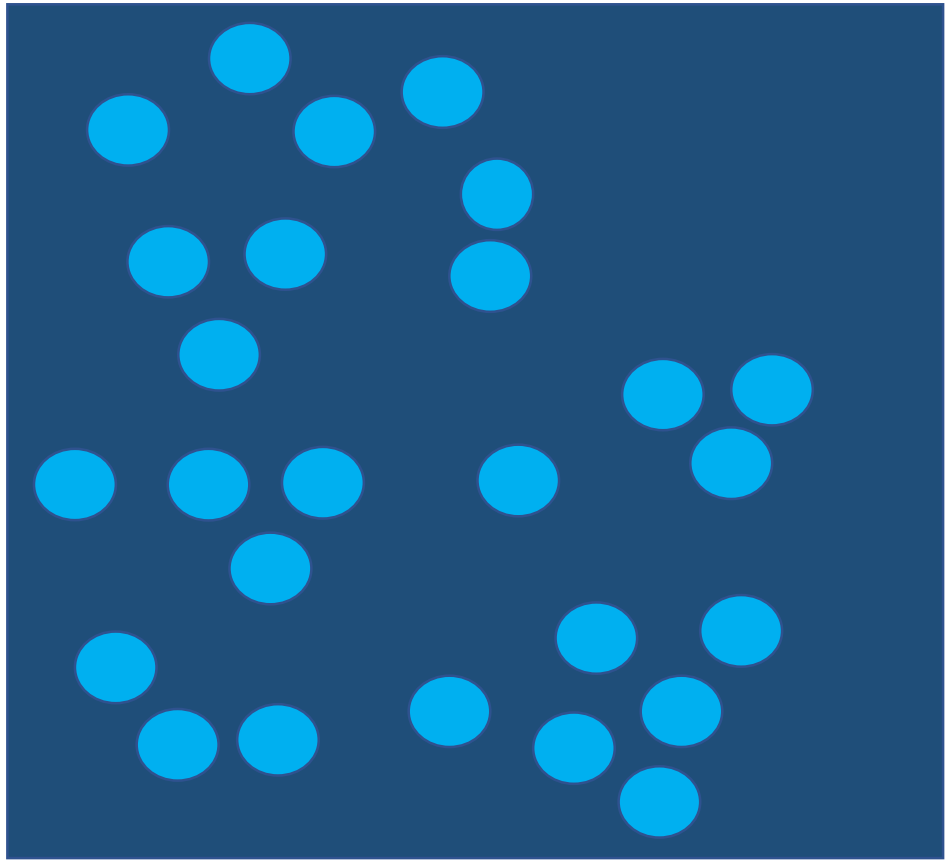
Parallelization-III

Feb 1, 2019

The Chinese Sky

[Source: IDP]

