Memorial Sloan Kettering
Cancer Center

HPC User Group
Lilac Cluster

October 25, 2018

Agenda

- GPFS 5.0.1 storage available on Lilac
- New GPFS 5.0.1 storage, performance benchmarks
- July network upgrade
- LSF update
- How to execute jobs on Lilac (LSF)
- Job monitoring, LSF, and Grafana
- Documentation wiki
- How to get help on Lilac
- Q&A

Lilac computational resources as of October 25, 2018 1/1

name	#	model	CPU	GPU	cores	RAM	net	Access
ls01-18	18	SMC	2 Xeon 2.30GHz	4xGeForce GTX 1080	36	512	10GB	open
lt01-08	8	SMC	2 Xeon 2.30GHz	4xGeForce GTX 1080Ti	36	512	10GB	open
lt09-22	14	SMC	2 Xeon 2.30GHz	4xGeForce GTX 1080Ti	36	512	25GB	open
lv01	1	SMC	2 Xeon 2.30GHz	4xTeslaV100 PCIE-16GB	36	512	25GB	open
ld06-07	2	Nvidia DGX-1	2 Xeon 2.20GHz	8xTeslaV100-S XM2-16GB	40	512	10Gb	open

Lilac computational resources as of October 25, 2018 1/2

name	#	model	CPU	GPU	cores	RAM	net	Access
lg01,02, 06	3	Exxact	2 Xeon 2.30GHz	4xTitanX	36	512	10GB	SLA
lg03,04	2	Exxact	2 Xeon 2.30GHz	4xPascal	36	512	10GB	SLA
lp01-35	35	Super micro	2 Xeon 2.30GHz	4xGeForce GTX 1080Ti	24	512	10GB	SLA
ld01-05	5	Nvidia DGX-1	2 Xeon 2.20GHz	8xTeslaV100 SXM2-16GB	40	512	10GB	SLA
Total	88			380	3,196			

SLA: nodes purchased by partner PIs or departments. All users can run jobs with Walltime < 6 hours on the nodes in SLA, if the nodes are available.

GPFS 5.0.1 storage available on Lilac

name	host access	default size*	snap shots	network bandwidth	sequential 16MB performance	disaster recovery	cost
/home	all	100GB	7 days	40GB/s		Weekly copy	Free with account
/data 2.6PB	all	5TB	7 days	40GB/s	read: 27GB/s write: 14GB/s	N/A	$35/TB/ month
/warm 1.7PB	login nodes	-	7 days	8GB/s 4GB/s metadata	read: 5GB/s write: 4GB/s	N/A	$8/TB/ month

/home quota is per-user.
/data quota is per-lab.
We plan to add weekly copy for /warm on a fileset basis.

New GPFS 5.0.1 storage, performance benchmarks

Sequential Performance:

16M reads - 27GB/s

16M writes - 14GB/s

Random Performance:

16M reads - 27GB/s

16M writes - 14GB/s

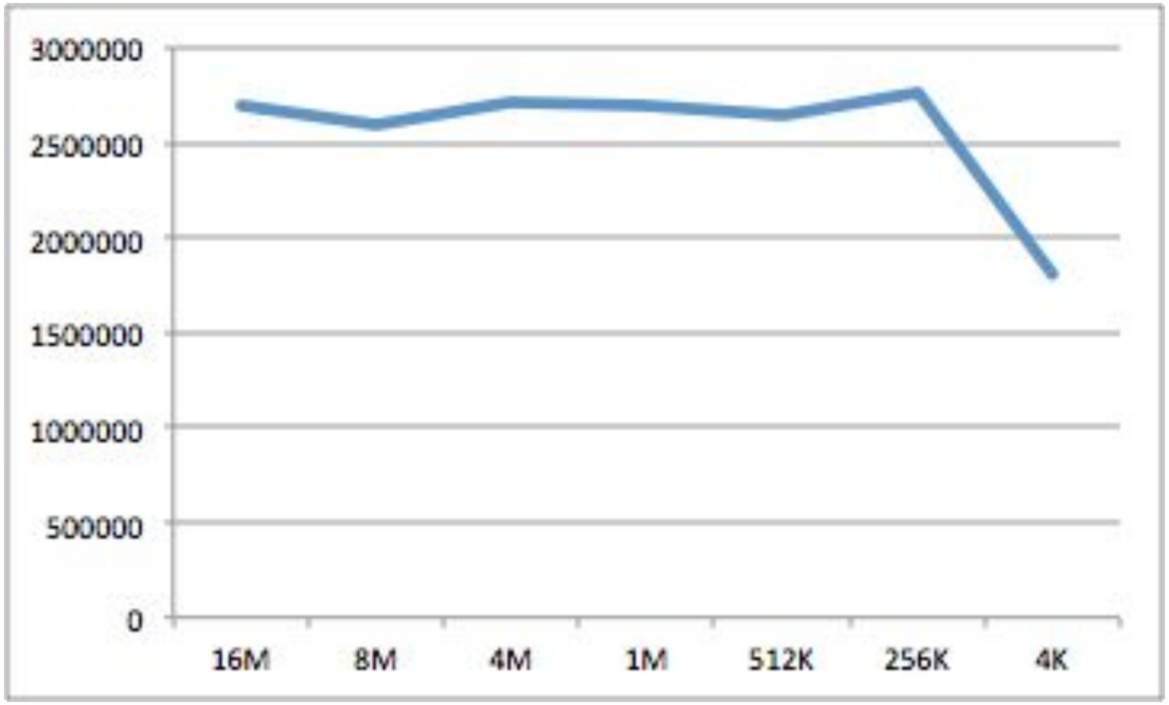
The above performance benchmarks are from GPFSperf tool, when run on multiple compute nodes on an idle cluster.

