BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY

SLURM Quick Start

Getting started on HPC

Tin Ho -- tin (at) lbl.gov

HPC Engineer

Science IT

LabTech 2018

Slurm talk overview

- What is Slurm
- How to submit job
- How to run job on GPU
- Simple troubleshooting

If you ever wondered about

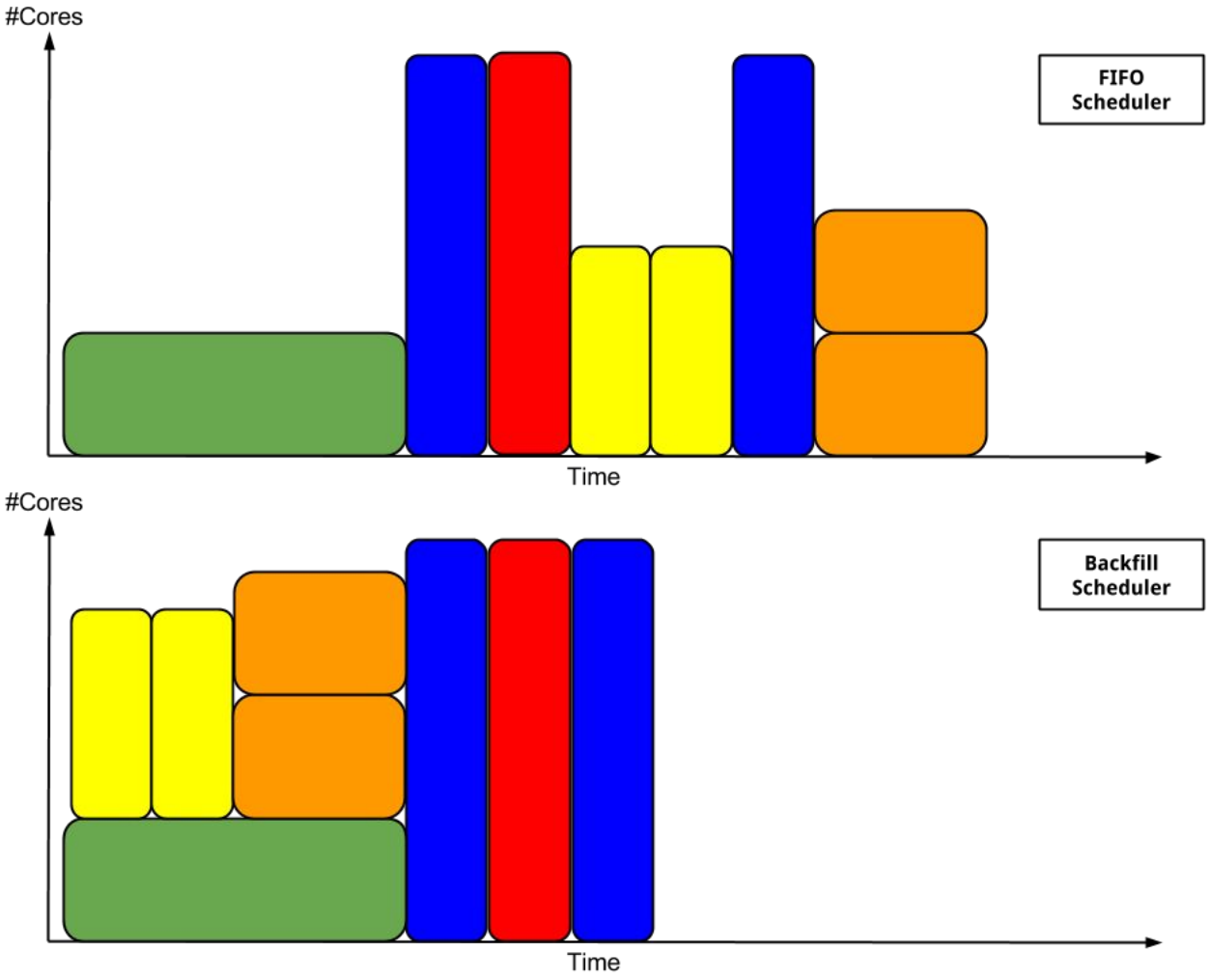
- Partitions
- QoS
- Scheduling and Priorities
- Condo vs Recharge

What is Slurm?

What is Slurm?? Why do I need to use it?

- “Supercomputer”
- HPC - High Performance Cluster
  - Many nodes
  - Δ CPU
  - Δ RAM
  - Δ GPU
- Workload manager
- Schedule, dispatch job

Slurm as a Batch Scheduler



Ref: https://wiki.auckland.ac.nz/pages/viewpage.action?pageId=88902184

What info is needed to use HPC?

From Scheduler's perspective:

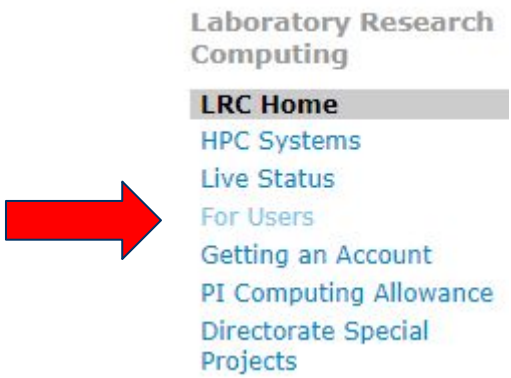
- Number of CPU cores
- Amount of RAM
- How long would it run

From User's Perspective:

- What kind of nodes are avail? How much are they?
- What QoS and account info do I use?
- Interactive vs Batch jobs?

What kind of CPU are available?

http://lrc.lbl.gov



Scroll down for recharge info.

LAWRENCIUM - High Performa

https://sites.google.com/a/lbl.gov/high-performance-computing-ser

BERKELEY LAB INFORMATION TECHNOLOGY
LAWRENCE BERKELEY NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Search this site

LRC Home | For Users | Getting Help | About | IT Division | Berkeley Lab

- Getting Started
  - New User Information
  - UserAgreement
  - LRC Supercluster Login
  - Environment Modules Usage Instructions
  - SLURM Usage Instructions
- User Guide
  - Jupyter Notebook
  - Remote Desktop / Remote Visualization
  - Data Transfer
  - Tips and Tricks
- Gateway & Authentication
  - LinOTP Usage
- HPCS Software Module Farm
  - Service Overview
  - Current Status
- Systems
  - ALSACC
  - BALDUR
  - CATAMOUNT
  - CUMULUS
  - DIRAC1
  - EXPLORER
  - ETNA
  - HBAR
  - JBEI
  - JCAP
  - **LAWRENCIUM**
  - MAKO
  - MHG
  - MUSIGNY
  - NANO
  - VOLTAIRE
  - VULCAN
  - XMAS
- UC Berkeley Systems

Cluster Description | Login and Data Transfer | HW Configuration | Storage and Backup | Recharge Model | Scheduler Configuration | Low Priority QoS Jobs | Job Script Examples

Hardware Configuration:

LAWRENCIUM is composed of multiple generations of hardware hence it is physically separated into several partitions to facilitate management and to meet the requirements to host Condo projects. The following table lists the hardware configuration for each individual partition.

