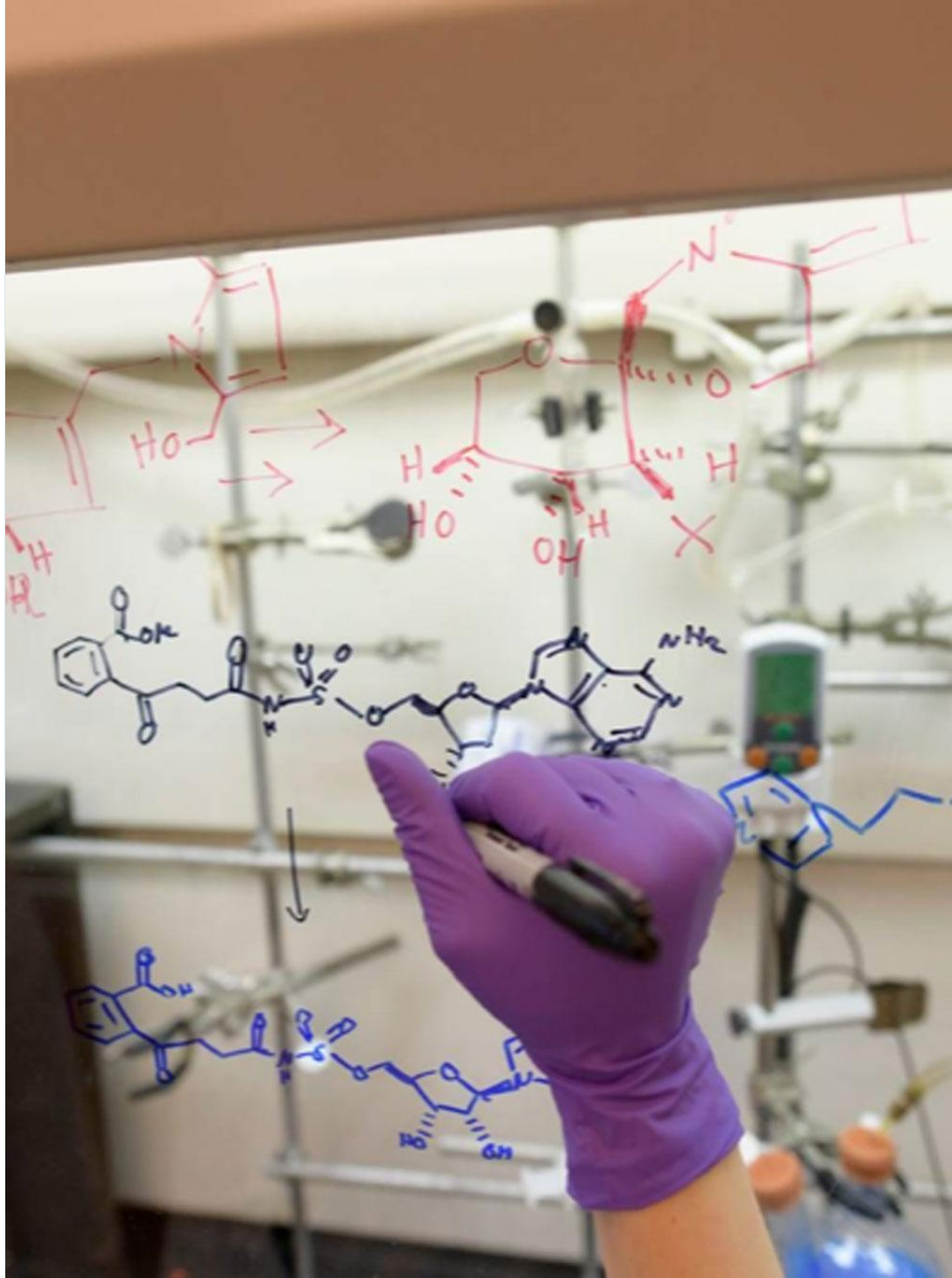
Memorial Sloan Kettering Cancer Center

DigITs

DIGITAL, INFORMATICS, TECHNOLOGY

HPC User Group

September 9, 2020



Memorial Sloan Kettering Cancer Center

DigITs

DIGITAL, INFORMATICS, TECHNOLOGY

Introduction

Juan Perin

September, 2020

HPC Group

Agenda for Today

Sept 9, 2020

1. DigITs Updates:
   a. The new team - Advanced Computing and Technology Group
   b. Cloud Center of Excellence - CCoE

2. HPC Infrastructure Updates - Sveta Mazurkova

3. Systems retirement and Data Management - Chris Pepper

4. Performance tuning and troubleshooting - Lohit Valleru

Memorial Sloan Kettering Cancer Center

DigITs

DIGITAL, INFORMATICS, TECHNOLOGY

Lilac and Juno Update

Sveta Mazurkova

September, 2020

HPC Group

Lilac Cluster

As of Sept, 2020

name	#	model	CPU	GPU	cores	RAM	Access
ld06-07	2	Nvidia DGX-1	2Xeon(R)2.20GHz	8xTeslaV100-SXM2-16GB	40	512	All users
ls01-18	18	Supermicro	2Xeon(R)2.30GHz	4xGeForce GTX 1080	36	512	All users
lt01-08	8	Supermicro	2Xeon(R)2.30GHz	4xGeForce GTX 1080Ti	36	512	All users
lt09-22	14	Supermicro	2Xeon(R)2.30GHz	4xGeForce GTX 1080Ti	36	512	All users
lv01	1	Supermicro	2Xeon(R)2.30GHz	4xTeslaV100 PCIE-16GB	36	512	All users
lu01-12	12	Supermicro	2Xeon(R)2.30GHz	4xGeForce RTX 2080	36	512	All users
lg01,02,03	3	Exxact	2Xeon(R)2.30GHz	4xTITANX	36	512	ACL
lg03,04	2	Exxact	2Xeon(R)2.30GHz	4xPascal	36	512	ACL
lp01-35	35	Supermicro	2Xeon(R)2.30GHz	4xGeForce GTX 1080Ti	24	512	ACL
ld01-05	5	Nvidia DGX-1	2Xeon(R)2.20GHz	8xTeslaV100 SXM2-16GB	40	512	ACL

