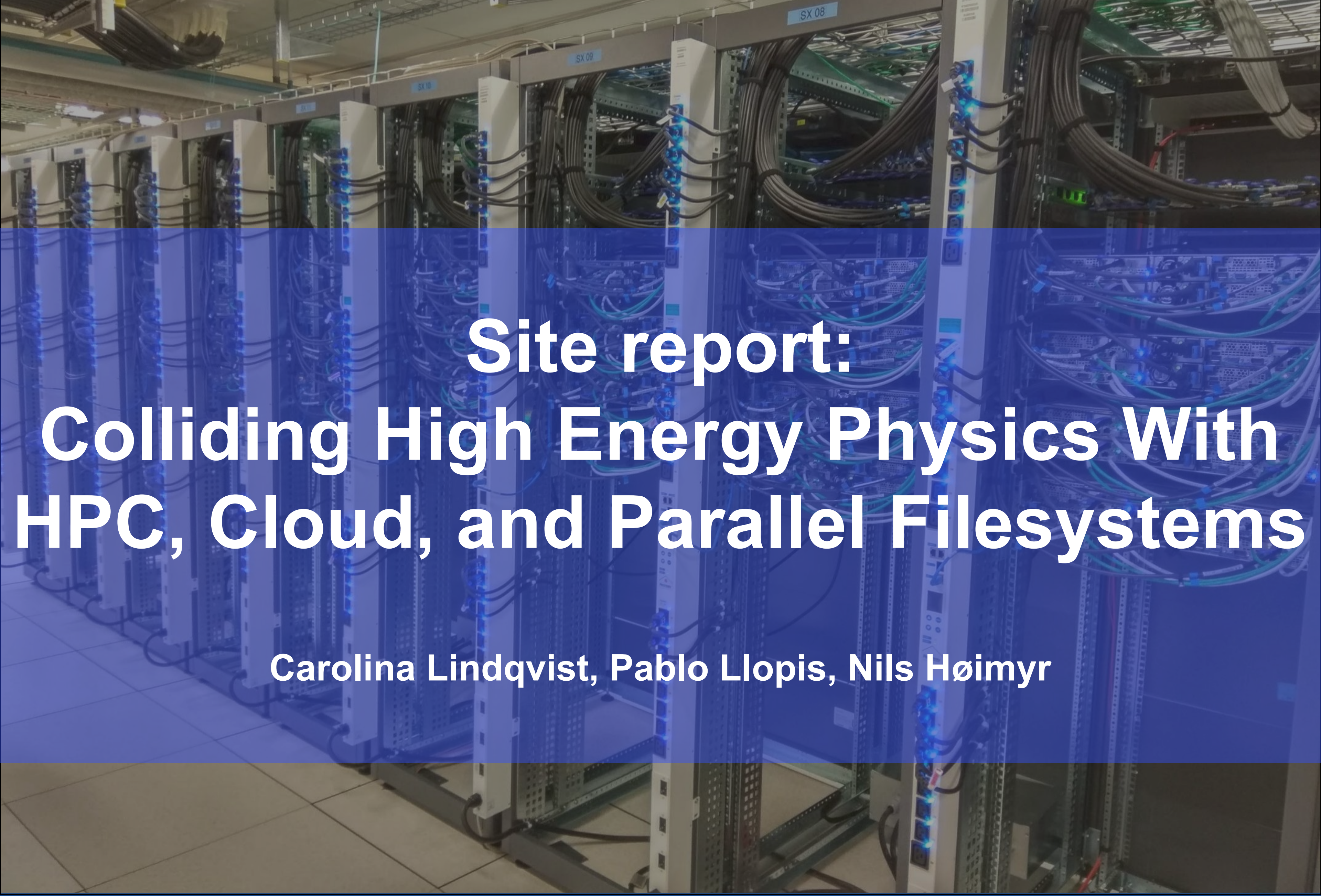
CERN



Site report: Colliding High Energy Physics With HPC, Cloud, and Parallel Filesystems

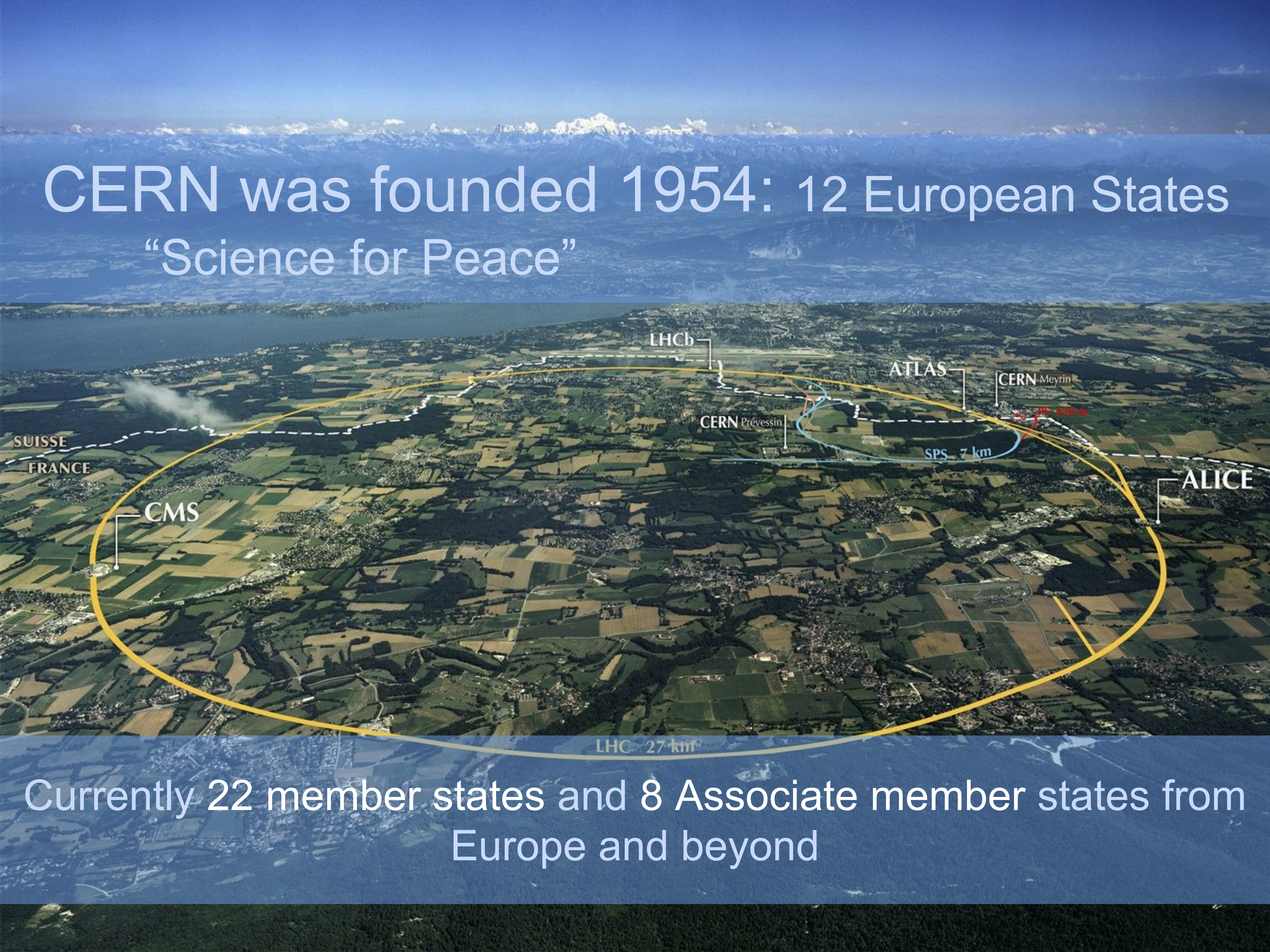
Carolina Lindqvist, Pablo Llopis, Nils Høimyr

