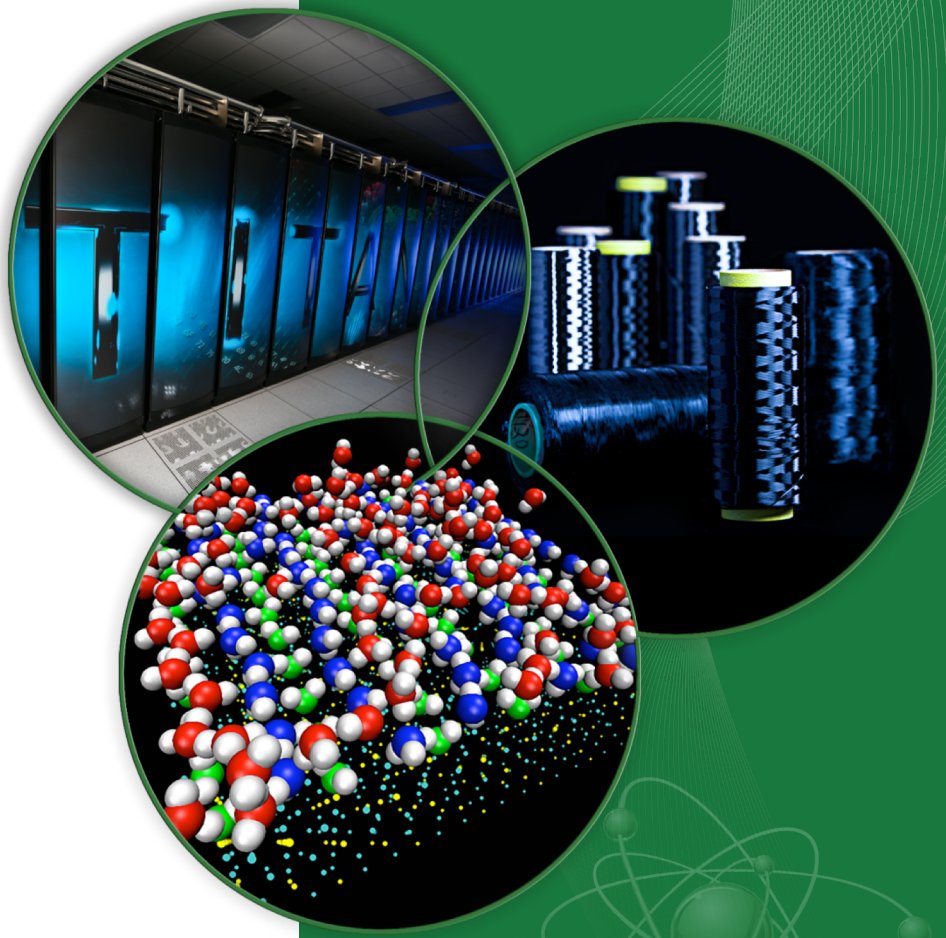
Titan Job Launch Introduction

OLCF Introduction to HPC Workshop

Chris Fuson

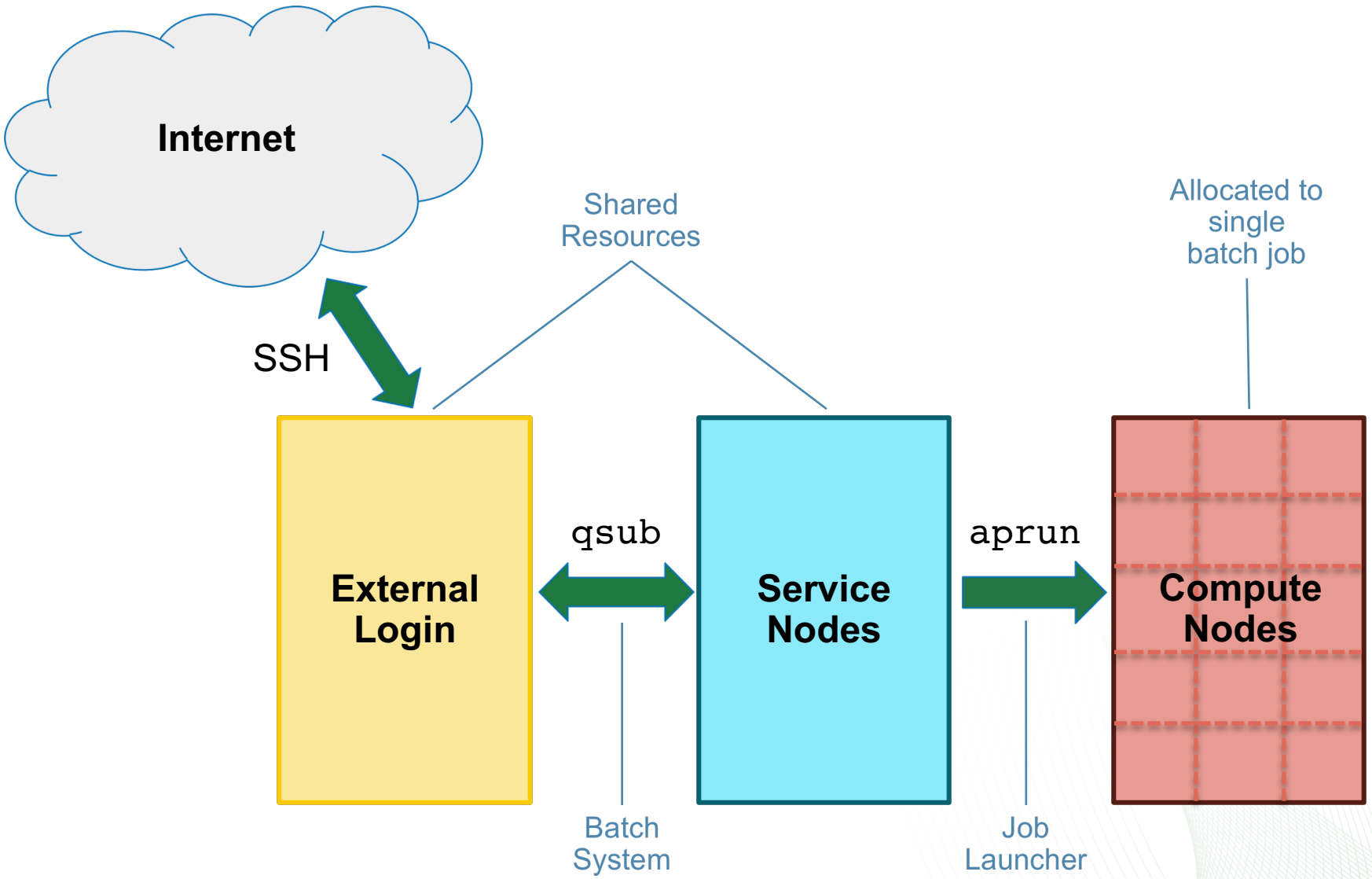
June 26, 2018



ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

Titan Login, Launch, Compute Nodes



Titan Parallel Job Execution

Batch System

Torque/MOAB

- Allocates compute resources
- Batch scheduler
