SLURM Community Meeting
SC 2009
November 2009

Morris Jette (jette1@llnl.gov)

Danny Auble (auble1@llnl.gov)

Don Lipari (lipari1@llnl.gov)

S&T Principal Directorate - Computation Directorate

LLNL-PRES-402832

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

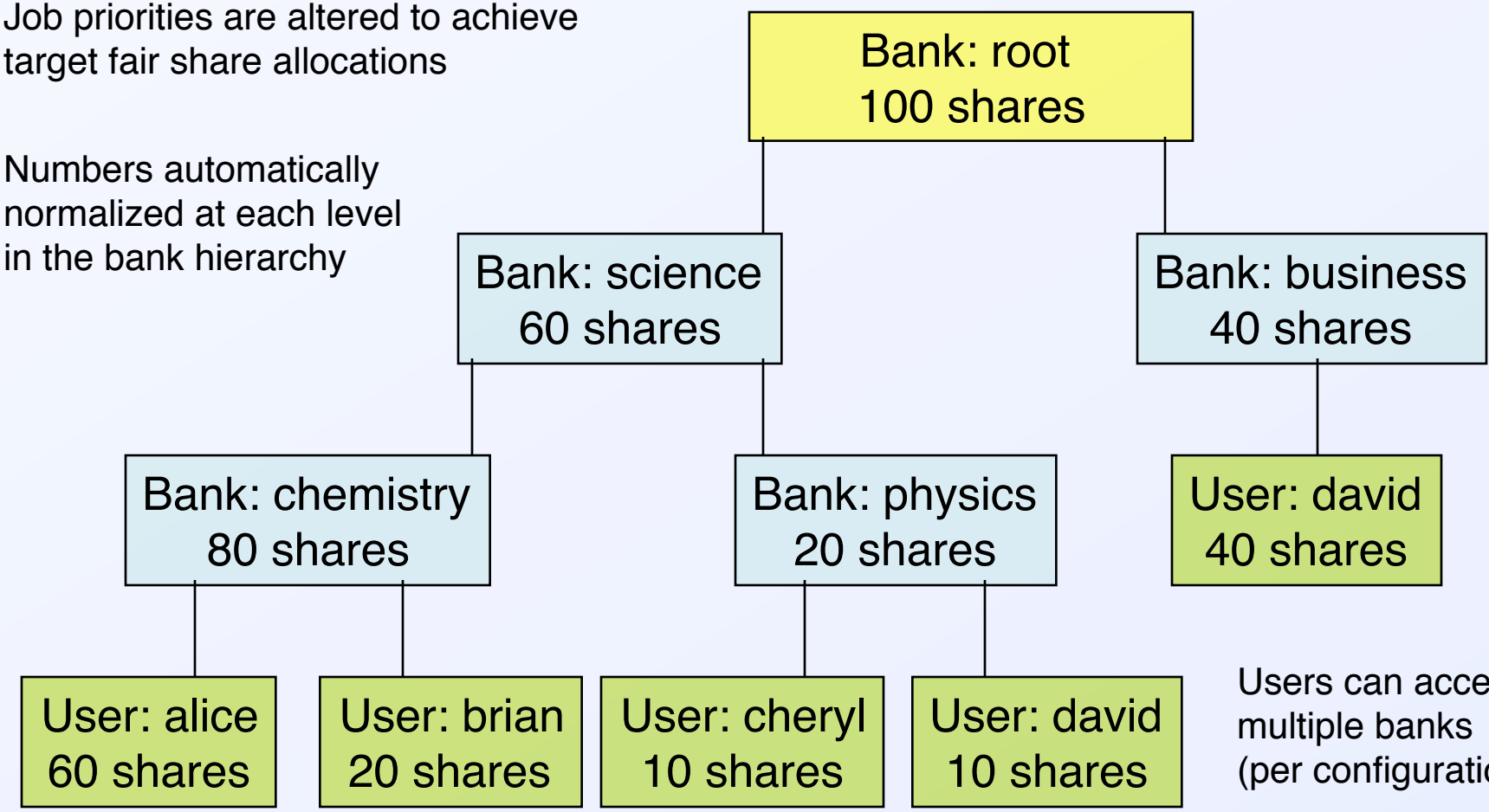
SLURM database integration

- 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 slurmctld
  - 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