The above performance metrics were taken during network upgrade. We will be tuning few other network parameters in near future that should increase the IO performance per server and will increase the overall performance metrics.

Benchmarks per compute node

Throughput - Sequential Reads:

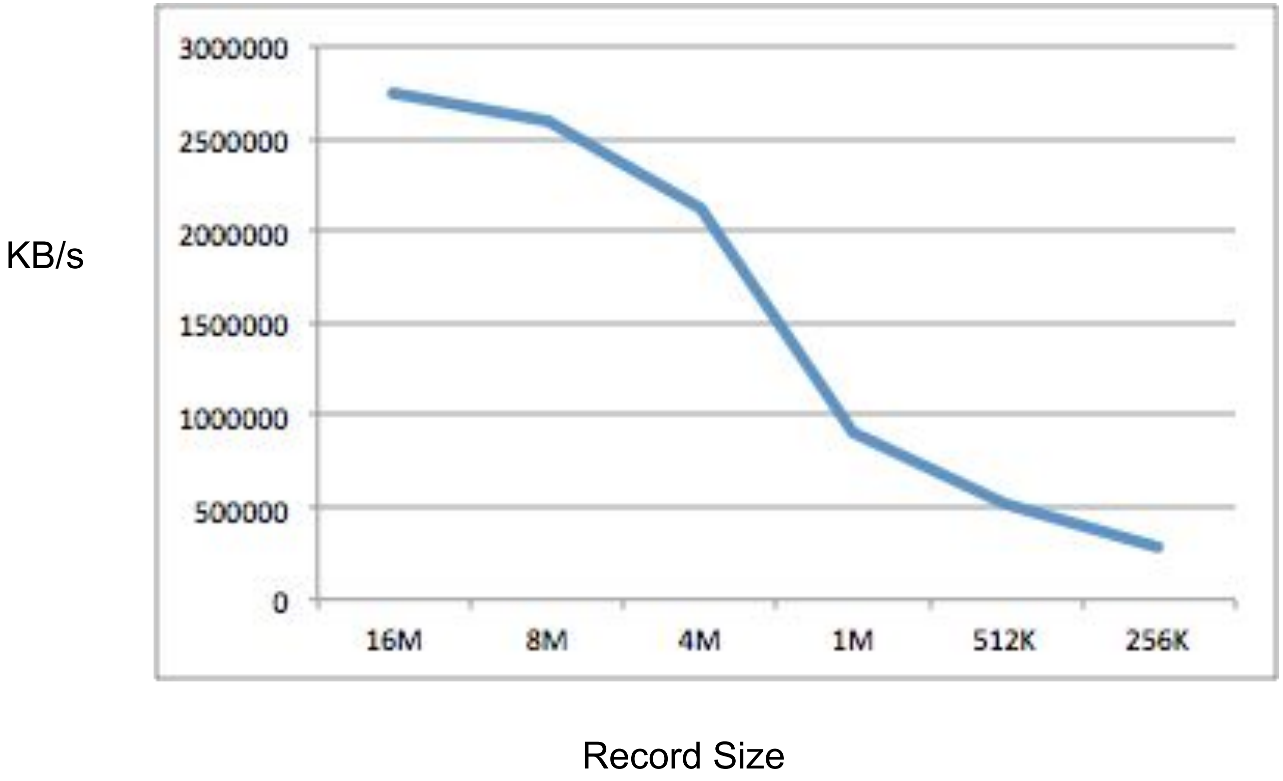
KB/s



Record Size

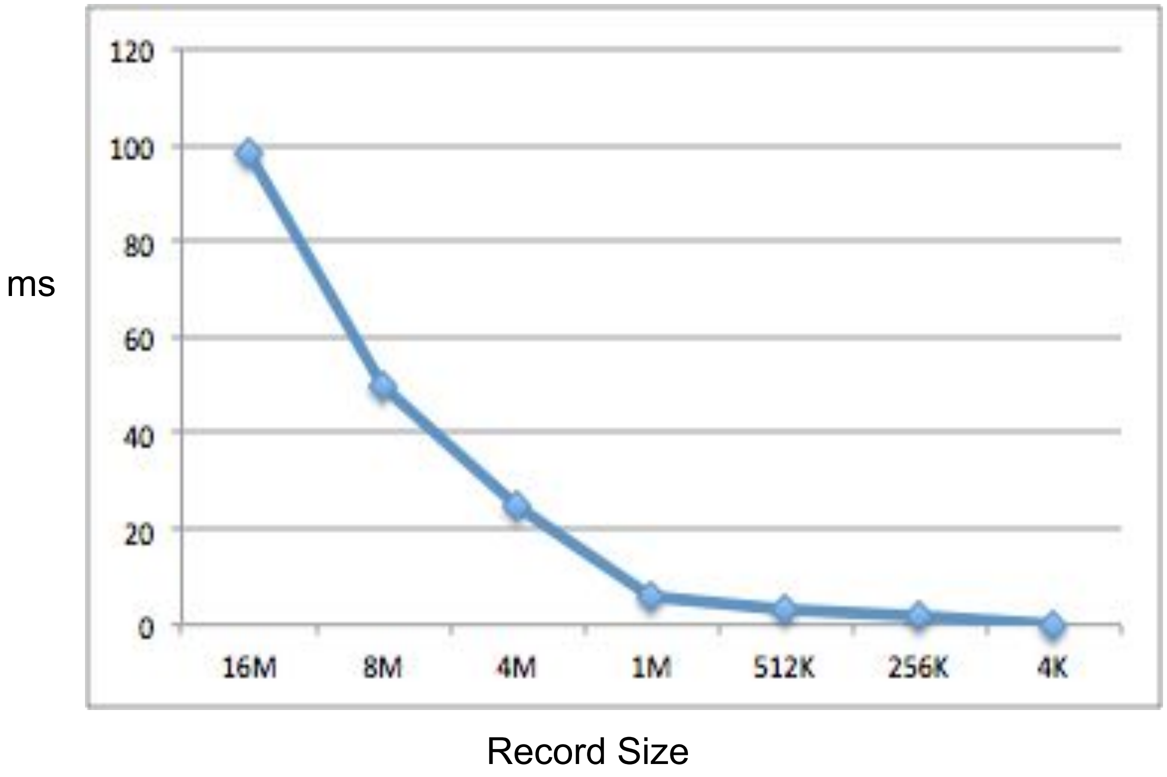
Benchmarks per compute node

Throughput - Random Reads:



Benchmarks per compute node

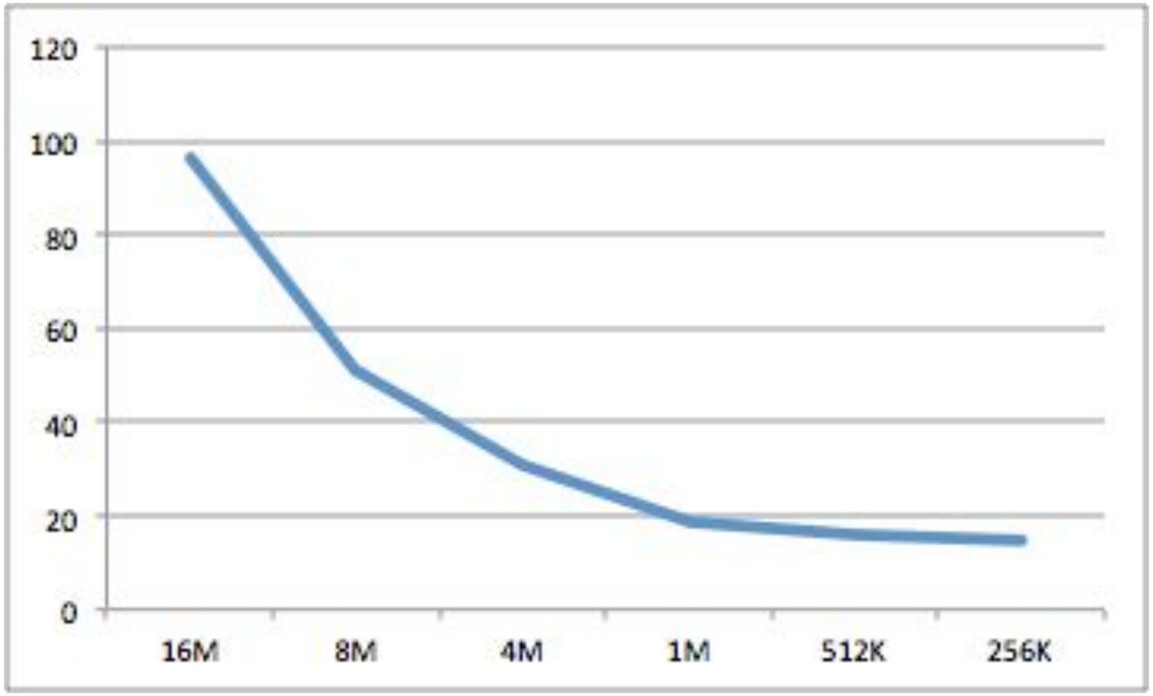
Latency - Sequential Reads:



Benchmarks per compute node

Latency - Random Reads:

ms



Record Size

Memorial Sloan Kettering
Cancer Center

July Network Upgrade, 1/3

- The HPC private network has been upgraded from 10gbps to 100gbps.
- The Lilac and Luna/Juno clusters have been consolidated onto a single Arista 7328x 100gbps network switch.
- We are working to make Lilac/Juno more similar, and exploring future changes to remove barriers between them to make it easier to leverage both clusters.

July Network Upgrade, 2/3

- The new Arista 7328x switch has 224 * 100 gigabit Ethernet ports.
- Each port can run at either 100gbps or 40gbps.
- Each port can also be broken out into 4 sub-ports, running at either 25gbps or 10gbps — for a total of 896 connections if we broke them all out.
- Most of our equipment still connects to our private network at 10gbps (compute nodes) or 20gbps (named servers), but recent equipment and new purchases use 25gbps/50gbps/100gbps with link aggregation.

July Network Upgrade, 3/3

We upgraded our newest equipment to 25gbps.

- Luna x## nodes have 25gbps connections.
- Lilac lt09..22, and lv01 have 25gbps connections.
- Lilac and Juno GPFS storage servers have been upgraded to 100gbps.
- All new equipment will connect to the private network at 25gbps or better.

LSF update

- LSF FP6 was installed on Lilac during July.
- We are testing implementation of new futures related to GPUs (gmem, gmodel, gtile, and nvlink)
- LSF FP6 has a new limitation on requesting GPUs as ngpu_physical with || (OR) conditions. This only affects GPU MPI jobs. IBM has issued a patch which requires changing LSF configuration/syntax for GPU jobs. We are testing it internally.
- LSF/GPFS jobs test suite. User input.

LSF Queues

queue	hosts	max slots per host	max RAM per host	max walltime	access
cpuqueue (default)	ls01-18 lt01-22 lg01-06 lp01-35	68	466GB	7 days	all
gpuqueue	ls01-18 lt01-22 lg01-06 lp01-35 ld01-07 lv01	ls/lt/lg/lv: 72; ld: 80	489GB	7 days	all
long	ld01-05 lg01-06	lg: 72; ld: 80	489GB	30 days	Fuchs Lab

Job Defaults on Lilac (LSF)

