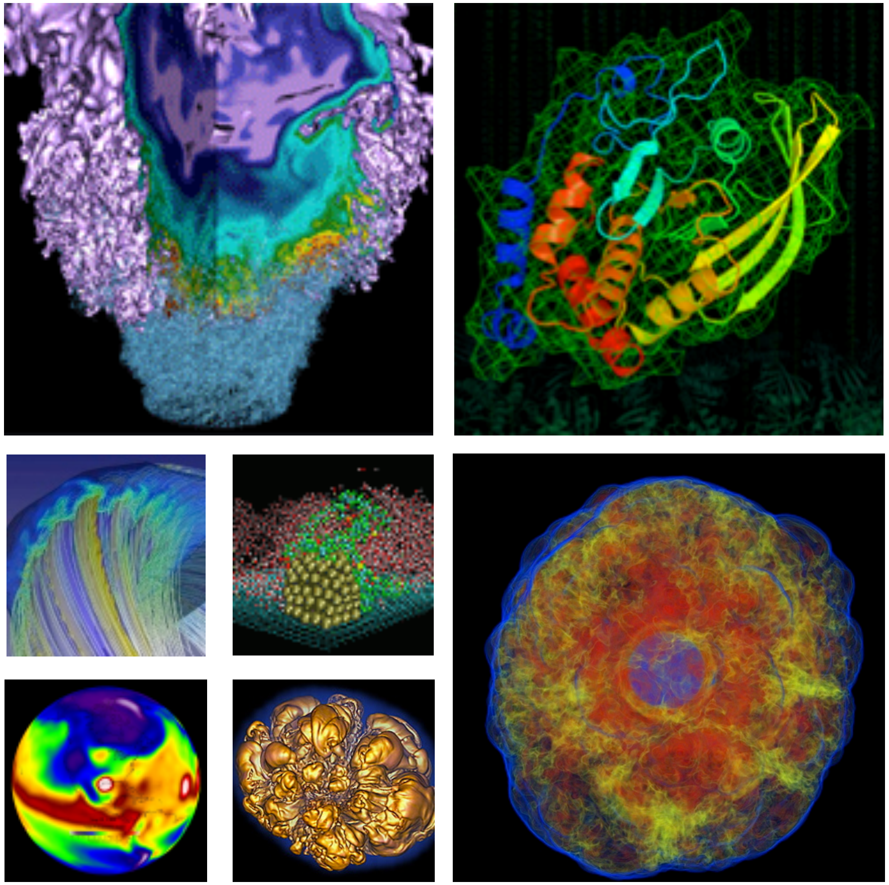
NERSC File Systems and Data Management



NERSC

Steve Leak

NERSC User Engagement Group

NUG New User Training

March 21, 2016

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

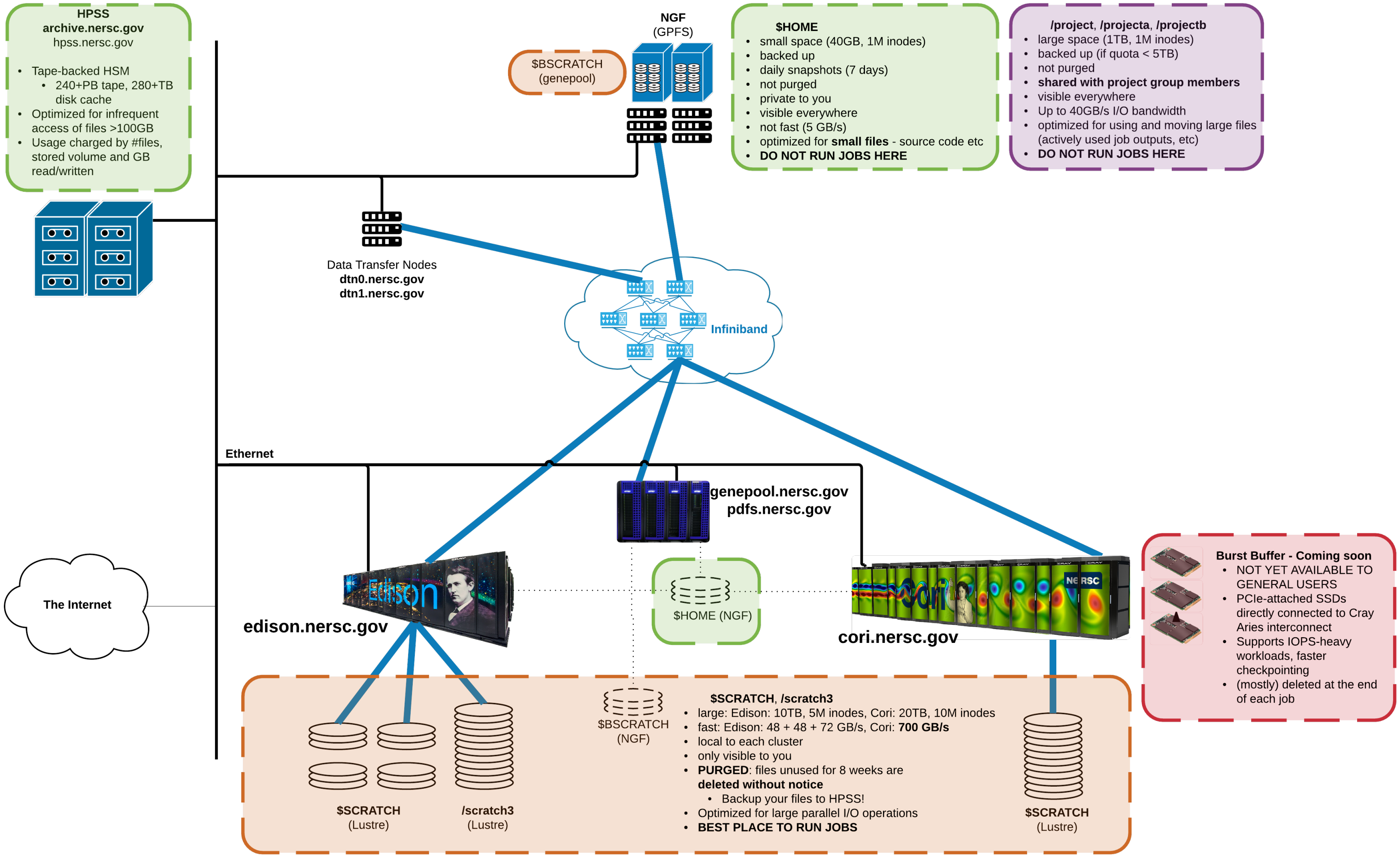
Topics

- **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)

$BSCRATCH
(genepool)

$HOME

- 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**

/project, **/projecta**, **/projectb**

- 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 (actively used job outputs, etc)
- **DO NOT RUN JOBS HERE**

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

genepool.nersc.gov
pdfs.nersc.gov

edison.nersc.gov

$HOME (NGF)

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

$BSCRATCH
(NGF)

$SCRATCH, **/scratch3**

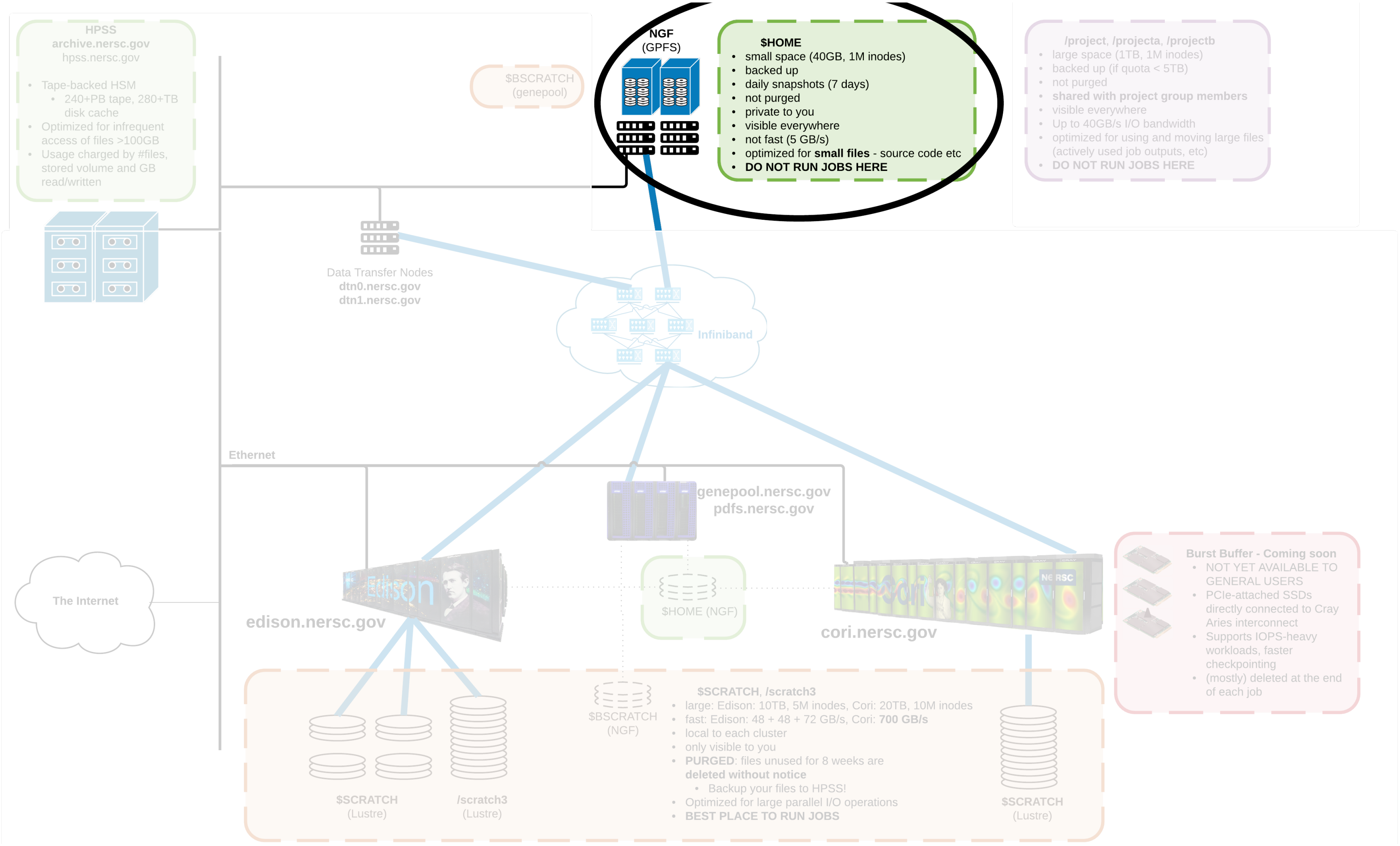
- large: Edison: 10TB, 5M inodes, Cori: 20TB, 10M inodes
- fast: Edison: 48 + 48 + 72 GB/s, Cori: **700 GB/s**
- local to each cluster
- only visible to you
- **PURGED**: files unused for 8 weeks are **deleted without notice**
  - Backup your files to HPSS!
- Optimized for large parallel I/O operations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

$SCRATCH
(Lustre)

/scratch3
(Lustre)

$SCRATCH
(Lustre)

NERSC Global $HOME



NGF
(GPFS)

$HOME

- 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**

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

$BSCRATCH
(genepool)

/project, /projecta, /projectb

- 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 (actively used job outputs, etc)
- **DO NOT RUN JOBS HERE**

Data Transfer Nodes
dtn0.nersc.gov
dtn1.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

$BSCRATCH
(NGF)

$SCRATCH, /scratch3

- large: Edison: 10TB, 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**

$SCRATCH
(Lustre)

/scratch3
(Lustre)

$SCRATCH
(Lustre)