Agenda

- What is CERN ?
- The CERN IT agile environment
- HPC at CERN
- How we use SLURM
- Future work, our plans for our HPC infrastructure

CERN was founded 1954: 12 European States

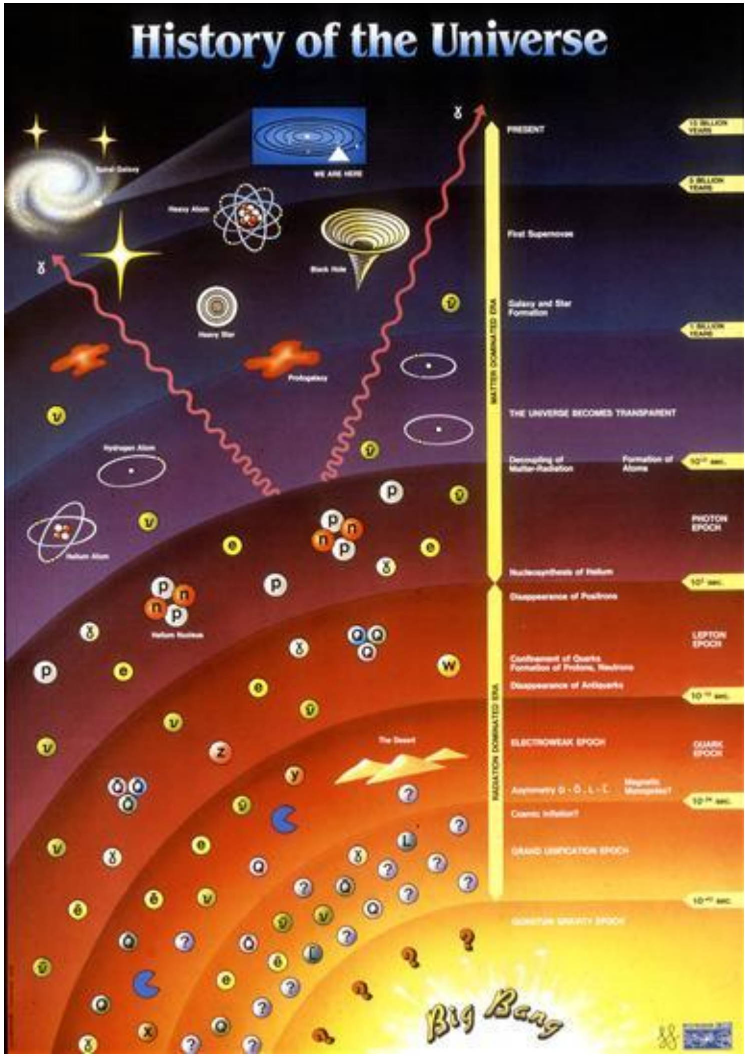
"Science for Peace"



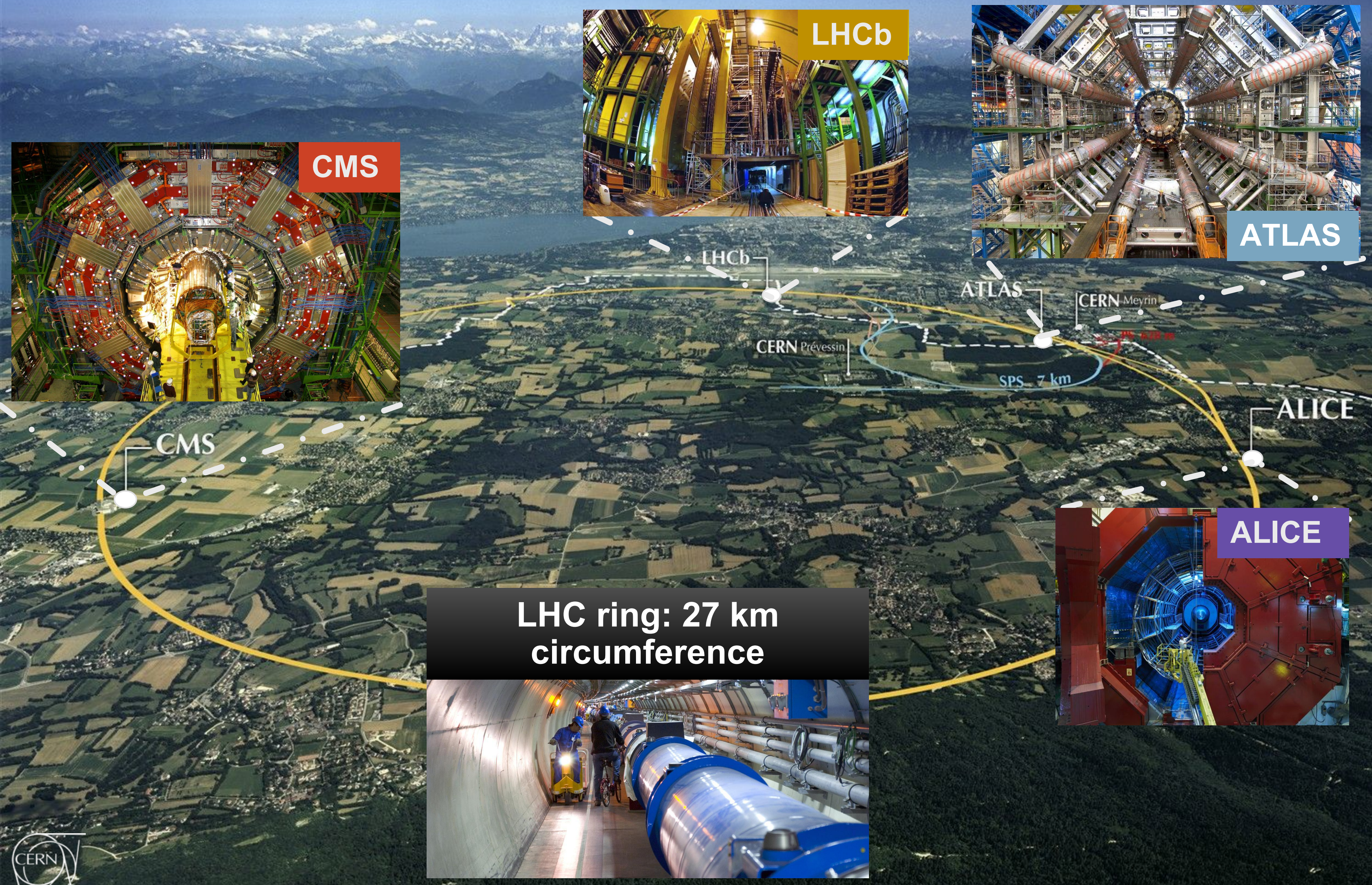
Currently 22 member states and 8 Associate member states from Europe and beyond

The mission of CERN

Probing the fundamental structure of the universe using the world's largest and most complex scientific instruments to study the basic constituents of matter – the fundamental particles.

