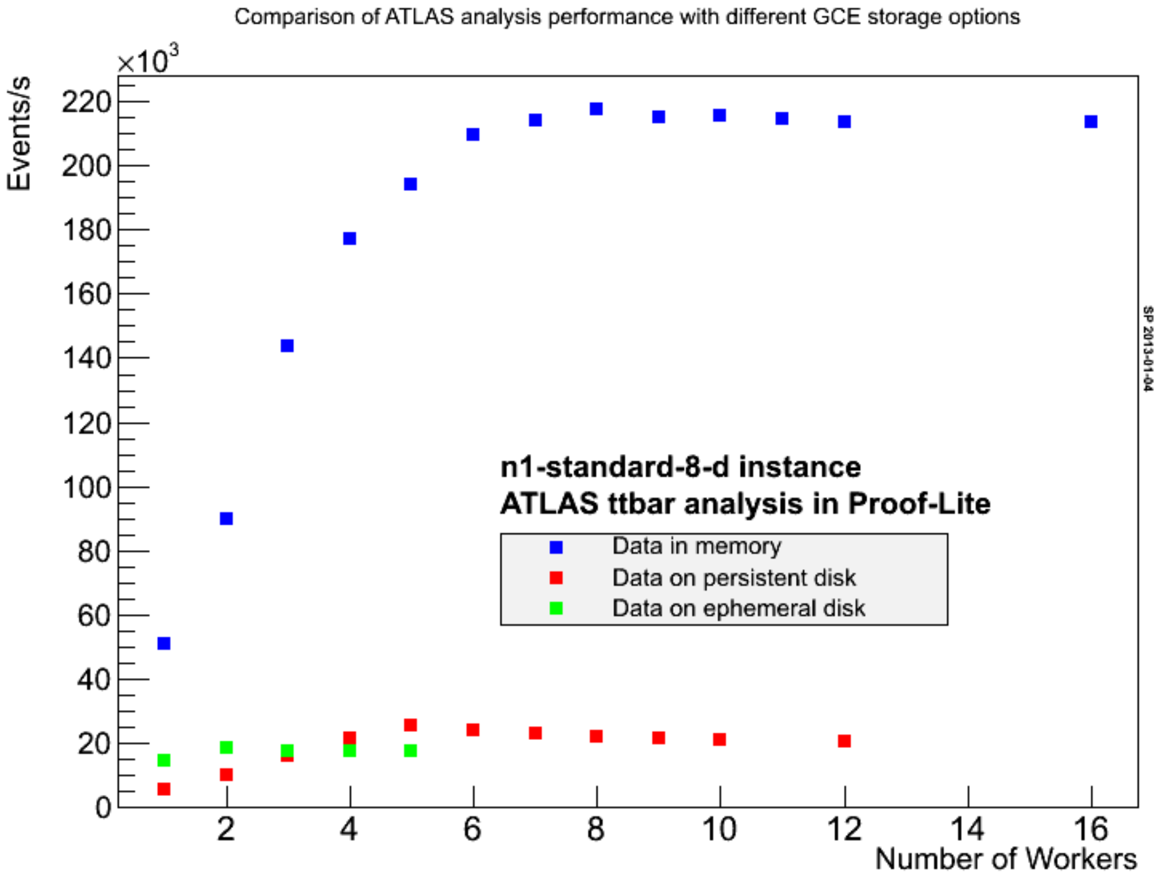


PROOF Tests on GCE

- Access to GCE allowed us to build and test large PROOF clusters, up to 1000 workers
- Figure shows scalability test for 500 worker PROOF cluster
  - n1-standard-8-d type instances



