CERN

HPC / Slurm service

CSC on IT services

Nils Høimyr / IT-CD

7th of November 2024

Outline

High Performance Computing (HPC) - CERN context

- User community
- Examples of HPC use cases at CERN

Message Passing Interface (MPI)

HPC clusters - hardware

HPC – software and OS

HPC Batch cluster – user environment

- Running a job
- HPC – Slurm partitions and queues

Slurm architecture

HPC backfill

Possible user issues

Future plans

High Performance Computing (HPC)

CERN context – reminder for newcomers

- **Motivation**: Address needs of parallel MPI applications and use cases that do not fit the standard batch High Throughput Computing (HTC) model
- SLURM MPI clusters as complement to HTCondor batch service
- **Theory and ATS sector main users**
  - Restriced HPC service (**KB0004975**) and user community
- Batch HTC under HTCondor (~400k cores) main compute service
  - Worker nodes with up to 96 cores
  - A few “bigmem” nodes (1TB of memory) for special use cases
  - Some GPU capacity (T4,V100 and A100)
- For ML use cases: K8S and Kubeflow

HPC and HTC at CERN

- Any application that fit in a single physical server => Use HTCondor
  - Multi-core CPU jobs (also MPI or OpenMP within a box)
  - “Bigmem” Condor jobs (1TB of RAM)
    - Detector calibration runs
    - Engineering (ANSYS Mechanical, CST Field Calculations)
  - GPU enabled applications
    - Batch GPU nodes under Condor, or K8S
- Parallelized MPI applications that can scale out on multi-node clusters => run on Slurm

User community

BE

- Plasma simulations for Linac 4
- Beam simulations for LHC, CLIC, FCC…
- Xtrack, PyOrbit etc

TH

- Lattice QCD simulations

HSE

- Safety/fire simulations (FDS, OpenFOAM)

SY

- Gdfdl (field calculations for RF cavities)
- Field calculations (CST...)

TE

- Picmc
- Engineering (Ansys and Comsol)

EN

- CFD (Ansys-Fluent, OpenFOAM)
- Structural analysis (Ansys, LS-Dyna...)

Other users, HTC and batch service please!

