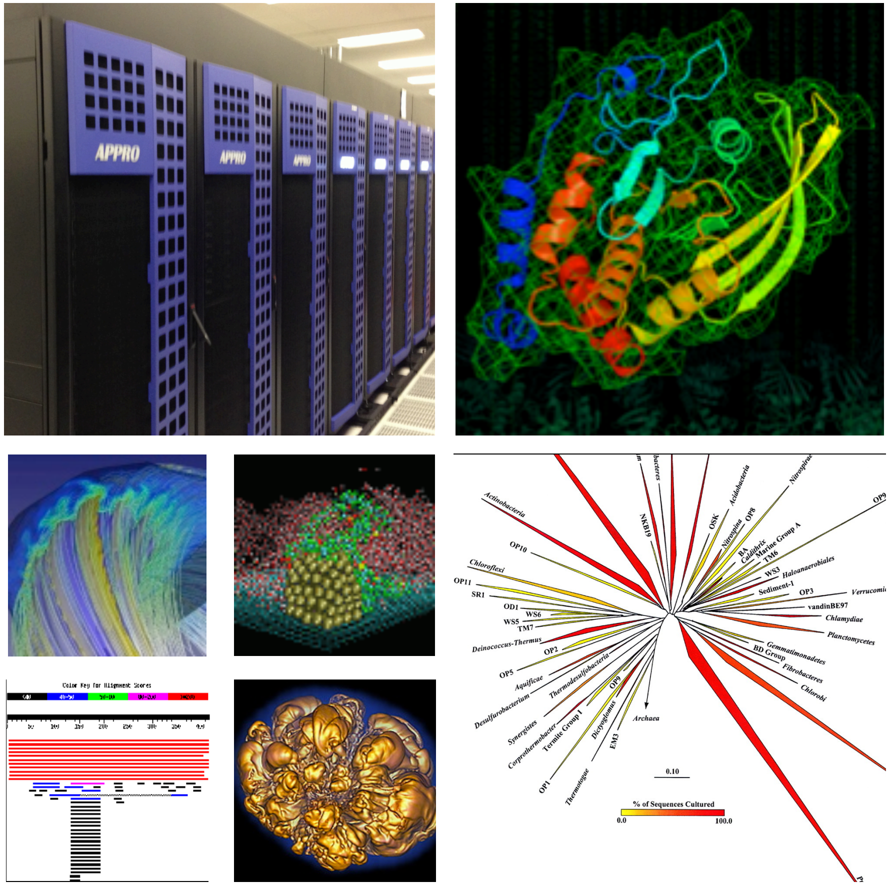
Running Jobs on Genepool



Douglas Jacobsen

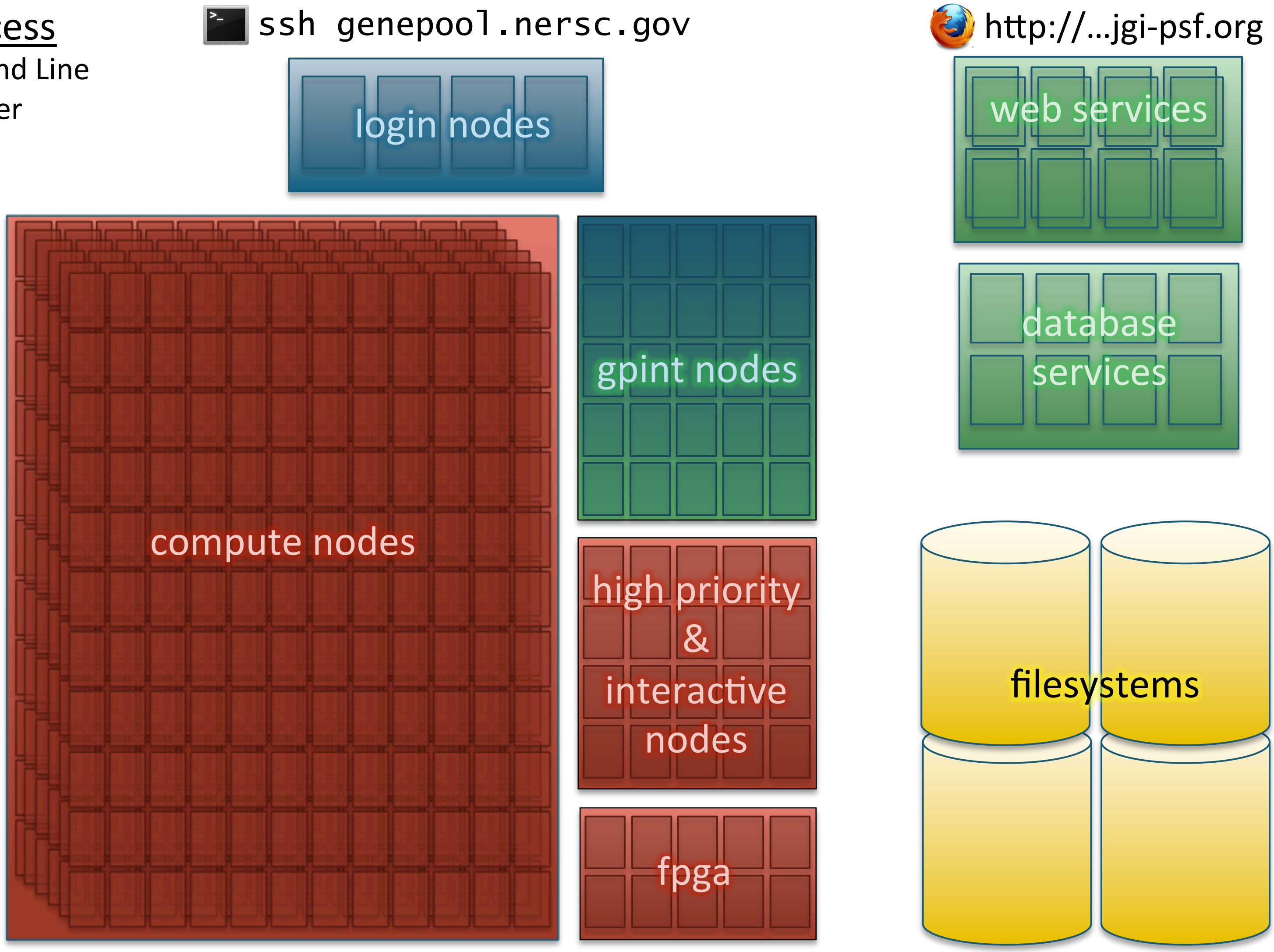
NERSC Bioinformatics Computing Consultant

