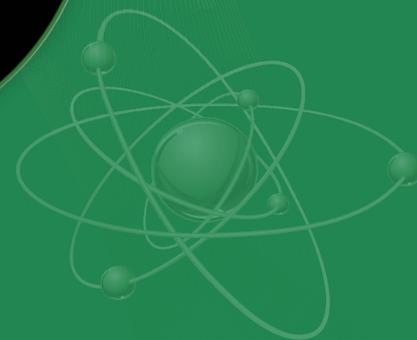
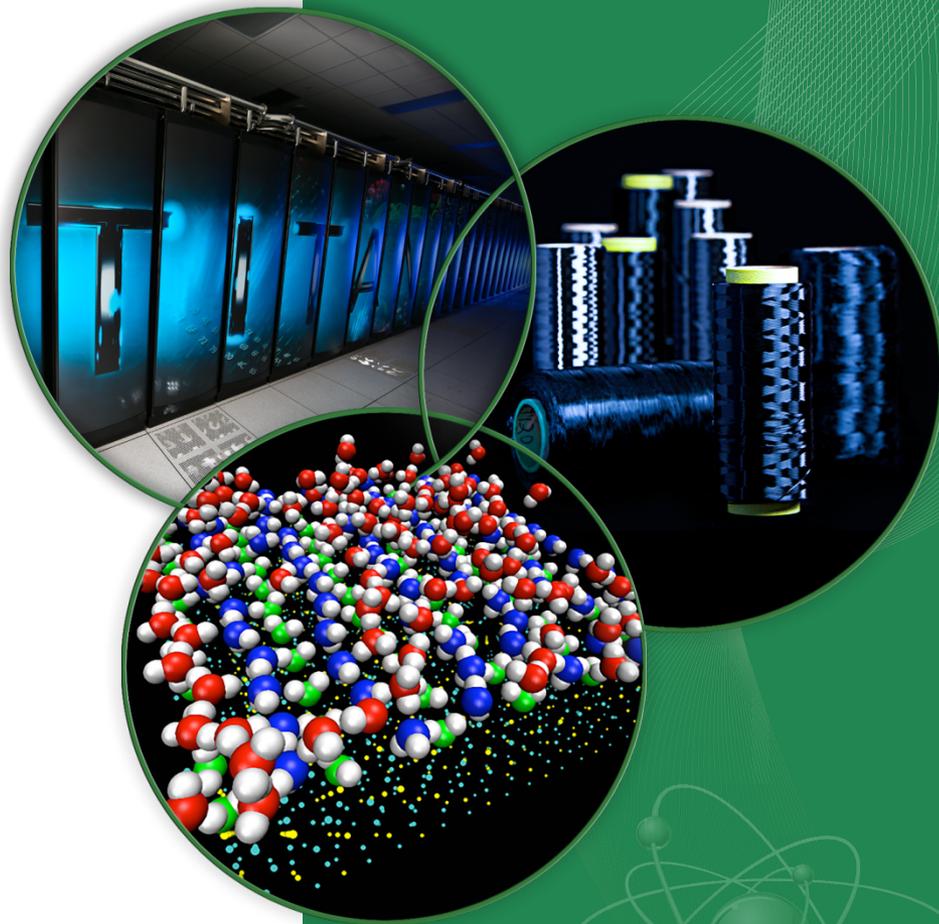


OLCF New User Tips

Bill Renaud
OLCF User Support



Introduction

- OLCF systems share many similarities to other HPC systems you may have used, but there are also some critical differences on the Cray systems
- This will focus on Titan and Eos
 - Most tips and operations are very similar
 - I'll note the differences
- Systems like Rhea are more traditional Linux clusters
 - They'll use the same filesystems as Titan/Eos
 - Compiling and running jobs will differ

Physical Cores vs. Core Hours (Titan)

- Each node has a 16-core processor and a GPU
 - Actually, the processor has 8 “Bulldozer” modules, each with 2 integer cores and a shared FPU
 - The processor contains 2 “NUMA Nodes” of 4 Bulldozer modules each
 - See the Titan User Guide for more information
- Jobs request a number of nodes with
`#PBS -l nodes=N`
- Jobs are charged for $30 * N$ core-hours
 - Why 30? 16 cores + 14 GPU Streaming Multiprocessors
 - You’re still limited to 16 MPI tasks per node (1 per core)

Physical Cores vs. Core Hours (Eos)

- Each node has 2 8-core processors
 - Without Hyper-Threading, you can run 16 tasks per node
 - With Hyper-Threading, you can run 32 tasks per node
- Jobs request a number of nodes with
`#PBS -l nodes=N`
- Jobs are charged for $30 * N$ core-hours