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

NERSC Global $HOME

- **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  my_stuff
.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:
                  ---------- Space (GB) ---------  --------------- Inode ---------------
FileSystem             Usage      Quota    InDoubt        Usage        Quota      InDoubt
---------------- --------- --------- ---------  ----------- ----------- -----------
/global/cscratch          0      20480          -           51    10000000            -
HOME                      6         40          0       133431     1000000            0
```

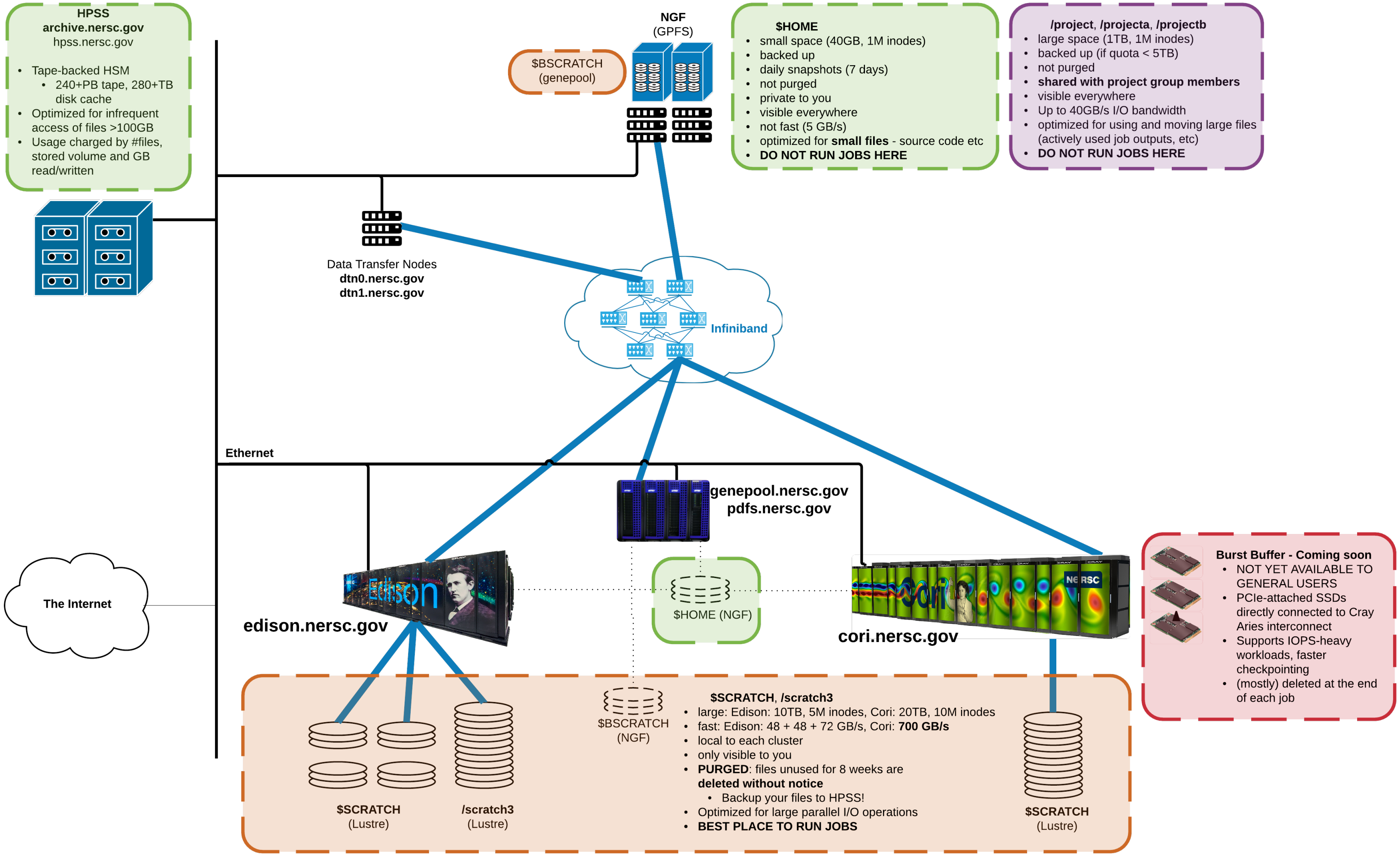
NERSC Global $HOME

- **Help! I deleted some large files, but my usage according to `myquota` 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
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)

$BSCRATCH
(genepool)

$HOME

- 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**

/project, /projecta, /projectb

- 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 (actively used job outputs, etc)
- **DO NOT RUN JOBS HERE**

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

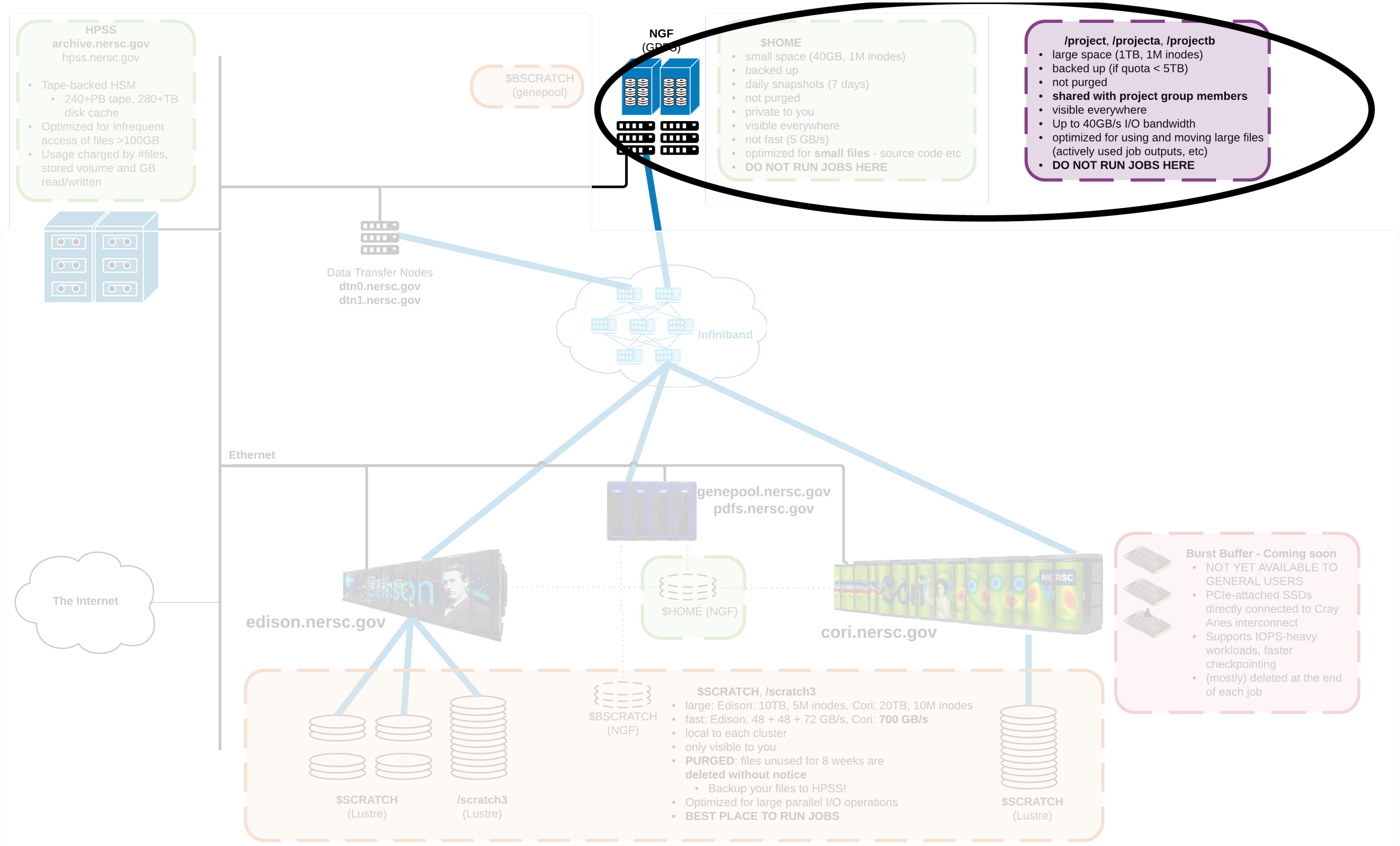
Burst Buffer - Coming soon

- 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: 10TB, 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**

Project File Systems



/project, /projecta, /projectb

- 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 (actively used job outputs, 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 (5 GB/s)
- optimized for **small files** - source code etc
- **DO NOT RUN JOBS HERE**

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

$SCRATCH, /scratch3

- large: Edison: 10TB, 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**

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

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

Project File Systems

- **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
  - – See http://www.nersc.gov/users/data-analytics/science-gateways/

Project File Systems

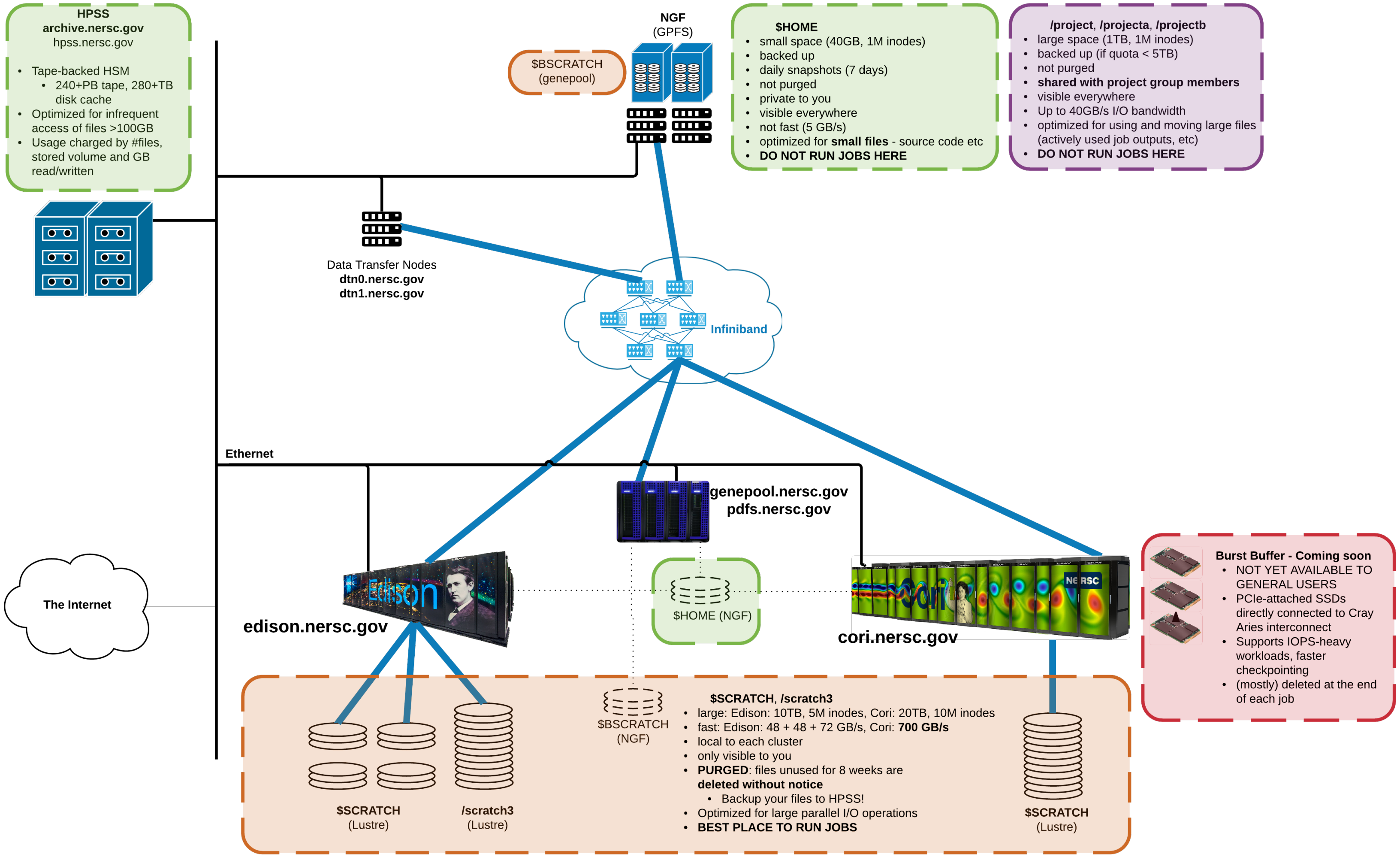
- **Quotas**
  - 1 TB
  - 1,000,000 inodes (i.e. files and directories)
  - Quota increases considered
    - http://www.nersc.gov/users/storage-and-file-systems/file-systems/disk-quota-increase-request/
  - Monitor your usage with `prjquota <yourproject>`
    - Also visible in NIM