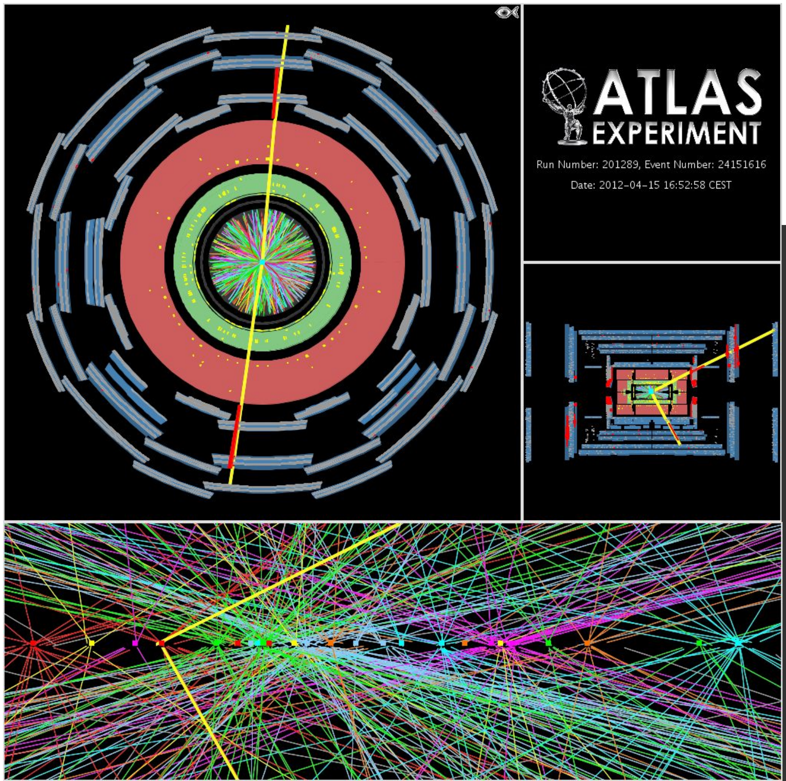


LHC accelerator and detectors

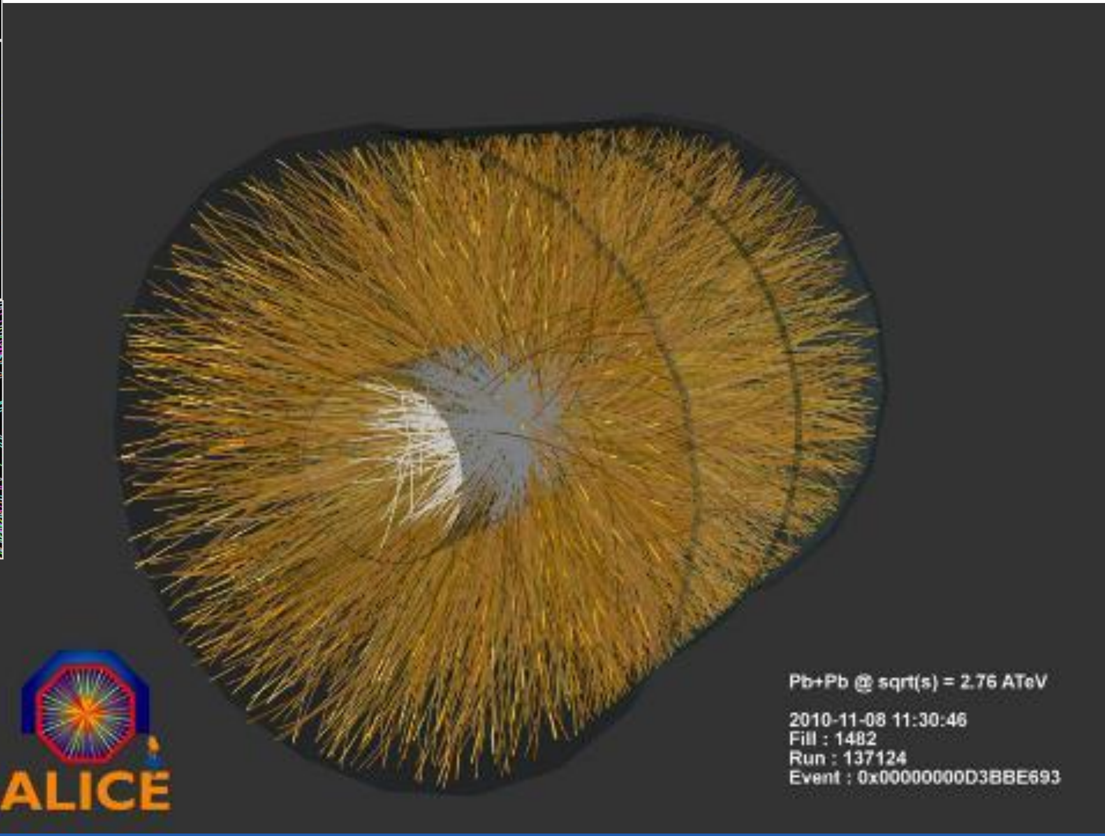


LHC ring: 27 km circumference

Collisions Produce 1PB/s


ATLAS EXPERIMENT
Run Number: 201289, Event Number: 24151616
Date: 2012-04-15 16:52:58 CEST

