Slurm Version 15.08

Danny Auble (SchedMD)
da@schedmd.com

SC15

Version 15.08

- Version 15.08.0 released on August 31
- Massive changes from version 14.11
  - Diff file >250,000 lines

Trackable Resources (TRES)

- Tracks utilization of memory, GRES, burst buffer, license, and any other configurable resources in the accounting database
- Any TRES can be used as a factor in computing a job's billing weight as used in calculating its resource utilization
- Any TRES can be used as a factor on calculating a job's priority
- Separate presentation with more details

Per-Partition QOS

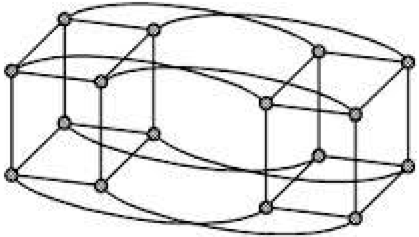
- Each partition can now have an associated QOS
- Each partition now has all of the limits available in a QOS
- Separate presentation with more details

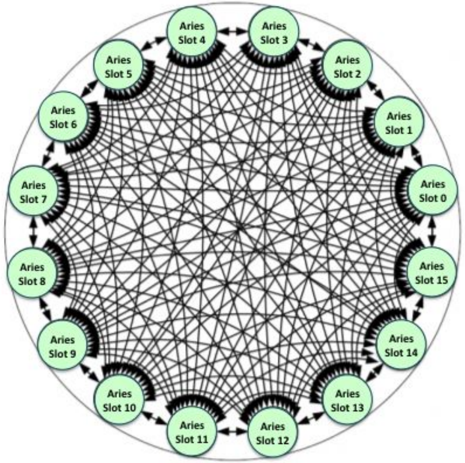
Burst Buffers

- Burst buffers are a cluster-wide high-performance file system
- Slurm can allocate resources, stage-in files before a job starts, stage-out files after a job completes and otherwise manage burst buffer resources
- Separate presentation with more details

Network Topologies

- Optimized resource allocations for
  - SGI Hypercube (work by SGI)
  - Dragonfly





Advanced Reservations

- New flag “replace”
  - As resources are allocated, replace them with idle resources to the extent possible
  - Maintains constant size of available resources
- Replace flag "License_only" with flag "Any_Nodes".
  - Used to indicate the advanced reservation resources (licenses and/or burst buffers) can be used with any compute nodes

Job Preemption

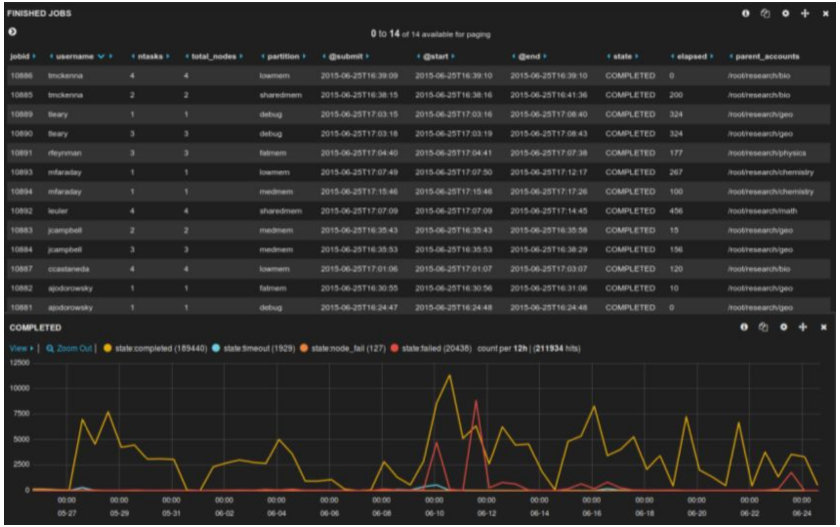
- Permit “PreemptMode=suspend,gang” and “PreemptType=qos” and to be used together
  - A high-priority QOS job will now oversubscribe resources and gang schedule, but only if there are insufficient resources for the job to be started without preemption

Single User Per Node

- Compute nodes can be allocated to multiple jobs, but restricted to a single user
  - New job option “--exclusive=user”
  - New partition configuration parameter “ExclusiveUser=yes”
  - Only accounts for resources allocated to jobs, idle resources currently not changed to the user who has been allocated the compute node

Elasticsearch Job Records

- New job completion plugin records a job's details into Elasticsearch database
- Many Elasticsearch tools available for analysis



