Runing MPI jobs

Open MPI 2.0.1 is currently installed in /share/apps on Prince. There are some notable differences between this version and those installed on Prince cluster:

- From the open-mpi.org web site:
  Please note that mpiexec/mpirun automatically binds processes as of the start of the v1.8 series. Two binding patterns are used in the absence of any further directives:
  - Bind to core - when the number of processes is <= 2
  - Bind to socket - when the number of processes is > 2

- Slurm is plugged into this Open MPI installation as an environment-specific service. Open MPI automatically obtains directly from Slurm the list of hosts and the number of processes to start on each host. So it is unnecessary to specify the --hostfile, --host, or -np options to mpiexec/mpirun. Open MPI will also use SLURM-native mechanisms to launch and kill processes.

In the mpiexec PI calculation example below, it is expressed that this job requires 3 nodes, 4 tasks on each node, and 7 CPU cores for each task.
#!/bin/bash
#
#SBATCH --job-name=mpiexec-test
#SBATCH --nodes=3
#SBATCH --tasks-per-node=4
#SBATCH --cpus-per-task=7
#SBATCH --time=12:00:00
#SBATCH --mem=120GB

module purge
module load openmpi/intel/2.0.1

cd /share/apps/examples/mpi-openmp
mpiexec ./pi

After the job is submitted and running, check its status:

$ squeue -u $USER

JOB ID  PARTITION  USER   ST   TIME  NODES  NODELIST(REASON)
6848    normal     wd35  R    0:07  3 c06-[01-03]

$ scontrol show job -dd 6848  # the command generates many line result including the following
NodeList=c06-[01-03]
BatchHost=c06-01
NumNodes=3 NumCPUs=84 NumTasks=12 CPUs/Task=7 ReqB:S:C:T=0:0:*:*
TRES=cpu=84,mem=360G, node=3
Socks/Node=* NtasksPerN:B:S:C=4:0:*:* CoreSpec=*
Nodes=c06-[01-03] CPU_IDS=0-27 Mem=122880