

SLURM Database Use Accounting and Limits



Morris Jette
jette@schedmd.com

SchedMD LLC

Outline



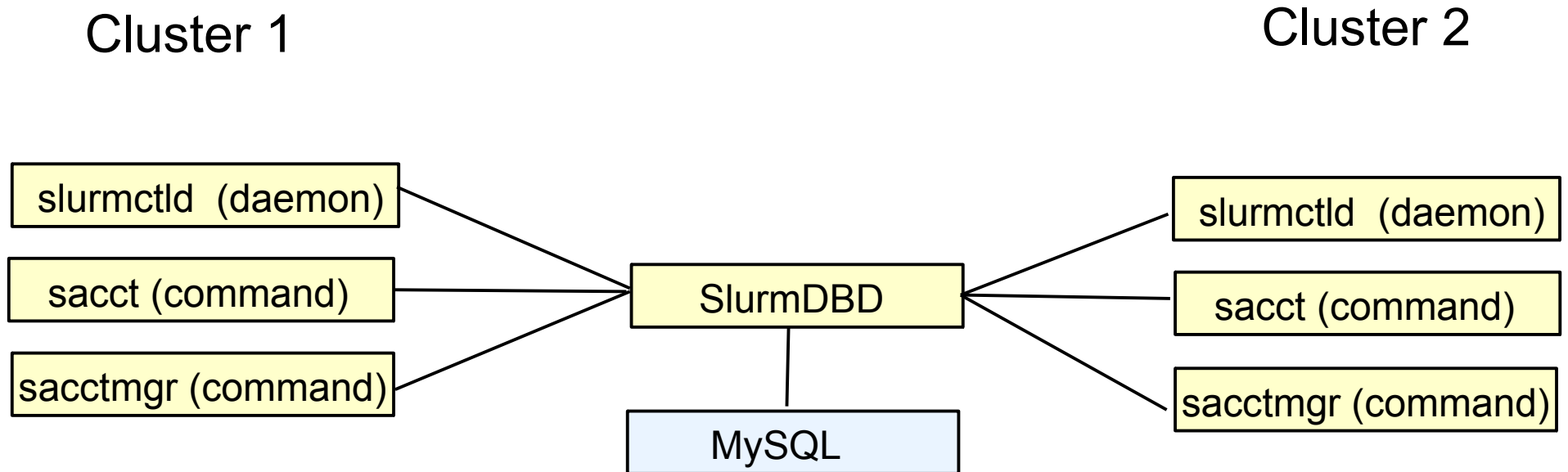
- System architecture for database use
- Accounting commands
- Resource limits
- Fair-share scheduling
- Accounting configuration

Architecture



- It is recommended to maintain one database containing the information about all computers and users at a site
 - One database per computer is possible, but increases the maintenance effort
- MySQL is strongly recommended
- Data maintained by user name
 - A uniform mapping of user name to ID across all computers is strongly recommended

Architecture



SlurmDBD



- SlurmDBD == SLURM DataBase Daemon
- An intermediary between user and the database
 - Avoid granting users direct database access
 - Authenticate communications between user and slurmdbd (using Munge)
 - Only slurmdbd needs permission to read/write the database
- Pushes update information out to slurmctld on the clusters
- slurmctld daemon will cache data if slurmdbd not responding

Association



- Association is a combination of cluster, account, user name and (optional) partition name
- Each association can have a fair-share allocation of resources and a multitude of limits
- NOTE: Each account name must be unique. The name can not be repeated at different points in the hierarchy

