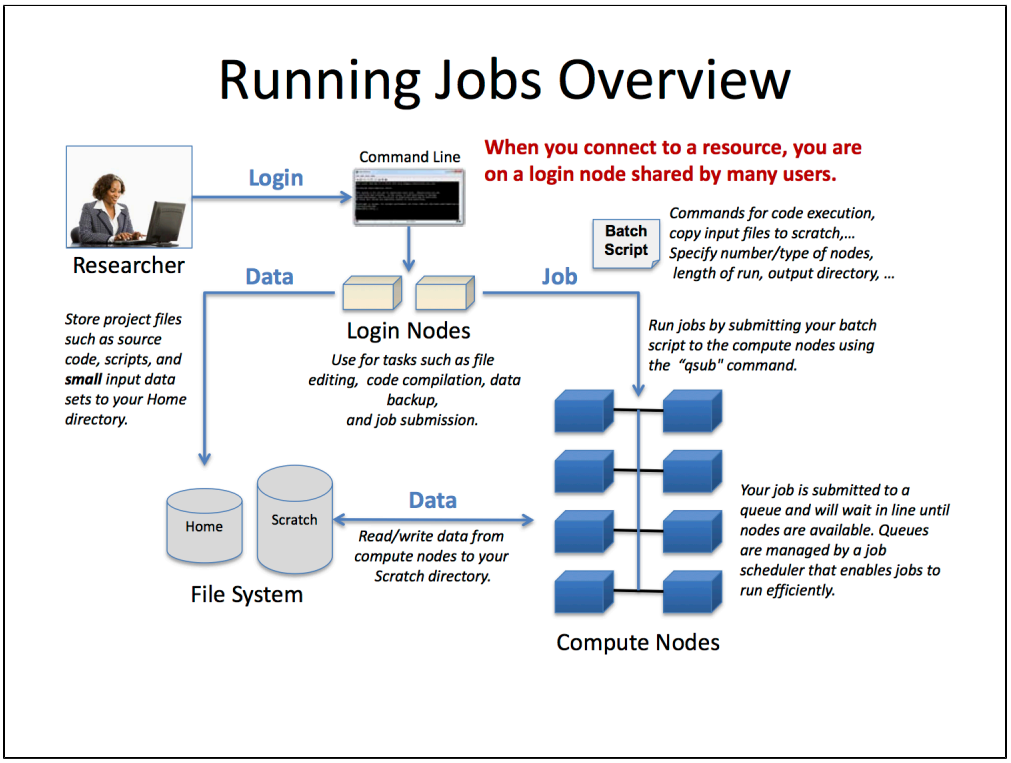
Running Jobs



qsub

1. queues
    a. Blue Waters has 2 different node types (#PBS -l ... resource specification of xe, xk, or x which matches either).
        i. sample batch scripts
        ii. interactive job demo

```
terri@meadowlark: ~
        MPI_Comm_rank(MPI_COMM_WORLD, &rank);
        MPI_Comm_size(MPI_COMM_WORLD, &size);

        MPI_Get_processor_name(name, &len);

#pragma omp parallel private(myid,nthreads, core)
    {
        nthreads = omp_get_num_threads();

        myid = omp_get_thread_num();
        core= sched_getcpu();
        printf("rank %d of %d on %s core %d", rank, size, name, core);
        printf(" (thread id= %d of %d)\n",myid, nthreads);
    }

        MPI_Finalize();
}
arnoldg@nid27641:~/c> _D_-  1-Cray-_ $ [INTERACTIVE JOB]  \
> cc -g -o hello_world hello_world.c
arnoldg@nid27641:~/c> _D_-  1-Cray-_ $ [INTERACTIVE JOB] \
> module swap PrgEnv-cray PrgEnv-pgi
arnoldg@nid27641:~/c> _D_-  1-PGI-_ $ [INTERACTIVE JOB] \
> cc -g -o hello_world hello_world.c
arnoldg@nid27641:~/c> _D_-  1-PGI-_ $ [INTERACTIVE JOB]
```

1. For metadata and repeatability, it's a good idea to load the modules you want into the job's environment. You may also build or manipulate data from small (1-node) jobs and the launch node which runs your script.

