NERSC File Systems and Data Management

Steve Leak
NERSC User Engagement Group

NUG New User Training
March 21, 2016
Topics

• What filesystems and storage do we have?
  – And how/when to use it

• How to share data with colleagues

• How to move data to, from and around NERSC systems
Key Points

• Variety of storage types available to meet different needs
  – Be aware of strengths and limitations of each, use each accordingly

• BACK UP YOUR IMPORTANT FILES TO HPSS (archive)

• Many ways to move data to/from NERSC
  – And most of them are better than ‘scp’

• If in doubt, ask for help
  – www.nersc.gov -> “For Users”
  – ServiceNow (help.nersc.gov) or email (consult@nersc.gov)
NERSC File Systems in a nutshell

HPSS
archive.nersc.gov
hpss.nersc.gov
- Tape-backed HSM
  - 240+PB tape, 280+TB disk cache
- Optimized for infrequent access of files >100GB
- Usage charged by files, stored volume and GB read/written

NGF (GPFS)

SHOME
- small space (40GB, 1M nodes)
- backed up
- daily snapshots (7 days)
- not purged
- private to you
- visible everywhere
- not fast (5 GB/s)
- optimized for small files - source code etc
- DO NOT RUN JOBS HERE

$SCRATCH (genepool)

$HOME (NGF)

Project, /projecta, /projectb
- large space (17TB, 1M nodes)
- backed up if quota < 5TB
- not purged
- shared with project group members
- visible everywhere
- Up to 40GB/s I/O bandwidth
- optimized for using and moving large files
  - (actively used job outputs, etc)
- DO NOT RUN JOBS HERE

Burst Buffer - Coming soon
- NOT YET AVAILABLE TO GENERAL USERS
- PCIe-attached SSDs directly connected to Cray Aries interconnect
- Supports IOPS-heavy workloads, faster checkpointing
- (mostly) deleted at the end of each job

$SCRATCH, /scratch3
- large: Edison: 10TB, 5M nodes, Cori: 20TB, 10M nodes
- fast: Edison: 48 + 48 + 72 GB/s, Cori: 700 GB/s
- local to each cluster
- only visible to you
- PURGED: files unused for 8 weeks are deleted without notice
  - Backup your files to HPSS!
- Optimized for large parallel I/O operations
- BEST PLACE TO RUN JOBS

data transfer nodes
dtn0.nersc.gov
dtn1.nersc.gov

edison.nersc.gov
cori.nersc.gov

The internet

infiniband

ethernet
NERSC Global $HOME

• Home directory shared across all NERSC clusters
• Small space (40GB, 1M inodes)
• Backed up to tape, and daily snapshots for last 7 days
• Never purged
• Private to you
• Visible everywhere
• Suitable for source code, configuration files, etc

• **DO NOT RUN JOBS HERE**
• **Served from NERSC Global Filesystem (NGF)**
  
  – Based on IBM GPFS

• **Provided by two ~100 TB file systems**
  
  – `/global/u1/`
  
  – `/global/u2/`
  
  – Users assigned randomly to one of them
    
    • Symbolic link on the other
      
      `/global/u1/s/sleak`
      
      `/global/u2/s/sleak` → `/global/u1/s/sleak`

• **Access it with $HOME or ~/**
  
  – Underlying name might change, “$HOME” will not
NERSC Global $HOME

• Served from NERSC Global Filesystem (NGF)
  – Based on IBM GPFS
• 5 GB/s aggregate bandwidth
  – To $HOME, shared by all users
• Shared by ~6000 active NERSC users
  – Inefficient use affects others
• Don’t run jobs here!
  – Neither space nor I/O bandwidth are suitable
• Don’t send Slurm stderr/stdout here
  – Submit jobs from $SCRATCH, or redirect output to there
NERSC Global $HOME

• $HOME daily snapshots (last 7 days)
  – Extra-hidden folder $HOME/.snapshots

sleak@cori03:~$ ls -a
.                  .bashrc.ext .globus .local .pyhistory .udiRoot .zprofile.ext
..                 .cache       .history .login .python-eggs .vim .zshenv
.Xauthority       .config     .inputrc .login.ext .ssh .viminfo .zshenv.ext
.bash_history    .cshrc      .intel .netrc .subversion .vimrc .zshrc
.bash_profile    .cshrc.ext  .kshrc  .odbc.ini .swp .zlogin .zshrc.ext
.bash_profile.ext .fontconfig .kshrc.ext .profile .tcshrc .zlogin.ext .zshenv.ext
.bashrc          .gitconfig  .lesshst .profile.ext .tcshrc.ext .zprofile

sleak@cori03:~$ ls .snapshots
2016-03-09 2016-03-10 2016-03-11 2016-03-12 2016-03-13 2016-03-14 2016-03-15 2016-03-16

sleak@cori03:~$ ls .snapshots/2016-03-12
NESAP Tools Training UserSupport aaa bin intel log.lammps xtnodestat

• Mistakes, hardware failures happen!