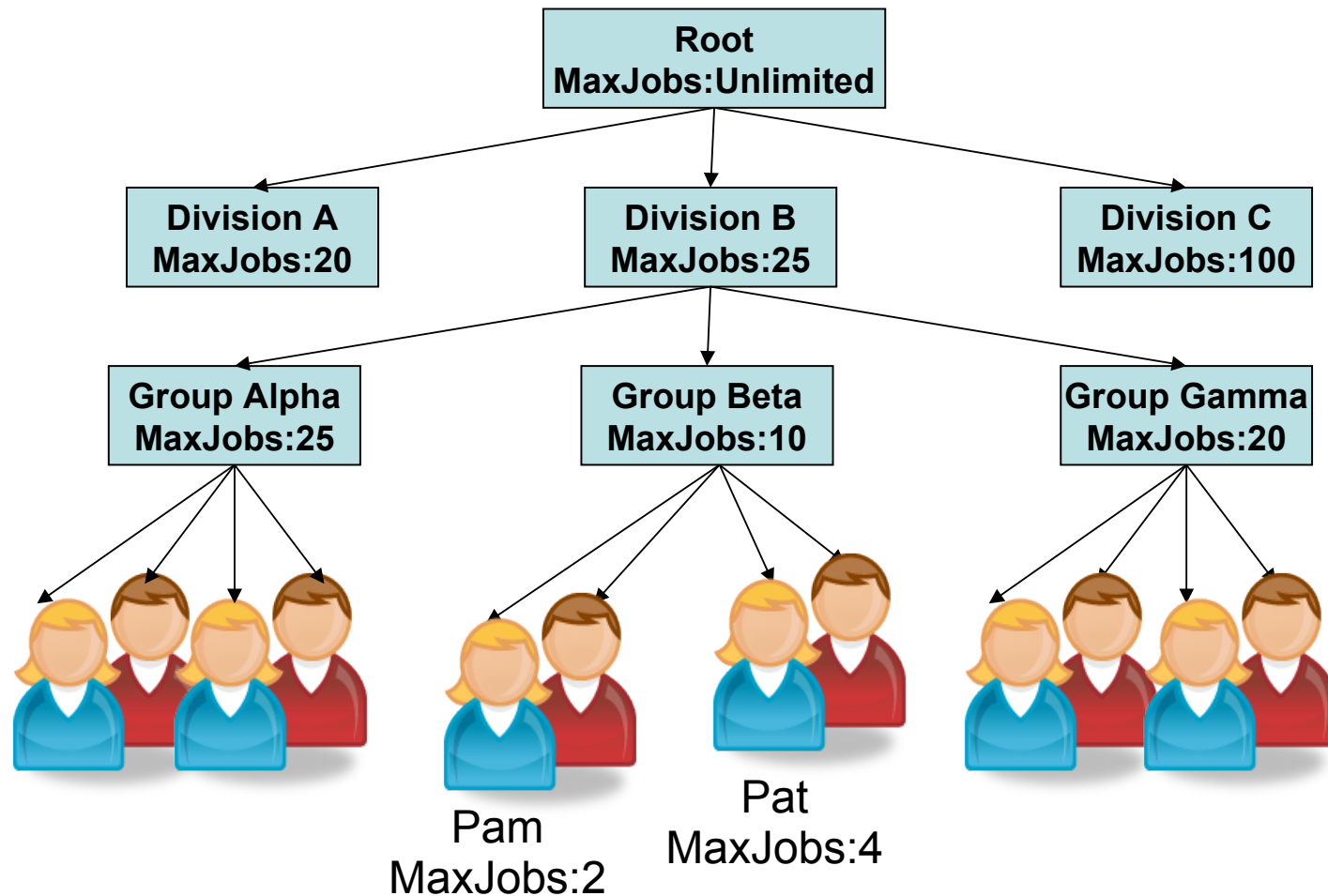
User=Pam Account=Beta FairShare=20% MaxTime=2hours MaxJobs=2 etc.

Account Coordinator



- Users given permission to add users or sub-accounts, modify fair-share and limits to the accounts and users they are coordinator over
 - Limits of their association can not be increased, but users and sub-accounts of that association can increased up to that limit

Hierarchical account example



sacctmgr command



- sacctmgr used by privileged users to view and modify the SLURM database
 - Manage clusters
 - Manage accounts
 - Manage users
 - Manage limits

sacctmgr examples of use

```
sacctmgr add cluster tux
```

```
sacctmgr add account science Description="science" Organization=science
```

```
sacctmgr add account chemistry,physics parent=science \  
Description="physical sciences" Organization=science
```

```
sacctmgr add user adam DefaultAccount=chemistry
```

```
sacctmgr show account
```

```
sacctmgr modify association where user=adam account=chemistry set MaxJobs=2
```

