

SLURM Version 1.3

May 2008



Morris Jette (jette1@llnl.gov)

Danny Auble (auble1@llnl.gov)

S&T Principal Directorate - Computation Directorate

Disclaimer



Major Changes in Slurm Version 1.3 Include

- Major changes in user commands
- Job accounting logic largely re-written and integrated with a database
- Major enhancements to job scheduling including support for gang scheduling (time-sharing for parallel jobs)
- See `RELEASE_NOTES` for a more complete description of changes

Command Changes

- 's , , and options removed. Use and commands instead. Most options are consistent across commands
- command removed. Use command instead
- option added for job steps
 - Provides resource management within job allocation for multiple concurrent job steps
- Feature counts added for job constraints
 -

Command Changes (continued)

- option added for pseudo-terminal support
- Time specification is more flexible:
 - <minutes> OR
 - <minutes>:<seconds> OR
 - <hours>:<minutes>:<seconds> OR
 - <days>-<hours>:<minutes>:<seconds>
- Much richer job dependency support:
 - Each job can be dependent upon many other jobs
 - Several dependency types added: Wait for other job to begin, complete successfully (exit code of zero), fail or complete (any exit status)

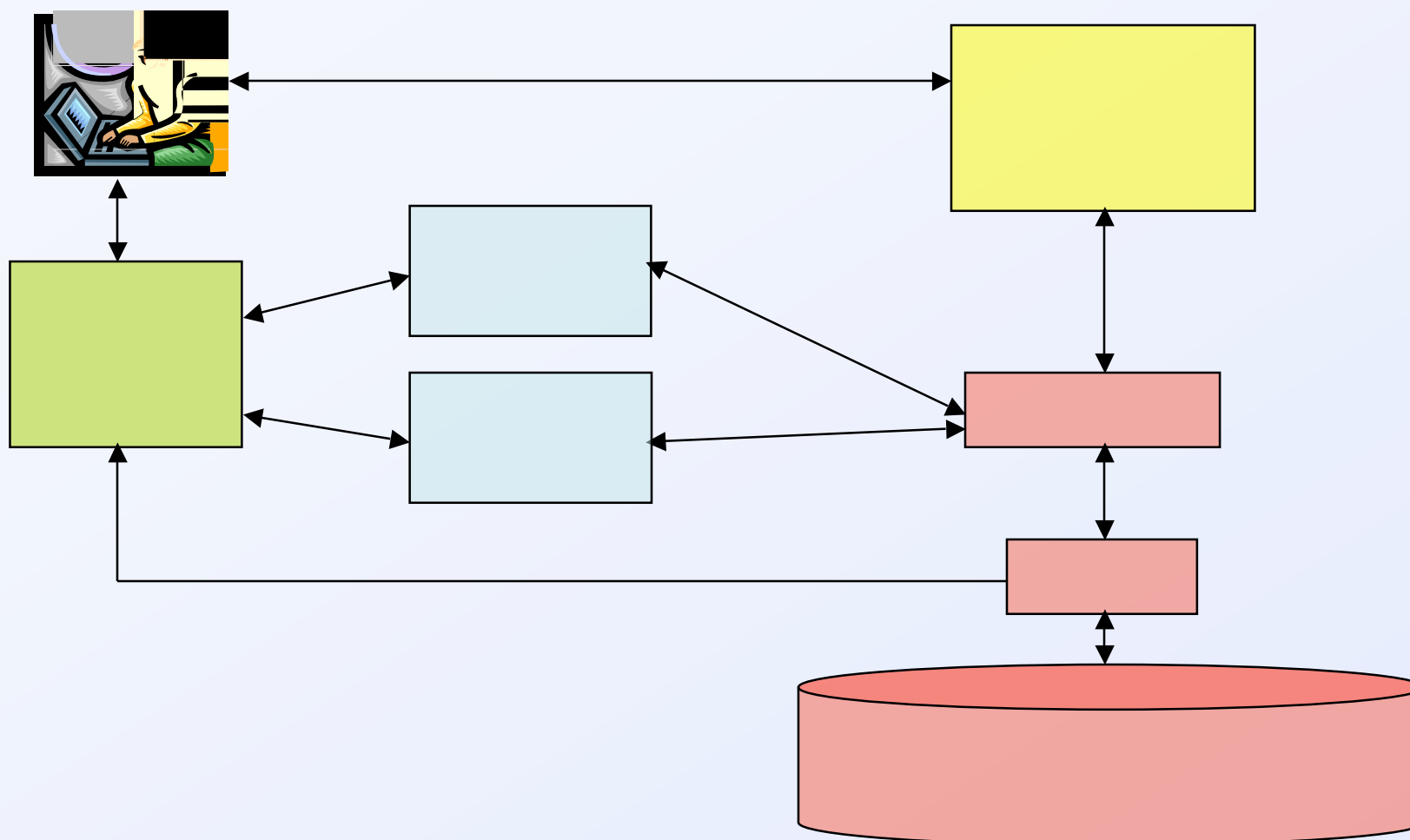
Accounting Changes



- Job accounting split into two components

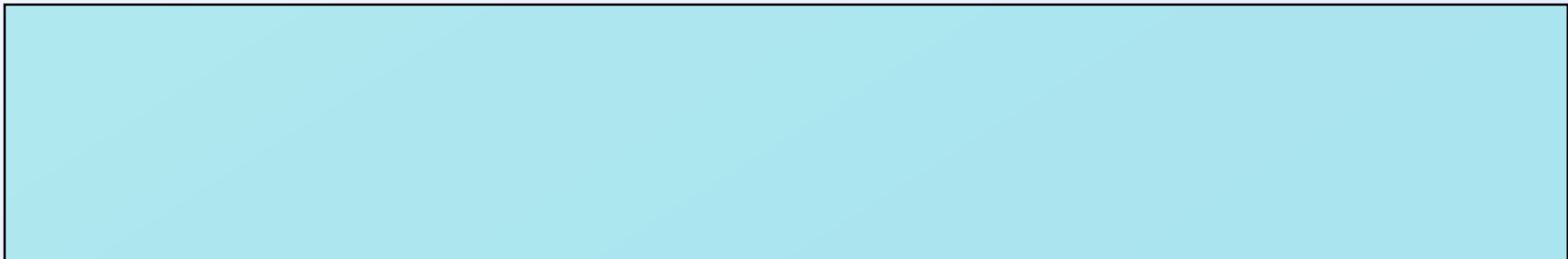


Sample Accounting and Workload Scheduling Architecture



Scheduling Changes

- Backfill scheduler plugin re-written to support all configurations and job options
- Partitions have parameter
