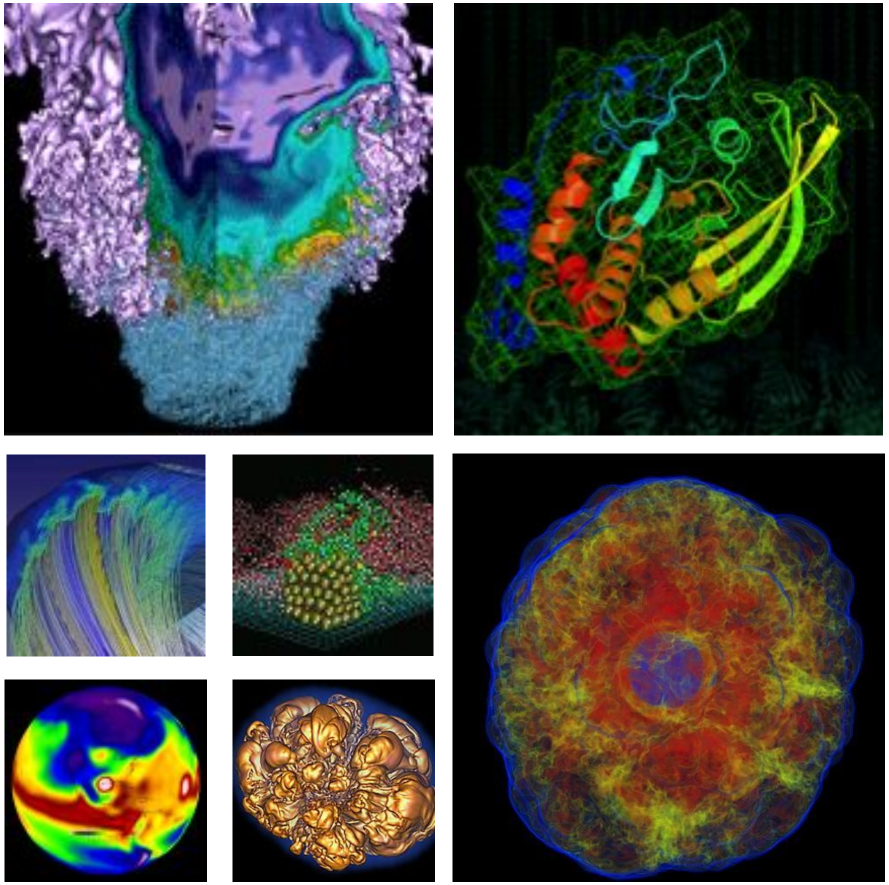
NERSC File Systems

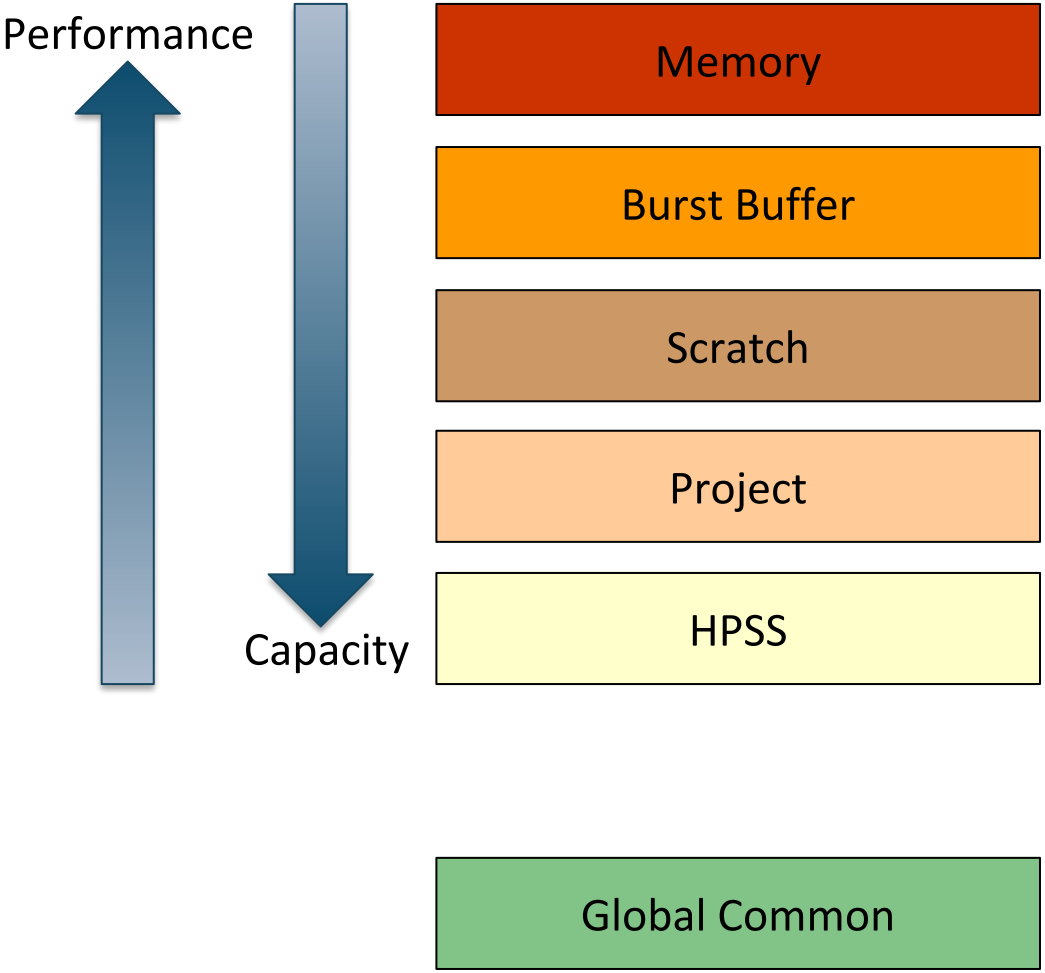


Jialin Liu
Data and Analytics Group

NERSC New User Training

June 21 2019

Simplified NERSC File Systems



1.8 PB SSD Burst Buffer on Cori
Cray Datawarp
1.8 TB/s,
temporary for job or campaign

28 PB (Cori) HDD Scratch
Lustre
700 GB/s,
temporary (12 wk purge)

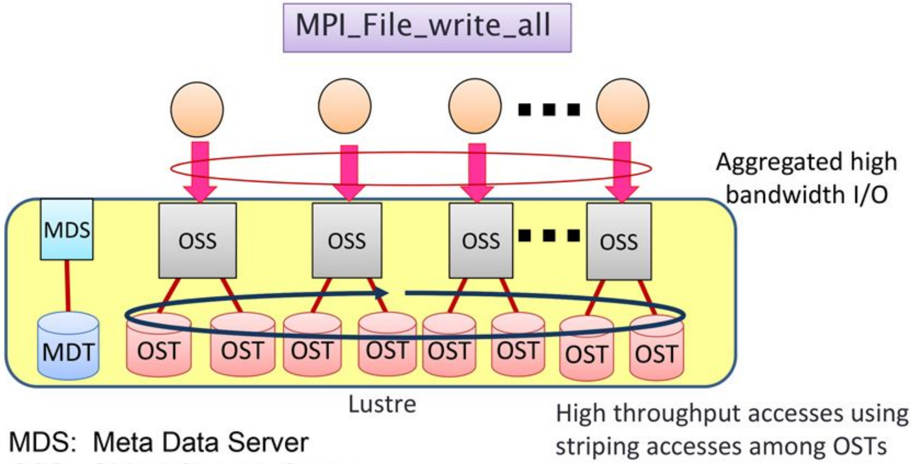
10.7 PB HDD and SSD Project
Spectrum Scale (GPFS)
150 GB/s, permanent

150 PB Tape Archive
HPSS
Forever

20 TB SSD Software
Spectrum Scale
Permanent,
read-only on computes

Where Do I Put My Data? Scratch

- **Scratch**
