Navigating NERSC File Systems

May 3, 2011

David Turner
NERSC User Services Group

U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC National Energy Research Scientific Computing Center

BERKELEY LAB Lawrence Berkeley National Laboratory

Overview

- **Focus on user-*writable* file systems**
- **Global file systems**
- **Local file systems**
- **Policies**
- **Performance**
- **Data movement**
- **Data sharing**
- **Platform summary**
  - Hopper, Franklin, Carver/Magellan, Euclid, Datatrans (dtn01/dtn02)

Protect Your Data!

- Some file systems are backed up for disaster recovery (of entire file system).
- Some file systems are not backed up at all.
- Restoration of individual files/directories is *not* possible.
- Hardware failures and human errors *will* happen.

BACK UP YOUR FILES TO HPSS!

Global File Systems

- **NERSC Global Filesystem (NGF)**
  - Based on IBM's General Parallel File System (GPFS)
  - Architected and managed by NERSC's Storage Systems Group
  - Provides directories for home, global scratch, and project
  - Also provides /usr/common/

`/usr/common->/global/common/<platform>`

Global Homes Overview

- **Provided by two 20TB file systems**

  `/global/u1`

  `/global/u2`

- **Low-level name**

  `/global/u1/d/dpturner`

  `/global/u2/d/dpturner -> /global/u1/d/dpturner`

- **Better name**

  `/global/homes/d/dpturner`

- **Best name**

  `$HOME`

Global Homes Use

- **Shared across all platforms**
- **One common approach**
  - `$HOME/Hopper`, `$HOME/Euclid`, etc.
- **Another approach**
  - `$HOME/my_code/hopper_build`
  - `$HOME/my_code/euclid_build`
- **Tuned for small file access**
  - Compiling/linking
  - Job submission
    - But probably not job execution

Shell Initialization Files

- **Also known as "dot files"**
  - Set default PATH, MODULEPATH, etc
- **Read-only**

  `.bashrc->/global/homes/skel/read-only/.bashrc`

  `.cshrc->/global/homes/skel/read-only/.cshrc`
- **Customizable**

  `.bashrc.ext, .cshrc.ext`
- **Restore with `fixdots` command**
  - Makes `KeepDots.<timestamp>`
  - Might need `/usr/common/usg/bin/fixdots`
- **Debug with `.dbgdot` file**

  `touch .dbgdot`

Platform Detection

- **Environment variable or command**

```
$NERSC_HOST
nersc_host
```

- **csh example**

```
if ($NERSC_HOST == "carver") then
  module load openmpi
endif
```

- **bash example**

```
if [ $NERSC_HOST == "carver" ]; then
  module load openmpi
fi
```

Global Homes Policies

- **Quotas enforced**
  - 40GB
  - 1,000,000 inodes
  - Quota increases rarely granted
  - Monitor with **`myquota`** command
- **"Permanent" storage**
  - No purging
  - Hardware failures and human errors *will* happen.

BACK UP YOUR FILES TO HPSS!

Global Scratch Overview

- **Provides 873TB high-performance disk**
- **Primary scratch file system for Carver/ Magellan, Euclid, and Datatrans**
  - Also mounted on Hopper
- **Low-level name**

  `/global/scratch/sd/dpturner`
- **Better names**
  - **`$SCRATCH`** or **`$GSCRATCH`** on Carver/Magellan, Euclid, and Datatrans
  - **`$GSCRATCH`** on Hopper

Global Scratch Use

- **Shared across many platforms**
  - Example:
    - Generate data on Carver
    - Post-process data on Euclid
- **Tuned for large streaming file access**
  - Running I/O intensive batch jobs
  - Data analysis/visualization
  - 12GB/s aggregate bandwidth

Global Scratch Policies

- **Quotas enforced**
  - 20TB
  - 2,000,000 inodes
  - Quota increases may be requested
  - Monitor with `myquota` command
- **Temporary storage**
  - Weekly purges of *all* files that have not been accessed in over 12 weeks
    - Beginning in May, 2011

BACK UP YOUR FILES TO HPSS!

Project Overview

- **Provides 873TB high-performance disk**
  - First NGF deployment (70TB Oct 2005)
  - Recently redeployed on new hardware
  - Will expand to about 1.5PB by summer 2011
    - **Several hundred TB dedicated to two projects**
- **Widely available, including PDSF!**
- **Intended for sharing data between platforms, between users, or with the outside world**
- **Example names**

  `/project/projectdirs/m9999`

  `/project/projectdirs/bigsci`

Project Use

- **Tuned for large streaming file access**
  - Running I/O intensive batch jobs
  - Data analysis/visualization
  - Expect 12GB/s aggregate bandwidth with additional tuning
- **Access controlled by Unix file groups**
  - Group name usually same as directory
  - Requires administrator (usually the PI or PI Proxy)
    - **Project directories must be requested**
    - **Use NIM to add users to file group**

