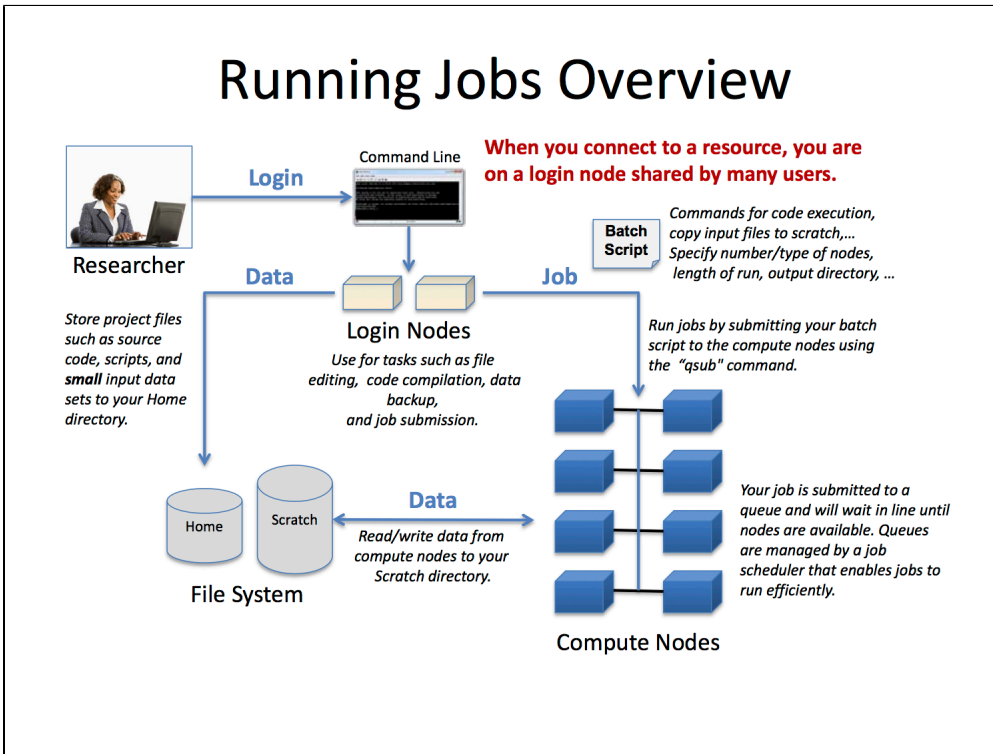


# Running Jobs



## 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. 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 ?

```
arnoldg@sony: ~
arnoldg@h2ologin2:~> _D_- 1-Cray-_ $ showbf -f xe -p nid11293
Partition      Tasks    Nodes    Duration    StartOffset    StartDate
-----
nid11293      1056     33      00:59:52     00:00:00     13:53:41_01/12
nid11293       864     27      1:43:07     00:00:00     13:53:41_01/12
nid11293       832     26      4:00:00     00:00:00     13:53:41_01/12
nid11293       832     26      7:06:19     00:00:00     13:53:41_01/12
arnoldg@h2ologin2:~> _D_- 1-Cray-_ $ qstat -u $USER

nid11293: Blue_Waters

  Req'd   Req'd   Elap
Job ID   Memory Time   S   Username   Queue   Jobname   SessID  NDS   TSK
-----
1334224.nid11293      arnoldg   normal   waitamin   --      8    25
6 -- 08:00:00 C --
1334227.nid11293      arnoldg   normal   waitamin   23323  30   96
0 -- 00:30:00 R 00:00:54
arnoldg@h2ologin2:~> _D_- 1-Cray-_ $ showbf -f xe -p nid11293
Partition      Tasks    Nodes    Duration    StartOffset    StartDate
-----
nid11293      1536     48      4:00:00     00:00:00     13:56:18_01/12
nid11293      1536     48      4:35:02     00:00:00     13:56:18_01/12
nid11293       128      4      1:07:03:42  00:00:00     13:56:18_01/12
1. arnoldg@h2ologin2:~> _D_- 1-Cray-_ $ █
```

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 ?

```
terri@meadowlark:~
Begin Torque Prologue Mon Jan 12 09:27:00 CST 2015
Job Id:      111690.nid00030
Username:    arnoldg
Group:       bw_staff
Job name:    mpi
Requested resources:  neednodes=2:ppn=16:xe,nodes=2:ppn=16:xe,walltime=03:55:00
Queue:       normal
Account:
End Torque Prologue: 0.596 elapsed
-----

"module improvements" functions loaded into your shell. Commands to turn features
on and off are of the form modimp_*

to enable dynamic prompts, use modimp_prompt_* commands. Run modimp_prompt_help to
get more info.

arnoldg@nid00014:~> _D_- 1-Cray-_ $ [INTERACTIVE JOB] hostname
nid00014
arnoldg@nid00014:~> _D_- 1-Cray-_ $ [INTERACTIVE JOB] \
> aprun -n 2 -N 1 hostname
nid00002
nid00003
Application 248003 resources: utime ~0s, stime ~2s, Rss ~4036, inblocks ~38, outblocks ~56
arnoldg@nid00014:~> _D_- 1-Cray-_ $ [INTERACTIVE JOB]
```

i. The node

- i. running the job script is not part of the MPI process. **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,