~9000 cores for HPC

~400 000 cores for batch

Examples of HPC use cases at CERN

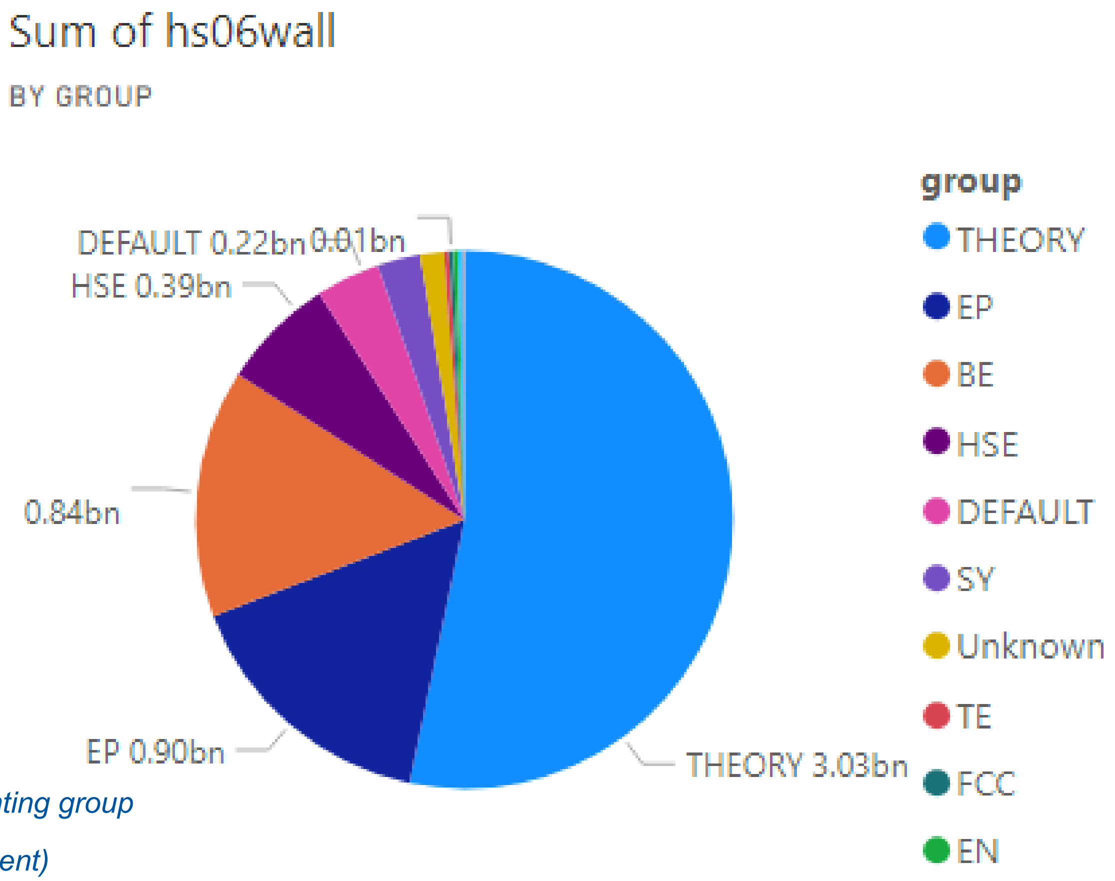
- Theoretical Physics: Perturbative Quantum Chromodynamics (QCD)
  - Lattice QCD ⇐ largest HPC users at CERN
  - Development of Latttice-QCD simulation codes: OpenQCD, Grid and others
  - Running on external supercomputers with research grants
- Numerical search for optimal damper settings for beam quality in LHC and HL-LHC
  - Optimisation of beam luminosity using the hybrid MPI-OpenMP application COMBIp
- Beam formation simulation in LINAC4 ion source
  - Running ONIX (Orsay Negative Ions eXtraction) 3D Particle-in-Cell Monte Carlo Collision code
- Self-consistent electron cloud simulations to study coupled bunch instabilities
  - Electron cloud can cause beam instabilities through the electromagnetic coupling of the electron motion and the proton beam dynamics
  - E-cloud instabilities regularly occur e.g. in the LHC, important to study them with respect to upgrades (HL-LHC)
  - Using the PyECLOUD-PyHEADTAIL suite, developed and maintained at CERN

HPC use cases at CERN - 2

- Plasma simulations applied to superconducting thin film coating processes for RF cavities
  - Superconducting radiofrequency (SRF) cavities **accelerate** the charged particles of a beam through an RF **electric field**
  - Superconductive film coating increasing the energy coupling to the beam
  - Simulations using PICMC (Particle-In-Cell Monte Carlo) code from Fraunhofer institute
- HPC for dynamic, thermo-mechanical and CFD simulations on Beam Intercepting Devices
  - FEM simulation for thermomechanical behavior with Ansys LS-DYNA
  - CFD simulations with AnsysCFX and Fluent
- Fire and smoke dynamics simulations in underground accelerator installations
  - Fire Induced Radiological Integrated Assessment (FIRIA project), risk and consequence analysis
  - Using FDS – Fire Dynamics Simulator developed by NIST
  - Also near-field dispersion with ANSYS Fluent CFD simulations

For more information and examples, please refer to the **HPC user workshop session 1 and session 2** held in 2020 with presentations of applications and details of HPC use cases in different teams