Backup important files to HPSS
NERSC Global $HOME

• Quotas
  – 40 GB
  – 1,000,000 inodes (i.e. files and directories)
  – Quota increases for $HOME are almost never granted
    • (why do you need more than 40GB of source code? May need to reconsider what you are storing in $HOME)
  – Monitor your usage with `myquota`
    • Also visible in NIM

    sleak@cori03:~$ myquota
    Displaying quota usage for user sleak:

    | FileSystem       | Usage | Quota | InDoubt | Usage  | Quota    | InDoubt |
    |------------------|-------|-------|---------|--------|----------|---------|
    | /global/cscratch | 0     | 20480 | –       | 51     | 10000000 | –       |
    | HOME             | 6     | 40    | 0       | 133431 | 1000000  | 0       |
• Help! I deleted some large files, but my usage according to my quota stayed the same
  – Check for any running processes that are using the deleted files. The space will not be returned until these processes finish or are killed
    • The process may be on a different login node, or part of a batch job you have running
NERSC Global $HOME

• **Backups and retention**
  – Nightly backups to tape
    • Kept for 90 days
    • Last 7 days accessible via hidden $HOME/.snapshots folder
    • Recovering from tape is possible but slow, contact us via ServiceNow (help.nersc.gov) or email (consult@nersc.gov)
  – Data is kept on tape for 1 year after your account is deactivated
NERSC File Systems in a nutshell

HPSS
- Tape-backed HSM
  - 240+PB tape, 280+PB disk cache
- Optimized for infrequent access of files >100GB
- Usage charged by hours, stored volume and GB read/written

NGF (GPFS)
- $SCRATCH (genepool)
  - Large space (178PB, 1.1M nodes)
  - Backed up
  - Daily snapshots (7 days)
  - Not purged
  - Private to you
  - Visible everywhere
  - Not fast (5GB/s)
  - Optimized for small files - source code etc.
  - DO NOT RUN JOBS HERE

$HOME (NGF)
- Large space (178PB, 1.1M nodes)
- Backed up
- Not purged
- Shared with project group members
- Visible everywhere
- Up to 40GB/s IO bandwidth
- Optimized for using and moving large files
- (actively used job outputs, etc)
- DO NOT RUN JOBS HERE

Data Transfer Nodes
- dtn0.nersc.gov
- dtn1.nersc.gov

Infiniband

edison.nersc.gov
- SSCRATCH (/scratch3)
  - Large: Edison: 10TB, 5M nodes, Cori: 20TB, 10M nodes
  - Fast: Edison: 48 + 48 + 72 GB/s, Cori: 700 GB/s
  - Local to each cluster
  - Only visible to you
  - PURGED: files unused for 8 weeks are deleted without notice
    - Backup your files to HPSS!
  - Optimized for large parallel I/O operations
  - BEST PLACE TO RUN JOBS

- scratch3
  - Lustre

Burst Buffer - Coming soon
- NOT YET AVAILABLE TO GENERAL USERS
- PCIe-attached SSDs directly connected to Cray Aries interconnect
- Supports IOPS-heavy workloads, faster checkpointing
- (mostly) deleted at the end of each job

The Internet
Project File Systems

• Shared across all NERSC clusters
• Large space (1TB, 5M inodes)
• Backed up to tape, and daily snapshots for last 7 days
  – If quota <= 5 TB
• Never purged
• Shared with project group members
• Visible everywhere
• Web-accessible via science gateways
• Best for holding and sharing actively-used data
• **DO NOT RUN JOBS HERE**
• Served from NERSC Global Filesystem (NGF)
• 5.1 PB high-performance disk
  – 50GB/s aggregate bandwidth
• Every MPP repo has a project space
  – /project/projectdirs/m9999
• Tuned for large streaming file access
  – Not the place to run jobs .. But jobs could read large input files directly from here
Project File Systems

• **Sharing data**
  – Access control is via Unix groups
  – PI manages membership
    • (http://www.nersc.gov/users/accounts/nim/nim-guide-for-pis/)
  – More on sharing soon

• **Science gateways**
  – Web portals for sharing data with external collaborators
    mkdir /project/projectdirs/yourproject/www
    chmod -R 755 /project/projectdirs/yourproject/www
  – Corresponds to http://portal.nersc.gov/project/yourproject
## Project File Systems

- **Quotas**
  - 1 TB
  - 1,000,000 inodes (i.e. files and directories)
  - Quota increases considered
  - Monitor your usage with `prjquota <yourproject>`
    - Also visible in NIM

```
sleak@cori03:$ prjquota acme

<table>
<thead>
<tr>
<th>Project</th>
<th>Usage</th>
<th>Quota</th>
<th>InDoubt</th>
<th>Usage</th>
<th>Quota</th>
<th>InDoubt</th>
</tr>
</thead>
<tbody>
<tr>
<td>acme</td>
<td>1014</td>
<td>1024</td>
<td>0</td>
<td>899382</td>
<td>1000000</td>
<td>0</td>
</tr>
</tbody>
</table>
```
• **Backups and retention**
  