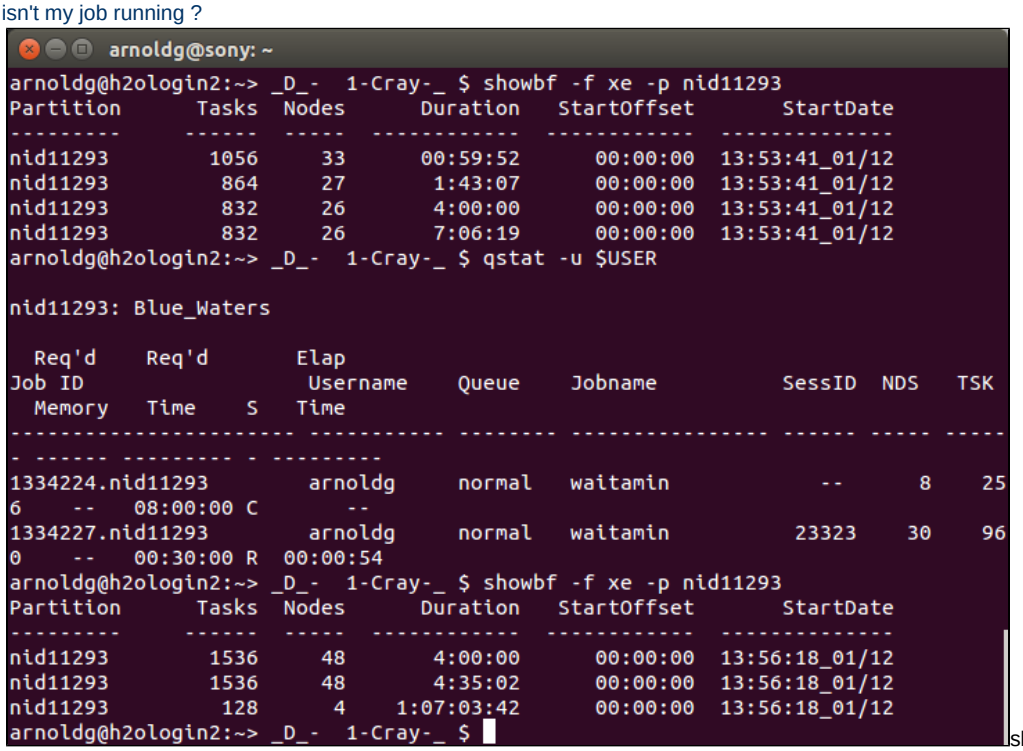
        iii. best practices
            1. Load the modules you want in effect with your job script (see the sample job scripts).
            2. Do not wildly overestimate the wall time of your job (runs your job as soon as possible and helps maintain good

system utilization).

3. Request all the cores on a node (32 for xe, 16 for xk) and place your job's processes explicitly with aprun flags (avoids "claim exceeds reservation...").

b. A fair share policy is in effect for scheduling jobs.

i. So why isn't my job running ?

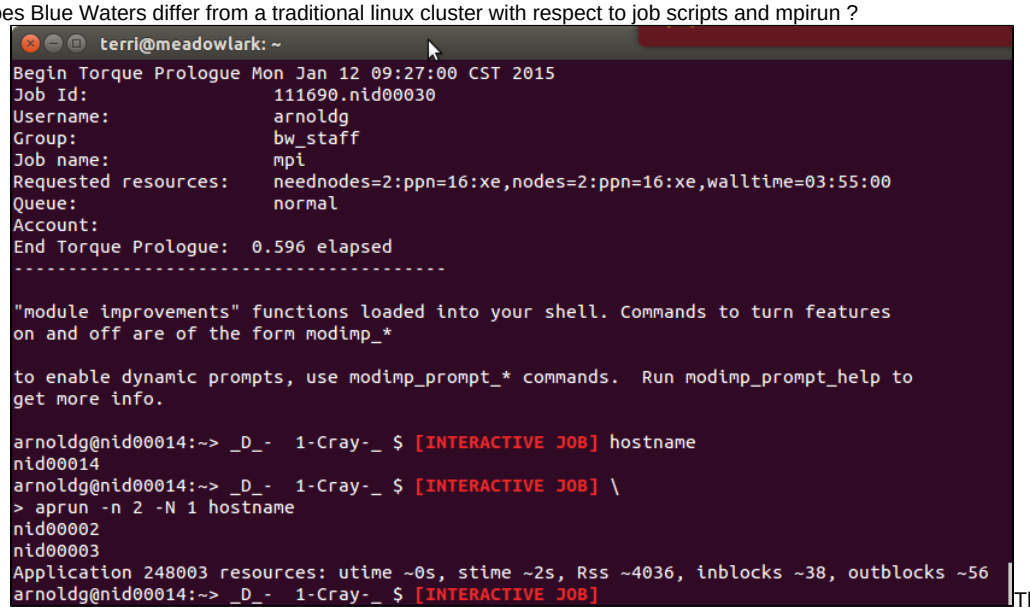
1. 