HPC usage by main accounting group 2023


Sum of hs06wall

BY GROUP

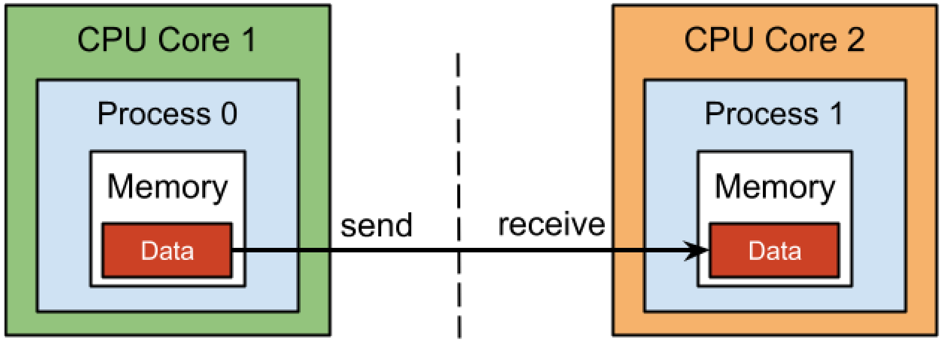
Remarks:

- *Atlas backfill jobs filtered out*
- *Default: User not yet in an accounting group*
- *EP: Engineering users (Ansys Fluent)*

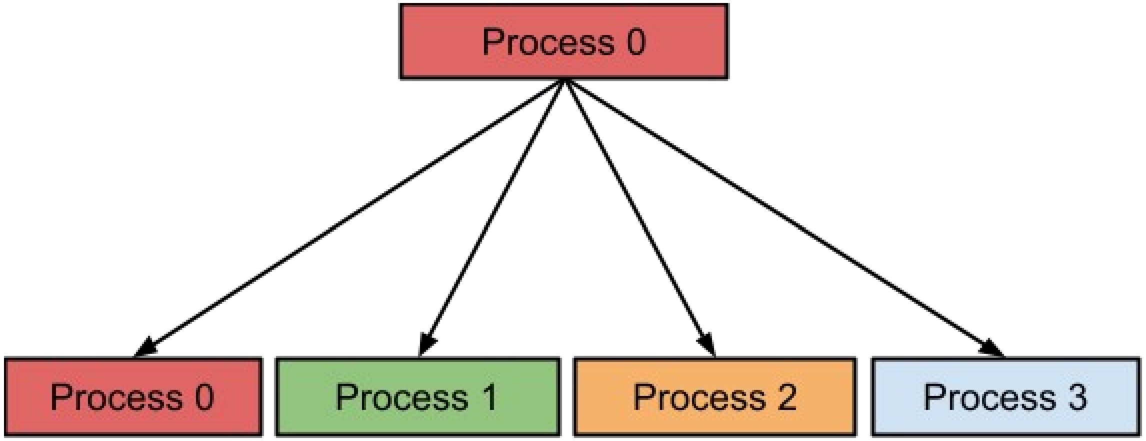
Message Passing Interface (MPI) - 1

- The Message Passing Interface (MPI) is a standardized and portable message-passing standard designed to function on parallel computing architectures
- Library of functions to be called from C, C++ or Fortran code

MPI Point to point communication



MPI Collective communication



Message Passing Interface (MPI) - 2

- Documentation of MPI features: https://www.mpi-forum.org
- Several MPI implementations:
  - MVAPICH2 https://mvapich.cse.ohio-state.edu/
  - OpenMPI https://www.open-mpi.org/
  - Also commercial: e.g. Intel MPI and HP MPI
- Low latency and high memory bandwidth for performance
- Benchmark tests:
  - OSU Latency (Point to Point)
  - OSU Allgather (Collective)
- Run application: `mpirun` / `mpiexec`, (`srun` with Slurm)

HPC MPI clusters - hardware

- 4 Infiniband clusters, each on different Slurm partitions:
  - 2x72 nodes with 2 x Xeon(R) CPU E5-2630/20 cores (40HT), Infiniband FDR (partitions “inf-short” and “inf-long”)
  - 72 nodes with 2x AMD EPYC 7302 32 cores, Infiniband EDR (partition “photon”)
  - 80 nodes with 2x Intel® Xeon® Gold 6442Y – 48 cores (96HT) Infiniband HDR (“muon” partition)

- All nodes with shared CephFS file system `/hpcscratch`

HPC – software and OS

- Clusters now running EL9 Linux (RHEL 9.4)