Partition	Nodes	Node List	CPU	Cores	Memory	Infiniband	Accelerator
lr2	198	n0[000-141].lr2 n0[146-153].lr2 n0[161-180].lr2 n0[182-208].lr2	INTEL XEON X5650	12	24GB	QDR	-
		n0181.lr2			96GB		
lr3	300	n0[000-163].lr3	INTEL XEON E5-2670	16	64GB	FDR	-
		n0[164-203].lr3 n0[213-308].lr3	INTEL XEON E5-2670 v2	20	64GB		
lr4	108	n0[000-095].lr4 n0[099-110].lr4	INTEL XEON E5-2670 v3	24	64GB	FDR	-
lr5	144	n0[000-143].lr5 n0[148-171].lr5	INTEL XEON E5-2680 v4 INTEL Xeon ES-2640 v4	28 20	64GB 128GB	FDR QDR	-
lr_amd	4	n0[157-160].lr2	AMD OPTERON 6276	32	64GB	QDR	-
lr_bigmem	5	n0[154-155].lr2	AMD OPTERON 6174	48	256GB	QDR	-
		n0156.lr2	AMD OPTERON 6180 SE	48	512GB		
		n0210.lr3	INTEL XEON E5-4620	32	1024GB		
		n0211.lr3	INTEL XEON E7-4860 v2	48	1024GB		
lr_manycore	12	n0[204-207].lr3	INTEL XEON E5-2603	8	64GB	FDR	INTEL XEON PHI 7120
		n0[208-209].lr3	INTEL XEON E5-2603	8	64GB		NVIDIA KEPLER K20
		n0[096-098].lr4	INTEL XEON E5-2623 v3	8	64GB		4X NVIDIA KEPLER K80
		n0[136-138].lr4	INTEL XEON E5-2623 v3	8	64GB		4X NVIDIA GTX 1080TI
mako	272	n0[000-271].mako0	INTEL XEON E5530	8	24GB	QDR	-
mako_ manycore	4	n0[272-275].mako0	INTEL XEON X5650	12	24GB	QDR	2X NVIDIA TESLA C2050
cf1	20	n0[000-019].cf1	INTEL XEON Phi 7210	64	192GB	FDR	

Storage and Backup:

What kind of CPU are available?

```
[tin@n0003]> scontrol show partition | grep Partition
PartitionName=alice
PartitionName=alsacc
PartitionName=baldur1
PartitionName=catamount
PartitionName=cf1
PartitionName=cf1-hp
PartitionName=dirac1
PartitionName=etna
PartitionName=etna-shared
PartitionName=etna_gpu
PartitionName=explorer
PartitionName=hbar1
PartitionName=jbei1
PartitionName=lr2
PartitionName=lr3
PartitionName=lr4
PartitionName=lr_amd
PartitionName=lr_bigmem
PartitionName=lr_manycore
PartitionName=lr5
PartitionName=lr6
PartitionName=mako
PartitionName=mako_manycore
PartitionName=mhg
PartitionName=musigny
PartitionName=nano1
PartitionName=voltaire
PartitionName=vulcan
PartitionName=vulcan_gpu
PartitionName=vulcan_c20
PartitionName=xmas
PartitionName=lr
```

```
tin@n0003:~
File  Edit  View  Search  Terminal  Help
**^ tin n0003.scs00 ~ ^**> scontrol show partitions=lr6
PartitionName=lr6
   AllowGroups=ALL AllowAccounts=ALL AllowQos=ALL
   AllocNodes=ALL Default=NO QoS=N/A
   DefaultTime=NONE DisableRootJobs=NO ExclusiveUser=NO GraceTime=0 Hidden=NO
   MaxNodes=UNLIMITED MaxTime=UNLIMITED MinNodes=1 LLN=NO MaxCPUsPerNode=UNLIMITED
   Nodes=n0[000-019].lr[6]
   PriorityJobFactor=1 PriorityTier=1 RootOnly=NO ReqResv=NO OverSubscribe=EXCLUSI
   OverTimeLimit=NONE PreemptMode=REQUEUE
   State=UP TotalCPUs=640 TotalNodes=20 SelectTypeParameters=NONE
   DefMemPerNode=93000 MaxMemPerNode=UNLIMITED

**^ tin n0003.scs00 ~ ^**>
```

```
tin@n0003:~
File  Edit  View  Search  Terminal  Help
**^ tin n0003.scs00 ~ ^**> scontrol show node=n0001.lr6
NodeName=n0001.lr6 Arch=x86_64 CoresPerSocket=16
   CPUAlloc=0 CPUErr=0 CPUTot=32 CPULoad=0.01
   AvailableFeatures=lr6_c32,lr6
   ActiveFeatures=lr6_c32,lr6
   Gres=(null)
   NodeAddr=10.0.41.1 NodeHostName=n0001.lr6 Version=17.11
   OS=Linux 3.10.0-693.11.6.el7.x86_64 #1 SMP Wed Jan 3 18:09:42 CST 2018
   RealMemory=95308 AllocMem=0 FreeMem=92055 Sockets=2 Boards=1
   State=IDLE ThreadsPerCore=1 TmpDisk=7935 Weight=1 Owner=N/A MCS_label=N/A
   Partitions=lr6
   BootTime=2018-09-13T10:55:41 SlurmdStartTime=2018-09-13T10:56:36
   CfgTRES=cpu=32,mem=95308M,billing=32
   AllocTRES=
   CapWatts=n/a
   CurrentWatts=0 LowestJoules=0 ConsumedJoules=0
   ExtSensorsJoules=n/s ExtSensorsWatts=0 ExtSensorsTemp=n/s
```

What account and QoS do I use?

QoS = Quality of Service = Priority
Account = Project code for Service Unit tracking

```
> sacctmgr show associations -p user=dromps format=acc,part,qos |\
  sed 's/|/\t/g'

Account         Partition       QOS
ac_cumulus      lr5             lr_debug,lr_lowprio,lr_normal
ac_cumulus      lr4             lr_debug,lr_lowprio,lr_normal
ac_cumulus      lr3             lr_debug,lr_lowprio,lr_normal
ac_cumulus      lr2             lr_debug,lr_lowprio,lr_normal
ac_cumulus      lr_manycore     lr_lowprio,lr_normal
ac_cumulus      lr_bigmem       lr_lowprio,lr_normal
ac_cumulus      lr_amd          lr_lowprio,lr_normal
ac_cumulus      cf1             cf_debug,cf_normal
ac_cumulus      mako_manycore   mako_lowprio,mako_normal
ac_cumulus      mako            mako_debug,mako_lowprio,mako_normal
lr_cumulus      lr2             condo_cumulus
lr_cumulus      lr6             condo_lr6_normal
```

Submit a batch job

```
bash /home/tin
File  Edit  View  Search  Terminal  Help
[tin@n0003]> cat slurm-job.sh
#!/bin/sh
#SBATCH --job-name=test
#SBATCH --partition=lr6
#SBATCH --qos=lr_normal
#SBATCH --account=scs
#SBATCH --nodes=2
#SBATCH --mem-per-cpu=2G
#SBATCH --time=02:30:00

hostname
date
uptime
[tin@n0003]> sbatch slurm-job.sh
Submitted batch job 14485093
[tin@n0003]> cat slurm-14485093.out
n0003.lr6
Fri Oct 19 16:06:47 PDT 2018
 16:06:47 up 37 days,  5:00,  0 users,  load average: 0.00, 0.01, 0.05
[tin@n0003]>
```

Interactive Job

```
fish /home/ti...   tin@n0003:~
**^ tin n0003.scs00 ~ ^**>  srun  --pty  -p lr2  -A scs  --qos=lr_normal  -N 1
-t 0:0:05  /bin/bash
**^ tin n0124.lr2 ~ ^**>  hostname
n0124.lr2
**^ tin n0124.lr2 ~ ^**>  w
 20:43:18 up 59 days, 11:38,  0 users,  load average: 0.08, 0.03, 0.05
USER     TTY      FROM             LOGIN@   IDLE   JCPU   PCPU WHAT
**^ tin n0124.lr2 ~ ^**>  date
Thu Oct 18 20:43:20 PDT 2018
**^ tin n0124.lr2 ~ ^**>  srun: Force Terminated job 14460607
srun: error: n0124.lr2: task 0: Killed
srun: Terminating job step 14460607.0
srun: Force Terminated job step 14460607.0
**^ tin n0003.scs00 ~ ^**>  date
Thu Oct 18 20:44:47 PDT 2018
**^ tin n0003.scs00 ~ ^**>
```