  – Nightly backups to tape
    • Kept for 90 days
    • Last 7 days accessible via hidden $HOME/.snapshots folder
    • Recovering from tape is possible but slow, contact us via ServiceNow (help.nersc.gov) or email (consult@nersc.gov)
  
  – Data is kept on tape for 1 year after project becomes inactive (no allocation, no activity)
NERSC File Systems in a nutshell

- **HPSS**
  - NERSC archiving system
  - Tape-backed HSM
  - 240+PB tape, 280+TB disk cache
  - Optimized for infrequent access of files >100GB
  - Usage charged by files, stored volume and GB read/written

- **NGF (GPFS)**
  - Large space (17TB, 11M nodes)
  - Backed up
  - Daily snapshots (7 days)
  - Not purged
  - Private to you
  - Visible everywhere
  - Not fast (5 Gbps)
  - Optimized for large parallel I/O operations
  - BEST PLACE TO RUN JOBS

- **SHOME**
  - Small space (40GB, 1M nodes)
  - Backed up
  - Daily snapshots (7 days)
  - Not purged
  - Private to you
  - Visible everywhere
  - Not fast (5 Gbps)
  - Optimized for small files - source code etc
  - DO NOT RUN JOBS HERE

- **/project, /projecta, /projectb**
  - Large space (17TB, 11M nodes)
  - Backed up if quota < 5TB
  - Not purged
  - Shared with project group members
  - Visible everywhere
  - Up to 400GB I/O bandwidth
  - Optimized for using and moving large files
  - (actively used job outputs, etc)
  - DO NOT RUN JOBS HERE

- **/scratch, /scratch3**
  - Large: Edison: 10TB, 5M nodes, Cori: 20TB, 10M nodes
  - Fast: Edison: 48 + 48 + 72 GB/s, Cori: 700 GB/s
  - Local to each cluster
  - Only visible to you
  - PURGED files unused for 8 weeks are deleted without notice
  - Backup your files to HPSS!
  - Optimized for large parallel I/O operations

- **Burst Buffer - Coming soon**
  - NOT YET AVAILABLE TO GENERAL USERS
  - PCIe-attached SSDs directly connected to Cray Aries interconnect
  - Supports IOPS-heavy workloads, faster checkpointing
  - (mostly) deleted at the end of each job

- **Data Transfer Nodes**
  - dtn0.nersc.gov
  - dtn1.nersc.gov

- **edison.nersc.gov**

- **genepool.nersc.gov**

- **cori.nersc.gov**

- **Infiniband**

- **Ethernet**

- **The Internet**
Local $SCRATCH$

- Tape-backed HSM
  - 240-1PB tape, 280+TB disk cache
  - Optimized for infrequent access of files >100GB
  - Usage charged by slice, stored volume and GB read/written

$SCRATCH$ (genepool)
- NGF (GPFS)
  - small space (40GB, 1M inodes)
  - backed up
  - daily snapshots (7 days)
  - not purged
  - private to you
  - visible everywhere
  - not fast (5 GB/s)
  - optimized for small files - source code etc.
  - DO NOT RUN JOBS HERE

SHOME
- large space (1TB, 1M inodes)
- backed up (if quota < 5TB)
- not purged
- shared with project group members
- visible everywhere
- Up to 40GB/s I/O bandwidth
- optimized for using and moving large files
  (active use job outputs, etc.)
- DO NOT RUN JOBS HERE

$project$ (GeneProject)
- large space (1TB, 1M inodes)
- backed up (if quota < 5TB)
- not purged
- shared with project group members
- visible everywhere
- Up to 40GB/s I/O bandwidth
- optimized for using and moving large files
  (active use job outputs, etc.)
- DO NOT RUN JOBS HERE

HPSS
- archive.nersc.gov
- hpss.nersc.gov
- Tape-backed HSM
- 240-1PB tape, 280+TB disk cache
- Optimized for infrequent access of files >100GB
- Usage charged by slice, stored volume and GB read/written

Data Transfer Nodes
dtn5.nersc.gov
dtn1.nersc.gov

edison.nersc.gov

Burst Buffer - Coming soon
- NOT YET AVAILABLE TO GENERAL USERS
- PCIe-attached SSDs directly connected to Cray Aries interconnect
- Supports IOPs-heavy workloads, faster checkpointing
- (mostly) deleted at the end of each job

$SCRATCH$, /scratch3
- large: Edison: 30TB, 5M inodes, Cori: 20TB, 10M inodes
- fast: Edison: 48 = 48 + 72 GB/s, Cori: 700 GB/s
- local to each cluster
- only visible to you
- PURGED: files unused for 8 weeks are deleted without notice
  - Backup your files to HPSS!
- Optimized for large parallel I/O operations
- BEST PLACE TO RUN JOBS
Local $\text{SCRATCH}$

