Gaea

November 1, 2011

Tara McQueen

Overview

- Terminology
- Hardware
- System Architecture
- Nodes
- Partitions
- Queues
- Filesystems
- Jobs
- Data Transfers
- Modules
- Software
- Compilers
- Do’s & Don’ts
- Programming Environments
- Help

Terminology

- Moab – workload manager, scheduler for all new NOAA R&D systems
- Torque PBS – resource manager, Moab relies on Torque PBS
- Partition – section of Gaea that has its own scheduler. It is a logical unit in Moab.
- DTN – data transfer node

Gaea - current hardware

Cray XT6 LC

30,912 cores

2,576 Socket G34 AMD 2.1 GHz 12-core Magny-Cours processors

4 eslogin nodes

8 remote data transfer nodes (RDTN)

16 local data transfer nodes (LDTN)

Peak performance: 260 TF

14 cabinets in a 2x7 cabinet configuration

Filesystems

- home
- fast scratch
- long term scratch

Seastar interconnect

Gaea – future hardware

Cray XE6 LC (Separate System Partition c2)

78,336 additional compute cores (2,448 nodes)

32 cores / node

4,896 AMD Interlagos processors

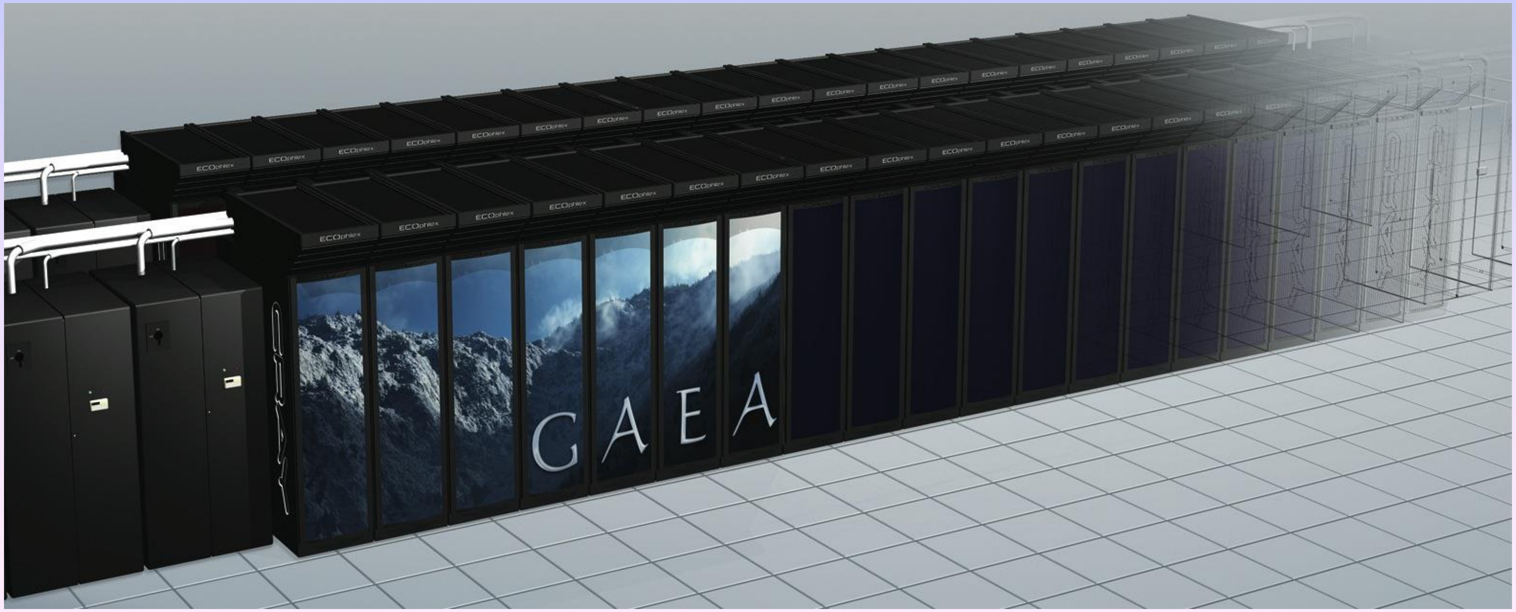
Gemini High Speed Interconnect Infrastructure

26 cabinets cabinets

4 additional eslogin nodes

c1ms will be upgraded to match c2 in Spring of 2012

* NOTE – c2 will not bit wise reproduce with c1ms



Node types

- Compute nodes = 24 cores 64GBs
  - Run model executable
- Batch nodes = 2 cores 8GBs
  - Runs scripts only
- Login nodes = 16 cores 128 GBs
  - Interactive, Matlab
- LDTN = 8 cores 24 GBs
  - Moves data from fs to ltfs, I/O intensive applications
- RDTN = 16 cores 48 Gbs
  - Moves data from ltfs to your center