Lilac Cluster (continued)

As of Sept, 2020

name	#	model	CPU	GPU	cores	RAM	Access
lx01-12	12	Supermicro	2Xeon(R)2.10GHz	4xGeForce RX2080Ti	40	754	All users
ly01-09	9	Supermicro	2Xeon(R)2.10GHz	4xQuadro RTX 6000	40	754	ACL
lw01-02	2	Supermicro	2Xeon(R)2.10GHz	8xGeForce RTX 2080Ti	40	1.5TB	All users
li01	1	Supermicro	2Xeon(R)2.10GHz	4xGeForce RTX 2080Ti	52	1TB	ACL
Total	**124**			**508**	**4,180**		

Juno Cluster

As of Sept, 2020

name	#	model	CPU	cores	RAM	NVMe
ju01-34	34	HPE DL160G9	2Xeon®2.60GHz	16	256	
jv01-04	4	HPE DL160G9	2Xeon®2.60GHz	16	256	
jx01-34	34	Supermicro	2Xeon®2.40GHz	20	256	2x2TB
jy01-03	3	Supermicro	2Xeon®2.40GHz	36	512	
ja01-10	10	Supermicro	2Xeon®2.30GHz	36	512	2TB
jb01-24	24	Supermicro	2Xeon®2.30GHz	36	512	2TB
jc01-02	2	Supermicro, 4*2080Ti GPUs	2Xeon®2.30GHz	36	512	2TB
jd01-04	4	Supermicro	2Xeon®2.30GHz	36	512	2TB
jf01-02	2	Supermicro, 4*2080Ti GPUs	2Xeon®2.10GHz	40	768	2x2TB
je01-24	24	Supermicro	2Xeon®2.10GHz	40	768	2x2TB
Total	**140**			**3,924**		

GPFS Storage on Lilac

As of Sept, 2020

name	host access	default quota	snaps	bandwidth	performance: sequential 16MB access	Disaster recovery	Cost
/home	all	100GB	7 days	50gbyte/s	read 50gbyte/s write 40gbyte/s	Weekly copy	Free with account
/data 4.1PB	all	1TB	7 days	50gbyte/s	read 50gbyte/s write 40gbyte/s	N/A	$35/TB/m
/warm 3.1PB	login nodes	-	7 days	8gbyte/s 4gbyte/s metadata	read 8gbyte/s write 5gbyte/s	N/A	$8/TB/m
/allen 120TB	ld-gpu (DGX-1)	-		20gbyte/s	Test (DGX-1)	N/A	Testing