- 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.

```

❌ ⓘ ⓘ terri@meadowlark: ~
arnoldg@nid00014:~/c> _D_ 1-Cray_ $ [INTERACTIVE JOB] \
> aprun -n 4 ./hello_world
rank 0 of 4 on nid00002 core 0 (thread id= 0 of 1)
rank 1 of 4 on nid00002 core 1 (thread id= 0 of 1)
rank 3 of 4 on nid00002 core 3 (thread id= 0 of 1)
rank 2 of 4 on nid00002 core 2 (thread id= 0 of 1)
Application 247998 resources: utime ~0s, stime ~1s, Rss ~3944, inblocks ~6421, outblocks ~15585
arnoldg@nid00014:~/c> _D_ 1-Cray_ $ [INTERACTIVE JOB] \
> aprun -n 4 -d 2 ./hello_world
rank 0 of 4 on nid00002 core 0 (thread id= 0 of 1)
rank 2 of 4 on nid00002 core 4 (thread id= 0 of 1)
rank 3 of 4 on nid00002 core 6 (thread id= 0 of 1)
rank 1 of 4 on nid00002 core 2 (thread id= 0 of 1)
Application 247999 resources: utime ~0s, stime ~1s, Rss ~3944, inblocks ~6421, outblocks ~15585
arnoldg@nid00014:~/c> _D_ 1-Cray_ $ [INTERACTIVE JOB] \
> aprun -n 4 -cc 1,3,5,7 ./hello_world
rank 3 of 4 on nid00002 core 7 (thread id= 0 of 1)
rank 0 of 4 on nid00002 core 1 (thread id= 0 of 1)
rank 1 of 4 on nid00002 core 3 (thread id= 0 of 1)
rank 2 of 4 on nid00002 core 5 (thread id= 0 of 1)
Application 248000 resources: utime ~0s, stime ~1s, Rss ~3944, inblocks ~6421, outblocks ~15585
1. arnoldg@nid00014:~/c> _D_ 1-Cray_ $ [INTERACTIVE JOB] █

```

with 4 ranks

```

❌ ⓘ ⓘ terri@meadowlark: ~
arnoldg@nid00014:~/c> _D_ 1-Cray_ $ [INTERACTIVE JOB]
arnoldg@nid00014:~/c> _D_ 1-Cray_ $ [INTERACTIVE JOB] \
> OMP_NUM_THREADS=2 aprun -n 2 ./hello_world
rank 0 of 2 on nid00002 core 0 (thread id= 0 of 2)
rank 1 of 2 on nid00002 core 1 (thread id= 0 of 2)
WARNING: Requested total thread count and/or thread affinity may result in
oversubscription of available CPU resources! Performance may be degraded.
Set OMP_WAIT_POLICY=PASSIVE to reduce resource consumption of idle threads.
Set CRAY_OMP_CHECK_AFFINITY=TRUE to print detailed thread-affinity messages.
rank 0 of 2 on nid00002 core 0 (thread id= 1 of 2)
rank 1 of 2 on nid00002 core 1 (thread id= 1 of 2)
WARNING: Requested total thread count and/or thread affinity may result in
oversubscription of available CPU resources! Performance may be degraded.
Set OMP_WAIT_POLICY=PASSIVE to reduce resource consumption of idle threads.
arnoldg@nid00014:~/c> _D_ 1-Cray_ $ [INTERACTIVE JOB] \
> OMP_NUM_THREADS=2 aprun -d 2 -n 2 ./hello_world
rank 0 of 2 on nid00002 core 0 (thread id= 0 of 2)E JOB] OMP_NUM_THREA
rank 1 of 2 on nid00002 core 2 (thread id= 0 of 2)
rank 1 of 2 on nid00002 core 3 (thread id= 1 of 2)
rank 0 of 2 on nid00002 core 1 (thread id= 1 of 2)
Application 247985 resources: utime ~0s, stime ~1s, Rss ~3944, inblocks ~6254, outblocks ~15585
arnoldg@nid00014:~/c> _D_ 1-Cray_ $ [INTERACTIVE JOB]
arnoldg@nid00014:~/c> _D_ 1-Cray_ $ [INTERACTIVE JOB]
arnoldg@nid00014:~/c> _D_ 1-Cray_ $ [INTERACTIVE JOB]
2. arnoldg@nid00014:~/c> _D_ 1-Cray_ $ [INTERACTIVE JOB] █

```

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

```

terri@meadowlark: ~
arnoldg@nid25335:~/debug> _D_- 1-Cray-_ $ [INTERACTIVE JOB] \
> ATP_ENABLED=1 aprun -n 4 ./bugc
Hello world! I'm 0 of 4
Hello world! I'm 1 of 4
Hello world! I'm 2 of 4
Hello world! I'm 3 of 4
Application 6771549 is crashing. ATP analysis proceeding...

ATP Stack walkback for Rank 1 starting:
_start@start.S:113
__libc_start_main@libc-start.c:226
main@bugc.c:40
abug@bugc.c:62
ATP Stack walkback for Rank 1 done
Process died with signal 7: 'Bus error'
Forcing core dumps of ranks 1, 0
View application merged backtrace tree with: stat-view atpMergedBT.dot
You may need to: module load stat

_pmiu_daemon(SIGCHLD): [NID 00131] [c0-2c0s1n1] [Mon Jan 12 12:32:57 2015] PE RA
NK 1 exit signal Bus error
[NID 00131] 2015-01-12 12:32:57 Apid 6771549: initiated application termination
Application 6771549 exit codes: 135
Application 6771549 resources: utime ~0s, stime ~0s, Rss ~10072, inblocks ~6740,
outblocks ~16694
ugc.c | grep --before-context=10 62
52 {
53     double a[1];
54     if (rank == 1)
55     {
56         a[1]=3.5;
57         sum += a[1];
58         a[100]=4.0;
59         sum += a[100];
60         a[1000]=4.0;
61         sum += a[1000];
62         a[10000]=4.0;
arnoldg@nid25335:~/debug> _D_- 1-Cray-_ $ [INTERACTIVE JOB]

```

- a.
2. DDT debugger