N-body Simulation



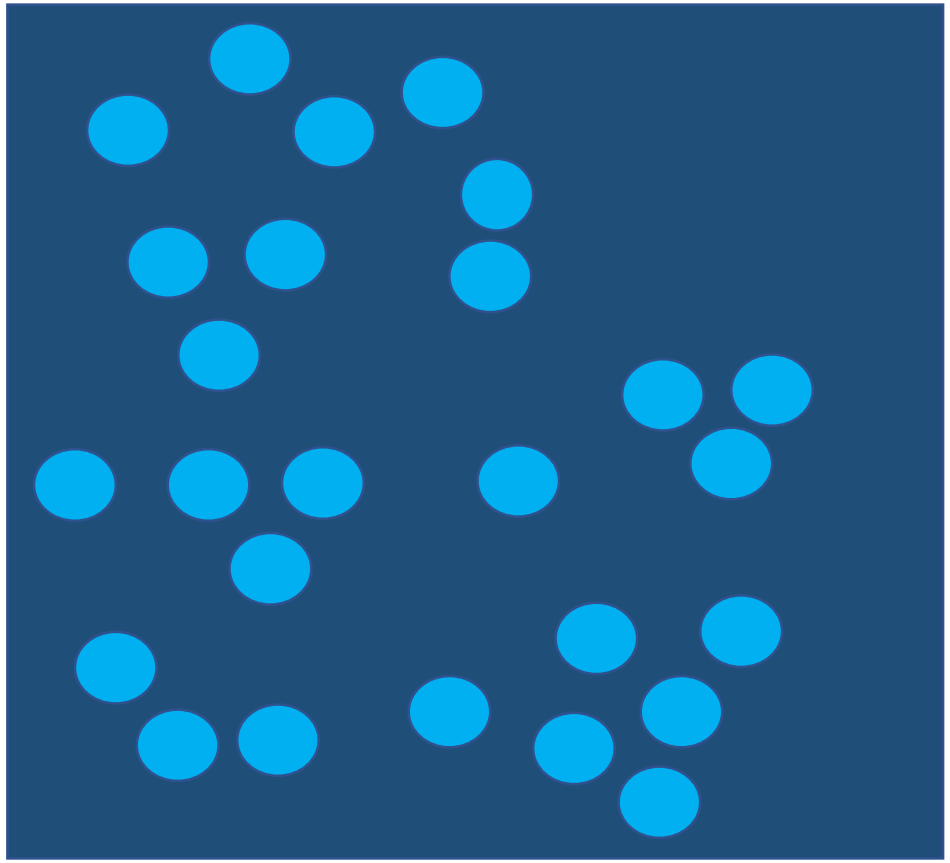
Problem

- N bodies exert force on each other
- Model positions of the particles over time