GPFS Storage on Juno

As of Sept, 2020

name	host access	default quota	snaps	bandwidth	performance: sequential 16MB access	Disaster recovery	Cost
/home	all	100GB	7 days	40gbyte/s		Weekly copy	Free with account
/juno 8.1PB	all	1TB	7 days	100gbyte/s	read 100gbyte/s write 80gbyte/s	N/A	$35/TB/m
/warm 3.1PB	login nodes	-	7 days	8gbyte/s 4gbyte/s metadata	read 8gbyte/s write 5gbyte/s	N/A	$8/TB/m

Lilac and Juno Clusters

New preemptable queue

Lilac user request: "The CPU jobs should be able to use all available resources and when priority jobs are in PEND state the preemptable jobs would be killed automatically.”

Lilac cluster. Queue: preemptable

Hosts: all, except DGX

```
bsub -n 5 -W 90 -q preemptable …
```

Juno cluster. Queue: preemptable

Hosts: all, except jw02 (control)

```
bsub -n 5 -W 90 -q preemptable …
```

Lilac and Juno Clusters

LSF status and work in progress

Juno: LSF 10.1 Fix_Pack 9. Custom patch to fix job exit signal 19. Custom patch to fix mbatchd high memory usage issue.

Lilac: LSF 10.1 Fix_Pack 9

LSF 10.1 Fix_Pack 10 available. Work in progress with IBM on singularity job termination.

Why Doesn’t My Job Run?

LSF job status, resource utilization

> bjobs -p3 -l ..

Tue Sep 8 16:01:28: Submitted from host <lilac>, CWD <$HOME>,
Requested Resources <select[gpu_model0=='GeForceGTX1000']>,
Requested GPU;

PENDING REASONS:
Candidate host pending reasons (0 of 123 hosts).

Non-candidate host pending reasons (123 of 123 hosts):
Job's resource requirements not satisfied: lp35, lx10, lx11, lx12, lx13, lx14, boson, lt01, lt02, lt03, lt04, lt05, lt06, lt07, lt08
…..
Not specified in job submission: ld01, ld02, ld03, ld04, ld05, ld07, lv01, l
i01, lila-sched01, lila-sched02;
Load information unavailable: ld06, lg05, lp08, lp09, ls10, ls18, lp21, lp26

Closed by LSF administrator: lw01, lw02, ls05, lu04, lu05, lx09;

RUNLIMIT
10.0 min

This job won’t run because the gpu_model0 is not correct in bsub and this resource is not available on Lilac cluster :
Candidate host : 0
Correct name is GeForceGTX1080

>lshosts -gpu

HOST_NAME	gpu_id	gpu_model	gpu_driver	gpu_factor	numa_id
ls01	0	GeForceGTX1080	440.33.01	6.1	0
	1	GeForceGTX1080	440.33.01	6.1	0
	2	GeForceGTX1080	440.33.01	6.1	1
	3	GeForceGTX1080	440.33.01	6.1	1

Why Doesn’t My Job Run?

cont.

>bjobs -p3 -l …

Tue Sep 8 16:05:23: Submitted from host <lilac>, CWD <$HOME>, 4 Task(s), Requested Resources <rusage[mem=200]>;
PENDING REASONS:
Candidate host pending reasons (99 of 123 hosts):
Resource limit defined on host(s) and/or host group has been reached (Resource: mem, Limit Name: limit11, Limit Value: 95): lt15, lt17,lt18, lt19, lx14….
Job's requirements for resource reservation not satisfied (Resource: mem): l x10, lx12, lx13, boson, lt05, lt08, lt09, lx08, lx07….
Host is reserved to honor SLA guarantees: lp34, lp32, lp01, lp03, lp05, lp06…..
Non-candidate host pending reasons (24 of 123 hosts):
Not specified in job submission: li01, lv01...
Load information unavailable: lp21, ls18, lp26, ls10, lp09, lp08, lg05, ld06
Closed by LSF administrator: lu04, lu05, ls05, lx09, lw02, lw01;
MEMLIMIT
200 G
RESOURCE REQUIREMENT DETAILS:
Combined: select[(healthy=1) && (type == local)] order[!-slots:-maxslots] rusage[mem=200.00] span[hosts=1] same[model] affinity[thread(1)*1]

This job won’t run on Lilac cluster, because it requested 4x200=800GB of RAM on the same host span[hosts=1]
There is no host with 800GB on Lilac in cpuqueue.