- Local to each cluster
- Large
  - Edison: 10 TB, 5,000,000 inodes
  - Cori: 20 TB, 10,000,000 inodes
- FAST
  - Edison $\text{SCRATCH}$: 48 GB/s aggregate per filesystem
  - Edison /scratch3: 72 GB/s aggregate
  - Cori $\text{SCRATCH}$: 700 GB/s aggregate
- Optimized for large parallel I/O workloads
- **BEST PLACE TO RUN JOBS**
Local $SCRATCH

• Not backed up
• Subject to purging
  – Files not actively used in last 8 weeks are deleted without notice
    • Purged files are listed in $SCRATCH/.purged.<timestamp>

BACK UP IMPORTANT FILES TO HPSS!
• **Quotas**
  – Edison: 10 TB, 5,000,000 inodes
  – Cori: 20 TB, 10,000,000 inodes
  – Quota increases considered
  – Monitor your usage with `myquota`
    • Also visible in NIM

```bash
sleak@cori03:~$ myquota
Displaying quota usage for user sneak:

<table>
<thead>
<tr>
<th>FileSystem</th>
<th>Usage</th>
<th>Space (GB)</th>
<th>Quota</th>
<th>InDoubt</th>
<th>Usage</th>
<th>Quota</th>
<th>InDoubt</th>
</tr>
</thead>
<tbody>
<tr>
<td>/global/cscratch</td>
<td>0</td>
<td>20480</td>
<td>-</td>
<td>-</td>
<td>51</td>
<td>10000000</td>
<td>-</td>
</tr>
<tr>
<td>HOME</td>
<td>6</td>
<td>40</td>
<td>0</td>
<td></td>
<td>133431</td>
<td>1000000</td>
<td>0</td>
</tr>
</tbody>
</table>
```
Local $\text{SCRATCH}$

- **Lustre filesystem**
  - Edison: provided by two 2 PB filesystems
    - Users assigned randomly to one of them
  - Cori: single 28 PB filesystem
  - **Access it with $\text{SCRATCH}$**
  - Edison /scratch3: access considered by request
    - [http://www.nersc.gov/users/computational-systems/edison/file-storage-and-i-o/](http://www.nersc.gov/users/computational-systems/edison/file-storage-and-i-o/)
    - Access it by name (/scratch3/scratchdirs/$\text{USER}$)
    - /scratch3 has greater I/O bandwidth
Local $\$SCRATCH$

- $\$SCRATCH$ is configured to provide high-bandwidth I/O for many simultaneous users
  - How does it work?

MDS == "MetaDataServer" == "which OSS to talk to"

OST == "Object Storage Target" == "bunch of disks"

- Striping == spread file over multiple disks improves available bandwidth (if reading/writing enough data)

- MDS

- OSS

- High-speed data network directly to/from storage

- Many clients

just enough to find which OSS
• **Tip: Cray MPI-IO is Lustre-aware**
  – Aggregator MPI tasks communicate each with 1 OST

10 disks / OST
4 OST / OSS

0.5 GB/s per OST

**-default striping of 2
== two full OSTS**

**x12**

48 GB/s aggregate bandwidth for each of /scratch1, /scratch2
- I/O striped over 8 OSTs of 40 disks each
  - high I/O bandwidth
• Large space, highly parallel
  – Eventually will become global scratch space
Optimizing I/O Performance

• You can view/change the stripe size
  – lfs getstripe $SCRATCH/my_file.dat
  – lfs setstripe --s 4m --c 4 $SCRATCH/my_file.dat

• Some shortcuts for single-shared-file I/O:
  – stripe_small $SCRATCH/my_folder
    • Files >1 GB
  – stripe_medium $SCRATCH/my_folder
    • Files >10 GB
  – stripe_large $SCRATCH/my_folder
    • Files >100 GB

• Use with care: can make performance worse
Lustre tips and gotchas

• Don’t keep 100,000 files in the same folder
  – Hard work for OSS, affects performance for other users
  – 100 folders with 1000 files each is much faster

• ‘ls’ vs ‘ls –l’
  – Passing options to ‘ls’ invokes an inquiry on each inode in the folder – occupies OSS/OST with small transfers, non-optimal
  – Basic ‘ls’ needs only information kept in MDS, much faster

• ‘lfs find’ vs ‘find’
  – Same principle: special (limited) version of find that only uses data on MDS, not OSS/OST
NERSC File Systems in a nutshell

- **HPSS**
  - archive.nersc.gov
  - hpss.nersc.gov
  - Tape-backed HSM
  - 240+PB tape, 280+TB disk cache
  - Optimized for infrequent access of files >100GB
  - Usage charged by files, stored volume and GB read/written

- **NGF**
  - $SCRATCH (genepool)
  - $HOME (NGF)
  - Large space (20TB, 10M nodes)
  - Backed up (if quota < 5TB)
  - Not purged
  - Shared with project group members
  - Visible everywhere
  - Up to 400GB I/O bandwidth
  - Optimized for using and moving large files

