

October 15, 2015

**LCPQ CLUSTER
LPQSV26**



Why ?

- **Replace old clusters.**
- **Up-to-date software and operating system (goodbye centos 5).**
- **Add modern hardware (xeon haswell, hp moonshot).**
- **Upgrade network stack.**
- **Shared filesystem for each nodes.**
- **1 cluster to run them all.**



History

- **Phase 1 (April 2015)**
 - A new cluster is born.
 - sv3 integrated.
- **Phase 2 (September 2015)**
 - sv7 integrated, end of sv9.
 - Moonshot arrives.
 - 10 GbE everywhere.
- **Phase 3 (End of 2015?)**
 - sv6 integration.
 - End of Lustre, rise of BeeGFS.



LCPQ cluster's today



System software

- **GNU/Linux Centos 6.7.**
- **Linux kernel 2.6.32.**
- **Rocks clusters 6.2.**
- **Slurm 14.11.**



Storage

- **Home directory : 50GB/user. Backed up (after phase 3).**
- **/mnt/nas/\$USER : 15TB for the LAB (incoming).**
- **/mnt/beegfs : *TBA* TB shared between nodes (lustre replacement) (cluster phase 3).**
- **\$TMPDIR : today on local disks, soon on beegfs.**
- **To use a private local disk, on sbatch just add :**
`#SBATCH --gres=ioperso`



Network

- **Large part of the cluster run only on Gb/10Gb ethernet.**
- **ex-sv6 will stay on infiniband.**



Modules

- **No module loaded by default.**
- **Available modules :**
 - module avail
- **Load a module :**
 - module load my_module
- **Unload module :**
 - module unload any_module



Modules

- **Available modules today :**

- compilers/gcc-4.8.4
- compilers/gcc-4.9.2
- compilers/gcc-5.1.0
- rocks-openmpi (1.6)
- rocks-openmpi_ib (1.6)
- openmpi-i386 (1.8)
- openmpi-x86_64 (1.8)
- openmpi-moonshot (1.8)
- g09/d01 (gaussian)
- blas/openblas-v0.2.14-int32
- blas/openblas-v0.2.14-int32-openmp
- blas/openblas-v0.2.14-int64
- blas/openblas-v0.2.14-int64-openmp



Software collections

- **If you need python 2.7 and/or 3.3**

- `scl enable python27 bash`
- `scl enable python33 bash`

- **How to use in a job**

`source /opt/rh/python27/enable`

`source /opt/rh/python33/enable`



SLURM

- **Replace SGE.**
- **Queues becomes partitions.**
- **Few changes in commands and scripts.**



SLURM commands

- **Submit a job :**

`sbatch myscript`

- **Cancel a job :**

`scancel jobid`

- **Watch partition/node status :**

`sinfo`

- **How many cpus are allocated on a node :**

`scontrol show node compute-X-X`



SLURM script example

```
#!/bin/bash
#SBATCH --job-name=nom_du_job    # job name on queue
#SBATCH --nodes=4                # I want 4 nodes
#SBATCH --ntasks=32             # I need 32 MPI tasks
#SBATCH --cpus-per-task=2        # How many cpus per MPI task
# ( openMP )
#SBATCH --mem-per-cpu=2048      # 2Gb wanted per core
#SBATCH -t 48:00:00             # for 2 days (wall time, not CPU time)
#SBATCH -p xeonv3                # run on xeonv3 partition

module load openmpi-x86_64
export OMP_NUM_THREADS=2
mpirun ./my_mpi_executable > out
```



SLURM - sbatch

- **How many nodes :**
#SBATCH -N or --nodes
- **How many cpus :**
#SBATCH -c or --cpus-per-task
- **How many tasks :**
#SBATCH -n or --ntasks
- **To reserve an entire node :**
#SBATCH --exclusive
- **To share a reserved node :**
#SBATCH -s or --share



SLURM - sbatch

- **Time duration**

`#SBATCH -t` or `--time` = **days-HH:MM:SS**

- **To validate a script without calculation :**

`#SBATCH --test-only`



SLURM - scripts examples

- **Share your submission scripts !**

Here :

/share/apps/common/Examples_slurm/



Partitions

- **#SBATCH -p my_partition**
- **DefMemPerCpu=between 1 and 15Gb.**
- **11M cpu hours available per year.**



Partitions - New version

- **CLUSTER contains only xeonv1 & v2**
- **Other partitions needs to be precised**

	xeonv1	xeonv2	xeonv3	moonshot	amd128	xeonv1_ib	xeonv2_ib
CPUUs per node	16	20	24	8	48	8	20
Sockets/Cores/ Threads	2/8/2	2/10/2	2/12/2	1/8/1	4/12/1	2/4/2	2/10/2
Hyperthreading	Yes	Yes	Yes	No	No	Yes	Yes
Frequency	Average	Average	Average	Low	Low	High	Average
Memory	64Gb	64Gb	128Gb	16Gb	128Gb	128Gb	128Gb
Quantity	14	7	5	64	5	4	1
Shared	NO	NO	NO	NO	NO	YES	NO
Walltime	14 days	14 days	14 days	14 days	14 days	14 days	14 days



Partitions - Old version

- **Until today, it was :**
 - CLUSTER (unlimited walltime) and ...

	xeonv1	xeonv2	xeonv3	moonshot	amd128	xeonv1_ib	xeonv2_ib
CPUUs per node	16	20	24	8	48	8	20
Sockets/Cores/ Threads	2/8/2	2/10/2	2/12/2	1/8/1	4/12/1	2/4/2	2/10/2
Hyperthreading	Yes	Yes	Yes	No	No	Yes	Yes
Frequency	Average	Average	Average	Low	Low	High	Average
Memory	64Gb	64Gb	128Gb	16Gb	128Gb	128Gb	128Gb
Quantity	14	7	5	64	5	4	1
Shared	YES	NO	NO	YES	NO	YES	NO
Walltime	7 days	7 days	7 days	7 days	7 days	7 days	7 days



More infos

- [**http://lpqsv26.ups-tlse.fr/lcpq/**](http://lpqsv26.ups-tlse.fr/lcpq/)
- [**http://lpqsv26.ups-tlse.fr/ganglia/**](http://lpqsv26.ups-tlse.fr/ganglia/)



Questions ?

