
CERN

Clouds and Science: OpenStack at CERN

Domenico Giordano (CERN)

1st ASTERICS-OBELICS Workshop

12-14 December 2016, Rome, Italy

H2020-Astronomy ESFRI and Research Infrastructure Cluster (Grant Agreement number: 653477).

CERN

International organisation close to Geneva, straddling Swiss-French border, founded 1954 by 12 Member States

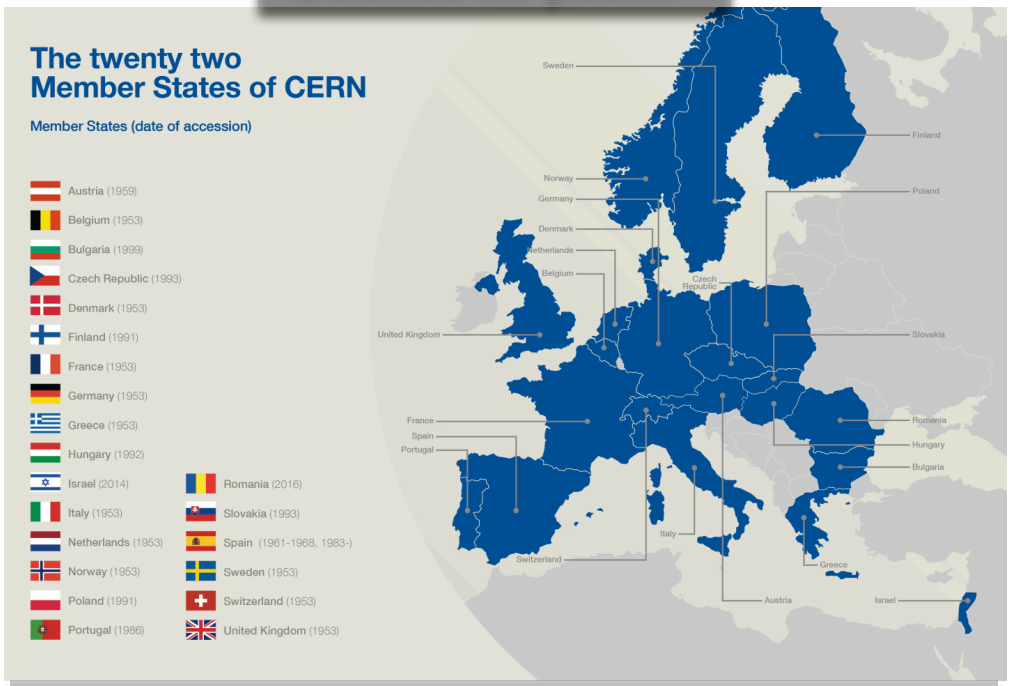
Facilities for fundamental research in particle physics

22 member states,
1 B CHF budget

~3'200 staff, fellows, apprentices,…

~13'100 associates

"Science for peace"

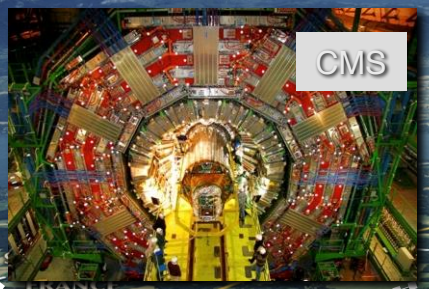


Associate member: Serbia, Cyprus, Turkey, Pakistan, Ukraine
Observers: European Commission, India, Japan, Russia, UNESCO, United States of America
Numerous **non-member states with collaboration agreements**

Large Hadron Collider & Detectors

ATLAS & CMS:
General Purpose experiments w/ pp and heavy ions
Discovery of new physics: Higgs, SuperSymmetry

pp, B-Physics, CP Violation
(matter-antimatter symmetry)



Exploration of a
new energy frontier
in p-p and Pb-Pb collisions

Heavy ions, pp
(state of matter
of early universe)

LHC ring:

- 27 km circumference
- Thousands of superconducting magnets (1.9 K)
- Beams Energy
  - Run 1 (2010-2013): 4 + 4 TeV
  - Run 2 (2015-2018): 6.5 + 6.5 TeV

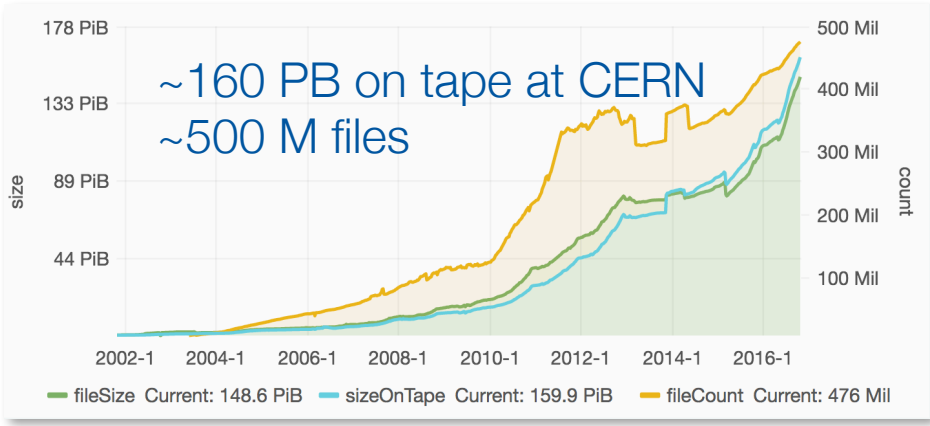
ATLAS experiment (as example)

- Collaboration: ~3'000 scientists
- Detector: 25 m diameter, 46 m length, 7'000 tons
  ~100 million electronic channels
- Data rate: 1 PB/s filtered down
  to 1-2 GB/s stored @ Tier-0

