Partitions and QoS

Partitions and QoS

- Mea Trahan
- *Email: Daniel.Trahan@Colorado.edu*
- *RC Homepage: https://www.colorado.edu/rc*
- *RC Email: rc-help@colorado.edu*

- Slides available for download at:
- *https://github.com/ResearchComputing/Partition_and_QoS_Fall_2021*

Outline

- What is a Partition
- Summit Partitions
  - shas
  - sgpu
  - smem
  - Others
- Summit QoS
- Using Partitions and QoS in a Job

Quick note

- Clusters come in all shapes and sizes, so much of the information here may not apply to other systems. HPC is a very diverse landscape so make sure you check with the system administrators of whichever cluster you are using.

What is a Partition

- Research Computing offers a variety of different node types for users to utilize.
- A **partition** is a grouping of node types that offer specialized features that are not found in our most common node.
  - Each node type offers different resource limits.
- By default the partition is "`shas`" is selected and should be used for most jobs.
- For a full guide of Partitions, checkout our docs!
  - https://curc.readthedocs.io/en/latest/running-jobs/job-resources.html#partitions

The `shas` partition

- Summit's most common partition for users.
  - `shas` = Summit Haswell Node
  - Selected by default if no partition is provided
- Node Details
  - Node Count: 380 Nodes
  - Core Count: 24 cores
  - Memory limit: 4.84 GB per Core (Max: 116 GB)
- Queue wait: Usually very active but little wait if Allocation is not heavily used.

The `sgpu` partition

- Summit's specialized GPU partition.
- Each node is equipped with a 2 Tegra K80 Nvidia GPUs
- Not specialized for Machine learning!
- Node Details:
  - Node count: 10
  - Core count: 24
  - Memory: 4.84 GB per core
- GPU Details:
  - 4992 CUDA Cores
  - 24 GB of Graphical RAM
- Queue wait: Can vary substantially depending on time of year.

The `smem` partition

- Summits specialized High Memory resource
- Very useful for applications that demand high memory resource usage with no cross node communication.
- Max wall time is 7 Days by default.
- Node Details:
  - Node count: 5
  - Core count: 48
  - Memory: 42.7 GB per core (Max 2 TB)
- Queue wait: Usually congested, one of Summit's most demanded resources.

Partition Variants

- The Shas partition has several variants that we make available for users wanting for faster Queue times. These partitions have specialized limits of use:
  - `shas-testing`: Special high priority partition for shas testing jobs
  - `shas-interactive`: Special high priority partition for shas interactive jobs
  - `sgpu-testing`: Special high priority partition for sgpu testing jobs
  - `sknl-testing`: Special high priority partition for smem testing jobs
- Testing Partitions: Max 30 min runtime @ 24 cores
- Interactive Partitions: Max 4 hour runtime @ 1 core

Other types of Partitions

- Summit Condo Buy-Ins: ssky and ssky-preemptable
  - Summit Skylake nodes
  - Node Count: 5 nodes and 15 nodes on ssky and ssky-preemptable respectively.
  - Core Count: 24 cores per node.
  - Memory: 7.68 GB per core.
- Summit KNL Nodes
  - Summit Phi (KNL) nodes
  - GPU in a CPU: allows high cores per node.
  - Performance was not as good as expected and Intel discontinued.
  - Still around for your use!

What is a QoS

- Summit also leverages a mechanism called QoS to change certain limitations imposed on your Jobs.
- **QoS** or **Quality of Service** is a field that constrains or modifies certain parameters of your job script.
  - Changes Priority of jobs so you may wait shorter or longer depending on the application.
  - Can also change limitation on how many nodes can be run on the system
- Do not set a QoS unless you need the QoS modifications.
- Ex: Biggest limitation
  - By default Summit jobs cannot run longer than 24 hours.
  - Fix: The long QoS!

The Normal QoS

- Summit's base QoS set by default.
- QoS Parameters:
  - Sets a max walltime on all partitions (except smem) to 24 Hours.
  - 1000 Jobs can be submitted per user.
  - Up to 256 Nodes can be requested per Job.

The Long QoS

- Specialized QoS allowing for longer runs on Summit
- QoS Parameters:
  - Sets a max walltime on shas, sknl, and ssky to 7 Days
  - Maximum of 200 Jobs can be submitted at a time
  - Only 22 nodes maximum can be requested per job. Maximum of 40 Nodes total throughout Summit.
- Generally try to avoid this! Queue times can be quiet substantial so keep your jobs smaller to avoid the wait.

Blanca Buy-ins

- Condo buy in cluster
- Load the `slurm/blanca` module to see
- To access your node set QoS to: `blanca-<name-of-partition>`
- QoS Parameters:
  - Max priority against anyone outside of other owners of the nodes.
  - Max wall time varies but usually 7 days.
  - No limit on max jobs, but limited to your nodes.

The Preemptable QoS

- Specialized QoS allowing running on ssky condo node on Summit
- QoS Parameters:
  - Sets a max walltime on all partitions to 24 hours
  - Maximum of 1000 Jobs can be submitted at a time
  - No limit on max jobs, but jobs can be preempted if condo user requires the node.
- Very useful for jobs with progressive checkpointing that need Skylake instructions.
- Blanca users can utilize preemptable to run jobs on idle nodes.

Using QoS and Partition in Jobs

- To utilize a Summit Partition or QoS, simply add the following sbatch directives to your job scripts:

```
#SBATCH --partition=<desired-partition>
#SBATCH --qos=<desired-qos>
```

- On interactive jobs, this can be set as such:

```
sinteractive --partition=<desired-partition  \
--qos=<desired-qos>
```

Reminder: Never set QoS unless you wish to utilize different resource limits for your desired partition.

Example 1:

```
#!/bin/bash

#SBATCH --nodes=1
#SBATCH --ntasks=24
#SBATCH --time=04:00:00
#SBATCH --partition=shas
#SBATCH --job-name=gpu-job-ex
#SBATCH --output=gpu.%j.out

# Load Modules
module load cuda/10.1

# Run Application
./GPUApp
```

Example 2:

```
#!/bin/bash

#SBATCH --nodes=5
#SBATCH --ntasks=120
#SBATCH --time=48:00:00
#SBATCH --partition=shas
#SBATCH --qos=long
#SBATCH --output=long.%j.out

# Load Modules
module load intel impi

# Run Application
./longapp
```

Some Common issues

- **Problem:** My shas-interactive or shas-testing jobs won't run!
- **Solution:** Usually this is because you have reached the job limits of that partition. Since those two partitions have single job limits, try checking to see if you have an existing job running.

- **Problem:** I'm trying to run a high memory node job for 7 days but it taking forever!
- **Solution:** You might have set QoS to long in your job script. Remember high memory node jobs allow for 7 day run time by default.

Questions?

Thank you!

- Please fill out the survey: http://tinyurl.com/curc-survey18
- Contact information: rc-help@Colorado.edu
- Slides:
  https://github.com/ResearchComputing/Partition_and_QoS_Fall_2020
- Documentation:
  https://curc.readthedocs.io/en/latest/running-jobs/job-resources.html