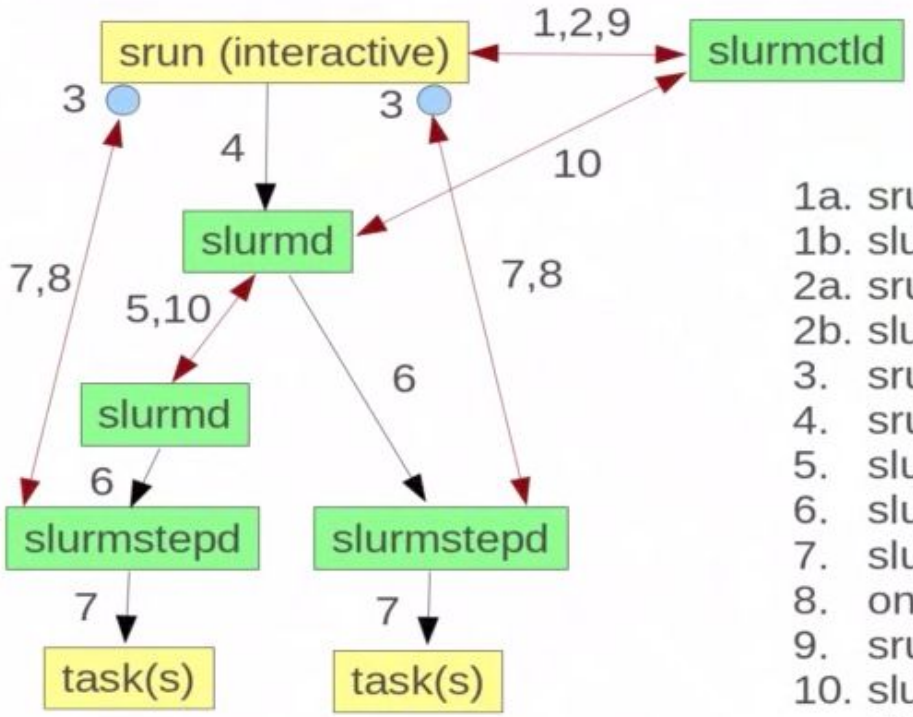
Interactive Job

fish /home/ti... | tin@n0003:~

```
**^ tin n0003.scs00 ~ ^**> srun --pty -p lr2 -A scs --qos=lr_normal -N 1
-t 0:0:05 /bin/bash
**^ tin n0124.lr2 ~ ^**> hostname
n0124.lr2
**^ tin n0124.lr2 ~ ^**> w
 20:43:18 up 59 days, 11:38,  0 users,  load average: 0.08, 0.03, 0.05
USER     TTY      FROM             LOGIN@   IDLE   JCPU   PCPU WHAT
**^ tin n0124.lr2 ~ ^**> date
Thu Oct 18 20:43:20 PDT 2018
**^ tin n0124.lr2 ~ ^**> srun: Force Terminated job 14460607
srun: error: n0124.lr2: task 0: Killed
srun: Terminating job step 14460607.0
srun: Force Terminated job step 14460607.0
**^ tin n0003.scs00 ~ ^**> date
Thu Oct 18 20:44:47 PDT 2018
**^ tin n0003.scs00 ~ ^**>
```

```
**^ tin n0003.scs00 ~ ^**> srun -p lr5 -A scs --qos=lr_debug -N 2 -t 5 bash
hostname
n0102.lr5
n0096.lr5
uptime
 21:05:01 up 58 days,  9:21,  0 users,  load average: 0.10, 0.09, 1.17
 21:05:01 up 58 days,  9:21,  0 users,  load average: 0.15, 0.12, 1.88
exit
**^ tin n0003.scs00 ~ ^**>
```

Job Execution Sequence

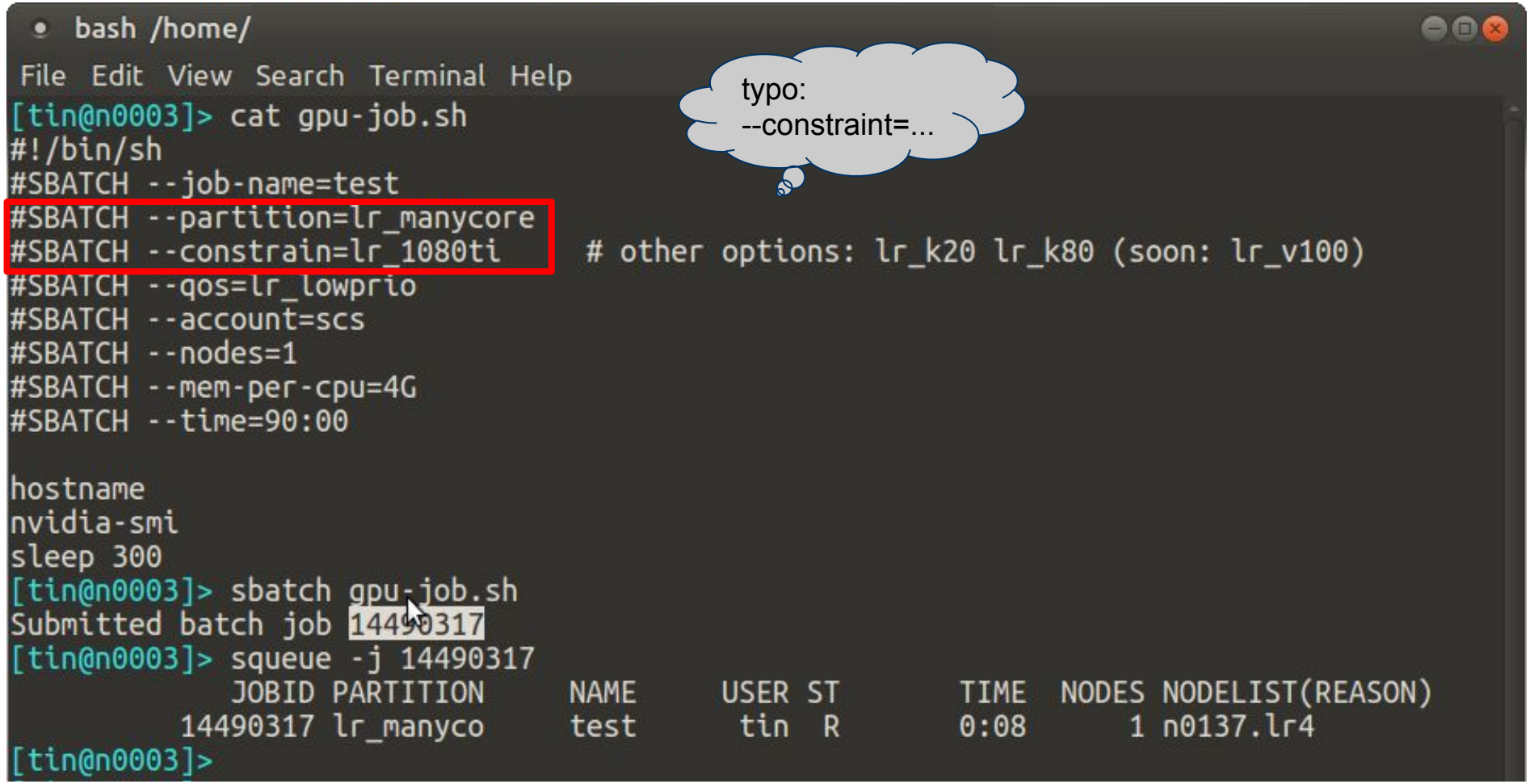


1a. srun sends job allocation request to slurmctld
1b. slurmctld grant allocation and returns details
2a. srun sends step create reqeust to slurmctld
2b. slurmctld responds with step credential
3. srun opens sockets for I/O
4. srun forwards credential with task info to slurmd
5. slurmd forward request as needed (per fanout)
6. slurmd forks/execs slurmstepd
7. slurmstepd connects I/O to run & launches tasks
8. on task termination, slurmstepd notifies srun
9. srun notifies slurmctld of job termination
10. slurmctld verifies termination of all processes via slurmd and releases resources for next job