```
sleak@cori03:~$ prjquota acme
                ------ Space (GB) -------     ----------- Inode -----------
      Project    Usage    Quota   InDoubt      Usage      Quota     InDoubt
     --------   -------  -------  -------     -------    -------    -------
         acme      1014     1024        0      899382    1000000          0
```

Project File Systems

- **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
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

$HOME

- 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**

/project, /projecta, /projectb

- 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 (actively used job outputs, etc)
- **DO NOT RUN JOBS HERE**

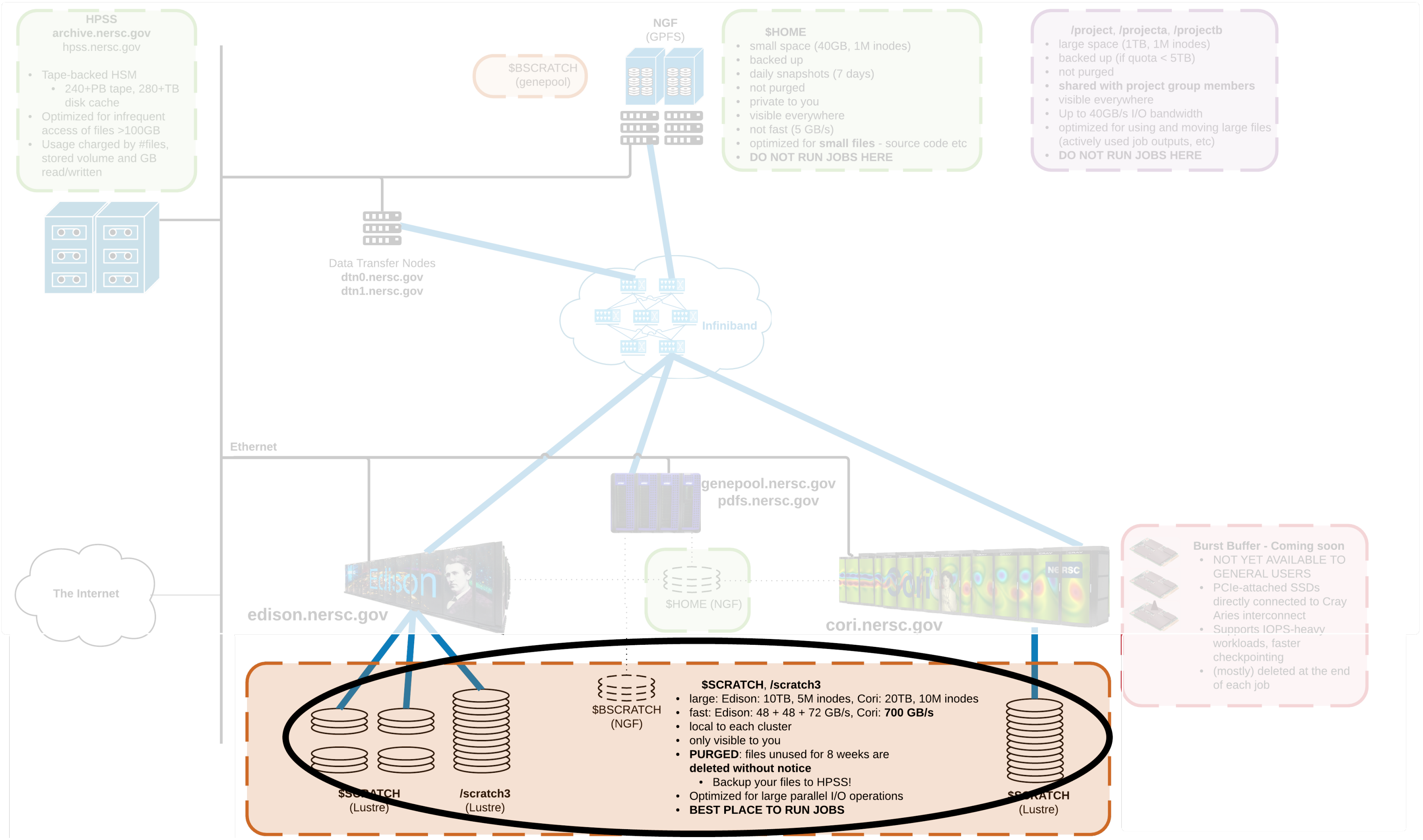
Burst Buffer - Coming soon

- 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 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 $SCRATCH



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

$HOME

- 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**

/project, /projecta, /projectb

- 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 (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 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 $SCRATCH

- **Local to each cluster**
- **Large**
  - Edison: 10 TB, 5,000,000 inodes
  - Cori: 20 TB, 10,000,000 inodes
- **FAST**
  - Edison $SCRATCH: 48 GB/s aggregate per filesystem
  - Edison /scratch3: 72 GB/s aggregate
  - Cori $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!

Local $SCRATCH

- **Quotas**
  - Edison: 10 TB, 5,000,000 inodes
  - Cori: 20 TB, 10,000,000 inodes
  - Quota increases considered
    - http://www.nersc.gov/users/storage-and-file-systems/file-systems/disk-quota-increase-request/
  - Monitor your usage with `myquota`
    - Also visible in NIM

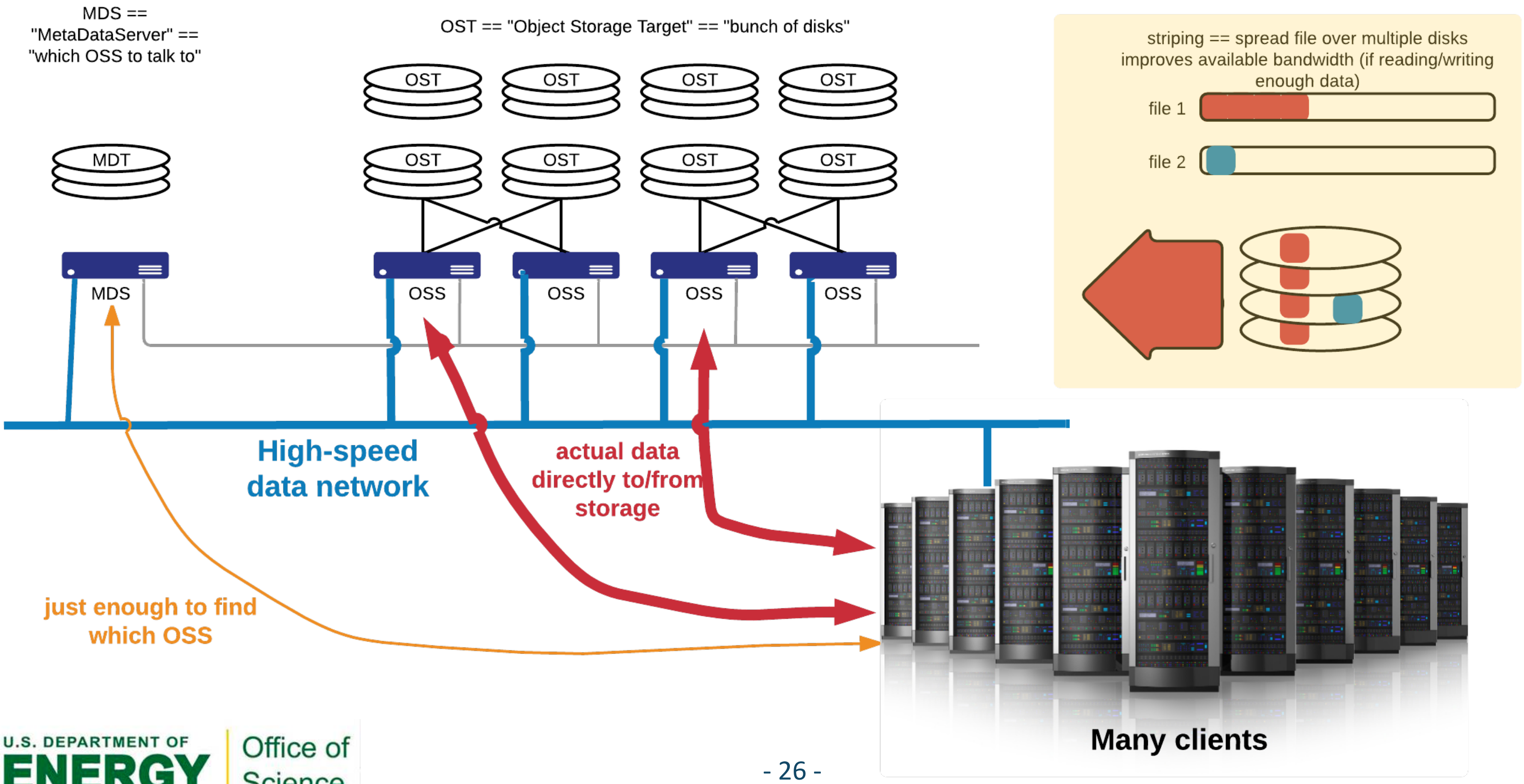
```
sleak@cori03:~$ myquota
Displaying quota usage for user sleak:
                  ---------- Space (GB) ---------  --------------- Inode ---------------
FileSystem           Usage      Quota     InDoubt      Usage        Quota        InDoubt
----------------  ---------  ---------  ---------  -----------  -----------  -----------
/global/cscratch          0      20480          -           51     10000000            -
HOME                      6         40          0       133431      1000000            0
```

Local $SCRATCH

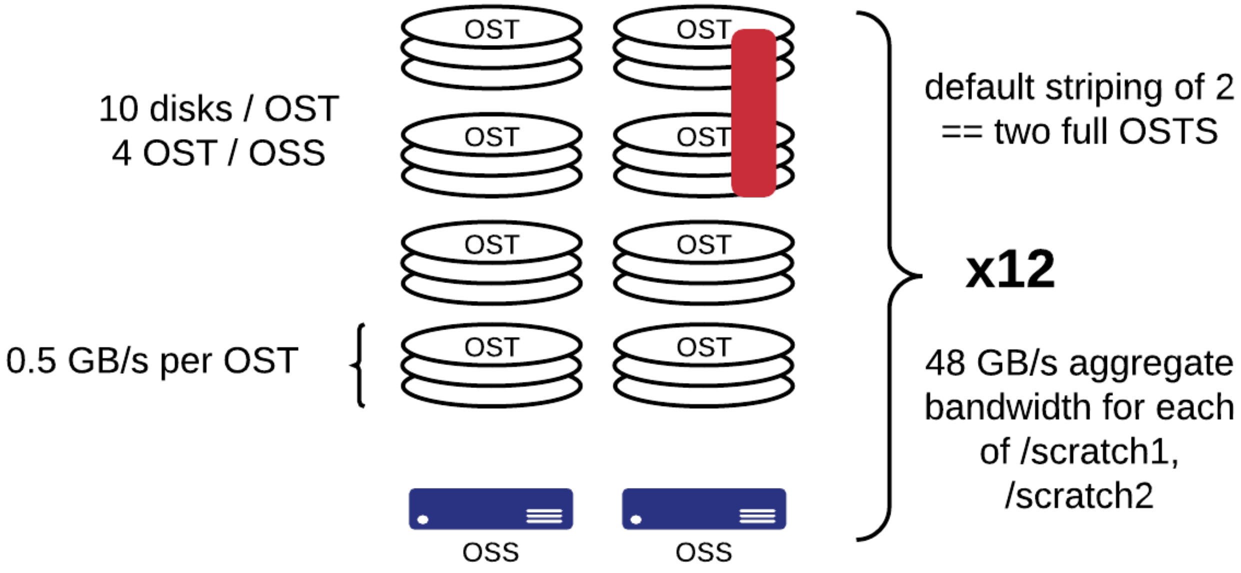
- **Lustre filesystem**
  - Edison: provided by two 2 PB filesystems
    - Users assigned randomly to one of them
  - Cori: single 28 PB filesystem
  - Access it with $SCRATCH
  - Edison /scratch3: access considered by request