Partitions & Queues

- c1ms – compute partition
  - batch (default)
- es – support partition
  - eslogin queue - compiles & data processing
  - ldtn queue - system data movement & I/O intensive applications
  - rdtn queue - remote data transfers
- Submission examples:
  - Command line = msub –l partition=c1ms *scriptname*
  - Script directives = #PBS –l partition=c1ms
  - Command line wins

Queue Policies

- Persistent – jobs that run continuously
- Urgent – heightened priority
  - These job priorities are allocated by the center and group administration.
- Novel – Jobs that require more than 25% of the system
  - Novel jobs are held until after a PM.
- Debug – 10% of the c1ms during business hours

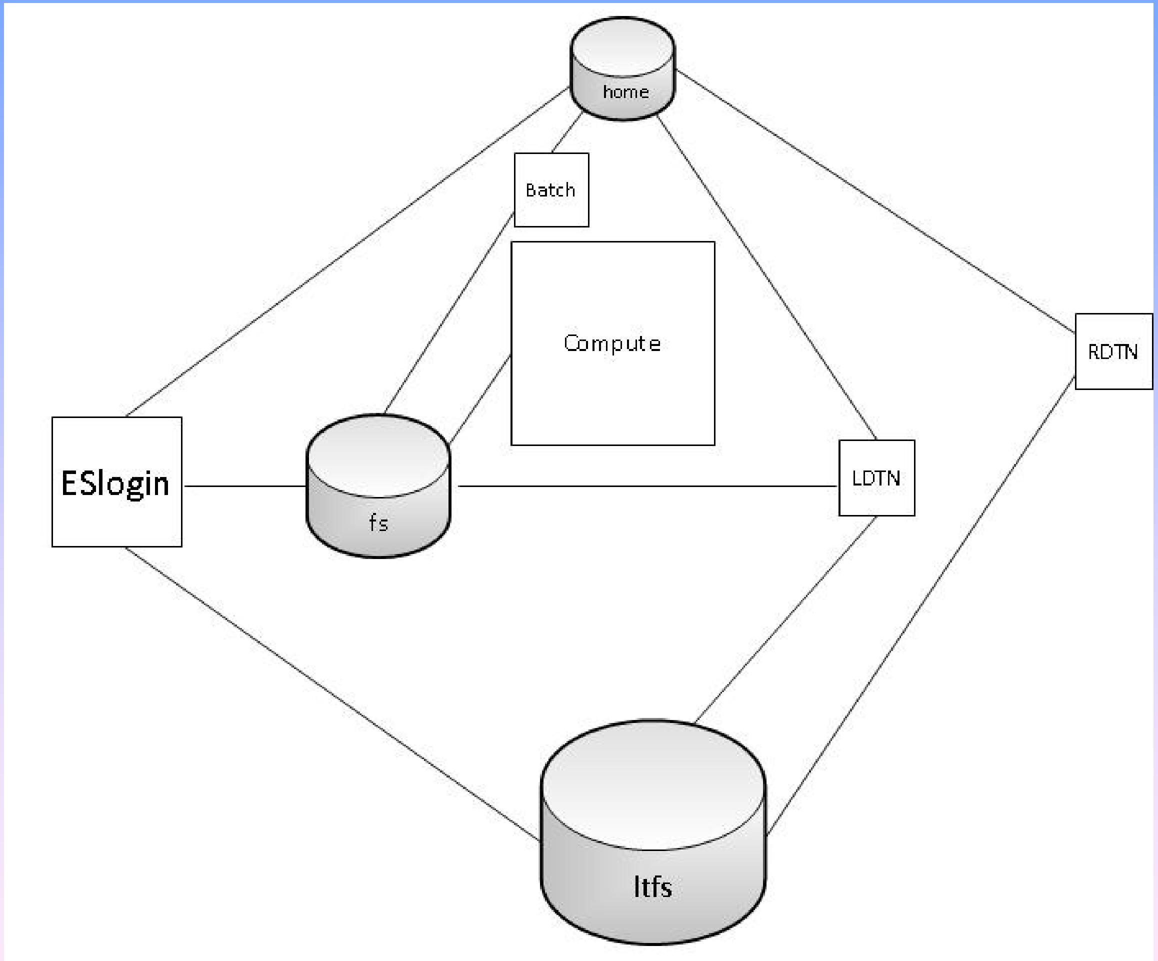
Filesystems

- Home – 5GB limit
  - separated into home1 and home2
- Fast Scratch(FS) – 1PB lustre filesystem
  - /lustre/fs/scratch/$USER
  - swept every 2 weeks
  - non swept pdata space will be available, please see your group head
- Long Term Scratch(LTFS) – 3PB lustre filesystem
  - /lustre/ltfs/scratch/$USER
    - Not swept, but monitored for usage, allocated by center and group
  - /lustre/ltfs/stage/
  - Stage is swept every two weeks
- Lustre filesystems are not backed up

