Some of the links listed below may require login on the ORNL network via VPN

ORNL Institutional Clusters User Guide

This is the user guide for the ORNL Institutional Cluster a.k.a. OIC, managed by Information Technology Services Division (ITSD). This guide is not meant to be an exhaustive guide on computer use or parallel programming in general, but instead focuses on information that is specific to the OIC.

Contents:

OIC Announcements and News

News, announcements and other information concerning the OIC is posted in the OICdate blog.

Outages and critical announcements are shared via the email list <oic-users@ornl.gov>. When you become a user of the OIC, you are automatically subscribed to this list. Instructions on using the list and unsubscribing will be sent to you. It is a low-volume list.

Ganglia is a monitoring tool that you can view to see "health" information across the entire OIC.

Back to the Contents

Support

If you ever need help with the cluster, please email helpline@ornl.gov or call helpline at 241-ORNL. All calls are automatically assigned to Jennifer Tippens, but if she is out, her backup, Jim Trater, will be able to assist you. If you email Jennifer directly and she is out of the office, the call will not be answered until Jennifer returns to the office. If you have an urgent problem on a night or weekend, let helpline know that it is a critical problem and they will know how to get in touch with Jen, Jim or another backup to help you from home.

Hardware Failure. When a node breaks and cannot be used for a period of time, it will be unmapped from the cluster. Other nodes will *not* be renamed. We have one “extra” node per VXRACK, but no “extra” nodes on the Altix Clusters. If a hardware failure occurs on a node that the user's job is running on, the user will need to resubmit the job. In other words, it will not automatically migrate to another node.

Back to the Contents

Accountability

Users are accountable for their actions and may be held accountable to applicable administrative or legal sanctions.

Security. Users must notify the Helpline (helpline@ornl.gov or 241-ORNL) immediately if they become aware that any of the accounts used to access OIC have been compromised.

Misuse/abuse. OIC personnel and users are required to address, safeguard against and report misuse, abuse and criminal activities. Misuse of OIC resources can lead to temporary or permanent disabling of accounts, loss of DOE allocations, and administrative or legal actions.

Software Use. All software used on OIC computers must be appropriately acquired and used according to the appropriate licensing. Possession or use of illegally copied software is prohibited. Likewise, users shall not copy copyrighted software, except as permitted by the owner of the copyright.

In general, the use of export controlled codes is prohibited, but special circumstances are considered and accommodated on a case by case basis. The use of an export controlled code must be approved and a security plan in place before the administrator can upload it to the export controlled area of the OIC. Export controlled code is not backed up. No export controlled code can be stored for any length of time in a user's home directory. All Principal Investigators using OIC resources and OIC staff members are responsible for knowing whether their project generates any of these prohibited data types or information that falls under Export Control. OIC users are required to restrict Export Controlled Information from foreign nationals of DOE sensitive countries. Contact the Helpline (helpline@ornl.gov or 241-ORNL), to set up a meeting to discuss your export controlled project.

Data Use. The OIC computer systems are operated as research systems and can only contain data related to scientific research. The OIC cannot store personally identifiable information (PII - data that falls under the Privacy Act of 1974 5U.S.C. 552a) for any length of time, however short. Use of OIC resources to store, manipulate, or remotely access any sensitive or national security information is prohibited. This includes, but is not limited to classified information, unclassified controlled nuclear information (UCNI), naval nuclear propulsion information (NNPI), the design or development of nuclear, biological, or chemical weapons or any weapons of mass destruction. The use of OIC resources for personal or non-work-related activities is also prohibited.

Back to the Contents

Architecture

Currently there are three kinds of nodes on the OIC. Access to each kind of node is restricted to users who belong to departments who have purchased these nodes. All nodes are running RedHat WS 4.

Original OIC. These blocks were acquired and brought online in 2006. They consist of a bladed architecture from Ciara Technologies (http://www.ciara-tech.com) called VXRACK. Each VXRACK contains two login nodes, three storage nodes and 80 compute nodes. Each compute node has Dual Intel 3.4GHz Xeon EM64T processors, 4GB of memory and dual Gigabit Ethernet Interconnects. Each VXRACK and its associated login and storage nodes are called a block. There are a total of nine blocks of this type.