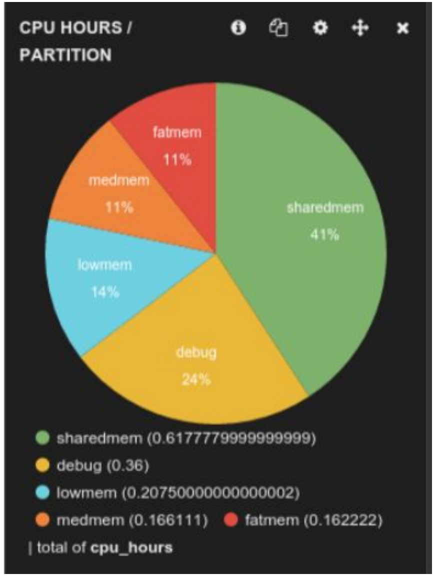
FINISHED JOBS

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	/root/research/geo
10890	tleary	3	3	debug	2015-06-25T17:03:18	2015-06-25T17:03:19	2015-06-25T17:08:43	COMPLETED	324	/root/research/geo
10891	rfeynman	3	3	fatmem	2015-06-25T17:04:40	2015-06-25T17:04:41	2015-06-25T17:07:38	COMPLETED	177	/root/research/physics
10893	mfaraday	1	1	lowmem	2015-06-25T17:07:49	2015-06-25T17:07:50	2015-06-25T17:12:17	COMPLETED	267	/root/research/chemistry
10894	mfaraday	1	1	medmem	2015-06-25T17:15:46	2015-06-25T17:15:46	2015-06-25T17:17:26	COMPLETED	100	/root/research/chemistry
10892	leuler	4	4	sharedmem	2015-06-25T17:07:09	2015-06-25T17:07:09	2015-06-25T17:14:45	COMPLETED	456	/root/research/math
10883	jcampbell	2	2	medmem	2015-06-25T16:35:43	2015-06-25T16:35:43	2015-06-25T16:35:58	COMPLETED	15	/root/research/geo
10884	jcampbell	3	3	medmem	2015-06-25T16:35:53	2015-06-25T16:35:53	2015-06-25T16:38:29	COMPLETED	156	/root/research/geo
10887	ccastaneda	4	4	lowmem	2015-06-25T17:01:06	2015-06-25T17:01:07	2015-06-25T17:03:07	COMPLETED	120	/root/research/bio
10882	ajodorowsky	1	1	fatmem	2015-06-25T16:30:55	2015-06-25T16:30:56	2015-06-25T16:31:06	COMPLETED	10	/root/research/geo
10881	ajodorowsky	1	1	debug	2015-06-25T16:24:47	2015-06-25T16:24:48	2015-06-25T16:24:48	COMPLETED	0	/root/research/geo

COMPLETED

View ▸ | Zoom Out | ● state:completed (189440) ● state:timeout (1929) ● state:node_fail (127) ● state:failed (20436) count per **12h** | (**211934** hits)



CPU HOURS / PARTITION

New Job Options

- sbast command can operate on nodes associated with either a job step (new) or an entire job
- Job “--mail” options now apply to a job array as a whole rather than each task of the job array
- OR'ed job dependencies
  - For example: “--depend=afterok:123?afternotok:124”
- Job “--kill-on-invalid-dep” to control behavior if the job's dependency can never be satisfied

New Environment Variables

- New job array environment variables
    - For example: --array=1-99:2
        - SLURM_ARRAY_TASK_MIN=1
        - SLURM_ARRAY_TASK_MAX=99
        - SLURM_ARRAY_TASK_STEP=2
- Other new environment variables
    - SLURM_JOB_ACCOUNT
    - SLURM_JOB_QOS
    - SLURM_JOB_RESERVATION

Task Layout Options (1 of 3)

- Job “--dist” option has “pack” and “nopack” flags
  - “pack” fills nodes, resulting in uneven distribution
  - “nopack” distributes tasks evenly across nodes (default)

	Node 0	Node 1
Pack	Task 0	Task 4
	Task 1	Task 5
	Task 2	Empty
	Task 3	Empty

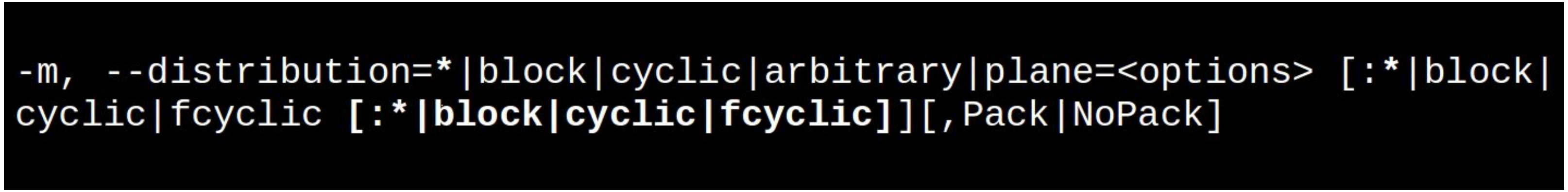
	Node 0	Node 1
NoPack	Task 0	Task 3
	Task 1	Task 4
	Task 2	Task 5
	Empty	Empty

