SchedMD

Slurm Development and Support

http://www.schedmd.com
sales@schedmd.com
925-695-7782

SchedMD is the core company behind the Slurm workload manager software, a free open-source workload manager designed specifically to satisfy the demanding needs of high performance computing. Slurm is in widespread use at government laboratories, universities and companies world-wide. As of the November 2013 Top 500 computer list, Slurm was performing workload management on five of the ten most powerful computers in the world including the number 1 system, Tianhe-2 with 3,120,000 computing cores.

Our vision is to extend workload management functionality to address Exascale requirements and beyond in an open and collaborative fashion. We are also working to bring HPC technology to the world of Big Data, with early results yielding orders of magnitude improvement in performance.

Slurm

Slurm is a highly configurable open-source workload manager. In its simplest configuration, it can be installed and configured in a few minutes. Use of optional plugins provide the functionality needed to satisfy the needs of demanding HPC centers. More complex configurations rely upon a database for archiving accounting records, managing resource limits by user or bank account, and supporting sophisticated scheduling algorithms. Notable features include:

- **Scalability:** High scalability was of paramount importance from Slurm's initial design work in 2002. All daemons and most use commands are extensively multi-threaded with separate read and write locks on the various data structures. Many of the largest computers in the world today use Slurm, and systems orders of magnitude larger have been validated using virtualization.
- **Performance:** Up to 1,000 job submissions and 500 job executions per second.
- **Flexibility:** Slurm is highly configurable using a building block approach with about 100 plugins available to support various interconnects, scheduling algorithms, MPI versions, etc. These plugins are documented and permit extensive customization using C or the Lua scripting language.
- **Power Management:** Jobs can specify desired CPU frequency while power consumption is included in job accounting records. Idle resources can be powered down until needed to conserve energy. Each job's power consumption can be recorded in the accounting database.
- **Free and Open Source:** Source code freely available under the GNU General Public License.
- **Fault Tolerant:** Backup daemons eliminate any single point of failure. Applications can optionally continue to run after failure of compute nodes and

request additional resources to replace those which have failed.

- **Topology Optimized Allocations:** Resource allocations can be optimized with respect to network topology to minimize communication latency. Jobs can specify the maximum number of leaf switches desired and how long it is willing to wait for such an allocation.
- **Processor Binding:** Slurm maintains detailed information about a node's architecture including NUMA, sockets, cores and hyperthreads. Applications are automatically bound to resources designed to optimize performance and user options provide complete control over task binding to resources.
- **Resizable Jobs:** Job allocations can grow and shrink on demand. Job submissions can specify size and time limit ranges. This can also used by an application to replace failed resources.
- **Advanced Reservations:** Resources can be reserved for future use by specific users and/or bank accounts. Resources can also be reserved for scheduled maintenance.
- **Flexible Scheduling Policies:** Sophisticated scheduling policies include hierarchical fair-share scheduling and an assortment of constraints by user, bank account and Quality Of Service (QOS). Management of these limits can also be delegated to bank "coordinates" to minimize system administration work. Gang Scheduling (time-slicing of parallel jobs) and preemption can be configured to optimize responsiveness while maintaining high utilization.
- **Generic Resource Support:** Slurm supports the allocation of generic resources, which can be used to manage GPUs and MIC processors.
- **Job Profiling:** Periodically record each task's CPU, memory, power, network, and I/O usage to an HDF5 file for analysis.
- **Hadoop Integration:** Slurm eliminates the need for a dedicated Hadoop cluster and provides dramatically better scalability without application changes.

For additional information about Slurm, see http://slurm.schedmd.com. For additional information about SchedMD, including employment opportunities, see http://www.schedmd.com

"In 2010, when we embarked upon our mission to port Slurm to our Cray XT and XE systems, we discovered first-hand the high quality software engineering that has gone into the creation of this product. From its very core Slurm has been designed to be extensible and flexible. Moreover, as our work progressed, we discovered the high level of technical expertise possessed by SchedMD who was very quick to respond to our questions with insightful advice, suggestions and clarifications. In the end we arrived at a solution that more than satisfied our needs. The project was so successful we have now migrated all our production science systems to Slurm, including our 20 cabinet Cray XT5 system. The ease with which we have made this transition is testament to the robustness and high quality of the product but also to the no-fuss installation and configuration procedure and the high quality documentation. We have no qualms about recommending Slurm to any facility, large or small, who wish to make the break from the various commercial options available today."

Colin McMurtrie, Head of Systems, Swiss National Supercomputing Centre