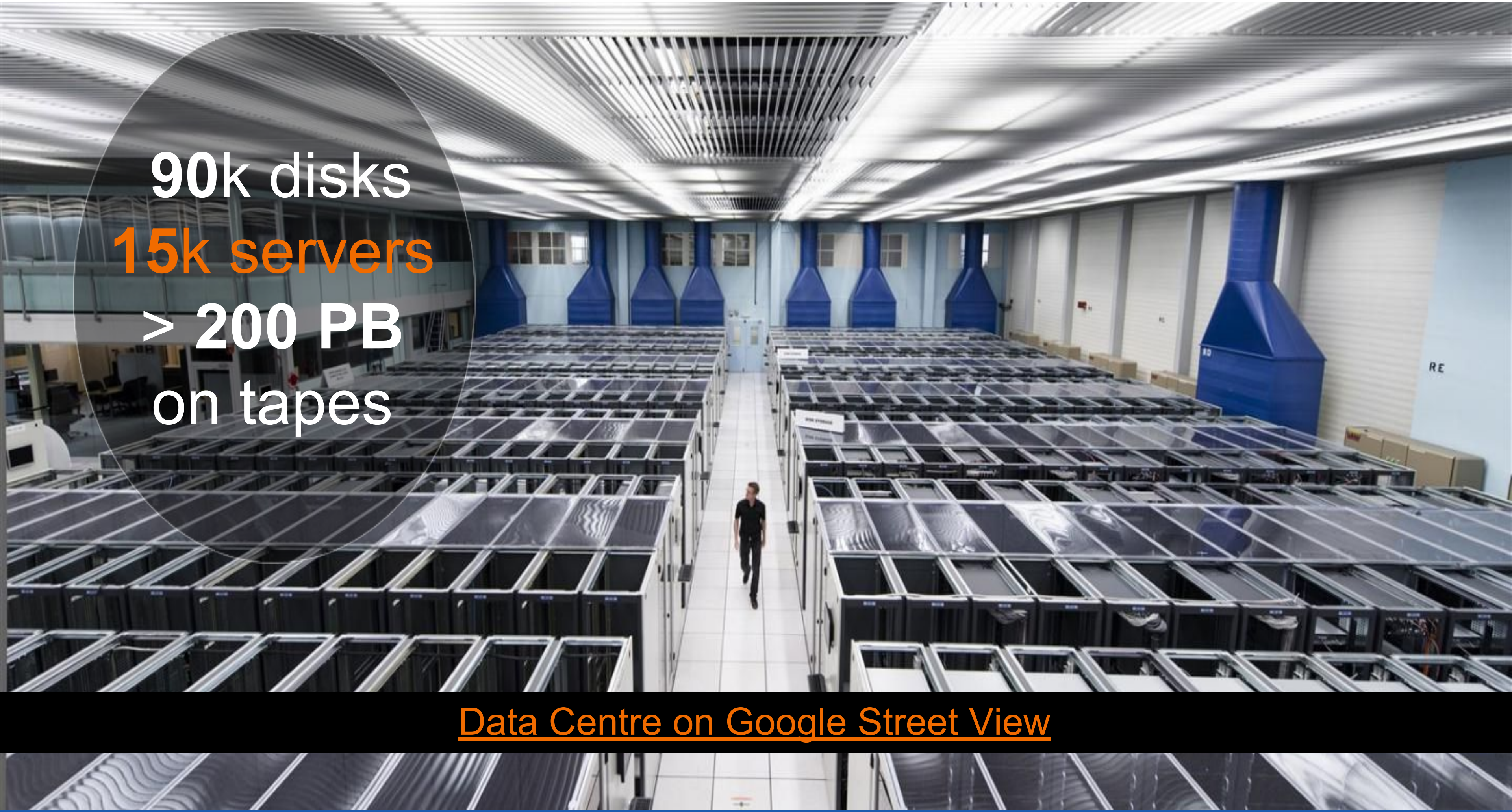
- Event filtering – down to 6Gb/s today
- Data reconstruction
- Data analysis
- Find the interesting events



Simulations

- Particle beam trajectories
- Theory behind events
- Events and detectors...

CERN Data Centre: Primary Copy of LHC Data



90k disks

15k servers

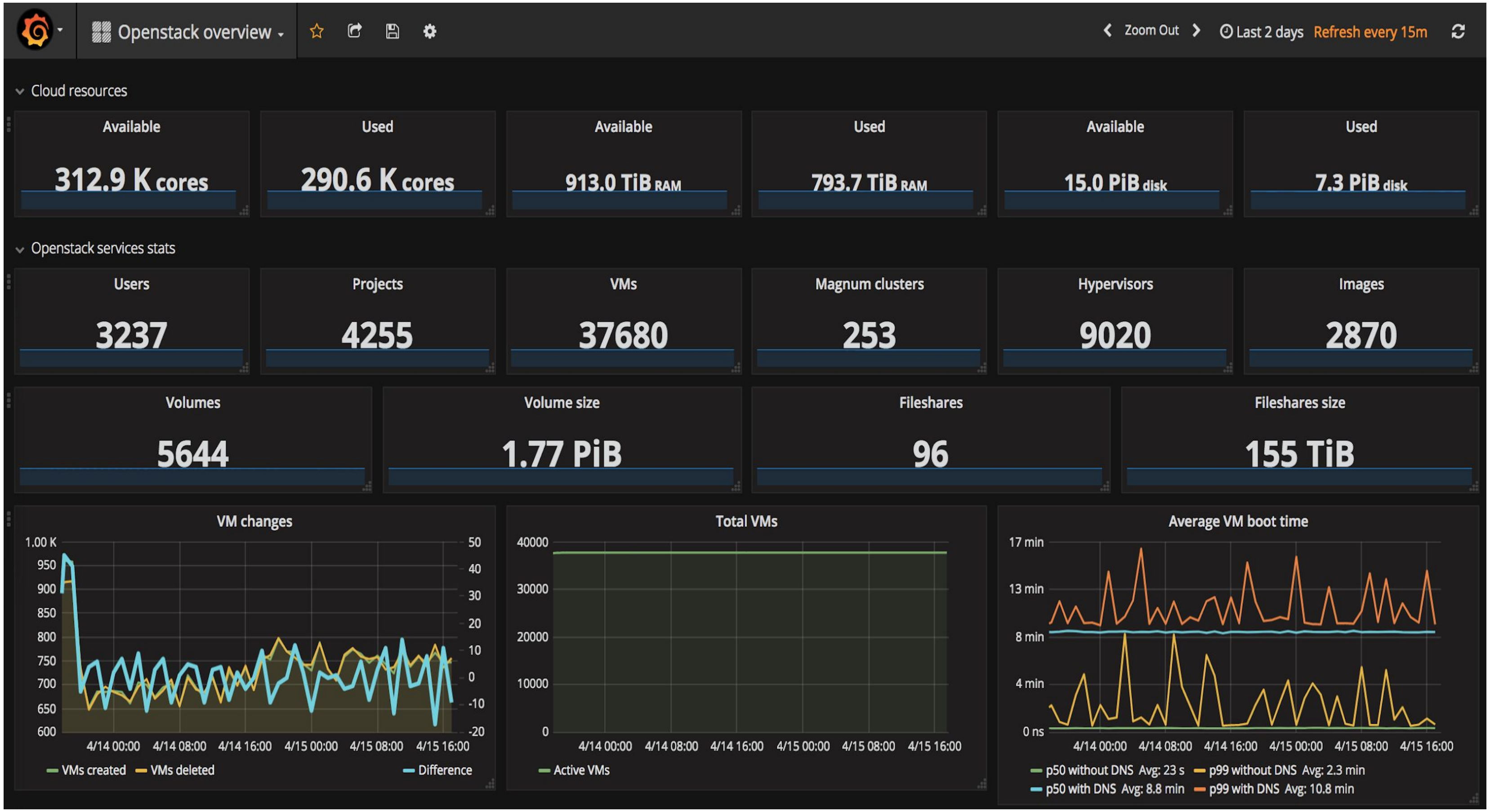
> 200 PB

on tapes

Data Centre on Google Street View

CERN Data Centre: Private OpenStack Cloud

- Over 500 000 physics jobs/day on over 300 000 cores



Cloud resources

Available
312.9 K cores

Used
290.6 K cores

Available
913.0 TiB RAM

Used
793.7 TiB RAM

Available
15.0 PiB disk

Used
7.3 PiB disk

Openstack services stats

Users
3237

Projects
4255

VMs
37680

Magnum clusters
253

Hypervisors
9020

Images
2870

Volumes
5644

Volume size
1.77 PiB

Fileshares
96

Fileshares size
155 TiB

WLCG: LHC Computing Grid

About WLCG:

