Managing Data in the HEC Environment

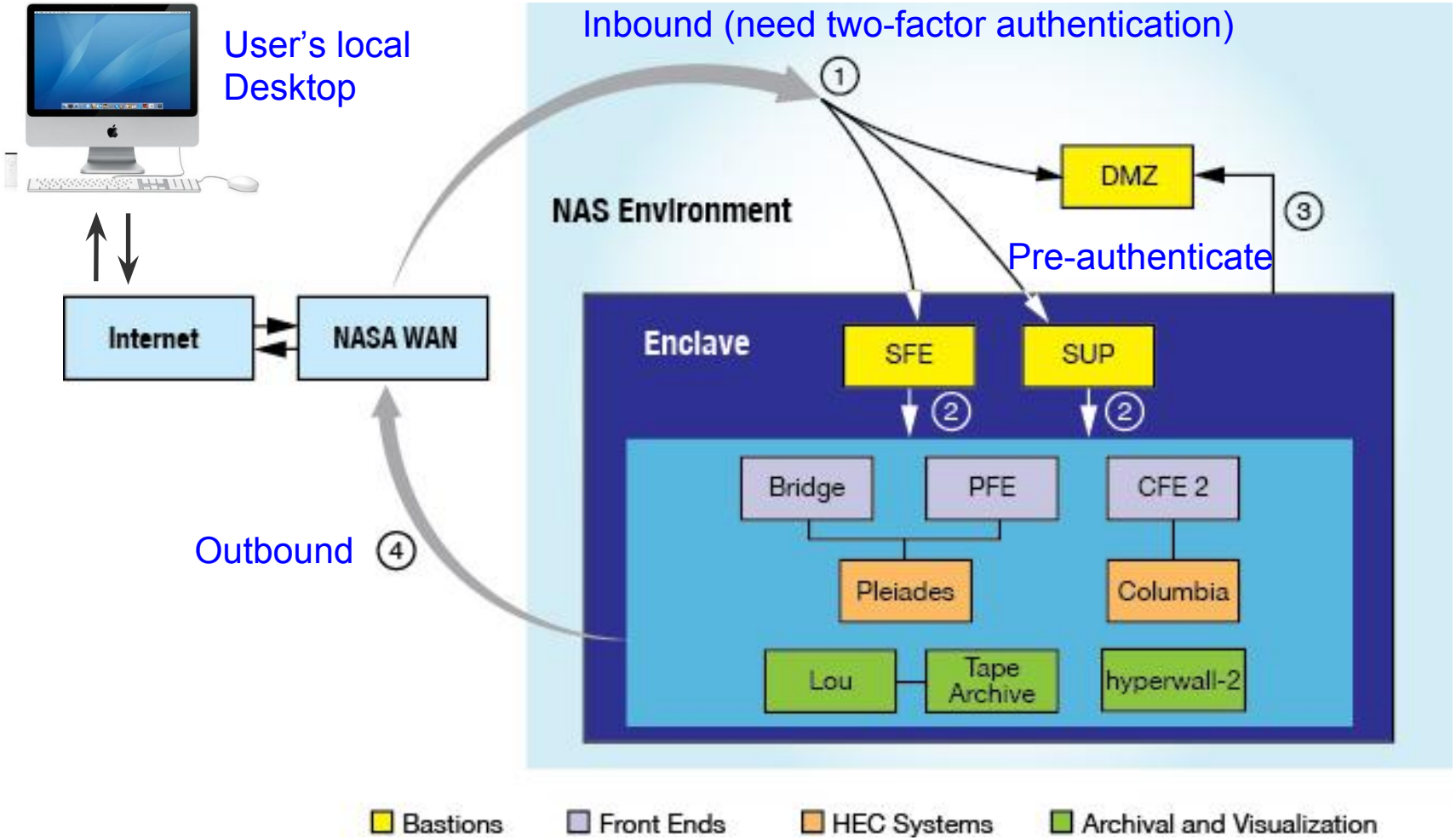
Sherry Chang
NAS Division
NASA Ames Research Center

2012 Summer Short Course for
Earth System Modeling and Supercomputing

Outline

- Handling Remote File Transfers
  - various transfer methods; ease of use; security; performance
- Managing Data for Active Use
  - different filesystems; quota; striping; tar; sharing files
- Managing Archive
  - archive servers; file integrity; DMF commands
- References in NAS Knowledge Base

Handling Remote File Transfers



Handling Remote File Transfers (cont'd)

- Download data from Web
  - use wget on a NAS host if you know the correct URL of the file

    local% ssh nas_username@pfe1.nas.nasa.gov

    pfe1% wget http://abc.gsfc.nasa.gov/file.zip
  - web browser -> go to URL -> click on file -> save to desktop

    -> remote transfers to pfe[1-12], bridge[1-4], lou[1-2]

- Transfer between localhost and pfe[1-12], bridge[1-4], lou[1-2]
  - Outbound (transfer command initiated at NAS)
    - If your localhost allows it: *scp, bbftp*
  - Inbound (transfer command initiated at your localhost)
    - Authenticate through sfe[1,2]: *scp, bbftp*

      Better to set up SSH Passthrough to allow one-step transfers
    - Pre-authenticate through sup: *sup scp, sup bbftp, sup shiftc* (use scp, bbftp, rsync underneath)

Handling Remote File Transfers (cont'd)

- scp (~ 2 MB/sec)
  - Use it for small files (< 1GB)
  - Encrypt password, passcode, and data
  - *local% scp foo pfe1.nas.nasa.gov:foo* (if nas_username is in .ssh/config)

    scp foo nas_username@pfe1.nas.nasa.gov:

    pfe1% scp local_username@localhost.domain:foo .

    pfe1% scp –r localhost.domain:mydir .

