Job scheduling

Sandeep Agrawal
C-DAC Pune

What is a job?

- The user's program is not simply the name of an executable. It has input data and parameters, environment variables, descriptions of computing resources needed to run the application, and output directives. All of these specifications collectively are called a job
- The user submits the job to an HPC system, usually in the form of a job script
- The job script contains a formal specification that requests computing resources, identifies an application to run along with its input data and environment variables, and describes how best to deliver the output data
- The term job size refers to the number of nodes the job requests
- The term job length refers to the total time that a job want to run

Modes of program execution

- Interactive mode
  - Similar way as of using a personal computer
  - Not enabled on all HPC systems
  - Job may have to wait in queue

- Batch mode
  - refers to program execution in background
  - most HPC jobs can use resources of multiple nodes
  - User jobs are accepted and typically wait in queues before adequate resources are available to the job and then it is scheduled for execution.
  - Most resource policies on the HPC systems tend to be exclusive

Batch System

- A typical batch system is composed of three functionalities:
  - Job Scheduler or Workload Manager
  - Resource Manager
  - Execution Manager
- Job Scheduling:
  - The batch system is responsible for receiving the job script.
  - If the job cannot be executed immediately, the job script is added to a queue.
  - The job waits in the queue until the job's requested resources are available.
  - The batch system then runs the job.
- The software component within the batch system which identifies jobs to run, selects the resources for the job, and decides when to run the job is called the scheduler or the workload manager.
- The software component within the batch system which identifies the compute resources and keeps track of their usage and feeds back this information to the workload manager, is called the resource manager.
- Once a scheduler selects the resources for a job and decides to schedule it, the actual job initiation and start of execution is co-ordinated by the execution manager of the batch system.

Batch Scheduling

- Scheduling policies:
  - FCFS:
    - The last job submitted is added to the bottom of the queue. The rule for a FCFS scheduler is that no other job in the queue will run before the job that is at the top of the queue. That top job waits in the queue until enough jobs finish to free up the resources that the top priority job needs.
- Back-filling:
  - If the scheduler has the intelligence to launch jobs lower in the queue, on resources that are currently idle, it is called a back-fill scheduler. The back-fill scheduler follows a strict rule to only schedule lower priority jobs on idle resources if it will not delay the start of the top priority job.
- Fair-share:
  - A method to allocate resource shares to a user or groups of users or a project
  - A fair method for ordering jobs based on their usage history
  - The job to be run next is selected from the set of jobs belonging to the most deserving entity
- Preemptive:
  - A job with higher priority can signal currently running job to stop and release resources to allow the high priority job to run.
  - Necessary requirement for pre-emption is support for job checkpoints and restart ability.

Job Queues

- While a single job queue could service an entire system, an HPC cluster is typically partitioned into pools of node resources each with its own job queue.
- While most of a cluster's node resources are dedicated to running production jobs, some nodes are typically set aside to use in debugging applications.
- These two uses, production and debugging, are at cross purposes.
  - Production runs tend to be full sized jobs that last multiple hours.
  - Debugging sessions are typically smaller sized runs and are more short
  - lived.
  - Hence the batch queue is configured with wall clock and job size limits that favor production jobs.
  - The debug queue has shorter time and smaller size limits that ensure more
  - immediate access to resources for smaller, quick running jobs.
- Users specify the appropriate queue when the job is submitted.

Batch system commands

- The batch system provides a collection of commands for users to interact with the scheduler and resource manager.
- There are commands to submit a job, display the job queue, and see status and details of the job itself.
- In addition, there are commands which provide information on the computing resources, showing which resources are allocated, which are idle, and which are off-line or down.

SLURM: sample job script

- #!/bin/bash

- #SBATCH -N 1
- #SBATCH --ntasks-per-node=10
- #SBATCH --mem=20GB
- #SBATCH --time=01:00:00
- #SBATCH --job-name=Test
- #SBATCH --error=job.%J.err
- #SBATCH --output=job.%J.out
- #SBATCH --partition=gpu

cd $SLURM_SUBMIT_DIR

mpiexec.hydra -n $SLURM_NTASKS <exe>

SLURM: some commands

- sbatch: submit a batch script to Slurm
- sinfo: display node partition (queue) summary information
- srun: launch one or more tasks of an application across requested resources
- scancel: cancel a job or job step or signal a running job or job step
- squeue: display the jobs in the scheduling queues, one job per line
- scontrol: display (and modify when permitted) the status of jobs, nodes, partitions, reservations, etc.
- salloc: request an interactive job allocation
- sacct: display accounting data for all jobs and job steps in the Slurm database
- sview: a graphical tool for displaying jobs, partitions, reservations etc.

SLURM: some more features

- Job dependencies
  - sbatch –dependency=<type:job id> <job script>
    - type = after | afterok | singleton
- Job arrays
  - #SBATCH --array=1-4

Popular Batch scheduling software

- Slurm
- PBSPro/Torque
- LSF
- LoadLeveller
- Condor

References

- Understanding HPC Job Schedulers, Dr. J. LAKSHMI, SERC, IISc
- https://its.tntech.edu/display/MON/HPC+Job+Scheduling
- https://slurm.schedmd.com/quickstart.html
- https://hpc.llnl.gov/banks-jobs/running-jobs/slurm-commands

- Acknowledgements: Some of the slides are adapted from: Understanding HPC Job Schedulers, Dr. J. LAKSHMI, SERC, IISc