- **SHOME**
  - Small space (40GB, 1M nodes)
  - Backed up
  - Daily snapshots (7 days)
  - Not purged
  - Private to you
  - Visible everywhere
  - Not fast (5 GB/s)
  - Optimized for small files - source code etc.
  - DO NOT RUN JOBS HERE

- **/project, /projecta, /projectb**
  - Large space (17TB, 1M nodes)
  - Backed up (if quota < 5TB)
  - Not purged
  - Shared with project group members
  - Visible everywhere
  - Up to 400GB I/O bandwidth
  - Optimized for using and moving large files
  - Actively used job outputs, etc
  - DO NOT RUN JOBS HERE

- **Infiniband**
  - Data Transfer Nodes
  - dtn0.nersc.gov
  - dtn1.nersc.gov

- **Ethernet**
  - edison.nersc.gov
  - cori.nersc.gov

- **Burst Buffer - Coming soon**
  - NOT YET AVAILABLE TO GENERAL USERS
  - PCIe-attached SSDs directly connected to Cray Aries interconnect
  - Supports IOPS-heavy workloads, faster checkpointing
  - (mostly) deleted at the end of each job

- **edison.nersc.gov**
  - SSCRATCH, /scratch3
    - Large: Edison: 10TB, 5M nodes, Cori: 20TB, 10M nodes
    - Fast: Edison: 48 + 48 + 72 GB/s, Cori: 700 GB/s
    - Local to each cluster
    - Only visible to you
    - PURGED: files unused for 8 weeks are deleted without notice
    - Backup your files to HPSS!
    - Optimized for large parallel I/O operations
    - BEST PLACE TO RUN JOBS

- **cori.nersc.gov**
  - SSCRATCH
    - Large: Edison: 50TB, 10M nodes, Cori: 20TB, 10M nodes
    - Fast: Edison: 48 + 48 + 72 GB/s, Cori: 700 GB/s
    - Local to each cluster
    - Only visible to you
    - PURGED: files unused for 8 weeks are deleted without notice
    - Backup your files to HPSS!
Burst Buffer

- Tape-backed HSM
  - 240+PB tape, 260+TB disk cache
- Optimized for infrequent access of files >100GB
- Usage charged by slice, stored volume and GB read/written

- NGF (GPFS)
- $SCRATCH$ (genepool)
  - 33GB
  - not fast (5GB/s)
  - optimized for small files - source code etc.
  - DO NOT RUN JOBS HERE

- HOME
  - small space (40GB, 1M inodes)
  - backed up
  - daily snapshots (7 days)
  - not purged
  - private to you
  - visible everywhere
  - not fast (5GB/s)
  - DO NOT RUN JOBS HERE

- $SCRATCH$ (lustre)
  - large: Edison 20TB, 5M inodes
  - fast: Edison 48 + 48 + 72 GB/s, Cori: 700 GB/s
  - local to each cluster
  - only visible to you
  - PURGED: files unused for 8 weeks are deleted without notice
  - Backup your files to HPSS!
  - Optimized for large parallel I/O operations
  - BEST PLACE TO RUN JOBS

- cori.nersc.gov
- edison.nersc.gov
- dtn5.nersc.gov
dtn1.nersc.gov
- genepool.nersc.gov
- pdfs.nersc.gov
- Burst Buffer - Coming soon
- NOT YET AVAILABLE TO GENERAL USERS
- PCIe-attached SSDs directly connected to Cray Aries interconnect
- Supports IOPS-heavy workloads, faster checkpointing
- (mostly) deleted at the end of each job
Burst Buffer

• Coming soon! (not yet available to general users)
• SSD-equipped nodes (and supporting software) for high-IOPS, high-throughput, “job-local” storage
  – Directly attached to XC-40 interconnect (Aries)
• Pre/post-job stage in and stage out
• Current configuration:
  – 144 BB nodes (2 SSDs per BB node)
  – 900 TB @ 900 GB/s, 12.5M IOPS (measured)
• Cori phase 2:
  – ~2x
Burst Buffer

• Why?
  – Limitations of $SCRATCH$:
    • Relies on large, throughput-oriented I/O for performance
  – Checkpointing – extreme bandwidth requirements
    • 1000’s of nodes each writing 10’s of GB
    • Mostly not required again
  – For large parallel jobs, I/O is often “bursty”
    • Most cores waiting while few cores do I/O

• How?
  – #BB job directives passed to sbatch
Burst Buffer

Burst buffer nodes

Storage Fabric

Storage Servers

I/O nodes present $SCRATCH, $HOME
Burst Buffer

compute nodes

BB nodes

LNET/DVS

IO nodes

service nodes
NERSC File Systems in a nutshell

HPSS
archive.nersc.gov
hpss.nersc.gov
- Tape-backed HSM
  - 240+PB tape, 280+TB disk cache
- Optimized for infrequent access of files >100GB
- Usage charged by effort, stored volume and GB read/written