Applications

- Evolution of the universe
- Crack propagation in a material

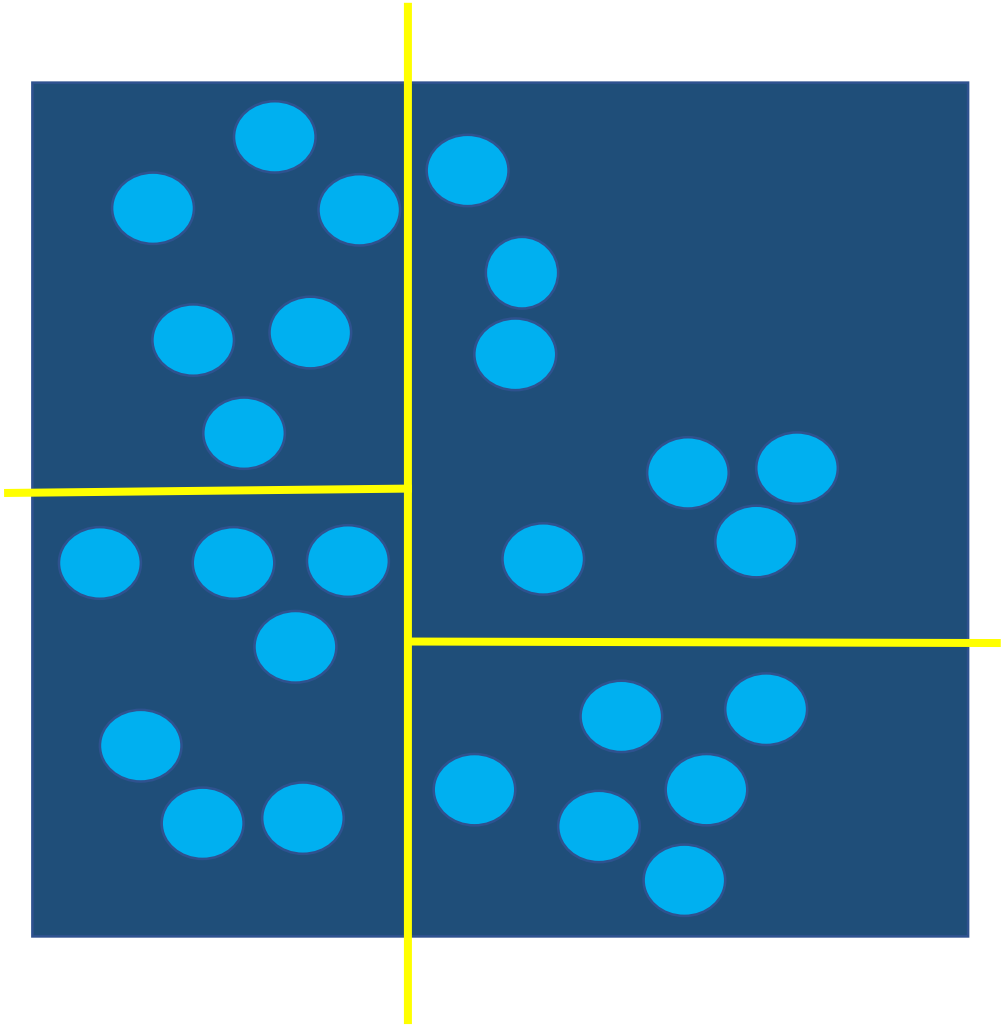
N-body Simulation – Example

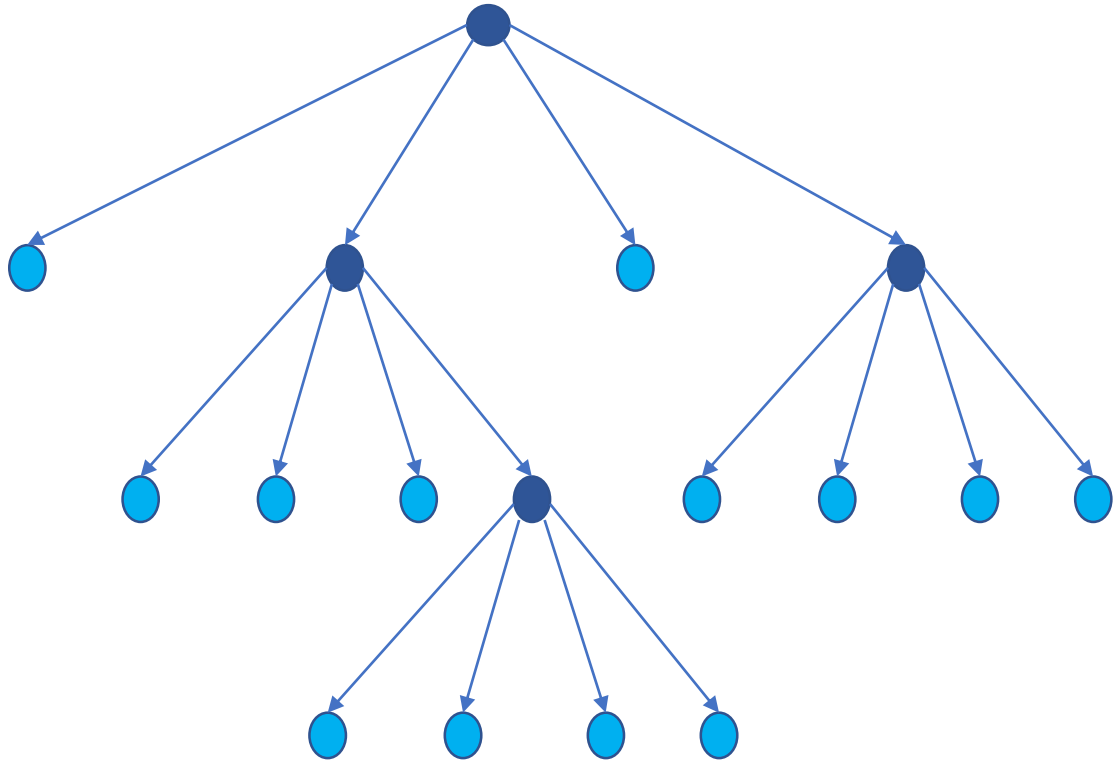


Cosmological simulations

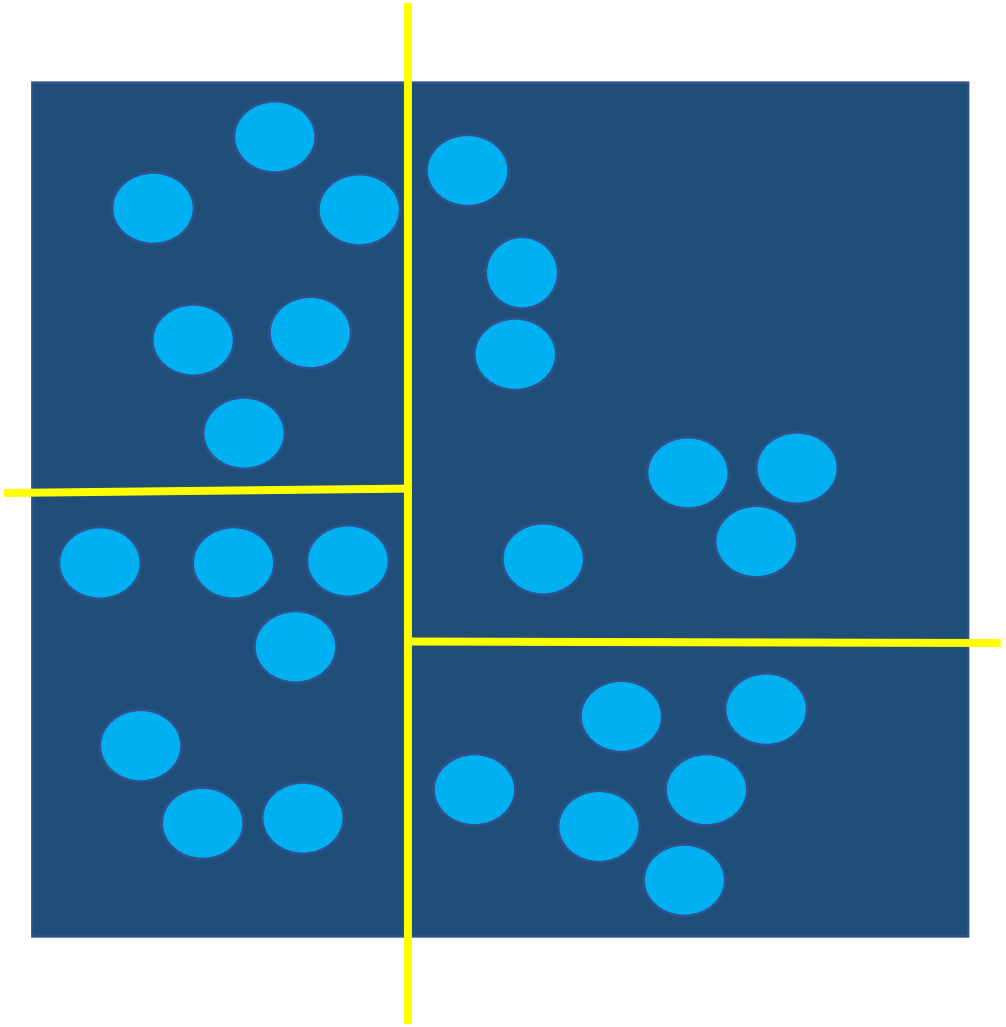
- Net force on every celestial object estimated
- Fields – three-dimensional position, velocity, acceleration, and mass
- Thousands of time steps
- Positions updated, evolving system
- 8000 hrs → 20 hrs [Appel et al., 1985]
- [1986] Barnes and hut
  - Center of mass approximation for distant bodies
  - O(N log N) force computations

N-body Simulation – Force computations





N-body Simulation



Performance considerations

- Positions change across time steps
- Dynamic decomposition
- Irregular communication
- Synchronization between steps
- Repartitioning
- Data reuse

HPC2010

- 369 in top500 in June 2010
- 376 nodes – 368 compute nodes
- Intel Xeon (8 cores per node), later some more nodes were added
- Connected by Infiniband
- Home and scratch file system
- PBS scheduler
- Submit to “courses” queue