Task Layout Options (2 of 3)

- Job “--thread-spec” option can reserve a hyper-thread for system use rather than an entire core (the “--core-spec” option)
- Job “--accel-bind” option binds tasks to nearest GPU and NIC

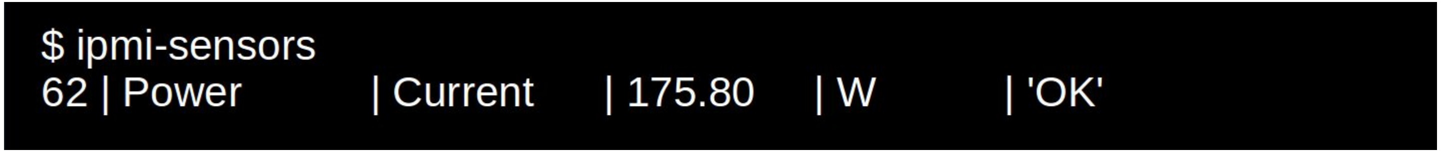
Task Layout Options (3 of 3)

- To better support the Xeon Phi processor architecture and OpenMP
  - New parameter to control the distribution of allocated threads across cores for binding to tasks
- Allows greater control over the placement of jobs
- The new syntax for the --distribution option is as follows. The new parameters are shown in bold

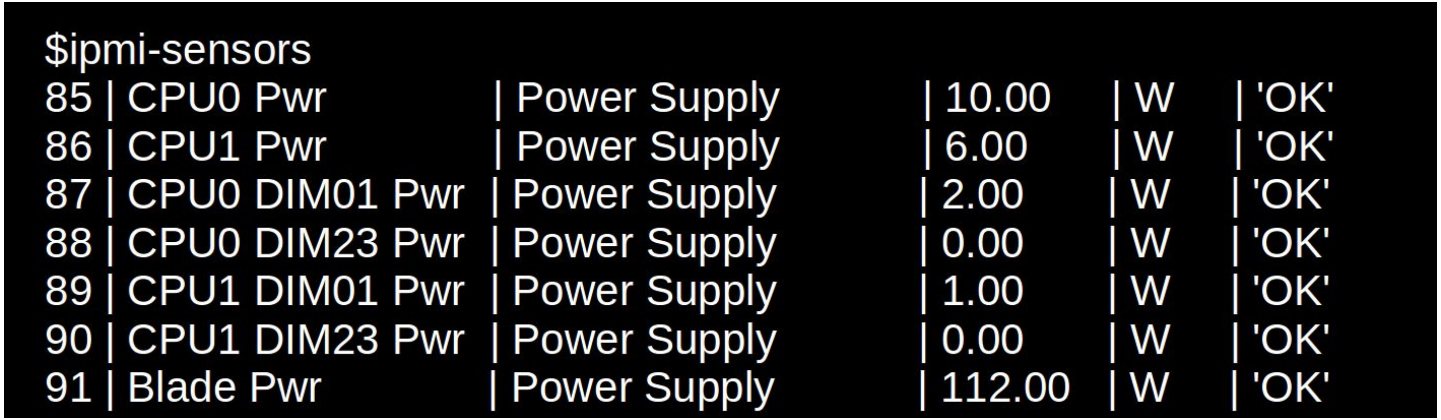


```
-m, --distribution=*|block|cyclic|arbitrary|plane=<options> [:*|block|
cyclic|fcyclic [:*|block|cyclic|fcyclic]][,Pack|NoPack]
```