What’s mounted where

- FS - /lustre/fs
  - login nodes, c1ms, ldtn
- LTFS - /lustre/ltfs
  - login nodes, ldtn, rdtn
- Home - /home1 | /home2
  - Login nodes, batch nodes, rdtn, ldtn

System Architecture



How do I login?

- ssh First.Last@gaea-rsa.rdhpcs.noaa.gov
- RSA token
- Certificates

Where I am?

- Run hostname

  -
    ```
    gaea1 4:20pm> hostname
    gaea1
    ```

- The system banner displays the system your on
- This is important for troubleshooting

Job Submission

- msub for command line submission
  - Options
    - -l partition=c1ms,size=48,walltime=10:00:00
    - -q ldtn
    - -I
    - -r
    - -v var=value
    - -V
    - -h
- Size is allocated on a node basis. You do not get less than a whole node.

Job Monitoring and Control

- showq
  - showq –u $USER
- checkjob
  - checkjob –v –v *jobid*
- mdiag
  - mdiag –j –v *jobid*
- mjobctl
  - mjobctl –h *jobid*

Job Types

- Batch jobs
  - Regular jobs – use msub
- Interactive/Debug jobs
  - Still use msub!
    - msub –I –X –l partition=c1ms,size=4000

Job States

- Running – the job is running
- Migrated – the job has been moved to the partition it will run on and is waiting to run
- Idle – the job is waiting to run
  - Jobs can be idle and blocked if you eligible job limit has been reached.
- BatchHold – the job has a system block
- UserHold – the job has a user hold
- SystemHold – the job has an administrative hold
