WAKE FOREST
UNIVERSITY

SLURM

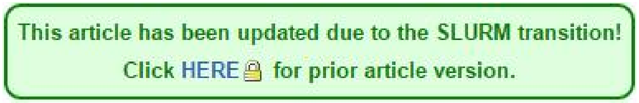
(Simple Linux Utility Resource Manager)

Status Update, Testing, and Timeline

High Performance Computing Team – Information Systems

Completed Tasks

- New scheduler server (rhel6slurm) fully operational
- User wiki updated:
  - Updated all pages with Torque content
  - Changed pages contain a banner linking to legacy versions during transition
  - Primary SLURM documentation pages:
    - https://wiki.deac.wfu.edu/index.php/SLURM
    - https://wiki.deac.wfu.edu/index.php/Category:SLURM
  - SLURM examples, commands, templates, and reference links available


This article has been updated due to the SLURM transition!
Click HERE 🔒 for prior article version.

- Conversion script
  - Convert any .pbs, .job, or .SC file to SLURM

    ```
    $ pbs2slurm.py -i <INPUT>
    ```
  - Full help showing options (-h), and preview mode available (-p)

Completed Tasks

- Group and user accounts created
- Partition integration successfully tested

Name	Priority	Node Limit	Time Limit
Debug	10	None	< 6 hours
Small	40	1 node	< 1 day
Medium	30	8 nodes	< 1 week
Large	20	None	< 1 year
Infiniband*	50	IB nodes only	< 1 year
GPU*	50	GPU nodes only	< 1 year

- Non-asterisk partitions submit to all compute nodes
- Submission testing of non-specialized software
- Server scripts and custom email notifications in place

Testing

- Ready for your testing!
- Submit jobs to SLURM only from **rhel6head4**

```
[bc103bl14 ~] $ sbatch myjob.slurm
```

- Current compute nodes available for testing

• 1G:	BC03
• 10G:	UCS Chassis 10/11
• Infiniband:	BC02BL02, BC02BL12
• GPU:	gpu01-05

- Need help with specialized software testing
  - NAMD
  - LS-Dyna
  - Infiniband specific jobs
  - GPU specific jobs

Changes: Job Prioritization

- Simplified job priority equation:

```
        (PriorityWeightAge) * (20) +
   (PriorityWeightJobSize) * (job_size_factor) +
 (PriorityWeightFairshare) * (100) +
       (PriorityWeightQOS) * (QOS_factor) +
 (PriorityWeightPartition) * (1000)
```

- Major change: Partition weight is primary determining factor
  - Encourages desired job types, many small form factor and parallelization
- Priority flag, “Small relative to time” enabled, favors full node consumption
- In-depth information:
  - http://slurm.schedmd.com/priority_multifactor.html

Other Changes

- Each user has an account created in the SLURM account database
  - Allows for individual fairshare weight assignment if desired
- Parent group fairshare inheritance
  - Users inherit fairshare weight from group accounts
- Topology mapping
  - SLURM has configured knowledge of chassis network connectivity
  - Allows for future expansion into multiple datacenters
- Infiniband job submission
  - Submit to “Infiniband” partition
  - Include the “switches=1” directive
  - https://wiki.deac.wfu.edu/index.php/SLURM_Job_Script_Templates#Parallel_Job_Type_2_-_Infiniband

Timeline

- Recommended testing procedure
  - Start with short, simple “hello world” type job submissions
  - Use the conversion script to test previously run jobs
    - Start simple, work up to complex
  - Run actual jobs when ready
- Test expansion
  - Compute nodes can be allocated for SLURM on demand
  - A pre-existing head node will be converted when required
- Transition
  - Goal is start of Spring Semester to be using SLURM
  - Flexible based upon feedback and results
- Process is in place to disable and remove Torque when ready
  - Tested successfully on UCS chassis 11

Questions?