SchedMD LLC
http://www.schedmd.com

slurm
workload manager

Requesting GPU



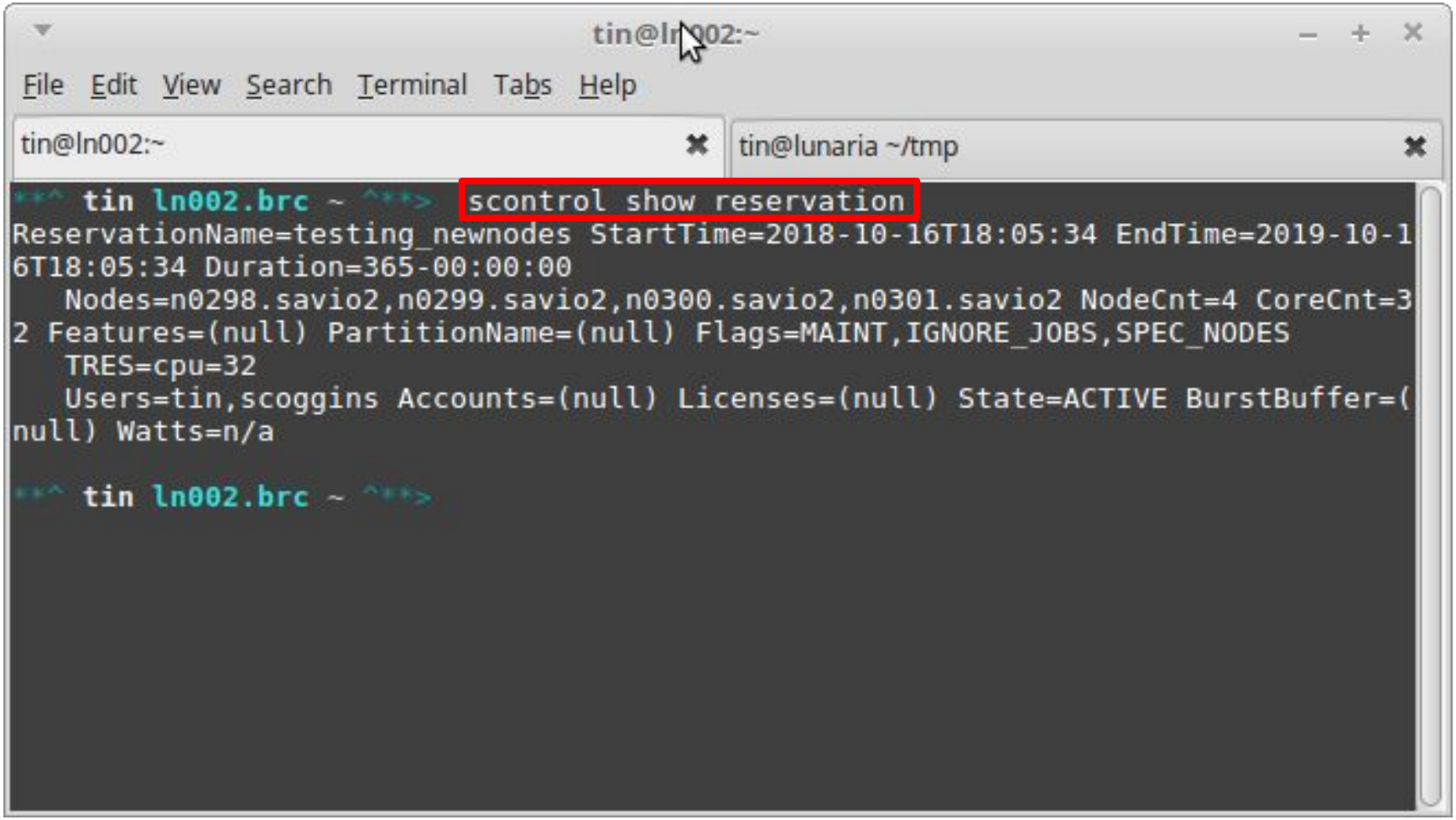
Requesting bigmem

```
bash /home/tin
File  Edit  View  Search  Terminal  Help
[tin@n0003]> cat bigmem-job.sh
#!/bin/sh
#SBATCH --job-name=bigmem_test
#SBATCH --partition=lr_bigmem
#SBATCH --qos=lr_normal
#SBATCH --account=scs
#SBATCH --nodes=1
#SBATCH --mem-per-cpu=16G
#SBATCH --time=90:00

hostname
free -h
sleep 300
[tin@n0003]> sbatch bigmem-job.sh
Submitted batch job 14491751
[tin@n0003]> cat slurm-14491751.out
n0156.lr2
              total        used        free      shared  buff/cache   available
Mem:           503G        3.7G        498G        1.2G        2.0G        497G
Swap:          3.8G        194M        3.6G
[tin@n0003]>
[tin@n0003]>
[tin@n0003]>
```

When will my job run?

```
bash /home/tin
File  Edit  View  Search  Terminal  Help
[tin@n0003]> squeue -u tin
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          14488010       lr6     test      tin PD       0:00     32 (PartitionNodeLimit)
          14488046       lr6     test      tin PD       0:00     16 (Resources)
          14488052       lr6     test      tin PD       0:00     32 (PartitionNodeLimit)
          14484776       lr6     test      tin  R      49:11      2 n0001.lr6,n0002.lr6
          14488039       lr6     test      tin  R       2:11      8 n0003.lr6,n0004.lr6,
n0005.lr6,n0006.lr6,n0007.lr6,n0008.lr6,n0009.lr6,n0010.lr6
          14488040       lr6     test      tin  R       2:11      8 n0011.lr6,n0012.lr6,
n0013.lr6,n0014.lr6,n0015.lr6,n0017.lr6,n0018.lr6,n0019.lr6
[tin@n0003]> date
Fri Oct 19 16:51:07 PDT 2018
[tin@n0003]> squeue --start -j 14488046
             JOBID PARTITION     NAME     USER ST          START_TIME  NODES SCHEDNODES
          NODELIST(REASON)
          14488046       lr6     test      tin PD 2018-10-19T19:18:49     16 n0001.lr6,n
0002.lr6, (Resources)
[tin@n0003]> squeue --start -j 14488052
             JOBID PARTITION     NAME     USER ST          START_TIME  NODES SCHEDNODES
          NODELIST(REASON)
          14488052       lr6     test      tin PD                 N/A     32 (null)
          (PartitionNodeLimit)
[tin@n0003]>
```

When will my job run?



sinfo -p lr_manycore
sprio

Cancel job

```
bash /home/tin
File  Edit  View  Search  Terminal  Help
[tin@n0003]> cat slurm-job.sh
#!/bin/sh
#SBATCH --job-name=test
#SBATCH --partition=lr6
#SBATCH --qos=lr_normal
#SBATCH --account=scs
#SBATCH --nodes=2
#SBATCH --mem-per-cpu=2G
#SBATCH --time=02:30:00

hostname
date
uptime
[tin@n0003]> sbatch -p lr5 --qos=lr_debug slurm-job.sh
Submitted batch job 14487581
[tin@n0003]> squeue -j 14487581
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          14487581       lr5     test      tin PD       0:00      2 (QOSMaxWallDurationPerJobLimit)
[tin@n0003]> scancel 14487581
[tin@n0003]>
```