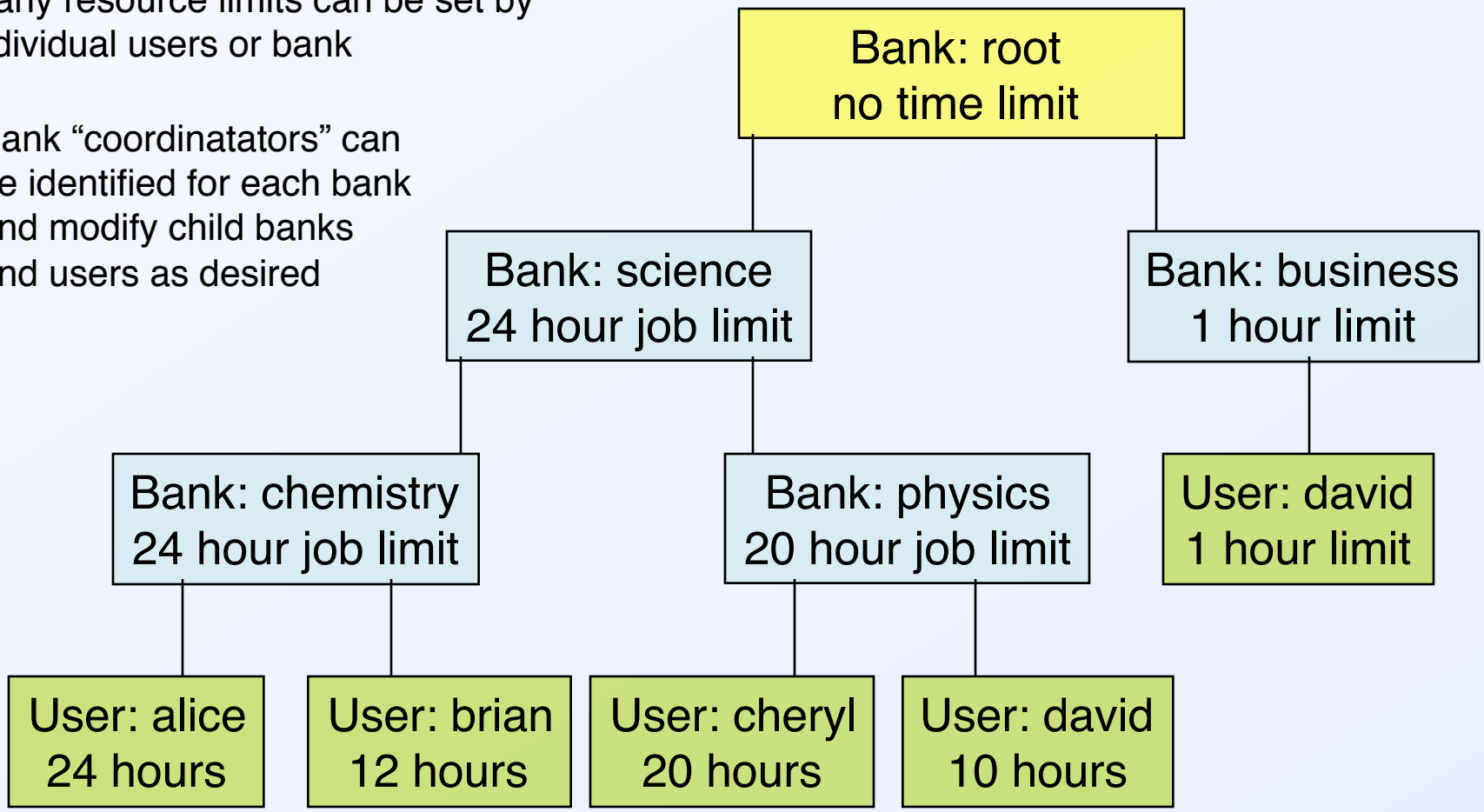


Users can access multiple banks (per configuration)