LHC Computing in 2016

LHC performance is above expectations

- 10.7 PB raw data recorded in July 2016
- June-Aug 2016 >500 TB / day

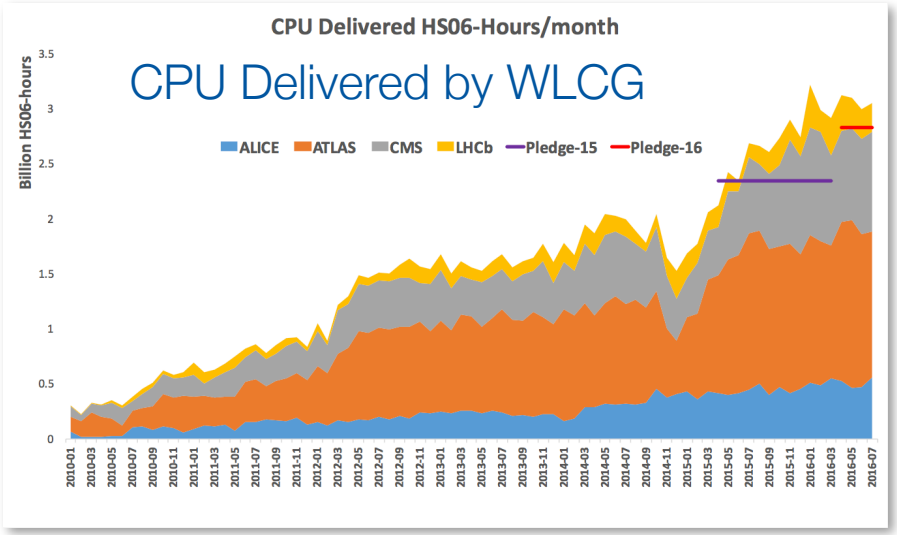


Data distribution to WLCG sites

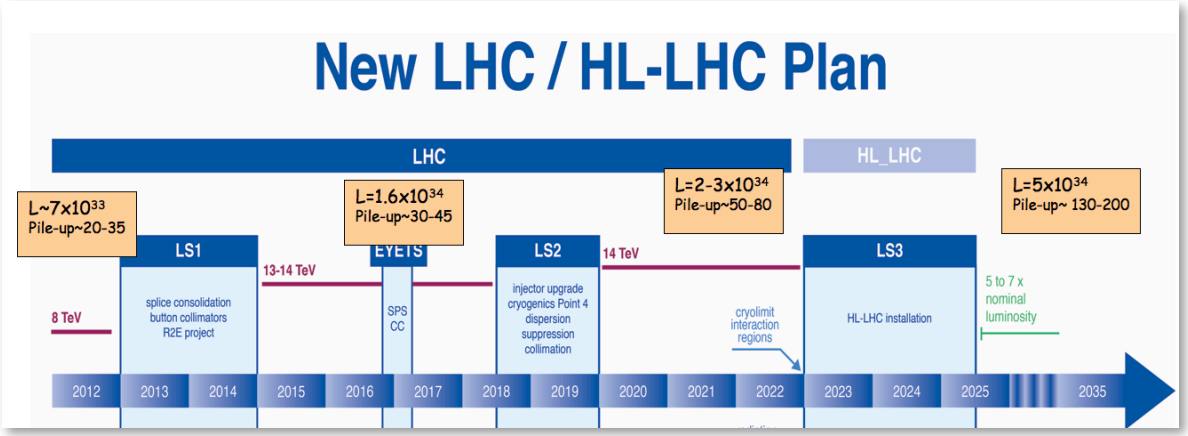
- Distributed >80 PB/month, rate > 50 GB/s
- GEANT has deployed additional capacity for LHC

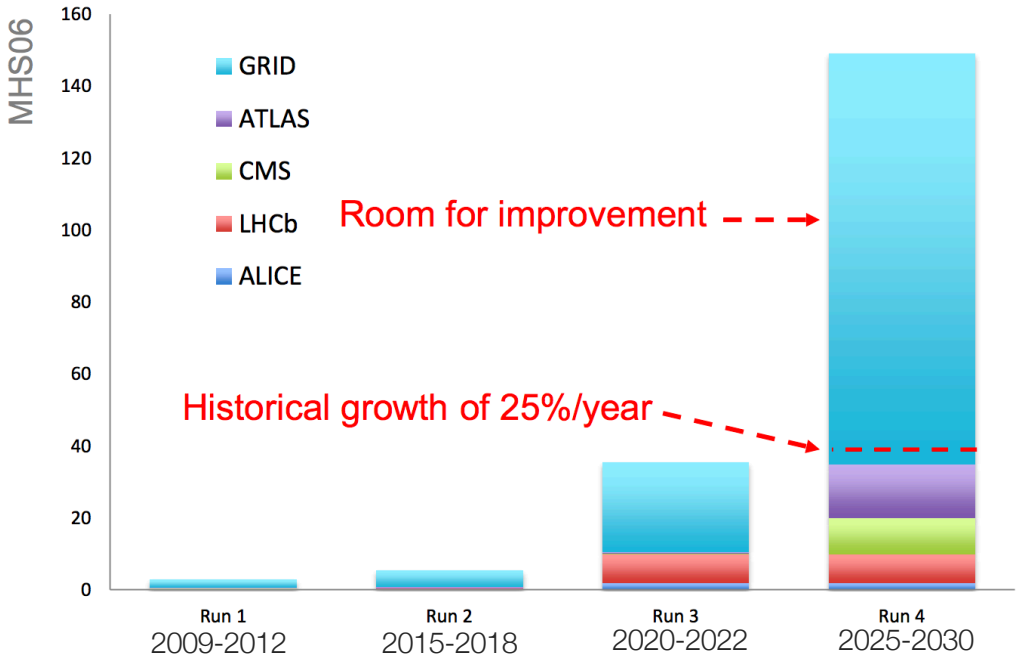
Compute resources (CPUs) are fully exploited by the Collaborations

- Systematically exceeding those formally pledged



Compute Growth Outlook





Data:

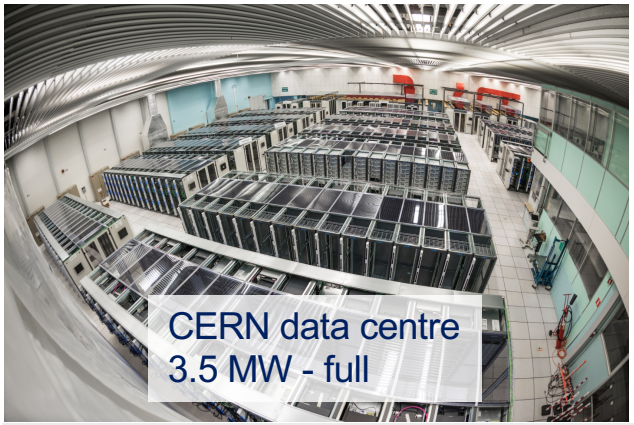
- Raw 2016: 50 PB → 2027: 600 PB
- Derived: 2016: 80 PB → 2027: 900 PB

CPU:

- x60 respect to 2016 resources

At least x10 above what is realistic to expect from technology with reasonably constant cost.

Tier-0: 20% of WLCG resources



MEYRIN DATA CENTRE

	last_value
Number of Cores in Meyrin	168,452
Number of Drives in Meyrin	85,729
Number of 10G NIC in Meyrin	11,153
Number of 1G NIC in Meyrin	23,452
Number of Processors in Meyrin	26,887
Number of Servers in Meyrin	14,245
Total Disk Space in Meyrin (TB)	165,848
Total Memory Capacity in Meyrin (TB)	723

WIGNER DATA CENTRE

	last_value
Number of Cores in Wigner	56,000
Number of Drives in Wigner	29,696
Number of 10G NIC in Wigner	2,981
Numer of 1G NIC in Wigner	6,579
Number of Processors in Wigner	7,002
Number of Servers in Wigner	3,504
Total Disk Space in Wigner (TB)	97,324
Total Memory Capacity in Wigner (TB)	221

NETWORK AND STORAGE

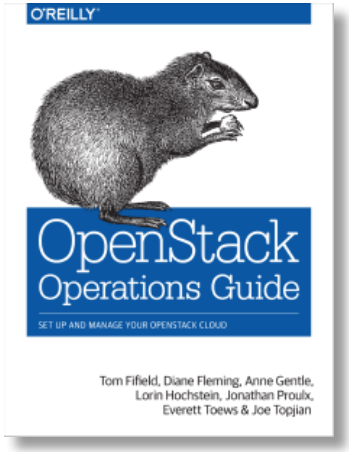
	last_value
Tape Drives	104
Tape Cartridges	22,437
Data Volume on Tape (TB)	184,774
Free Space on Tape (TB)	26,104
Routers (GPN)	145
Routers (TN)	30
Routers (Others)	103
Switches	3,713

Time to rethink

How can we avoid the sustainability trap ?
How can we learn from others and share ?