NGF (GPFS)
- Small space (40GB, 1M nodes)
- Backed up
- Daily snapshots (7 days)
- Not purged
- Private to you
- Visible everywhere
- Not fast (5 GB/s)
- Optimized for small files - source code etc
- DO NOT RUN JOBS HERE

SHOME
- Large space (17TB, 1M nodes)
- Backed up (if quota < 5TB)
- Not purged
- Shared with project group members
- Visible everywhere
- Up to 40GB/s I/O bandwidth
- Optimized for using and moving large files (actively used job outputs, etc)
- DO NOT RUN JOBS HERE

$SCRATCH (genepool)

$HOME (NGF)

$HOME (GPFS)

Data Transfer Nodes
dtn0.nersc.gov
dtn1.nersc.gov

Infiniband

Ethernet

edison.nersc.gov
cori.nersc.gov

$SCRATCH (Lustre)

$SCRATCH (GPFS)

$SCRATCH (Lustre)

$SCRATCH (Lustre)

Burst Buffer - Coming soon
- NOT YET AVAILABLE TO GENERAL USERS
- PCIe-attached SSDs directly connected to Cray Aries interconnect
- Supports IOPS-heavy workloads, faster checkpointing
- (mostly) deleted at the end of each job
• Data grows exponentially
  – 80% of stored data is never accessed again after 90 days
Memory

Burst Buffer

Disk

Tape

Speed, Cost

Space, Reliability
• archive.nersc.gov
  – HSM: disk cache, ultimately everything is stored on tape
  – Parallel connections over NERSC internal 10GbE network
• Available to all NERSC users
  – (a second system, hpss.nersc.gov, is for internal use such as system backups)
• No quota, but charged in “Storage Resource Units”
    • Like Amazon Glacier, etc
  – Monitor usage via NIM
## Accessing HPSS

<table>
<thead>
<tr>
<th>Tool</th>
<th>What it does</th>
<th>Where/why to use it</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>htar</td>
<td>Tar directly to/from HPSS</td>
<td>From NERSC hosts. Simple store/retrieve of large directories</td>
<td><code>$ htar cf results-for-publication.tar my_results/</code></td>
</tr>
<tr>
<td>hsi</td>
<td>CLI client</td>
<td>From NERSC hosts. Full featured client</td>
<td><code>$ hsi A:/home/s/sleak-&gt; put myfile</code></td>
</tr>
<tr>
<td>pftp, ftp</td>
<td>High performance (parallel) ftp</td>
<td>When need/prefer ftp-like interface</td>
<td><code>$ pftp archive.nersc.gov ftp&gt; pput results-for-publication.tar</code></td>
</tr>
<tr>
<td>gridFTP</td>
<td></td>
<td>External, gridFTP-enabled sites (you need a grid credential)</td>
<td><code>$ globus-url-copy file://${HOME}/myresults.tar gsfftp://garchive.nersc.gov/home/s/sleak/results-for-publication.tar</code></td>
</tr>
<tr>
<td>Globus Online</td>
<td>Data transfer service</td>
<td>Fire-and-forget transfers</td>
<td>See <a href="http://www.globusonline.org">www.globusonline.org</a></td>
</tr>
</tbody>
</table>
• **Tape storage performance and gotchas**
  
  – Tape is linear media
    
    • Data cannot be written anywhere, only appended at end
    • Reading and writing are sequential, not random-access
  
  – Very high latency:
    
    • Robot must fetch tape, load it into drive, read forwards until file is reached, then read file
    • Number-of-files has bigger impact on access performance than number-of-GB
  
  – Size matters
    
    • Sweet spot currently **100s of GB**
    • Files >1TB will cause trouble (too big for tapes)
Retrieving files in same order they were stored ...

.. vs in random order
• Best practices/Worst practices:
  – Store a few very large files, not many small files
    • htar or tar-first-in-$SCRATCH$
  – Recursively storing or fetching a directory tree will result in many unordered accesses
    • Use htar or tar instead
    • hpss_file_sorter.script => sorts a list of files into “tape order”
• **Best practices/Worst practices:**
  
  
  – HPSS has a single database instances, all user interactions trigger database activity
    
    • `hsi -q 'ls -l'` is database intensive, $O(N^2)$ with number of files in directory
      
      – Too many files in one folder can lock up system for everybody
  
  – Streaming data to pftp from Unix pipeline
    
    • HPSS does not know how big the data will be, likely to put it in wrong place
    
    • Vulnerable to network glitch
Checking my Usage

- nim.nersc.gov

My NGF Quotas & Usage

<table>
<thead>
<tr>
<th>Username</th>
<th>Full Name</th>
<th>Home Space Used (GiB)</th>
<th>Home Space Quota (GiB)</th>
<th>HSQ Def?</th>
<th>Home Inodes Used</th>
<th>Home Inode Quota</th>
<th>HIQ Def?</th>
<th>Home Quota End</th>
<th>Prop Chng</th>
</tr>
</thead>
<tbody>
<tr>
<td>sleek</td>
<td>Stephen Leak</td>
<td>6.1</td>
<td>40</td>
<td>Y</td>
<td>133,443</td>
<td>1,000,000</td>
<td>Y</td>
<td>Never</td>
<td>N</td>
</tr>
</tbody>
</table>