Job Accounting Commands



- `sacct` – Generates detailed accounting information about individual jobs or job steps
 - Filtering options by user, computer, partition, time, etc.
- `sreport` – Generates aggregated accounting reports
 - Reports resource usage by user, cluster, partition, etc
 - Data not reported about individual jobs or steps
- `sstat` – Generates very detailed accounting report about individual currently running job or job step

Resource Limits



- Association-level limits: Applies to an association and all children (e.g. account Beta)
 - GrpCPUMins, GrpCPUs, GrpJobs, GrpMemory, GrpNodes, GrpSubmitJobs, GrpWall
- Per job limits
 - MaxNodesPerJob, MaxWallDurationPerJob
- Can be used to set limits on a finer-grained basis than SLURM partition limits

Job Prioritization

- *PriorityType=priority/multifactor* enables job priorities to be based upon several factors with the weight of each tunable
 - Age
 - *PriorityMaxAge*
 - Job size
 - *PriorityFavorSmall*
 - Fair share
 - Partition/queue
 - QOS

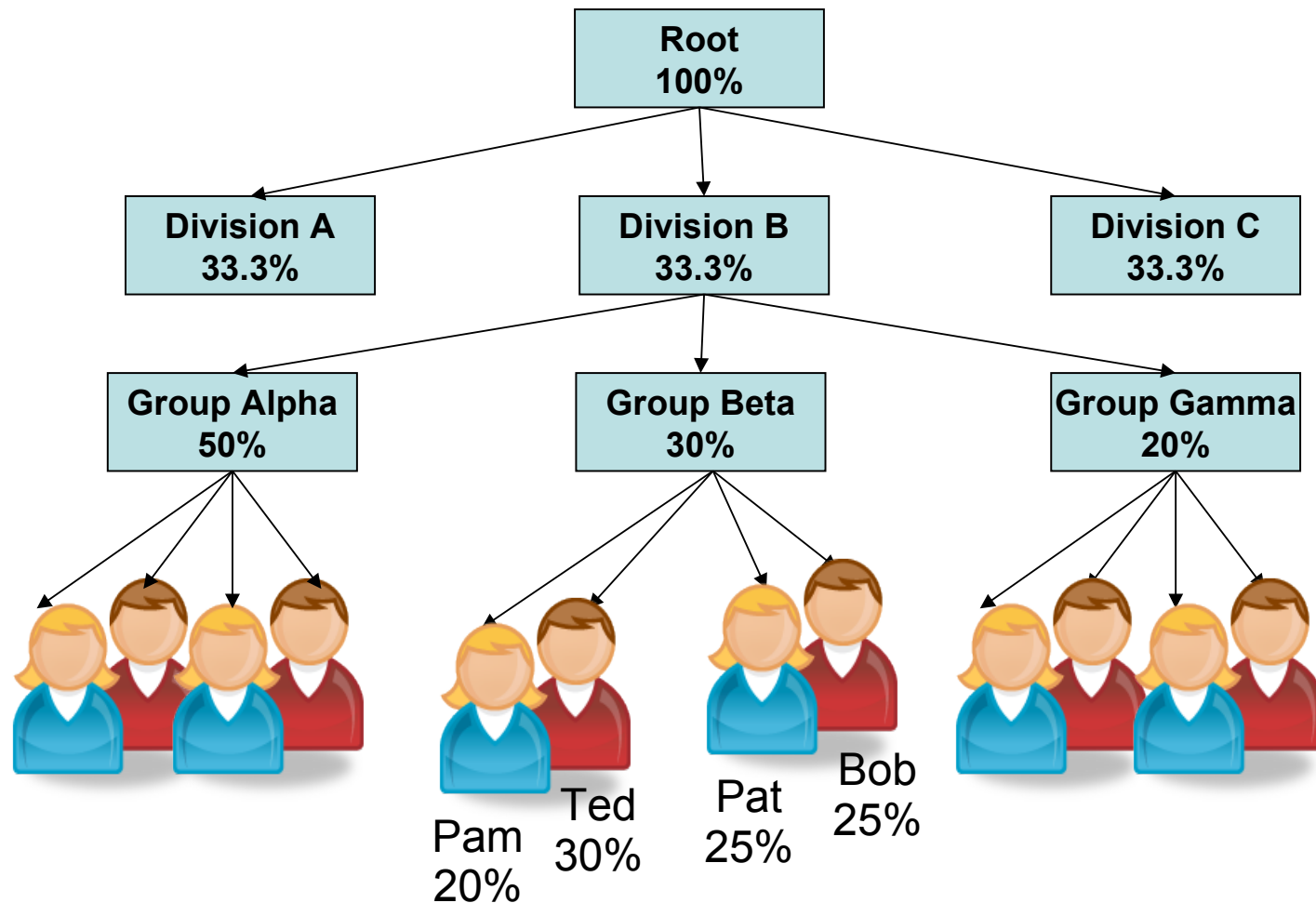
```
Job_priority =  
(PriorityWeightAge) * (age_factor) +  
(PriorityWeightFairshare) * (fair_share_factor) +  
(PriorityWeightJobSize) * (job_size_factor) +  
(PriorityWeightPartition) * (partition_factor) +  
(PriorityWeightQOS) * (QOS_factor)
```