February 12, 2013

Structure of the Genepool System

User Access

- 🟦 Command Line
- 🟥 Scheduler
- 🟩 Service



Types of Jobs on genepool

- **Batch – Scheduled** **(compute nodes, fpga)**
  - 8,320 cores for 72,953,280 compute hours per year in genepool
  - use “qsub” to submit a job
- **Interactive – Scheduled** **(compute nodes subset)**
  - 80 cores presently, increasing size
  - use “qlogin” to submit a job
- **Interactive – Unscheduled** **(login nodes, gpints)**
  - 4 login nodes, 27 gpint nodes
  - ssh to the host, direct-use
- **Services – Unscheduled** **(login nodes, gpints, gpweb, gpdb, gpodb)**
  - Web services
  - Database services
  - Automated job submission / control

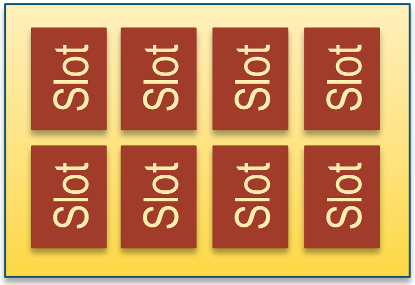
Basics of Batch Jobs

- Genepool is a shared resource
- Each calculation usually only takes a small portion of genepool
  - Every job is strictly limited on the consumption of genepool resources
  - The job description specifies the resource limits
- Univa GridEngine is used to schedule each calculation on genepool
  - The scheduler matches job resource limit requests with physical resources

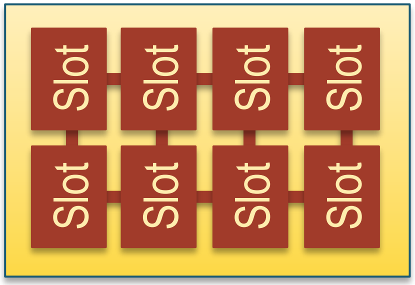
Basics of GridEngine

- **GridEngine schedules “slots”**
  - Not memory, nor processors, nor nodes
- **A *slot* is a portion of a node**
  - For most nodes on genepool, a slot is defined as a single processor plus (ram.c$_{nodeTotal}$/n$_{cores}$) memory
  - Some nodes are *exclusively scheduled* – all slots on the node are bonded together as one schedulable unit
- **Jobs are placed in *queues***
  - Queues manage the resources of disparate sets of nodes, and have distinct resource limits
    - normal.q has a 12 hour time limit
    - long.q has a 10 day time limit
- **Jobs are scheduled in order of a balance of:**
  - Resource availability
  - Job prioritization

Node



Exclusive Node



Compute Node Hardware

Count	Cores	Slots	Scheduleable Memory	Memory/ Slot	Interconnects
515	8	8	42G	5.25G	1Gb Ethernet
220	16	16	120G	7.5G	14x FDR Infiniband
8	24	24	252G	10.5G	1G Ethernet
9	32	32	500G	15.625G	4 have 10G Eth 5 have Infiniband
3	32	32	1000G	31.25G	1 has 10G Eth 2 have Infiniband
1	80	64	2000G	31.25G	10G Ethernet