Science Gateways

- **Make data available to outside world**

```
mkdir /project/projectdirs/bigsci/www
cp files_to_share* /project/projectdirs/bigsci/www
chmod o+rx /project/projectdirs/bigsci
chmod –R o+rx /project/projectdirs/bigsci/www
```

- **Access with web browser**

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

- **Use NERSC Web Toolkit (NEWT) to develop more sophisticated web interfaces**
  - Authentication, system status
  - File upload/download, directory listings
  - Remote command execution, submit/monitor batch jobs
  - Accounting information, persistent object storage

Project Policies

- **Quotas enforced**
  - 4TB
  - 4,000,000 inodes
  - Quota increases may be requested
  - Monitor with `prjquota` command
- **"Permanent" storage**
  - No purging
  - Hardware failures and human errors *will* happen.

BACK UP YOUR FILES TO HPSS!

Local Scratch Overview

- Scratch file systems based on Lustre
- Each Cray has 2 scratch file systems
  - `/scratch`
  - `/scratch2`
  - On Hopper, each has 1PB
  - On Franklin, each has 208TB
- Each user has 2 scratch directories
  - `$SCRATCH`
    - `/scratch/scratchdirs/dpturner`
  - `$SCRATCH2`
    - `/scratch2/scratchdirs/dpturner`

Local Scratch Use

- **Not shared across *any* platforms**
  - In particular, Hopper scratch and Franklin scratch are distinct
- **Tuned for large streaming file access**
  - Running I/O intensive batch jobs
  - 35GB/s aggregate bandwidth (each file system)

Lustre Striping

- **Improve performance by spreading I/O across multiple servers**
  - Default is 2
  - Too low: not take advantage of available bandwidth
  - Too high: unnecessary overhead and loss of performance
  - NERSC provides shortcut commands
- **Set striping on directory *before* file creation**
  - To change striping of existing file, must copy to directory with desired striping

Striping Recommendations

File Size	Single File Command	File-per-processor Command
< 1GB	Use default	Use default, or **stripe_fpp**
1GB – 10GB	**`stripe_small`**	Use default, or **stripe_fpp**
10GB – 100GB	**`stripe_medium`**	Use default, or **stripe_fpp**
100GB – 1TB	**`stripe_large`**	Ask consultants

Local Scratch Policies

- **Quotas enforced**
  - If combined ($SCRATCH and $SCRATCH2) usage exceeds quota, can't submit batch jobs
  - 5TB
  - 5,000,000 inodes
  - Quota increases may be requested
  - Monitor with `myquota` command
- **Temporary storage**
  - Weekly purges of *all* files that have not been accessed in over 12 weeks

BACK UP YOUR FILES TO HPSS!

Data Movement

- **Use NGF to minimize movement and reduce duplication**
  - global home, global scratch, project
- **Use `cp`**
  - Within or between any NGF file system
  - Within or between **`scratch`** and **`scratch2`** on *either* Cray
- **Use scp or bbcp**
  - From either scratch on one Cray to either scratch on *other* Cray
  - See website for bbcp examples

File System Availability

System	Hopper	Franklin	Carver	Euclid	Datatrans	PDSF
Global home	Yes	Yes	Yes	Yes	Yes	
Global scratch	Yes		Yes	Yes	Yes	
Global project	Yes	Yes	Yes	Yes	Yes	Yes
Local scratch	Yes	Yes				

File System Summary

File System	Home	Local Scratch	Global Scratch	Project
Environment Variable	$HOME	$SCRATCH $SCRATCH2	$GSCRATCH	none
Scope	Global	Local	Global	Global
Default Quota	40GB 1M inodes	5TB 5M inodes	20TB 2M inodes	4TB 4M inodes
Intended Purpose	• dot files • source codes • compiling • input files	• batch jobs • I/O intensive • temporary storage of large files	•batch jobs •shared access •temporary storage of large files	•batch jobs •shared access •permanent storage of large files
Performance	100MB/sec	35GB/sec	12GB/sec	12GB/sec
Purged?	No	Yes	Yes	No

More Information

http://www.nersc.gov/users/data-and-networking/hpss/

http://www.nersc.gov/users/data-and-networking/file-systems/global-home/

http://www.nersc.gov/users/data-and-networking/file-systems/global-scratch/

http://www.nersc.gov/users/data-and-networking/file-systems/project-directories/

http://www.nersc.gov/users/computational-systems/hopper/file-storage-and-i-o/

http://www.nersc.gov/users/computational-systems/franklin/file-storage-and-i-o/

http://www.nersc.gov/users/data-and-networking/transferring-data/bbcp/

http://www.nersc.gov/users/data-and-networking/file-systems/disk-quota-increase-form/

http://www.nersc.gov/users/data-and-networking/file-systems/project-directory-request-form/