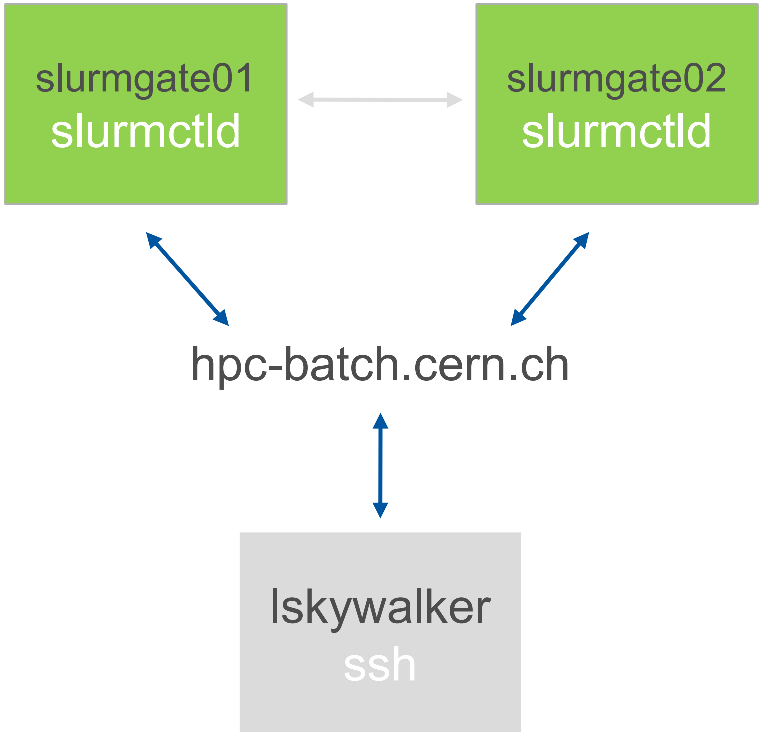
- Slurm 23.02.07
- MPI versions via modules
  - OpenMPI 316 and 411 (4.1.6 now also available)
  - Mvapich 2.3
- Same software and OS packages as lxplus and lxbatch for compability

HPC Batch cluster – user environment

- Login to submit node: “hpc-batch.cern.ch”
  - Users' home and scratch directories on /hpcscratch file system (CephFS)
  - AFS and EOS available, similar to lxplus
  - Applications on AFS or CVMS, (also local or EOS...)
  - EOS for data copy and project storage
- SLURM for HPC scheduling
  - Jobs typically run unauthenticated (run times up to several weeks)
  - Submission with Kerberos token supported via Auks, for copy back to EOS

Submit node

- Users compile their jobs against the MPI distribution they choose using module
- Users launch their jobs, check job status, cancel jobs…
- Similar to lxplus, but reserved for HPC



Running a job

- srun (process manager, interactive)

  ```
  $ srun -n 128 –cpus-per-task=2 -p inf-short -t 10 my_MPI_executable
  ```

- sbatch (script submit system, background)

  ```
  $ sbatch -t t20 -p inf-long my_MPI_script.submit
  ```

- salloc (allocation of nodes, interactive)

  ```
  $ salloc -n 256 --cpus-per-task=32 [bash|my_MPI_executable]
  ```

- More details: KB0004541

- Queues and submission parameters documented in: KB0004973

HPC – Slurm partitions and queues

Partition name	Max run time	Main users
inf-short	5 days	ATS, HSE, engineering
inf-long	21 days	ATS,HSE, engineering
photon	10 days	BE, TH, (ATS)
phodev	2 hours	BE, TH, (ATS)
muon	10 days	BE, TH, (ATS)
mudev	2 hours	BE, TH, (ATS)

Sample job submit and script

- sbatch (options: partition and time)

```
$ sbatch -p mudev -t 10 testm.sh
```

→ cat testm.sh

```
#!/bin/sh
#SBATCH -N 70
##SBATCH --ntasks-per-node=96  # would use this for other programs to run on all cores
echo "Running on `hostname` "
srun /usr/local/mpi/mvapich2/2.3/libexec/osu-micro-benchmarks/mpi/collective/osu_allgather
exit
```