- A community of 10,000 physicists
- ~250,000 jobs running concurrently
- 600,000 processing cores
- 700 PB storage available worldwide
- 20-40 Gbit/s connect CERN to Tier1s

Tier-0 (CERN)

- Initial data reconstruction
- Data recording & archiving
- Data distribution to rest of world

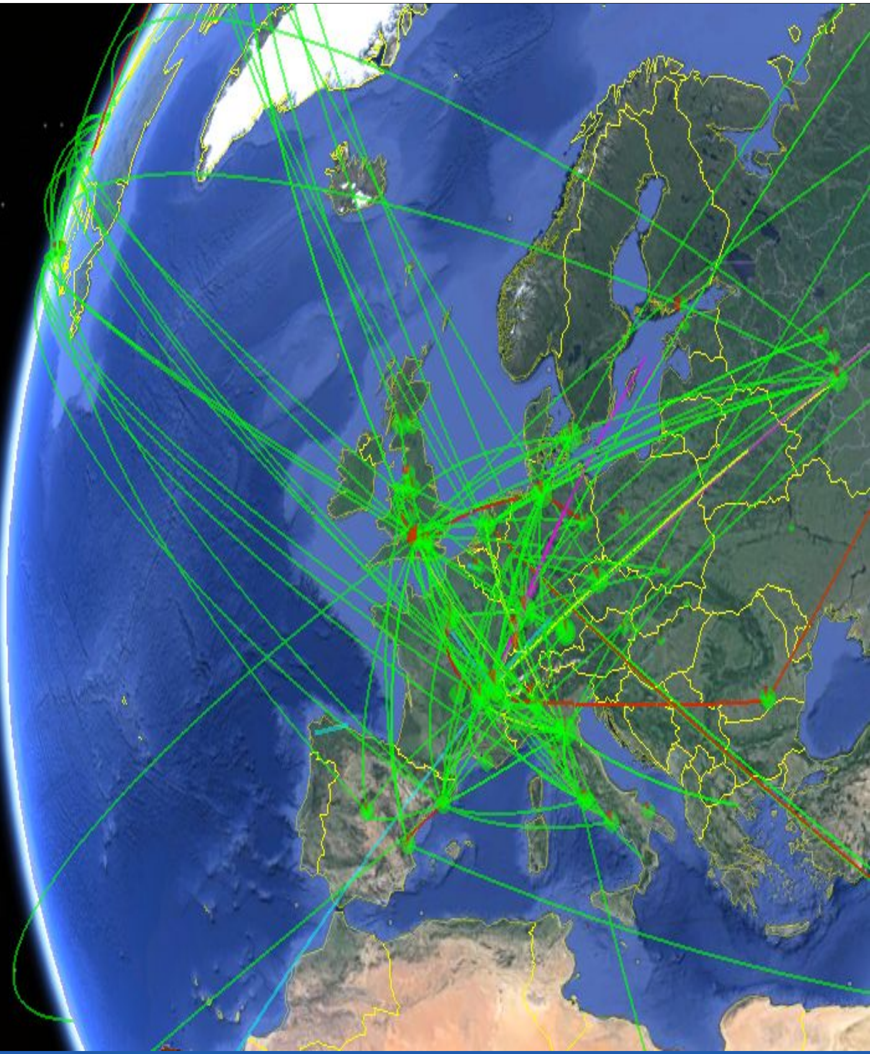
Tier-1s (14 centres worldwide)

- Permanent storage
- Re-processing
- Monte Carlo Simulation
- End-user analysis

Tier-2s (>150 centres worldwide)

- Monte Carlo Simulation
- End-user analysis

170 sites
WORLDWIDE
> 10000
users



CERN batch compute

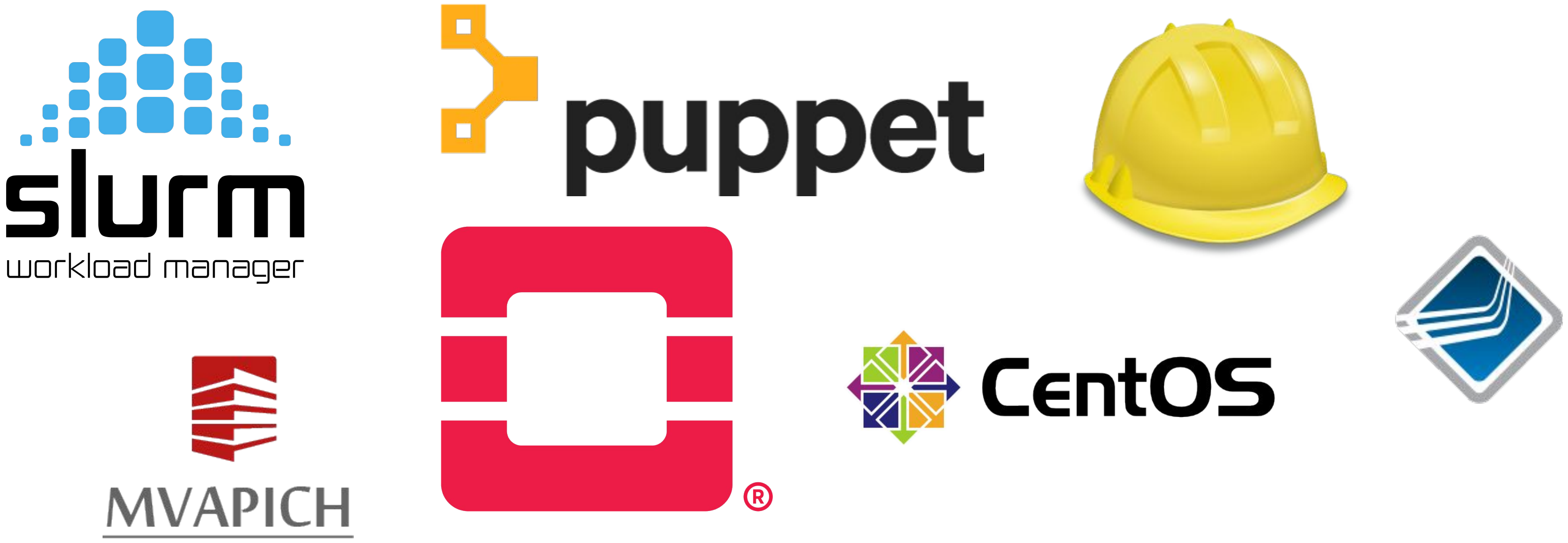
HTCondor
High Throughput Computing

- The bulk of computing at CERN is done via High Throughput Computing (HTC) facilities via Grid or local
- CERN local batch system
  - 1-8 cores for a single job for maximum efficiency
  - 16-48 cores for applications with special requirements
- Also volunteer computing (LHC@home) for high CPU/low I/O simulations

LHC @home
Volunteer computing for the LHC

HPC at CERN

- Applications and use cases that do not fit the standard batch High Throughput Computing (HTC) model.
- About 250 nodes, 5000 cores.
- Integration with Agile environment



HPC user community

Beams and technology

- Plasma and beam simulations for LHC and smaller experiments
  - Gdfdl - field calculations for RF cavities
  - Picmc - plasma simulation
  - PyOrbit - Objective Ring Beam Injection and Tracking

