Resource management with Slurm

Tobias Weßeler

Arbeitsbereich Wissenschaftliches Rechnen
Fachbereich Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften
Universität Hamburg

2016-01-18

UHH Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

informatik die zukunft

Agenda

1 Introduction

2 Slurm

3 Energy and Power Measurement

4 Upcoming

5 Summary

6 Literature

Introduction

- Resource manager
  - Monitors resource utilization (CPU, RAM, etc.) and allocates them to the users' jobs
  - Monitors power consumption
  - Switches off unused resources
  - Communicates with job scheduler
- Job scheduler
  - Uses information from resource manager to prioritize jobs
  - Schedules jobs to efficiently use resources
  - Informs resource manager about hardware needs

Introduction

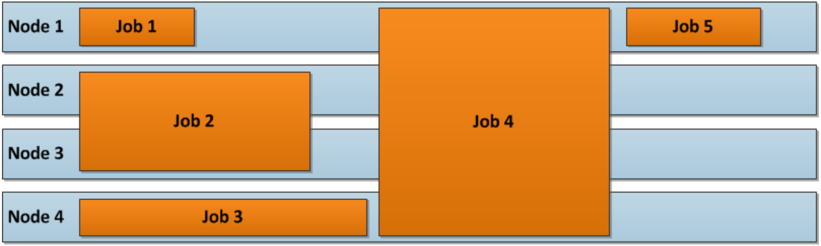


Figure: Naive queue, figure based on: [Ada14]

Introduction

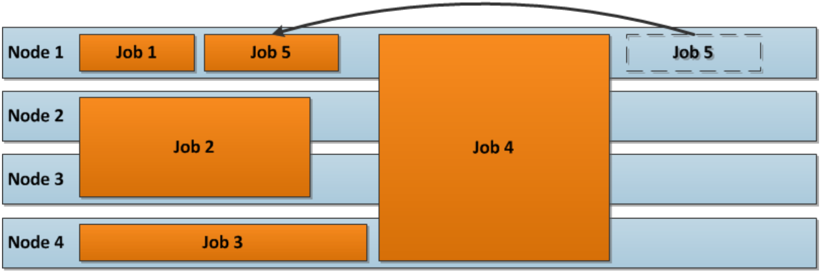


Figure: Backfill example, figure based on: [Ada14]

Introduction

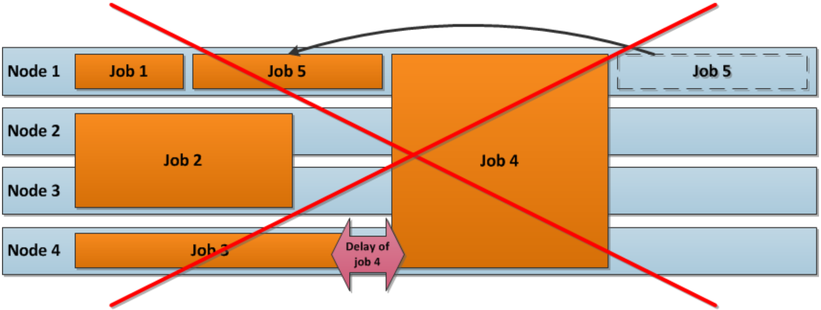


Figure: No backfill possible, figure based on: [Ada14]

SLURM

Simple **L**inux **U**tility for **R**esource **M**anagement

- Combined resource manager and job scheduler
- Open source, fault-tolerant, highly scalable
- Supports plugins (dynamically linked objects at runtime)
- Active development by community as well as SchedMD
- wide-spread use in HPC

"As of the June 2015 Top 500 computer list, Slurm was performing workload management on six of the ten most powerful computers in the world including the number 1 system, Tianhe-2 with 3,120,000 computing cores."
-SchedMD website

Daemons of slurm

Slurmctld – controller daemon

- Monitors and allocates resources
- Manages job queues
- Has optional backup with automatic fail-over

Slurmdbd – database daemon

- Stores accounting and configuration information
- Also has an optional automatic fail-over
- Attached database can be mysql, postgresql or text format

