Slurm Version 2.6

Morris Jette, Danny Auble (SchedMD)
Yiannis Georgiou (Bull)

Version 2.6

- Released 2.6.0 on 6 July
- Most stable major release to date
  - Version 2.6.1 not released until 16 August
  - Technical University of Dresden used version 2.6.0 for acceptance testing and helped to work out many of the bugs

Job Arrays

- Submit and manage collection of similar jobs easily
- To submit 50,000 element job array:

```
$ sbatch –array=1-50000 -N1 -i my_in_%a -o my_out_%a my.bash
```

- Submit time < 1 second
- “%a” in file name mapped to array index
- Additional environment variable with array index: SLURM_ARRAY_TASK_ID

Job Arrays (continued)

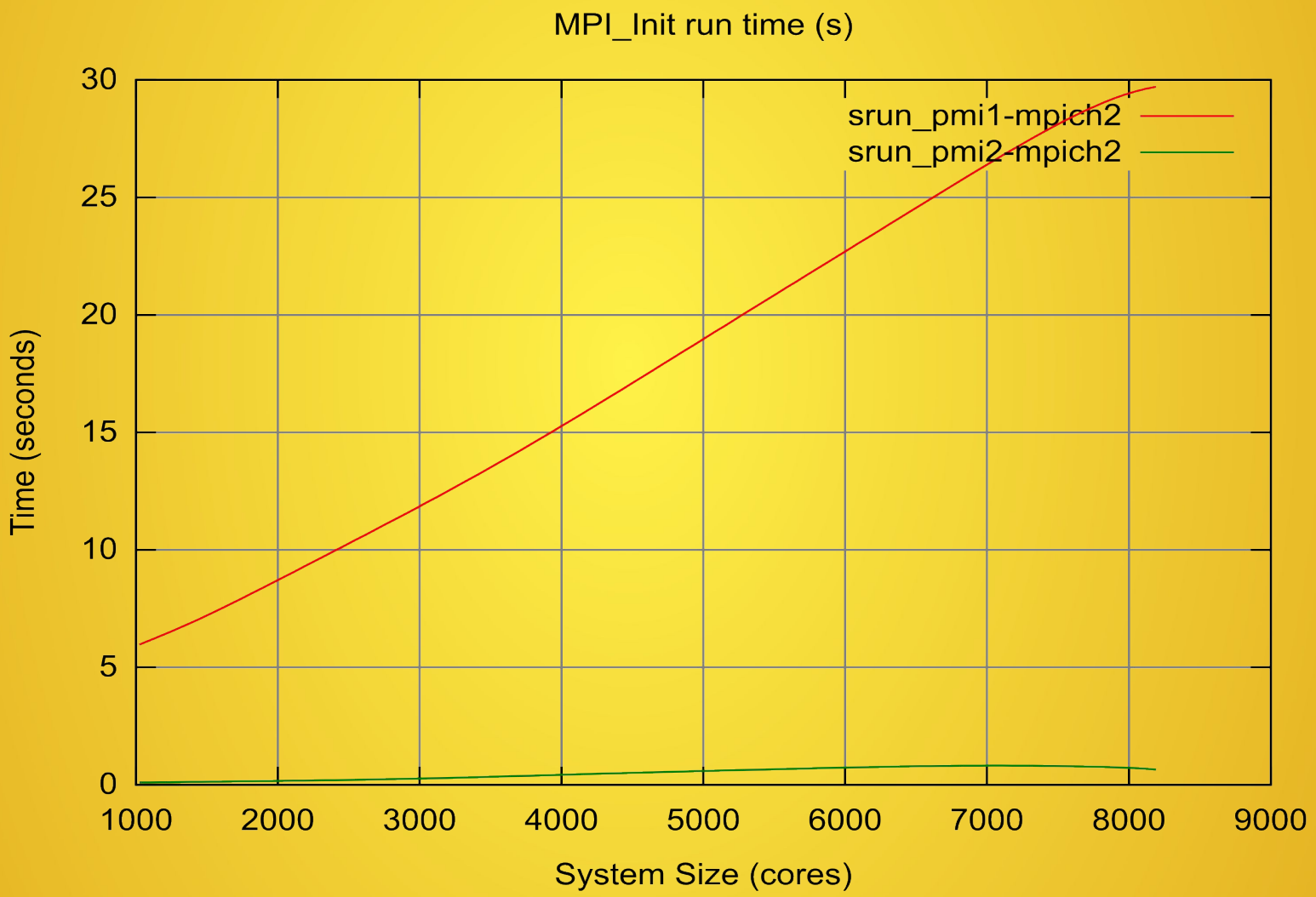
- squeue and scancel commands plus some scontrol options can operate on entire job array or select task IDs

```
$ squeue
JOBID      PARTITION  NODESLIST(REASON)
123_[2-50000]  debug   (Resources)
123_1          debug   tux0

$ scancel 123_[40000-50000]

$ scontrol hold 123
```

PMI2 Infrastructure

- Infrastructure adapted to support different MPI libraries (OpenMPI, BullxMPI)
- Provides key-value repository support needed by MPICH2
- Vastly more scalable than previous PMI infrastructure
- Work by NUDT: 30 seconds execution time for simple MPI_Init program on 30,000 tasks and 15,000 nodes