  - ***Lustre***, one of the most successful/mature HPC FS
  - 16 year research development
  - 60/100 of the top 100 fastest supercomputers in the world [1]
  - v2.7 (v2.12)

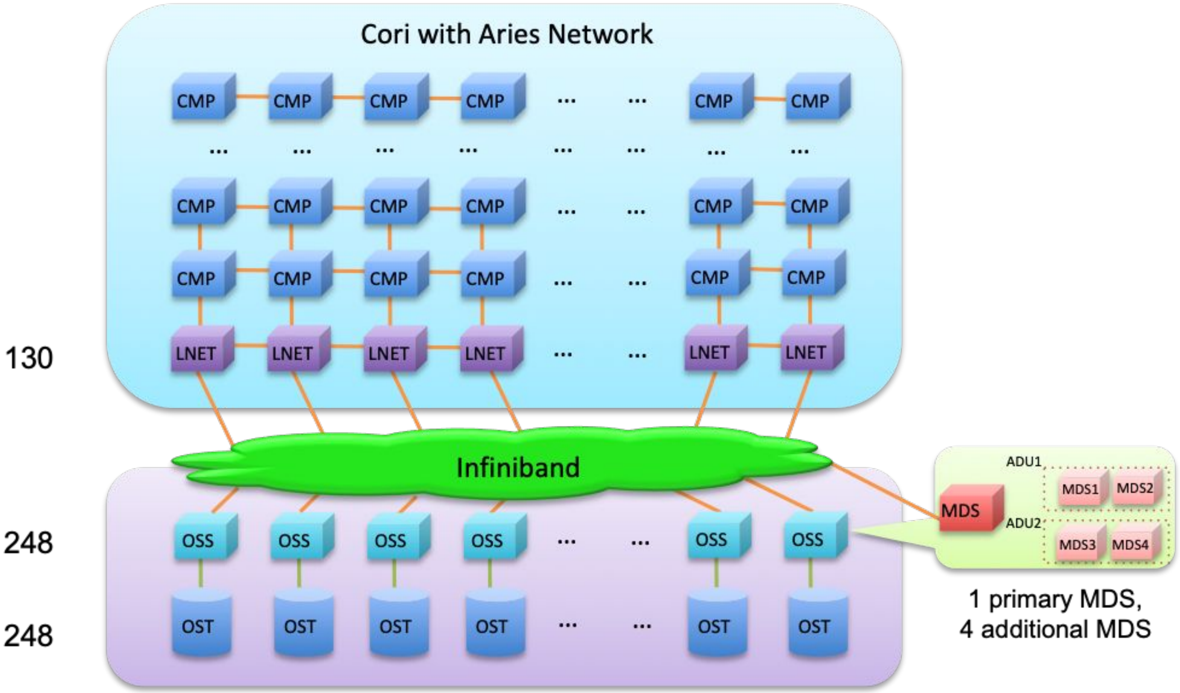

MPI_File_write_all
Aggregated high bandwidth I/O
MDS
OSS
OSS
OSS
OSS
MDT
OST
OST
OST
OST
OST
OST
OST
OST
Lustre
High throughput accesses using striping accesses among OSTs

MDS: Meta Data Server
OSS: Object Storage Server
MDT: Meta Data Target
OST: Object Storage Target

Using MPI-IO on Lustre[2]

[1] https://en.wikipedia.org/wiki/Lustre_(file_system)
[2] MPI-IO on Lustre: https://www.sys.r-ccs.riken.jp/ResearchTopics/fio/mpiio/