All CPUs are charged

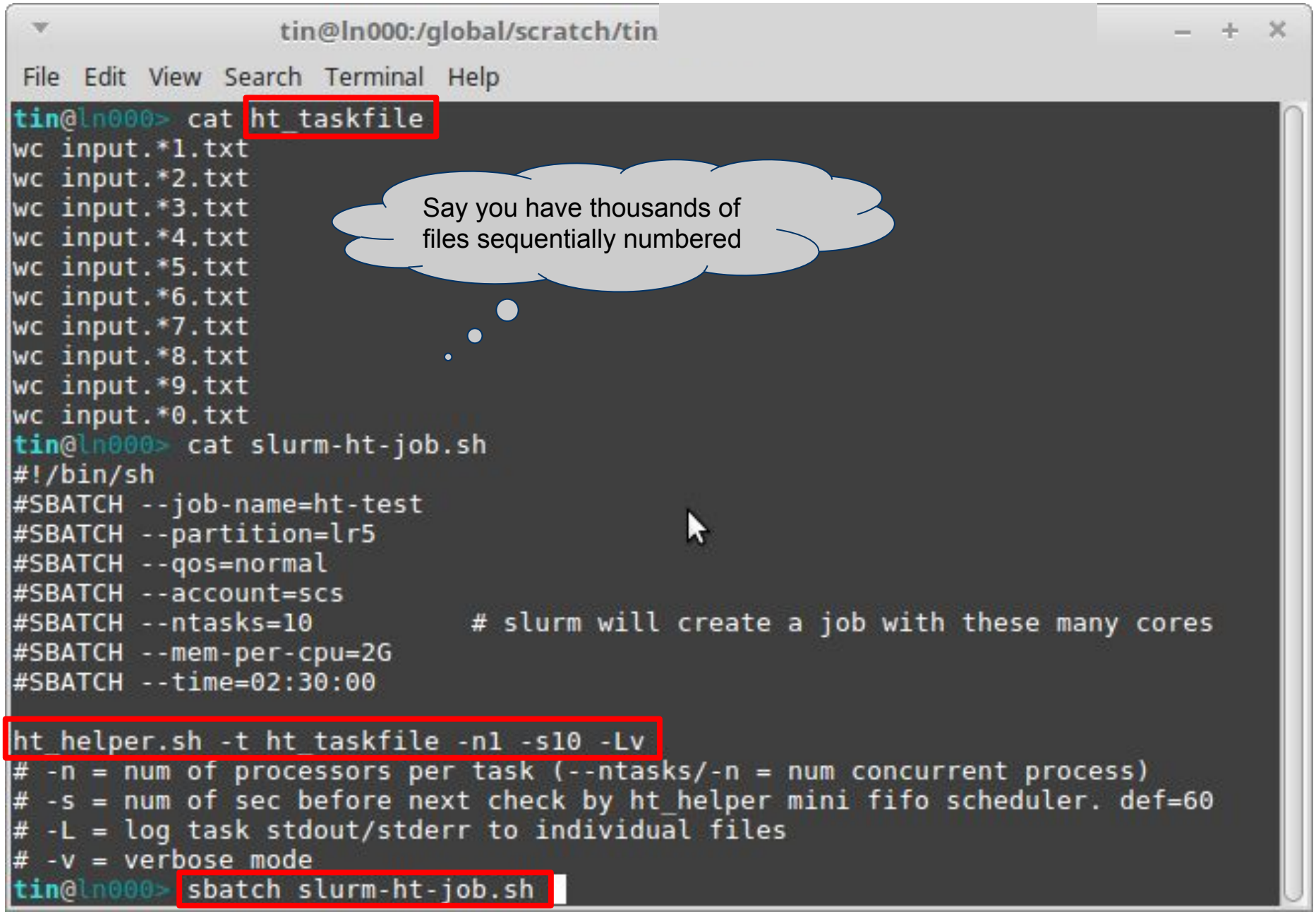
```
tin@n0003:~
File  Edit  View  Search  Terminal  Help
**^ tin n0003.scs00 ~ ^**>  scontrol show partitions=lr6
PartitionName=lr6
   AllowGroups=ALL AllowAccounts=ALL AllowQos=ALL
   AllocNodes=ALL Default=NO QoS=N/A
   DefaultTime=NONE DisableRootJobs=NO ExclusiveUser=NO GraceTime=0 Hidden=NO
   MaxNodes=UNLIMITED MaxTime=UNLIMITED MinNodes=1 LLN=NO MaxCPUsPerNode=UNLIMITED
   Nodes=n0[000-019].lr[6]
   PriorityJobFactor=1 PriorityTier=1 RootOnly=NO ReqResv=NO OverSubscribe=EXCLUSIVE
   OverTimeLimit=NONE PreemptMode=REQUEUE
   State=UP TotalCPUs=640 TotalNodes=20 SelectTypeParameters=NONE
   DefMemPerNode=93000 MaxMemPerNode=UNLIMITED

**^ tin n0003.scs00 ~ ^**>
```

All CPUs are charged

Partition	Nodes	Node List	CPU	Cores	Memory	Infiniband	Accelerator
lr5	144	n0[000-143].lr5 n0[148-171].lr5	INTEL XEON E5-2680 v4 INTEL Xeon ES-2640 v4	28 20	64GB 128GB	FDR QDR	-
lr_amd	4	n0[157-160].lr2	AMD OPTERON 6276	32	64GB	QDR	-
lr_bigmem	5	n0[154-155].lr2	AMD OPTERON 6174	48	256GB	QDR	-
		n0156.lr2	AMD OPTERON 6180 SE	48	512GB		
		n0210.lr3	INTEL XEON E5-4620	32	1024GB		
		n0211.lr3	INTEL XEON E7-4860 v2	48	1024GB		
lr_manycore	12	n0[204-207].lr3	INTEL XEON E5-2603	8	64GB	FDR	INTEL XEON PHI 7120
		n0[208-209].lr3	INTEL XEON E5-2603	8	64GB		NVIDIA KEPLER K20
		n0[096-098].lr4	INTEL XEON E5-2623 v3	8	64GB		4X NVIDIA KEPLER K80
		n0[136-138].lr4	INTEL XEON E5-2623 v3	8	64GB		4X NVIDIA GTX 1080TI
mako	272	n0[000-271].mako0	INTEL XEON E5530	8	24GB	QDR	-
mako_ manycore	4	n0[272-275].mako0	INTEL XEON X5650	12	24GB	QDR	2X NVIDIA TESLA C2050
cf1	20	n0[000-019].cf1	INTEL XEON Phi 7210	64	192GB	FDR	

Partition	Nodes	Node List	SU to Core CPU Hour Ratio	Effective Recharge Rate
lr2	198	n0[000-141].lr2 n0[146-153].lr2 n0[161-208].lr2	0.50	$0.0050 per Core CPU Hour
lr3	300	n0[000-203].lr3 n0[213-308].lr3	0.75	$0.0075 per Core CPU Hour
lr4	108	n0[000-095].lr4 n0[099-110].lr4	1.00	$0.0100 per Core CPU Hour
lr5	144	n0[000-143].lr5 n0[148-171].lr5	1.00	$0.0100 per Core CPU Hour
lr_amd	4	n0[157-160].lr2	0.50	$0.0050 per Core CPU Hour
lr_bigmem	5	n0[154-156].lr2 n0[210-211].lr3	0.75	$0.0075 per Core CPU Hour
lr_manycore	9	n0[204-209].lr3 n0[096-098].lr4 n0[136-138].lr4	1.00	$0.0100 per Core CPU Hour
mako	272	n0[000-271].mako0	0.50	$0.0050 per Core CPU Hour
mako_manycore	4	n0[272-275].mako0	0.50	$0.0050 per Core CPU Hour
cf1	20	n0[000-019].cf1	0.40	$0.0040 per Core CPU Hour

ht_helper - High Throughput Computing



Closing Reminder

Storage and Backup:

LAWRENCIUM cluster users are entitled to access the following storage systems so please get familiar with them.