New cultural approach

- CERN computing need are common to other domains

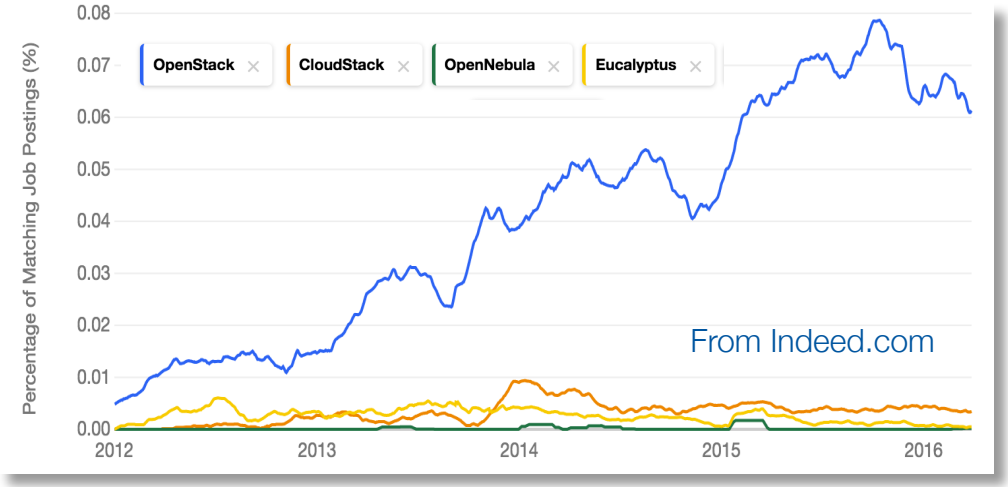


Tooling choice

- Technical requirements
- Open source with supporting community ecosystem
- Helps to train new staff

Job trend considerations

- Attract talents and return value to the funding member states
- Build skills that help for finding future job opportunities



OpenStack Community

One of the largest open source communities

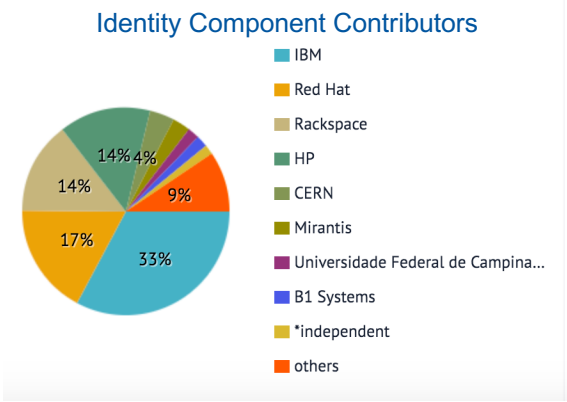
- 2,300+ developers contributed code at the last release

Technical committee guiding development direction, elected by the contributors

User committee covering working groups such as Telecom NFV, Large Deployments and Scientific Working group

Community Collaboration

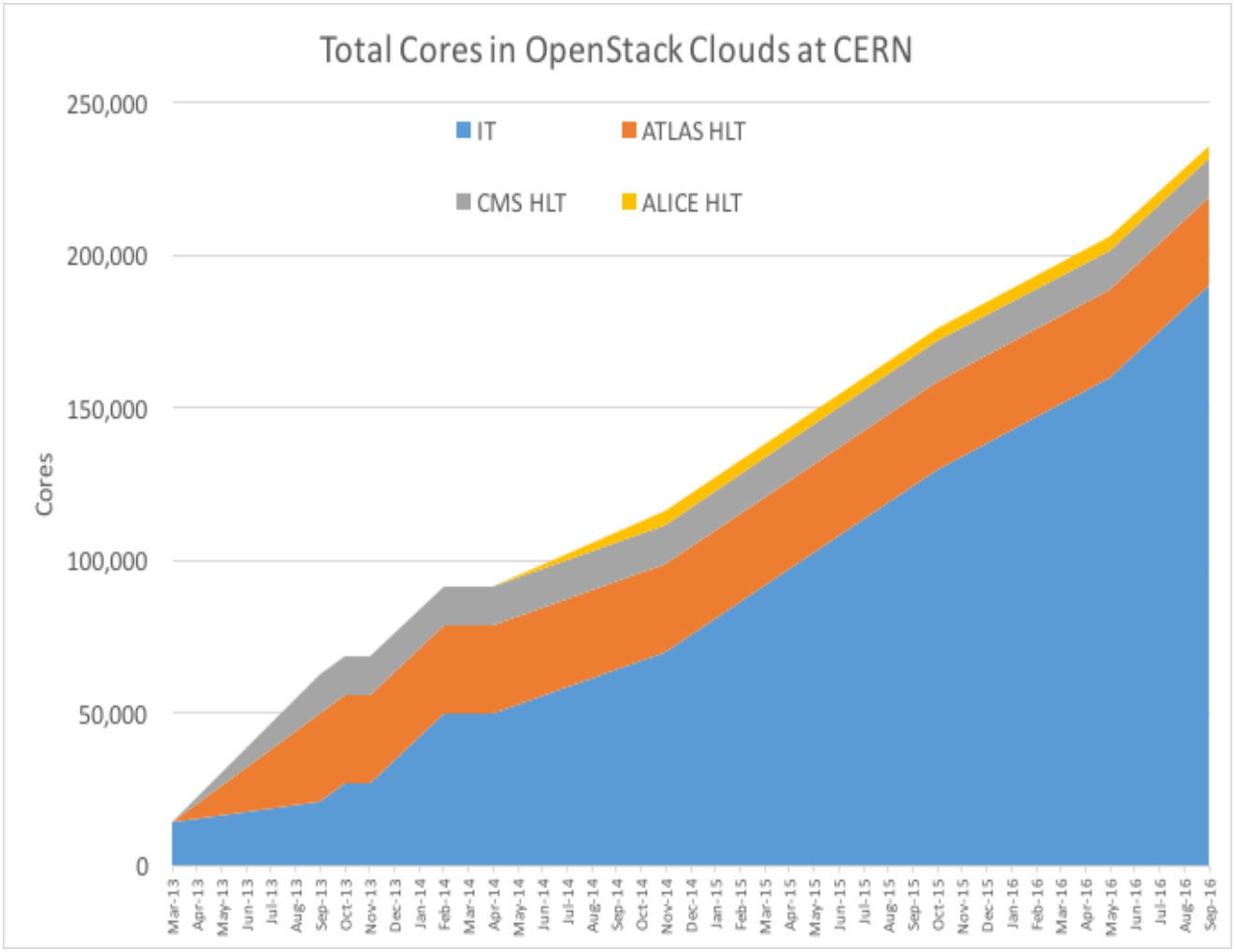
- Open source collaboration sets model for in-house teams
- Reviews and being reviewed is a constant learning experience
- External recognition by the community is highly rewarding for contributors