Please, check resources:
>lshosts
HOST_NAME type model cpuf ncpus maxmem maxswp
ls01 X86_64 GTX1080 60.0 72 512G ..

To check limits on resources:
>bresource
Begin Limit
NAME = limit11
QUEUES = cpuqueue
PER_HOST = ls-gpu/ lt-gpu/ lg-gpu/ lu-gpu/ lx-gpu/ lw-gpu/ ly-gpu/
SLOTS = 68
MEM = 95%
ngpus_physical = 0
End Limit

Why Doesn't My Job Run?

cont.

bjobs -p3 -l…..

#BSUB -n 1;#BSUB -gpu 'num=1';#BSUB -R 'span[ptile=1] rusage[mem=30]';#BSUB -q gpuqueue;

PENDING REASONS:
Candidate host pending reasons (92 of 123 hosts):
Job's requirements for resource reservation not satisfied (Resource: ngpus_physical): lx12, lx14, lt01, lt02, lt03, lt04, lt05, lt07, lt08, lt09, lu10, lx05, lx04, lx03, lg02, lu08, lt12, lt19...
Affinity resource requirement cannot be met because there are not enough processor units to satisfy the job affinity request: lt10, lt11, lx10, lt13, lt14…
Host is reserved to honor SLA guarantees: lp31, lp01, lp03, lp04, lp05, lp06, lp07, lp27, lp33, lp30, lp34, lp25, lp24, lp35, lp10…
Non-candidate host pending reasons (32 of 123 hosts):
Job's resource requirements not satisfied: lu02, ls13, lv01, lg06, ld07, ld05.
Load information unavailable: ls18, lp21, ls10, lp26, lp09, lp08, lg05, ld06
Closed by LSF administrator: lu04, lu05, ls05, lx09, lw02, lw01;
Not enough GPUs on the hosts: ly03, lx01, lx02, lx06, lx11;
ESTIMATION:
Tue Sep 8 17:00:21: Started simulation-based estimation;
Tue Sep 8 17:00:39: Simulated job start time <Tue Sep 8 17:29:08> on host(s) <1*lt03>

This job is waiting for mainly gpu resources to be available in Lilac cluster.

Estimated start time is :
Simulated job start time <Tue Sep 8 17:29:08> on host(s) <1*lt03>

bhosts -l lt03
HOST lt03
STATUS CPUF JL/U MAX NJOBS RUN SSUSP USUSP RSV DISPATCH_WINDOW
ok 60.00 - 72 4 4 0 0 0 -
CURRENT LOAD USED FOR SCHEDULING:

	r15s	r1m	r15m	ut	pg	io	ls	it	tmp	swp	mem	slots	ngpus
Total	0.0	0.0	0.0	8%	0.0	19	1	76	48.7G	0G	269G	68	4.0
Reserved	0.0	0.0	0.0	0%	0.0	0	0	0	0G	0G	22G	-	0.0

	ngpus_physical	healthy	gpu_shared_avg_ut	gpu_shared_avg_mut
Total	0.0	1.0	44.0	1.0
Reserved	4.0	0.0	0.0	0.0

Memorial Sloan Kettering Cancer Center

DigITs

DIGITAL, INFORMATICS, TECHNOLOGY

Retirement & File Transfers

Chris Pepper

September, 2020

HPC Group

solisi & luna Retirement

/ifs has reached end of life

The solisi storage system, used for /ifs/res and /ifs/work, as reached the end of its usable life. Our support contracts have begun to expire, and parts are no longer available for most of the system from the manufacturer.

The /juno GPFS filesystem is currently 8PB, and the /res & /work directories on the juno cluster are *really* /juno/res & /juno/work. Let us know if you need a quota adjustment or help with the migration, but the /ifs filesystem should no longer be considered reliable.