- **Scratch**



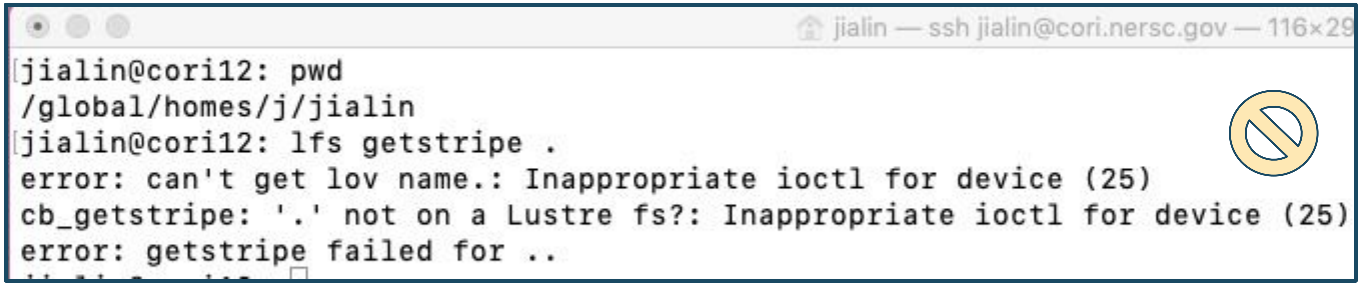
Each OSS controls one OST. The Infiniband connects the MDS, ADUs and OSSs to the LNET routers on the Cray XC System. The OSTs are configured with GridRAID, similar to RAID6, (8+2), but can restore failure 3.5 times faster than traditional RAID6. Each OST consists of 41 disks, and can deliver 240TB capacity.

Cori File System at NERSC

- Too many files and very deep hierarchy?
  - MDS->ADU?
- Too big file?
  - Striping?

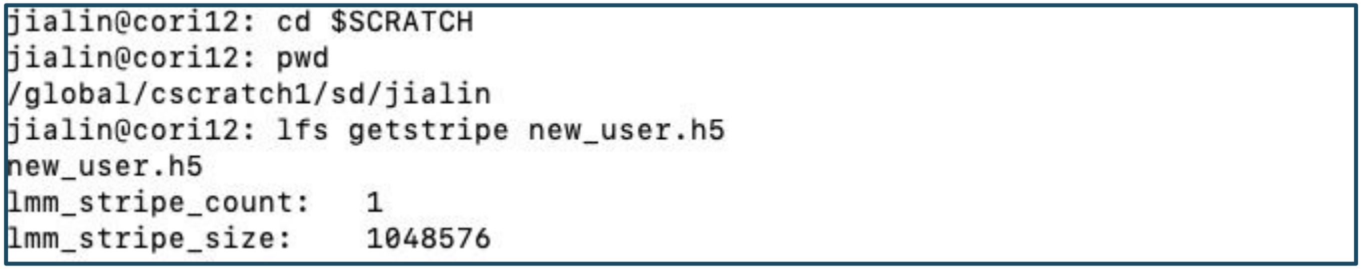
Send us email!

Demo with lfs cmd



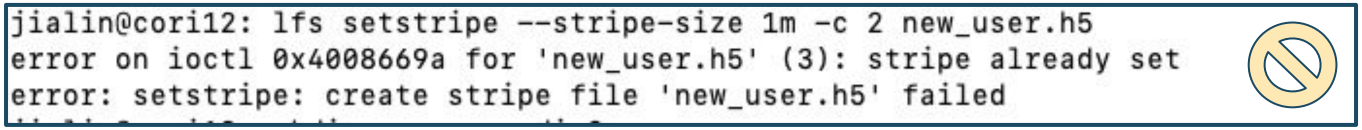
```
jialin@cori12: pwd
/global/homes/j/jialin
jialin@cori12: lfs getstripe .
error: can't get lov name.: Inappropriate ioctl for device (25)
cb_getstripe: '.' not on a Lustre fs?: Inappropriate ioctl for device (25)
error: getstripe failed for ..
```

lfs cmd doesn't work on HOME



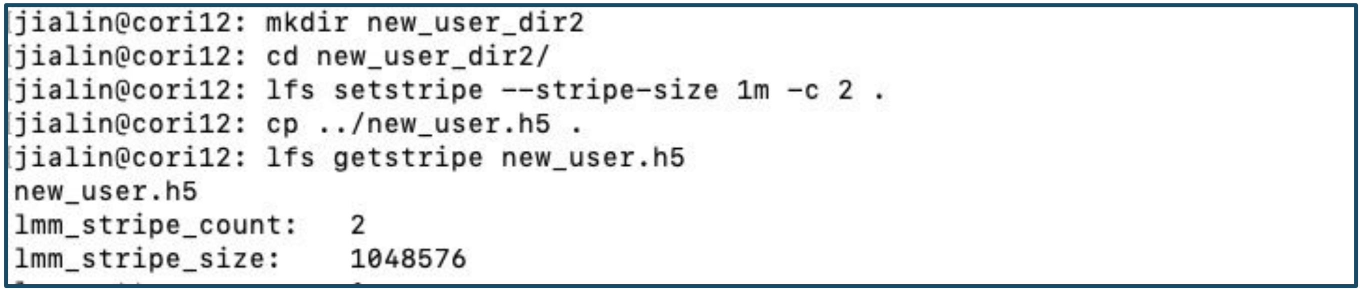
```
jialin@cori12: cd $SCRATCH
jialin@cori12: pwd
/global/cscratch1/sd/jialin
jialin@cori12: lfs getstripe new_user.h5
new_user.h5
lmm_stripe_count:   1
lmm_stripe_size:    1048576
```

lfs cmd only works on Lustre fs



```
jialin@cori12: lfs setstripe --stripe-size 1m -c 2 new_user.h5
error on ioctl 0x4008669a for 'new_user.h5' (3): stripe already set
error: setstripe: create stripe file 'new_user.h5' failed
```