Hands-on

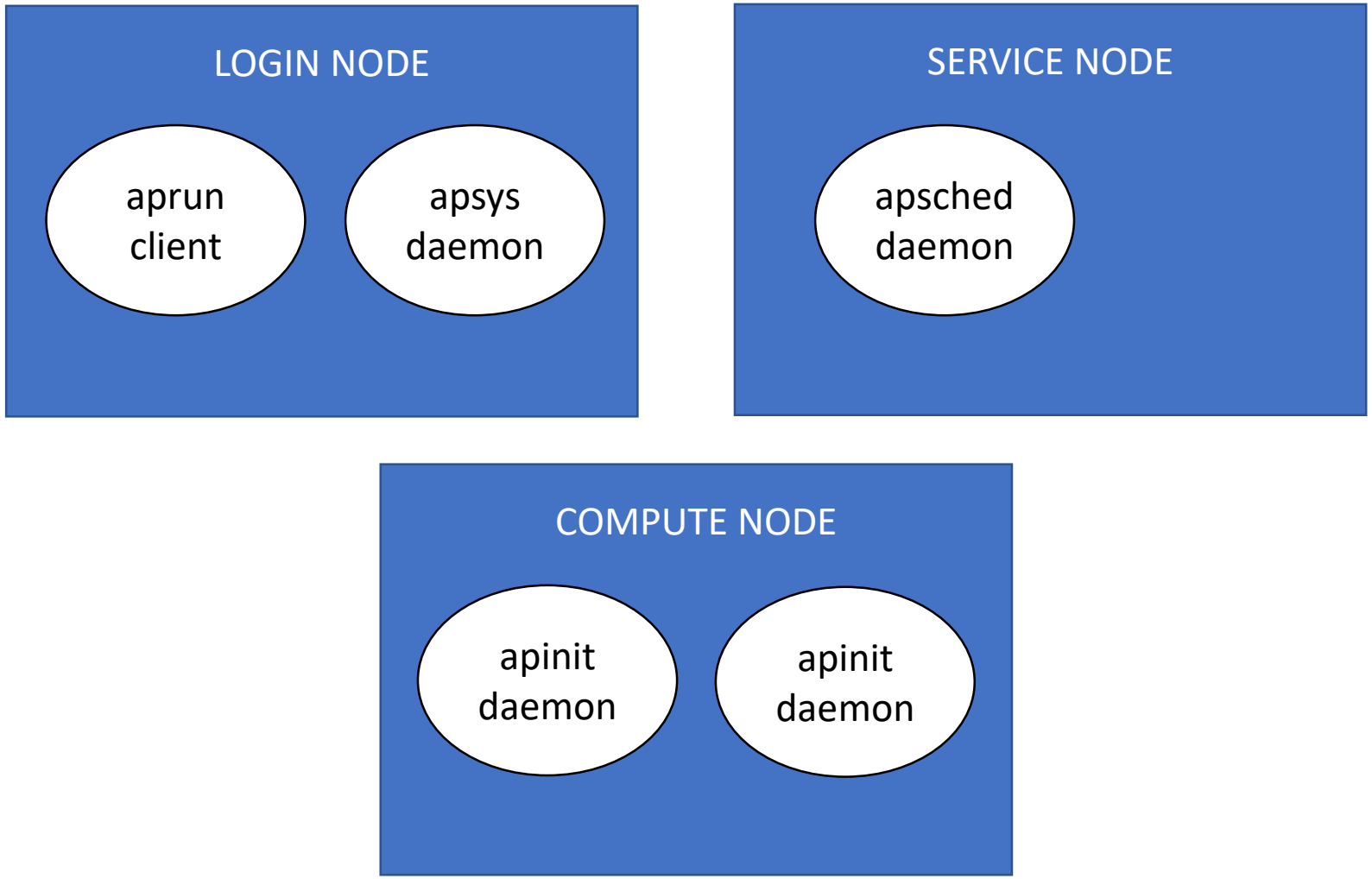
Getting Started

- Login to HPC2010
  - ssh –X <username>@hpc2010.hpc.iitk.ac.in
- Interactive shell to compile, submit job
  - qsub –I –X
- Basic commands (man for all options)
  - qsub <jobscript>
  - qstat –u <username>
  - qdel
  - qhold
  - qrls