Keystone 2013

Paris 2014
Superuser Award Winner

OpenStack@CERN Status



In production:

- 4 clouds
- >220K cores
- >9,000 hypervisors

~100,000 additional cores being installed in next 6 months

90% of CERN's compute resources are now delivered on top of OpenStack

Cloud Infrastructure by numbers

2 Data Centres
- 44 compute cells
- 5 availability zones

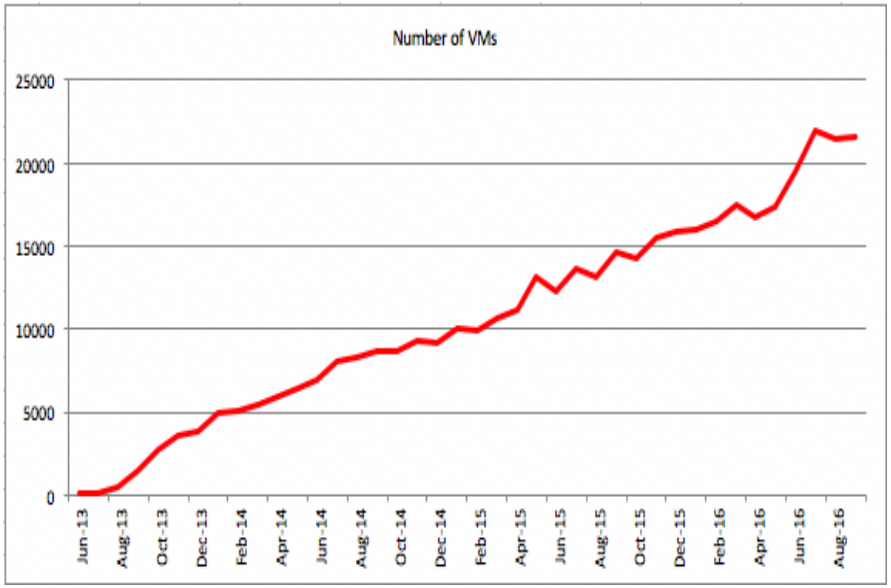
~ 22000 VMs running

~ 7000 Compute Nodes (193k cores)
- KVM
- Hyper-V

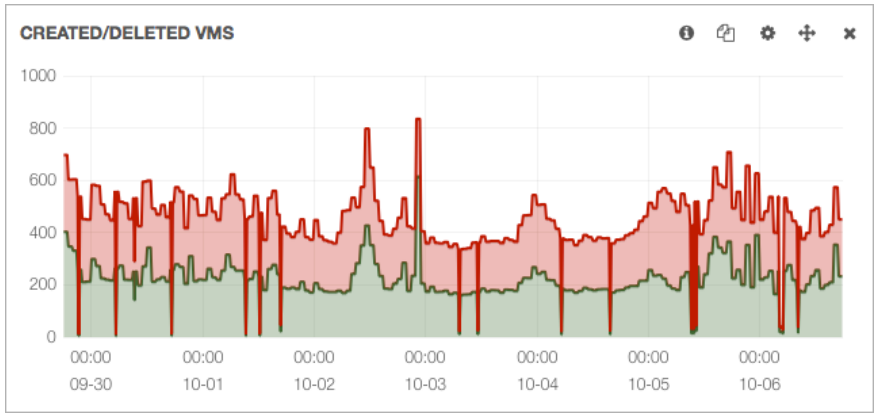
~ 3400 Images (~ 50 TB in use)

~ 2800 Volumes (~ 1 PB allocated)

~ 2400 Users
~ 2800 Projects

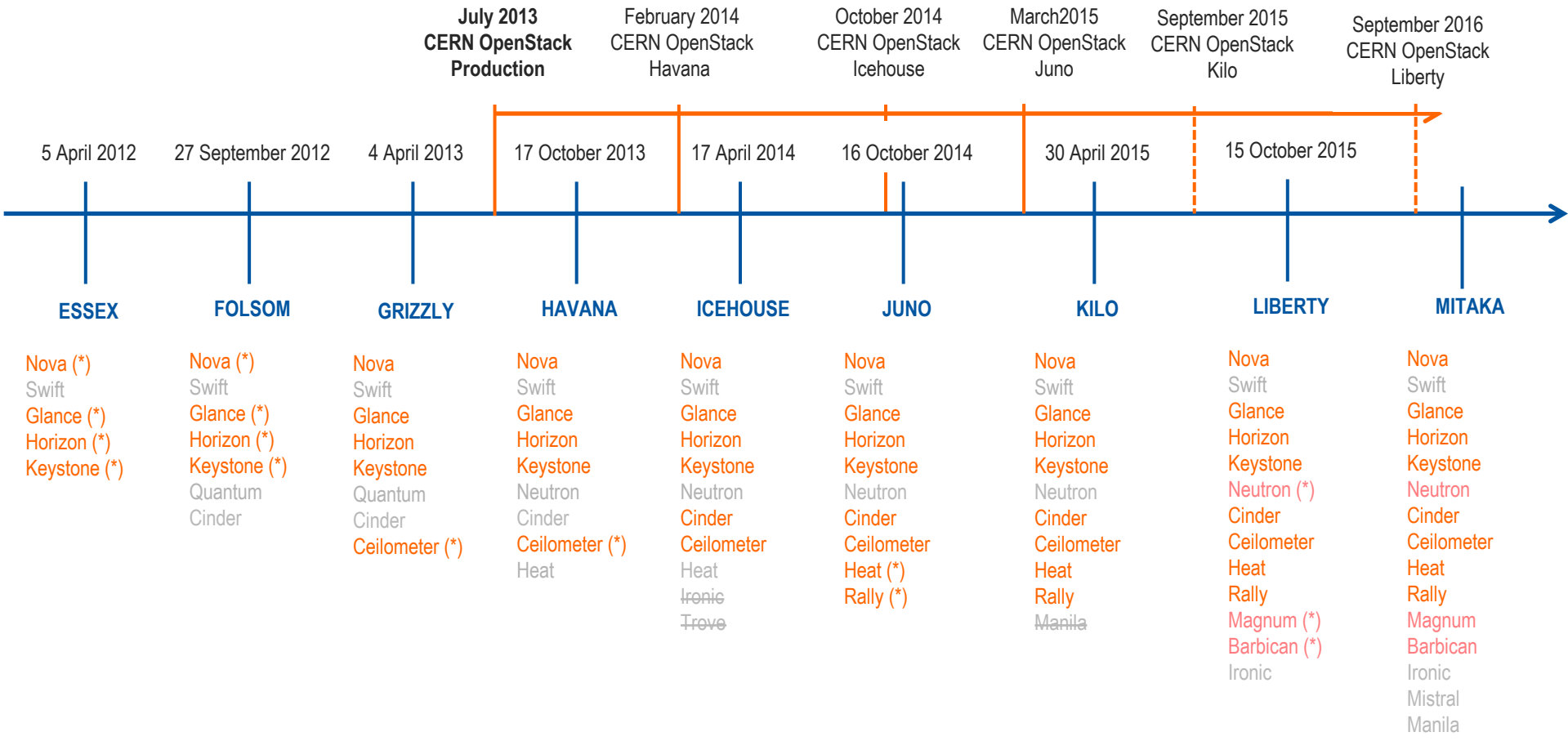


Number of VMs running



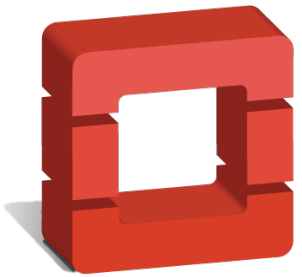
Number of VMs created (green) and VMs deleted (red) every 30 minutes

CERN OpenStack Project

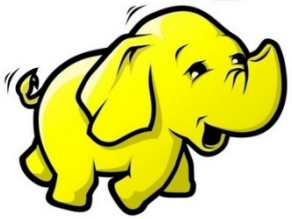


(*) Pilot Trial

CERN Tool Chain



elasticsearch.



