Gaea System Details

From GFDL

Gaea System Details

To view the current state of gaea: please reference **http://www.ncrc.gov/**. This site will display the current status of compute nodes, login nodes, ldtn's, rdtn's and the network file system. It also provides information regarding the last time a particular section was down.

Contents

Terminology

Moab - The workload manager, the scheduler for all new NOAA research and development systems

Partition - A section of gaea that has its own scheduler. It is a logical unit in Moab.

Class - A Moab term for queue

DTN - data transfer node

Torque PBS - The resource manager, Moab will rely on torque PBS on the cray system

CMRS - Another name for Gaea. You may see/hear this some place.

Current Hardware

- c1ms
  - 30912 Cores
- 4 Login Nodes (Gaea1-4)
  - RDTNS (remote data transfer nodes)
  - LDTNS (local data transfer nodes)
  - 3 Files Systems
  - Users can request the system in increments of 24
  - The system is expected to be upgraded in March 2012 to be bitwise reproducible to the second system
- c2ms
  - A second system c2ms, is to be installed by December 7, 2011, The second system is not expected to bit-wise reproduce with the first system

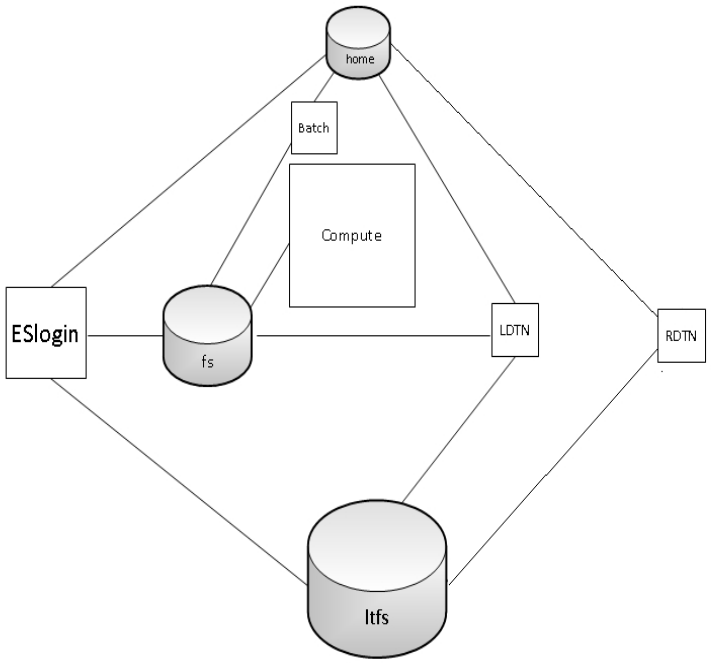
System Architecture

Node types:

Compute Nodes
: 24 cores, 48GB memory, run model executable, filesystem mounts - FS

Batch Nodes
: 2 cores, 8GB memory, run scripts only (cores are not charged, Note: Batch Nodes are not very powerful. Do not write code/jobs that will use Batch nodes to do CPU intensive work)

ESLogin Nodes
: 16 cores, 128GB memory, run interactive sessions, data jobs, Matlab



Partitions

c1ms - Gaea's current large compute partition. Future systems will have different queue nodes

es -login nodes, local data transfer nodes (ldtn) and remote data transfer nodes (rdtn)

t1ms - Testing and Development System (TDS) Partition is a partition that users usually will not use. As its name implies, it is used to test and develop new hard- and software. A few users may occasionally be asked to test models/codes/scripts on these test partitions. The TDS is currently split into two partitions. Currently one section is using the Gemini fabric for testing while the other is still using the C-Star fabric.

Examples-

```
msub -l partition=c1ms scriptname
#PBS -l partition=c1ms
```

Filesystems

Gaea has 3 filesystems. Home, FS, and LTFS

Home

The home filesystem is split into two sections which are backed up. There is a home1 and home2 for load balance purposes. Each user has a 5GB limit.

Home is mounted on:

- Batch nodes
- LDTN
- Login nodes
- RDTN

A nightly snapshot can be accessed at /ncrc/home1|2/.snapshot/nightly.0/$USER

You can use this path to restore any files or sub-directories that are contained within that directory from last night. Use nightly.1 for files from 2 nights ago. All files and sub-directories contained there will carry the same permissions as the originals. Users can simply copy from that location to any destination.

Restores from tape are performed on a best-effect basis, typically next business day.

Fast Scratch (FS)

The FS is a 1PB lustre filesystem. User allocations are available at /lustre/fs/scratch/$USER/. All files over 2 weeks old will be scrubbed within the /lustre/fs/scratch/$USER/ directories. This means files that have not been accessed or used in at least 2 weeks will be scrubbed. FS is NOT backed up. Users are responsible for monitoring their files and transferring what they do not want to lose to a location without a scrubbing policy.

FS is mounted:

- c1ms (batch and compute nodes)

- LDTN
- Login nodes

Long Term Scratch (LTFS)

The LTFS is a 3PB lustre filesystem. The LTFS scratch directories are not scrubbed. User allocations are available at /lustre/ltfs/scratch/$USER/. LTFS is NOT backed up. /lustre/ltfs/stage used for data transfers. The stage directory on the LTFS follows the same scrubbing policy as the FS scratch, files over 2 weeks old will be scrubbed.

LTFS is mounted on:

- LDTN
- Login nodes
- RDTN

Job Submission

Please see the following Moab details.

There are two job types.

- Batch
  - Regular jobs - use msub
- Interactive/Debug
  - msub -I -X -l partition=c1ms,size=48
    - Note: the size is the number of desired cores in increments of 24

Queues and Job States

There are currently 4 different queues.

- batch - no specification needed
- eslogin - compiles and data processing jobs
- ldtn - data movement queue (local)
- rdtn - data movement (remote)

Examples:

```
msub -q eslogin scriptname
msub -l partition=es -q ldtn scriptname
```

Queue polices are as follows:

- Persistent - jobs that run continuously Please see your group head for access to this queue.
- Urgent - heightened priority. Please see your group head for access to this queue.

- Novel - Jobs that require more than 25% of the system are held until a post preventive maintenance.
- Debug - 10% of c1ms during M-F business hours

Job Monitoring

The following are job monitoring commands with examples:

- showq - displays the queues. It is split into running jobs, eligible and blocked.

```
showq
showq -u $USER
showq -u $USER -c
```

- checkjob - provides job information. You can use up to 3 -v's for more info.

```
checkjob -v -v jobid
```

- mdiag - system state information

```
mdiag -j -v jobid
```

- mjobctl - cancel and control holds on jobs

```
mjobctl -h jobid
mjobctl -c jobid
```