- Two blocks are maintained by CCS and are separate from the OIC infrastructure.
- One block is separately reserved for use only by SNS Fermi and Accelerator groups. Half of this block has an additional Infiniband interconnect.
- Half a block is reserved for ITMS testing and development.
- Half a block is allocated to ORNL general users.
- Four blocks are allocated to CCSD.
- One block is allocated to CNMS.

OIC - Phase 2. These blocks were acquired and brought online in 2008. They are SGI Altix

machines. There are two types of blocks in this family:

Thin Nodes. Each Altix contains one login node, one storage node and 28 compute nodes within 14 chassis. Each node has eight cores. There are 16GB of memory per node. The login and storage nodes are XE240 boxes from SGI. The compute nodes are XE310 boxes from SGI. Browse the specs of the XE310s here (middle row): http://www.sgi.com/products/servers/altix/xe/configs.html.

There are a total of four blocks of this type.

- One block is allocated to MST.
- One block is allocated to NSSD.
- One block is allocated to CNMS.
- One block is allocated to ESD.

Fat Nodes. Each Altix contains one login node, one storage node and 20 compute nodes within 20 separate chassis. 3GHz. Each node has 8 cores. There are 16GB of memory per node. These nodes contain larger node-local scratch space and a much higher I/O to this scratch space because the scratch space is a volume from 4 disks. These are XE240 nodes from SGI. There are a total of two blocks of this type.

- One block is allocated to CHEM.
- One block is allocated to CNMS.

Back to the Contents

Requesting Accounts

Request an OIC account through UCAMS: https://ucams.ornl.gov. Log into UCAMS, click on “Main Menu”, click on “Access Management”, click the radio button that correlates to “Request” and “Resource by Owner” and hit “OK”. Scroll down to “ORNL Institutional Clusters”, highlight it and press “Select”. Choose the kind of access you need in the function section --this is your department or sponsor and you may select more than one function if you belong to more than one group. If you don't recognize the departments, or don't belong to any of the departments, you may request “ORNL General User”. Press next. The rest of the pages should be self-explanatory, but if you need help, please contact helpline at helpline@ornl.gov or 241-ORNL. Your account will not be active until it is approved by the assigned UCAMS approver.

Who are my UCAMS Approvers?

Admin User (very few people will require this access):

Suzanne Willoughby

Benchmark User (very few people will require this access):

Suzanne Willoughby

CCSD User:

Buddy Bland

David Hetrick

Jeff Nichols

Richard Reid

Becky Verastegui

CHEM User:

Phill Britt

CNMS User:

Peter Cummings

Bobby Sumpter, sumpterbg@ornl.gov

ESD User:

Mac Post

ESMG User:

Mac Post

ESSG User:

Mac Post

General ORNL User:

Suzanne Willoughby

MST User:

Malcolm Stocks

NSSD User:

Stephen Miller

SNS Accelerator:

Andrei Shishlo

SNS Neutronics:

Erik Iverson

Back to the Contents

Logging in

To login on OIC, you must be on the ORNL internal network. If you are off campus or behind a different firewall, you will either need to use the VPN, or ssh into login1.ornl.gov before sshing into any of the OIC machines.

More information on the VPN: http://home.ornl.gov/general/vpn/index.html

Only ssh is allowed. To copy data to and from OIC, use scp. Use your UCAMS user name and password. If you have never logged into the OIC before, you will see ssh generating keys for you the first time you log in. This will only happen once.

Login Nodes.

CCSD User:

ccsd[1-8].oic.ornl.gov

ccsd8.oic.ornl.gov must be used for interactive jobs.

CHEM User:

bes[0-5].ornl.gov -- Recommended. New Technology

bes[6-13].ornl.gov – Not recommended, but you can submit from these if it is more convenient (like if you are logged in from using another part of the cluster – these alias to the CCSD machines). These are OLDER technology and do not have Infiniband.

bes-inter.ornl.gov – This is the node from which you can submit interactive jobs. It is also older technology, but PBS requires that interactive jobs be submitted from this node.