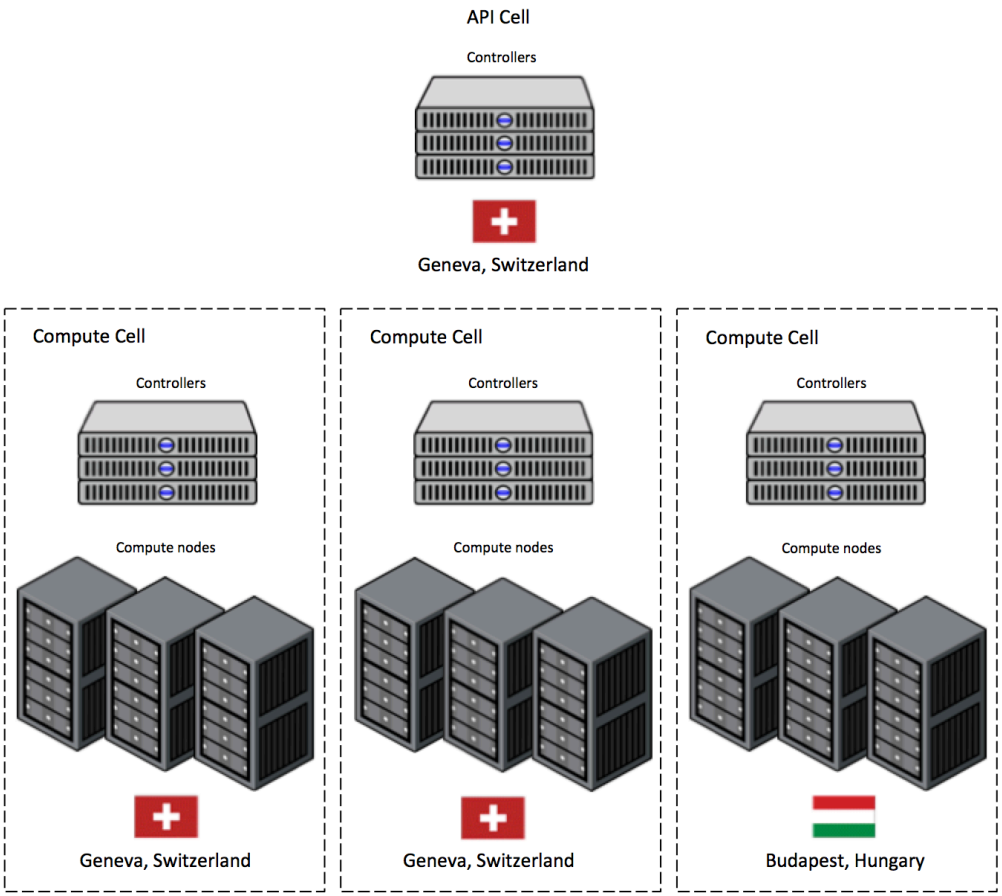
FOREMAN

Jenkins

Nova Cells

Top level cell
- Runs API service
- Top cell scheduler

Child cells run
- Compute nodes
- Scheduler
- Conductor
- ~40 cells



Performance optimizations

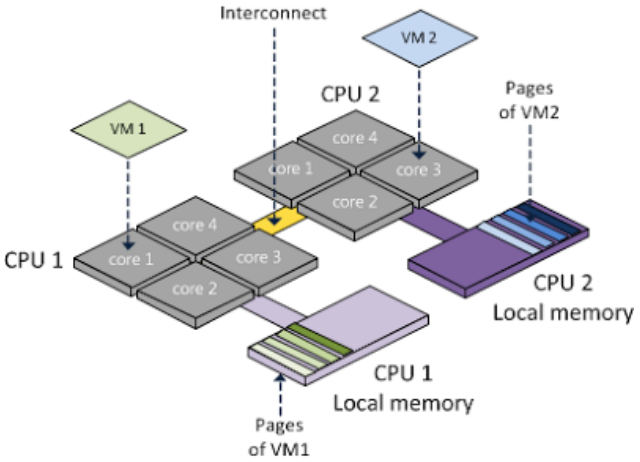
Developed an optimized configuration

- NUMA-aware VM scheduling, 2MB huge page

Virtualization overhead pushed below 5%

Rollout in all batch dedicated cells

- ~2000 compute nodes to upgraded
- Batch team recreated 6k VMs



[Frank Denneman]

Continue to investigate Memory and IO performance in cloud environments

Service Growth and Operations

Enabled Federated access to CERN Cloud

- EduGain credentials, Indigo WP3 team

Better resource utilization

- Expire Personal VMs
- Investigate how to run opportunistic workloads in temporary unused resources

Introducing new OpenStack components

- Manila (File Share as a Service), as a potential evolution of the Filer service
- Ironic (Bare Metal as a Service), to deploy physical servers likes VMs

