
CERN

HPC Service – update for BE/ABP 23/11-2017

Carolina Lindqvist, Pablo Llopis, Philippe Ganz* and Nils Høimyr

* Tech in the HPC team until 08-2017

Contents

- High Performance Computing (HPC) context at CERN
- HPC versus batch use cases
- HPC hardware resources
  - Batch cluster pilot cluster
  - New Infiniband clusters
- HPC batch cluster
  - User access
  - SLURM commands
- Work in progress and future plans
- Questions

High Performance Computing (HPC)

Applications and use cases that do not fit the standard batch High Throughput Computing (HTC) model. Typically parallel MPI applications

- 32-2000 cores for a single job
- Applications that scale well with parallelization
  - MPI application performance requires fast interconnects with low latency between nodes in a cluster
  - Stability of OS and environment critical
  - Some applications also require fast access to shared storage

Ref. KB0004192 for more information

User community

BE
- gdfdl (field calculations for RF cavities)
- Plasma simulations for Linac 4
- *Beam simulations for LHC, CLIC, FCC…*
- *PyOrbit etc*

TH
- Lattice QCD simulations

HSE
- Safety/fire simulations (CFD)

TE
- Picmc
- *Potentially also engineering (Ansys)*

EN
- *CFD (Ansys-Fluent, OpenFOAM)*
- *Structural analysis (Ansys)*

Other users, HTC and batch service please!

~5000 cores for HPC (soon)

~160 000 cores for batch

HPC service for accelerator and technology sector

- Batch HPC facility using SLURM “HPC-batch”
  - Pilot cluster with core batch nodes with low-latency 10Gb Ethernet interconnects
  - New HPC cluster resources: 2x72 nodes with Infiniband, being installed and added to the cluster this week

HPC Batch pilot cluster

- ~78 Quanta 16 core / 128Gb boxes with low latency 10Gb ethernet
- Access autorized by e-group (Ref. KB0004975)
- Running CC7.4 and setup for MPI applications
  - Modules for Mvapich2 (recommended) and OpenMPI
  - Users' home directories on /hpcscratch CephFS file system (avoid token issues, allow parallel I/O for results)
- SLURM for HPC scheduling
  - SLURM scheduler setup for HPC jobs
    - HTCondor good for within one node scheduling, SLURM most popular open source HPC scheduler
    - Cluster to be backfilled with batch jobs when idle

New Infiniband HPC cluster

- 2x72 nodes with 20 cores (40 with HT)
  - Intel “Broadwell” CPUs, 128GB RAM, DDR4, 2.4GHz; E5-2630v4, 2.20GHz; 4*960GB Intel DC S3520 SSD; ConnectX-3; 10GbE and Infiniband “Fat tree”.
- In the process of being added to HPC batch facility
- Same user login and environment as pilot HPC cluster
- SLURM partitions “be-short” and be-long”
  - Additional cluster scratch space to be added for faster I/O

HPC service – access authorization

- Access to batch HPC facility is authorized by e-group (ref. KB0004975)
  - Users in BE/ABP should ask the BE/ABP computing WG admins to be added to the e-group: **service-hpc-be** that will authorize login to the hpc-batch cluster
  - Once in the e-group, you can login to hpc-batch with your CERN account (ref: KB0004541)

HPC Batch cluster - access

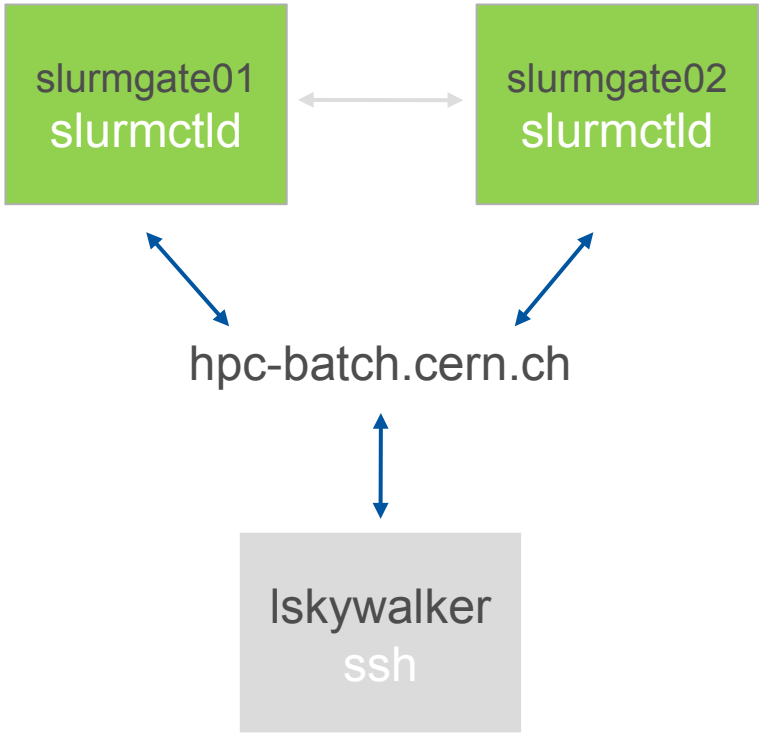
- Login to head node: “hpc-batch.cern.ch”
  - Users' home directories on /hpcscratch file system (avoid token issues)
  - Allows for compilation of MPI code if required
    - Mvapich2 or OpenMPI via modules (command: `module avail` to list)
  - Applications on AFS or EOS
  - EOS for data copy and storage
- SLURM for HPC scheduling
  - Interactive run with “srun -n 128” (-p -t) myapplication
  - Batch run with sbatch script
  - https://slurm.schedmd.com/documentation.html