Accounting – Profiling
Support of multiple energy sensors



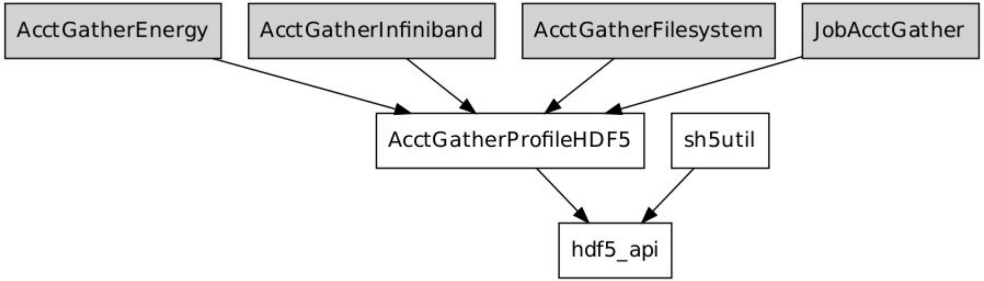
```
$ ipmi-sensors
62 | Power          | Current     | 175.80     | W          | 'OK'
```



```
$ipmi-sensors
85 | CPU0 Pwr          | Power Supply         | 10.00    | W     | 'OK'
86 | CPU1 Pwr          | Power Supply         | 6.00     | W     | 'OK'
87 | CPU0 DIM01 Pwr    | Power Supply         | 2.00     | W     | 'OK'
88 | CPU0 DIM23 Pwr    | Power Supply         | 0.00     | W     | 'OK'
89 | CPU1 DIM01 Pwr    | Power Supply         | 1.00     | W     | 'OK'
90 | CPU1 DIM23 Pwr    | Power Supply         | 0.00     | W     | 'OK'
91 | Blade Pwr         | Power Supply         | 112.00   | W     | 'OK'
```

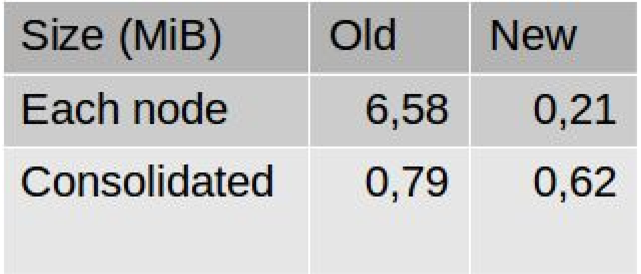
Profiling
HDF5 Architecture Optimizations

- New more scalable and flexible architecture
  - AcctGatherProfile operates as a service
- Based upon the high level HDF5 API
  - Features such as data compression (TODO add new parameter)
- Update sh5util (kept backward compatibility)
  - Calculate statistics during merge and not during processing



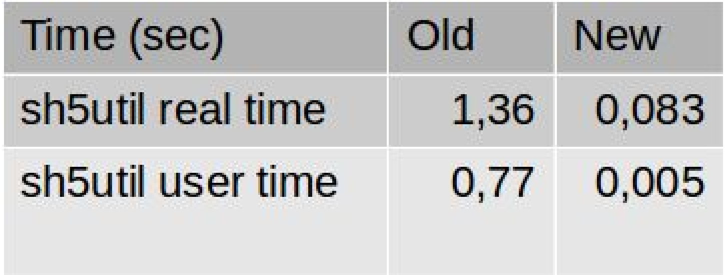
Profiling
HDF5 Architecture Optimizations

- Results: Profiling of a medium instance of HPLinpack upon 3 nodes (24min)
- Size of the profiling files:



Size (MiB)	Old	New
Each node	6,58	0,21
Consolidated	0,79	0,62

- Time to merge per node:



Time (sec)	Old	New
sh5util real time	1,36	0,083
sh5util user time	0,77	0,005

Power Adaptive Scheduling

- Provide **centralized mechanism** to dynamically **adapt the instantaneous power consumption** of the whole platform
  - Reducing the number of usable resources or running them with lower power
- Based upon the layouts framework
- Separate presentations with more details

Power Management

- Job “--cpu-freq” option now supports minimum frequency (in addition to maximum frequency and governor) and supported for salloc and sbatch (for power adaptive scheduling)
- --cpu-freq =<p1[-p2[:p3]]>
  - p1 is current options or minimum frequency
  - optional p2 is maximum
  - optional p3 is scaling governor
- New configuration parameter “CpuFreqGovernors” identifies allowed governors

Message Aggregation

- Improve performance by aggregating messages (such as epilog complete, node registration,etc) into a smaller number of composite messages,
- To reduce the number of incoming TCP connections to serve.
- To decrease the processing time of messages
- Based upon the reverse of the message forwarding/fanout mechanism
- Separate presentation with more details

WARNINGS (1 of 2)

- **The database schema has changed. Updating slurmdbd will take time. No records will be lost while upgrading, but the slurmdbd may not be responsive. It will not be possible to automatically revert the database for an earlier version of Slurm**
- In preparation for inter-cluster jobs, the MaxJobID has been reduced from 4,294,901,760 to 2,147,463,647. **Any job with an ID above 2,147,463,647 will be purged when upgrading**
- MVAPICH plugin now requires Munge for authentication
- Every plugin except SPANK must be built against the same version of Slurm (major and minor version number) to be loaded

WARNINGS (2 of 2)

- **HDF5 node-step file format changed** for improved performance
  - sh5util command has changed
  - Both file formats supported for next few releases

Questions?