Slurmd – compute node daemon

- Launches and manages tasks
- Very light-weight
- Quiet (except for optional accounting)

Slurmstepd

- Manages job steps and I/O
- Spawned for each jobstep and terminated after

Daemons of slurm

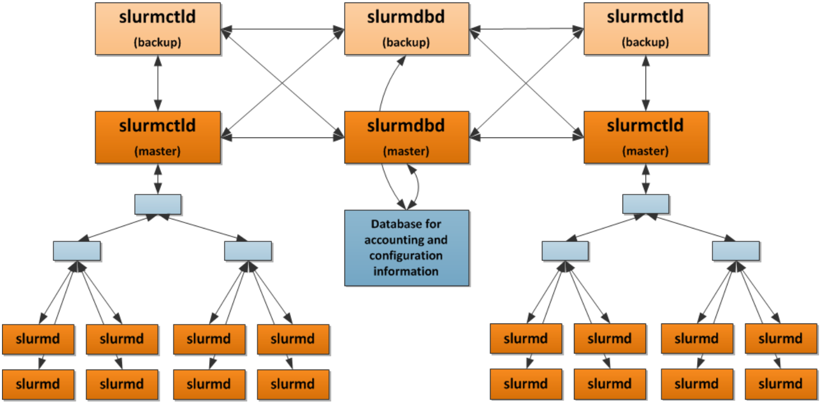


Figure: Multi-cluster environment

Figure based on Introduction to Slurm: [Sch16]

What is a plugin?

- Also called addins, addons or extensions
- A plugin is an optional software module that can extend or change the functionality of an existing program
- Usually plugins are very specific and only work with a certain program - just like a puzzle piece only fits into its own puzzle
- Software gets more customizable and becomes extensible
- Often represented as a puzzle piece
- Described via interface or API

Plugins

- Over 80 plugins (as of dec 2012)
- Objects that are dynamically linked during runtime
- Currently 26 well-defined APIs / programmer's guides

Energy accounting

- Measure the power and energy consumed by nodes or jobs
- Power profiling:
  analyse power demands of cluster and utilization of resources
- Improve energy efficiency

- Power:
  $P = I \cdot V$ (Product of Current and Voltage)
  SI: watt (1 joule over 1 second)
- Energy consumption:
  $P \cdot t$ (Product of Power and Time)
  SI: watt-hours (3600 joule)

Motivation

- Money
- DKRZ consumes over 17 GWh per year
- It costs over 1,850,000 €
- That is roughly 11 cents per KWh
- For comparison:
  Average energy consumption per Person - 2,000 KWh
  Factor: 8,500,000
- Environmental awareness

How to measure energy

- PM = power meter
  different possible locations
- Library / API
- Program (or plugin) to use library or API
- Samples are collected by client and then processed

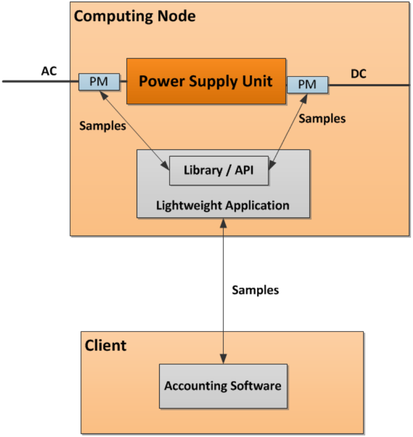


Figure: Concept drawing [Wes]

Slurm Energy Plugin - RAPL

- Running Average Power Limit
- Samples are estimated from a power consumption model based on hardware counters
- Estimates seem to be very accurate
- Only 2 readings necessary -> low overhead

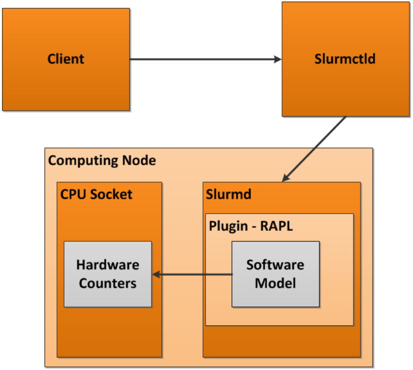


Figure: Concept drawing [Wes]