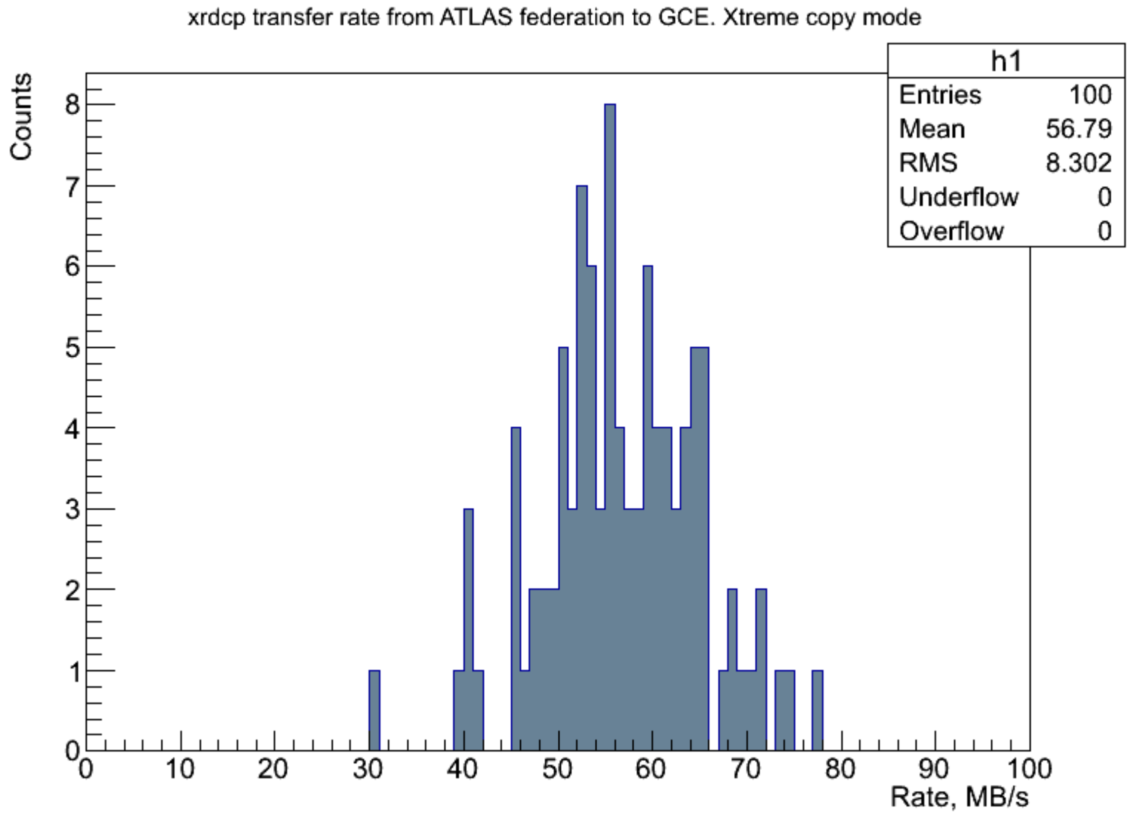
GCE storage performance comparison



Single PROOF node
Single disk
1 TB persistent
1.7 TB ephemeral

Persistent disk shows better scaling and peak performance.
Note that ephemeral disk has better single worker performance
RAID is needed for better performance

Data transfers from FAX to GCE



- Data transfer from Federated ATLAS Xroot (FAX) to GCE in multisource/multi-stream mode
- GCE Xroot cluster using ephemeral storage with 1.7 TB volumes per node
- Average transfer rate: 57 MB/s (single source xrdcp rate 40 MB/s)
- Note, this is over public networks