show bf example with job filling the first slot, elapsed time approx. 5 min.

aprun

1. ## aprun options and examples

a. How does Blue Waters differ from a traditional linux cluster with respect to job scripts and mpirun ?



i. The node running the job script is not part of the MPI proces. **It's not rank 0.**

ii. aprun is the only command that will run tasks on your reserved compute nodes (specified from qsub).

iii. If you need $PBS_NODEFILE :

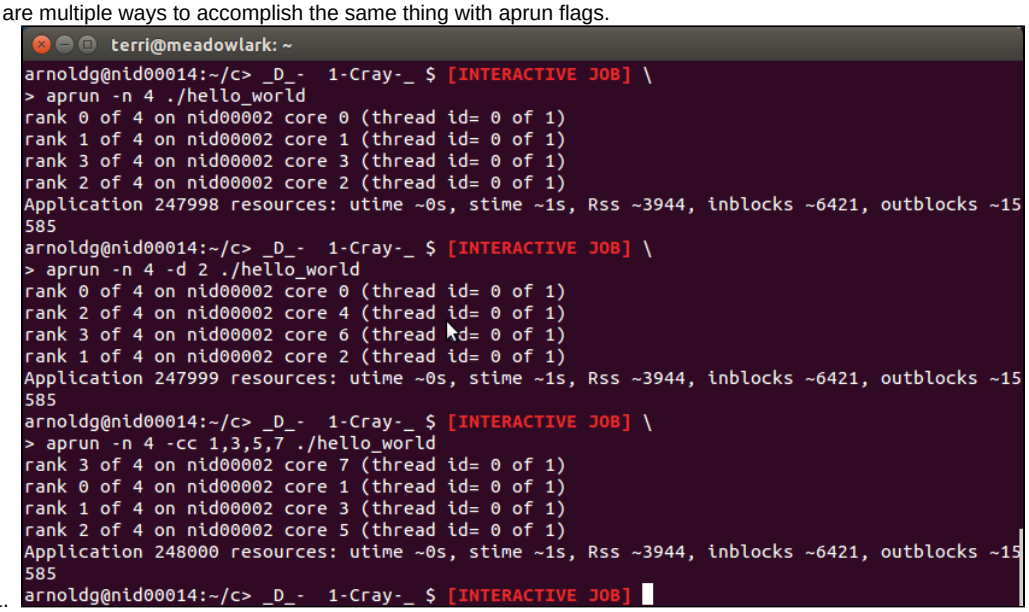
1. You may be heading for a job failure on the cray.
2. Use this and remember the node running the job script will not participate:
   a. aprun -n <nodes> -N 1 hostname > PBS_NODEFILE; export PBS_NODEFILE=`pwd`/PBS_NODEFILE
   b. aprun is still the only way to access the compute nodes--you cannot ssh to compute nodes (instead,

aprun your shell or python script)

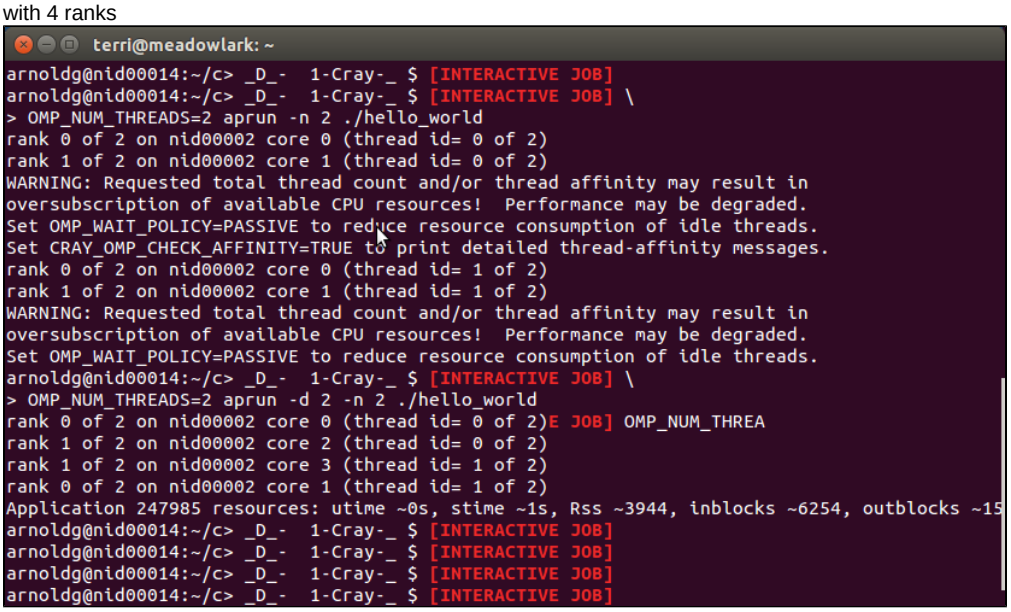
iv. The node running the job script may be shared with other users in the system–it's one of 64 pbs mom nodes.

