Job Scheduling, Resource Management,and Accounting

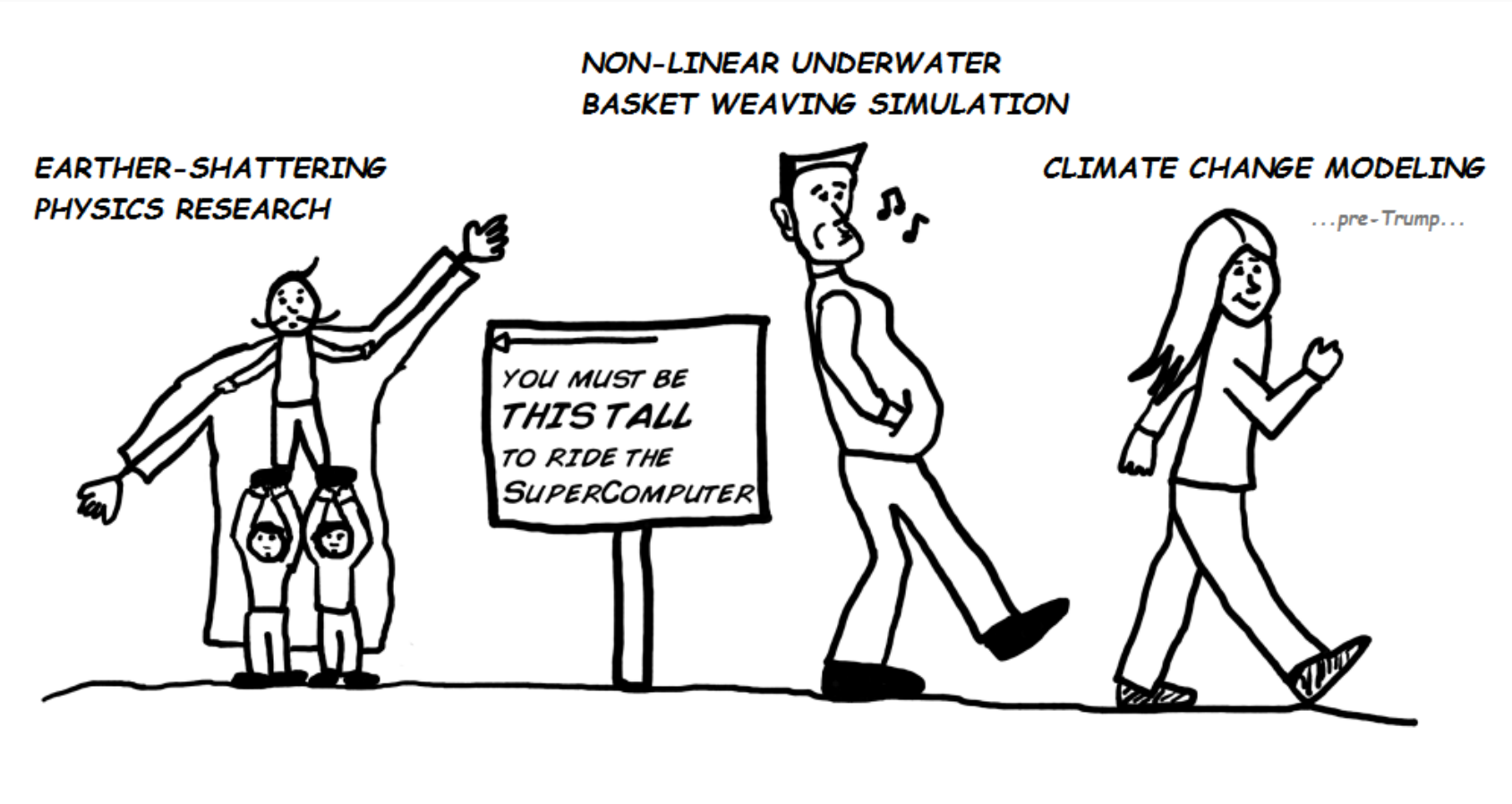
Cluster Computing Basics; How to use What You've Learned.

April 20, 2017

If we've got 30 compute-nodes, and 50 Projects...

It's a Hard Knock Life

- HPC Cluster ⇒ Many nodes ⇒ Many Jobs, often large
- How are resources allocated?
- How do you actually run you code?
- Who goes first?



Job Scheduler and Resource Management

HPC Clusters need a way to distribute jobs

BUT! It's a complicated task...

- Jobs have various requirements. i.e. CPU,memory,disk-space,Network transportation...
- Some jobs are more important
- Large jobs need lots of nodes.
- Some nodes may be down, Some may have insufficient resources.

BUT WAIT...There's more....

We don't want one person using all of the resources, so we also need something to take care of that.

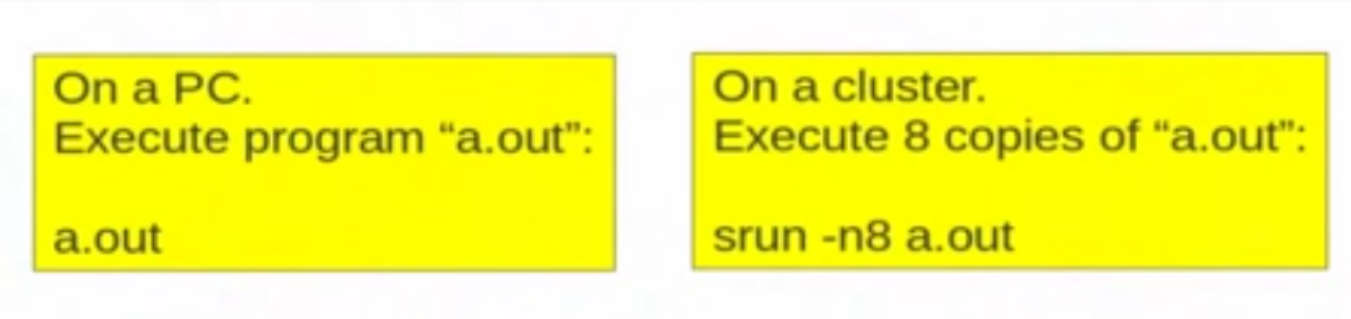
All of this is handled by a piece of software,or software suite.