Name	Location	Quota	Backup	Allocation	Description
HOME	/global/home/users/$USER	12GB	Yes	Per User	HOME directory for permanent data storage
GROUP-SW	/global/home/groups-sw/$GROUP	200GB	Yes	Per Group	GROUP directory for software and data sharing with backup
GROUP	/global/home/groups/$GROUP	400GB	No	Per Group	GROUP directory for data sharing without backup
SCRATCH	/global/scratch/$USER	none	No	Per User	SCRATCH directory with Lustre high performance parallel file system
CLUSTERFS	/clusterfs/axl/$USER	none	No	Per User	Private storage for AXL condo
CLUSTERFS	/clusterfs/cumulus/$USER	none	No	Per User	Private storage for CUMULUS condo
CLUSTERFS	/clusterfs/esd/$USER	none	No	Per User	Private storage for ESD condos
CLUSTERFS	/clusterfs/geoseq/$USER	none	No	Per User	Private storage for CO2SEQ condo
CLUSTERFS	/clusterfs/nokomis/$USER	none	No	Per User	Private storage for NOKOMIS condo

NOTE: HOME, GROUP, GROUP-SW and CLUSTERFS directories are located on a highly reliable enterprise level BlueArc storage device. Since this appliance also provides storage for many other mission critical file systems, and it is not designed for high performance applications, running large I/O dependent jobs on these file systems could greatly degrade the performance of all the file systems that are hosted on this device and affect hundreds of users, thus this behavior is explicitly prohibited. HPCS reserves the right to kill these jobs without notification once discovered. Jobs that have I/O requirement should use the SCRATCH file system which is designed specifically for that purpose.

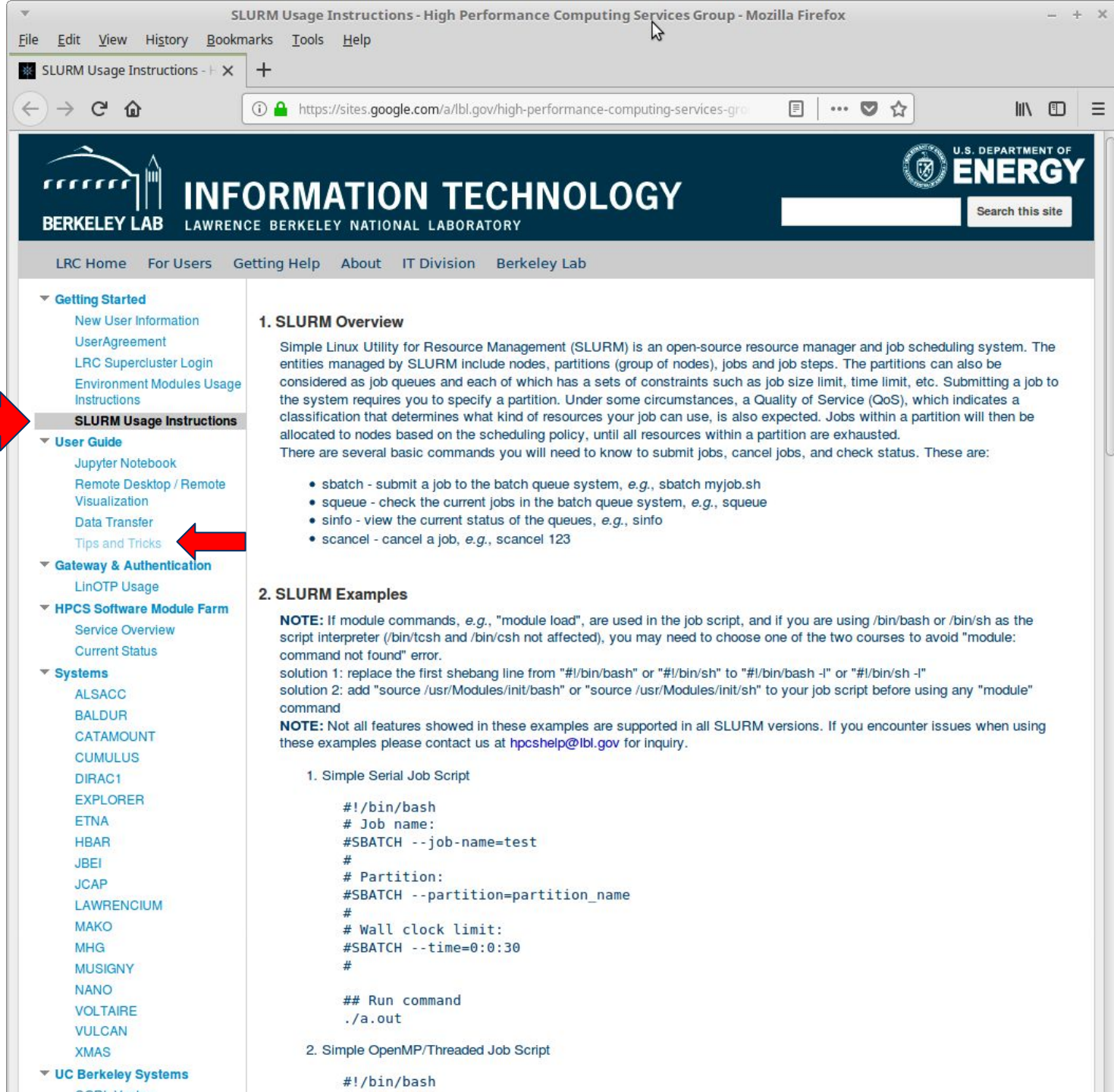
More Info

[lrc.lbl.gov](lrc.lbl.gov)

Laboratory Research Computing

- **LRC Home**
- HPC Systems
- Live Status
- For Users
- Getting an Account
- PI Computing Allowance
- Directorate Special Projects

http://tiny.cc/slurm



SCIENCE ACCELERATOR

U.S. DEPARTMENT OF ENERGY

Thank You!

tin (at) lbl.gov

http://tiny.cc/slurm

UNIVERSITY OF CALIFORNIA

Backup Slides

U.S. DEPARTMENT OF
ENERGY

UNIVERSITY OF
CALIFORNIA

How many SU are left?

