

Advanced Cluster Usage

FAS Research Computing

Outline

- Job Submission
- Job Resource Requirements
- Job/Partition/Queue Monitoring
- Job Checkpointing
- Fairshare
- Storage Workflow



Job Submission - Interactive








- `salloc -p test --mem=4G -t 0-01:00`
 - Gives back a shell prompt on a compute node
 - Uses
 - Testing code
 - Working interactively on the cluster without resource contention
 - Limitations
 - Session stall
 - Ties up prompt
 - Not great for GUI applications
 - If submitting to a busy partition `salloc` may take a while to respond

Job Submission - Interactive

OnDemand provides an integrated, single access point for all of your HPC resources.

Pinned Apps A featured subset of all available apps

Interactive Apps

 Jupyter notebook / Jupyterlab System Installed App	 Matlab System Installed App	 Postgresql db System Installed App	 RStudio Server System Installed App
 Remote Desktop System Installed App	 SAS System Installed App	 Stata System Installed App	

Welcome to FAS-RC Cluster

The Computing Cluster is a resource for the research community, hosted by Research Computing at Harvard University's Faculty of Arts and Sciences.

To apply for an account please refer to [this webpage](#).

From this web service you can submit your jobs, check running jobs, and open interactive graphical sessions to run your favorite applications.

These are some examples of the things you will be able to do :

- Open an interactive remote desktop session to a compute node
- Run Jupyter Notebooks
- Run Rstudio Server sessions
- Browse and edit your files
- Open a terminal connection to a login node

Check out our documentation at [this page](#).

<https://vdi.rc.fas.harvard.edu> (VPN Required)

Job Submission - Interactive

Home / My Interactive Sessions / Remote Desktop

Interactive Apps
Desktop Apps
Matlab
SAS
Stata
Desktops
Containerized FAS-RC Remote Desktop
Remote Desktop
FAS CGA
Postgresql db
Web Apps
HeavyAI
Jupyter notebook / Jupyterlab
RStudio Server

Remote Desktop version: 3066999

This app will launch an interactive desktop session on a compute node in a partition of your choice. This app is useful to have graphical user interface (GUI) on the FAS-RC cluster.

You can select the amount of RAM (in GB), number of cores, and Timelimit (in hrs).

Upon request you can receive email notification when the job starts. You must specify your email address.

See [Remote Desktop VDI app documentation](#) for how to use the app, update the web browser within the app, copy and paste between your computer and the remote desktop app.

See [How to launch software](#) for how to launch various software, like Abaqus, Comsol, ParaView, etc.

Resolution

width 1024 px height 768 px

Reset Resolution

Partition

test

`sbatch -p, --partition=<partition_names>`

Slurm partition name (e.g., **shared**), or comma-separated list of partition names (e.g., **shared,test**)