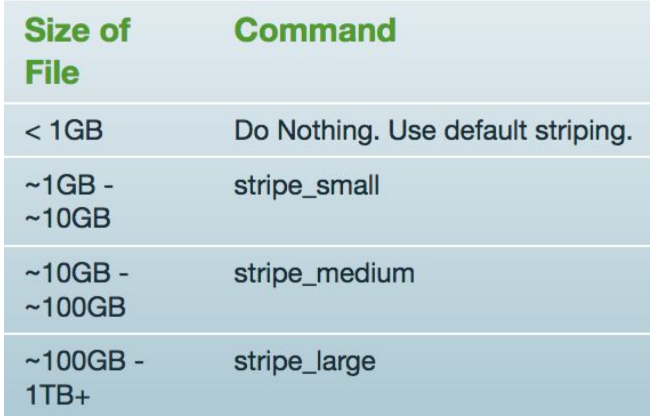
Can't change stripe for existing files



```
jialin@cori12: mkdir new_user_dir2
jialin@cori12: cd new_user_dir2/
jialin@cori12: lfs setstripe --stripe-size 1m -c 2 .
jialin@cori12: cp ../new_user.h5 .
jialin@cori12: lfs getstripe new_user.h5
new_user.h5
lmm_stripe_count:   2
lmm_stripe_size:    1048576
```

set stripe first

- **Scratch Striping Recommendations**
  - Optimize IO by controlling striping
    - By default data is striped across 1 OST, ideal for file per process IO
    - Single shared file IO should be striped according to its size

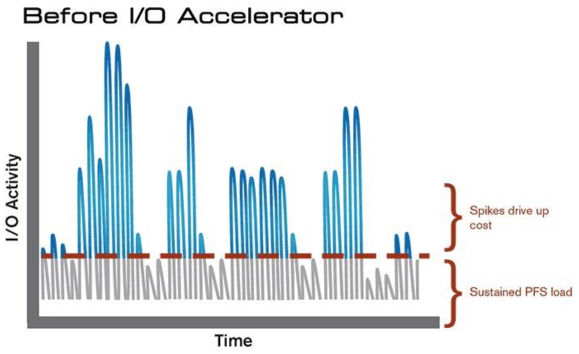


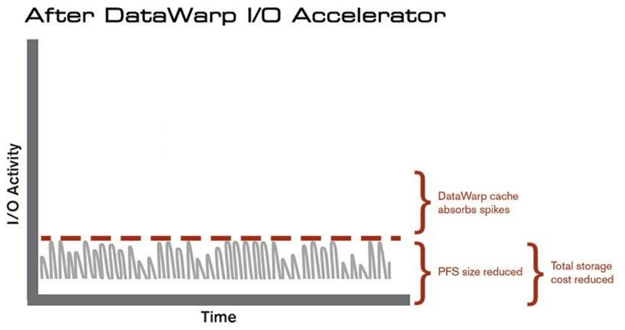
Size of File	Command
< 1GB	Do Nothing. Use default striping.
~1GB - ~10GB	stripe_small
~10GB - ~100GB	stripe_medium
~100GB - 1TB+	stripe_large

    - Manually query with “lfs getstripe <file_name>” and manually set with “lfs setstripe -S 1m -c 2 <empty_folder>”
  - Purged! Files not accessed for more than 12 weeks are automatically deleted

Where Do I Put My Data? Burst Buffer

- **Burst Buffer**
  - Datawarp: Cray's applications I/O accelerator
  - For: Data read in/out by any high IO b/w or IOPS application
  - Truly transient, you must stage in and out from Cori scratch, control with slurm directives (see details later)
  - Scale BW by sizing request, unique MDS so can serve high IOPS workloads
- **Why Burst Buffer**





https://www.cray.com/products/storage/datawarp

Example of Using Burst Buffer

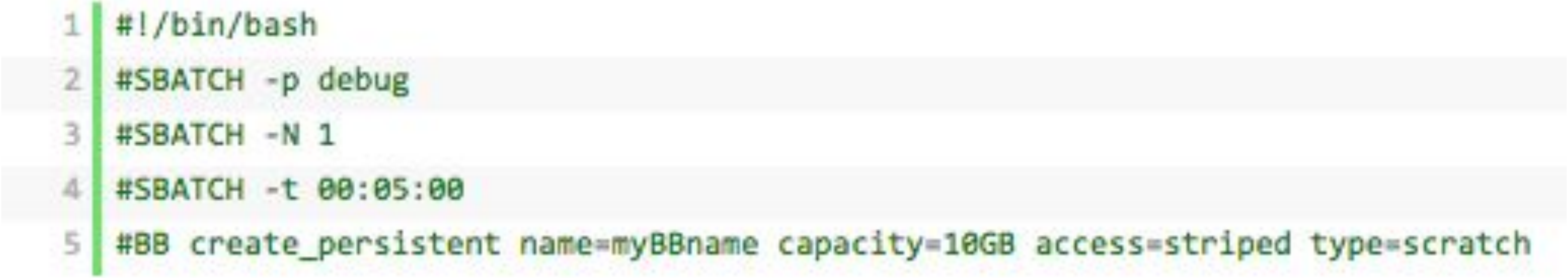
```
#!/bin/bash
#SBATCH -q regular -N 10 -C haswell -t 00:10:00
#DW jobdw capacity=1000GB access_mode=striped type=scratch
#DW stage_in source=$SCRATCH/inputs destination=$DW_JOB_STRIPED/inputs \ type=directory
#DW stage_in source=$SCRATCH/file.dat destination=$DW_JOB_STRIPED/ type=file
#DW stage_out source=$DW_JOB_STRIPED/outputs destination=/lustre/outputs \  type=directory
srun my.x --indir=$DW_JOB_STRIPED/inputs --infile=$DW_JOB_STRIPED/file.dat \
--outdir=$DW_JOB_STRIPED/outputs
```

❑ 'type=scratch' – duration just for compute job (i.e. not 'persistent')

❑ 'access_mode=striped' – visible to all compute nodes and striped across multiple BB nodes

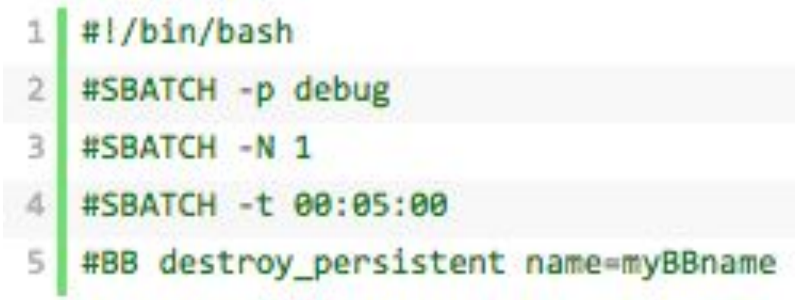
❑ Data 'stage_in' before job start and 'stage_out' after

- **Creating a Persistent Reservation on Busrt Buffer**
  - Need a batch job to create one PR first
  - Don’t forget to delete the PR after 6 weeks.



