



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"

CMS

CERN Prévessin

ATLAS

LICE

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

27 km

LHC

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

LHCh

CERN Prévessin

CMS

CMS

HCb

ATLAS

SPS

ATLAS

ALICE

ALICE

LHC ring: 27 km circumference

Collisions Produce 1PB/s





ALICE



Simulations

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



- Data reconstruction
- Data analysis
- Find the interesting events



2010-11-08 11:30:46 Fill : 1482 137124 0x00000000D3BBE693



CERN Data Centre: Primary Copy of LHC Data



Data Centre on Google Street View



CERN Data Centre: Private OpenStack Cloud

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

🔞 🔹 🕼 Openstack overview 🗸 🕆 🗈 🛊 🗘 🖓 🖓 🖓 🖒 🖓 🖓 🖓 🖓 🖓 🖓 🖓 🖓 🖓 🖓 🖓 🖓 🖓					
Cloud resources					
Available	Used	Available	Used	Available	Used
312.9 K cores	290.6 K cores	913.0 ТІВ кам	793.7 TiB RAM	15.0 PiB disk	7.3 PiB disk
 Openstack services stats 					
Users	Projects	VMs	Magnum clusters	Hypervisors	Images
3237	4255	37680	253	9020	2870
Volumes		Volume size Fileshares		Fileshares size	
5644		1.77 PiB 96			155 TiB
VM changes		Total VMs		Average VM boot time	
950 - 40 900 - 30		30000			
		20000			
		0			
4/14 00:00 4/14 08:00 4/14 1 — VMs created — VMs deleted	16:00 4/15 00:00 4/15 08:00 4/15 16:00 — Difference	4/14 00:00 4/14 08:00 4/14 16:00 4/15 00:00 4/15 08:00 4/15 16:00 — Active VMs		4/14 00:00 4/14 08:00 4/14 16:00 4/15 00:00 4/15 08:00 4/15 16:00 — p50 without DNS Avg: 23 s — p99 without DNS Avg: 2.3 min — p50 with DNS Avg: 8.8 min — p99 with DNS Avg: 10.8 min	



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



- •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 (<u>LHC@home</u>) for high CPU/low

I/O simulations



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

WLCG

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

- Structural analysis
 - Ansys
 - LS-Dyna



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





Repository snapshotting to control changes





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 puppet module

- Configurable and customisable setup for SLURM
- Supports SLURM versions 16.* onwards
- Available at: <u>https://github.com/cernops/puppet-slurm</u>
- 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: <u>https://github.com/mej/nhc</u>
- Tried Slurm-web: <u>https://github.com/edf-hpc/slurm-web</u>



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)





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