CNMS User:

cnms.oic.ornl.gov

ccsd[1-8].oic.ornl.gov

ccsd8.oic.ornl.gov must be used for interactive jobs.

ESD, ESSG and ESMG Users:

bes[0-5].ornl.gov -- Recommended. New Technology

bes[6-13].ornl.gov – Not recommended, but you can submit from these if it is more convenient (like if you are logged in from using another part of the cluster – these alias to the CCSD machines). These are OLDER technology and do not have Infiniband.

bes-inter.ornl.gov – This is the node from which you can submit interactive jobs. It is also

older technology, but PBS requires that interactive jobs be submitted from this node.

General ORNL User:

ccsd[1-8].oic.ornl.gov

ccsd8.oic.ornl.gov must be used for interactive jobs.

MST User:

bes[0-5].ornl.gov -- Recommended. New Technology

bes[6-13].ornl.gov – Not recommended, but you can submit from these if it is more convenient (like if you are logged in from using another part of the cluster – these alias to the CCSD machines). These are OLDER technology and do not have Infiniband.

bes-inter.ornl.gov – This is the node from which you can submit interactive jobs. It is also older technology, but PBS requires that interactive jobs be submitted from this node.

NSSD User:

bes[0-5].ornl.gov -- Recommended. New Technology

bes[6-13].ornl.gov – Not recommended, but you can submit from these if it is more convenient (like if you are logged in from using another part of the cluster – these alias to the CCSD machines). These are OLDER technology and do not have Infiniband.

bes-inter.ornl.gov – This is the node from which you can submit interactive jobs. It is also older technology, but PBS requires that interactive jobs be submitted from this node.

SNS Accelerator:

acclr1.oic.ornl.gov - must be used for interactive jobs.

acclr2.oic.ornl.gov

SNS Neutronics:

fermi1.oic.ornl.gov

fermi2.oic.ornl.gov - must be used for interactive jobs.

Back to the Contents

Shell and Environment

The OIC runs RedHat WS 4.

Shell. Your login shell is set to bin/bash by default. Your login shell can be easily changed by using the command ldapchsh. Caution: If you change your shell some software may not work the same way and some automated tasks may not be performed. There is limited support for users who change their shells from the default.

Path and Modules. The cluster is set up with several modules for your convenience to help simplify your dot files. The convenience of the modules approach is that the user is not required to explicitly specify paths for different software versions. Modules keep the path, manpath and related environment variables coordinated. With the modules approach, you simply ``load'' and ``unload'' modules to control your environment and path. Type module avail to see which modules are available. If you would like to request a new module, drop a note to helpline@ornl.gov. To load a module, type module load [module-name]. Modules have to be loaded for each session to take effect for that shell, so if you will be needing a module to be available to a parallel job, append module load [module-name]to the end of your .bashrc (or shell equivalent dotfile). More information on modules and what one can do with modules (and how) can be found here: http://modules.sourceforge.net/

MPI Switcher. A slightly different approach is taken with the MPI modules. Since only one MPI can be loaded at a time and users generally wish to use the same MPI implementation for each session, a program called switcher.

- You can list all available MPI implementations by switcher mpi –list
- You can show the MPI implementation that you are currently set up for by switcher mpi --show
- If you want to switch to another MPI implementation, for instance, to use the Intel compiler suite with mpich, say switcher mpi = mpich-ch_gm-icc-1.2.5.9

Back to the Contents

Compiling

When you have selected the mpi implementation and the compiler version (using module) you wish to use, you can use the scripts, mpicc, mpiCC, mpif77 or mpif90 to compile your code. To see which compilers, libraries and versions are available, use module list.

Other libraries without modules may be installed in /opt. Check in /opt to see if you can find what you are looking for. If you can't find a particular library, please contact helpline@ornl.gov. Compilers we provide are:

Portland Group Compilers (PGI)

Intel Compilers and Math Kernel Libraries

GNU Compilers – both version 3.x and 4.x

Back to the Contents

Editors

Emacs and XEmacs is available. What more could you possibly ask for? Ok. vi is available also.