- Allocates entire nodes
- Torque based on PBS

Job Launcher

aprun

- Similar functionality mpirun
- Cray specific
- Used on Titan and Eos
- Only way to reach the compute nodes

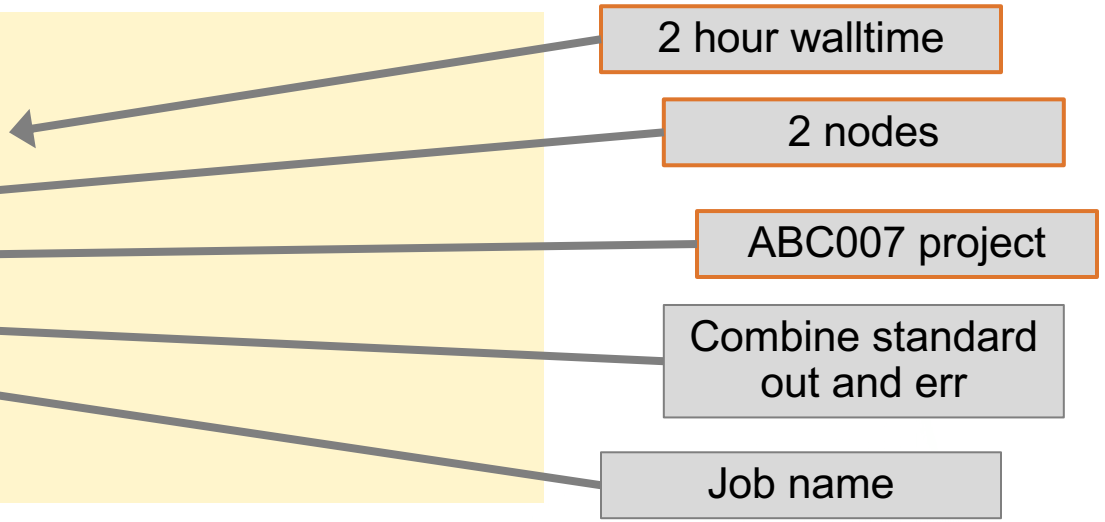
Example Batch Script

Batch script example

```
#!/bin/bash

#PBS –l walltime=02:00:00
#PBS –l nodes=2
#PBS -A abc007
#PBS –j oe
#PBS -N example

aprun -n2 hostname
```