Slurm Energy Plugin - IPMI

- Intelligent Platform Management Interface
- Protocol to read from sensors
- Lights Out Management:
  enables remote control and management of machine
- Baseboard Management Controller
  Special microcontroller connected to sensors on hardware
- Phyiscal Interfaces:
  SM Buses, Serial Port, IMPB
- BMC communicates with BMU
  (**B**aseboard Management Controller **M**anagement **U**tility)

Slurm Energy Plugin - IPMI

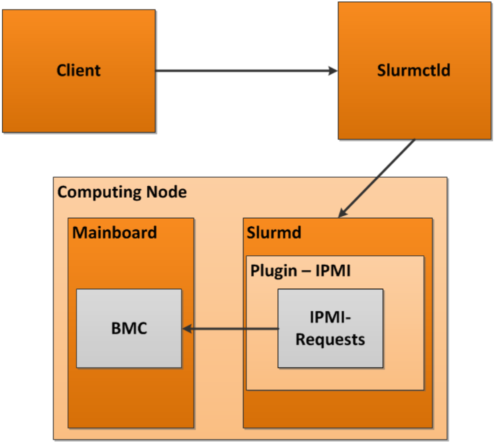


Figure: Concept drawing [Wes]

Slurm Energy Plugin - Config

- Slurm.conf (main config file)
  - AcctGatherEnergyType
    Specifies which plugin should be used.
  - AcctGatherNodeFreq
    Time interval between pollings in seconds.
- Acct_gather.conf (same dir as slurm.conf)
  - Contains configuration for acct_gather related plugins
  - E.g. EnergyIPMIFrequency:
    number of seconds between BMC access samples

Slurm Energy Plugin - ext_sensors

- New infrastructure -> HDEEM
- High Definition Energy Efficiency Monitoring
- High resolution: ~1000 samples per second
- Plugin needs to be written to utilize functionality
- Ext_sensors plugin works independently from acct_gather plugins
- Standard config files only allows up to 1 call per second

HDEEM

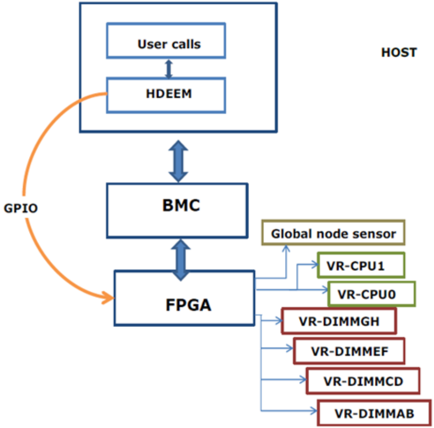


Figure: Concept drawing [unk]

Challenges

- CPU current and voltage are highly dynamic
- Frequency is much higher than sampling rate
- Power meter returns instant values - need to be converted
- Many conversion steps required - sources of inaccuracy:
  - Voltage and current sensors
  - Analog-digital-converter
  - Lowpass filters
  - data formats
  - average calculations
- Calculation of energy correct avg power values over time period

Configuration cases

- node energy monitoring

  AcctGatherEnergyType=acct_gather_energy/ipmi or rapl
  AcctGatherNodeFreq=<seconds>
  or
  ExtSensorsType=ext_sensors/rrd

  ExtSensorsFreq=<seconds>

- job/step energy accounting

  JobAcctGatherType=jobacct_gather/linux or cgroup
  AcctGatherEnergyType=acct_gather_energy/ipmi or rapl
  JobAcctGatherFrequency=task=<seconds>
  or
  JobAcctGatherType=jobacct_gather/linux or cgroup