Update User Quotas

Usage for My Project Directories

<table>
<thead>
<tr>
<th>Project Directory</th>
<th>Owner</th>
<th>Group Name</th>
<th>ERCAP Project</th>
<th>Space Usage</th>
<th>Space Quota</th>
<th>Default Space Quota?</th>
<th>Space%</th>
<th>Inode Usage</th>
<th>Inode Quota</th>
<th>Default Inode Quota?</th>
<th>Inode%</th>
<th>Quota Expiration Date</th>
<th>Projdir Status</th>
<th>Status Effective Date</th>
<th>Projdir ID</th>
<th>Group ID</th>
<th>Project ID</th>
<th>Prop Chng</th>
</tr>
</thead>
<tbody>
<tr>
<td>carver</td>
<td>d paul</td>
<td>mpc cc</td>
<td>staff</td>
<td>8</td>
<td>1.0</td>
<td>Y</td>
<td>0.8</td>
<td>63,918</td>
<td>1,000,000</td>
<td>Y</td>
<td>6</td>
<td>Never</td>
<td>Active</td>
<td>Jan-06-2016</td>
<td>43906</td>
<td>11988</td>
<td>13439</td>
<td>N</td>
</tr>
<tr>
<td>dirac</td>
<td>whitney</td>
<td>mpc cc</td>
<td>staff</td>
<td>165</td>
<td>1.0</td>
<td>Y</td>
<td>16</td>
<td>15,576</td>
<td>1,000,000</td>
<td>Y</td>
<td>1.6</td>
<td>Never</td>
<td>Active</td>
<td>Jan-06-2016</td>
<td>43946</td>
<td>11988</td>
<td>13439</td>
<td>N</td>
</tr>
<tr>
<td>gene pool</td>
<td>jay</td>
<td>mpc cc</td>
<td>staff</td>
<td>130</td>
<td>1.0</td>
<td>Y</td>
<td>13</td>
<td>900,469</td>
<td>1,000,000</td>
<td>Y</td>
<td>90</td>
<td>Never</td>
<td>Active</td>
<td>Jan-06-2016</td>
<td>43970</td>
<td>11988</td>
<td>13439</td>
<td>N</td>
</tr>
</tbody>
</table>

- myquota
- prjquota
Sharing Data
Sharing Data

• **Security matters!**
  – Never share passwords

• **With other NERSC users**
  – Project directories (/project) are designed for sharing files with colleagues
    • Not $HOME
  – Unix groups, FACLs (“file access control lists”)
  – give, take commands

• **With external collaborators**
  – Science gateways (on /project)
Sharing Data

• Unix groups
  – What groups am I in?
    • groups
  – New files are associated with your default group
  – To change which group the file is associated with:
    • chgrp my_other_group myfile.txt
    • chgrp –R my_other_group whole_directory_tree/
  – To ensure users in my_other_group can read/write a file or folder:
    • chmod g+rw myfile.txt
    • chmod g+rws my_new_folder/
      – “s” – setgid
“setgid” ??

• setgid “set group id”
  – File mode, set with `chmod`
  – When set on a folder, it means “things added to this folder should inherit the group of the folder”
    • (so I don’t need to keep typing `chgrp` for each new file)
  – NOTE: only things added, not things that were already there
**FACLs**

- **Finer-grain control of access**
  
  - `getfacl`, `setfacl`
  
  - `setfacl -m u_or_g:who:what_perms myfile.txt`
  
  - `setfacl -x`

  - Remove a FACL

```plaintext
getfacl some_file.txt
# file: some_file.txt
# owner: sneak
# group: sneak
user::rw-
group::r--
other::---

setfacl -m u:rjhb:rw some_file.txt

getfacl some_file.txt
# file: some_file.txt
# owner: sneak
# group: sneak
user::rw-
user:rjhb:rw-
group::r--
mask::rw-
other::---
```
My colleague still can’t see my file?

