QMCPACK Training 2016
Introduction to ALCF Systems

Anouar Benali (ANL)

Ye Luo (ANL)

ALCF resources

- **Mira** (Production) – IBM Blue Gene/Q
  - 49,152 nodes / 786,432 cores
  - 768 TB of memory
  - Peak flop rate: 10 PF
  - Linpack flop rate: 8.1 PF
- **Cetus** (Test & Devel.) – IBM Blue Gene/Q
  - 4,096 nodes / 65,536 cores
  - 64 TB of memory
  - 838 TF peak flop rate
- **Vesta** (Test & Devel.) – IBM Blue Gene/Q
  - 2,048 nodes / 32,768 cores
  - 32 TB of memory
  - 419 TF peak flop rate
- **Cooley** (Visualization) – Cray + NVIDIA
  - 126 nodes / 1512 x86 cores (Haswell)
  - 126 NVIDIA Tesla K80 GPUs
  - 47 TB x86 memory / 3 TB GPU memory
  - 293 TF peak flop rate

Detail shot of Mira

Mira and her cables

IBM Blue Gene/Q

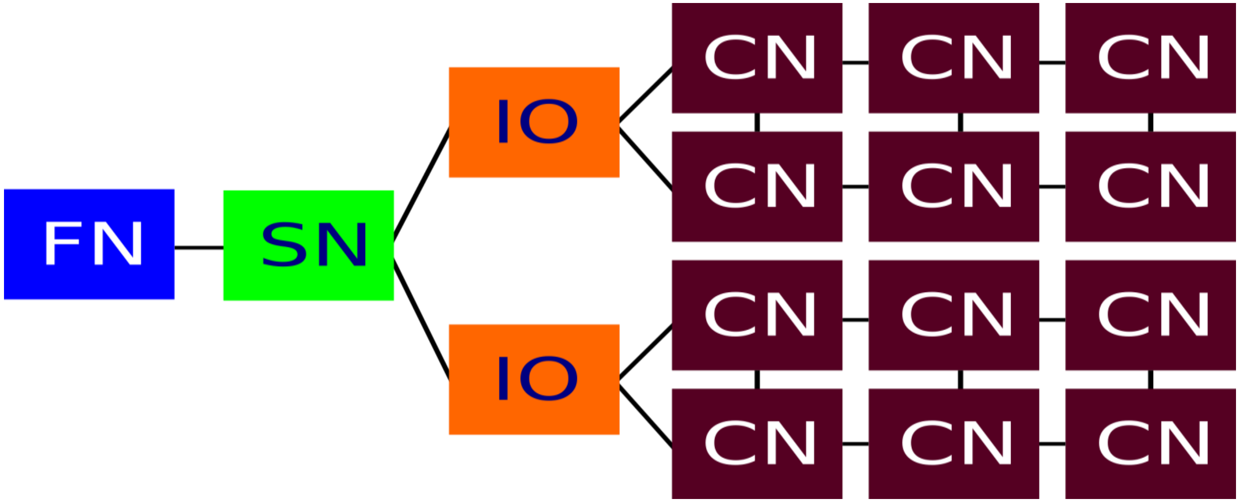
- **Storage**
  - Scratch: 27 PB usable capacity,
    330 GB/s bw (GPFS)
    aggregate over 2 file systems
  - Home: 1.1 PB usable capacity,
    45 GB/s bw (GPFS)

Blue Gene Features

- **Low speed, low power**
  - Embedded PowerPC core with custom SIMD floating point extensions
  - Low frequency: 1.6 GHz on Blue Gene/Q
- **Massive parallelism**
  - Many cores: 786,432 on Mira, 32,768 on Vesta
- **Fast communication network(s)**
  - 5D Torus network on Blue Gene/Q
- **Balance**
  - Processor, network, and memory speeds are well balanced
- **Minimal system overhead**
  - Simple lightweight OS (CNK) minimizes noise
- **Standard programming models**
  - Fortran, C, C++ & Python languages supported
  - Provides MPI, OpenMP, and Pthreads parallel programming models
- **System-on-a-Chip (SoC) & Custom designed ASIC (Application Specific Integrated Circuit)**
  - All node components on one chip, except for memory
  - Reduces system complexity and power, improves price / performance
- **High reliability**
  - Sophisticated RAS (Reliability, Availability, and Serviceability)
- **Dense packaging**
  - 1024 nodes per rack