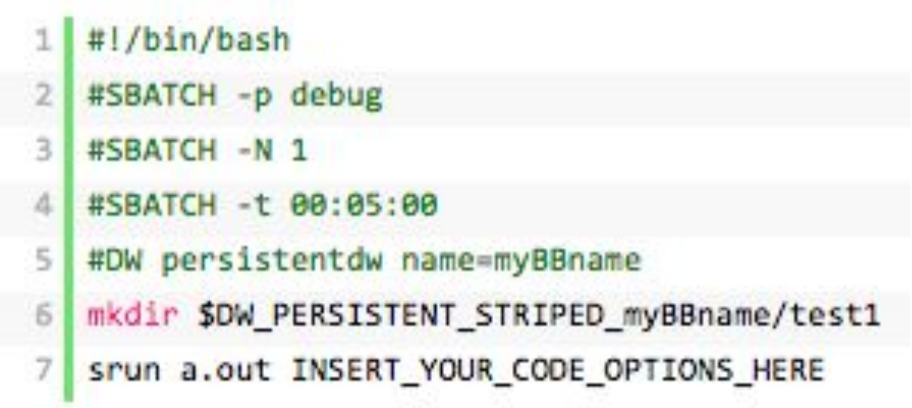
```
#!/bin/bash
#SBATCH -p debug
#SBATCH -N 1
#SBATCH -t 00:05:00
#BB create_persistent name=myBBname capacity=10GB access=striped type=scratch
```

Job0: Create a PR



```
#!/bin/bash
#SBATCH -p debug
#SBATCH -N 1
#SBATCH -t 00:05:00
#BB destroy_persistent name=myBBname
```

Job_6weeks_later: Delete PR



```
#!/bin/bash
#SBATCH -p debug
#SBATCH -N 1
#SBATCH -t 00:05:00
#DW persistentdw name=myBBname
mkdir $DW_PERSISTENT_STRIPED_myBBname/test1
srun a.out INSERT_YOUR_CODE_OPTIONS_HERE
```

Use PR in your jobs

Where Do I Put My Data? Project

- **Project**
  - For: Large data that you need for the next few years
  - Set up for sharing with group read permissions by default
  - **Snapshots** automatically back up for the last 7 days. Accidentally delete something? Get it back at /project/projectdirs/<reponame>/.snapshots/<date>
  - Can share data externally by dropping it into a www directory
    - Example: **/global/project/projectdirs**/das/www/jialin
    - https://portal.nersc.gov/project/das/jialin/
  - Data is never deleted, usage is managed by quotas
  - Can request quota increases and custom directory names

https://docs.nersc.gov/filesystems/project/

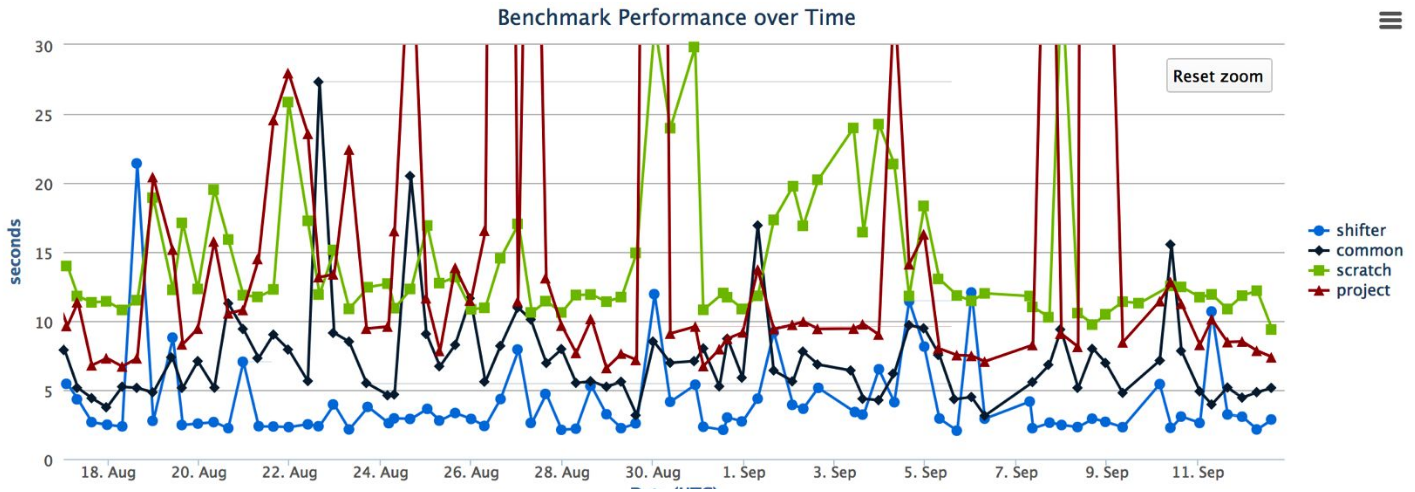
Where Do I Put My Data? HPSS

- **HPSS**
  - For: Data from your finished paper, raw data you might need in case of emergency, really hard to generate data
- **HPSS is tape!**
  - Data first hits a spinning disk cache and gets migrated to tapes
  - Files can end up spread all over, so use htar to aggregate into bundles
  - Archive the way you intend to retrieve the data

https://docs.nersc.gov/filesystems/archive/
https://docs.nersc.gov/filesystems/archive_access/