- The 42G nodes are scheduled “by slot”
  - Multiple jobs can run on the same node at the same time
- Higher memory nodes are exclusively scheduled
  - Only one job can run at a time on a node

Basics of Batch Job Submission

Example Batch Script

```
#!/bin/bash
module load blast+
input=$1
database=$2
blastn -query $input -db $database <more options>
```

Submitting the example

```
genepool$ qsub -cwd example.sh queries.fa myDB
Your job 347283 (“example.sh”) has been submitted.
```

- “qsub” submits the job for batch processing
- “-cwd” directs the job to work out of the present location in the filesystem
  - the current working directory
- Default resource limits will be applied, since none were specified
  - **1 slot**
  - **5.25GB memory/slot**
  - **12 hours**

Resource Limits Request User Interface

- **Basic Resources:**
  - Cores/Processors
    - Default: 1 slot (unspecified pe)
    - -pe pe_slots n (n cores on a single machine, e.g. for threaded job)
    - -pe pe_m n*m (n nodes with m cores per node, e.g. for MPI job)
  - Memory – in units of memory/core
    - Default: 5.25G (if unspecified)
    - -l ram.c=42G (request 42 gigabytes / core)
    - -l ram.c=500 (request 500 bytes / core)

    At present memory enforcement is on virtual memory; so the entire virtual memory requirements of your job must be considered!
  - Time – specified as HH:MM:SS or in seconds
    - Default: 12:00:00 (12 hours)
    - -l h_rt=2:00:00 (request 2 hours)
    - -l h_rt=300 (request 300s = 5 minutes)

Resource Limits Request User Interface

- **Additional Resources that can be specified:**
  - -l exclusive.c request an exclusive node
  - -l infiniband.c *future:* specify a node with infiniband
- **User-Requestable Queues**
  - -l high.c **OR** -q high.q
    - put job in high-priority queue

Queues on Genepool

Queue Name	Walltime Limit	Nodes* (Slots)	Slot Limits	Memory/ slot	Other Limits
normal.q	12:00:00	443 (3544)	None	5.25G	N/A
long.q	240:00:00	70 (560)	320 per user	5.25G	N/A
normal_excl.q	12:00:00	**170 (2720)**	None	7.5G	Whole-node scheduling
long_excl.q	240:00:00	**50 (750)**	None	7.5G	Whole-node scheduling
high.q	240:00:00	10 (80)	8 per user	15G	N/A

* These numbers do not count the high-memory resources
All the high memory resources are in both normal_excl.q and long_excl.q

Submitting Jobs: Mapping Resources to Slots

- User interface is focused on machine resources required: cores, memory/core, time
- GridEngine is best able to schedule uniform-sized slots per machine-class
- NERSC automatically “re-shapes” your request to get optimally scheduled:

```
qsub -l ram.c=40G myScript.sh
```

```
qsub -l ram.c=5.25G -pe pe_slots 8 myScript.sh
```

Total memory automatically inflated to 42G

Job Prioritization: Fair Share

- **Genepool was originally created by merging together a variety of legacy systems**
- **Each group was assigned a "share" proportional to its contribution to genepool**
- **GridEngine tries to ensure that each group *on average* uses just that share**
  - When the system is idle, any group can use the **whole** cluster
- **If you belong to multiple projects, make sure you attribute the job to the correct project with:**
  - qsub -P <project>.p …

Job Submission Recommendations

- **If at all possible use 12 hours or less!**
  - The long queue has few nodes, and usage is constrained
- **Requesting more than 42G results in getting an exclusive node**
  - Unless you need the new nodes, this can significantly drain your project's share
- **Do specify –cwd or –wd <directory> with qsub**
  - Writing output to your **home directory** (the default) from the cluster *en masse* can slow everybody down
- **Specify a *meaningful* name for your job**
  - qsub -N eColi_BlastSeg11 will make things easier on you later as you try to monitor your jobs or pick up the pieces after a crash

NERSC

JGI

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Resource Limits Request User Interface

- **Examples:**
  - **Number of "slots"** – effectively processors for most of genepool
    - Request 8 processors on one node

      -
        ```
        genepool$ qsub -pe pe_slots 8 …
        ```

      -
        ```
        genepool$ qsub -pe pe_8 8 …
        ```
    - Request two 16-processor nodes

      -
        ```
        genepool$ qsub -pe pe_16 32 …
        ```
  - **Memory per slot**

      -
        ```
        genepool$ qsub -l ram.c=8G …
        ```
  - **"walltime" limit**, total execution time limit

      -
        ```
        genepool$ qsub -l h_rt=5:00:00 …
        ```