Workload managers/Schedulers

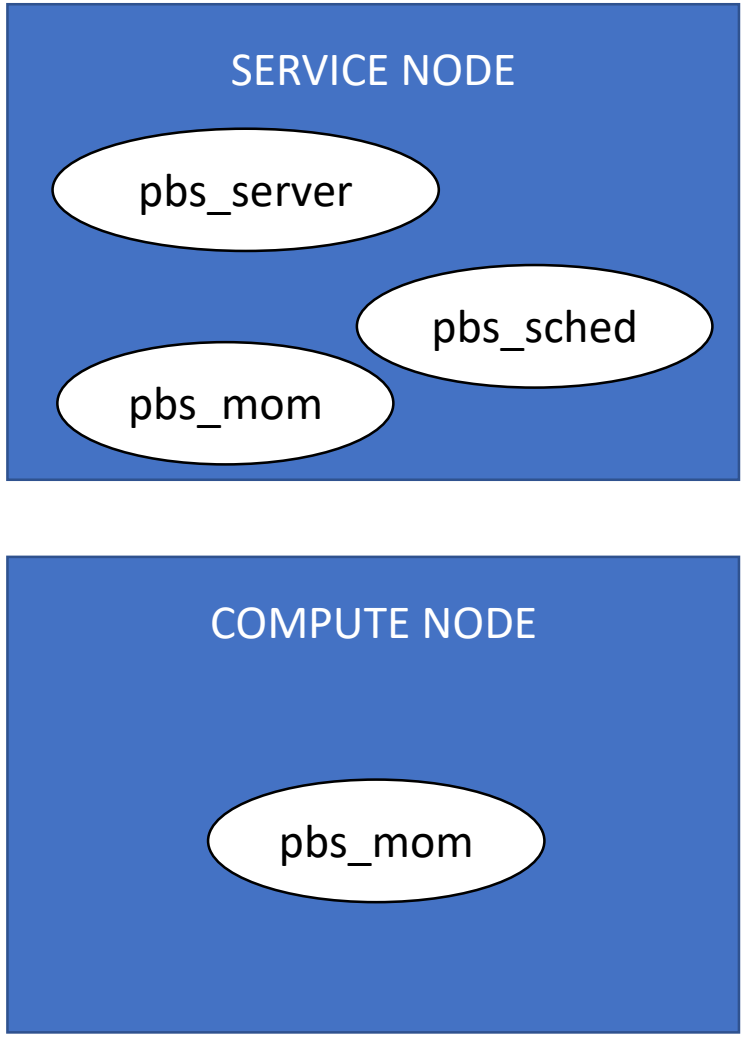
- Portable Batch System (PBS)
- LoadLeveler
- Application Level Placement Scheduler (ALPS)
- Load Sharing Facility (LSF)
- Moab/Torque
- Simple Linux Utility for Resource Management (SLURM)