Sample job submit and script - 2

sbatch (options can also be in job script)

```
$ sbatch mmixer-testN2-8.sh
```

→ cat mmixer-testN2-8.sh

```
#!/bin/bash
#SBATCH -p muon
##SBATCH -p photon
#SBATCH --time 24:00:00
#SBATCH -N 2
#SBATCH --exclusive
cd $SLURM_SUBMIT_DIR
export PATH=/cvmfs/projects.cern.ch/engtools/comsol/comsol62/multiphysics/bin:$PATH
export LD_LIBRARY_PATH=/cvmfs/projects.cern.ch/engtools/comsol/comsol62/multiphysics/lib/glnxa64:$LD_LIBRARY_PATH
echo $SLURM_JOB_NUM_NODES
echo $SLURM_CPUS_ON_NODE
echo $SLURM_NTASKS
srun comsol batch --usebatchlic -nn 16  -np 8  -nnhost 2 -mpifabrics tcp -mpibootstrap slurm -mpipath "$MPI_LIB" -mpiroot "$MPI_HOME"  -configuration "/tmp/config_@process.id" -tmpdir "/tmp" -prefsdir "/tmp/prefs" -data "/tmp/data_@process.id" -inputfile ./inputfile.mph -outputfile ./outputfile.mph
```

Queues and cluster status

- squeue (check jobs and queues, “-u” for user)

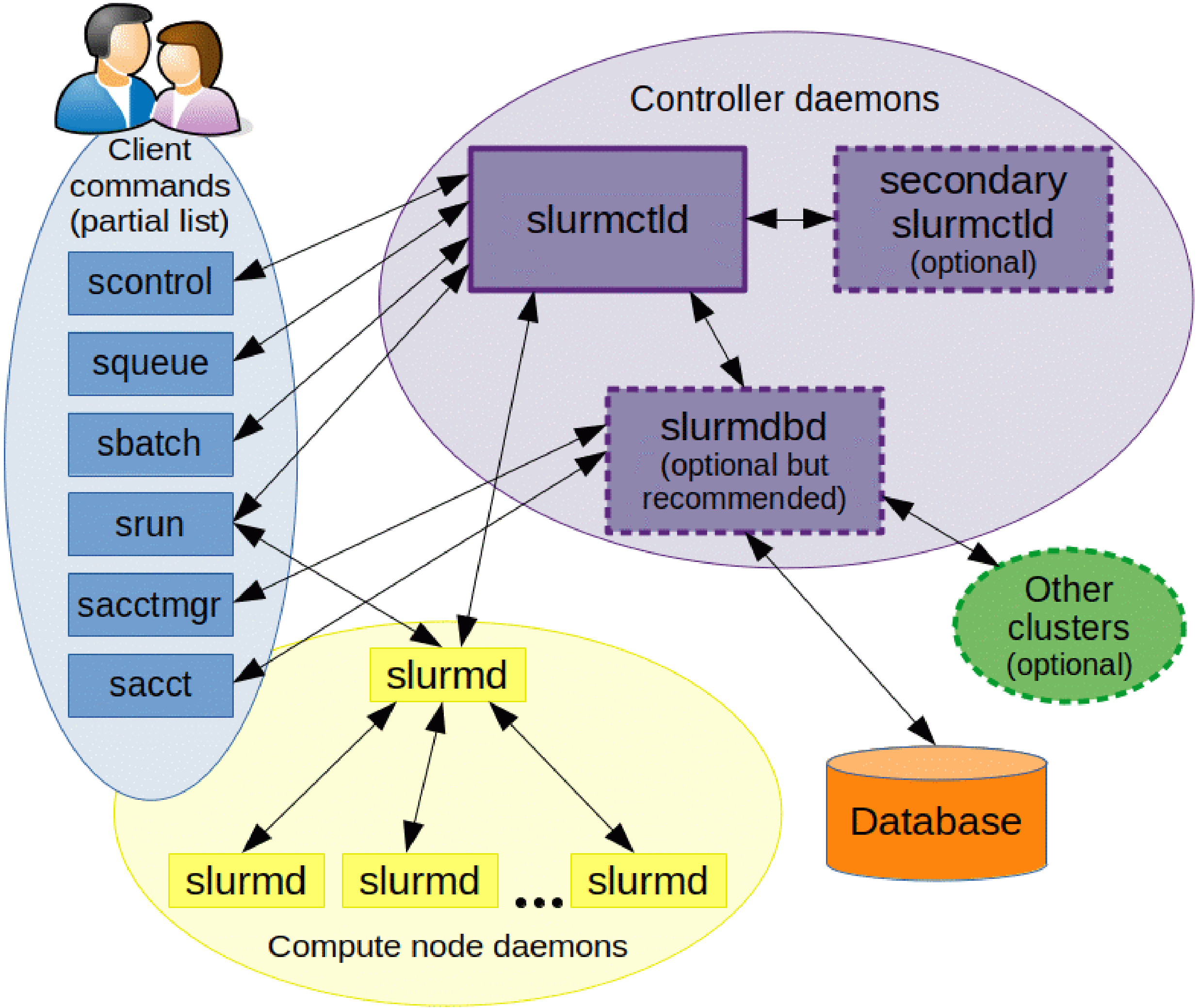
```
$ → squeue -u nils
   JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
  328207     mudev testm.sh     nils  R       0:01     70 hpc-muon[001-004,006-012,017-020,026-080]
  328198      muon mmixer-t     nils  R      17:03      2 hpc-muon[024-025]
```