Running a job

- srun (process manager, interactive)
  $ srun -n 128 –cpus-per-task=2 -p batch-short -t 10 my_MPI_executable
- sbatch (script submit system, background)
  $ sbatch -t t20 -p batch-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

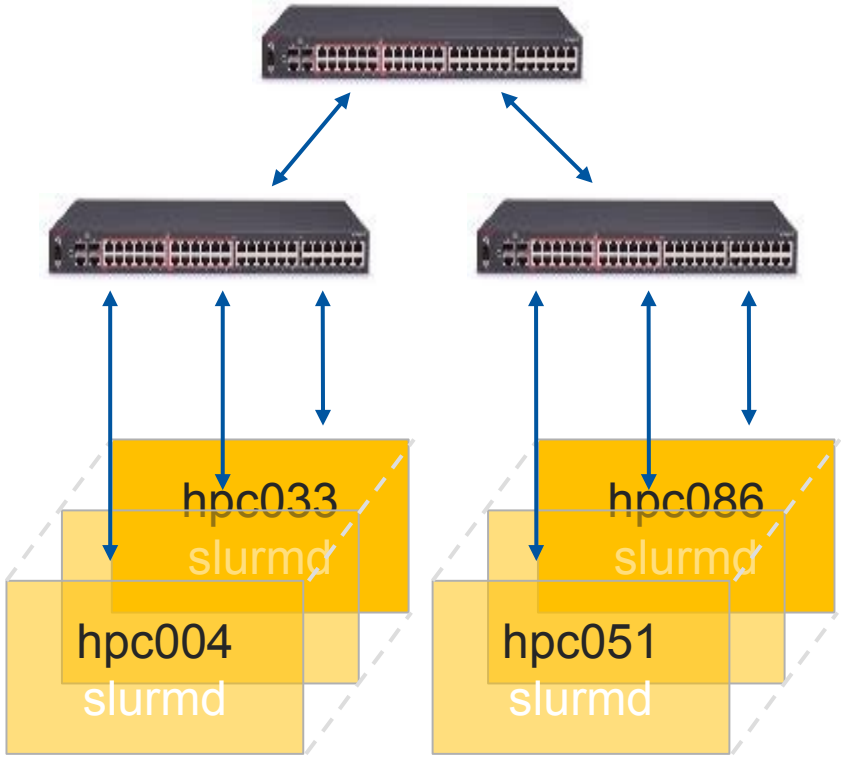
Submit node

- 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



Workernode

- Currently 78 nodes on 3 switches
  - Will be extended with 2x72 nodes on 2 Infiniband trees
- Same version of OFED/MPI as on headnodes
- Users should not log in on workernodes



Monitoring

- Lemon sensors (Ceph mount, RDMA...)
- Kibana dashboards:
  - Time since last puppet runs, RDMA broken
- Nice to have: slurm-web
- Useful commands: sinfo, squeue...
- SLS availability

HPC Batch applications and storage

- For the moment continue with applications on AFS, later EOS (awaiting new Eos-fuse client)
- Home-directory for simulations (/hpcscratch) for results and job data
  - AFS home directory also accessible from head-nodes
- Copy results and input to EOS for storage or post-processing
- Distributed applications spanning many physical hosts should use local I/O as far as possible for performance reasons
  - E.g. application settings:
    - Project-dir: `/hpcscratch/user/{username}/myproject`
    - Temporary directory: `/tmp/{user}/myapp`
    - On new Infiniband nodes also: `/scratchfs`

Pilot HPC cluster experience

- 14 users active since cluster launched in spring
  - Activity of BE/ABP limited after the departure of S. Mattei
  - Long term running jobs of TE/VCS (Plasma studies) and HSE/SEE (Fire simulations)
  - The shorter queue (batch-short) is under-used
- Mvapich2 more stable MPI implementation
- Overall stable running cluster
  - 4 scheduled interventions for SLURM and cluster upgrades (pilot phase)
  - A few node crashes due to applications using up all the memory, protection mechanism added

Possible user issues

- MPI environment errors (ref. KB0004541 and how to load MPI modules)
- Job does not start (lack of free nodes): KB0004837
- SLURM queues and job parameters: KB0004973
- Again, the commands: sinfo, squeue are useful!

Will add more KB entries as we gain experience

Commissioning of new cluster

- Entering test phase in next weeks
- Would be happy to get test cases for cluster validation and performance tuning
- Hyperthreading performance v.s. physical core performance with BE applications?
  - 2 threads per core by default for MPI jobs with HT
- Scriptable examples of BE applications and benchmarks welcome!

Future plans

- Job priorities (Fairshare, check of user/accounting groups)
  - E.g. BE/ABP, BE/RF, others like TE and HSE
  - Input needed regarding reservation of resouces
- Extend cluster with new hardware (in progress)
- Improve monitoring (log files, node health….)
- Application benchmarks and templates
- Tuning of applications when sources available
- User feedback

Questions?

Contact & more info: *Service Portal and HPC service*

www.cern.ch