Batch submission

```
titan-ext3> qsub example.pbs
4106766
titan-ext3>
```

Common qsub Options

Option	Example Usage	Description
-l walltime	#PBS –l walltime=01:00:00	Requested Walltime hours:minutes:seconds
-l nodes	#PBS –l nodes=1024	Number of nodes
-A	#PBS –A *ABC123*	Project to which the job should be charged
-N	#PBS –N MyJobName	Name of the job. If not specified, will be set to name of batch job script.
-j oe	#PBS –j oe	Combine STDOUT and STDERR
-e	#PBS –e joberr	File into which job STDERR should be directed
-o	#PBS –o jobout	File into which job STDOUT should be directed
-m	#PBS –m b #PBS –m e	Send job report via email once job completes (e) or begins (b)
-V	#PBS -V	Exports all environment variables from the submitting shell into the batch job shell. Since the login nodes differ from the service nodes, using the '-V' option is **not recommended**. Users should create the needed environment within the batch job.

**More details and flags can be found in the qsub man page*

Interactive Batch Job

- Allows access to compute resources interactively
- Through batch system similar to batch script submission, but returns prompt on launch node
- Run multiple apruns with only one queue wait, very useful for testing and debugging
- Syntax
  - -I
  - Most other batch flags valid
  - Add batch flags to command line

Presentation examples use the following to allocate resources

```
titan-ext3> qsub –I –lnodes=2 –lwalltime=01:00:00 –A prj123
qsub: waiting for job 4106869 to start
qsub: job 4106869 ready
titan-batch2> aprun -n2 –N1 hostname
nid00339
nid00332
titan-batch2>
```

Common Torque/MOAB Commands

Function	Command
Submit	qsub
Monitor Queue	showq/qstat
Alter Queued Job	qalter
Remove Queued Job	qdel
Hold Queued Job	qhold
Release Held Job	qrls

