Agenda

- Accounting and resource management capabilities using SlurmDBD (SLURM DataBase Daemon)

- SLURM Version 2.1 capabilities and deployment plan

- SLURM Version 2.2 and beyond

- Round table discussion
  - Feature requests
  - Usage models
SLURM database integration

- Store job accounting information
  - Use sacct and sreport tools to view

- Manage user/bank account information
  - Grant users access to bank resources
  - Apportion resources by user and bank
  - Many limits available by user and bank
  - Changes propagate in real-time to slurmcctld
  - Use sacctmgr tool to view and modify information

- One database can serve all computers at one site
Hierarchical banks, fair share example

Job priorities are altered to achieve target fair share allocations.

Numbers automatically normalized at each level in the bank hierarchy.

- **Bank: root**
  - 100 shares

- **Bank: science**
  - 60 shares

- **Bank: business**
  - 40 shares

- **Bank: chemistry**
  - 80 shares

- **Bank: physics**
  - 20 shares

- **User: alice**
  - 60 shares

- **User: brian**
  - 20 shares

- **User: cheryl**
  - 10 shares

- **User: david**
  - 40 shares

Users can access multiple banks (per configuration).

S&T Principal Directorate - Computation Directorate
Hierarchical banks, resource limits

Many resource limits can be set by individual users or bank.

Bank “coordinators” can be identified for each bank and modify child banks and users as desired.

Bank: root
  - no time limit
Bank: science
  - 24 hour job limit
Bank: business
  - 1 hour limit
Bank: chemistry
  - 24 hour job limit
Bank: physics
  - 20 hour job limit
User: alice
  - 24 hours
User: brian
  - 12 hours
User: cheryl
  - 20 hours
User: david
  - 10 hours
User: david
  - 1 hour limit
Hierarchical banks, resource limits (continued)

Some resource limits apply to all child banks and users (they are aggregated)

- **Bank: root**
  - no job limit

- **Bank: science**
  - 20 job limit

- **Bank: business**
  - 100 job limit

- **Bank: chemistry**
  - 20 job limit

- **Bank: physics**
  - 12 job limit

- **User: alice**
  - 20 jobs

- **User: brian**
  - 20 jobs

- **User: cheryl**
  - 12 jobs

- **User: david**
  - 5 jobs

One user can have different limits in different banks
Sample Job Scheduling Architecture

SLURM (cluster 1) -> SLURM administration tools

User and bank info -> Accounting data

SlurmDBD -> MySQL

User account and limit info
Job and step accounting info

Jobs & status -> SLURM (cluster N)
SLURM Version 2.1 - Highlights for Users

- “—signal=<int>@<time>” option added all job submit commands. Notifies job when time limit approaches

- “—start” option added to squeue to report expected initiation times of pending jobs (requires backfill scheduler plugin to set times)

- New job wait reason added: ReqNodeNotAvail, required node not available (down or drained)

- “—detail” option added to scontrol show job to display CPU/memory allocation information node-by-node
Support for QOS (Quality of Service) added to accounting database with configurable limits, priority and preemption rules

Gang scheduler moved into slurmctld, can be used with backfill scheduler

Preempted jobs can be suspended/resume, requeued, checkpointed, or cancelled

Pam_slurm Pluggable Authentication Module now distributed directly with SLURM
SLURM Version 2.1 - More for System Administrators

- Sacctmgr show problems command added to report database anomalies (e.g. banks with no users, etc.)
- Support added for overlapping advanced reservations
- Support added for OpenSolaris
- Scalability of svview dramatically improved
- Upward compatibility for RPCs and state save files for future major releases
- Many enhancements for BlueGene systems
SLURM Version 2.1 – Release plans

- Development wrapping up
- Testing at scale underway this week
- Stable versions available from the “under_development” folder
- Release planned for December
SLURM Version 2.2 – Plans

- To be released in the summer of 2010

- Cross-cluster command support
  - Login to one cluster and view state of other clusters and submit jobs to other clusters
  - Destination cluster must be explicitly named (e.g. “sbatch –cluster=tux my.job”)
  - NOT enterprise-wide scheduling

- Improved support for task affinity
Other work planned

- Port to BlueGene/Q
- Improvements in fault-tolerance for jobs (e.g. hot-spare nodes)
- Linux containers to better control a job’s available memory
Customer feedback

- Feature requests
- Usage models
- Open discussion

One more slide…
For more information

- Visit SLURM web site
  - https://computing.llnl.gov/linux/slurm

- Visit NNSA boot. Demos scheduled at
  - Day, Time, General SLURM use
  - Day, Time, SLURM database use
Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process discloses, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trademark, manufacture, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.