Memory Allocation in GB

8

Number of cores

2

Number of Cpus to allocate

Number of GPUs

0

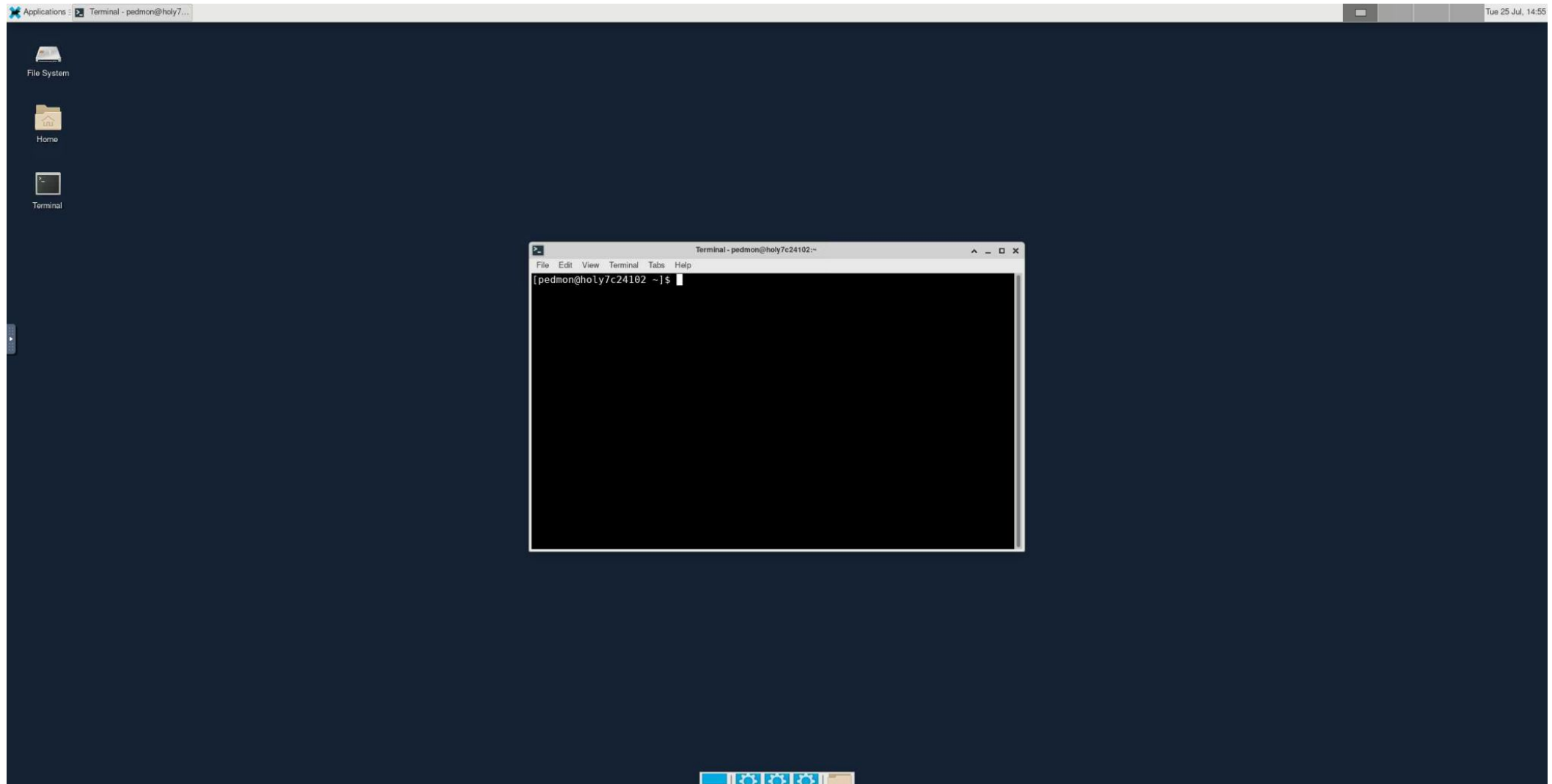
Number of GPUs to allocate. Available only on GPU enabled partitions

Allocated Time (expressed in MM, or HH:MM:SS, or DD-HH:MM)

04:00:00

`sbatch -t, --time=<time>`

Job Submission - Interactive



Job Submission - Batch

```
#!/bin/bash
#SBATCH -J hybridtest
#SBATCH -n 32
#SBATCH -c 4
#SBATCH -p shared
#SBATCH -t 1-12:00:00
#SBATCH --mem-per-cpu=8G
#SBATCH -o hybrid_%A.out
#SBATCH -o hybrid_%A.err

module load intel/23.0.0-fasrc01 openmpi/4.1.4-fasrc01

srun -c $SLURM_CPUS_PER_TASK -n $SLURM_NTASKS --mpi=pmix ./wombat.x
```

sbatch runscript.slurm

Submits list of instructions and commands as a script to the scheduler

Does not require an open prompt

Types

- Serial
- Array (--array)
- Thread
- Rank
- Hybrid

Useful Options (not exhaustive)

- contiguous
- constraint/--prefer
- dependency
- exclusive[={user | mcs}]
- gpu/--gres



