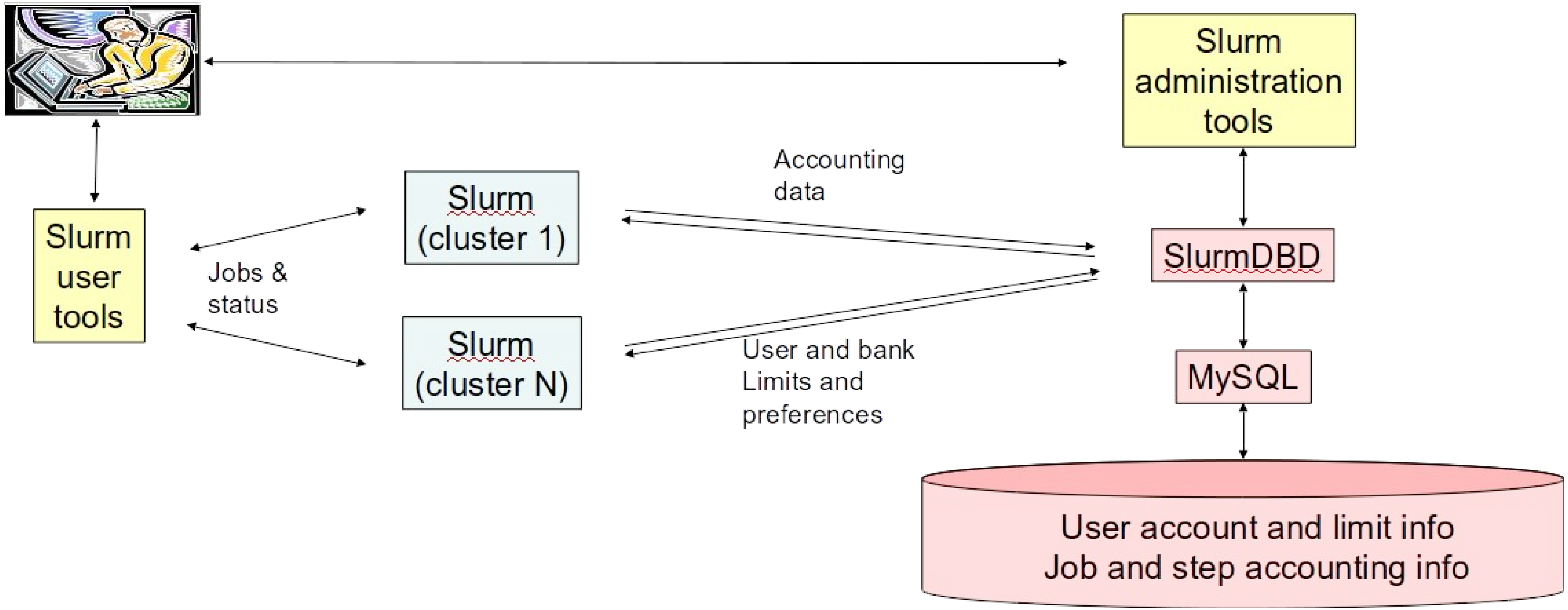
Slurm Workload Manager Overview
SC15

Alejandro Sanchez
alex@schedmd.com

Slurm Workload Manager Overview

- Originally intended as simple resource manager, but has evolved into sophisticated batch scheduler
- Able to satisfy scheduling requirements for major computer centers with use of optional plugins
- No single point of failure, backup daemons, fault-tolerant job options
- Highly scalable (3.1M core Tianhe-2 at NUDT)
- Highly portable (autoconf, extensive plugins for various environments)
- Open source (GPL v2)
- Operating on many of the world's largest computers
- About 500,000 lines of code today (plus test suite and documentation)