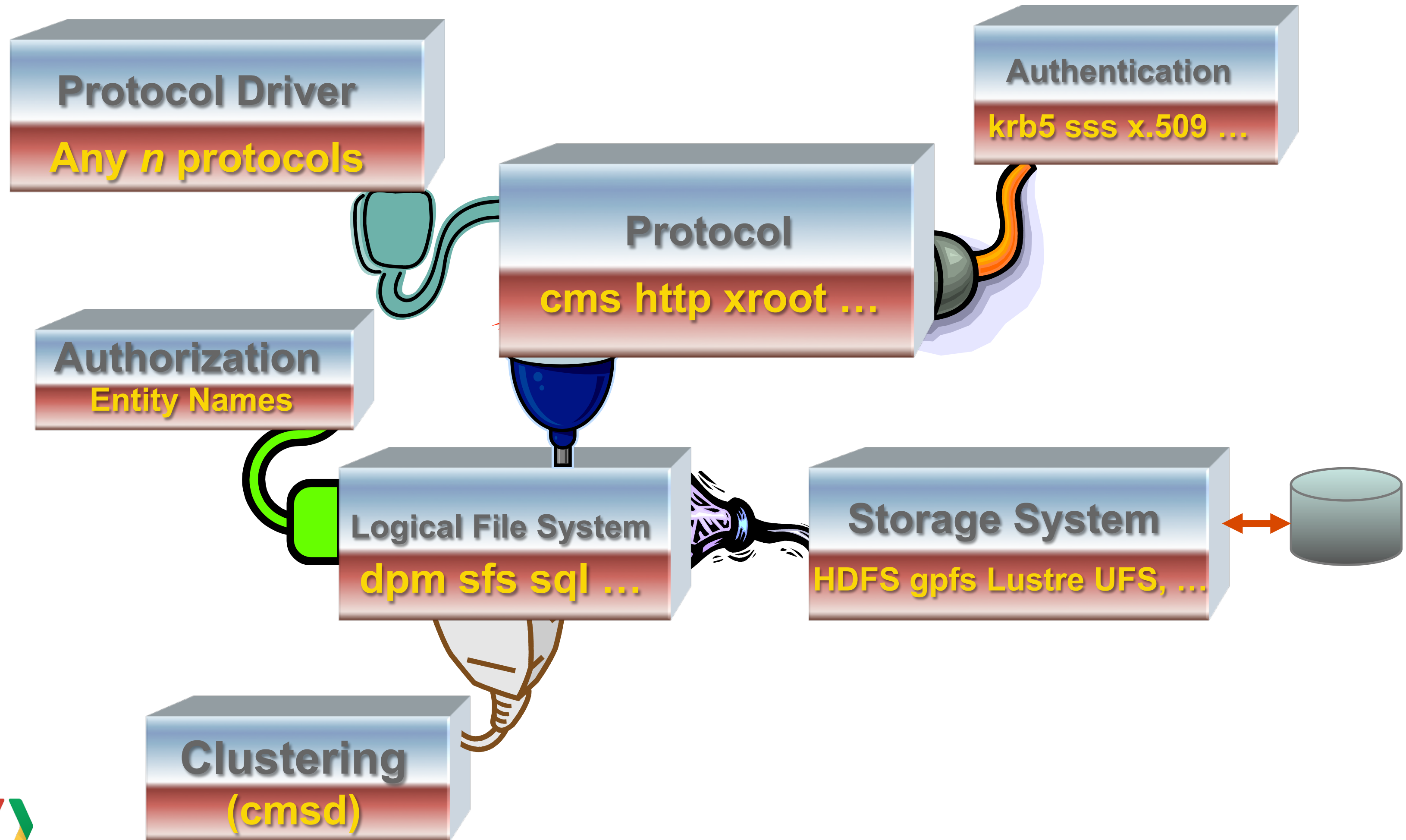
Xroot in more details

What Is **XRootD**

- A system for scalable cluster data access

- Not a file system
- Not *just* for file systems
- If you can write a plug-in you can cluster it