Theoretical Physics

- OpenQCD - Lattice QCD simulations

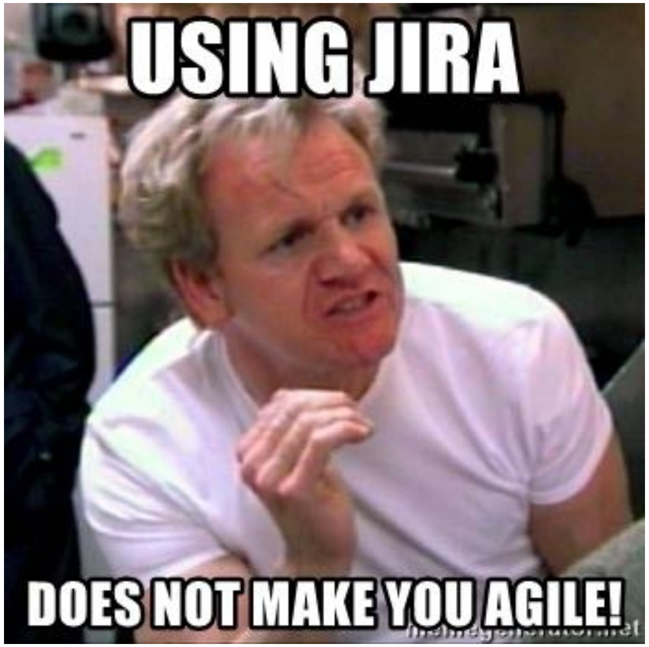
Safety and Engineering

- Safety and fire simulations
  - FDS (Fire Dynamics Simulator)
- Computational Fluid Dynamics
  - Ansys-Fluent
  - OpenFOAM
- Structural analysis
  - Ansys
  - LS-Dyna

WLCG

- Worldwide LHC Computing Grid
- Backfill with Grid jobs via HTCondor to increase cluster utilization

Agile vs. HPC



Agile Methodologies

- High automation and frequent changes
- Shared configuration
- No room for special cases

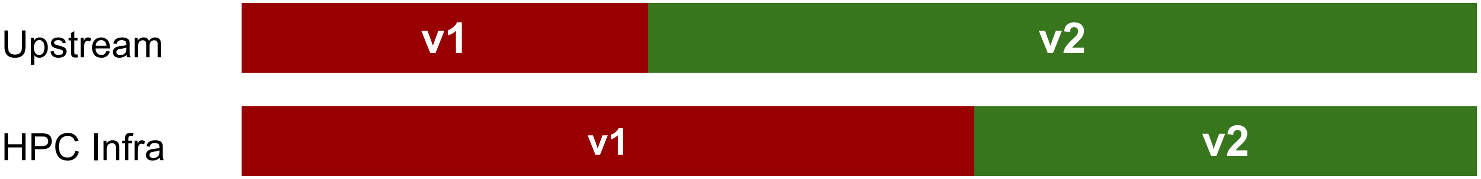
HPC

- Long-running jobs (several weeks)
- Stability
- Few interventions and changes
- Performance tuning

Agile + HPC

- Keep high level of automation, frequent changes
- Separate testing and production environments
- Perform extensive testing before rolling out to production
- Almost never need to drain all nodes

- Repository snapshotting to control changes



Upstream v1 v2

HPC Infra v1 v2

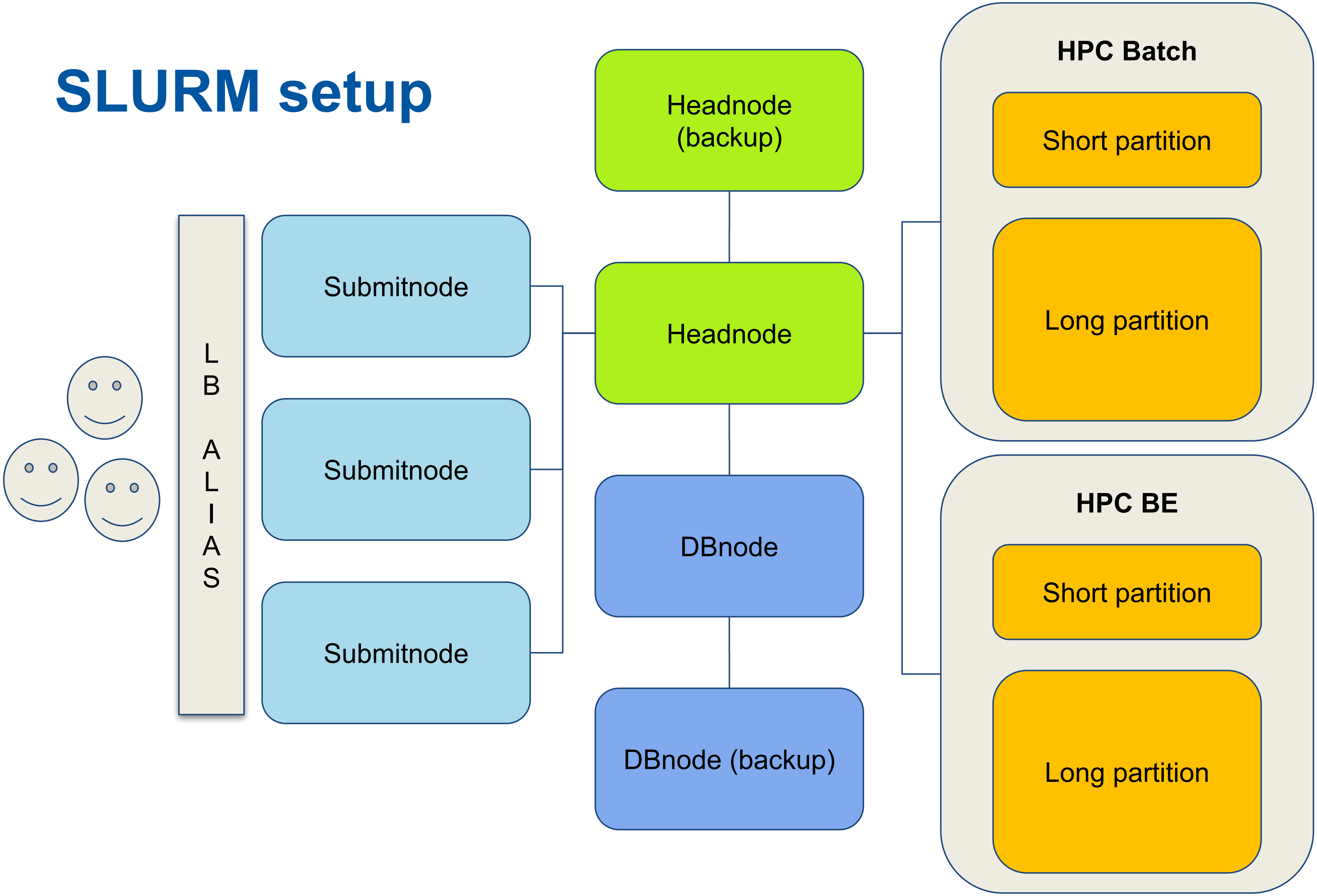
SLURM setup

- Four partitions covering two clusters
- Configuration done by puppet module
- Smaller replicated setup for QA/testing
    - Management nodes (VMs) + 2-5 QA workernodes

Challenges

- Automating the setup and choosing plugins
- Integrating with HTCondor for backfill

SLURM setup


Headnode (backup)
Headnode
DBnode
DBnode (backup)
L B ALIAS
Submitnode
Submitnode
Submitnode
HPC Batch
Short partition
Long partition
HPC BE
Short partition
Long partition

SLURM puppet module

- Configurable and customisable setup for SLURM
- Supports SLURM versions 16.* onwards
- Available at: https://github.com/cernops/puppet-slurm
- Contributions welcome!