Some Examples

- Torque(Previously PBS)
- Slurm – Relative Newcomer from LLNL
- There are loads more, but these are very common in science.

Role of the Resource Manager and Job Scheduler

The Resource Manager is like the glue for a parallel computer to execute jobs. It should make using a parallel system as easy as a PC.



Responsibilities of the Resource Manager

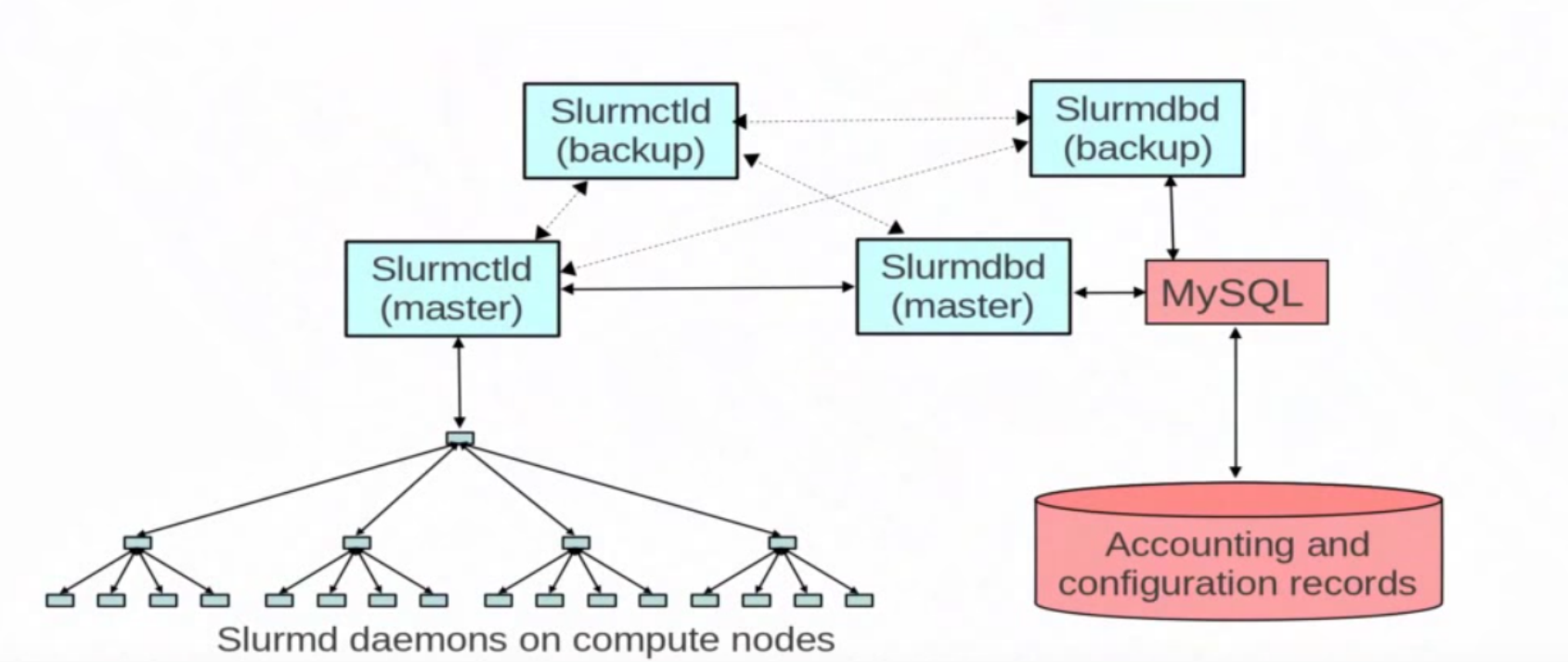
- Hardware. i.e. nodes, sockets,cores,hyperthreads,memory,switches
- Launch and otherwise manage jobs.

Responsibilities of the Job Scheduler

- When there is more work than resources it’s in charge of managing the line(Queue).
  - Complex algorithms that are optimized for any number of things.

Architechture:SLURM

Simple Linux Utility for Resource Management



Daemons

- Slurmctld - Central Controller
  - typically one per cluster
  - Moniters state of resources
  - Manages Job Queue
  - Allocates Resources
- Slurmd - Compute Node Daemon
  - Typically one per compute node
  - Launches and manages tasks
  - Small-light weight
  - Hierarchical communications with configurable fanout
- Slurmdbd - Database daemon
  - Typically one per enterprise
  - Collects accounting info
  - Uploads configurations(limits, fair-share, etc)

Using SLURM

Commands	Description
Sinfo	reports the state of partitions and nodes managed by Slurm (it has a variety of filtering, sorting, and formatting options)
Squeue	reports the state of jobs (it has a variety of filtering, sorting, and formatting options), by default, reports the running jobs in priority order followed by the pending jobs in priority order
Scancel	cancel a pending or running job
Sacct	report job accounting information about active or completed jobs
Sbatch	submit a job script for later execution (the script typically contains one or moresrun commands to launch parallel tasks)
salloc	allocate resources for a job in real time (typically used to allocate resources and spawn a shell, in which the srun command is used to launch parallel tasks)
srun	used to submit a job for execution in real time

Questions?