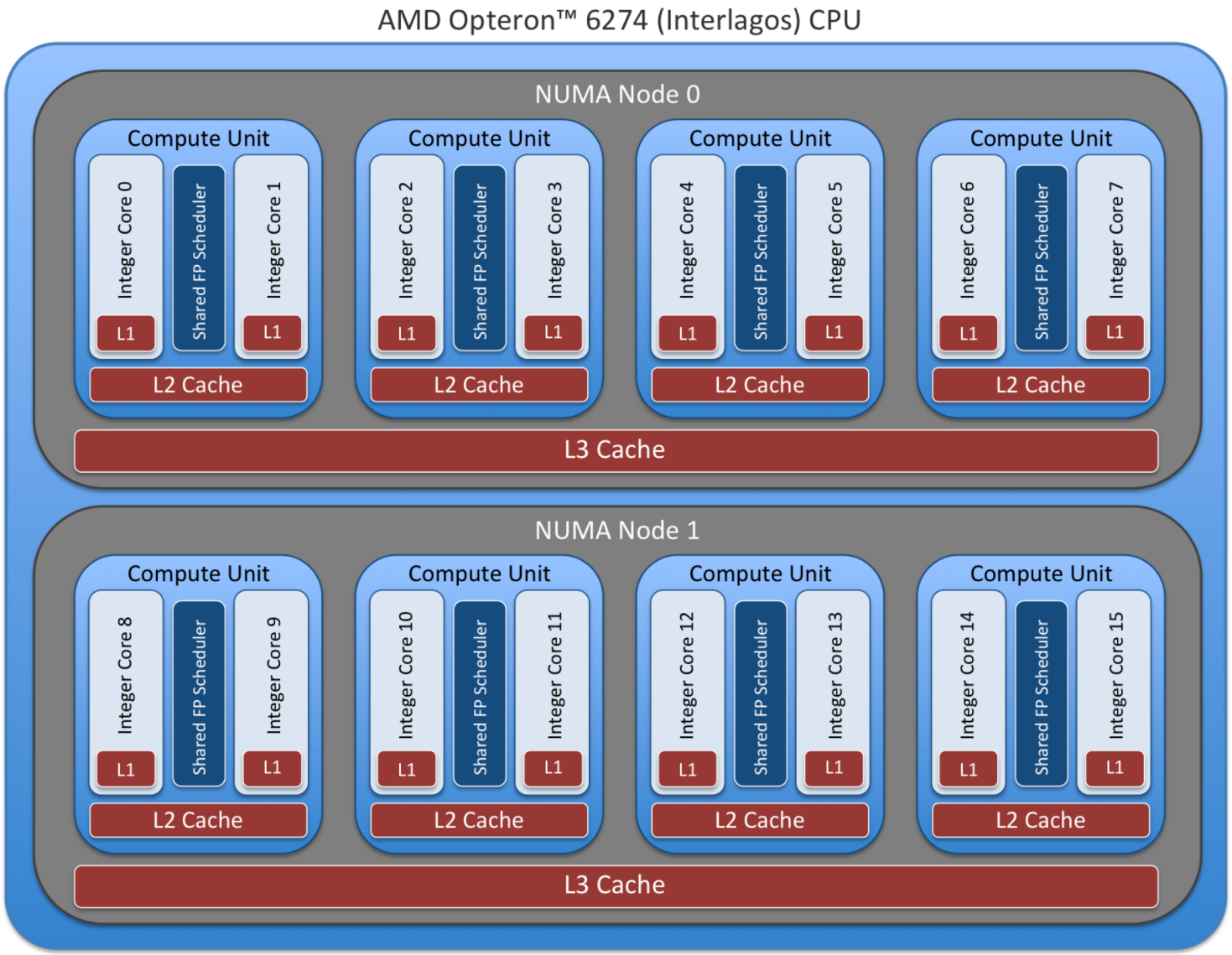
Viewing the Batch Queue

- ‘qstat’
  - Display all queued jobs. Basic output.
- ‘showq’
  - Will also show all queued jobs, but with more useful detail
  - Queue organized into three high level categories
    - 1) Running 2) Pending Eligible 3) Pending Ineligible
- ‘checkjob <*jobID*>’
  - Display more details about given job
  - MOAB
- ’qstat –f <*jobID*>’
  - Display details about given job
  - Torque