We have removed most of the compute nodes from the luna cluster, and migrated them into the juno cluster, which now contains over 100 nodes. We will disable access to luna soon — you should be able to do all the same work on juno instead.

lilac Work Next Wednesday (9/16)

We are upgrading software on the lilac cluster

We need to perform disruptive upgrades on the lilac login node. We will begin at noon on 9/16 and expect to be finished by 1pm. We will disconnect logged-in users, but LSF jobs will not be disrupted. Please do not log into lilac until this work is finished (which we will announce in the #hpc_mskcc Slack channel), and instead use lilac2.mskcc.org, our alternative login server. You can do the same work on lilac2 as you would on lilac.

Large Network Transfers

Please warn us in advance of large network transfers to or from the juno & lilac clusters

MSKCC monitors for large data flows, and large downloads and uploads can disrupt access for other users or be blocked as possible malicious activity. If you intend to perform large data transfers (larger than 10 gigabytes), please let us know in advance via Jira ticket. Provide some detail about what you're transferring, why, and when, and we'll warn the Networking team to make sure your transfers aren't impeded.

If you perform large or frequent data transfers, consider using juno-xfer01 and lilac-xfer01. These are dedicated data transfer servers. They're not LSF submit hosts and don't have GPUs or a large amount of computational power, but they have fast network connections and are lightly used. They have access to the same data as juno and lilac, respectively, and are the fastest tools for data transfers.

Memorial Sloan Kettering Cancer Center

DigITs

DIGITAL, INFORMATICS, TECHNOLOGY

Performance Troubleshooting

Lohit Valleru

September, 2020

HPC Group

Login Server Performance Limitations

General purpose login servers limit CPU and memory resource usage with cgroups.

Resource limits are applied per-user, not per ssh session or per-application.

All a user's ssh sessions count against the same per-user resource limits.

Login server limits

Lilac: 4 CPUs and 2 GB memory per user

Juno: 1 CPU and 5 GB memory per user

Due to the above limits, running any compute, memory, or IO intensive tasks on login nodes can drastically slow down all usage, including data housekeeping.

IO Performance Issues

Login Server:

Although cgroups limit the CPU/memory usage per user - cgroups cannot limit IO or Network utilization per user.

Login server is a shared resource, and thus if any user reaches the limits of GPFS on the login server - Storage IO could slow down for the rest of the users.

Compute Node:

Compute server IO performance is constrained by GPFS and network limits per node. Each compute node is limited to the network card throughput available for that node, and the storage hardware bandwidth.

IO Performance Debugging

Resources available for independent debugging, to confirm IO issues:

Grafana:

https://hpc-grafana.mskcc.org

Check for high load on the node, or any other heavily used resource.

Linux Tools:

ls:

Is `ls` very slow on a particular path, or an entire filesystem?

top:

See whether your applications are waiting for storage or network IO, which is shown with “D” status.

systemd-cgtop:

See whether cgroups are at resource limits on the login node.

IO Performance Debugging (continued)

strace:

strace on an application can show where it is spending most of its time:

Examples:

```
strace -tTfp <pid> -o <outputfile>

strace -tTf <applicationcommand> -o <outputfile>

strace -cp <pid> -o <outputfile>
```

IO Performance Debugging (continued)

Please include as much information as possible when opening a ticket for an IO performance issue:

1. Time of the slowdown.
2. Server on which the slowdown occurred. Remember that login servers are shared, and restricted using cgroups. Please include any other relevant details for involved servers.
3. Full command that seems to be responding slower than usual, along with full output and any other details.
4. LSF job ID, if applicable.
5. Time taken to list the path of the dataset or command:

```
date; time ls <path>

date; time /usr/bin/ls <path>.
```

6. Process status output:

```
date; ps auwx | grep -i <applicationname>
```

IO Performance Debugging (continued)

7. For interactive applications :

```
strace -tTf <applicationcommand> -o <outputfile>
```

8. For non-interactive applications, run the commands below for 1 minute, and attach the output:

```
strace -tTfp <pidofapplication> -o <outputfile>
strace -cfp <pidofapplication> -o <outputfile>
```

IO Performance Debugging (continued)

Demonstration of Grafana and additional tools.

https://hpc-grafana.mskcc.org

Zoom Poll

Zoom Q&A Discussion

Please either “Raise Hand” in Zoom to ask out loud, or ask questions in Zoom Chat.

What do you want from juno/lilac or a new cluster?

How has COVID-19 and the lockdown affected your usage of HPC resources? xbio and VPN have been much more heavily utilized, and many people have new laptops, but what else should we know?