WAKE FOREST
UNIVERSITY

SLURM

(Simple Linux Utility Resource Manager)

DEAC transition and implementation

High Performance Computing Team – Information Systems

Why use it?

- Experiencing repeated service crashes and errors with current installation
  - SLURM replaces both the Resource Manager (Torque) and Scheduler (Maui) service
- Open source and widely supported, with Enterprise level backing from SchedMD
  - Torque/Maui support has been lacking for many years
- Multiple benefits:
  - Enhanced Scheduler Performance
    - Up to 1000 job submissions per second and 500 job executions per second
  - Flexible Scheduling
    - More options for job prioritization
  - Topology Optimized Allocations
    - Can define network level configurations to minimize network latency
  - Generic Resource Support (GPUs)

Who uses it?

- Becoming the standard in HPC
- Used in 60% of the world's TOP500 super computers
  - Including the world's fastest supercomputer, Tianhe-2
- Many Universities:
  - Texas A&M
  - University of Michigan
  - NC State
  - BYU
  - Many, many, more!
- Coming soon: Wake Forest University

When will we do it? (tentatively)

- Desired implementation before start of Fall Semester (August 25)
  - *Requesting feedback on dates, time desired before classes, conflicts, etc.*
- Transition pre-requisites
  - Complete Admin testing (June 15)
    - Using 1 head node, two infiniband nodes, two tengig nodes, one 1G node, and GPU node
  - Complete User testing (July 31)
    - Includes testing the SLURM-torque plugin
  - Deliver documentation to all users (July 17)
    - At least one-month before transition date
- Push SLURM to all remaining nodes (August 17)
- Torque/Maui removal/retirement (September 18)

What we need to know?

- Need volunteers for testing behavior of scheduler
  - Starting in June
  - Submit jobs to test nodes
- Need to determine the following things with your input:
  - What to do with Torque scripts
  - Desired behaviors for job prioritization
    - Considering currently used settings
    - Considering previous discussions
    - Considering options available

What we need to know? (Torque scripts)

- When SLURM is in place, what to do with pre-existing Torque scripts?
  - Utilize SLURM-torque plugin?
    - Allows torque scripts to be submitted to queue
      - No additional steps necessary
    - Compatible with ~95% of torque syntax
    - Cannot be installed when Torque is installed
  - Utilize a Torque to Slurm conversion script?
    - Create SLURM script from provided input Torque script
      - Not compatible with some torque syntax
      - Not compatible with multi-line commands
    - Once created, can be submitted like any other SLURM script
    - Allows Torque to remain installed
  - Other thoughts?

What we need to know? (Job prioritization)

- SLURM job prioritization considers five categories:
  - Fairshare
    - Determined by funding contributions to the cluster
    - Considers usage over the previous 30 days
  - Age
    - Age of job since submission
    - Up to specified number of day max, currently 7 days in Torque
  - QOS
    - Allows user specified ranking of importance for job
    - Not currently implemented in Torque
  - Job Size
    - Considers number of nodes or CPU a job requests
    - Not currently implemented in Torque
  - Partition
    - Configured as a specific group of nodes configured to only execute certain job types
    - Can define many partitions, each with different weights, that share the same nodes
    - Not currently implemented in Torque

What we need to know? (Job prioritization)

- What prioritization behavioral options do you want?
  - Previous repcom meeting referenced separation based on walltime
- Behavior examples:
  - Set job size preference
    - Large vs small job size (based on CPUs or Nodes requested)
  - Reward efficient use of resources
    - Consume all CPUs on a node
    - CPUs consumed per node
  - Favor short walltime jobs
  - Number of nodes requested
  - Limit enforcement on research groups
    - Set max jobs and/or CPUs consumed by group at a given time
  - Node exclusivity
    - Separate nodes for specific job types and/or functions

Questions?

Feel free to:

- Send in follow-up questions after the meeting
- Request documentation be added to the wiki
- Ask for supplemental links
- Etc.