    - http://www.nersc.gov/users/computational-systems/edison/file-storage-and-i-o/
    - Access it by name (/scratch3/scratchdirs/$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?

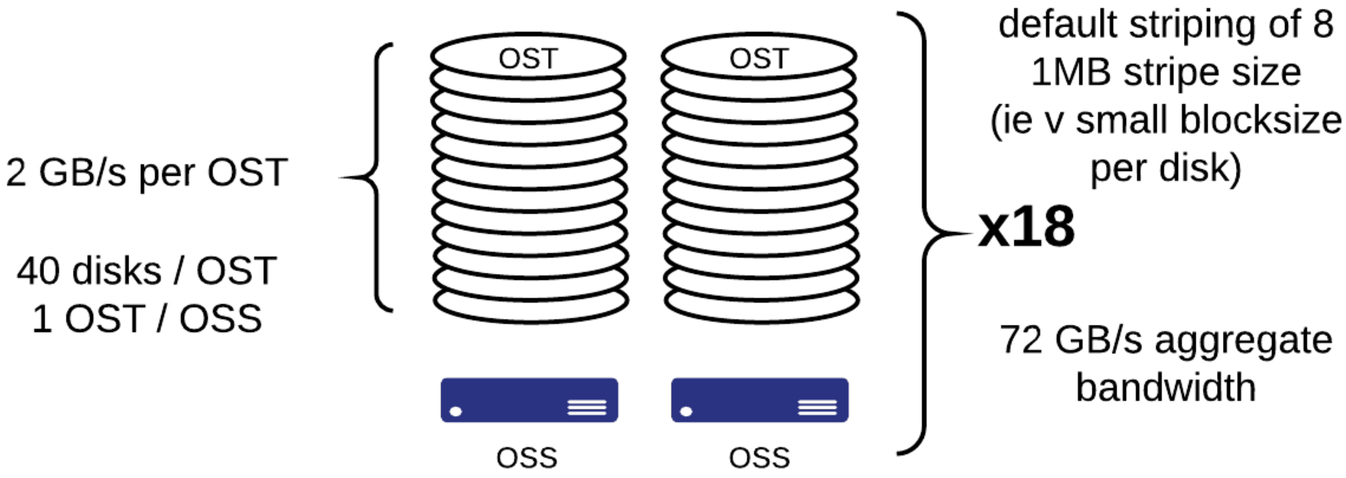

MDS == "MetaDataServer" == "which OSS to talk to"
OST == "Object Storage Target" == "bunch of disks"
OST
OST
OST
OST
OST
OST
OST
OST
MDT
MDS
OSS
OSS
OSS
OSS
striping == spread file over multiple disks improves available bandwidth (if reading/writing enough data)
file 1
file 2
High-speed data network
actual data directly to/from storage
just enough to find which OSS
Many clients

Edison $SCRATCH



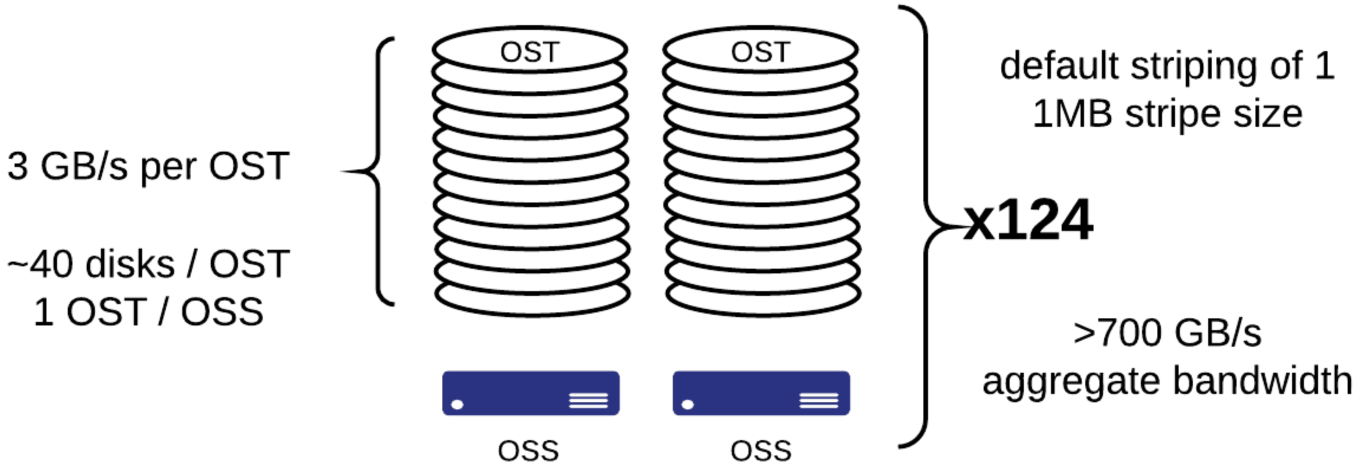
- **Tip: Cray MPI-IO is Lustre-aware**
  - Aggregator MPI tasks communicate each with 1 OST

Edison /scratch3



- **I/O striped over 8 OSTs of 40 disks each**
  - high I/O bandwidth

Cori $SCRATCH



- **Large space, highly parallel**
  - Eventually will become global scratch space

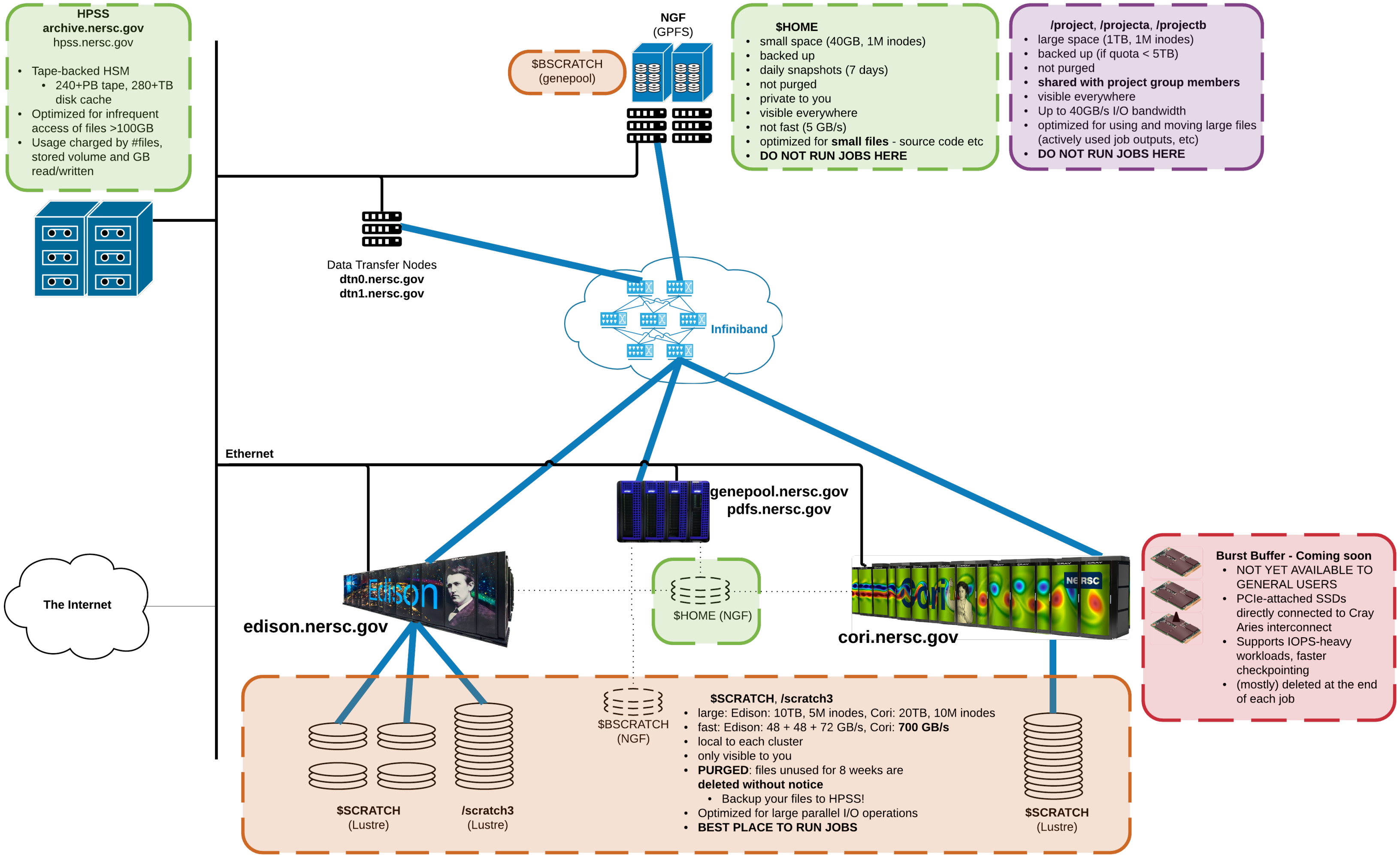
Optimizing I/O Performance

- **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

$HOME

- 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**

/project, **/projecta**, **/projectb**

- 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 (actively used job outputs, etc)
- **DO NOT RUN JOBS HERE**

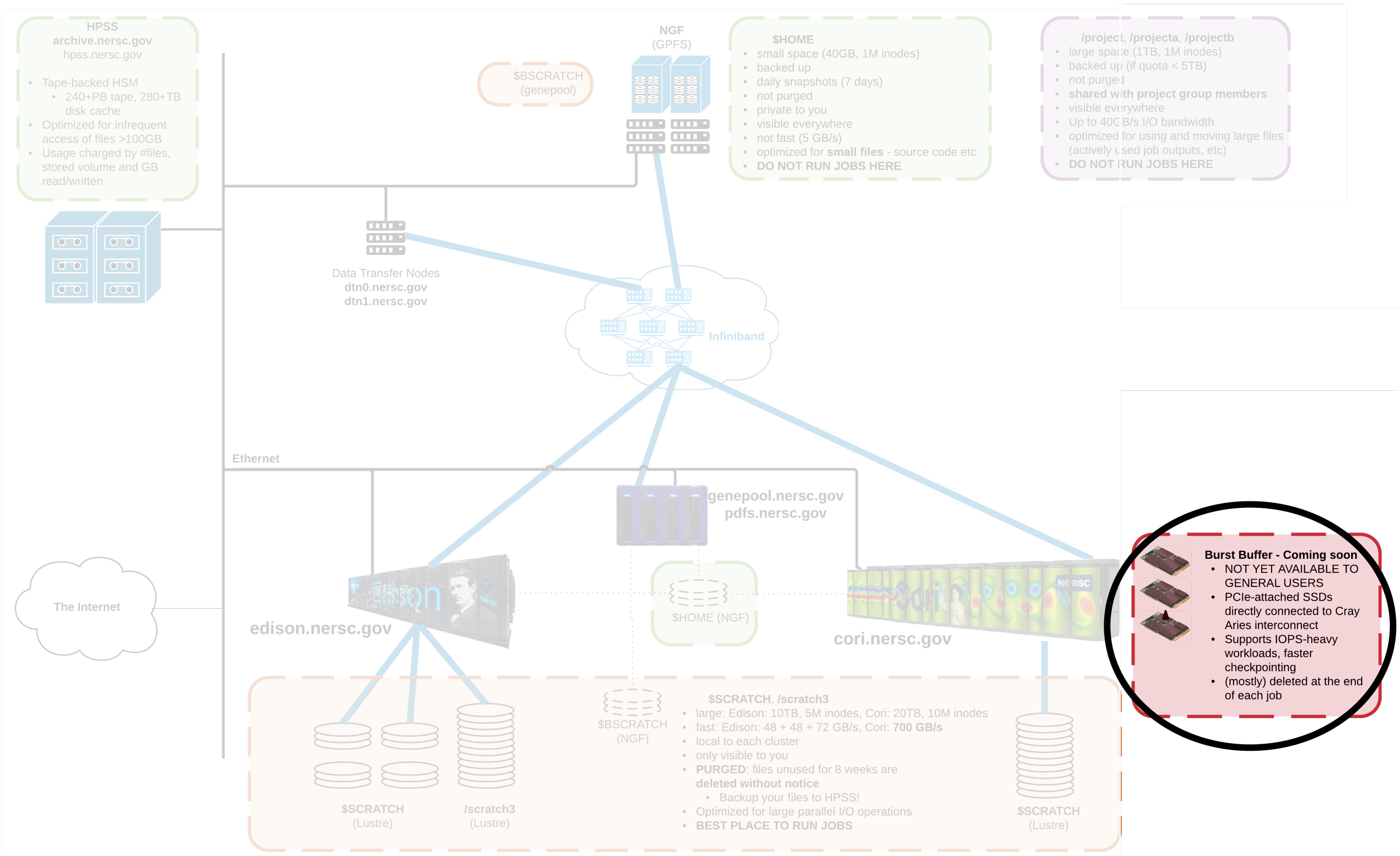
Burst Buffer - Coming soon

- 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 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**

Burst Buffer



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

$HOME

- 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**

/project, /projecta, /projectb

- 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 (actively used job outputs, etc)
- **DO NOT RUN JOBS HERE**

$SCRATCH, /scratch3

- large: Edison: 10TB, 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**

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

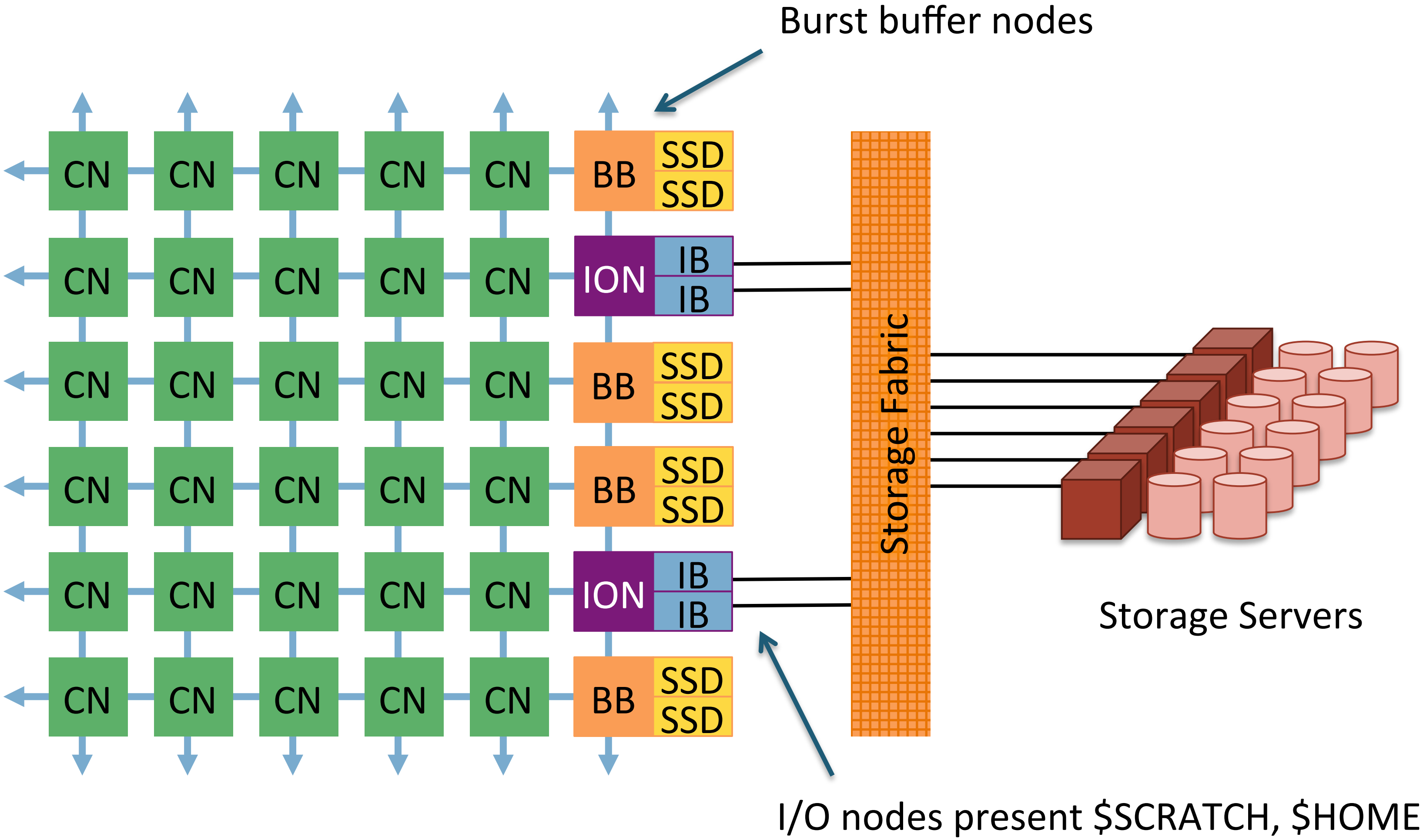
Burst Buffer

- **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

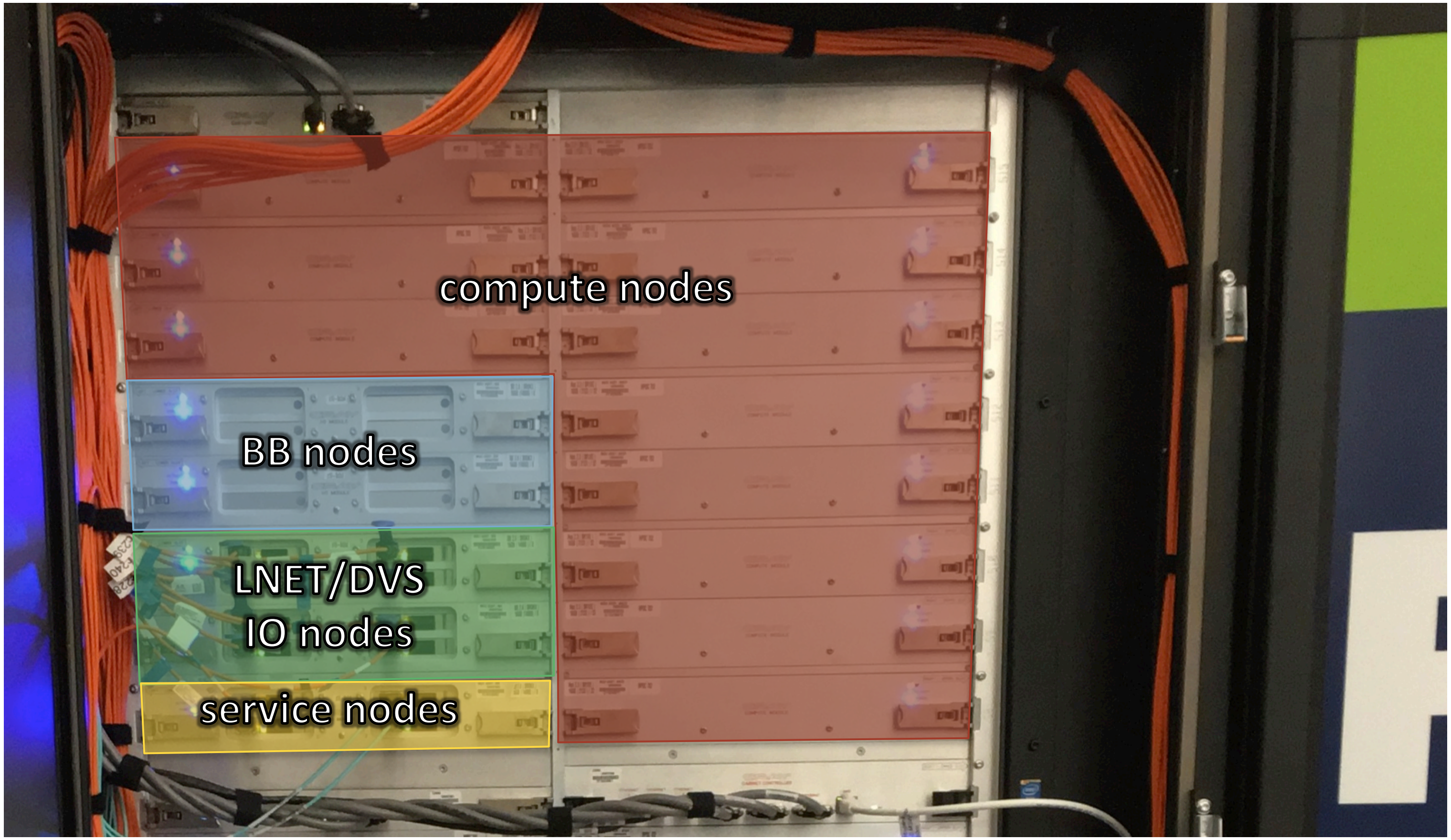
Burst Buffer

- **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

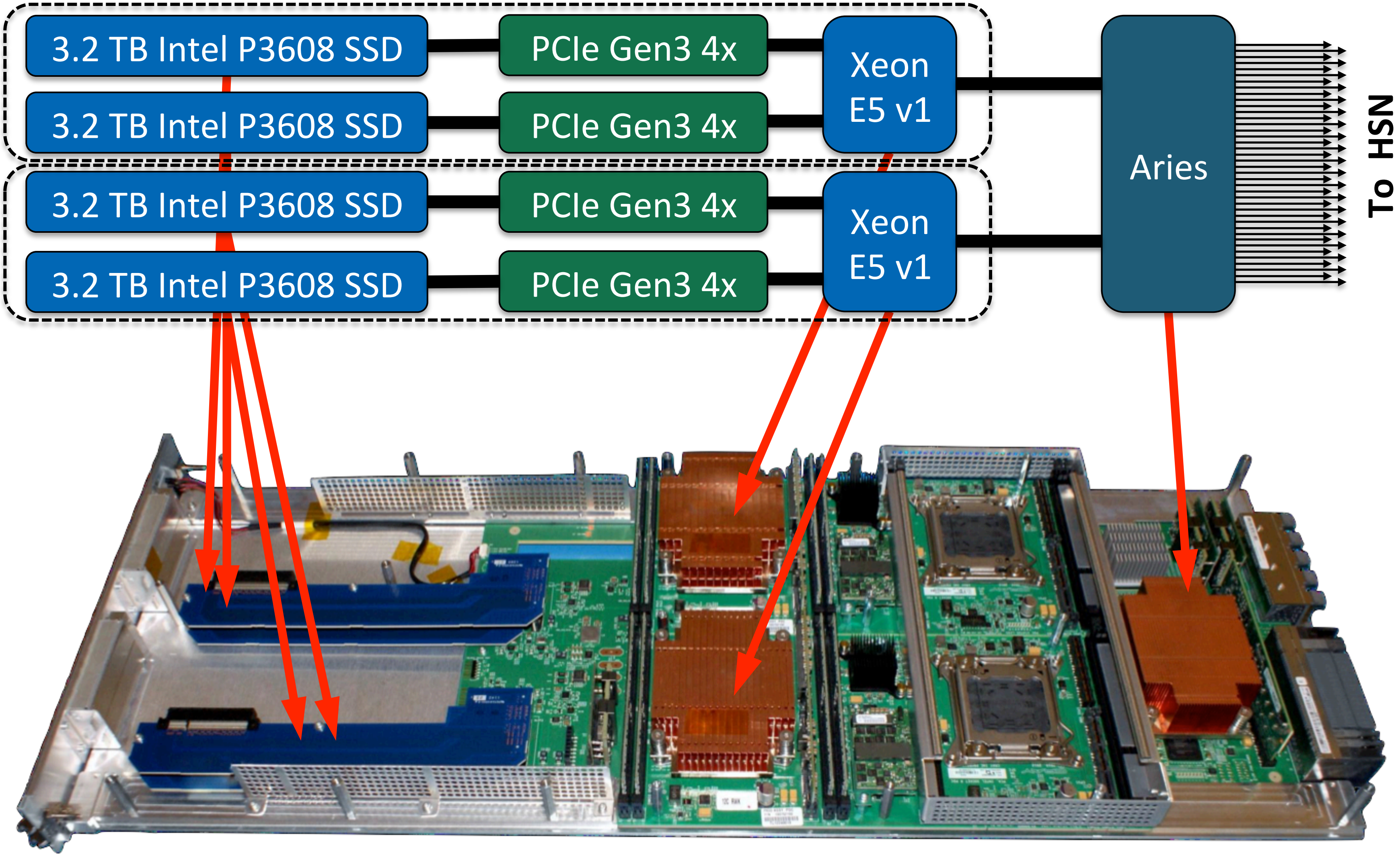
Burst Buffer



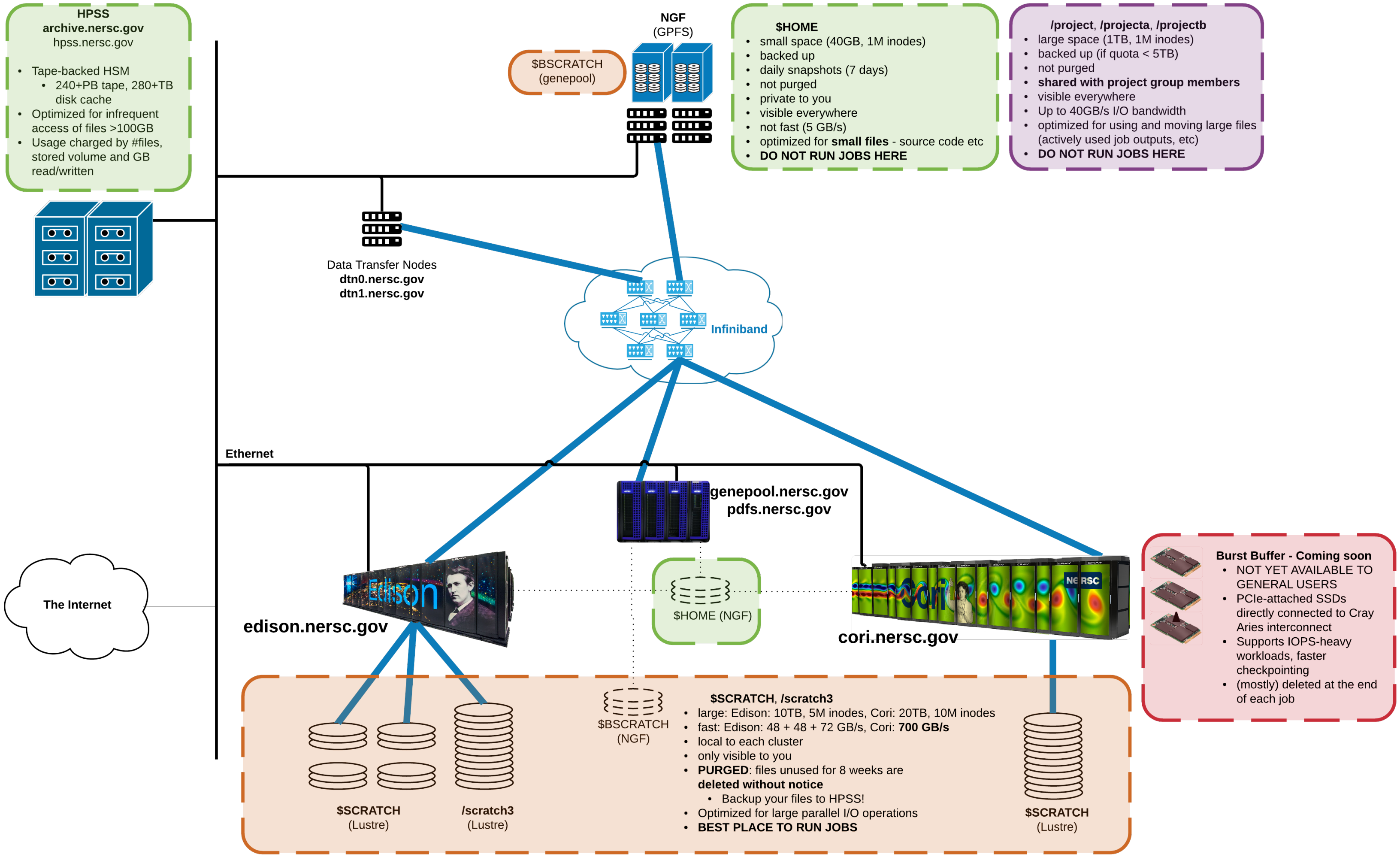
Burst Buffer



Burst Buffer



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

$HOME

- 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**

/project, /projecta, /projectb

- 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 (actively used job outputs, etc)
- **DO NOT RUN JOBS HERE**

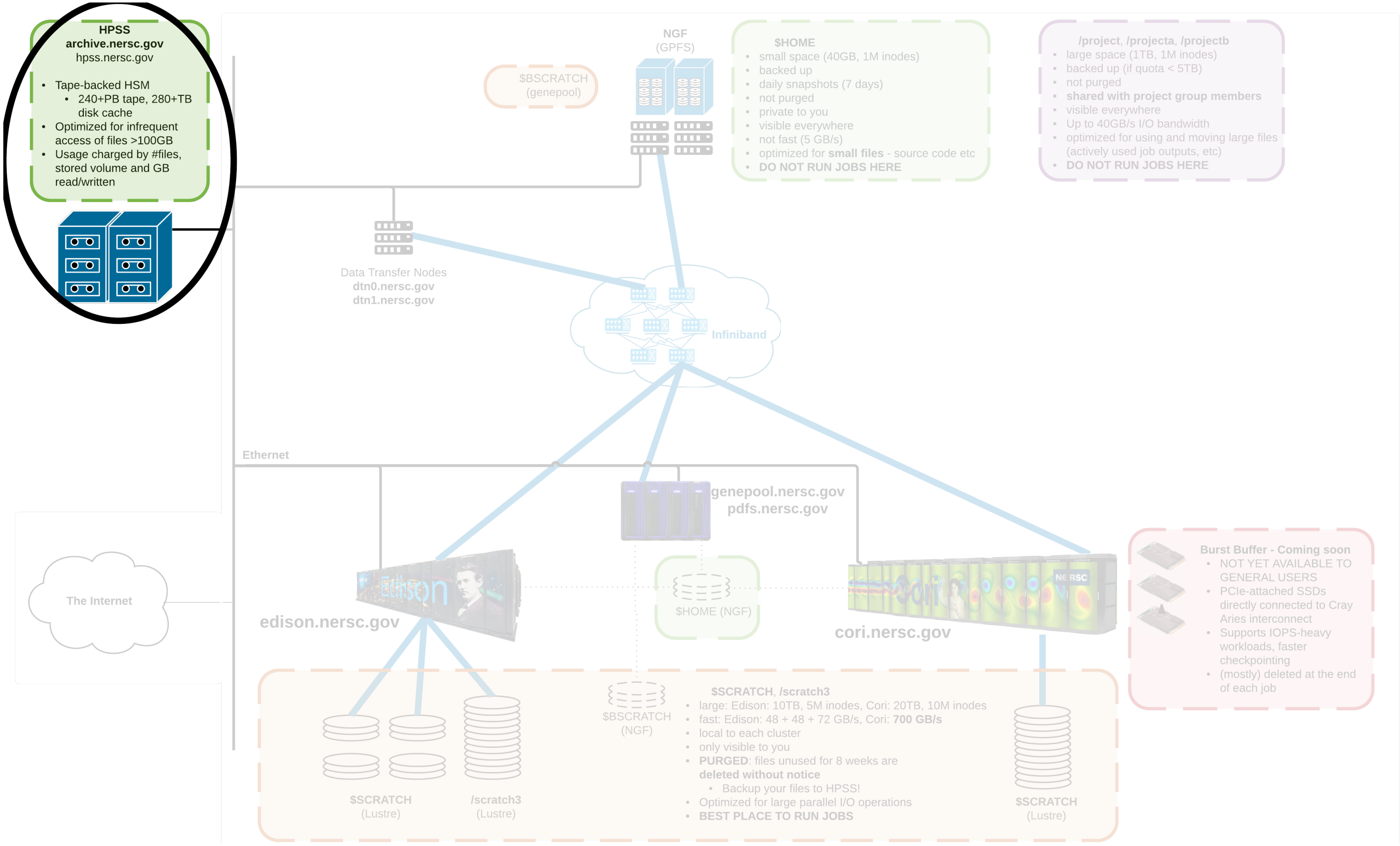
Burst Buffer - Coming soon