- sinfo (cluster status)

```
$  sinfo

PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST

inf-short    up 5-00:00:00      2   plnd hpc-be[010,038]

inf-short    up 5-00:00:00    133  alloc hpc-be[001-009,011-025,027-034,036-037,039,041-108,110-116,118-129,131-139,141-142]

inf-short    up 5-00:00:00      7   idle hpc-be[026,035,040,109,117,140,144]

inf-long     up 21-00:00:0      2   plnd hpc-be[010,038]

inf-long     up 21-00:00:0     69  alloc hpc-be[003-005,009,015,018,024,027-029,033-034,036,039,043-044,046,048,050,052,057-058,060-063,065-066,068-071,073-079,081-084,086,089-093,097-098,100-101,105,110-111,114-115,118-119,121-122,125,127,131-133,137-138]

photon       up 10-00:00:0      1    mix hpc-photon001

photon       up 10-00:00:0     70  alloc hpc-photon[002-071]
```

Slurm architecture

- Headnodes/ controllers: `slurm01,02`
- DB nodes: `slurmdb01,02`
- Client/submit nodes: `slurmgate01-09`
- Compute/worker nodes: `hpc-{cluster}001-072`


Client commands (partial list)
scontrol
squeue
sbatch
srun
sacctmgr
sacct
Controller daemons
slurmctld
secondary slurmctld (optional)
slurmdbd (optional but recommended)
Other clusters (optional)
Database
slurmd
slurmd
slurmd
...
slurmd
Compute node daemons