SLURM plugins and tools

- Fairly basic setup with VMs and bare-metal
  - Separate MySQL instance for accounting
  - Munge, X11, cgroups, multifactor priority...
- STUBL tools: https://github.com/ubccr/stubl
- NHC: https://github.com/mej/nhc
- Tried Slurm-web: https://github.com/edf-hpc/slurm-web

HPC Containers in SLURM

Singularity containers

- Environment and libraries shipped with application
- Fulfill specific application requirements
- Easier to reuse, refer to and share job configurations

HPC ❤ OpenStack

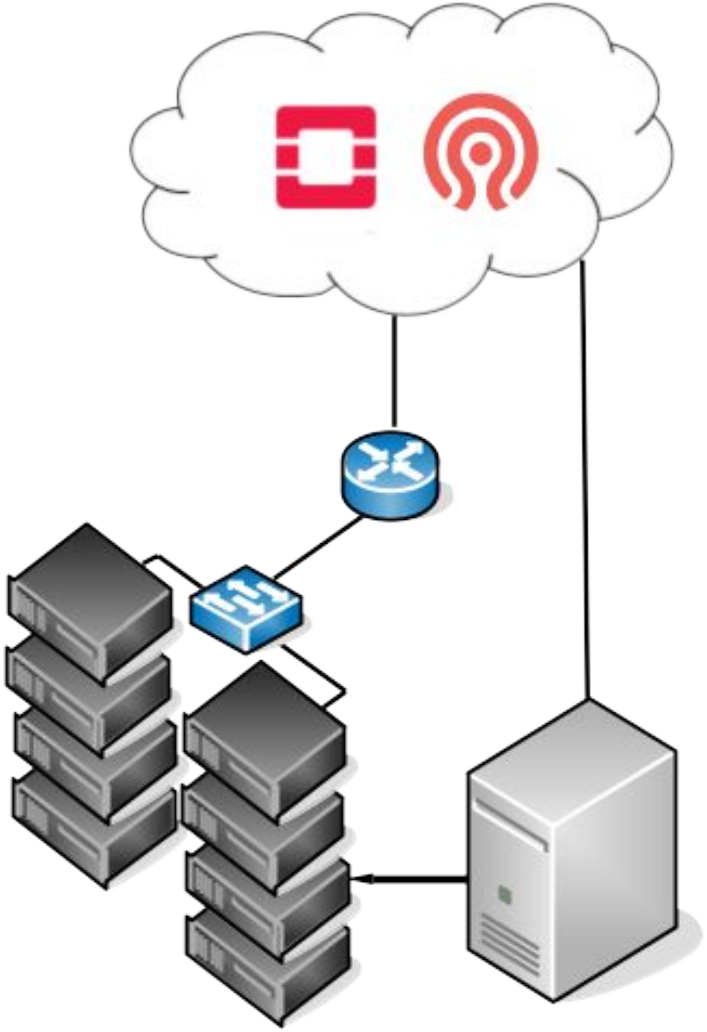
OpenStack Ironic bare-metal provisioning

- Access to raw resources without hypervisor isolation or overhead
- No resource sharing among tenants
- Faster context switching, no hypercalls, less cache flushes, less overhead (latency!)
- PMU access
- Possibility to optimize low-level BIOS and kernel settings
- Full advantage of fast Infiniband interconnects

HPC ❤ CephFS

HPC workernodes

- Intel Xeon E5 2630 v3
- 128GB Memory 1600Mhz
- RAID 10 SATA HDDs
- Low-latency Chelsio T520-LL-CR
- Communication iWARP/RDMA CM



CephFS Jewel

- 3x replication
- Per-host replication
- Shared file POSIX consistency model
- Mon, MDS live in cloud

Legacy bare-metal provisioning

VMs on OpenStack

HPC ❤ CephFS

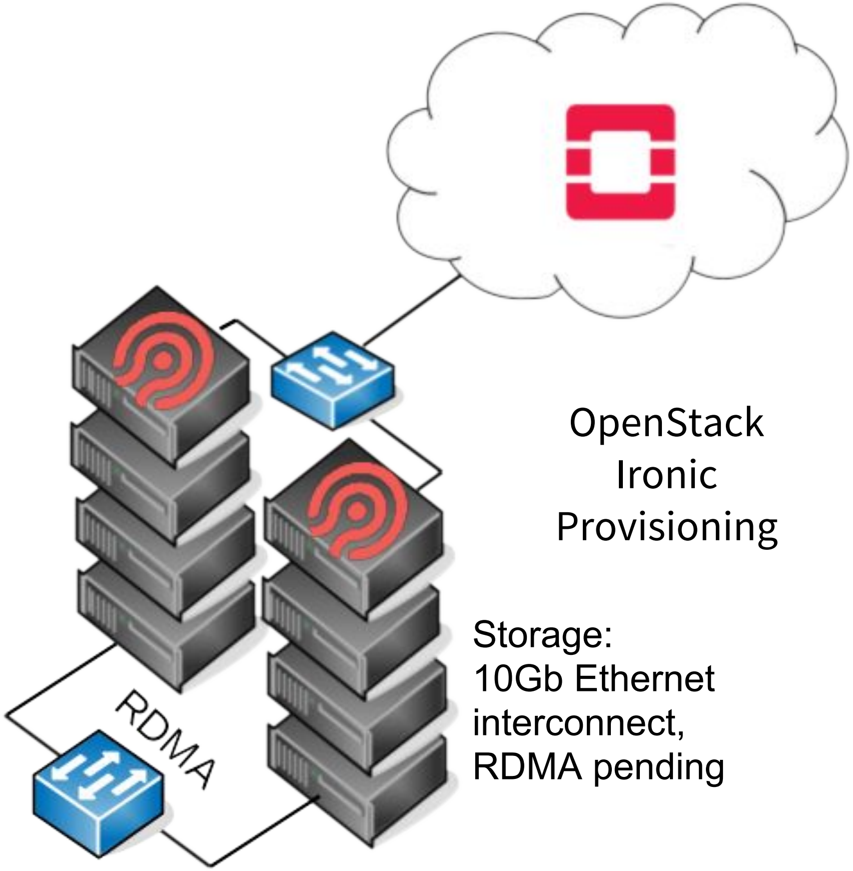
Hyperconverged
Compute + Storage

- Intel Xeon E5 2630 v4
- 128GB 2400Mhz 18ASF2G72PDZ-2G3B1
- 4x 960GB Intel S3520 SATA3
- RDMA Interconnect (compute)
- Mellanox MT27500 ConnectX-3 56Gb/FDR
- 10Gb Ethernet (storage)