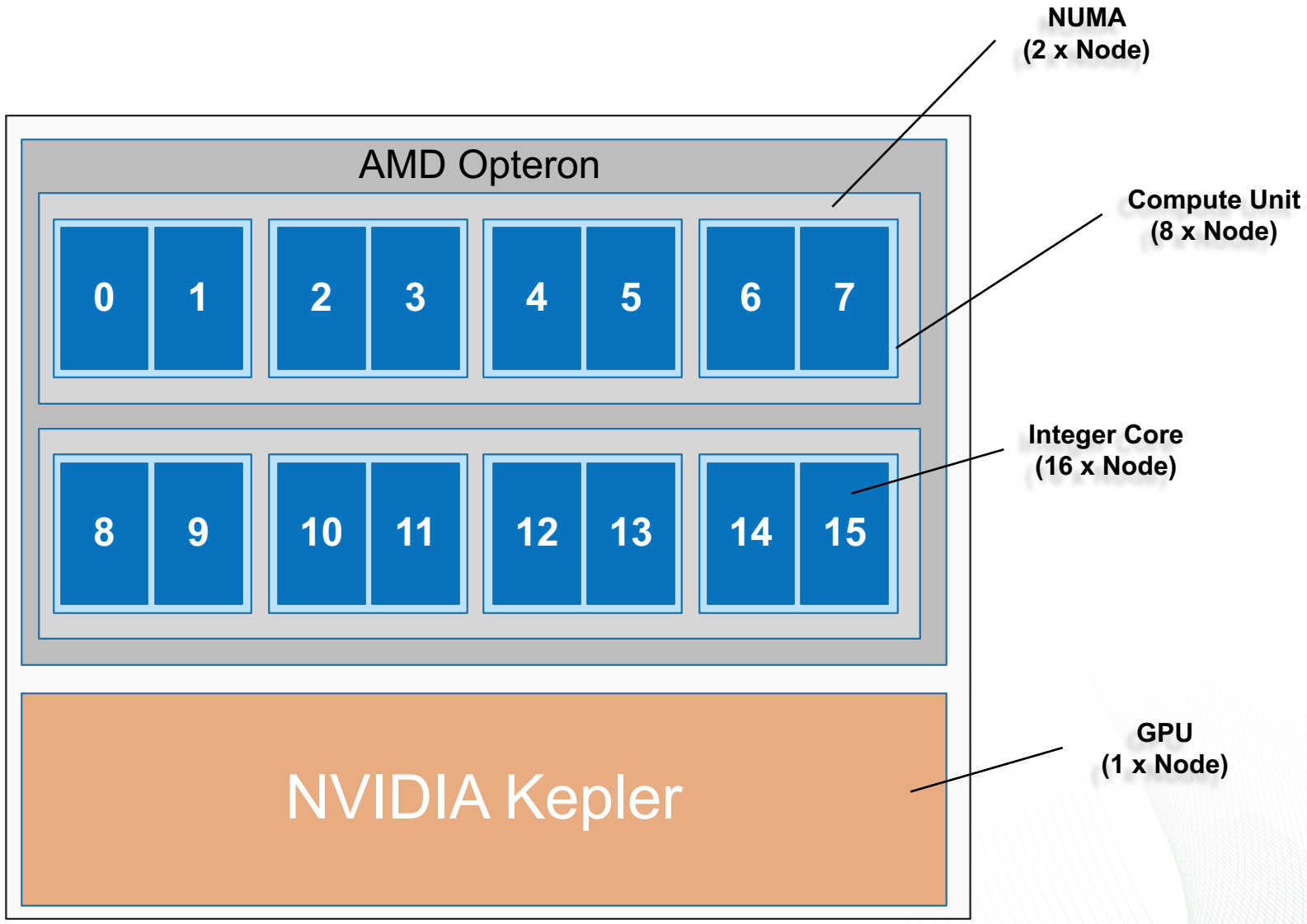
showq Example

```
titan-ext3> showq
active jobs------------------------
JOBID            USERNAME      STATE   PROCS   REMAINING         STARTTIME
4106888             user1       Running    2112      5:59:57         Sun Jun 24 10:25:57
4106872             user2       Running      16     00:14:16         Sun Jun 24 10:20:16
29 active jobs       243584 of 300448 processors in use by local jobs (81.07%)
                  15359 of 18667 nodes active      (82.28%)
eligible jobs----------------------
JOBID            USERNAME      STATE   PROCS    WCLIMIT         QUEUETIME
4106917             user6        Idle      32000     12:00:00      Sat Jun 23 16:24:12
4106917             user5        Idle         32      2:00:00      Sun Jun 24 10:36:47
137 eligible job
blocked jobs-----------------------
JOBID            USERNAME      STATE    PROCS    WCLIMIT         QUEUETIME
3699467             user8      UserHold   118592    1:00:00:00     Thu Jun 21 01:44:47
3929851             user1     BatchHold       16       1:00:00      Fri Jun 22 17:43:22
362 blocked jobs
Total jobs:  528
titan-ext3>
```

Titan Compute Node

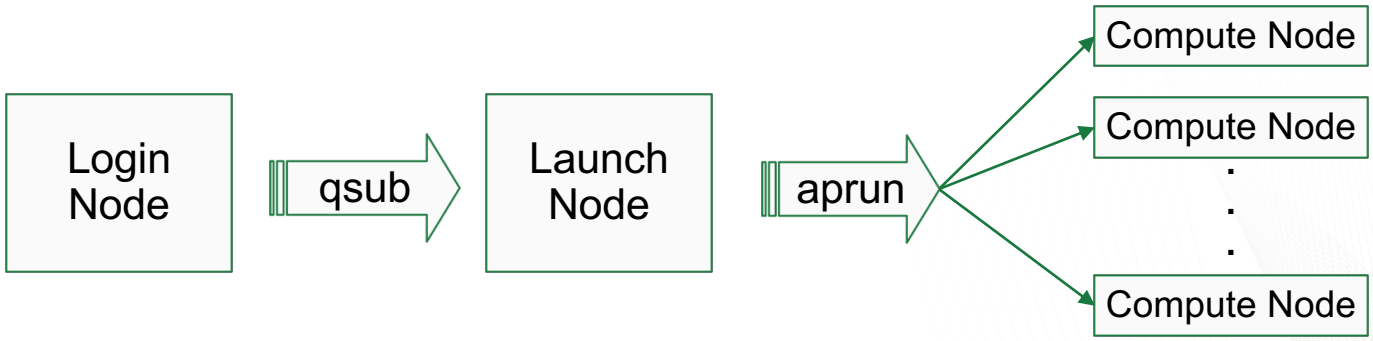


Titan Compute Node



aprun Introduction

- Launch job across compute resources
  - Compute nodes can only be reached via the aprun command
- Similar functionality to mpirun
- Non-aprun commands executed on launch node
- Single simultaneous aprun per node
- Compute nodes can not see home areas