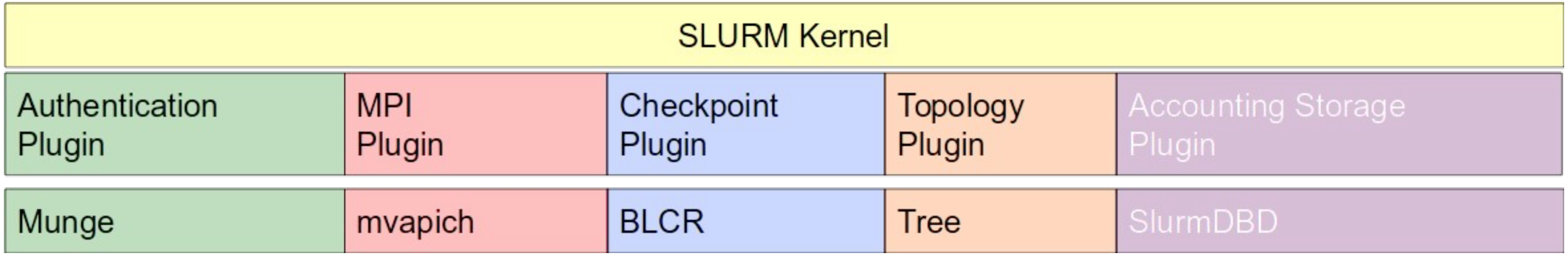
Enterprise Architecture


Slurm administration tools
Slurm user tools
Slurm (cluster 1)
Slurm (cluster N)
Jobs & status
Accounting data
User and bank Limits and preferences
SlurmDBD
MySQL
User account and limit info
Job and step accounting info

http://www.schedmd.com

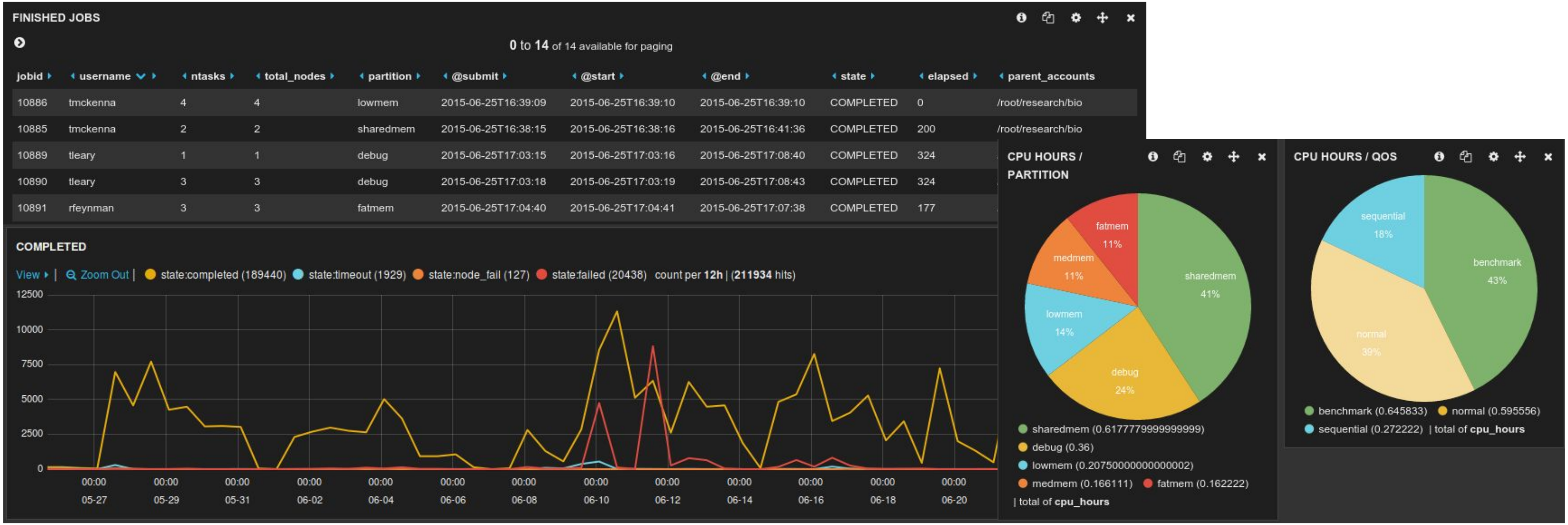
Architecture

- Kernel with core functions plus about 100 plugins to support various architectures and features
- Easily configured using building-block approach
- Easy to enhance for new architectures or features, typically just by adding new plugins