Blue Gene/Q system components

- **Front-end nodes** – dedicated for user's to login, compile programs, submit jobs, query job status, debug applications. **RedHat Linux OS.**
- **Service nodes** – perform partitioning, monitoring, synchronization and other system management services. Users do not run on service nodes directly.
- **I/O nodes** – provide a number of Linux/Unix typical services, such as files, sockets, process launching, signals, debugging; run Linux.
- **Compute nodes** – run user applications, use simple **compute node kernel (CNK)** operating system, ships I/O-related system calls to I/O nodes.



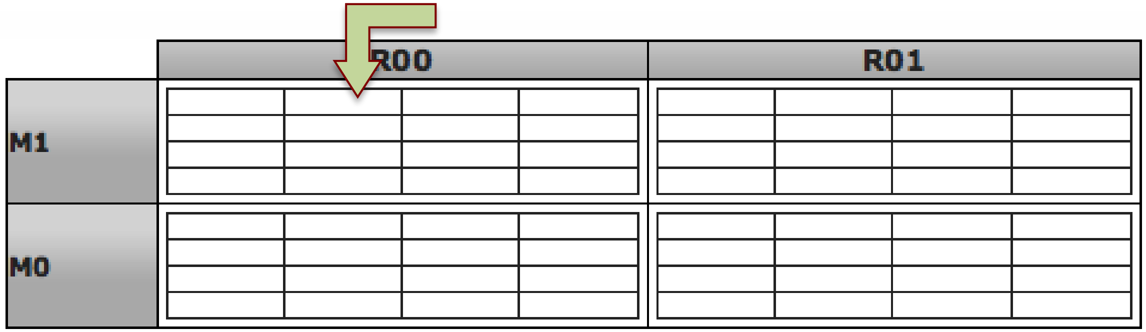
Partition dimensions on Blue Gene/Q systems

Vesta

32 nodes = minimum partition size on Vesta

Vesta

2 racks



Nodes
32
64
128
256
512
1024
2048(*)

http://status.alcf.anl.gov/mira/activity (beta, a.k.a. The Gronkulator)

(*) Partition not active.

SoftEnv

- A tool for managing a user's environment
  - Sets your PATH to access desired front-end tools
  - *Your compiler version can be changed here*
- Settings:
  - Maintained in the file ~/.soft (Mira/Cetus & Vesta) or ~/.soft.cooley (Cooley)
  - Add/remove keywords from ~/.soft or ~/.soft.cooley to change environment
  - ***Make sure @default is at the very end***
- Commands:
  - **softenv**
    - A list of all keywords defined on the systems
  - **resoft**
    - Reloads initial environment from ~/.soft or ~/.soft.cooley file

http://www.mcs.anl.gov/hs/software/systems/softenv/softenv-intro.html

QMCPACK, GAMESS and PWSCF on Vesta

- ALCF supports all software we will be using during the school. Binaries are locate

However, we will use “Nexus” during the training to manage job submission and binaries locations.

Section:

Considerations before you run

Accounts, projects, allocations, etc.