 - Partitions can have overlapping nodes, but differing user, time, and size limits so they are really queues
- Partitions have a count of how many jobs can share an allocated resource (node, socket, core, etc. depending upon and)



Scheduling Changes (continued)

- Added support for cluster-wide consumable resources (e.g. licenses, added in v1.3.1)
- Many enhancements for Moab and Maui schedulers
 - New job and node state information managed
 - Slurm partitions and their jobs can be scheduled without Moab or Maui interaction for better responsiveness without scheduling policy support)

Gang Scheduling Support Added

- Gang scheduling support added to time-slice parallel jobs for improved responsiveness and utilization
- Jobs in the same partition sharing resources will alternately be suspended and resumed so all jobs make progress
- Jobs in lower priority partitions can be preempted (suspended) to execute jobs in higher priority partitions. Suspended jobs will automatically be resumed when idle resources are available
- Options and configuration parameters added to avoid memory over-subscription

Gang Scheduling Example

Time	Node 0	Node 1	Node 2	Node 3
0	Job A	Job A	Job A	Job A
1	Job B	Job B	Job C	Job C
2	Job D	Job D	Job D	Job E

Other Recent Changes

- Added support for periodic node health check (see [\[link\]](#) and [\[link\]](#))
- Added response logic for non-killable processes (see [\[link\]](#) and [\[link\]](#))
- Configurable default job behavior on node failure (requeue or kill, see [\[link\]](#))
- Perl APIs and PBS/Torque command wrappers added (in v1.2.13)
- Event trigger support added (e.g. run some script when specific or any nodes goes DOWN, added in v1.2.2)