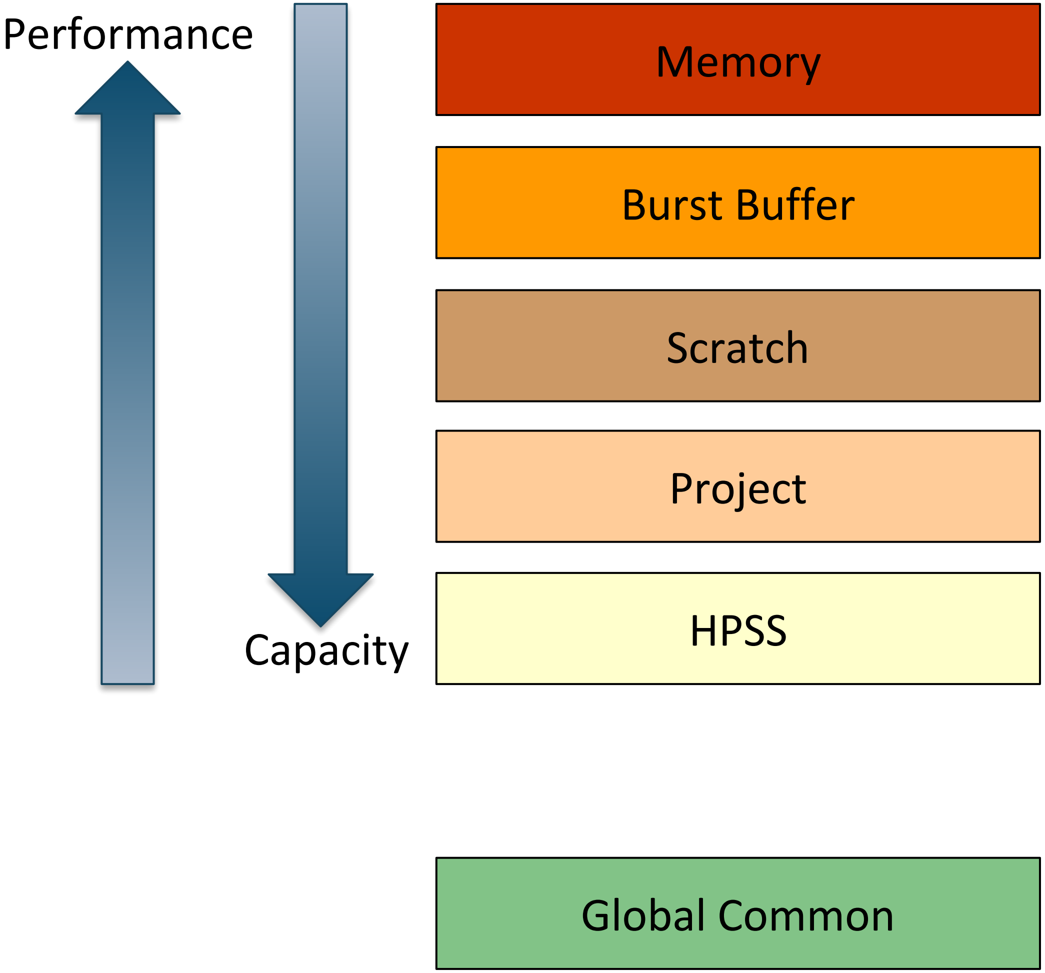
Where Do I Put My Data?

- **Global Common**
  - For: Software stacks
  - Why? Library load performance



- **Group writable directories similar to project, but with a 10 GB quota**
- **Smaller block size for faster compiles than project**

Simplified NERSC File Systems



Performance

Capacity

Memory

Burst Buffer

Scratch

Project

HPSS

Global Common

1.8 PB SSD Burst Buffer on Cori

Cray Datawarp
1.8 TB/s,
temporary for job or campaign

28 PB (Cori) HDD Scratch

Lustre
700 GB/s,
temporary (12 wk purge)