PMI vs PMI2 Performance



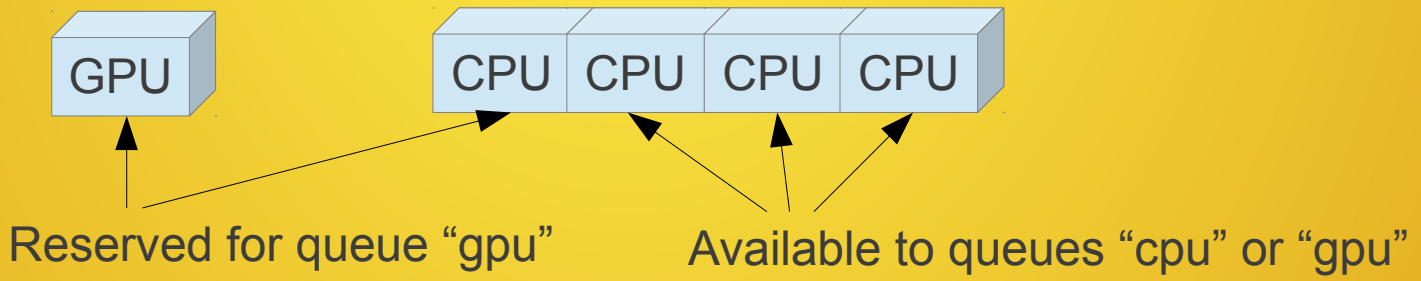
GPU Scheduling Support

- Added partition configuration parameter of MaxCPUsPerNode

- Configure partition/queue for GPU use and prevent jobs not using GPUs from consuming all of the CPUs

GPU Scheduling
Example

- For cluster with 1 GPU and 4 CPUs per node
- Configure 2 queues: “cpu” and “gpu”
- Configure queue “cpu” with MaxCPUsPerNode=3
- Queue “gpu” use restricted to jobs using GPUs (enforce using job_submit plugin)



Faster Throughput

- New logic supports pending job steps and slurmctld daemon sends message to srun when the job step allocation is available, replaces polling logic

- Modified locking to prevent high rate of batch job submissions from blocking job scheduling or state read commands (e.g. squeue)

Backfill Scheduling

- Added new “SchedulerParameters” option of “bf_continue” which permits the backfill scheduler to continue operating on the same job queue even if new jobs are added during its periodic release of locks (sleeps after every 2 seconds of execution to permit other work to happen)
- Considers more jobs for scheduling
- Ignores jobs submitted during it’s release of locks, so newly submitted jobs might be started later than ideal

Scheduling

- Added SelectTypeParamter value of CR_ALLOCATE_FULL_SOCKET
- Added PriorityFlags of TICKET_BASED and merged priority/multifactor2 plugin into priority/multifactor plugin
- Added PriorityFlags configuration parameter of SMALL_RELATIVE_TO_TIME. If set, job's size factor is computed by dividing it's size by it's time limit rather than by using the job's size alone

Advance Reservations

- Added Prolog and Epilog support
  - Permits customization of behavior when an advanced reservation begins or ends (e.g. killing or requeuing user jobs)
  - Configuration parameters ResvProlog and ResvEpilog
- Added support for reservations at the core level rather than whole nodes

Health Check Program

- Added HealthCheckNodeState to identify nodes states on which the HealthCheckProgram should run (e.g. partially or entirely idle nodes)

PBS/Torque Support

- New in version 2.6.3 (Working with NASA)
- Added job_submit and SPANK PBS plugins
- Support added for more options in qsub and sbatch #PBS structures
  - “before” job dependencies
- Sets many PBS environment variables

Cray and BlueGene Systems

- Batch jobs executed on front-end. No slurmd daemon on the compute nodes

- Added front end node configuration options: AllowGroups, AllowUsers, DenyGroups and DenyUsers

MapReduce+

- Added SlurmctldPlugstack configuration and support for generic stack of slurmctld daemon plugins (on the plugin's init and fini functions are called)
- Added slurmctld/dynalloc plugin with MapReduce+ support
- Additional infrastructure required not yet available from Greenplum/EMC

Energy Monitoring Infrastructure

- Added Support for External Sensors Plugins to allow out-of-band monitoring of cluster sensors
  - Possibility to Capture energy usage and temperature of various components (switches, rack-doors, etc)
  - Support for RRD databases for collection of energy/temperature data
  - Plugin to be used with real wattmeters or out-of-band IPMI capturing
  - Power data captured used for per node power monitoring (scontrol show node) and per job energy accounting (Slurm DB)
- Improved support for in-band monitoring of energy data
  - Stable Support of IPMI and RAPL mechanisms

Enhanced Accounting

- Added Support for local disk IO usage accounting per task in Slurm DB
  - Capturing through /proc
  - Average and Maximum values written in Slurm DB
- Added new field for Requested Cpu Frequency (besides Average Cpu Frequency) in Slurm DB

Job profiling Infrastructure

- Added support for job profiling to periodically capture the task's usage of various resources like CPU, Memory, Lustre, Infiniband and Power per node
- Resource Independent polling frequency configuration
- Based on hdf5 file format
  - Profiling per node (one hdf5 file per job on each node)
  - Aggregation on one hdf5 file per job (after job termination)
  - Slurm built-in tools for extraction of hdf5 profiling data