- 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 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**

HPSS



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

$HOME

- 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**

/project, /projecta, /projectb

- 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 (actively used job outputs, etc)
- **DO NOT RUN JOBS HERE**

Burst Buffer - Coming soon

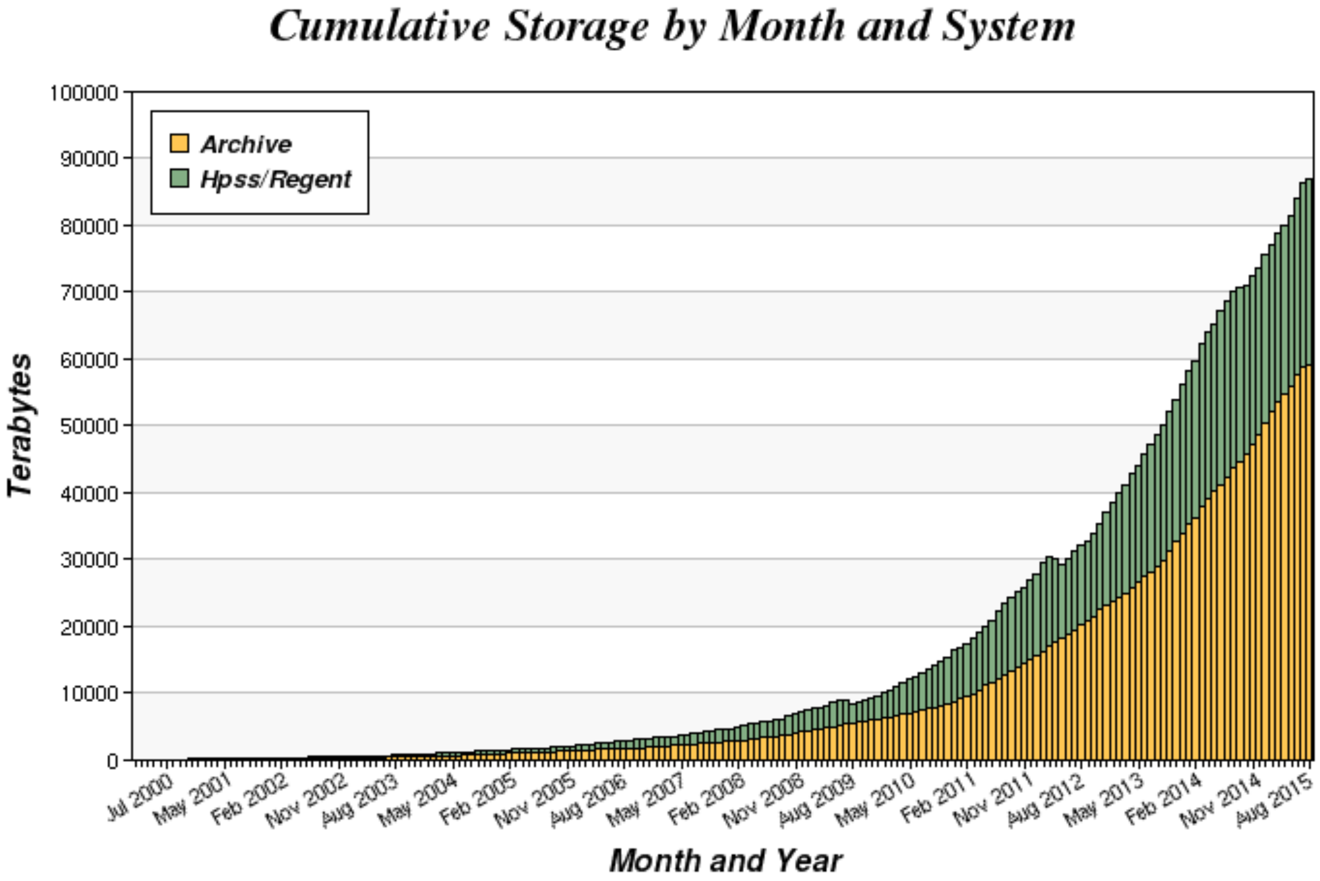
- NOT YET AVAILABLE TO GENERAL USERS
- PCIe-attached SSDs directly connected to Cray Aries interconnect
- Supports IOPS-heavy workloads, faster checkpointing
- (mostly) deleted at the end of each job

$SCRATCH, /scratch3

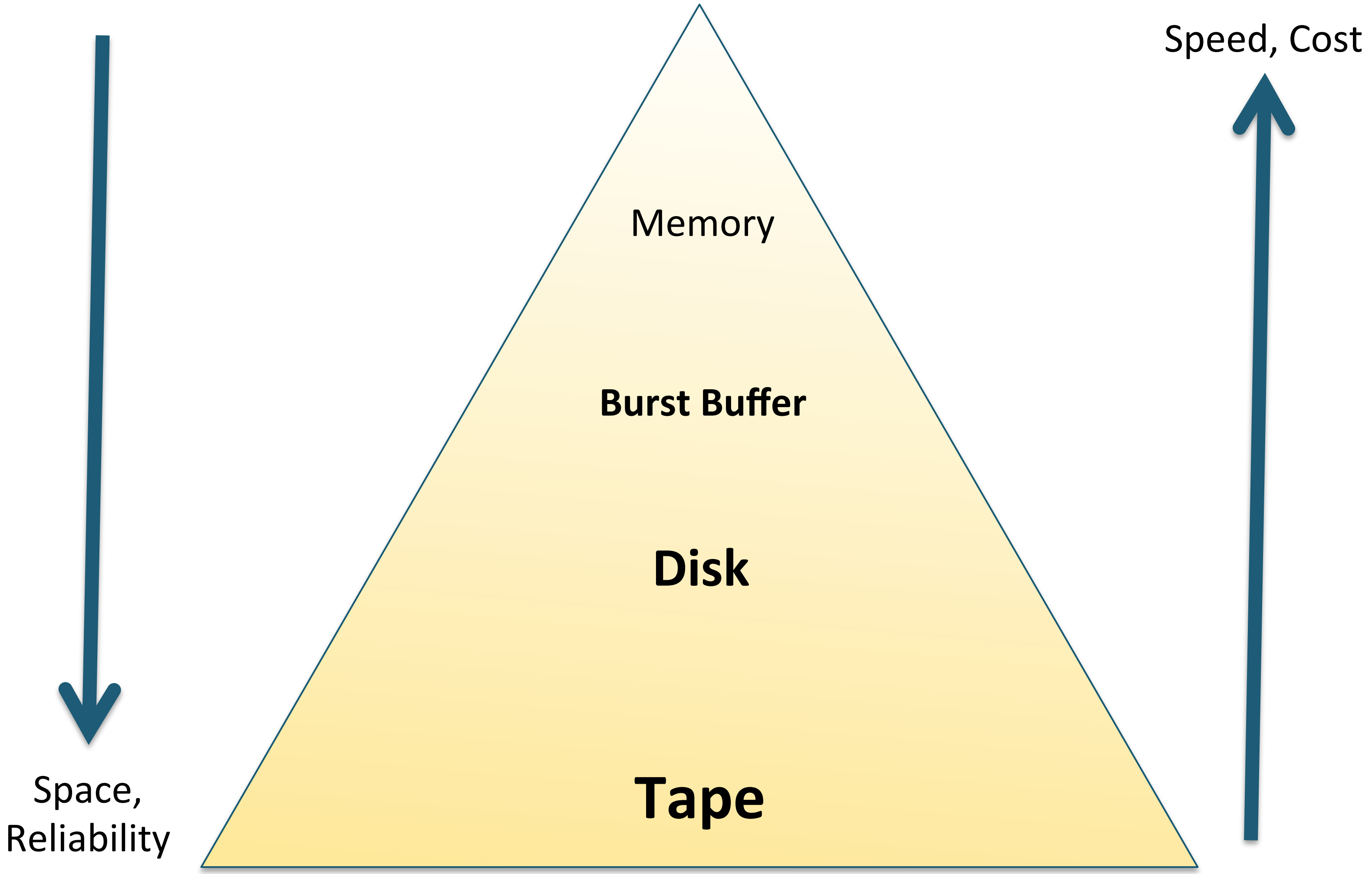
- large: Edison: 10TB, 5M inodes, Cori: 20TB, 10M inodes
- fast: Edison: 48 + 48 + 72 GB/s, Cori: **700 GB/s**
- local to each cluster
- only visible to you
- **PURGED**: files unused for 8 weeks are **deleted without notice**
  - Backup your files to HPSS!
- Optimized for large parallel I/O operations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

HPSS

- **Data grows exponentially**
  - 80% of stored data is never accessed again after 90 days



HPSS



Speed, Cost

Memory

Burst Buffer

Disk

Tape

Space, Reliability

HPSS

- **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”**
  - Function of volume-of-data-in-storage, number-of-files-in-storage and volume-of-data-transferred
    - Like Amazon Glacier, etc
  - Monitor usage via NIM

Accessing HPSS

Tool	What it does	Where/why to use it	Example
htar	Tar directly to/from HPSS	From NERSC hosts. Simple store/retrieve of large directories	`$ htar cf results-for-publication.tar my_results/`

hsi	CLI client	From NERSC hosts.
		Full featured client
			`$ hsi`
			`A:/home/s/sleak-> put myfile`

pftp, ftp	High performance (parallel) ftp	When need/prefer ftp-like interface	`$ pftp archive.nersc.gov`
			`ftp> pput results-for-publication.tar`

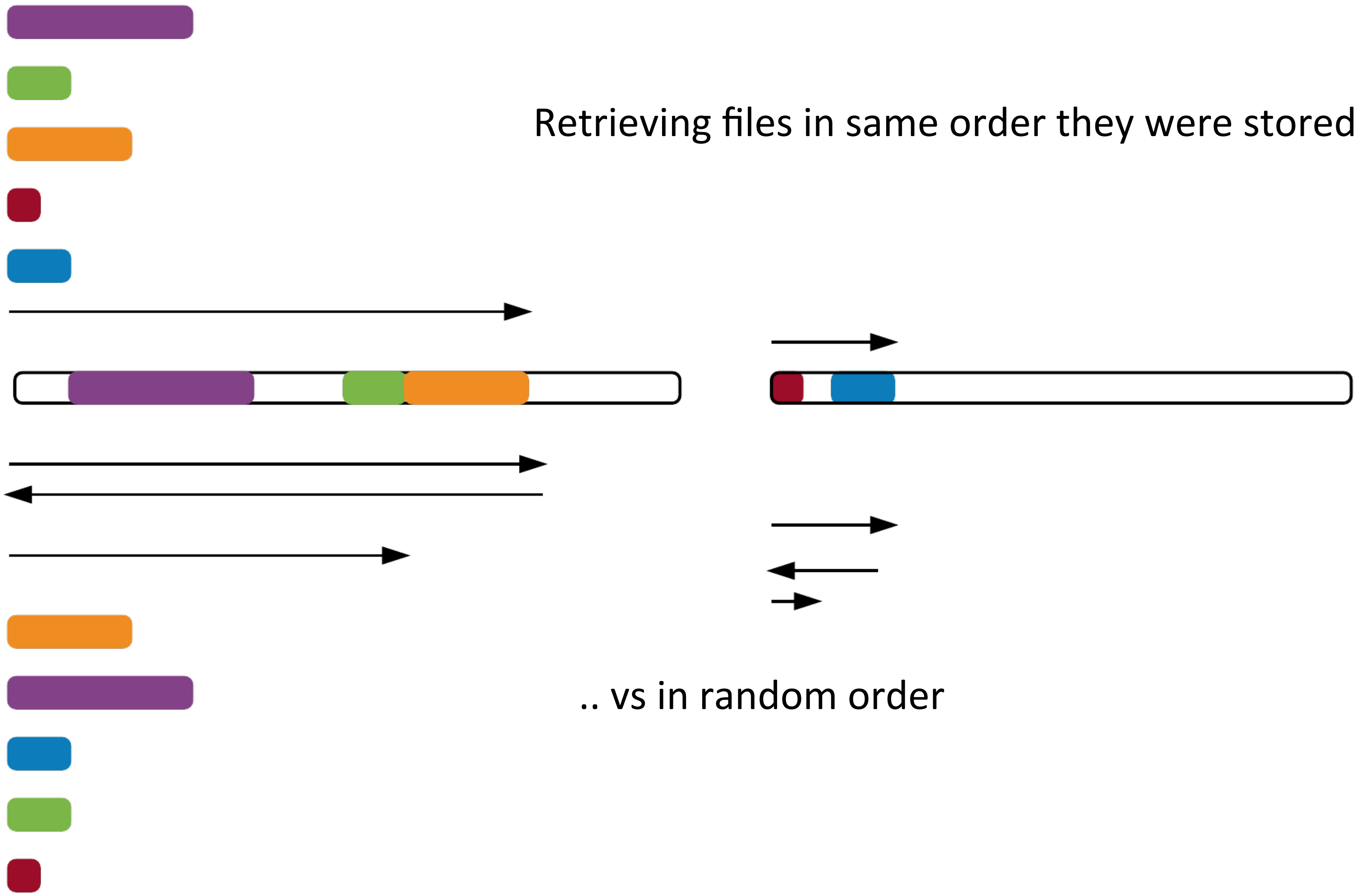
gridFTP		External, gridFTP-enabled sites (you need a grid credential) Note: **g**archive.nersc.gov	`$ globus-url-copy file://${HOME}/myresults.tar gsiftp://garchive.nersc.gov/home/s/sleak/results-for-publication.tar`
Globus Online	Data transfer service	Fire-and-forget transfers	See www.globusonline.org

HPSS

- **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)

HPSS

Retrieving files in same order they were stored …



.. vs in random order

HPSS

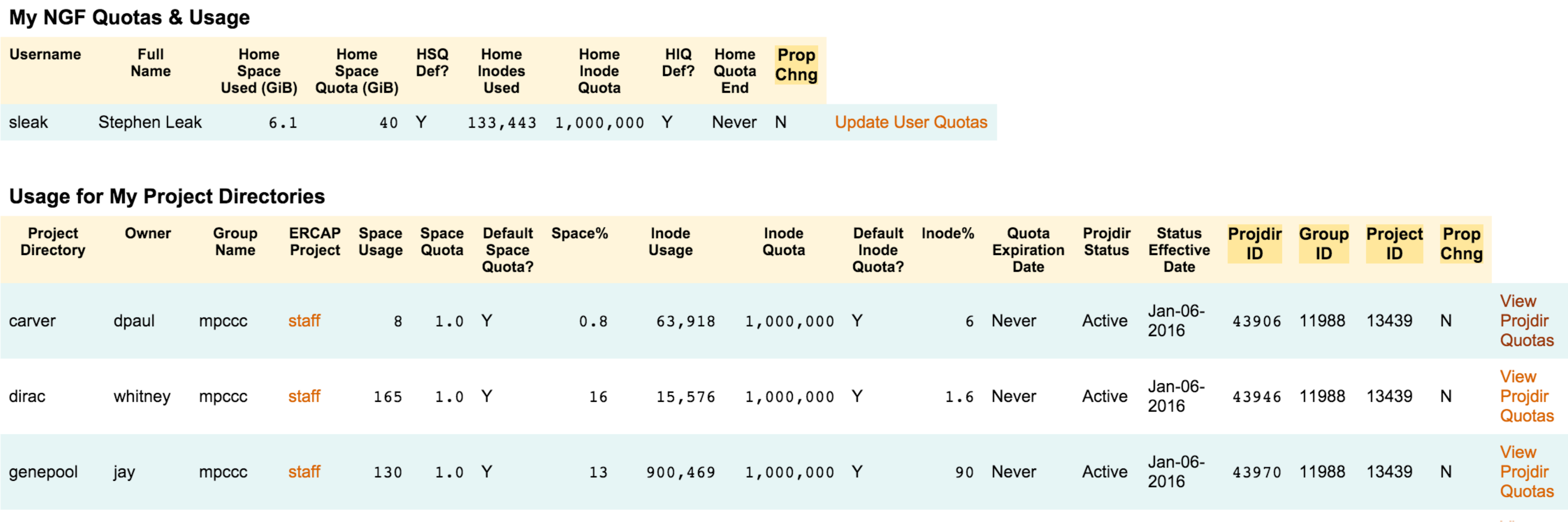
- **Best practices/Worst practices:**
  - http://www.nersc.gov/users/storage-and-file-systems/hpss/storing-and-retrieving-data/mistakes-to-avoid/
  - Store a few very large files, not many small files