- Deferred – an error prevented the job from running. The scheduler will reattempt the run. After multiple deferred attempts, a system hold is put in place on the job.

Job Limits

- There are currently no limits on running jobs
- There is a limit of 24 eligible jobs per user
  - Jobs exceeding the 24 job limit will be placed in the blocked section of the queue, they gain priority but will not become eligible until your eligible jobs are below 24.

Modules

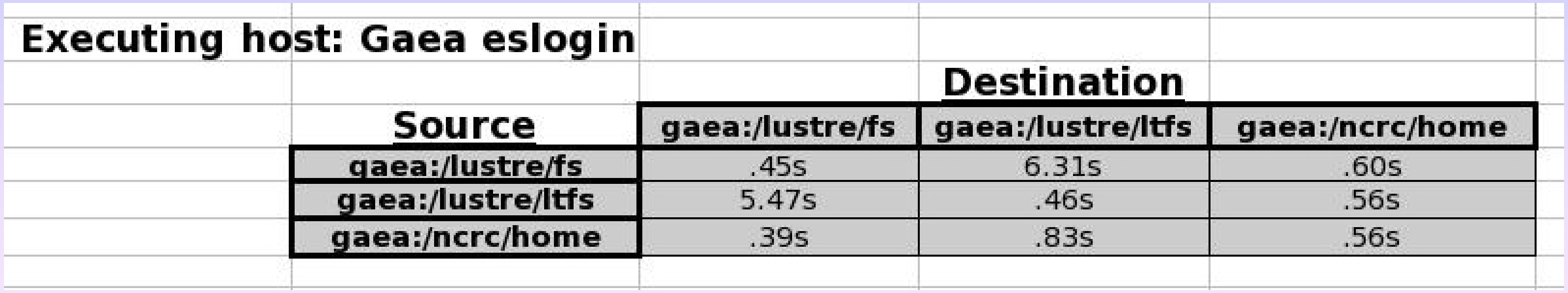
- module avail
- module load
  - module load gcp
- module list
- module unload
  - module unload gcp
- module swap
  - module swap gcp/1.4.4 gcp/1.5.0

Software

- Matlab
- Ferret
- R
- Nedit
- Tau
- Xdiff
- Meld
- Nco
- Ncview
- Module list for all applications

Data Transfers (gcp)

- gcp selects the appropriate transfer mechanism depending on source and destination
- module load gcp



Executing host: Gaea eslogin

	Destination		
Source	**gaea:/lustre/fs**	**gaea:/lustre/ltfs**	**gaea:/ncrc/home**
gaea:/lustre/fs	.45s	6.31s	.60s
gaea:/lustre/ltfs	5.47s	.46s	.56s
gaea:/ncrc/home	.39s	.83s	.56s

Data Transfers (gcp) cont.

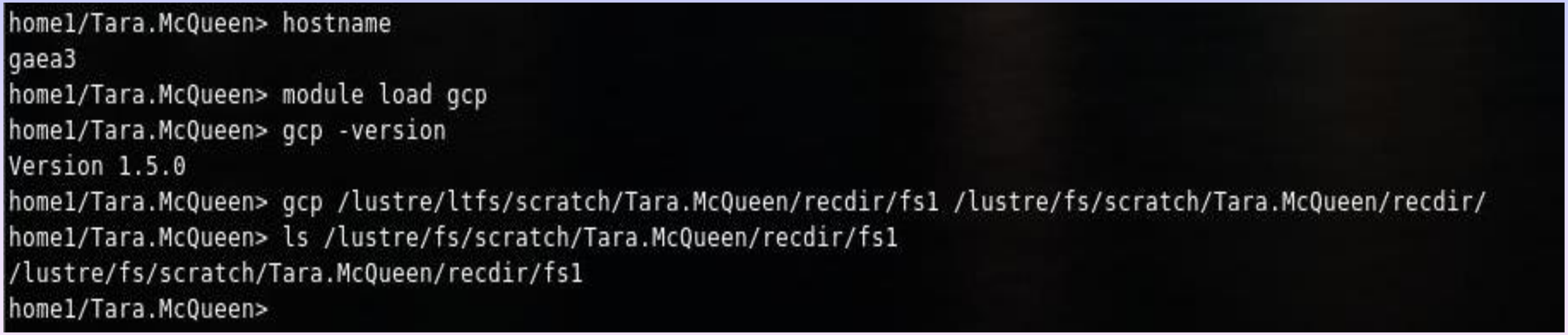
- Options
  - --help or –h
  - --version
  - --recursive or –r
  - --debug or –d
  - --verbose or –v
  - --create-dirs or –cd
  - --not-world

Data Transfers (gcp) cont.

local to local

gcp /path/to/source /path/to/destination

gcp /lustre/ltfs/scratch/$USER/file /lustre/fs/scratch/$USER/

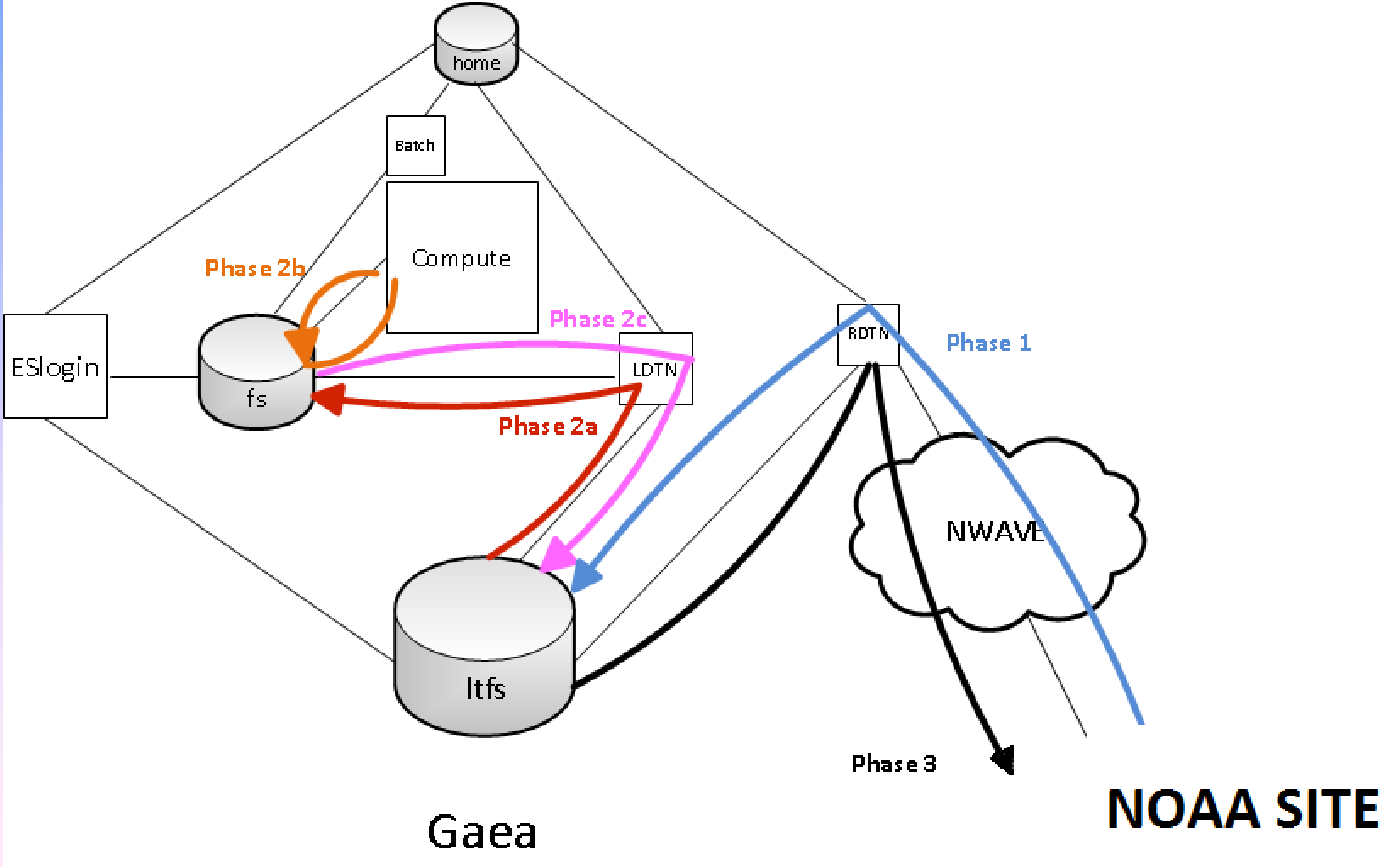