Job Resource Requirements

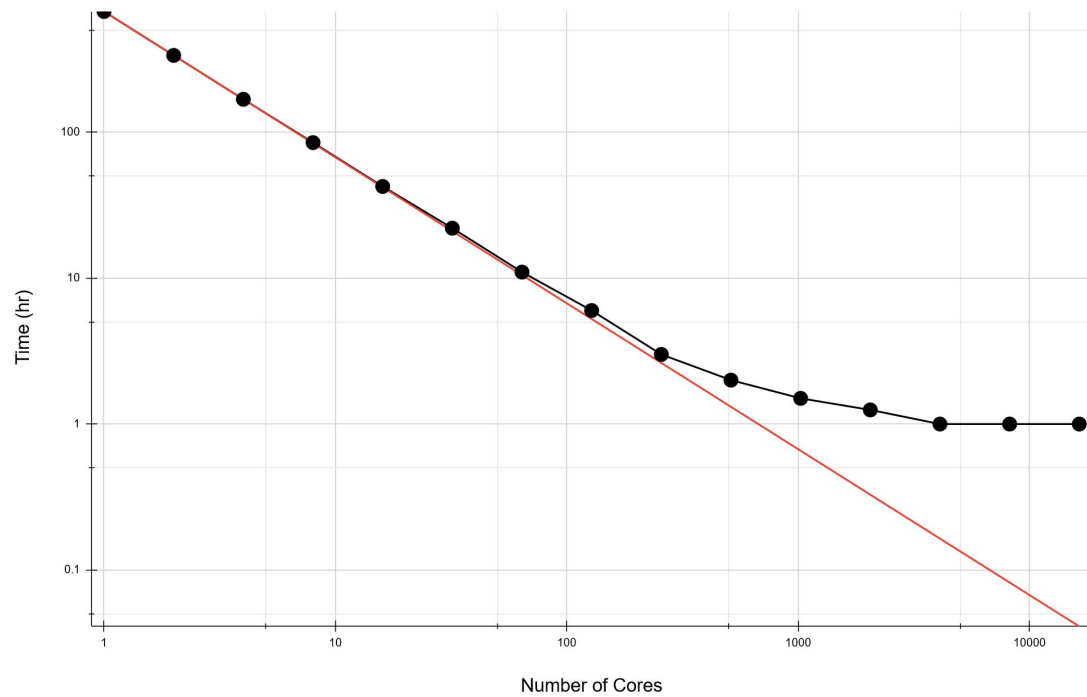
```
[user@boslogin01 home]# seff 1234567
Job ID: 1234567
Cluster: odyssey
User/Group: user/user_lab
State: COMPLETED (exit code 0)
Nodes: 8
Cores per node: 64
CPU Utilized: 37-06:17:33
CPU Efficiency: 23.94% of 155-16:02:08 core-walltime
Job Wall-clock time: 07:17:49
Memory Utilized: 1.53 TB (estimated maximum)
Memory Efficiency: 100.03% of 1.53 TB (195.31 GB/node)
```

- Know your Code
 - Numerical Methods
 - Size of Data
 - Type of Parallelism
- Experimentation
 - Validate Memory Size Requirements
 - Scaling Tests
 - Profiling
 - sstat vs. sacct
- Select Appropriate Partition and Hardware

