# Slurm
# Recent Developments and Roadmap

Morris Jette - SchedMD LLC
CUG 2017

# Slurm Overview

- Workload management system
  - Open source (GPL)
  - Fault tolerant
  - Highly scalable
  - Sophisticated scheduling capabilities (backfill, gang, limits, accounting)
- Used on 5 of top 10 systems from TOP 500 list
  - #1 - Sunway TaihuLight - National Supercomputer Center, China
  - #2 - Tainhe-2 - National Supercomputer Center, China
  - #4 - Sequoia - Lawrence Livermore National Laboratory (LLNL)
  - #5 - Cori (Cray) - NERSC
  - #8 - Piz Daint (Cray) - Swiss National Supercomputing Centre (CSCS)

# Cray Capabilities

- Eliminates need for ALPS
  - Uses Cray APIs to manage network and launch MPI jobs
  - Can run multiple jobs from multiple users on each node
    - Run one job per core or thread in its own container
- Support for KNL
  - User selection of NUMA and MCDRAM modes
  - User ability to change mode and reboot (configurable permissions)

# Cray Capabilities (continued)

- Support for DataWarp
  - DataWarp and compute resource scheduling coordinated
    - File stage-in scheduled before compute resources
    - File stage-out scheduled after compute resources released
- Power Capping
  - Per-node power caps adjusted in near real time
    - Any node not using its full power allotment has that power transferred to nodes that can use the power
    - Quickly adjusts to changing workload while allocating as much of the power cap as possible

# Other Capabilities

- Container for processes spawned outside of Slurm
  - Cray/cgroup container created on compute nodes at job allocation time
  - PAM module puts login shells into that container
  - Resource usage by external processes managed and accounted for
- High throughput computing
  - 500 jobs/second sustained

# Version 17.11

- Release November 2017
- Federation of clusters
  - Multiple clusters can be viewed as single compute resource
  - Workload scheduled across multiple clusters
  - Implemented using peer-to-peer model
- Heterogeneous job allocations
  - Heterogeneous memory specifications, CPUs per task, node features, etc.

```
srun --features=haswell --ntasks=1 master : --features=knl,a2a,flat --ntasks=72 slave
```

# Version 17.11 (continued)

- More flexible advanced reservations
  - Jobs able to use resources inside and outside of the reservation
  - Jobs able to start before and end after the reservation
- Add *scancel --hurry* option
  - Cancel job without staging-out DataWarp files

# Version 18.08

- Release August 2018
- Eliminate support for ALPS (tentative for version 18.08)
  - Transition to native Slurm mode

# Presentation Later Today

Scheduler Optimization for Current Generation Cray Systems

Wednesday 10 May, 3:00

Salon 3