Home Area Access

- Compute nodes can not see NFS home areas
- Needed input can not be in home areas
- Attempts to access home areas will result in an error similar to the following

```
titan-batch2> aprun hostname
[NID 17929] 2018-06-24 14:21:02 Exec /bin/hostname failed: chdir
/autofs/nccs-svm1_home No such file or directory
titan-batch2>
```

aprun Common Options

Flag	Description
-n	Number of MPI tasks/ranks
-N	MPI Tasks/ranks per node
-S	MPI Tasks/ranks per NUMA
-j	Cores per compute unit
-d	Threads per MPI rank/task
-r	Assign system services associated with your application to a compute core. Helps reduce jitter.

**for additional flags see the aprun man page*

Basic aprun Examples

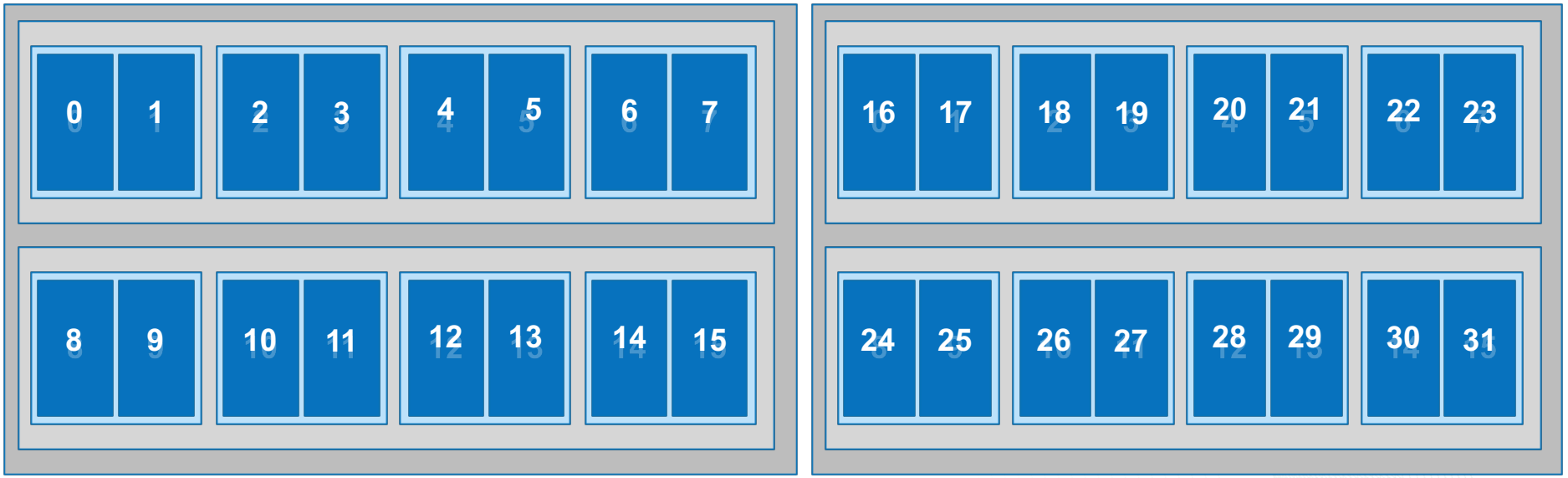
Description	Jsrun command	Layout notes
32 MPI tasks	***aprun –n 32 ./a.out***	2 nodes: 16 tasks node1, 16 tasks on node2
1 MPI task per compute unit	***aprun –n 16 –S 4 –j 1 ./a.out***	2 nodes, 4 tasks per NUMA, 1 task per compute unit
4 MPI tasks per node	***aprun –n 8 –S 2 –j 1 ./a.out***	2 nodes, 2 tasks per NUMA, 1 task per compute unit
8 threads per MPI task	***aprun –n 2 –N 1 –S 1 –d 8 ./a.out***	2 nodes, 8 threads per node

32 MPI Tasks

aprun –n 32 ./a.out

32 MPI tasks

MPICH_RANK_REORDER_DISPLAY can be used to view the layout.