Job Resource Requirements - Scaling

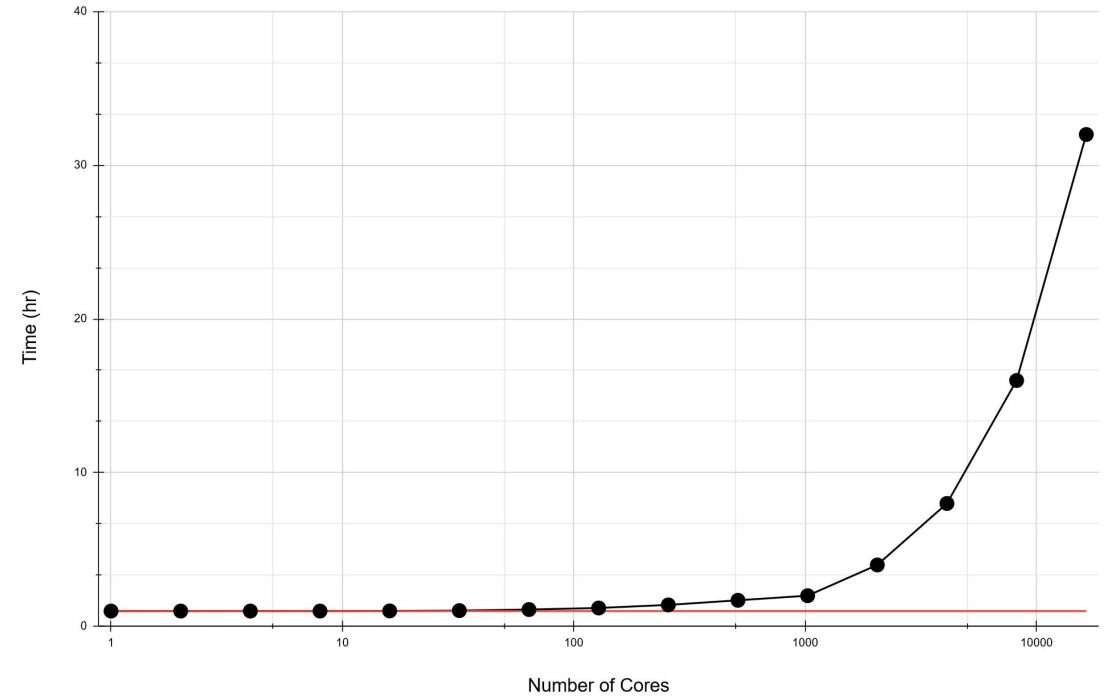
Strong Scaling

Size of Computation Constant While Number of Cores Increases (log-log)



Weak Scaling

Size of Computation Grows as Number of Cores Increases (log-linear)





Job Monitoring - sacct

```
[root@holy7c22501 general]# sacct -u mngo
JobID      JobName Partition      Account AllocCPUS   State ExitCode
-----
63911611   train_liif  gpu pfister_l+      14      TIMEOUT    0:0
63911611.ba+ batch      pfister_l+         14 CANCELLED 0:15
63911611.ex+ extern     pfister_l+         14 COMPLETED 0:0
63911755   train_liif  gpu pfister_l+      14      TIMEOUT    0:0
63911755.ba+ batch      pfister_l+         14 CANCELLED 0:15
63911755.ex+ extern     pfister_l+         14 COMPLETED 0:0
63971094   train_liif  gpu pfister_l+      14      COMPLETED 0:0
63971094.ba+ batch      pfister_l+         14 COMPLETED 0:0
63971094.ex+ extern     pfister_l+         14 COMPLETED 0:0
64063319   train_liif  gpu pfister_l+      14      RUNNING    0:0
64063319.ba+ batch      pfister_l+         14      RUNNING    0:0
64063319.ex+ extern     pfister_l+         14      RUNNING    0:0
64063323   train_liif  gpu pfister_l+      14      RUNNING    0:0
64063323.ba+ batch      pfister_l+         14      RUNNING    0:0
64063323.ex+ extern     pfister_l+         14      RUNNING    0:0
64063331   train_liif  gpu pfister_l+      14      RUNNING    0:0
64063331.ba+ batch      pfister_l+         14      RUNNING    0:0
64063331.ex+ extern     pfister_l+         14      RUNNING    0:0
64078487   train_liif  gpu pfister_l+      14      RUNNING    0:0
64078487.ba+ batch      pfister_l+         14      RUNNING    0:0
64078487.ex+ extern     pfister_l+         14      RUNNING    0:0
64078502   train_liif  gpu pfister_l+      14      RUNNING    0:0
64078502.ba+ batch      pfister_l+         14      RUNNING    0:0
64078502.ex+ extern     pfister_l+         14      RUNNING    0:0
```

- Default shows data from last day
- Options
 - --starttime/--endtime
 - --format
 - --parsable2
 - --partition
 - --state