SLURM Kernel				
Authentication Plugin	MPI Plugin	Checkpoint Plugin	Topology Plugin	Accounting Storage Plugin
Munge	mvapich	BLCR	Tree	SlurmDBD

Elasticsearch Plugin



FINISHED JOBS

0 to 14 of 14 available for paging

jobid	username	ntasks	total_nodes	partition	@submit	@start	@end	state	elapsed	parent_accounts
10886	tmckenna	4	4	lowmem	2015-06-25T16:39:09	2015-06-25T16:39:10	2015-06-25T16:39:10	COMPLETED	0	/root/research/bio
10885	tmckenna	2	2	sharedmem	2015-06-25T16:38:15	2015-06-25T16:38:16	2015-06-25T16:41:36	COMPLETED	200	/root/research/bio
10889	tleary	1	1	debug	2015-06-25T17:03:15	2015-06-25T17:03:16	2015-06-25T17:08:40	COMPLETED	324	
10890	tleary	3	3	debug	2015-06-25T17:03:18	2015-06-25T17:03:19	2015-06-25T17:08:43	COMPLETED	324	
10891	rfeynman	3	3	fatmem	2015-06-25T17:04:40	2015-06-25T17:04:41	2015-06-25T17:07:38	COMPLETED	177	

COMPLETED

View ▸ | Zoom Out | ● state:completed (189440) ● state:timeout (1929) ● state:node_fail (127) ● state:failed (20438) count per **12h** | (**211934** hits)

CPU HOURS / PARTITION

● sharedmem (0.6177779999999999)
● debug (0.36)
● lowmem (0.20750000000000002)
● medmem (0.166111) ● fatmem (0.162222)
| total of **cpu_hours**

CPU HOURS / QOS

● benchmark (0.645833) ● normal (0.595556)
● sequential (0.272222) | total of **cpu_hours**

http://www.schedmd.com

Scheduling Capabilities

- Fair-share scheduling with hierarchical bank accounts
- Preemptive and gang scheduling (time-slicing parallel jobs)
- Integrated with database for accounting and configuration
- Resource allocations optimized for topology
- Advanced resource reservations
- Manages resources across an enterprise

Multifactor Prioritization Plugin

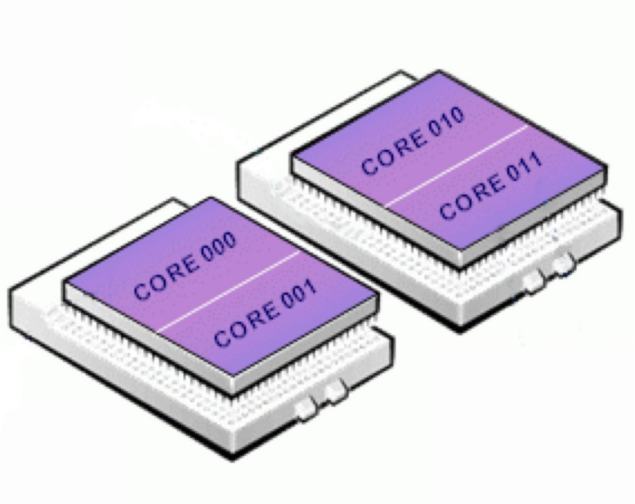
- Jobs can be prioritized using highly configurable parameters
  - Job Age
  - Job Partition
  - Job size
  - Job Quality Of Service (QOS)
  - User and account's fair-share allocation