- CephFS Luminous 12.2.5
- Network-local
- Pinned MDS
- OSDs on compute nodes
- 2x replication
- Rack-aware replication
- Lazy I/O relaxed POSIX

IO500 SCORE:
Throughput: 3.77 GB/s
Metadata: 8.20k IOPS
Best Score: 5.56
(On 10Gb Ethernet)

Openstack Pike + CephFS Luminous

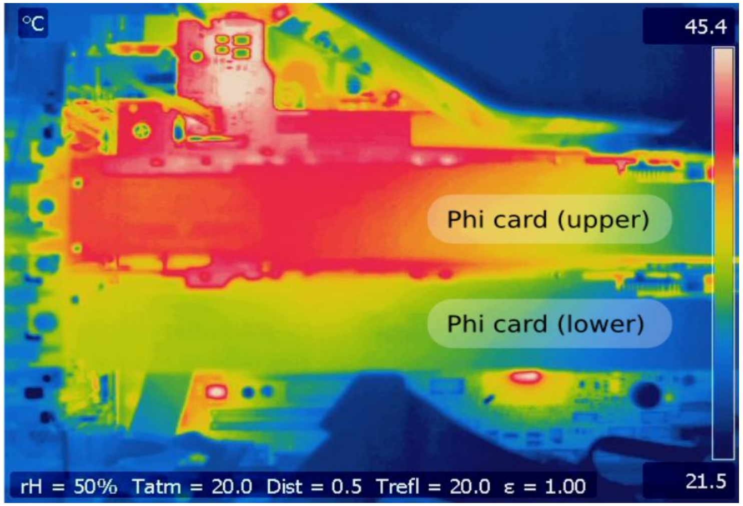


OpenStack
Ironic
Provisioning

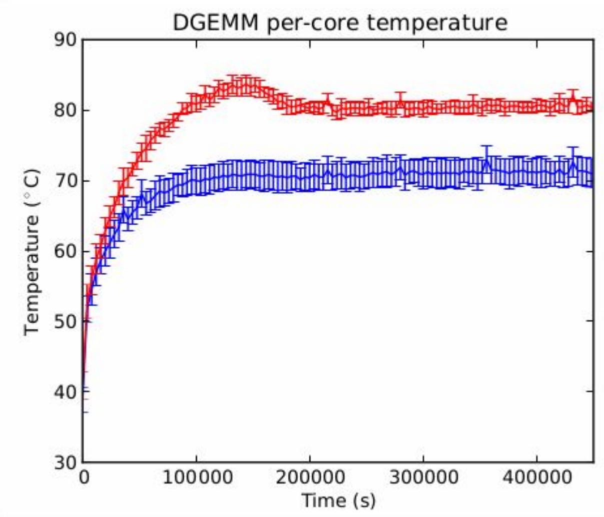
Storage:
10Gb Ethernet
interconnect,
RDMA pending

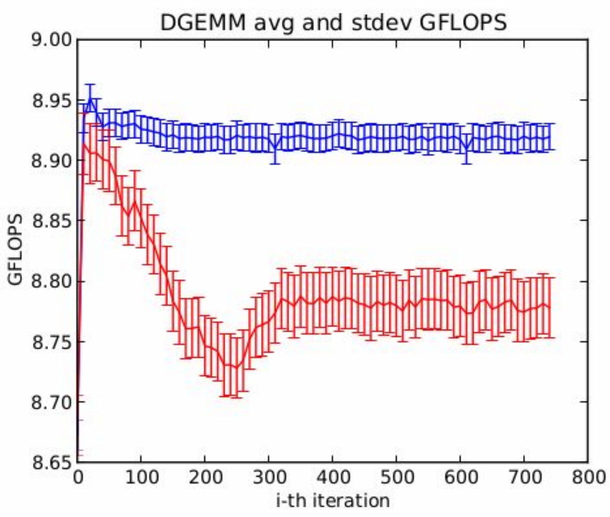
Future Work

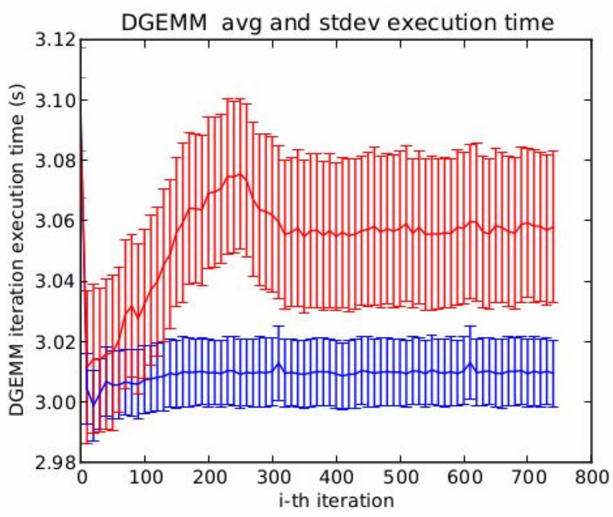
- Increase resource utilization
- Increase workload power and performance efficiency



- Improve data gathering and analysis of HPC workloads







Highlights

- CERN runs a relatively small HPC site that integrates with a very large HTC infrastructure
- We run an HPC facility on SLURM in an agile and cloud-based environment
- We're open sourcing our **puppet-slurm** module on GitHub.
- We are run CephFS as a shared and parallel filesystem for both production and experimental use cases.
- We look forward to discuss similar scenarios and use cases with you!

Our interests

- How to integrate engineering applications with SLURM?
  - Ansys-Fluent - how do you run on your site?
  - Commercial applications rely on ssh, do you restrict ssh in any way? pam_slurm_adopt or other solutions?
- Resource booking
  - Plugin or software for booking resources?
- Alerting and job performance statistics
  - Recommended solutions?

Questions and discussion

Credits

References:

Minimizing Thermal Variation Across System Components, Zhang et al., IPDPS 2015.

Enhancing the programmability and energy efficiency of HPC and virtualized environments, Thesis, Llopis et al. 2016.

Image sources:

HTCondor logo: https://research.cs.wisc.edu/htcondor/logos/

SLURM logo:https://commons.wikimedia.org/wiki/File:Slurm_logo.svg

Foreman logo: https://github.com/theforeman/foreman-graphics/blob/master/logo/foreman.png

Openstack logo:https://www.openstack.org/brand/openstack-logo/logo-download/

Centos logo: https://wiki.centos.org/ArtWork/Brand/Logo?action=AttachFile&do=get&target=centos-logo-light.png

Mvapich logo: http://mvapich.cse.ohio-state.edu/static/images/MVAPICH-Stacked.png

OpenMPI logo: https://www.open-mpi.org/images/open-mpi-logo.png

Using JIRA meme: https://memegenerator.net/img/instances/65567790/using-jira-does-not-make-you-agile.jpg

Testing in production meme: https://cdn.thenewstack.io/media/2018/07/8e60bbf1-one-does-not-y49d8t.jpg

Enjoy Slurm:

https://johnjohns1.fjcdn.com/comments/I+think+youre+confusing+clamps+and+slurms+mckenzie+_1e71e220a700567773186afa1e892b1e.jpg

If it fits it ships meme: https://media.makeameme.org/created/if-it-fits-5baacb.jpg