  ExtSensorsType=ext_sensors/rrd

- job/step power profiling

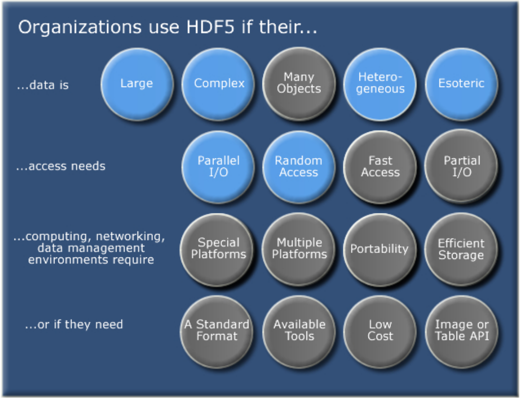
  AcctGatherEnergyType=acct_gather_energy/ipmi or rapl
  AcctGatherProfileType=acct_gather_profile/hdf5

  JobAcctGatherFrequency=energy=<seconds>

Slurm Energy Plugin - File Format

- Efficient file format needed to store collected data
- HDF5 (Hierarchical Data Format version 5)
- Represents a wide variety of data structures within a single file
- Supports very complex data
- High-level interfaces for C, C++, Fortran 90 and Java

HDF5



Organizations use HDF5 if their...

...data is: Large, Complex, Many Objects, Heterogeneous, Esoteric

...access needs: Parallel I/O, Random Access, Fast Access, Partial I/O

...computing, networking, data management environments require: Special Platforms, Multiple Platforms, Portability, Efficient Storage

...or if they need: A Standard Format, Available Tools, Low Cost, Image or Table API

Figure: Features [The16]

HDF5 - Requirements

- Shared filesystem for compute nodes
- Slurm.conf:
  Uses HDF5 Profile Plugin
- Acct_gather.conf:
  Root directory of profiling data (must be in shared filesystem)
- Each slurmstepd keeps his own file
- Files are merged after job completion

HDF5 Slurm integration

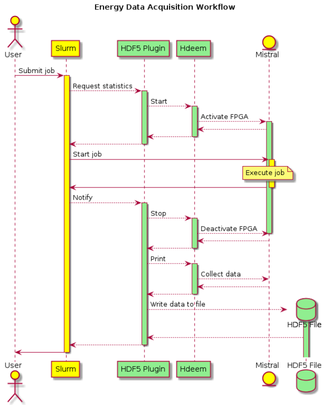


Figure: Workflow [unk]

What's new?

- Supports asymmetric resource allocation
  - Different amount of resources for each process / rank
- Enables MPMD approach
  - Classical approach: SPMD
  - Example call: mpirun -np 2 a.out : -np 2 b.out

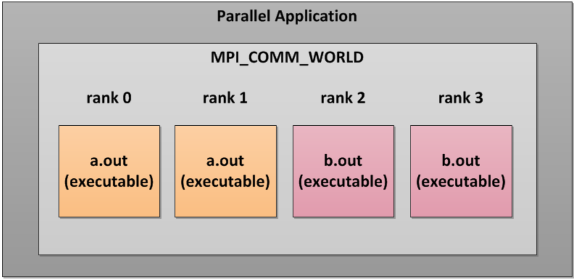


Figure: Concept drawing [Wes]

Summary

- HPC software stack
  - Slurm is a good choice
  - More possibilities for resource management in near future
- Energy Accounting
  - Range of available plugins is growing
  - Energy consumption and power profiles become increasingly important due to high costs in HPC
  - Accurate power profiling is difficult

Literature

[Ada14] Adaptive Computing Enterprises. Maui Scheduler Administrator's Guide, 1999-2014. http://docs.adaptivecomputing.com/maui/8.2backfill.php.

[ea14] Daniel Hackenberg et al. HDEEM: High Definition Energy Efficiency Monitoring. 2014.

[Mar13] Martin Perry (Bull). Energy Accounting and External Sensors Plugins, 2013. http://www.schedmd.com/.

[Sch16] SchedMD LLC. Slurm Commercial Support and Development, 2011-2016. http://www.schedmd.com/.

[The16] The HDF Group. Hierarchical Data Format, version 5, 1997-2016. http://www.hdfgroup.org/HDF5/.

[unk] unknown. materials provided by R. Heidari.

Thank you!

Questions?

Also read for information:

[ea14]
[Mar13]
[Yia12]