- ALCF Account
  - Login username
    - /home/username
  - Access to VESTA
  - CRYPTOCard token for authentication
    - PIN
    - Must call ALCF Help Desk to activate your token
- Project
  - Corresponds to allocation of core-hours on at least one machine
  - User can be member of one or more projects
    - /projects/ProjectName
- Logging in
  - ssh -Y username@vesta.alcf.anl.gov
    - Click button on CRYPTOCard
    - Password: PIN + CRYPTOCard display

Manage your account at

http://accounts.alcf.anl.gov

(password needed)

http://www.alcf.anl.gov/user-guides/accounts-access

HPC storage file systems at ALCF

Name	Accessible from	Type	Path	Backed Up to HPSS	*Daily Snapshots	Uses
vesta-home	Vesta	GPFS	/home or /gpfs/vesta-home	No	Yes	General use
projects	Vesta	GPFS	/projects	No	No	Intensive job output, large files

* Daily snapshots are stored for 1 week on-disk in /gpfs/{vesta,mira}-home/.snapshots/. These snapshots do NOT persist in the event of disk failure.

http://www.alcf.anl.gov/user-guides/bgq-file-systems

Backups and tape archiving

- **Backups**
  - On-disk snapshots of /home directories are done nightly
    - If you delete files accidentally, check:

      - /gpfs/mira-home/.snapshots on Mira

      - /gpfs/vesta-home/.snapshots on Vesta
  - **Only Mira/Cetus/Cooley home directories** are backed up to tape
    - The Vesta home directories are ***not*** backed up to tape (just daily snapshots)
    - Project directories are ***not*** backed up (/projects)

- **Manual data archiving to tape (HPSS)**
  - HSI is an interactive client
  - GridFTP access to HPSS is available
  - See http://www.alcf.anl.gov/user-guides/using-hpss

Hands-on!!

- Log into Vesta:

```
> ssh –X username@vesta.alcf.anl.gov
```

- Project:
  - Check that you are a member of the project used for this hands-on session:

```
> projects
    … QMCPACk-Training …
```

  - Check the allocation of core-hours available for this project:

```
> cbank allocations -p QMCPACk-Training
```

Hands-on session

- The reservation for today's event is: QMCPACK1
  - Check the name of the queue created for the hands-on session:

```
> showres
```

- Setting up your environment for the labs

```
> cp /projects/QMCPACk-Training/soft ~/.soft
> resoft
> mpixlc -qversion
```

Note: after editing your ~/.soft file, run command 'resoft' to refresh your environment.

Section:

Queuing and running

Vesta job scheduling

User Queue	Partition Sizes in Nodes	Wall-clock Time (hours)	Max. Running per User	Max. Queued Node-hours
default	32, 64, 128, 256, 512, 1024	0 - 2	5	1024
qmcpack	32, 64, 128, 256, 512, 1024	0 - 2	2	(10 jobs per user)

Remember!! We are 30 participants! Do not ask for 1024 nodes for yourself during the labs!

The queue “qmcpack” is active everyday from 12pm to 6pm. If you want to run the labs after hours, you will have to use the “default” queue

- **I/O to compute node ratio 1:32**

Cobalt resource manager and job scheduler

- Cobalt is the resource management software on all ALCF systems
  - Similar to PBS but not the same
- Job management commands:

  qsub: submit a job
  qstat: query a job status
  qdel: delete a job
  qalter: alter batched job parameters
  qmove: move job to different queue
  qhold: place queued (non-running) job on hold
  qrls: release hold on job
  qavail: list current backfill slots available for a particular partition size

- For reservations:

  showres: show current and future reservations
  userres: release reservation for other users

qsub Options

Syntax:

```
qsub [-d] [-v] -A <project name> -q <queue> --cwd <working directory>
        --env envvar1=value1:envvar2=value2 --kernel <kernel profile>
        -K <kernel options> -O <outputprefix> -t time <in minutes>
        -e <error file path> -o <output file path> -i <input file path>
        -n <number of nodes> -h --proccount <processor count>
        --mode <mode> -M <email> --dependencies <jobid1>:<jobid2> <command> <args>
```

- Standard options:

-A project	project to charge
-q queue	queue
-t <time_in_minutes>	required runtime
-n <number_of_nodes>	number of nodes
--proccount <number_of_cores>	number of CPUs
--mode <cX | script>	running mode
--env VAR1=1:VAR2=1	environment variables
<command> <args>	command with arguments
-O project <output_file_prefix>	prefix for output files (default **jobid**)
-M <email_address>	e-mail notification of job start, end
--dependencies <jobid1>:<jobid2>	set dependencies for job being submitted
-I or --interactive	run an interactive command

Further options and details may be found in the man pages (> man qsub) or at:

http://trac.mcs.anl.gov/projects/cobalt/wiki/CommandReference

Cobalt job control: basic method

- **Basic**: submit a BG/Q executable

  qsub -n *nodes* --proccount *P* --mode c*N* … *path/executable*

  - *N* is number of processes (MPI ranks) per node

  - Node has 16 cores
    --mode c1 — 1 rank/node
    --mode c2 — 2 rank/node
    …
    --mode c16 — 1 rank/core
    --mode c32 — 2 rank/core
    --mode c64 — 4 rank/core

  - Threads
    qsub --mode c1 --env OMP_NUM_THREADS=64
    qsub --mode c2 --env OMP_NUM_THREADS=32
    …
    qsub --mode c16 --env OMP_NUM_THREADS=4
    ….

Job dependencies

- Following job in sequence depends on completion of current job

Cobalt Job ID

Section:

After your job is submitted

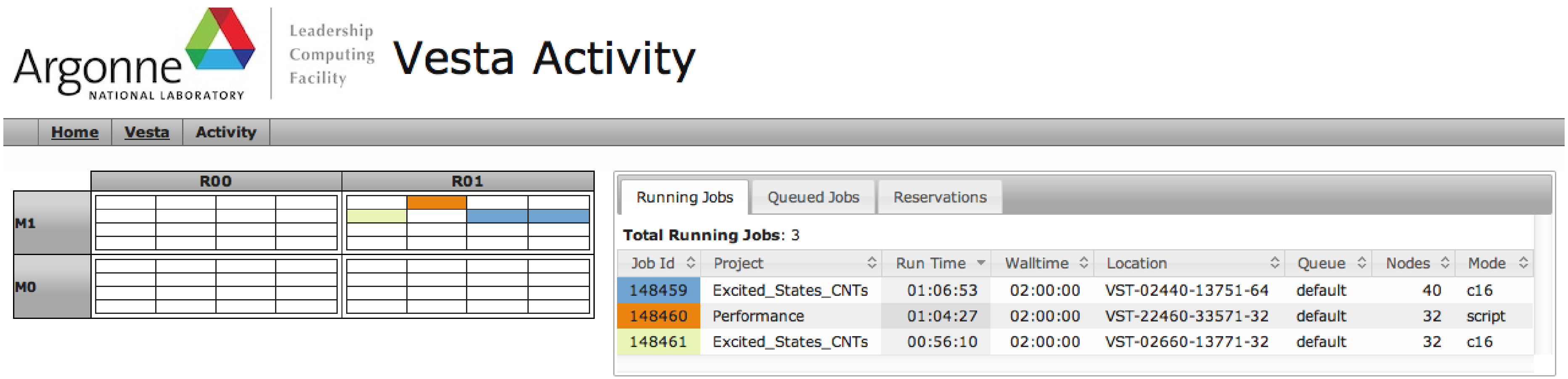
qstat: show status of a batch job(s)

- **qstat** # list all jobs

```
JobID   User    WallTime  Nodes  State    Location
=========================================================
301295  smith   00:10:00   16        queued  None
```

- About jobs
  - JobID is needed to kill the job or alter the job parameters
  - Common states: queued, running, user_hold, maxrun_hold, dep_hold, dep_fail
- **qstat -f <jobid>** # show more job details
- **qstat -fl <jobid>** # show all job details
- **qstat -u <username>** # show all jobs from <username>
- **qstat -Q**
  - Instead of jobs, this shows information about the queues
  - Will show all available queues and their limits
  - Includes special queues used to handle reservations