```
home1/Tara.McQueen> hostname
gaea3
home1/Tara.McQueen> module load gcp
home1/Tara.McQueen> gcp -version
Version 1.5.0
home1/Tara.McQueen> gcp /lustre/ltfs/scratch/Tara.McQueen/recdir/fs1 /lustre/fs/scratch/Tara.McQueen/recdir/
home1/Tara.McQueen> ls /lustre/fs/scratch/Tara.McQueen/recdir/fs1
/lustre/fs/scratch/Tara.McQueen/recdir/fs1
home1/Tara.McQueen>
```

Data Transfers (gcp) cont.

- Push/pull data
- Wildcards
- Single file, multiple file and directory
- Submits to ldtn when needed

Workflow Example



Compilers

- Standard Compiler:
  - Ftn/cc/CC
  - Need to run regardless of programming environment
- Also available:
  - Intel
  - pgi
  - Gnu
- OpenMP ported on all compilers
- To use a different compiler suite, change programming enviroment
  - PrgEnv-intel, PrgEnv-pgi, PrgEnv-gnu, etc.
- PrgEnv-pgi/3.1.29 is current default

Sample Compile / Link / Run

- Compile and Link:
  - ftn –O2 –g –r8 –i4 –o my_prog.exe my_prog.f90
- Launch an MPI executable
  - aprun –n <npes> my_prog.exe <my_prog_args>
- Launch openMP executable
  - aprun –n (mpi) –d omp_num_threads
- mpich2 is the mpi library

Debuggers

- Totalview & Allinea ddt
  - GUI debugger
    - Supports command line interface
    - Serial or parallel (MPI and OpenMP)
  - module load totalview
    - Must compile using –g and best to use –O0 (no optimization)
    - totalview aprun –a –n <npes> my_prog.exe
    - Totalviewtech.com
  - module load ddt
    - ddt –n (mpi) ./executable

Do's

- Put source files and commonly used files on ltfs
- Put transient data on fs
- Use gcp for transfers
- Compile on login nodes
- Use lfs (lustre) version of commands on lustre filesystems
- Copy data back to archive location (off gaea) using RDTN's

Don'ts

- Module purge
- Deep large scale use of “find” on lustre filesystems
- Recursive operations like ls -R
- cp
- Fs as permanent storage
- Transfers on batch
- Combines on batch will be killed
- Combines on compute nodes
- Run applications natively
- Compile on batch
- Unalias *
- Cron jobs

Help

- Kate Howard – NCEP's point of contact
  - Kate escalates to GFDL/Tara McQueen
  - Tara escalates to GFDL systems group or ORNL
- NCEP process escalation through Allan Darling
- Meetings to be aware of…
  - Integrated management team meeting – Monday afternoons
  - ORNL ticket review – Wednesday mornings
  - Gaea CM with ORNL – Thursday afternoons

Questions