Back to the Contents

Revision Control

The OIC currently offers the following revision control software:

Concurrent Versions System (CVS)

Subversion (svn)

Revision Control System (RCS)

Back to the Contents

Home Directories

Your home directory resides in a shared file system on a storage node. It is backed up on a schedule to HPSS. There may be several users sharing the same file system. Your home directory is shared out to all the compute nodes during your job run via Network File System protocol (NFS). Unfortunately, NFS is not a parallel file system and it definitely has limitations. NFS is a protocol originally developed as a file system to allow a computer to access files over a network as easily as if they were on its local disks. The problem with using NFS in the cluster is that if a program accesses or writes a lot of information to disk then this causes a large amount of network traffic (I/O bottleneck) and the program execution for all jobs on the cluster can take much longer than expected or it can crash the storage server that houses your home directory and the home directories of your colleagues. For this reason, your home directory should NOT be used as scratch space for your running jobs.

NOTE: OIC is *not* a storage facility. Generated output data should be downloaded to private computer system or discarded as soon as possible. Please, remove all files that are no longer needed and free the disk space for other jobs. If you have your own HPSS account, **hsi** is available for accessing the High Performance Storage System (HPSS).

Back to the Contents

Scratch Space

Since home directories will cause an I/O bottleneck if used by programs as scratch space, the OIC provides node-local scratch space to running jobs. This scratch space is available only during your job run. It is created by the PBS batch system prior to the running of your job and it is destroyed and cleaned up by the PBS batch system directly after your job exits. It can be accessed from within your PBS script by using the variable $PBS_SCRATCH and used by your programs as scratch space. If you have results written to the scratch space on each node, you will have to loop through the nodes assigned to your job and copy the desired data off of the scratch area before your job exits. The nodes assigned to your job are listed in the variable $PBS_NODE_FILE.

It is required that you use the scratch space if you run NWChem. To do this, cd to $PBS_SCRATCH before calling NWCHEM in your PBS script. Call NWChem by specifying the entire path to the executable and the entire path to your .nw file. If any code is run using the home directory as scratch space instead of node-local scratch, it will be prematurely terminated without prior notice.

Back to the Contents

Job Scheduling/PBS/Batch Launches

Jobs cannot be run directly on the OIC compute nodes. They must be submitted to the job queuing system. The OIC's queuing system is managed by Torque and Maui. Torque is a drop in replacement for PBS. Maui is the job scheduler that tells Torque (hereafter known as PBS) when to run the jobs. Maui enforces a “Fair Share Algorithm” that uses historical usage statistics to decide priority for launching jobs. Queues are defined in PBS with further limitations to carve out shares of the cluster to conform to department contributions. PBS ensures that your job will have available the resources you have requested for it.

The job submission process is as follows:

1. You submit your job and specify the required resources for it by either defining these in a PBS submit script (recommended) or right on the command line. Here's a help page on PBS submit commands.
2. You will receive a unique job identifier, such as 53678.b08l02
3. PBS queues your job.
4. Maui decides when to run your job. Maui defers to the queue definitions in PBS and also employs a “fair share algorithm” to assign a priority to your job. It compares your jobs priority to that of the other jobs in the queue. It reevaluates this priority every 30 seconds, and assigns a new priority – there is weight given to jobs as they sit and wait in the queue. It continues to compare your job with the other queued jobs in the queue until your jobs' priority rises above other job priorities and it sees that there are available resources. Maui also runs jobs to “backfill” the queue, i.e. small jobs of short duration will run if resources are available regardless of their priorities if Maui projects that those “backfill” jobs will complete before the other resources needed to run a higher priority job are free.
5. Output of your job is written to files – either default or user specified. You can specify a lot of options to control how and when your job writes its output. Look at this torque help page for available options.
6. The user or administrator can use commands to monitor the progress of jobs in the queue.
7. The job finishes and resources are released to be used by other jobs.