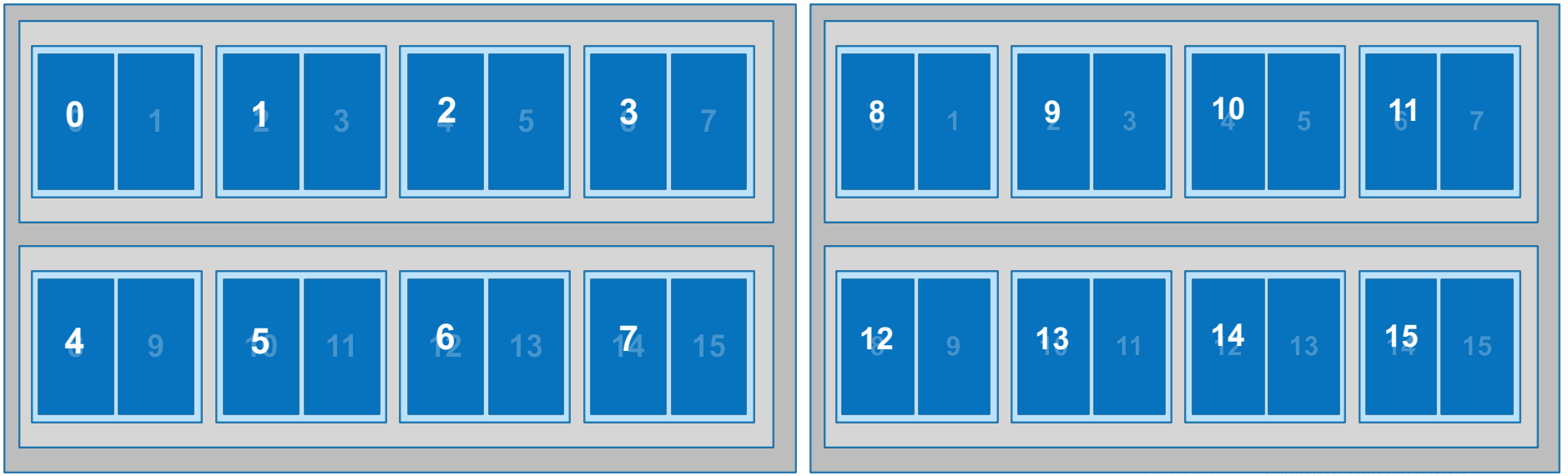


1 MPI task per compute unit

aprun –n 16 –S 4 –j 1 ./a.out

- 16 MPI ranks
- 4 ranks per NUMA
- 1 rank per compute unit

-S and –j provide layout control



4 MPI tasks per node

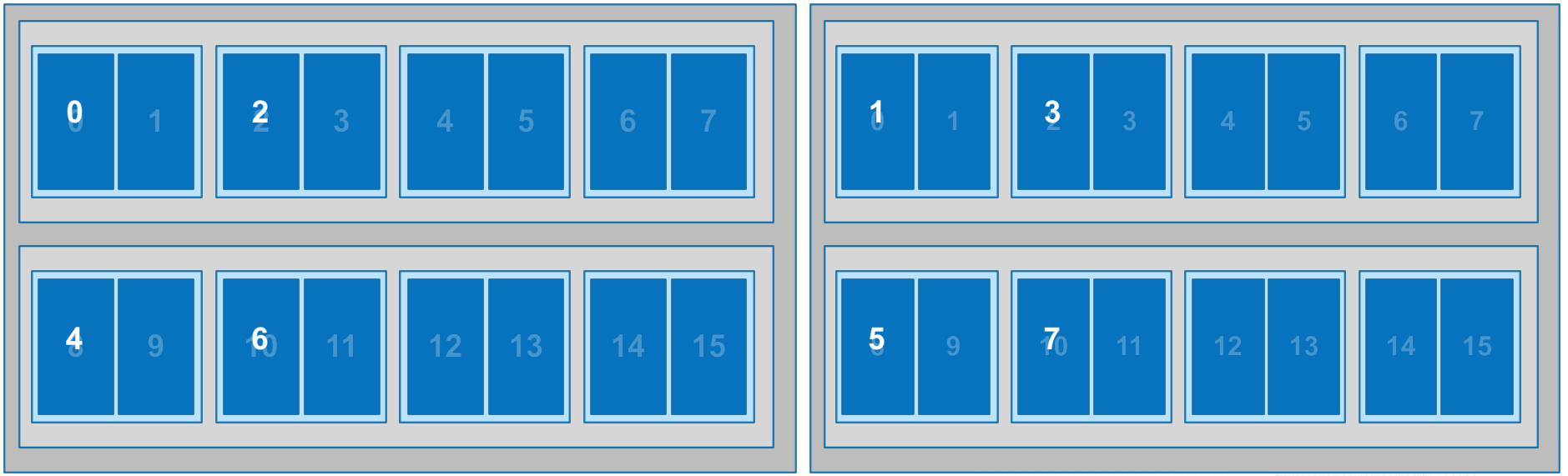
aprun –n 8 –S 2 –j 1 ./a.out

8 MPI ranks

2 ranks per NUMA

1 rank per compute unit

Setting MPICH_RANK_REORDER_METHOD = 0 will change layout to round robin



8 threads per MPI task