Job Monitoring - scontrol

```
[root@holy7c22501 general]# scontrol show job 64063319
JobId=64063319 JobName=train_liif
  UserId=mngo(63096) GroupId=pfister_lab(40134) MCS_label=N/A
  Priority=17583 Nice=0 Account=pfister_lab QOS=normal
  JobState=RUNNING Reason=None Dependency=(null)
  Requeue=1 Restarts=0 BatchFlag=1 Reboot=0 ExitCode=0:0
  RunTime=1-21:52:05 TimeLimit=3-00:00:00 TimeMin=N/A
  SubmitTime=2023-07-25T13:45:18 EligibleTime=2023-07-25T13:45:18
  AccrueTime=2023-07-25T13:45:18
  StartTime=2023-07-25T13:45:21 EndTime=2023-07-28T13:45:21 Deadline=N/A
  SuspendTime=None SecsPreSuspend=0 LastSchedEval=2023-07-25T13:45:21 Scheduler=Main
  Partition=gpu AllocNode:Sid=0.0.0.0:2677159
  ReqNodeList=(null) ExcNodeList=(null)
  NodeList=holygpu7c26101
  BatchHost=holygpu7c26101
  NumNodes=1 NumCPUs=14 NumTasks=1 CPUs/Task=14 ReqB:S:C:T=0:0:*:*
  TRES=cpu=14,mem=490G,node=1,billing=926,gres/gpu=4,gres/gpu:nvidia_a100-sxm4-40gb=4
  Socks/Node=* NtasksPerN:B:S:C=0:0:*:* CoreSpec=*
  MinCPUsNode=14 MinMemoryCPU=35G MinTmpDiskNode=0
  Features=(null) DelayBoot=00:00:00
  OverSubscribe=OK Contiguous=0 Licenses=(null) Network=(null)
  Command=/n/holyifs05/LABS/pfister_lab/Lab/coxf01/pfister_lab2/Lab/mngo/vu-master-thesis/liif/slurm/job_train_iter.sh
  WorkDir=/n/holyifs05/LABS/pfister_lab/Lab/coxf01/pfister_lab2/Lab/mngo/vu-master-thesis/liif
  StdErr=/n/holyifs05/LABS/pfister_lab/Lab/coxf01/pfister_lab2/Lab/mngo/vu-master-thesis/liif/slurm/outputs/myerrors_64063319.err
  StdIn=/dev/null
  StdOut=/n/holyifs05/LABS/pfister_lab/Lab/coxf01/pfister_lab2/Lab/mngo/vu-master-thesis/liif/slurm/outputs/myoutput_64063319.out
  Power=
  MemPerTres=gpu:100
  TresPerNode=gres:gpu:4
```



Partition Monitoring - showq

```
[root@holy7c22501 general]# showq -p intermediate -o
```

```
SUMMARY OF JOBS FOR QUEUE: <intermediate>
```

```
ACTIVE JOBS-----
```

JOBID	JOBNAME	USERNAME	STATE	CORE	GPU	REMAINING	STARTTIME
60897569	cm_afm.sh	joonholee	Running	48	0	167:46:27	Thu Jul 20 11:29:20
60897570	cm_afm.sh	joonholee	Running	48	0	196:29:26	Fri Jul 21 16:12:19
60897572	cm_afm.sh	joonholee	Running	48	0	247:54:04	Sun Jul 23 19:36:57
61583962	skin-WS_sy	csxue	Running	1	0	21:50:42	Fri Jul 21 09:33:35
62497782	.fasrcood/	verstyuk	Running	1	0	24:49:17	Fri Jul 14 17:32:10
62497784	.fasrcood/	verstyuk	Running	1	0	24:49:59	Fri Jul 14 17:32:52
63128424	sbatch	dverbart	Running	64	0	63:18:07	Sat Jul 22 03:01:00
63128430	sbatch	dverbart	Running	64	0	123:59:12	Mon Jul 24 15:42:05
63268520	sbatch	dverbart	Running	64	0	144:02:24	Tue Jul 25 11:45:17
63884205	doParallel	cdadams	Running	1	0	27:44:41	Tue Jul 25 07:27:34
63885162	BAYES2_N30	agarciasoto	Running	10	0	288:14:43	Tue Jul 25 11:57:36
63885172	BAYES2_N30	agarciasoto	Running	10	0	302:55:15	Wed Jul 26 02:38:08
63885194	BAYES2_N30	agarciasoto	Running	10	0	313:14:06	Wed Jul 26 12:56:59
63885224	BAYES2_N30	agarciasoto	Running	10	0	323:24:06	Wed Jul 26 23:06:59
63885226	BAYES2_N30	agarciasoto	Running	10	0	323:25:14	Wed Jul 26 23:08:07

```
28 active jobs : 539 of 576 cores ( 93.58 %) : 0 of 0 gpus ( 0.00 %) : 12 of 12 nodes (100.00 %)
```

```
WAITING JOBS-----
```