Launching Parallel Jobs (aprun)

- Largely the same for Titan and Eos
 - The `-j` option differs...see next slide
- Basic command
`aprun [options] executable`
- Numerous options...we'll discuss the common ones
 - For full description of all of them, see `man aprun`
- Must be done from a directory visible to compute nodes (e.g. Lustre)
 - An error similar to the following means you didn't do that:
`aprun: [NID 94]Exec a.out failed: chdir /autofs/na1_home/user1 No such file or directory`

Launching Parallel Jobs (aprun)

Option	Description
-n	Total number of MPI tasks
-N	Number of MPI tasks per physical node (≤ 16)
-S	Number of MPI tasks per NUMA node (≤ 8)
-d	Reserve this number of cores for each MPI task and its threads
-j	(Titan) Number of tasks per Bulldozer module (1 or 2) (Eos) Controls Hyper-Threading (-j0 and -j1 disable it, -j2 enables it)

- Floating-point intensive codes on Titan may perform better by “idling” half of the integer cores with `-j 1`
- Threaded codes use `OMP_NUM_THREADS` (or calls within the code) to **set** the number of threads and `-d` to **reserve** enough cores for the threads
- You must determine the total number of cores your aprun will need to determine the `#PBS -lnodes=X` request

Launching Parallel Jobs (aprun)

- A job must request enough nodes to “cover” its aprun request
 - If not, you’ll see an error similar to
`request exceeds max nodes alloc`
- For example, this will fail:

```
#PBS -l nodes=100  
aprun -n 800 -S 2 ./a.out
```
- Why?
 - 100 nodes should support up to 1600 tasks, right?
 - Yes; however, the aprun request actually needs 200 nodes
 - 800 tasks @ 2 per NUMA node = 400 NUMA nodes
 - 400 NUMA nodes @ 2 NUMA nodes per physical node = 200 nodes

Compiling (Titan and Eos)

- You're actually cross-compiling (different processors/OS on login nodes vs. compute nodes)
 - This can complicate utilities like `configure`
- Just use `cc` (C), `cc` (C++), or `ftn` (Fortran) to compile
 - The “back-end” compiler (PGI, Cray, GNU, Intel) is determined by the modules you have loaded
 - MPI/Math/Scientific libraries are automatically included
- Call the actual compiler (`ifort`, `pgif90`, `gcc`) to build for the login node
 - Actually, see §7.2 of the appropriate User Guide

Data Storage

- Multiple storage areas exist at OLCF
 - NFS (“home” areas)
 - Lustre (“work” areas, also called “scratch” areas)
 - HPSS (“archive” areas)
- We have both user-centric and project-centric storage locations
 - See the system user guides and data user guide (links on next slide) for more information

Data Storage

- It's important to be familiar with our Data Management Policy
https://www.olcf.ornl.gov/kb_articles/data-management-policy
- Different storage areas have different quotas, purge policies, and (post-project) data retention thresholds

Data Storage

Area	Path	Type	Permissions	Quota	Backups	Purged	Retention
User Home	\$HOME	NFS	User controlled	10GB	Yes	No	90 days
User Archive	/home/\$USER	HPSS	User controlled	2TB	No	No	90 days
Project Home	/ccs/proj/[projid]	NFS	770	50GB	Yes	No	90 days
Project Archive	/proj/[projid]	HPSS	770	100TB	No	No	90 days
Member Work	\$MEMBERWORK/[projid]	Lustre®	700	10TB	No	14 days	N/A
Project Work	\$PROJWORK/[projid]	Lustre®	770	100TB	No	90 days	N/A
World Work	\$WORLDWORK/[projid]	Lustre®	775	10TB	No	14 days	N/A

Data Storage

- We also provide tips for using Lustre® filesystems https://www.olcf.ornl.gov/kb_articles/spider-best-practices/
 - Edit/Build code in User/Project Home areas whenever possible
 - Use `ls -l` only where absolutely necessary
 - Open files as read-only whenever possible
 - Read small, shared files from a single task
 - Limit the number of files in a single directory
 - Place small files on a single OST
 - Place directories containing many small files on a single OST
 - `stat` files from a single task
 - Consider available I/O middleware libraries
 - Use large and stripe-aligned I/O whenever possible

For More Information

- See the more complete “Best Practices” presentation at:
<https://www.olcf.ornl.gov/wp-content/uploads/2014/05/Best-Practices-OLCF-and-more-Bill-Renaud.pdf>
- User Guides
<https://www.olcf.ornl.gov/support/system-user-guides/>
 - System User Guides for Titan, Eos, and Rhea
 - Accelerated Computing User Guide
 - Data Management User Guide