Slurm Roadmap

Versions 17.02 and beyond

Yiannis Georgiou, Andry Razafinjatovo (Bull)

Slurm User Group 2016

Version 17.02

- Version 17.02 to be released in February 2017
    - Maintaining 9 month release cycle
- Some key features already determined
- Other possible features still under discussion

Support for heterogeneous resources

- These developments have as goal to extend the job specification language of SLURM to support MPMD (Multiple Program Multiple Data)
- With the support of heterogeneous resources, the idea is to introduce a new type of jobs named job packs which will be described by a set of pack groups, each pack group having the same resources requirements.
- Examples of executions illustrating the targeted capability :
    - srun -n 2 -c2 ./app1 : -n 4 --mem-per-core 256 --gres=gpu:2 ./app2
    - Or
    - sbatch -n 2 -c 2 : -n 4 --mem-per-core 256 –gres=gpu:2 ./script.sh
    - cat script.sh
        - srun --pack-group 0 ./app1 : --pack-group 1 ./app2
        - srun –pack-group=[0,1] ./app

Powercap with RAPL

- Powercapping based on layouts and RAPL
    - RAPL ensures hardware powercap guarantee Run cluster within the power budget
    - RAPL provides a good estimate of socket power consumption -> Adapt layouts regularly to reflect real values
    - Adapt power based on real power consumption (capture the application behavior)
        Allow more jobs to take advantage of the unused power

Lustre and Infiniband accounting

- Capture accounting information for both infiniband network and lustre filesystem per step/job.
- Make the data available in the Slurm DB and through sstat and sacct commands.
- Based upon the new TRES (Trackable Resources) functionality

High Definition Power and Energy Monitoring

- Proposing a new plugin acct_gather_energy/hdeem
- Based on the new FPGA architecture supported through ipmi-raw
- Improved accuracy for both power profiling per components (100Hz) and nodes (1000Hz)
- Improved precision for energy consumption per job based on nodes (1000Hz) measurements
- Decrease overhead on the application (CPU and Memory) since the collection is done internally within the FPGA

Beyond 17.02

- Scheduling
  - Multi-objective resource selection
  - Machine Learning Optimizations
  - Towards energy budget control
- Hybrid environments
  - Tighter integration with Singularity for deploying of customized user environments
  - Deploy Big Data workflows and Cloud environments upon HPC clusters
- Enable SLURM simulation in very large scales
  - Proposal for flexible solution enabling comparisons with new generation RJMS

Thanks

Questions?

Towards Energy Budget Control

EnergyCap Scheduling

- Schedule jobs under particular energetic budgets for variable time durations.
- Extension of powercapping with the difference that we are interested to adapt the power consumption in a way that the final energy consumption of the particular time duration remains below the allowed energetic budget.
- The actual energy consumption reductions take place through coordinated techniques such as:
  - Dynamic CPU - Frequency scaling
  - Hardware power-capping (RAPL)
  - Keeping nodes idle
  - Shut-down nodes