JOBID	JOBNAME	USERNAME	STATE	CORE	GPU	WCLIMIT	QUEUE TIME
62277463	S110-6K	yrupan	Waiting	40	0	336:00:00	Thu Jul 13 03:45:48
63885231	BAYES2_N30	agarciasoto	Waiting	10	0	336:00:00	Sun Jul 23 11:23:48
63885232	BAYES2_N30	agarciasoto	Waiting	10	0	336:00:00	Sun Jul 23 11:23:48
63885233	BAYES2_N30	agarciasoto	Waiting	10	0	336:00:00	Sun Jul 23 11:23:48
63885234	BAYES2_N30	agarciasoto	Waiting	10	0	336:00:00	Sun Jul 23 11:23:48
63885235	BAYES2_N30	agarciasoto	Waiting	10	0	336:00:00	Sun Jul 23 11:23:48
63885236	BAYES2_N30	agarciasoto	Waiting	10	0	336:00:00	Sun Jul 23 11:23:49
63885237	BAYES2_N30	agarciasoto	Waiting	10	0	336:00:00	Sun Jul 23 11:23:49
63885238	BAYES2_N30	agarciasoto	Waiting	10	0	336:00:00	Sun Jul 23 11:23:49

- Shows queue state

- Options

- -p: partition
- -o: order by priority
- -U: username
- -s: only summary information