• Check permissions of the folder it is in, and the folder above that, etc
  – Missing permissions at any point in the tree will prevent access to the next level of the tree

• Don’t forget “x” on folders
Give and Take

• Appropriate for smaller files

joe% give -u bob coolfile
  – File copied to spool location
  – Bob gets email telling him Joe has given him a file

bob% take -u joe coolfile
  – File copied from spool location
Science Gateways

• Make data available to outside world
  
  mkdir /project/projectdirs/bigsci/www
  chmod o+x /project/projectdirs/bigsci
  chmod o+rx /project/projectdirs/bigsci/www

• Access with web browser
  
  http://portal.nersc.gov/project/bigsci

• More info:
  
Moving Data Around

• Don’t do it!
  – Ok, sometimes you need to
  – Don’t forget $HOME and /project are shared by all NERSC clusters

• Data transfer nodes
  – Fast network between all NERSC storage locations
  – Visible to internet
  – Dedicated to data transfer
    • Avoids adding load to Edison, Cori login nodes
NERSC File Systems Summary

HPSS
archive.nersc.gov
hpss.nersc.gov
- Tape-backed HSM
- 240+PB tape, 280+TB disk cache
- Optimized for infrequent access of files >100GB
- Usage charged by files, stored volume and GB read/written

NGF (GPFS)

$SCRATCH (genepool)

SHOME
- small space (40GB, 1M nodes)
- backed up
- daily snapshots (7 days)
- not purged
- private to you
- visible everywhere
- not fast (5 Gbps)
- optimized for small files - source code etc
- DO NOT RUN JOBS HERE

Data Transfer Nodes
dtn0.nersc.gov
dtn1.nersc.gov

Infiniband

Ethernet

The Internet

cori.nersc.gov

genepool.nersc.gov

pdfs.nersc.gov

edison.nersc.gov

$HOME (NGF)

$SCRATCH (NGF)

$SCRATCH, /scratch3
- large: Edison: 10TB, 5M nodes, Cori: 20TB, 10M nodes
- fast: Edison: 48 + 48 + 72 GB/s, Cori: 700 GB/s
- Local to each cluster
- only visible to you
- PURGED: files unused for 8 weeks are deleted without notice
- Backup your files to HPSS!
- Optimized for large parallel I/O operations
- BEST PLACE TO RUN JOBS

Burst Buffer - Coming soon
- NOT YET AVAILABLE TO GENERAL USERS
- PCIe-attached SSDs directly connected to Cray Aries interconnect
- Supports IOPS-heavy workloads, fast checkpointing
- (mostly) deleted at the end of each job
## Moving Data Around

<table>
<thead>
<tr>
<th>Tool</th>
<th>What it does</th>
<th>Where/why to use it</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>cp</td>
<td>Local copy</td>
<td>Between NERSC filesystems</td>
<td><code>$ cp $SCRATCH/output.dat /project/projectdirs/m9999/</code></td>
</tr>
</tbody>
</table>
| scp, rsync | Encrypted copy over network | Small amounts of data, collections of small files, over small distances. Use HPN version if available. | `$ scp my_code.f cori:`
  `$ scp -R my_folder/ cori:`
  `$ rsync -avr my_folder/ cori:`
  `$ ssh -V
  OpenSSH_7.1p1-hpn14v5NMOD_3.17, OpenSSL 0.9.8j-fips 07 Jan 2009` |
| bbcp   | Fast parallel network copy. Requires client program | Larger files, longer distances                                                     | `$ bbcp -T "ssh -x -a -oFallbackToRsh=no %I -l %U %H /usr/common/usg/bin/bbcp" /local/path/file
"user_name@dt01.nersc.gov:/remote/path/"` |

See https://www.nersc.gov/users/storage-and-file-systems/transferring-data/
## Moving Data Around

<table>
<thead>
<tr>
<th>Tool</th>
<th>What it does</th>
<th>Where/why to use it</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NERSC ftp upload</td>
<td>Temporary ftp account/server</td>
<td>Allow external collaborators to upload files for you to collect</td>
<td>See <a href="https://www.nersc.gov/users/storage-and-file-systems/transferring-data/nersc-ftp-upload-service/">https://www.nersc.gov/users/storage-and-file-systems/transferring-data/nersc-ftp-upload-service/</a></td>
</tr>
<tr>
<td>gridFTP</td>
<td>Fast network copy protocol, requires certificate</td>
<td>External, gridFTP-enabled sites (you need a grid credential)</td>
<td><code>$ globus-url-copy file://$HOME/myresults.tar</code></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Note: garchive.nersc.gov</td>
<td><code>gsiftp://garchive.nersc.gov/home/s/sleak/results-for-publication.tar</code></td>
</tr>
<tr>
<td>Globus Online</td>
<td>Fast data transfer service. Web or CLI</td>
<td>Fire-and-forget transfers (Especially between NERSC and other HPC centers)</td>
<td>See <a href="http://www.globusonline.org">www.globusonline.org</a></td>
</tr>
</tbody>
</table>

See https://www.nersc.gov/users/storage-and-file-systems/transferring-data/
Summary

• Variety of storage types available to meet different needs
  – Be aware of strengths and limitations of each, use each accordingly

• BACK UP YOUR IMPORTANT FILES TO HPSS (archive)

• Many ways to move data to/from NERSC
  – And most of them are better than ‘scp’

• If in doubt, ask for help
  – www.nersc.gov -> “For Users”
  – ServiceNow (help.nersc.gov) or email (consult@nersc.gov)