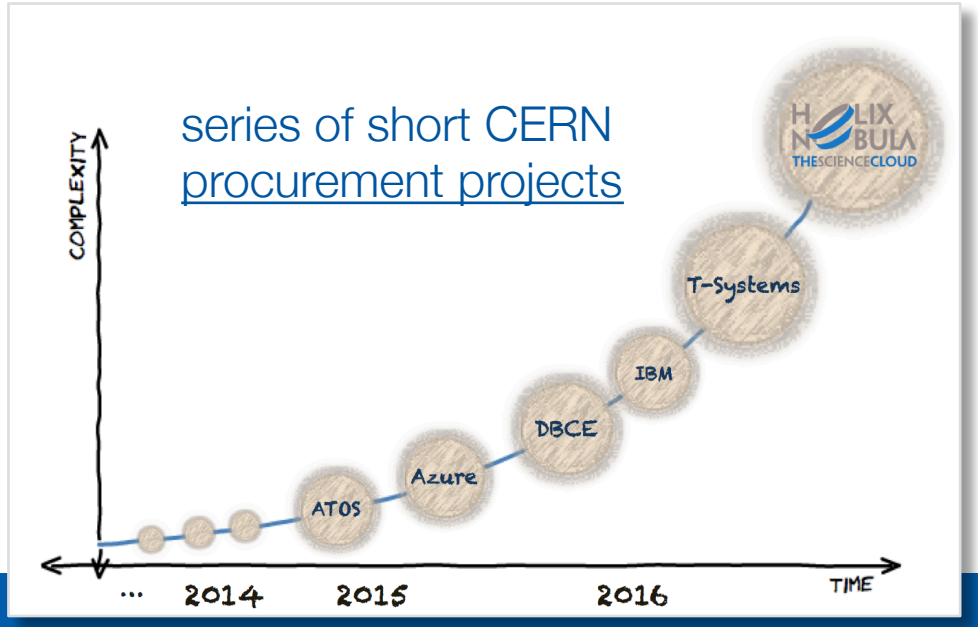
Expansion Options

Options to improve Meyrin CC are limited

Wigner contract will end in 2019

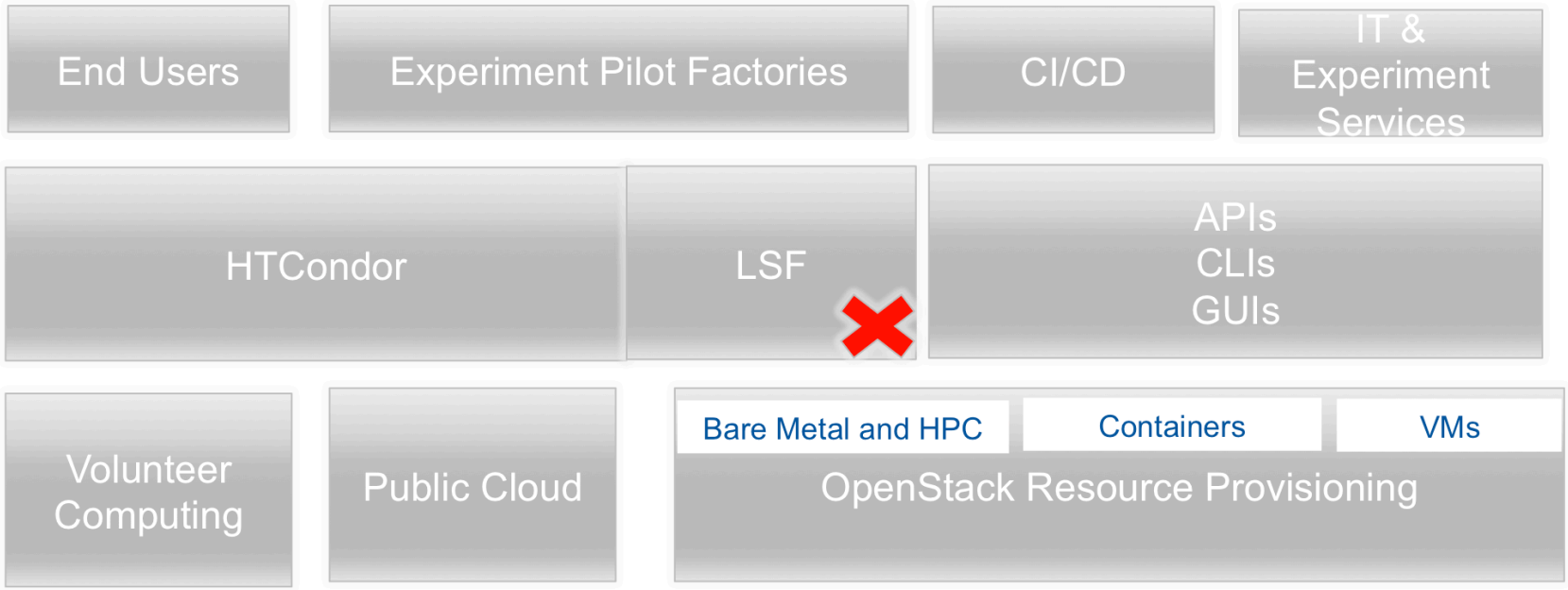
What next? On-premises Vs Hosted Vs Cloud resources

- Find a replacement centre like Wigner ?
- Public cloud reserved instance procurements ?
- Long term leasing of bare metal ?
- Spot?
- Combined procurement tests ongoing with HNSciCloud
  - See Bob Jones' talk on Wednesday: **H2020- HNSciCloud**



Tier-0 Compute Services 2017

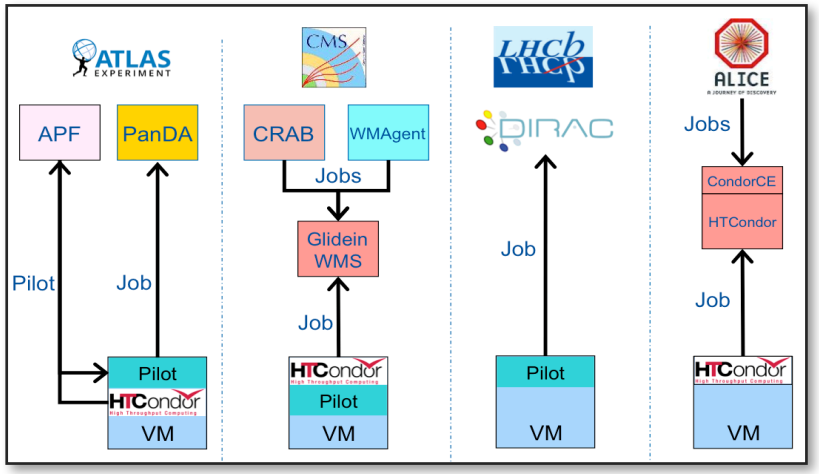
- Universal resource provisioning layer for bare metal, containers and VMs
- HTCondor as the single end user interface with LSF retirement by LS2
- Continue investing in automation and other communities for scaling with fixed staff
- Self service for end users within the policies and allocations



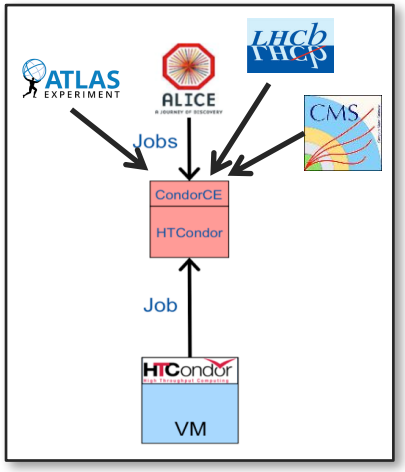
Transparent Extension in External Clouds

- Manage and exploit external resources using same toolset and entry points as CERN on premises resources
  - ***Puppet*** configuration
  - **HTCondor** for scheduling and match-making
  - Infrastructure **monitoring**
- Adopted ***Terraform*** for VM lifecycle management
  - Open source toolkit, supports several cloud providers