- Job default parameters
  - Queue name: cpuqueue
  - Number of slots (-n): 1
  - Walltime (max job runtime): 1 hours
  - Memory (RAM): 2GB
  - New! span[hosts=1]
- “bsub” will overwrite the default parameters
- Default GPU settings:

  num=1:mode=exclusive_process:mps=no:j_exclusive=yes
- Memory is per slot (bsub -n)
- To check queue configuration:

  bqueues -l

How to execute GPU jobs on Lilac (LSF)

- To submit a job with 2 slots(hyperthreads) and 20GB per slot, total 40GB RAM:

  ```
  bsub -q cpuqueue -n 2 -R “rusage[mem=20]”
  ```

- To submit a job for all slots and all 4 GPUs on one host:

  ```
  bsub -q gpuqueue -n 72 -gpu”num=4” -R”span[hosts=1]
  rusage[mem=6]”
  ```

- To submit a job with 6 slots and 3 GPUs(2 slots and 1 GPUs per host), distributed across 3 hosts:

  ```
  bsub -q gpuqueue -n 6 -gpu”num=1” -R”span[ptile=2]”
  ```

- To submit a job with one GPU job to two host partitions:

  ```
  bsub -q gpuqueue -gpu - -m "ls-gpu lt-gpu"
  ```

- To submit one slot, one GPU job to a specific GPU type:

  ```
  bsub -q gpuqueue -gpu
  -R "select[gpu_model0=='GeForceGTX1080Ti']"
  ```

- To specify non-default GPU settings for 2 GPUs job:

  ```
  bsub -q gpuqueue
  -gpu"num=2:mode=shared:j_exclusive=yes"
  ```

- To submit a job which request one DGX:

  ```
  bsub -q gpuqueue -gpu - -R V100
  ```

1884 Memorial Sloan Kettering
Cancer Center

LSF job monitoring

- Check all my jobs

```
bjobs
```

- Check my job’s stdout and stderr while the job is running

```
bpeek JID
```

- Check status of my job using JobID (JID)

```
bjobs -l JID
```

Why can’t my job run now? Check “PENDING REASONS” in bjobs output.

When will my job start to run? Check “ESTIMATION” in bjobs output.