10.7 PB HDD and SSD Project

Spectrum Scale (GPFS)
150 GB/s, permanent

150 PB Tape Archive

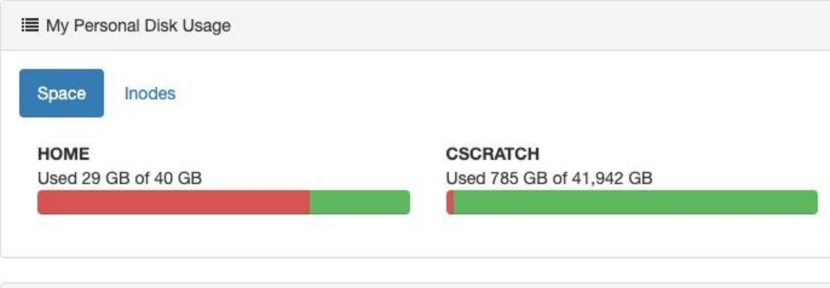
HPSS
Forever

20 TB SSD Software

Spectrum Scale
Permanent,
read-only on computes

Data Dashboard on Cori

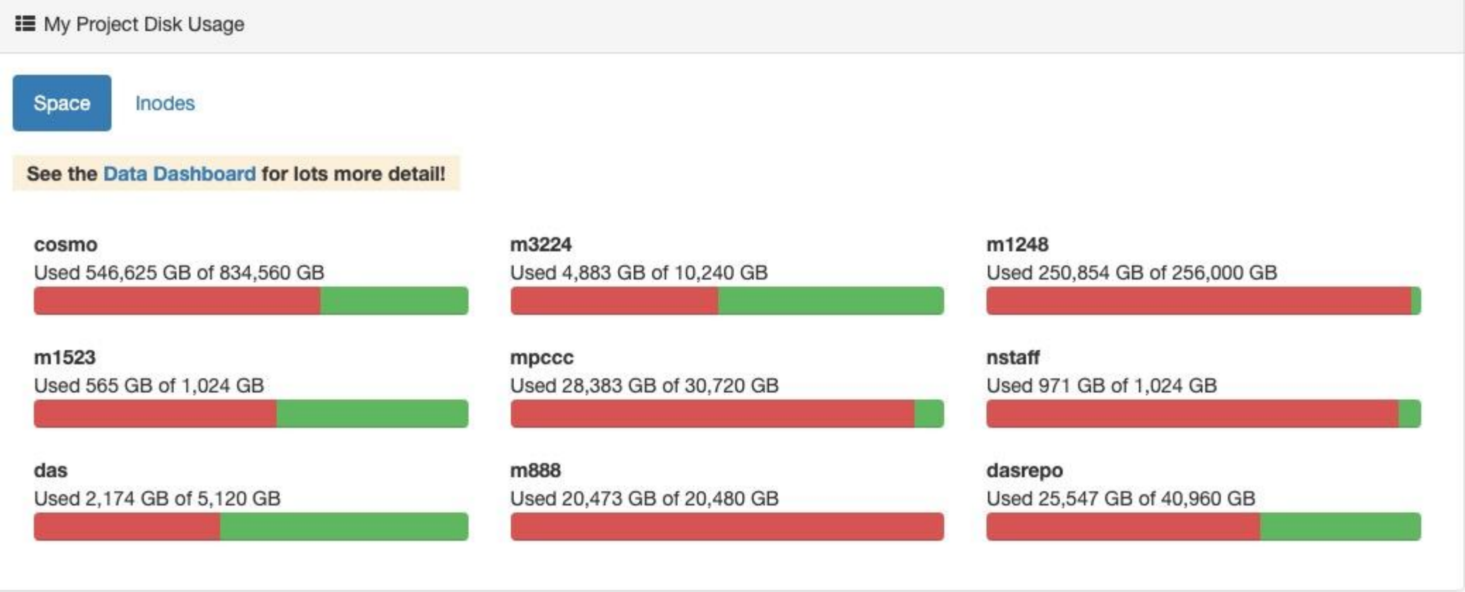
https://my.nersc.gov/



My Personal Disk Usage

HOME
Used 29 GB of 40 GB

CSCRATCH
Used 785 GB of 41,942 GB



My Project Disk Usage

See the Data Dashboard for lots more detail!

cosmo
Used 546,625 GB of 834,560 GB

m3224
Used 4,883 GB of 10,240 GB

m1248
Used 250,854 GB of 256,000 GB

m1523
Used 565 GB of 1,024 GB

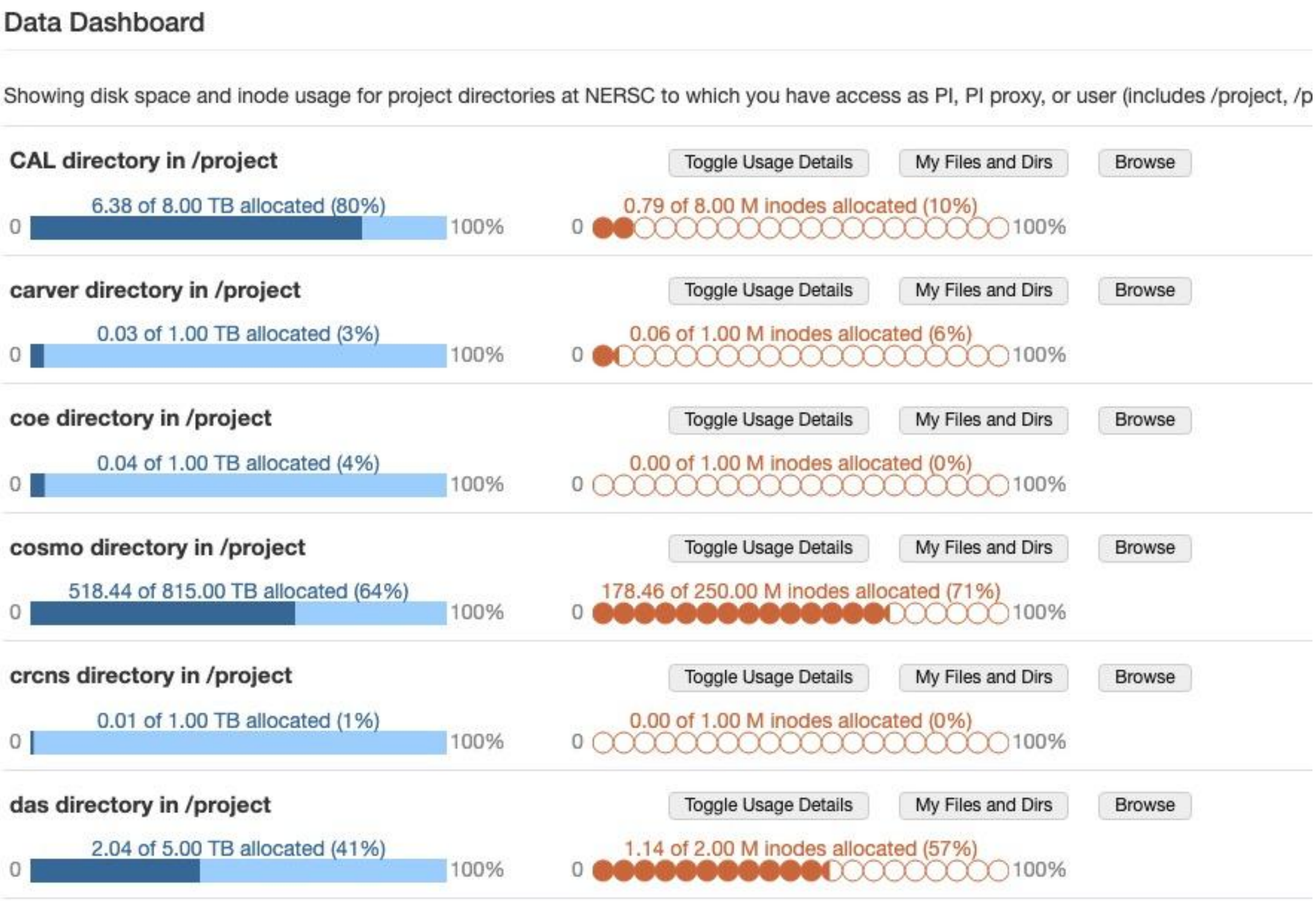
mpccc
Used 28,383 GB of 30,720 GB

nstaff
Used 971 GB of 1,024 GB

das
Used 2,174 GB of 5,120 GB

m888
Used 20,473 GB of 20,480 GB

dasrepo
Used 25,547 GB of 40,960 GB



Data Dashboard

Showing disk space and inode usage for project directories at NERSC to which you have access as PI, PI proxy, or user (includes /project, /p