Evaluated approaches

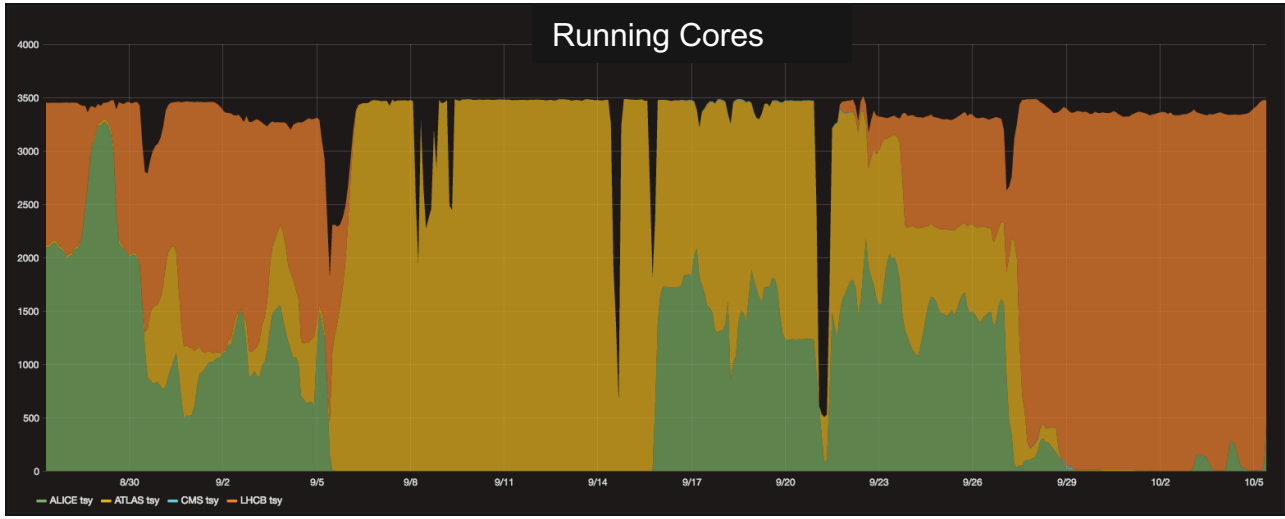


Selected approach



Recent activity: T-Systems

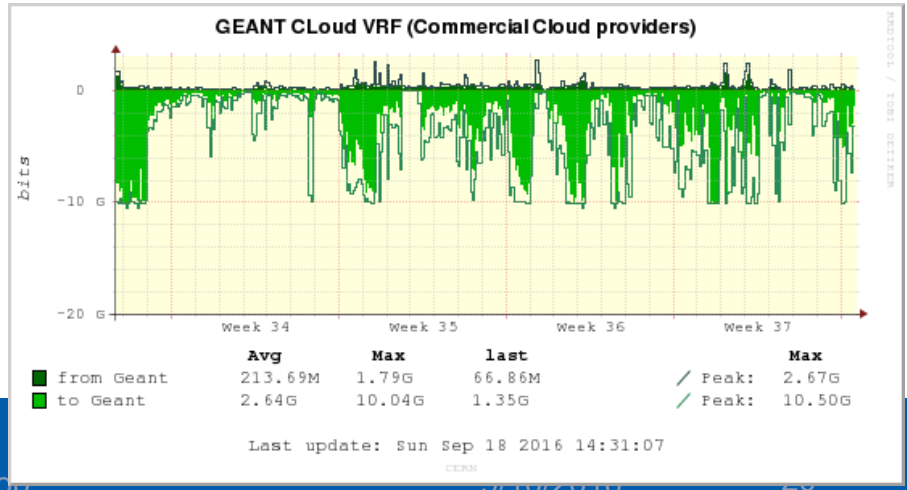
- Batch resources fully loaded
  - shared among VOs



- Mixture of "CPU-intensive" and "network-intensive" tasks
  - MC workloads easier to manage

	Max	Avg
LHCB tsy	99.05	85.04
ALICE tsy	93.83	75.98
ATLAS tsy	100.00	64.13

- WAN largely used
  - Sometimes even saturated

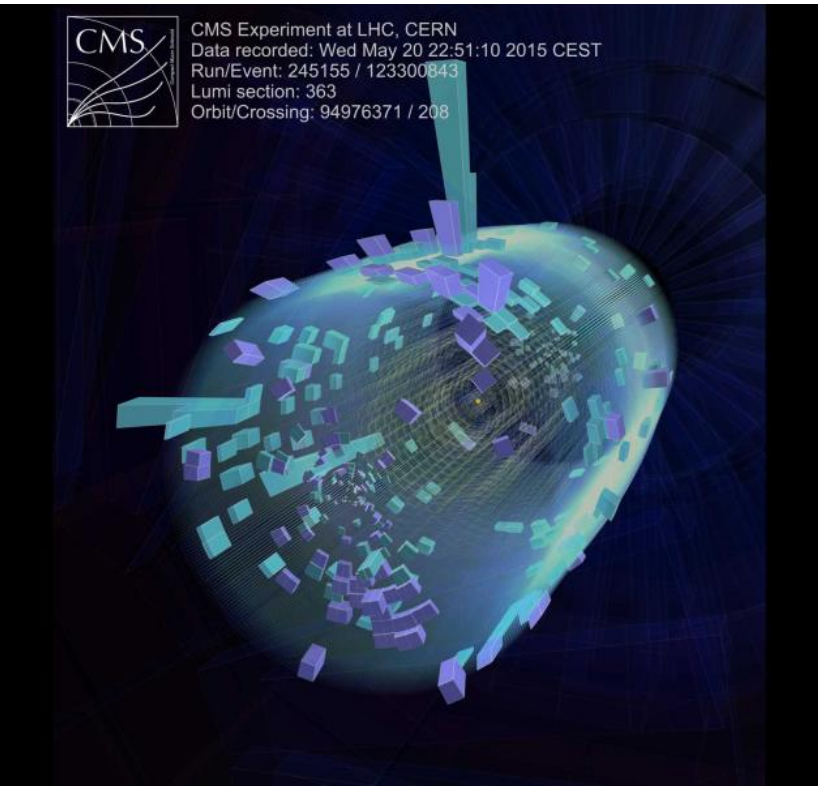


Summary

- OpenStack at CERN has been in production for 3 years to deliver LHC compute capacity
- Major cultural and technology changes have been successfully addressed
- Contributing back upstream has led to sustainable tools and effective technology transfer

This transformation would not have been possible without the OpenStack community

For Further Information



Technical details at http://openstack-in-production.blogspot.fr

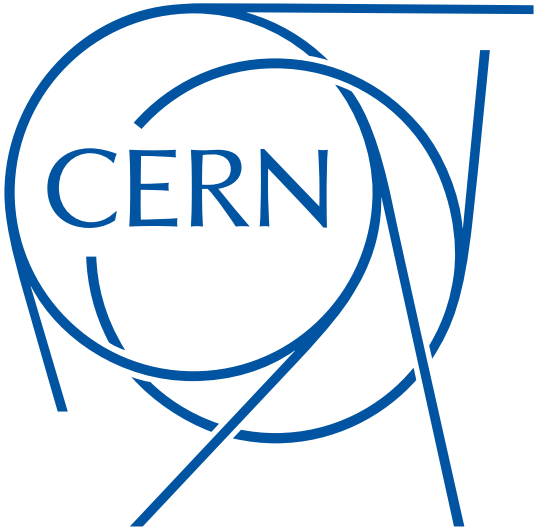
Scientific Working Group at https://wiki.openstack.org/wiki/Scientific_working_group

CERN Containers at https://indico.cern.ch/event/506245/

Other links at http://clouddocs.web.cern.ch/clouddocs/additional/index.html

CERN tools at http://github.com/cernops

Thank to
H2020-Astronomy ESFRI and Research Infrastructure Cluster (Grant Agreement number: 653477)
for the travel and accommodation support to participate to this Workshop