aprun –n 2 –N 1 –S 1 –d 8 ./a.out

2 MPI ranks

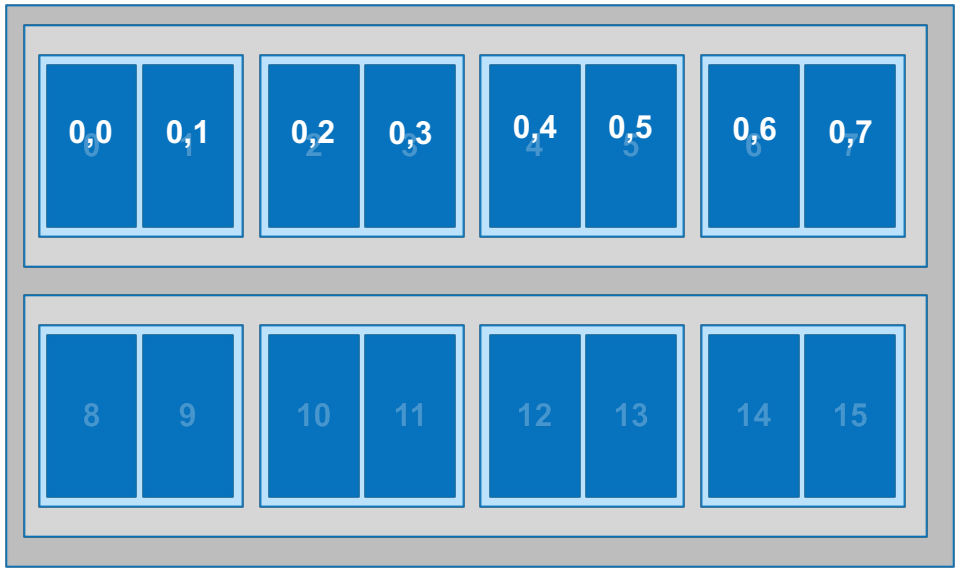
1 rank per node

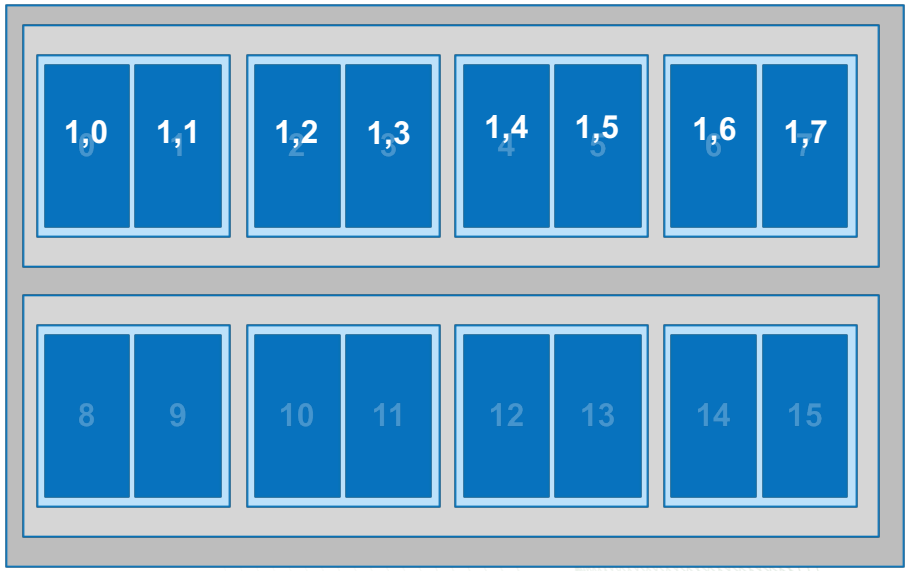
1 rank per NUMA

8 threads per rank

User should set OMP_NUM_THREADS = 8

Without –d, threads will be placed on same core.





Moving Forward

- Documentation
  - www.olcf.ornl.gov/for-users/system-user-guides/titan/running-jobs/
  - Man pages
    - aprun, qsub, showq, checkjob
- Help/Feedback
  - help@olcf.ornl.gov