Application Level Placement Scheduler

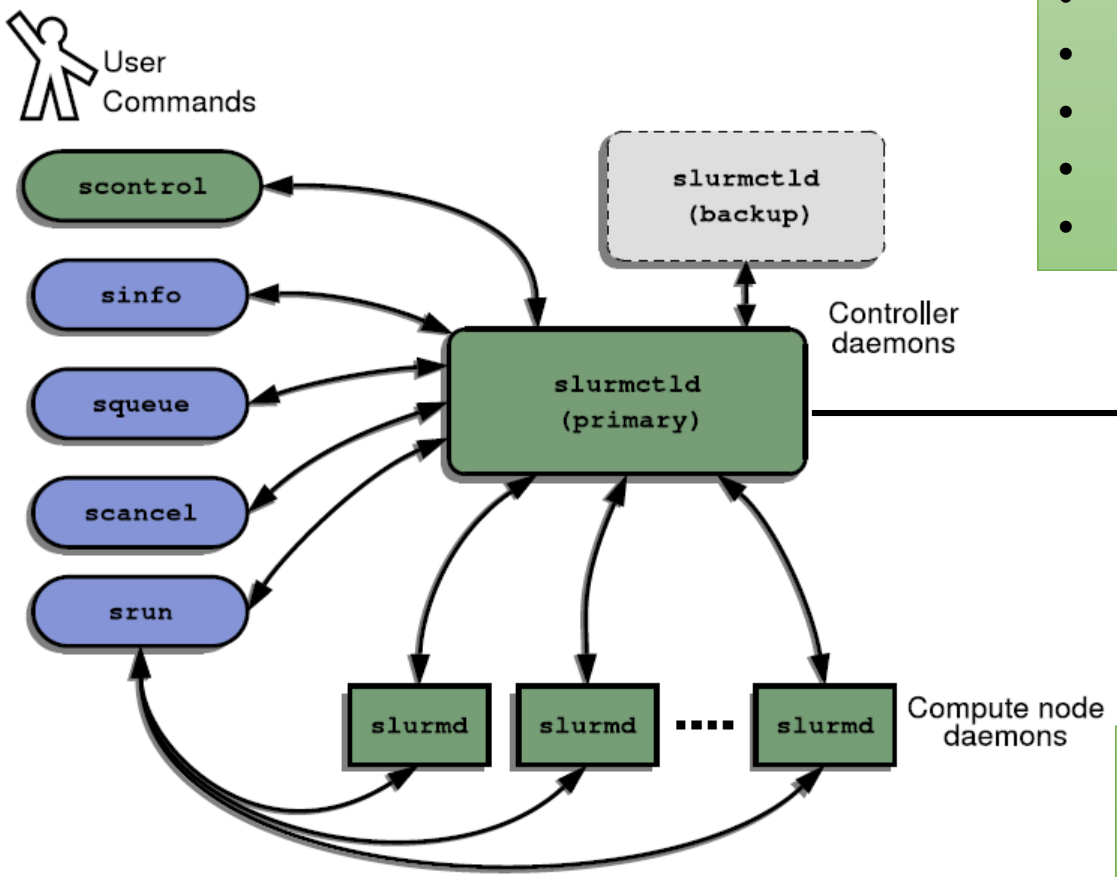


PBS daemons

- Server (pbs_server)
  - Handles PBS commands
  - Creates batch jobs
  - Sends jobs for execution
- Scheduler (pbs_sched)
  - Schedules jobs according to system policy
- MOM (pbs_mom)
  - Manages job execution on hosts
  - Notify server about job completion
  - Resource usage monitor
  - Record diagnostic messages



SLURM



Slurm architecture [Jette et al.]

- Monitors states of nodes
- Accepts job requests
- Maintains queue of requests
- Schedules jobs
- Initiates job execution and cleanup
- Polls slurmd periodically
- Maintains complete state information

- Responds to controller requests
- Maintains job state
- Initiate, manage, cleanup processes
- I/O handling

Queues

- Large
- Medium
- Small
- Debug
- Backfill

Running jobs

http://web.cse.iitk.ac.in/users/pmalakar/cs633/2019/code/feb1.tar.gz

- cd /home/username (you can run from /scratch/username as well)
- cd Feb1
- cd job1
- make
- qsub run1.sh (Job scheduler is Portable Batch System (PBS))
- qstat –u <username>
- Output → jobname.o<jobid>

Job submission script

Open run1.sh

- #PBS –N <jobname>
- #PBS -q courses
- #PBS -l nodes=4:ppn=8 *//resource list*
- #PBS -j oe
- cd $PBS_O_WORKDIR
- source /opt/software/intel/initpaths intel64
- #run the job
- mpirun -machinefile $PBS_NODEFILE -np 32 ./pname.x

Open output file

Running more jobs

- qsub run2.sh
- Open job output file – allocated based on sorted hostnames

- cd job2
- make
- qsub run1.sh
- Check the times in the output file (grep time | sort –k4n)

Job details

- qsub -l nodes=2:ppn=8 run2.sh
- Check output prepended with rank numbers
- tracejob <jobid>
- qstat –n –u <username>

PBS environment variables

- $PBS_NODEFILE
- $PBS_JOBID
- $PBS_JOBNAME
- PBS_O_PATH
- $PBS_NUM_NODES
- …

IMB Benchmarks

- cd job3
- cp -r /opt/software/intel_2015.u2/imb/4.0.2.031/src .
- cd src
- make
- cd ..
- qsub run1.sh
- qsub run2.sh

Structured Tracefile Format

- Tracefile can be in GBs
- Collect in multiple files (arbitrary number)
- Parallel reads and writes