Scalability

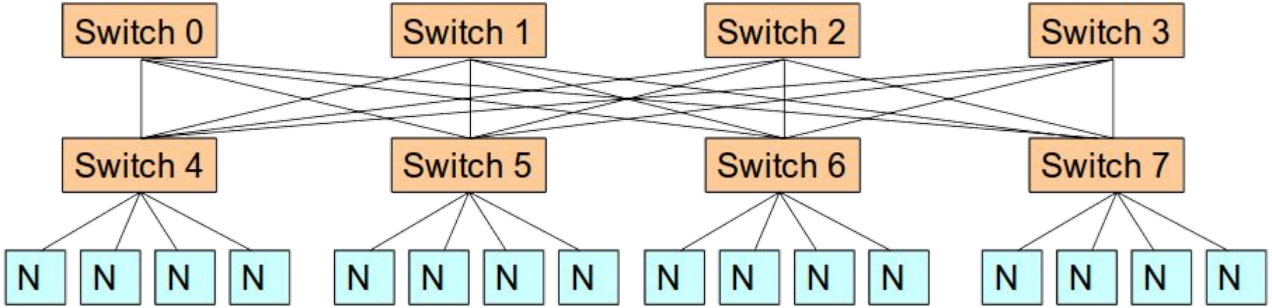
- Everything is multi-threaded
- Separate read and write locks on the various data structures in the daemons
- No single point of failure
- RPCs designed to minimize bottlenecks from control daemon as much as possible

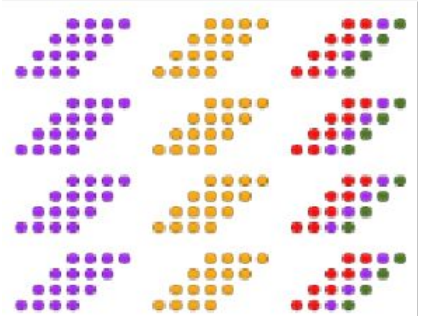
On-node Topology Optimization

- Users have complete control over task layout across the nodes, sockets, cores and threads to optimize application performance



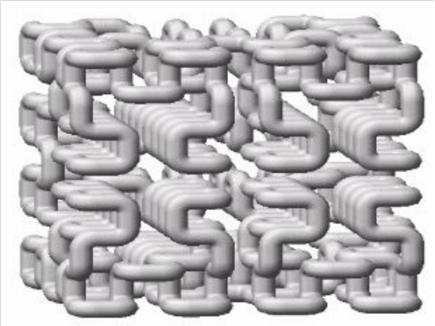
Topology Plugin Optimization





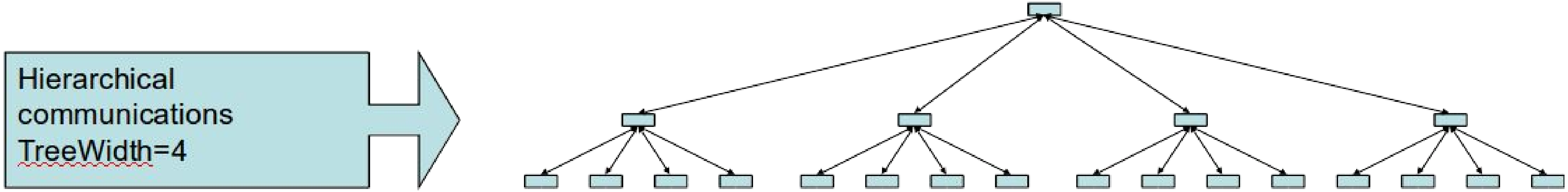
Job layout on BlueGene (sview output)

3-D Hilbert curve for Sun Constellation



Communications

- Hierarchical communications with configurable fanout and fault-tolerance



Hierarchical communications TreeWidth=4

Communications

- All commands and configuration files are designed to compress host names using a prefix and numeric suffix
- Easy to configure large systems