CAL directory in /project

0 6.38 of 8.00 TB allocated (80%) 100%

0 0.79 of 8.00 M inodes allocated (10%) 100%

carver directory in /project

0 0.03 of 1.00 TB allocated (3%) 100%

0 0.06 of 1.00 M inodes allocated (6%) 100%

coe directory in /project

0 0.04 of 1.00 TB allocated (4%) 100%

0 0.00 of 1.00 M inodes allocated (0%) 100%

cosmo directory in /project

0 518.44 of 815.00 TB allocated (64%) 100%

0 178.46 of 250.00 M inodes allocated (71%) 100%

crcns directory in /project

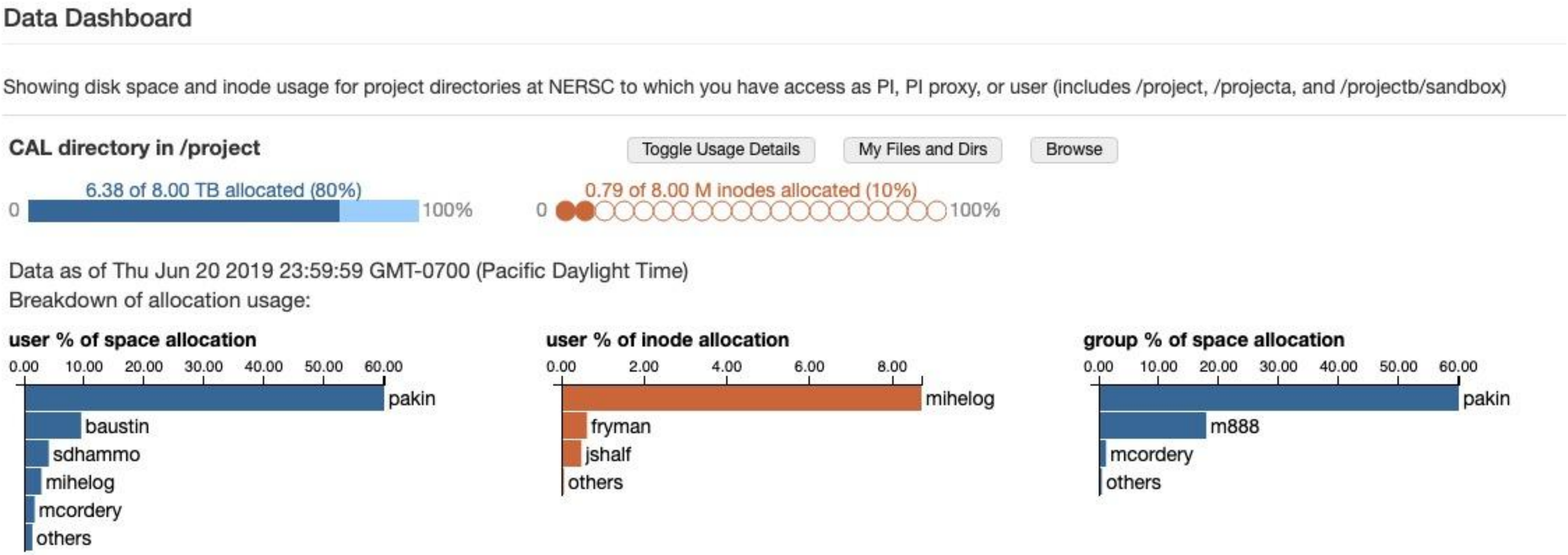
0 0.01 of 1.00 TB allocated (1%) 100%

0 0.00 of 1.00 M inodes allocated (0%) 100%

das directory in /project

0 2.04 of 5.00 TB allocated (41%) 100%

0 1.14 of 2.00 M inodes allocated (57%) 100%



Data Dashboard

Showing disk space and inode usage for project directories at NERSC to which you have access as PI, PI proxy, or user (includes /project, /projecta, and /projectb/sandbox)

CAL directory in /project

Toggle Usage Details
My Files and Dirs
Browse
0
6.38 of 8.00 TB allocated (80%)
100%
0
0.79 of 8.00 M inodes allocated (10%)
100%

Data as of Thu Jun 20 2019 23:59:59 GMT-0700 (Pacific Daylight Time)

Breakdown of allocation usage:

Resources

- **Cori File Systems**

  https://www.nersc.gov/users/computational-systems/cori/file-storage-and-i-o/

- **Optimizing I/O on Lustre**

  https://www.nersc.gov/users/storage-and-file-systems/i-o-resources-for-scientific-applications/optimizing-io-performance-for-lustre/

- **NERSC Burst Buffer Web Pages**

  http://www.nersc.gov/users/computational-systems/cori/burst-buffer/

- **Example batch scripts**

  http://www.nersc.gov/users/computational-systems/cori/burst-buffer/example-batch-scripts/

- **Crays DataWarp User Guide**

  http://docs.cray.com/PDF/XC_Series_DataWarp_User_Guide_CLE60UP04_S-2558.pdf

- **Burst Buffer Early User Program Paper**

  http://www.nersc.gov/assets/Uploads/Nersc-BB-EUP-CUG.pdf