- Why did my job exit abnormally?

```
bhist -l JID
bhist -n 0 -l JID
```

- Kill my job

```
bkill –l JID
```

- Kill all my jobs

```
bkill 0
```

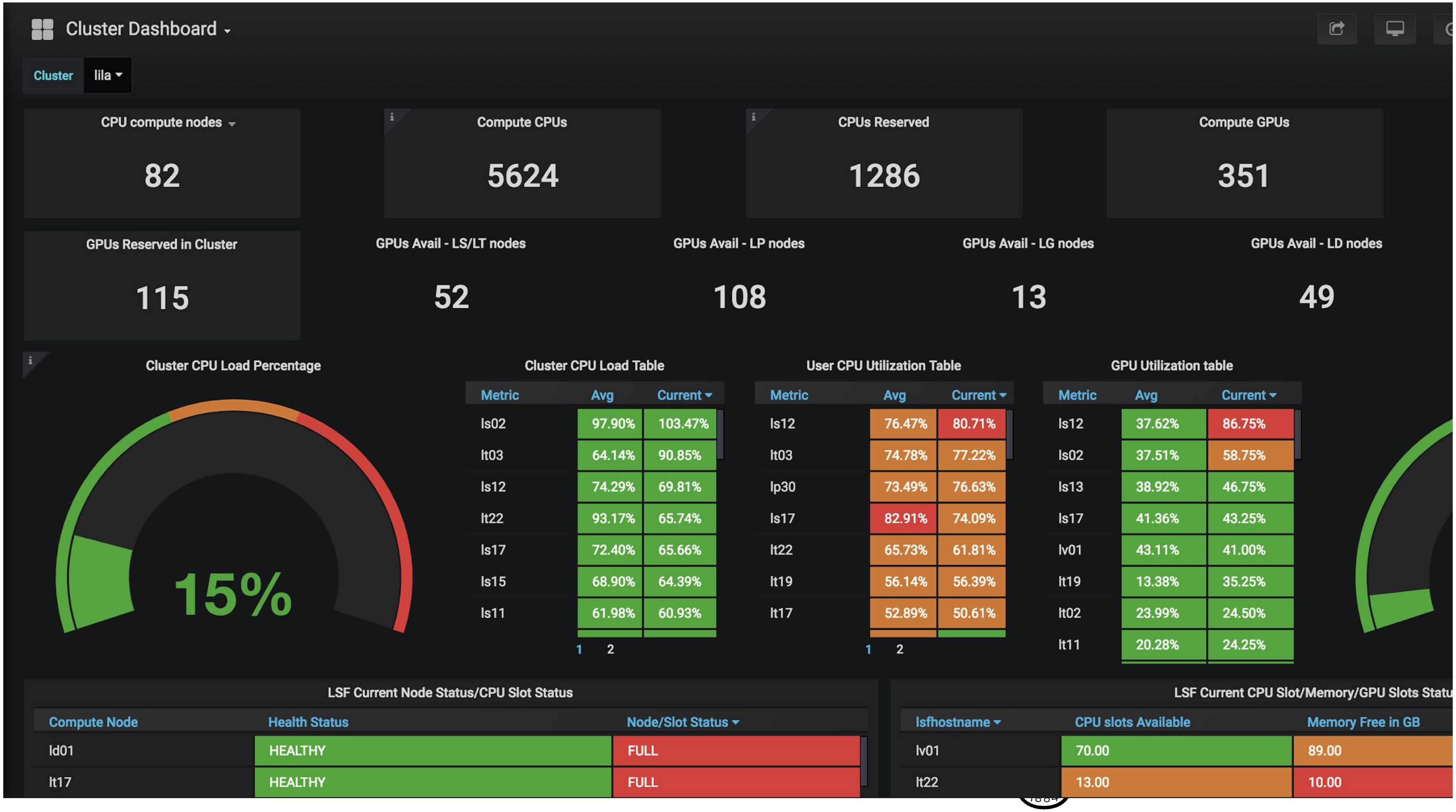
Useful LSF commands

Please use “man *command*”; or “*command* -h” to see all options

- bsub
- bhosts ; bhosts -l node_name
- bqueues; bqueues –l queue_name
- bmgroup
- lsload; lsload -l host_name
- lshosts
- bsla
- bjobs -uall -m host_name; bjobs –p
- bkill; bkill 0
- bhist -l JID; bhist –n 0 –l JID
- lsload -I healthy| grep 0.0
- lsload –s| grep gpu_model
- bhosts –s ngpus_physical

Grafana: Juno Cluster Dashboard

New URL for grafana:
https://hpc-grafana.mskcc.org/



Cluster: lila

CPU compute nodes	Compute CPUs	CPUs Reserved	Compute GPUs
82	5624	1286	351

GPUs Reserved in Cluster	GPUs Avail - LS/LT nodes	GPUs Avail - LP nodes	GPUs Avail - LG nodes	GPUs Avail - LD nodes
115	52	108	13	49

Cluster CPU Load Percentage

15%

Cluster CPU Load Table

Metric	Avg	Current
ls02	97.90%	103.47%
lt03	64.14%	90.85%
ls12	74.29%	69.81%