How to write a PBS submit script:

It is highly recommended that you write a submit script to tell PBS how to run your jobs. This submit script is just a shell script with PBS directives in it. You can do shell tasks from within the script and you can launch your job, redirect output, copy things from scratch space to your home directory, etc. PBS directives begin with “#PBS”. The lines that don't begin with “#PBS” are just run from the MOM node. There is integration between PBS and OpenMPI so that you don't have to specify the actual nodes to OpenMPI – it will automatically select the nodes assigned by PBS.

Here is an example PBS Submit script:

```
#!/bin/bash
#PBS -S /bin/bash
#PBS -m be
#PBS -M 2vt@ornl.gov
```

```
#PBS -N parallel-worlds-jen
#PBS -q nssd08q
#PBS -l nodes=1
#PBS -l walltime=00:20:30,mem=100mb

cd $PBS_SCRATCH
mpirun -v --mca mpi_leave_pinned 1 --mca mpool_base_use_mem_hooks 1 --mca bt1
openib,self -np 1 /home/2vt/parallel-worlds
```

Obviously, you will need to replace the queue name with your assigned queue. You may also need to change the mpirun line to suit the resources available to you and also to suit the version of mpi you chose to run. I'll define the queues below, based on your function:

CCSD User:

ccsdq – Ciarra VXRack queue. You can not use openib on this queue.

CHEM User:

chem08q - SGI Altix XE240 nodes with 4 local scratch disks that are striped to allow higher I/O.

CNMS User:

cnmsq - Ciarra VXRack queue. You can not use openib on this queue.
cnms08tq – SGI Altix XE310 nodes with one local scratch disk each.
cnms08fq – SGI Altix XE240 nodes with 4 local scratch disks that are striped to allow higher I/O. Recommended for NWChem runs.

ESD User:

esd08q - SGI Altix XE320 nodes with one local scratch disk each.

ESSG User:

essg08q - SGI Altix XE320 nodes with one local scratch disk each.

ESMG User:

esmg08q - SGI Altix XE320 nodes with one local scratch disk each.

General ORNL User:

ornlq – Ciarra VXRack queue. You can not use openib on this queue.

MST User:

mst08q - SGI Altix XE310 nodes with one local scratch disk each.

NSSD User:

nssdo8q - SGI Altix XE310 nodes with one local scratch disk each.

SNS Accelerator:

ibq - Ciarra VXRack queue. openib IS available on this queue.
lowibq - Ciarra VXRack queue. openib IS available on this queue. This is a lower priority queue than ibq

SNS Neutronics:

workq - Ciarra VXRack queue. You can not use openib on this queue.
lowibq - Ciarra VXRack queue. openib IS available on this queue. This is a lower priority queue than ibq

I recommend using OpenMPI for all jobs. It is required for using Infiniband, and optional for other queue types, but it is actively supported and allows both MPI1 and MPI2.

Back to the Contents

Running an interactive job.

Interactive jobs can only be launched from bes-inter.ornl.gov or ccsd8.ornl.gov. They will not run from any other login nodes. You may use your assigned batch queues to run interactive jobs.

Back to the Contents

Checking on your job in the Queues:

You can use most of the maui and pbs commands, such as:

xpbs, qstat, showq, diagnose, showbf, checkjob, etc. Some of these commands will return numbers that make little sense because they show the overall state for all queues, rather than just for your own queue. Using showq and comparing your priority to that of the other folks in your queue, and that of the other folks who share your physical resources will get you the most meaningful answer. Or you can call helpline if you would like someone to check on your job for you.

Back to the Contents

Finding Installed Software:

If there is a module for it, then we have it. List the modules with “module avail”.

There are a lot of packages installed in /opt.

Restricted programs and licensed programs are installed in /projects.

You can check for rpms with “rpm -qa”

Intel MKL is under /opt/intel/mkl.

This is a “living document”. Please feel free to suggest things you would like to see documented so that you can use the cluster more efficiently!

Back to the Contents