[tin@n0003]> check_usage.sh

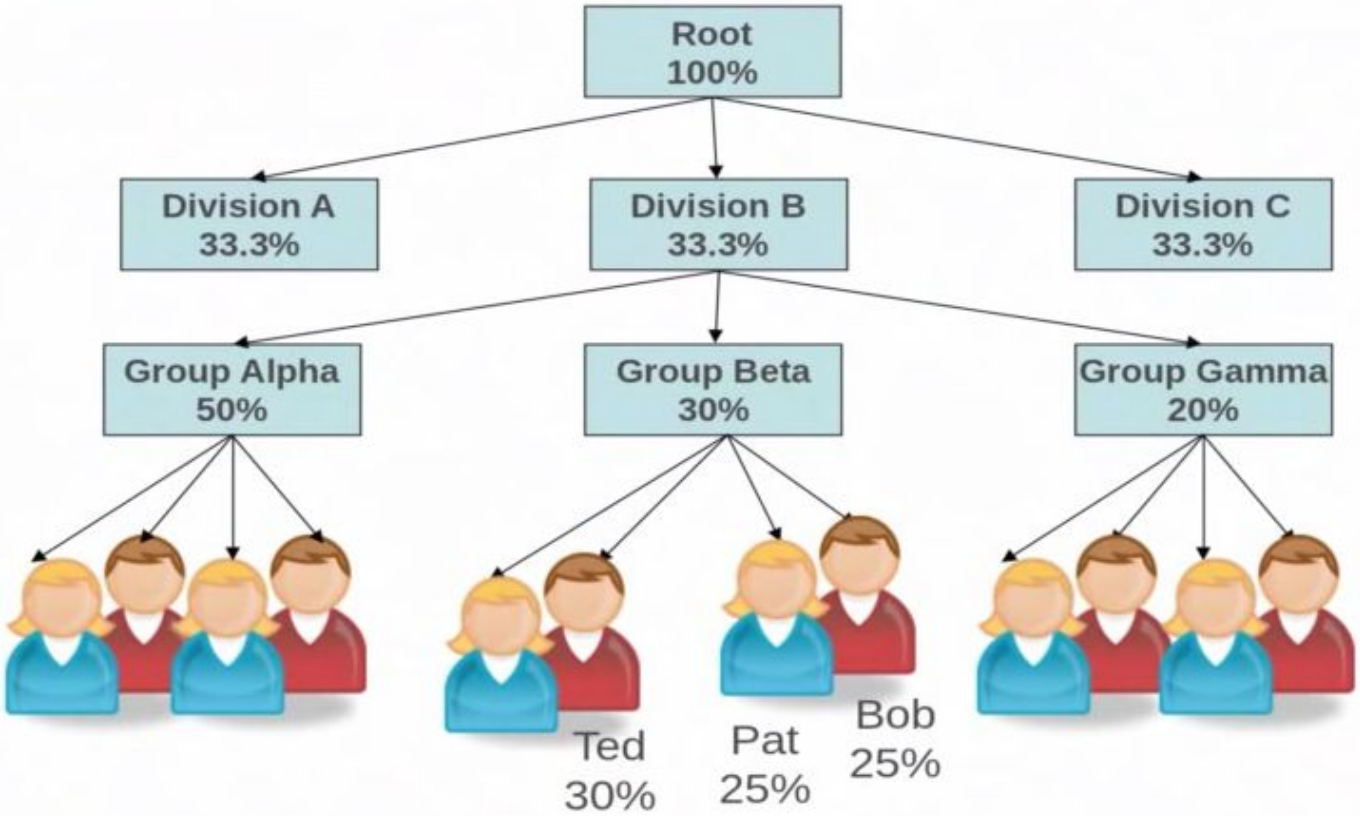
Usage for USER tin [2018-10-01T00:00:00, 2018-10-30T10:39:48]: 66 jobs, 1913.84 CPUHrs, 218.15 SUs

Exclusive vs Shared QoS

- Time Limit
- Node Limit

Partition	Nodes	Node List	Node Features	Shared	QoS	QoS Limit	Account
lr2	198	n0[000-141].lr2 n0[146-153].lr2 n0[161-180].lr2 n0[182-208].lr2	lr2	Exclusive	lr_normal	64 nodes max per job 72:00:00 wallclock limit	ac_* pc_*
					lr_debug	4 nodes max per job 4 nodes in total 00:30:00 wallclock limit	
					condo_co2seq	64 nodes max per group	lr_co2seq
					condo_cumulus	28 nodes max per group	lr_cumulus
		n0181.lr2	lr2 lr2_m96		condo_matgen	8 nodes max per group	lr_matgen
lr3	300	n0[000-163].lr3	lr3 lr3_c16	Exclusive	lr_normal	64 nodes max per job 72:00:00 wallclock limit	ac_* pc_*
					lr_debug	4 nodes max per job 4 nodes in total 00:30:00 wallclock limit	
					condo_axl	36 nodes max per group 30 nodes max per user	lr_axl
					condo_esd1	16 nodes max per group	lr_esd1
					condo_esd2	20 nodes max per group	lr_esd2
		n0[164-203].lr3 n0[213-308].lr3	lr3 lr3_c20		condo_nanotheory	4 nodes max per group	lr_nanotheory
					condo_nokomis	40 nodes max per group	lr_nokomis
lr4	108	n0[000-095].lr4 n0[099-110].lr4	lr4	Exclusive	lr_normal	64 nodes max per job 72:00:00 wallclock limit	ac_* pc_*
					lr_debug	4 nodes max per job 4 nodes in total 00:30:00 wallclock limit	
					condo_minnehaha	36 nodes max per group	lr_minnehaha
					condo_matminer	4 nodes max per group	lr_matminer
lr5	144	n0[000-143].lr5	lr5	Exclusive	lr_normal	64 nodes max per job 72:00:00 wallclock limit	ac_* pc_*
		n0[148-171].lr5	lr5_c20,lr5		lr_debug	4 nodes max per job 4 nodes in total 00:30:00 wallclock limit	
					condo_cedar		

Hierarchical bank example



slurm
workload manager

Common Problems and Solutions

Load all necessary modules (intel,openmpi,mkl) in .bashrc

Specify required memory with --mem-per-cpu

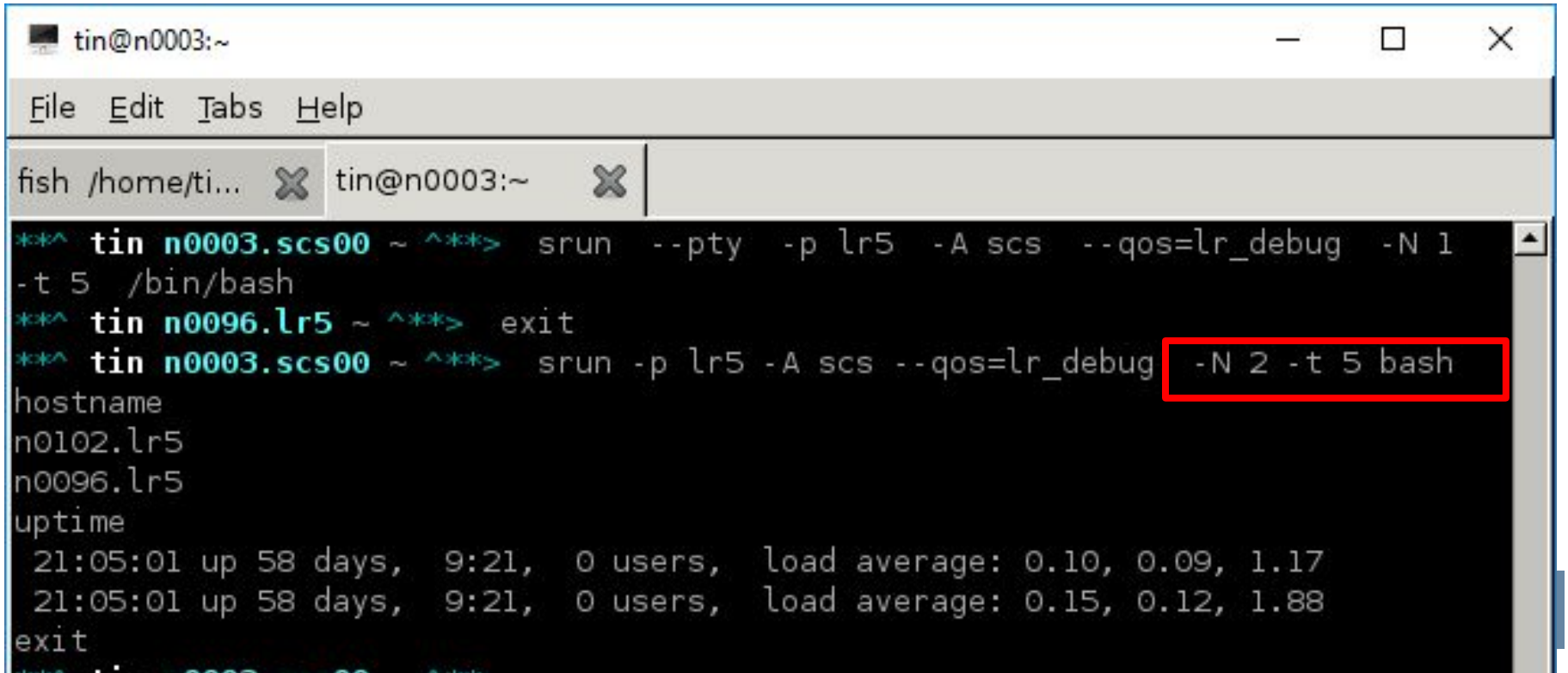
Process may get killed by OOM_killer.