HPC ❤ CephFS

Hyperconverged
Compute + Storage

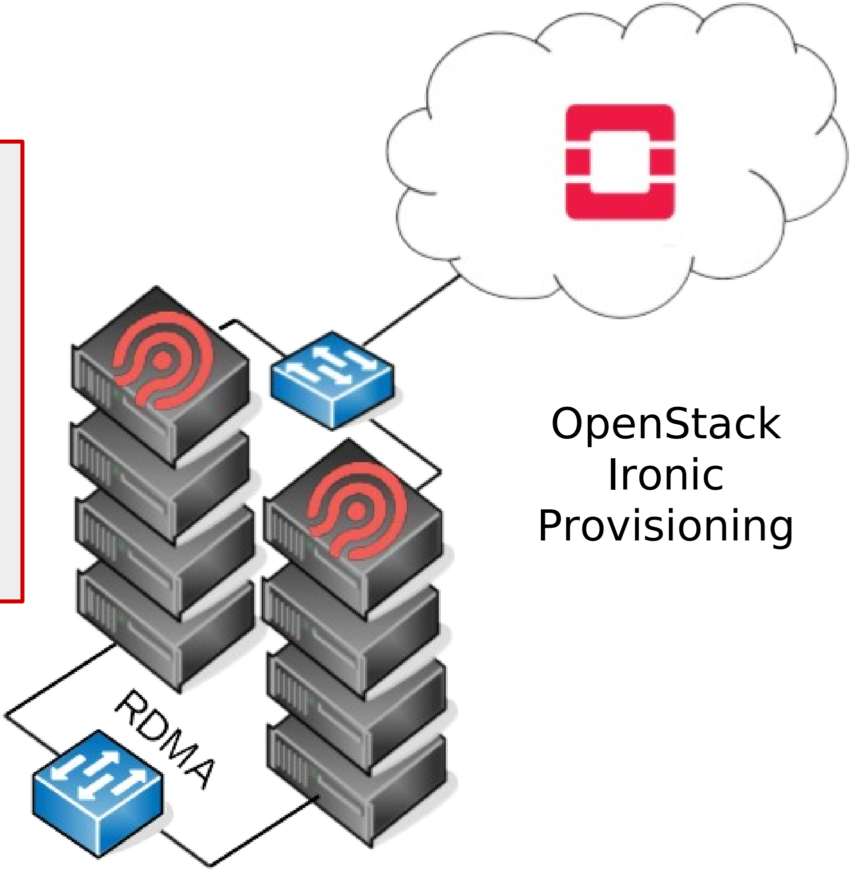
- Intel Xeon E5 2630 v4
- 128GB 2400Mhz 18ASF2G72PDZ-2G3B1
- 4x 960GB Intel S3520 SATA3
- RDMA Interconnect (compute)
- Mellanox MT27500 ConnectX-3 56Gb/FDR
- 10Gb Ethernet (storage)

- 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

Detailed info on numbers in the following contribution:
https://indico.cern.ch/event/587955/contributions/2936868/