Machine status web page



Vesta Activity

Total Running Jobs: 3

Job Id	Project	Run Time	Walltime	Location	Queue	Nodes	Mode
148459	Excited_States_CNTs	01:06:53	02:00:00	VST-02440-13751-64	default	40	c16
148460	Performance	01:04:27	02:00:00	VST-22460-33571-32	default	32	script
148461	Excited_States_CNTs	00:56:10	02:00:00	VST-02660-13771-32	default	32	c16

http://status.alcf.anl.gov/vesta/activity

Cobalt files for a job

- Cobalt will create 3 files per job, the basename <prefix> defaults to the jobid, but can be set with “qsub -O myprefix”
  - jobid can be inserted into your string e.g. "-O myprefix_$jobid"

- **Cobalt log file**: **<prefix>.cobaltlog**
  - created by Cobalt when job is submitted, additional info written during the job
  - contains submission information from qsub command, runjob, and environment variables
- **Job stderr file**: **<prefix>.error**
  - created at the start of a job
  - contains job startup information and any content sent to standard error while the user program is running
- **Job stdout file**: **<prefix>.output**
  - contains any content sent to standard output by user program

qdel: kill a job

- **qdel <jobid1> <jobid2>**
  - delete the job from a queue
  - terminate a running job

qalter, qmove: alter parameters of a job

- Allows user to alter the parameters of queued jobs without resubmitting
  - Most parameters may only be changed before the run starts

- Usage: **qalter** [options] <jobid1> <jobid2> …

- Example:

  > **qalter** -t 60 123 124 125

  (changes wall time of jobs 123, 124 and 125 to 60 minutes)

- Type '**qalter -help**' to see full list of options

- qalter cannot change the queue; use **qmove** instead:

  > **qmove** <destination_queue> <jobid>

qhold, qrls: holding and releasing

- **qhold** - Hold a submitted job (will not run until released)

  qhold <jobid1> <jobid2>

- To submit directly into the hold state, use qsub –h
- **qrls** - Release a held job (in the *user_hold* state)

  qrls <jobid1> <jobid2>

- Jobs in the dep_hold state released by removing the dependency

  qrls --dependencies <jobid>

  or qalter –dependencies none <jobid>

- Jobs in the *admin_hold* state may only be released by a system administrator

Section:

Potential problems

When things go wrong… logging in

- Check to make sure it's not a scheduled system maintenance day:
  - Login nodes on Blue Gene/Q and data analytics systems are often closed off during system maintenance to allow for activities that would impact users.
  - Look for reminders in the weekly maintenance announcement to users and in the pre-login banner message.
  - An all-clear email will be sent out to users at the close of maintenance.

- Remember that CRYPTOCard passwords:
  - Require a pin at the start
  - Are all hexadecimal characters (0-9, A-F). Letters are all **UPPER CASE**.

- On failed login, try in this order:
  - Try typing PIN + password again (without generating new password)
  - Try a different ALCF host to rule out login node issues (e.g., maintenance)
  - Push CRYPTOCard button to generate a new password and try that
  - Walk through the unlock and resync steps at: http://www.alcf.anl.gov/user-guides/using-cryptocards#troubleshooting-your-cryptocard
  - Still can't login?
    - Connect with **ssh -vvv** and record the output, your IP address, hostname, and the time that you attempted to connect.
    - Send this information in your e-mail to support@alcf.anl.gov

When things go wrong… running

- Cobalt jobs, by default, produce three files (*.cobaltlog, *.error, *.output)

- Only *.cobaltlog is generated at submit time, the others at runtime

- Boot status (successful or not) written to *.cobaltlog

- After booting, the *.error file will have a non-zero size:
  - *Note: If your script job redirects the stderr of cobalt-mpirun, it will not end up in the job's .error file*

- If you think there is an issue, it's best to save all three files:
  - Raise your hand and someone will come to help you!

Questions?