Hierarchical banks, resource limits

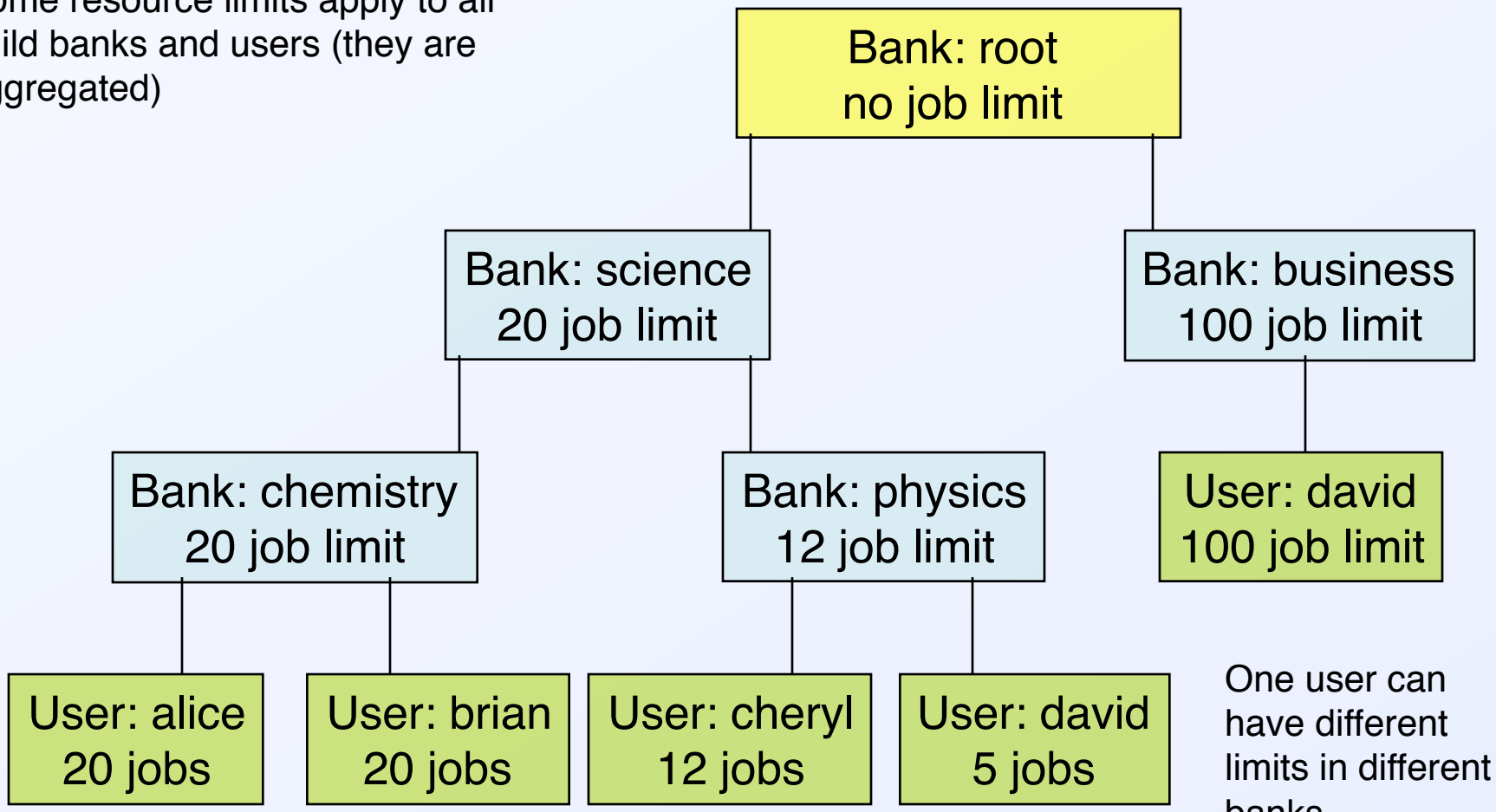
Many resource limits can be set by individual users or bank

Bank "coordinatators" can be identified for each bank and modify child banks and users as desired



Hierarchical banks, resource limits (continued)

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



One user can have different limits in different banks

Sample Job Scheduling Architecture

SLURM administration tools

SLURM (cluster 1)

User and bank info

Accounting data

SlurmDBD

Jobs & status

SLURM (cluster N)

MySQL

User account and limit info
Job and step accounting info

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

SLURM Version 2.1 - Highlights for System Administrators

- 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 sview 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 planed

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