wwall
wwstat

Multi-nodes interactive job

- funny shell

srun -p lr2 -A scs
--qos=lr_debug -N 2 -t 5 bash




```
tin@n0003:~
File  Edit  Tabs  Help
fish /home/ti...   tin@n0003:~
*** tin n0003.scs00 ~ ^***>  srun --label -p lr5 -A scs --qos=lr_debug  -N 2 -t 5
  bash
hostname
0: n0031.lr5
1: n0038.lr5
uptime
0:  21:15:17 up 59 days, 10:05,
1:  21:15:17 up 59 days, 10:05,
exit
*** tin n0003.scs00 ~ ^***>  echo $
0
*** tin n0003.scs00 ~ ^***>
*** tin n0003.scs00 ~ ^***>
*** tin n0003.scs00 ~ ^***>
*** tin n0003.scs00 ~ ^***>
*** tin n0003.scs00 ~ ^***>
*** tin n0003.scs00 ~ ^***>
*** tin n0003.scs00 ~ ^***>
*** tin n0003.scs00 ~ ^***>
*** tin n0003.scs00 ~ ^***>
*** tin n0003.scs00 ~ ^***>
*** tin n0003.scs00 ~ ^***>
*** tin n0003.scs00 ~ ^***>
```

```
tin@n0003:~
File  Edit  Tabs  Help
fish /home/ti...   tin@n0003:~
*** tin n0003.scs00 ~ ^***>
*** tin n0003.scs00 ~ ^***>  srun --label -p lr5 -A scs --qos=lr_debug  -N 2 -t 5
  bash -i
1: bash: cannot set terminal process group (-1): Inappropriate ioctl for device
1: bash: no job control in this shell
0: bash: cannot set terminal process group (-1): Inappropriate ioctl for device
0: bash: no job control in this shell

0: *** tin n0031.lr5 ~ ^***>
1: *** tin n0038.lr5 ~ ^***>
hostname
1: *** tin n0038.lr5 ~ ^***>  hostname
0: *** tin n0031.lr5 ~ ^***>  hostname
1: n0038.lr5
0: n0031.lr5
uptime
1: *** tin n0038.lr5 ~ ^***>  uptime
0: *** tin n0031.lr5 ~ ^***>  uptime
1:  21:17:09 up 59 days, 10:07,  0 users,  load average: 0.08, 0.12, 0.13
0:  21:17:09 up 59 days, 10:07,  0 users,  load average: 0.16, 0.12, 0.14

1: *** tin n0038.lr5 ~ ^***>
0: *** tin n0031.lr5 ~ ^***>
```

All CPUs are charged

Partition	Nodes	Node List	CPU	Cores	Memory	Infiniband	Accelerator
lr5	144	n0[000-143].lr5 n0[148-171].lr5	INTEL XEON E5-2680 v4 INTEL Xeon ES-2640 v4	28 20	64GB 128GB	FDR QDR	-
lr_amd	4	n0[157-160].lr2	AMD OPTERON 6276	32	64GB	QDR	-
lr_bigmem	5	n0[154-155].lr2	AMD OPTERON 6174	48	256GB	QDR	-
		n0156.lr2	AMD OPTERON 6180 SE	48	512GB		
		n0210.lr3	INTEL XEON E5-4620	32	1024GB		
		n0211.lr3	INTEL XEON E7-4860 v2	48	1024GB		
lr_manycore	12	n0[204-207].lr3	INTEL XEON E5-2603	8	64GB	FDR	INTEL XEON PHI 7120
		n0[208-209].lr3	INTEL XEON E5-2603	8	64GB		NVIDIA KEPLER K20
		n0[096-098].lr4	INTEL XEON E5-2623 v3	8	64GB		4X NVIDIA KEPLER K80
		n0[136-138].lr4	INTEL XEON E5-2623 v3	8	64GB		4X NVIDIA GTX 1080TI
mako	272	n0[000-271].mako0	INTEL XEON E5530	8	24GB	QDR	-
mako_ manycore	4	n0[272-275].mako0	INTEL XEON X5650	12	24GB	QDR	2X NVIDIA TESLA C2050
cf1	20	n0[000-019].cf1	INTEL XEON Phi 7210	64	192GB	FDR	

Partition	Nodes	Node List	SU to Core CPU Hour Ratio	Effective Recharge Rate

Partition	Nodes	Node List	SU to Core CPU Hour Ratio	Effective Recharge Rate
lr5	144	n0[000-143].lr5 n0[148-171].lr5	1.00	$0.0100 per Core CPU Hour
lr_amd	4	n0[157-160].lr2	0.50	$0.0050 per Core CPU Hour
lr_bigmem	5	n0[154-156].lr2 n0[210-211].lr3	0.75	$0.0075 per Core CPU Hour
lr_manycore	9	n0[204-209].lr3 n0[096-098].lr4 n0[136-138].lr4	1.00	$0.0100 per Core CPU Hour
cf1	20	n0[000-019].cf1	0.40	$0.0040 per Core CPU Hour

What account and QoS do I use?

QoS = Quality of Service = Priority
Account = Project code for Service Unit tracking

```
> sacctmgr show associations user=ljin format=acc,part,qos

Account       Partition      QOS
ac_seasonal   cf1            cf_debug,cf_normal
ac_seasonal   lr5            lr_debug,lr_lowprio,lr_normal
ac_seasonal   lr4            lr_debug,lr_lowprio,lr_normal
ac_seasonal   lr3            lr_debug,lr_lowprio,lr_normal
ac_seasonal   lr2            lr_debug,lr_lowprio,lr_normal
ac_seasonal   lr_manycore    lr_lowprio,lr_normal
ac_seasonal   lr_bigmem      lr_lowprio,lr_normal
ac_seasonal   lr_amd         lr_lowprio,lr_normal
ac_seasonal   mako           mako_debug,mako_lowprio,mako_normal
ac_seasonal   mako_manycore  mako_lowprio,mako_normal
```

What account and QoS do I use?

QoS = Quality of Service = Priority

```
> sacctmgr show associations -p  user=tin format=acc,part,qos \
| sed 's/|/\t\t/g'

Account    Partition       QOS

scs        lr6             lr_debug,lr_lowprio,lr_normal
scs        cf1-hp          condo_mp
scs        cf1             cf_debug,cf_normal
scs        lr5             lr_debug,lr_lowprio,lr_normal
scs        mako_manycore   mako_lowprio,mako_normal
scs        mako            mako_debug,mako_lowprio,mako_normal
scs        lr_manycore     lr_lowprio,lr_normal
scs        lr_bigmem       lr_lowprio,lr_normal
scs        lr_amd          lr_lowprio,lr_normal
scs        lr4             lr_debug,lr_lowprio,lr_normal
...
```

Pocket Survival Guide - HP | LAWRENCIUM - High Perfo | The Warewulf Cluster Mon

Not secure | metacluster.lbl.gov/warewulf

Apps | bookmark | UC Berkeley Events C | lbl18 | lbl-gdoc | gdoc | sn18 | ucb_18 | Other bookmarks

lr3

lr4

(End of backup slides)