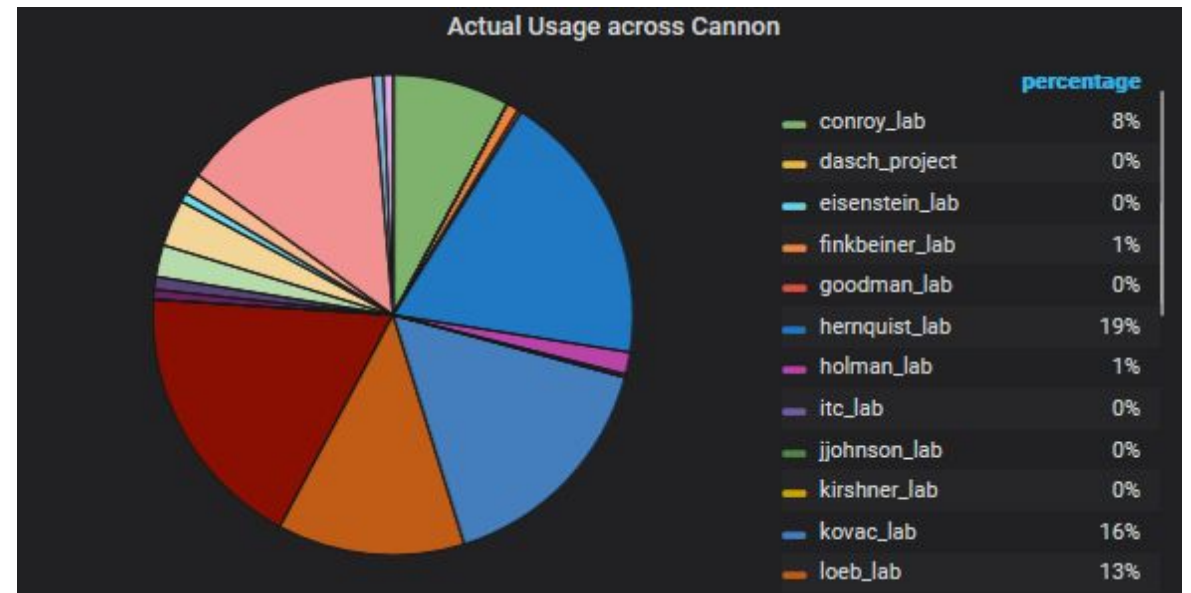
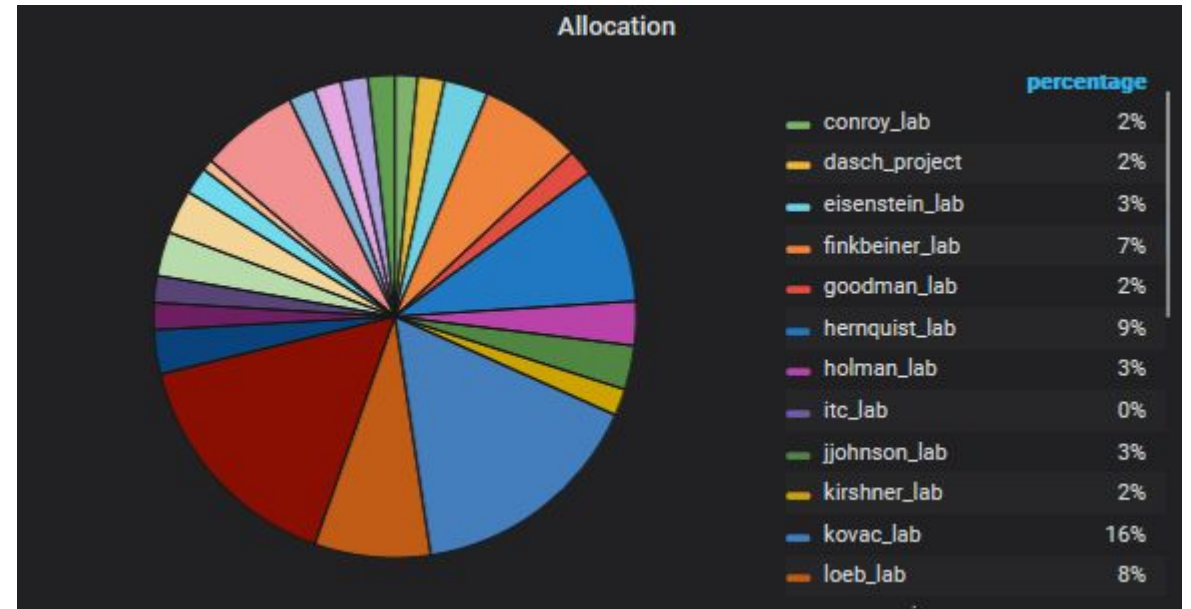


Job Checkpointing

- Creates a save point for your job to pick up from where it left off
 - Also Known As: Checkpointing, Save File, Restart File
- Useful for:
 - Long running jobs
 - Jobs that error out
 - Jobs that need midstream tweaking
 - Leveraging requeue partitions
- How?
 - Build it into your code
 - DMTCP: Distributed MultiThreaded Checkpointing
 - Leverage --dependency
 - Make code aware to check for checkpoint when requeued

Fairshare

1. A method for ensuring the equitable use of a cluster
2. The fraction of the cluster a user/group gets
3. The score assigned by Slurm to a user/group based on usage
4. Priority that users/groups get based on usage



Fairshare - sshare

```
[user1@holyitc01 ~]$ sshare --account=test_lab -a
Account  User  RawShares NormShares RawUsage  EffectvUsage FairShare
-----
test_lab      244      0.001363  45566082    0.000572    0.747627
test_lab user1 parent  0.001363  82028750.000572    0.747627
test_lab user2 parent  0.001363  248820 0.000572    0.747627
test_lab user3 parent  0.001363  163318 0.000572    0.747627
test_lab user4 parent  0.001363  18901027    0.000572    0.747627
test_lab user5 parent  0.001363  18050039    0.000572    0.747627
```