b. Why are there so many aprun options ?

i. There are multiple ways to accomplish the same thing with aprun flags.

1. 

mpi with 4 ranks

2. 

mpi with 2 ranks and 2 openmp threads

ii. interactive job demo with aprun and hello_world.c

c. best practices

i. Benchmark a test case with a couple variations of aprun ... in the same job.

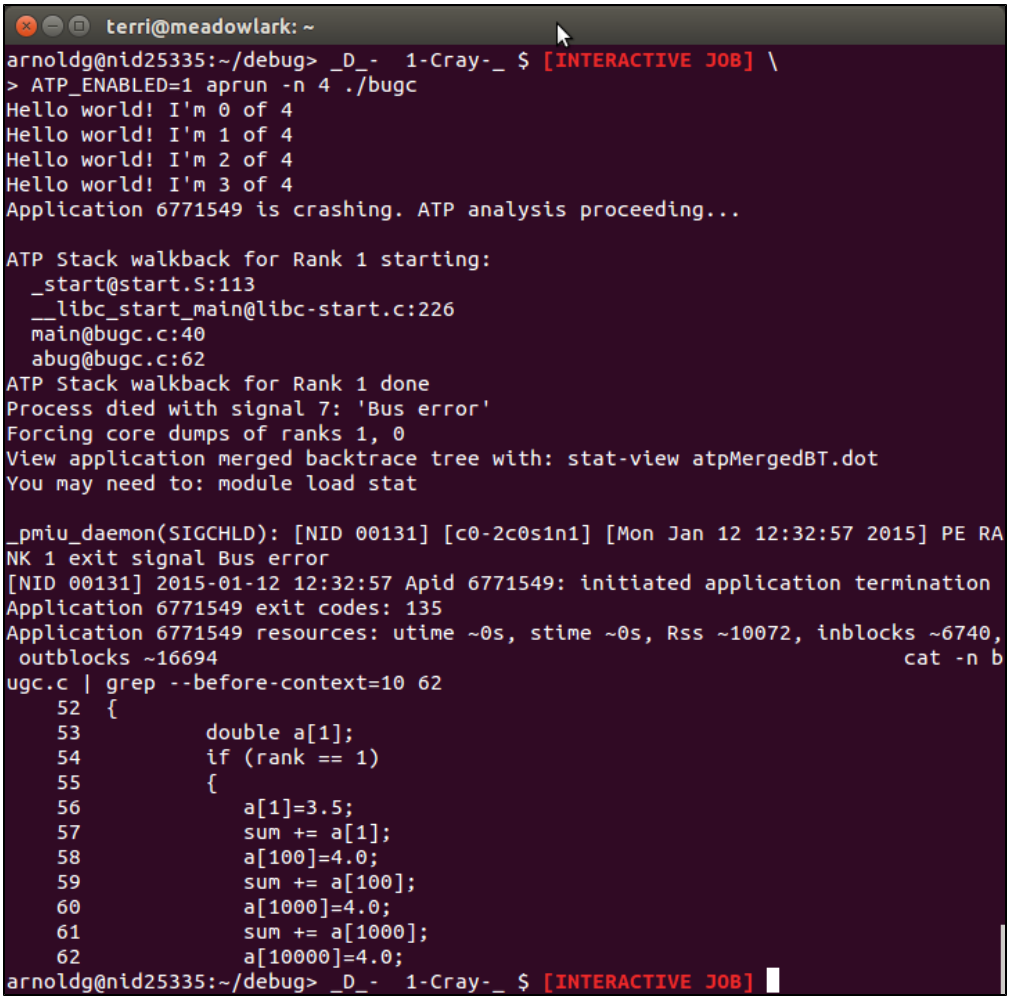
ii. Contact help+bw for advice (<-- *pro tip*).

iii. Use the fastest variant of aprun even if the syntax is a mess (time is money on Blue Waters).

troubleshooting

1. ATP: Abnormal Termination Processing

a.



2. DDT debugger