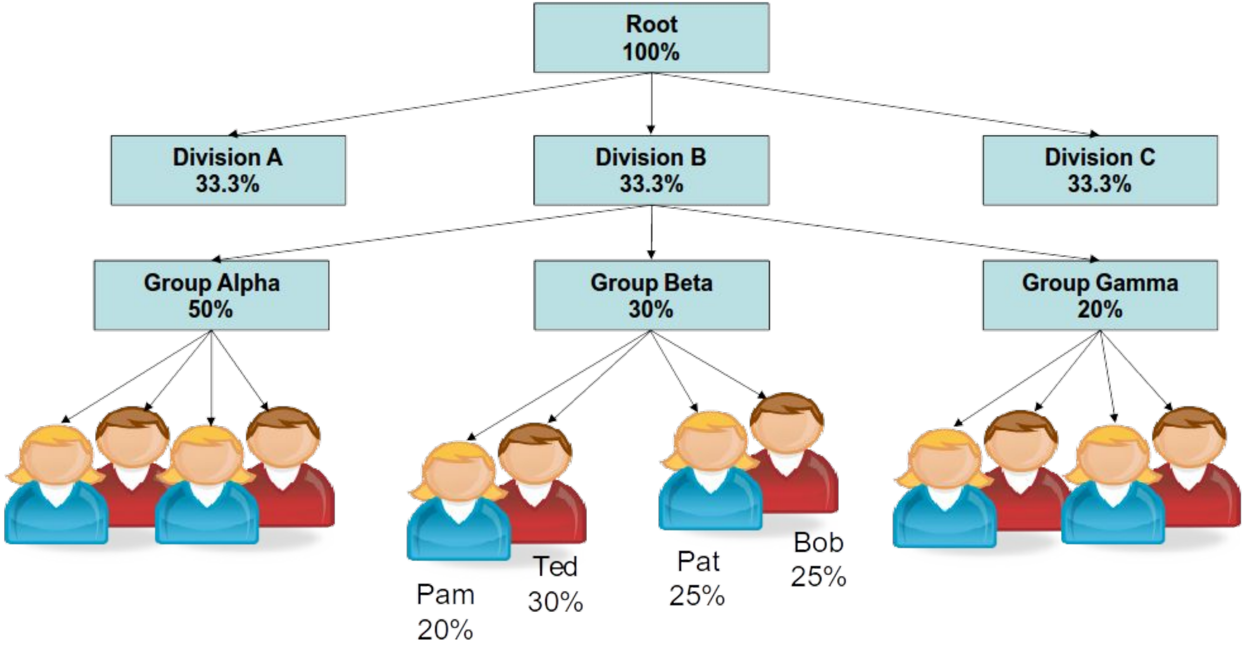
```
# Sample Slurm configuration file (excerpt)
#
NodeName=tux[0-1023] Sockets=4 CoresPerSocket=6
#
PartitionName=debug Nodes=tux[2-17] Default=yes
Maxtime=30
PartitionName=batch Nodes=tux[18-1023] MaxTime=24:00:00
```

Database Use

- Job accounting information written to a database plus
  - Information pushed out to scheduler daemons
  - Fair-share resource allocations
  - Many limits (max job count, max job size, etc)
  - Based upon hierarchical accounts
    - Limits by user AND by accounts

"All I can say is wow – this is the most flexible, useful scheduling tool I've ever run across."
Adam Todorski, Rensselaer Polytechnic Institute

Hierarchical Account Example



http://www.schedmd.com

Advanced Features

- Scheduling for generic resources (e.g. GPUs, MICs)
- User control over CPU frequency (performance and energy use)
- Real-time accounting down to the task level
  - Identify specific tasks with high CPU or memory usage
  - Record energy consumption by job
- Job profiling
  - Periodically capture each task's memory, CPU, power, network and I/O

15.08 Features

- Version 15.08.0 released on August 31
  - Massive changes from version 14.11
  - Diff file >250,000 lines
- Trackable Resources (TRES): Tracks utilization and/or limits enforce of memory, GRES, burst buffer, license, and any other configurable resources in the accounting database
- Per-Partition QOS
- Burst Buffers: a cluster-wide high-performance file system
- Network Topologies Optimizations, New parameters and environment variables...