Openstack Pike + CephFS Luminous



OpenStack
Ironic
Provisioning

CephFS scratch file system

Home directories for users: `/hpcscratch/user/`

Project areas: `/hpcscratch/project/`

Slurm run-time bitmap: `/hpcscratch/statesavelocation`

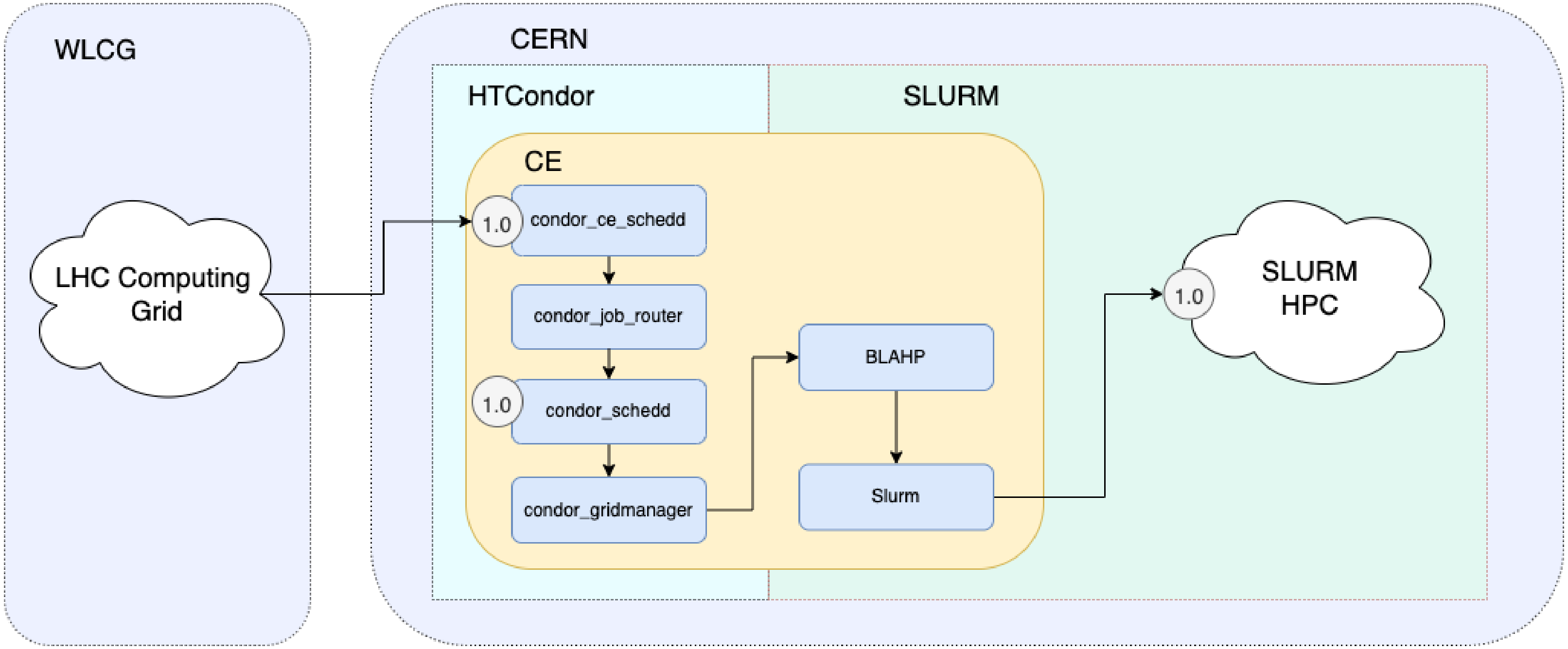
The shared file system is located on the Ceph cluster “Jim”, managed by the Ceph team in IT/SD

Also another Cephfs mount for the TH/QCD team: `/hpcqcd`

For more information, please refer to the Cephfs documentation and the Storage talk

HPC backfill

- In order to maximize use of the HPC resources, nodes not allocated to multi-node MPI user jobs are backfilled with grid jobs
  - Backfill is handled via a Condor Compute Element - **CEHPC**
  - When a user job starts, the backfill jobs are preempted with a SIGCONT and SIGTERM signal, and then after 5 minutes, SIGCONT, SIGTERM and SIGKILL



Possible user issues

- MPI environment errors (ref. KB0004541 and how to load MPI modules)
- Application runs out of memory (adjust cores/nodes)
- Job does not start (lack of free nodes): KB0004837
- SLURM queues and job parameters: KB0004973
- The commands: sinfo, squeue are useful!

Please refer to our user documentation:

https://batchdocs.web.cern.ch/linuxhpc/index.html

And Service Now Knowledge base:

HPC in the Service Portal

Proprietary applications

Some of the applications running on our clusters are proprietary software packages, delivered as “black box” binaries and distributed under a licence agreement.

- E.g. Engineering software like Ansys Fluent, Comsol, LS-Dyna, or Field calculations applications like CST and GDFiDL.
- Such applications have often been built with a proprietary MPI distribution (e.g. HP or Intel MPI) and are not necessarily optimised for a batch system environment.
  - E.g. not able to use `srun` under Slurm, need to apply workaround with `ssh` and to generate list of allocated hosts in the batch script.
  - Requires setting up ssh keys as a workaround
- Addressed by a set of step by step guides, e.g.
  - Guide for how to run Ansys Fluent (ref. KB0006084)
  - Instructions for CST: KB0005870

List of engineering software provisioned on Linux : KB0003575

More information in the Service Now Knowledge base:HPC in the Service Portal

Future plans

- Swan/notebook integration (For post-processing of results etc) WIP
- Extend cluster with new hardware (cluster renewal)
- Improve monitoring (log files, resource use)
- Slurm and OS upgrades
- Possible intergration with external cloud/HPC resources (if/when available)
- Evolve service with lxplus/batch

Questions?

CERN

www.cern.ch