Fair Share Scheduling



- Measure of how over- or under-served a user and each account is
 - Computed at each level in the account hierarchy and combined into a single value
 - If only one user in an account is active, that user will have access to all resources in that account
 - Configurable *PriorityDecayHalfLife* controls how long resource usage history is relevant
 - Fair share can be disabled at specific points in the hierarchy (e.g. to ignore the resource use by individual users of an account): *FairShare=parent*

Hierarchical account example



Hard Resource Limits

- Alternative to fair share
- Configuration parameters:
 - *PriorityHalfLife=0* (no decay)
 - *PriorityUsageResetPeriod=*
 - *Yearly*
 - *Quarterly*
 - *Monthly*
 - *Weekly*
 - *Daily*
 - *Now*
 - *None* (default)

Configuration – slurm.conf

- *ClusterName=...*
 - Name of cluster for accounting purposes
- *JobAcctGatherType=jobacct_gather/linux*
 - Define how to gather job accounting information
- *AccountingStorageType=accounting_storage/slurmdbd*
 - Define where to record job accounting data
- *JobCompType=jobcomp_none*
 - Define where to record job completion data
 - Redundant if job accounting enabled

Configuration – slurm.conf (continued)




- *AccountingStorageEnforce=...*
 - Associations – prevent the running of job unless user and account defined in the database
 - Limits – prevent user from exceeding user or account limits. Automatically sets associations to be enforced
 - QOS – Require all jobs to use valid QOS (Quality Of Service). Jobs must specify QOS or use their default.

Configuration – slurmdbd.conf

- *AuthType=auth/munge*
 - Define how authenticate communications
- *StorageType=accounting_storage/mysql*
 - Define where to record job accounting data
- *StorageUser=...*
 - User name used to access the database
- *StoragePass=...*
 - Password used to access the database

Configuration – slurmdbd.conf (continued)



- *PrivateData=...*
 - Limits access to accounting information to job owner, etc.
- *PurgeJobAfter=...*
 - Purge old job records, preserved indefinitely by default
- *PurgeStepAfter=...*
 - Purge old job step records, preserved indefinitely by default
- *StoragePass=...*
 - Password used to access the database

Upgrading



- slurmdbd can communicate with SLURM commands and daemons at the same or recent earlier versions (slurmdbd v2.5 can communicate with version 2.4 or 2.3, slurmdbd v2.4 will not recognize v2.5 RPCs)
- ALWAYS UPGRADE SLURMDBD FIRST

More Information



- SLURM documenation on line at:

<http://www.schedmd.com/slurmdocs/accounting.html>

http://www.schedmd.com/slurmdocs/resource_limits.html

http://www.schedmd.com/slurmdocs/priority_multifactor.html

<http://www.schedmd.com/slurmdocs/qos.com>