XRootD Plug-in Architecture

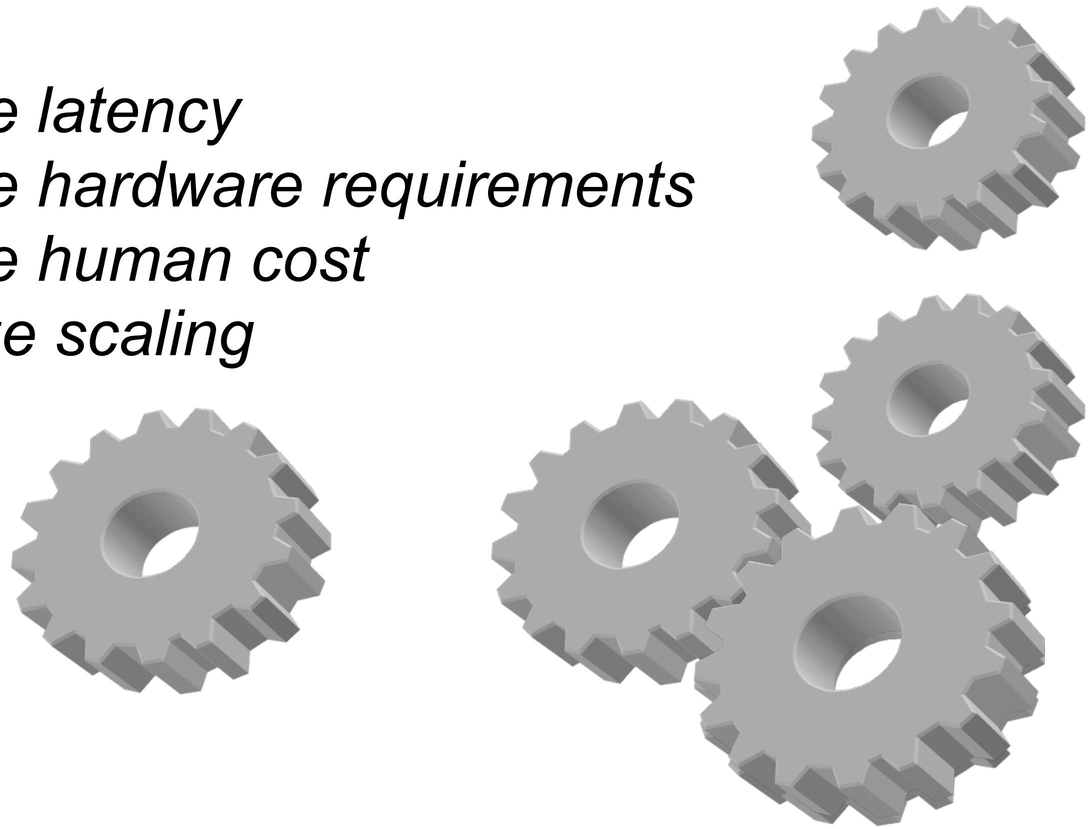


Data Access Problem

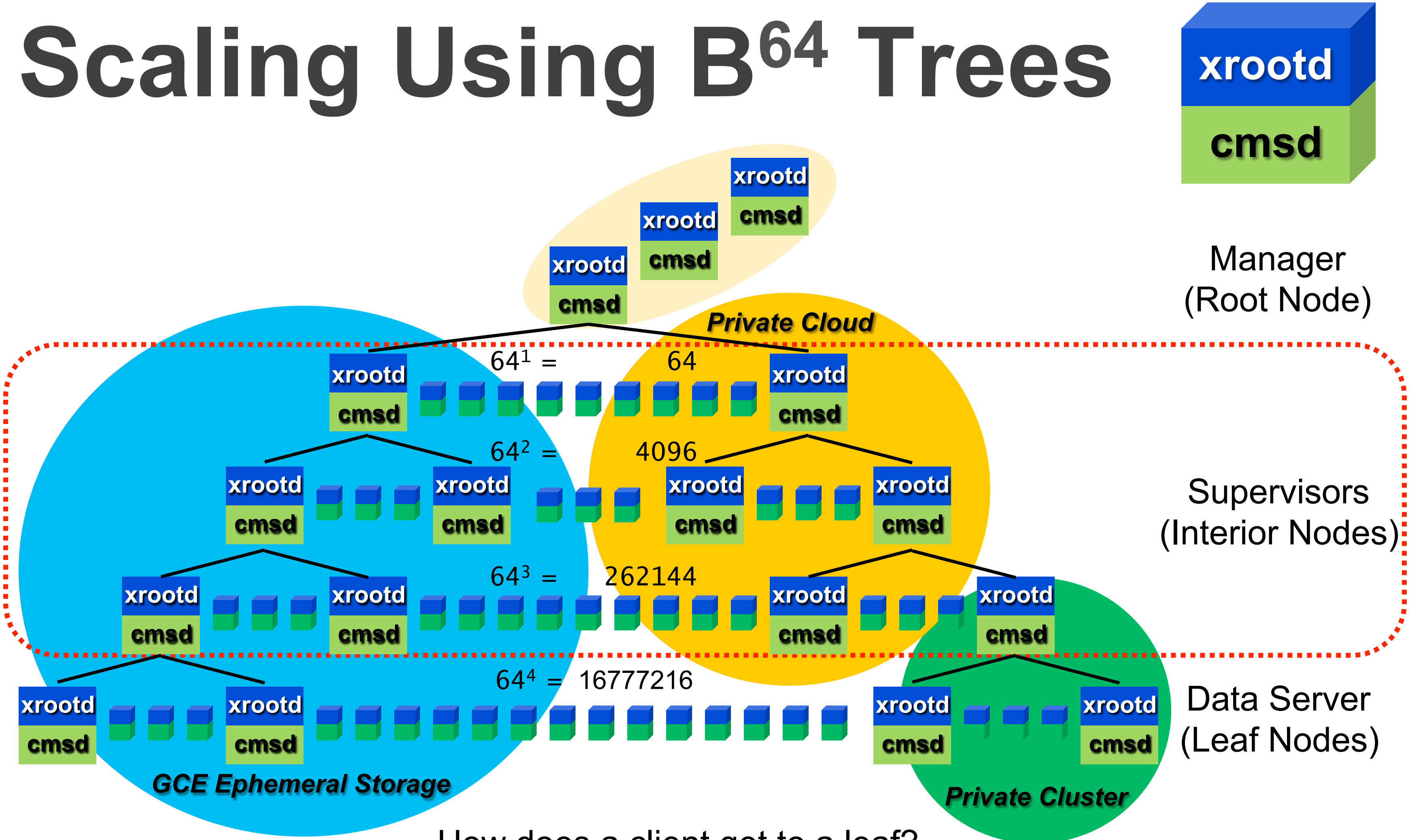
- The High Energy Physics analysis regime
  - Write once read many times access mode
  - Thousands of parallel batch jobs
  - Small block sparse random I/O

Synergistic Solution

Minimize latency
Minimize hardware requirements
Minimize human cost
Maximize scaling