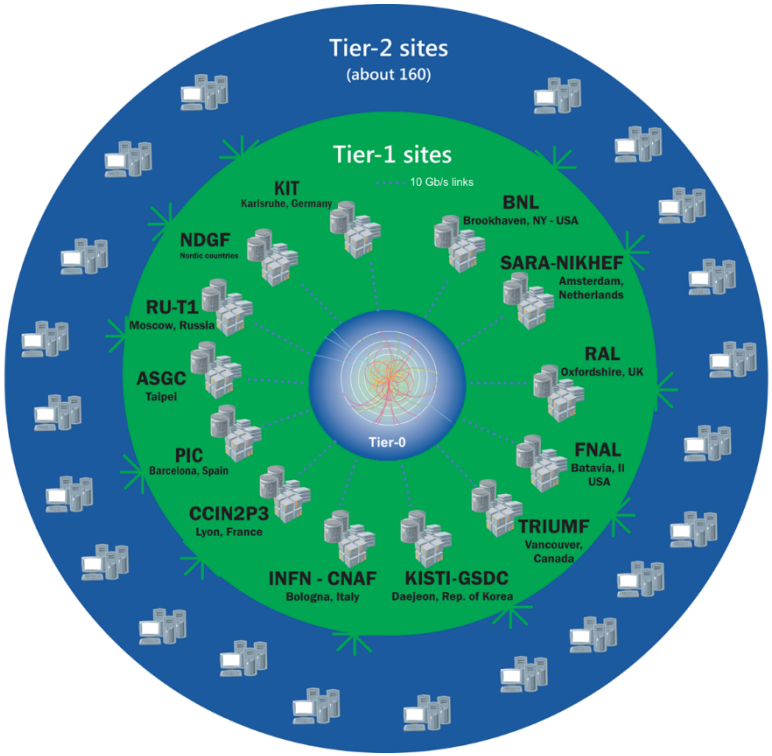

CERN

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



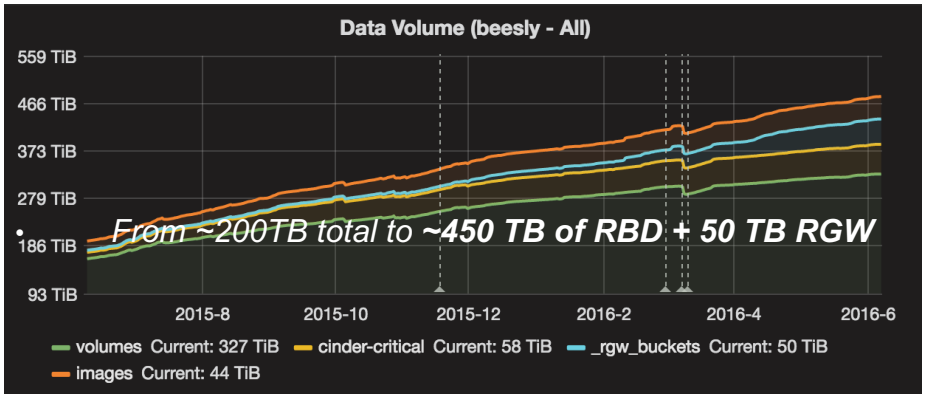
nearly 170 sites, 40 countries

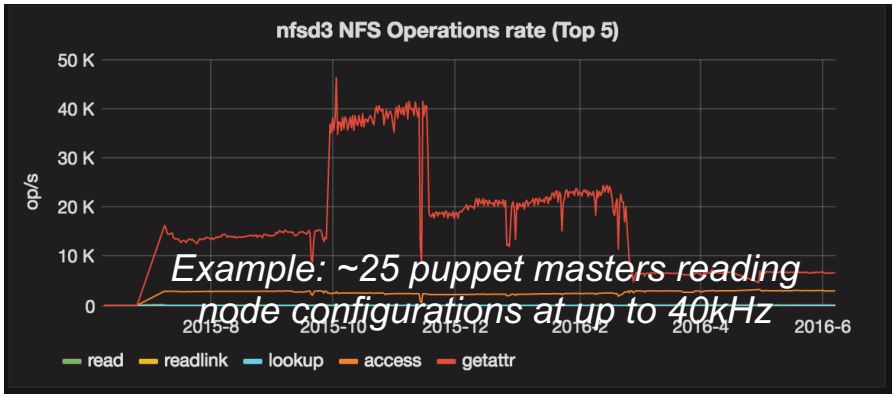
~350'000 cores

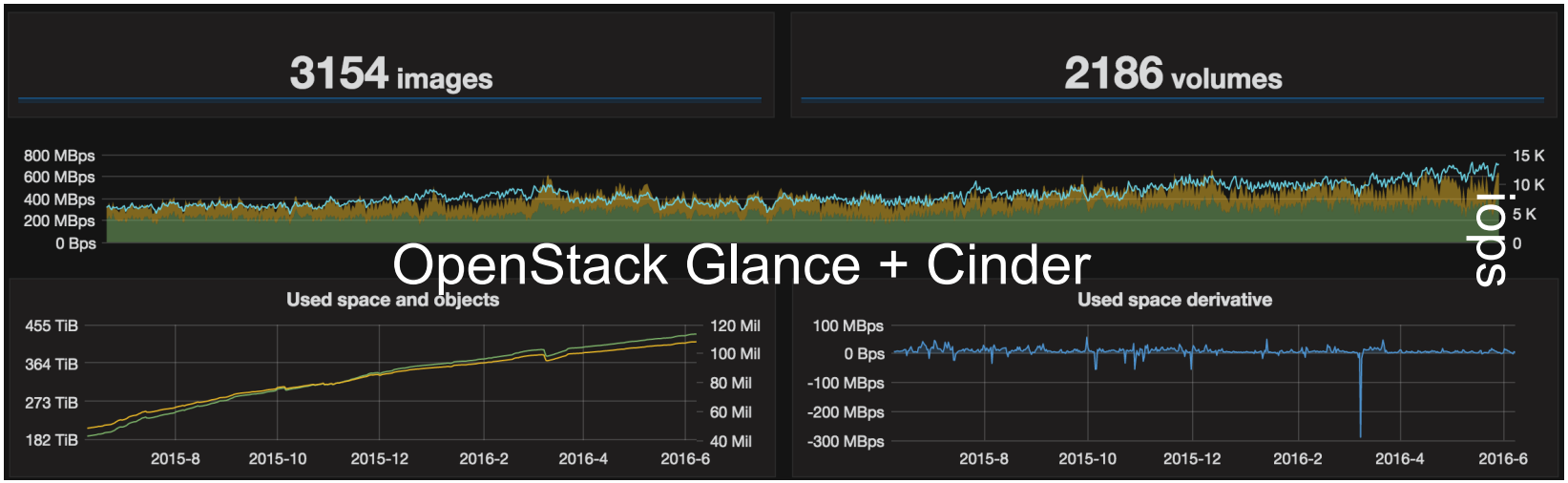
500 PB of storage

> 2 million jobs/day

10-100 Gb links







CephFS with Manila is now in pilot phase for cluster filesystems

Public Cloud Tests

Tests for 2-3 months on various European public resources