    - htar or tar-first-in-$SCRATCH
  - Recursively storing or fetching a directory tree will result in many unordered accesses
    - Use htar or tar instead
    - hpss_file_sorter.script => sorts a list of files into “tape order”

HPSS

- **Best practices/Worst practices:**
  - http://www.nersc.gov/users/storage-and-file-systems/hpss/storing-and-retrieving-data/mistakes-to-avoid/
  - 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

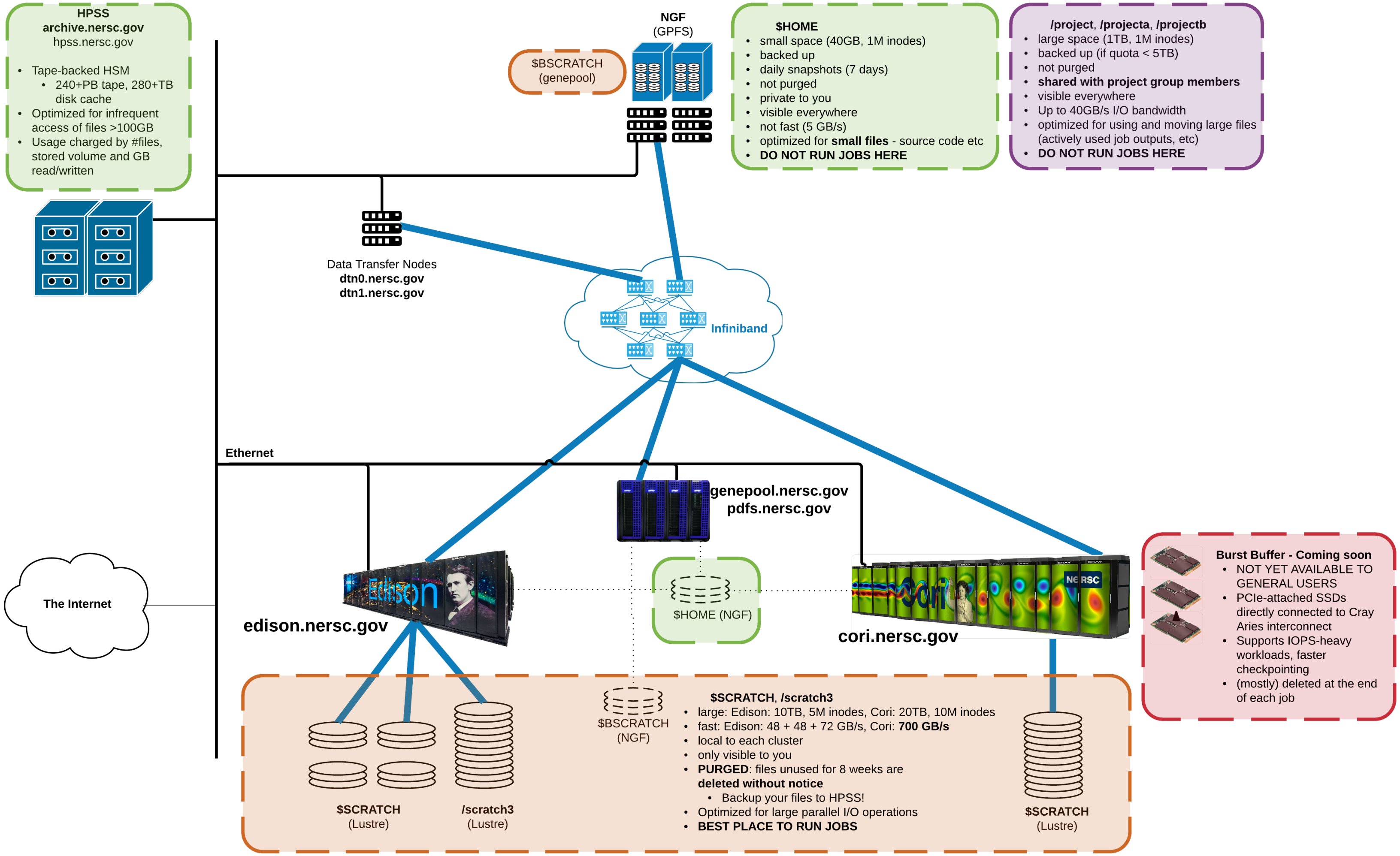
Username	Full Name	Home Space Used (GiB)	Home Space Quota (GiB)	HSQ Def?	Home Inodes Used	Home Inode Quota	HIQ Def?	Home Quota End	Prop Chng	
sleak	Stephen Leak	6.1	40	Y	133,443	1,000,000	Y	Never	N	Update User Quotas

Usage for My Project Directories

Project Directory	Owner	Group Name	ERCAP Project	Space Usage	Space Quota	Default Space Quota?	Space%	Inode Usage	Inode Quota	Default Inode Quota?	Inode%	Quota Expiration Date	Projdir Status	Status Effective Date	Projdir ID	Group ID	Project ID	Prop Chng	
carver	dpaul	mpccc	staff	8	1.0	Y	0.8	63,918	1,000,000	Y	6	Never	Active	Jan-06-2016	43906	11988	13439	N	View Projdir Quotas
dirac	whitney	mpccc	staff	165	1.0	Y	16	15,576	1,000,000	Y	1.6	Never	Active	Jan-06-2016	43946	11988	13439	N	View Projdir Quotas
genepool	jay	mpccc	staff	130	1.0	Y	13	900,469	1,000,000	Y	90	Never	Active	Jan-06-2016	43970	11988	13439	N	View Projdir Quotas

- myquota
- prjquota

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

$HOME

- 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**

/project, /projecta, /projectb

- 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 (actively used job outputs, etc)
- **DO NOT RUN JOBS HERE**

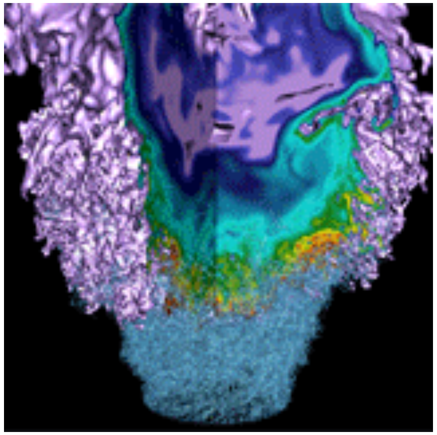
Burst Buffer - Coming soon

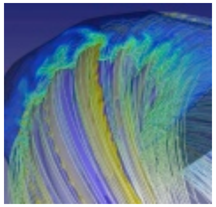
- NOT YET AVAILABLE TO GENERAL USERS
- PCIe-attached SSDs directly connected to Cray Aries interconnect
- Supports IOPS-heavy workloads, faster checkpointing
- (mostly) deleted at the end of each job

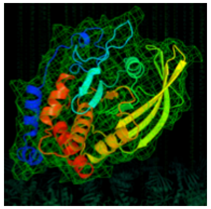
$SCRATCH, /scratch3

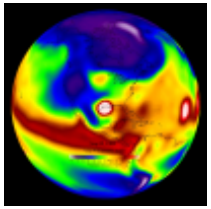
- large: Edison: 10TB, 5M inodes, Cori: 20TB, 10M inodes
- fast: Edison: 48 + 48 + 72 GB/s, Cori: **700 GB/s**
- local to each cluster
- only visible to you
- **PURGED**: files unused for 8 weeks are **deleted without notice**
  - Backup your files to HPSS!
- Optimized for large parallel I/O operations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

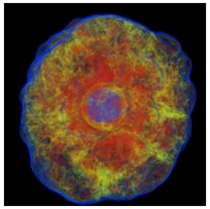
Sharing Data

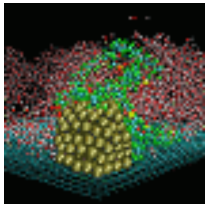












Sharing Data

- **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

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