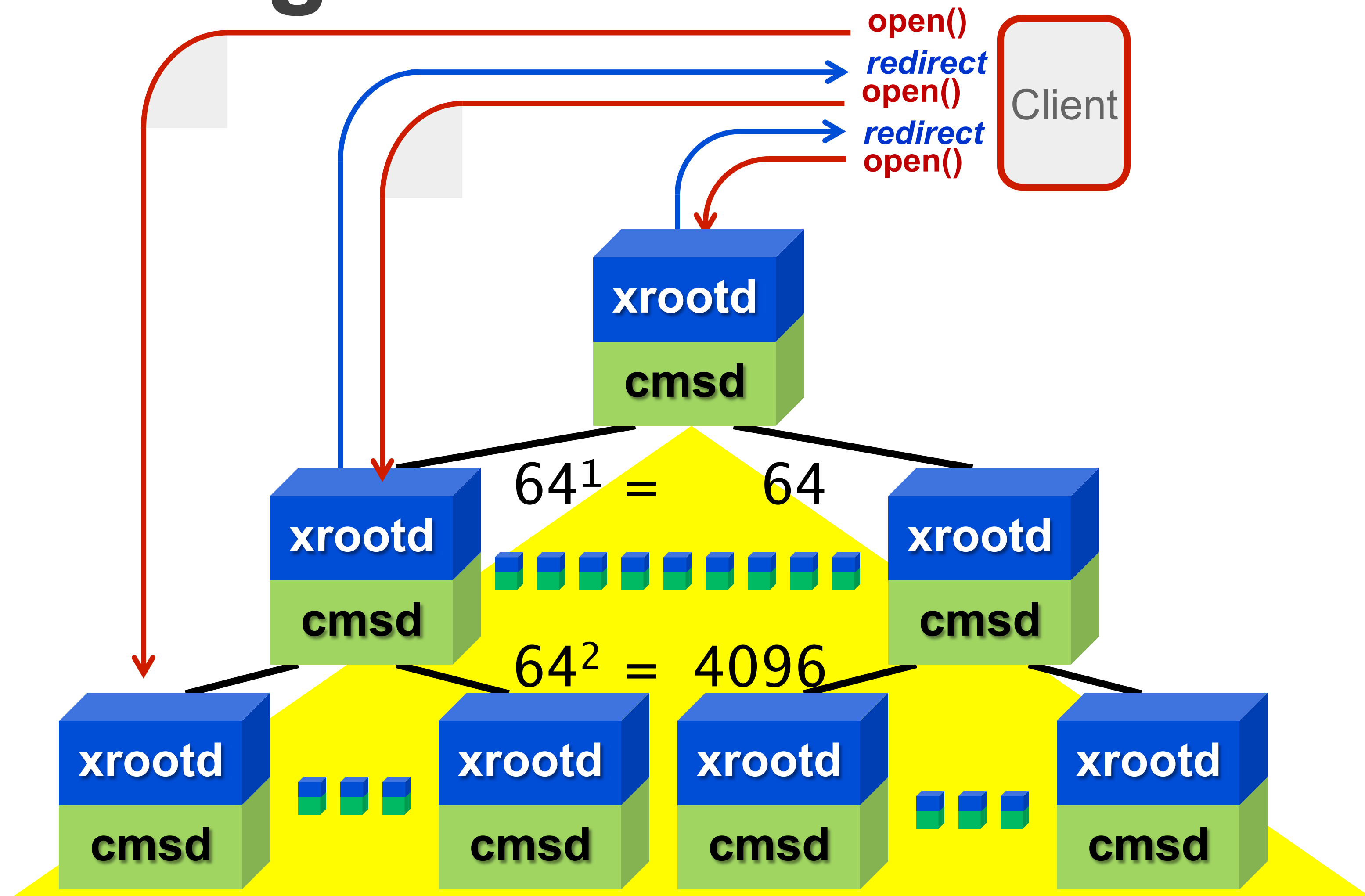


Scaling Using B^{64} Trees



How does a client get to a leaf?

Routing Clients To The Data



XRootD Bottom Line

- A simple, flexible, and effective system
  - But the devil is in the details
    - See "Scalla: Structured Cluster Architecture for Low Latency Access", IEEE IPDPSW 2012, Page(s): 1168 – 1175
- LGPL open-source
- Managed by the **XRootD** collaboration
  - SLAC, CERN, Duke, JINR, UCSD, & UNL (spring)
- Check out http://xrootd.org/

Summary

- Great experience with Google Compute Engine
- Tested several computational scenarios on GCE
  - PROOF clusters for data analysis
  - Xroot for cloud storage and federation
  - PanDA batch cluster for Monte Carlo Simulations
- Ran large scale Monte Carlo production on GCE
- We think that GCE is modern cloud infrastructure that can serve as a stable, high performance platform for scientific computing
- Tools developed by LHC community may be of interest for a broader community of developers working on GCE and other cloud platforms
  - Xroot, PROOF, ROOT, etc

Links to other HEP software

- Open source packages developed by High Energy Physics community
- ROOT - framework for data analysis, visualization, data persistency, etc
  - http://root.cern.ch/
- PROOF High performance implementation of MapReduce paradigm. http://root.cern.ch/drupal/content/proof