lt22	93.17%	65.74%
ls17	72.40%	65.66%
ls15	68.90%	64.39%
ls11	61.98%	60.93%

1 2

User CPU Utilization Table

Metric	Avg	Current
ls12	76.47%	80.71%
lt03	74.78%	77.22%
lp30	73.49%	76.63%
ls17	82.91%	74.09%
lt22	65.73%	61.81%
lt19	56.14%	56.39%
lt17	52.89%	50.61%

1 2

GPU Utilization table

Metric	Avg	Current
ls12	37.62%	86.75%
ls02	37.51%	58.75%
ls13	38.92%	46.75%
ls17	41.36%	43.25%
lv01	43.11%	41.00%
lt19	13.38%	35.25%
lt02	23.99%	24.50%
lt11	20.28%	24.25%

LSF Current Node Status/CPU Slot Status

Compute Node	Health Status	Node/Slot Status
ld01	HEALTHY	FULL
lt17	HEALTHY	FULL

LSF Current CPU Slot/Memory/GPU Slots Statu

lsfhostname	CPU slots Available	Memory Free in GB
lv01	70.00	89.00
lt22	13.00	10.00

How to get help on Lilac

- Please send email to: hpc-request@cbio.mskcc.org
- Additional options for help and contacting us:

  http://hpc.mskcc.org/contact-us/

Documentation wiki

- http://mskcchpc.org/display/CLUS/Lilac+Cluster+Primer
- http://hpc.mskcc.org/compute-accounts/

Questions & Answers