```
$ setfacl -m u:rjhb:rw some_file.txt
$ getfacl some_file.txt
# file: some_file.txt
# owner: sleak
# group: sleak
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**

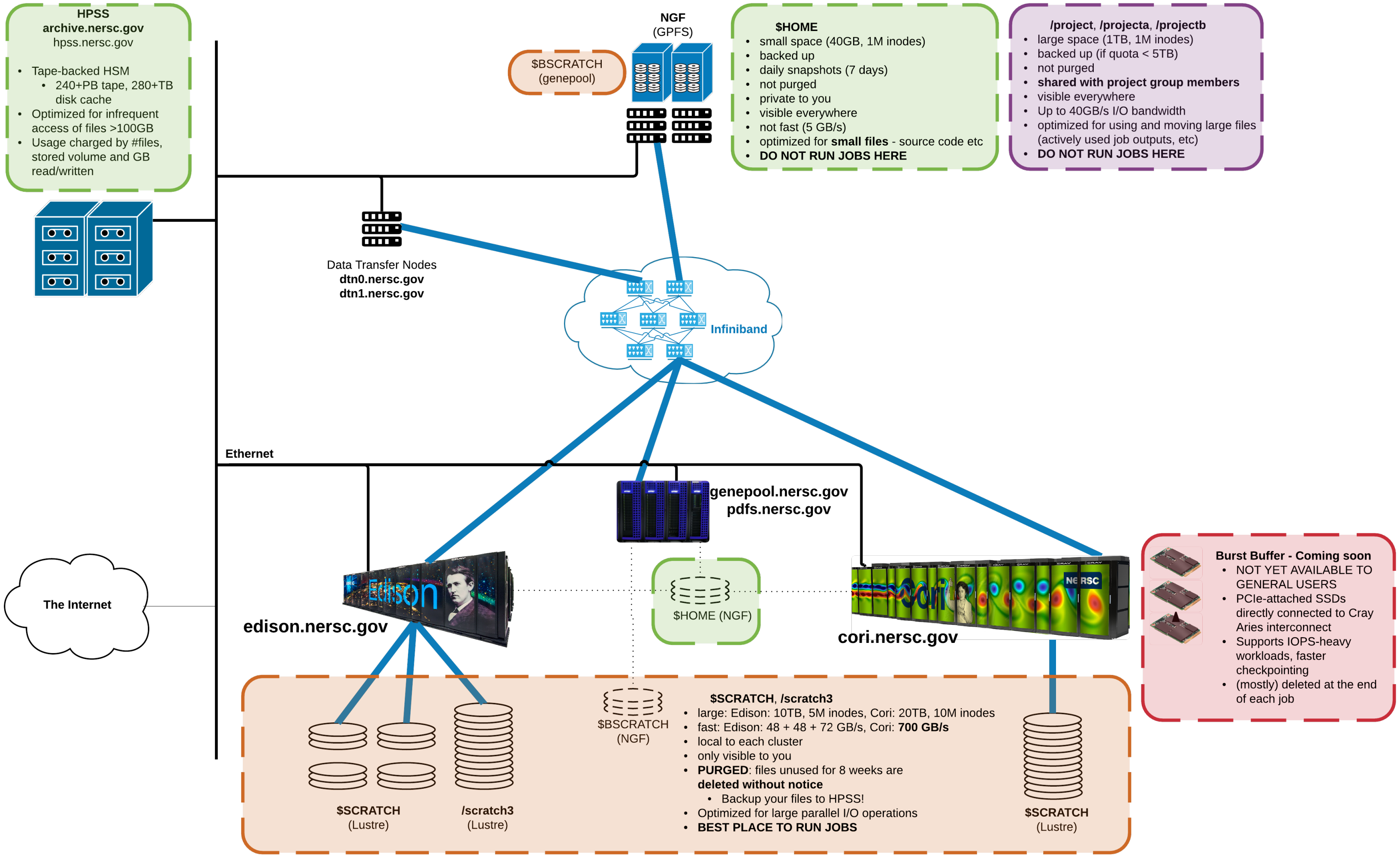
```
http://portal.nersc.gov/project/bigsci
```