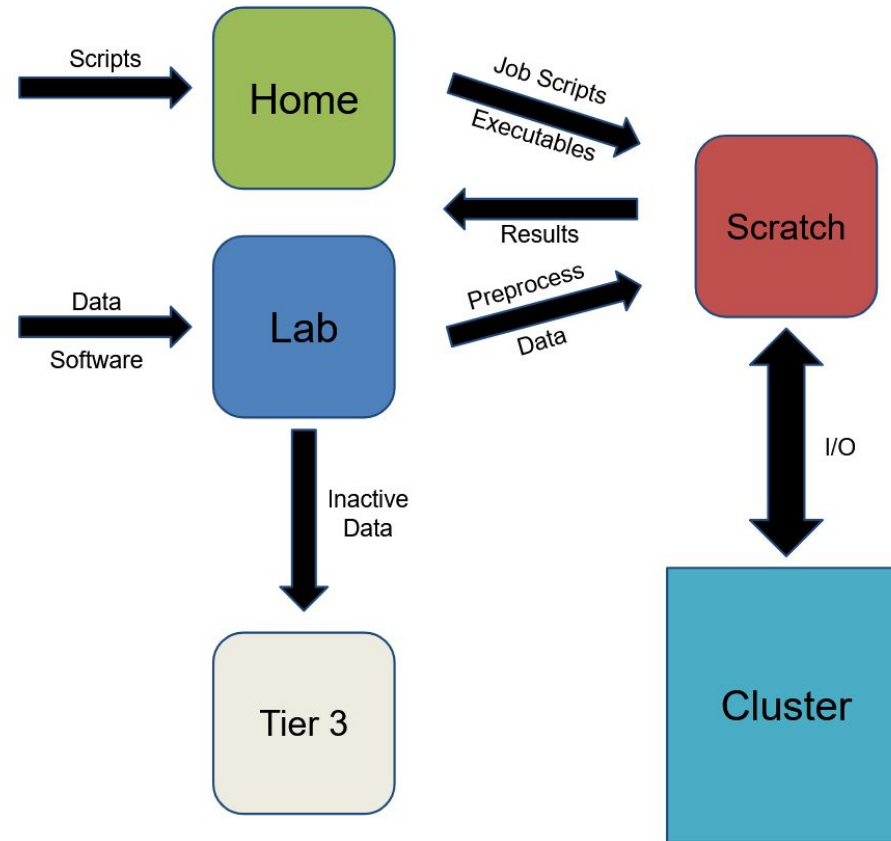
- Default Raw Shares
 - Cannon: 120
 - FASSE: 100
- Fairshare Regimes:
 - $f = 1$: Unused
 - $1.0 > f > 0.5$: Underutilized
 - 0.5: Average utilization
 - $0.5 > f > 0$: Over-utilized
 - $f = 0$: No share left



Fairshare - scalc

```
[root@holy7c22501 ~]# scalc
What do you want to calculate?
1) Projected FairShare Based on New RawShare
2) Additional RawShare Need for FairShare Score
3) Projected Time to Reach FairShare Score Assuming No New Jobs
4) Projected Usage and Fairshare Based on Job
5) Calculate New RawShare Based on Additional Hardware
Option: 4
4
We will now calculate how much TRES your jobs will cost as well as how it will impact the specified account's usage and fairshare.
First we need to know what account you want to calculate for: rc_admin
Next we need the partition you want to submit to: shared
How many cores will you use per job: 1024
How much memory in GB will you use per job: 4000
How many total GPUs will you user per job: 0
How long will the job run for (DD-HH:MM:SS): 1-00:00:00
How many jobs (or array elements) will you run of this type: 1
rc_admin has a current Raw Usage of 9725230 a Normalized Usage of 0.000026 a Normalized Allocation of 0.000759 and Fairshare of 0.976085
This partition has a TRES charge per second of CPU: 1.0 | Mem (per GB): 0.25 | GPU (per GPU): 0
This set of jobs has a total TRES usage of: 174873600.0
For rc_admin this will give a new Normalized Usage of 0.0004935173337802807 and a Fairshare of 0.6371829348127839
```


Storage Workflow



Questions, Comments, Concerns?