- **More info:**
  - https://www.nersc.gov/users/data-analytics/science-gateways/

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

$HOME

- 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**

/project, /projecta, /projectb

- 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 (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 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**

Moving Data Around

Tool	What it does	Where/why to use it	Example
cp	Local copy	Between NERSC filesystems	`$ cp $SCRATCH/output.dat /project/projectdirs/m9999/`

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@dtn01.nersc.gov:/remote/path/"`

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

Moving Data Around

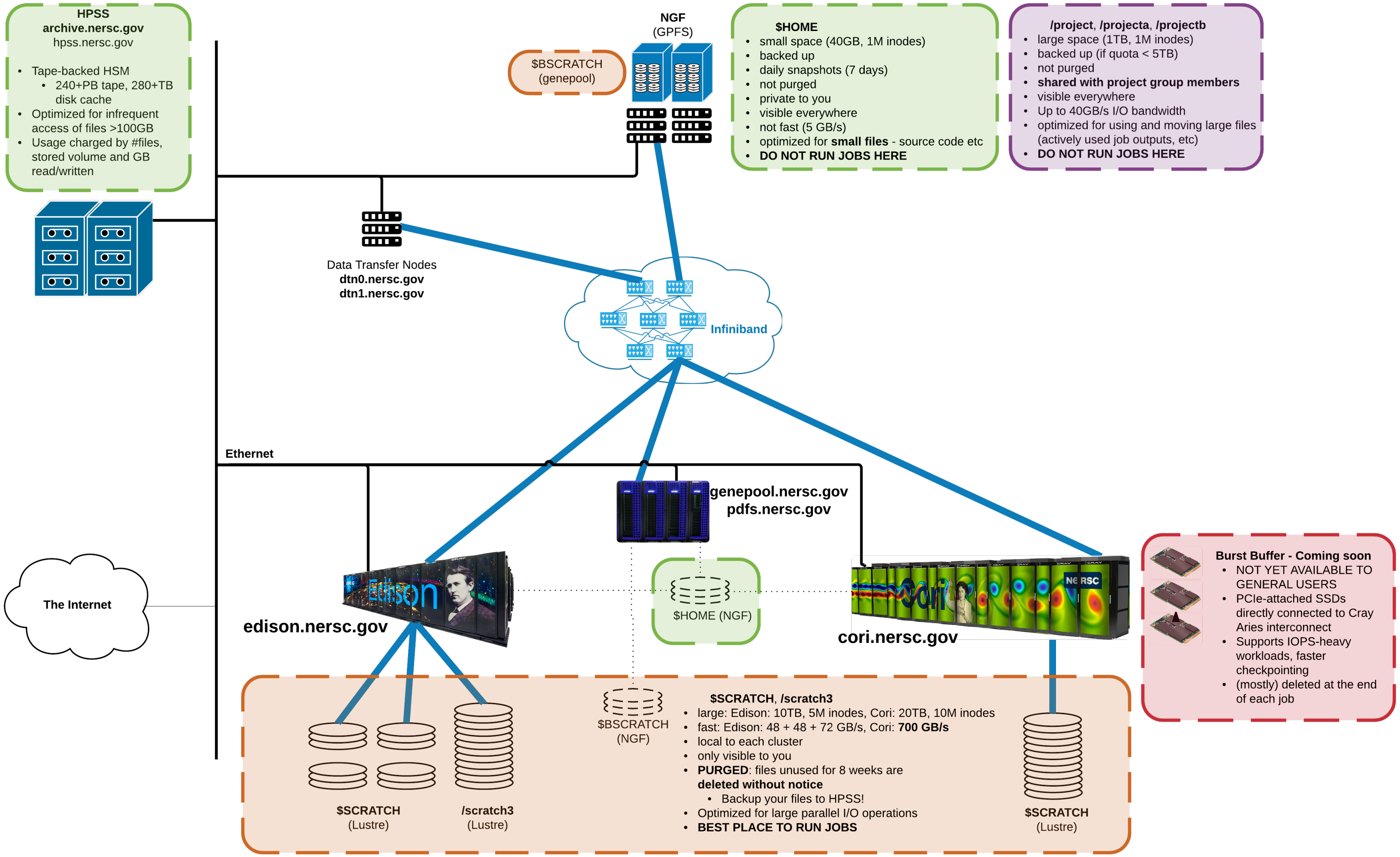
Tool	What it does	Where/why to use it	Example
NERSC ftp upload	Temporary ftp account/ server	Allow external collaborators to upload files for you to collect	See https://www.nersc.gov/users/storage-and-file-systems/transferring-data/nersc-ftp-upload-service/
gridFTP	Fast network copy protocol, requires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

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)

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)

$BSCRATCH
(genepool)

$HOME

- 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**

/project, /projecta, /projectb

- 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 (actively used job outputs, etc)
- **DO NOT RUN JOBS HERE**

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

genepool.nersc.gov
pdfs.nersc.gov

The Internet

edison.nersc.gov

$HOME (NGF)

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

$BSCRATCH
(NGF)

$SCRATCH, /scratch3

- large: Edison: 10TB, 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**

$SCRATCH
(Lustre)

/scratch3
(Lustre)

$SCRATCH
(Lustre)

NERSC

National Energy Research Scientific Computing Center