- bbftp (50 – 100 MB/sec)
  - Need bbftp server (for outbound) and client (for inbound) on localhost
  - Encrypt password, passcode, but NOT data
  - Recommend using multiple streams to get good transfer rates

    pfe1% bbftp –u local_username –e 'setnbstream 8; get foo'

    -E 'bbftpd -s -m 8' localhost.domain

    local% bbftp -u nas_username -e 'setnbstream 8; put foo'

    -E 'bbftpd -s -m 8' pfe1.nas.nasa.gov

    Cumbersome

    No recursive

http://www.nas.nasa.gov/hecc/support/past_webinars.html (Data Transfers, March 7, 2012)

Handling Remote File Transfers (cont'd)

- sup shiftc (for inbound only)
  - Need to download sup client to your localhost

    local% wget -O sup http://www.nas.nasa.gov/hecc/support/kb/file/9
  - Need to set up .meshrc on pfe's and lou's and authorize directories for writes

    pfe1% touch .meshrc

    pfe1% echo /u/username >>~/.meshrc (and also /nobackup/username)
  - Command is as simple as scp and can do recursive file transfers (even bbftp)

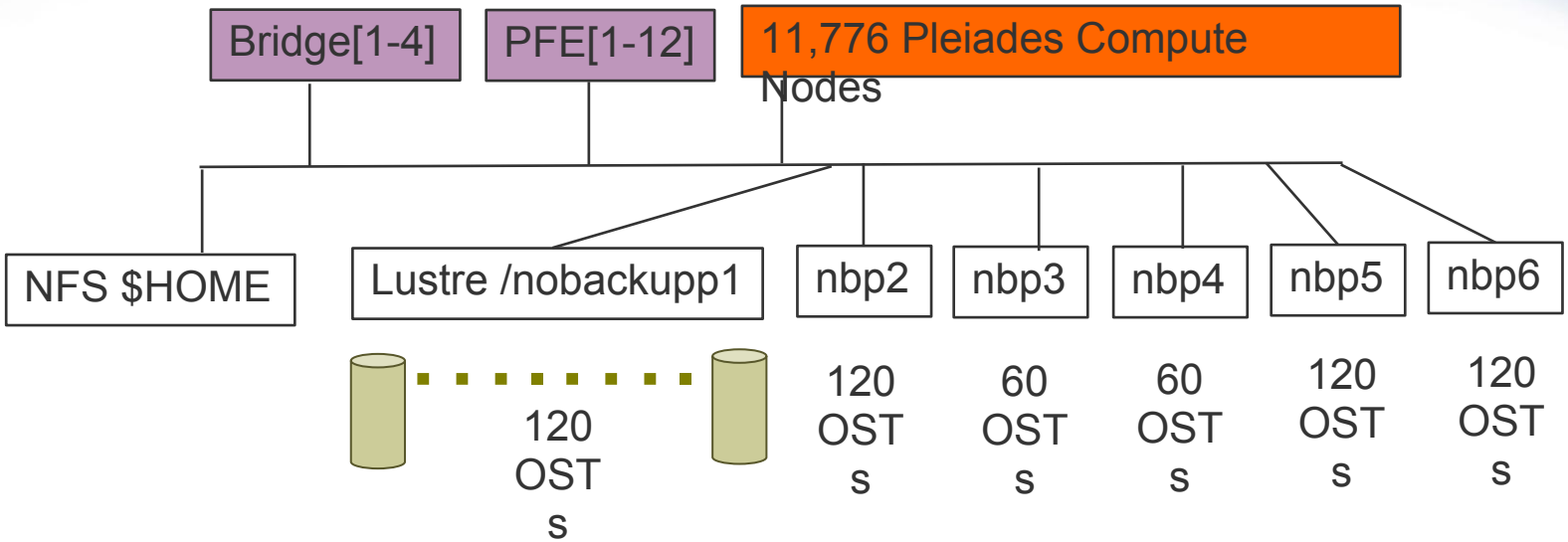
    local% sup shiftc –r mydir pfe1.nas.nasa.gov: (scp –r mydir pfe1.nas.nasa.gov:)
  - Automatic remote transport selection (sftp, rsync, bbftp with 8 streams

    [if installed on localhost, chosen as default under certain conditions, such as file size > 128 MB, no encryption asked, etc.])
  - Can include integrity verification (but the overall transfer task will be slower)

    local% sup shiftc --verify [--remote=bbftp] –r mydir nas_username@pfe1.nas.nasa.gov:
  - Once command is issued, you don't need to stay logged on localhost
  - Authentication is done automatically at first use and is not needed for subsequent transfers within 1 week

http://www.nas.nasa.gov/hecc/support/past_webinars.html (SUP/SHIFT, April 4, 2012)

Managing Data for Active Use

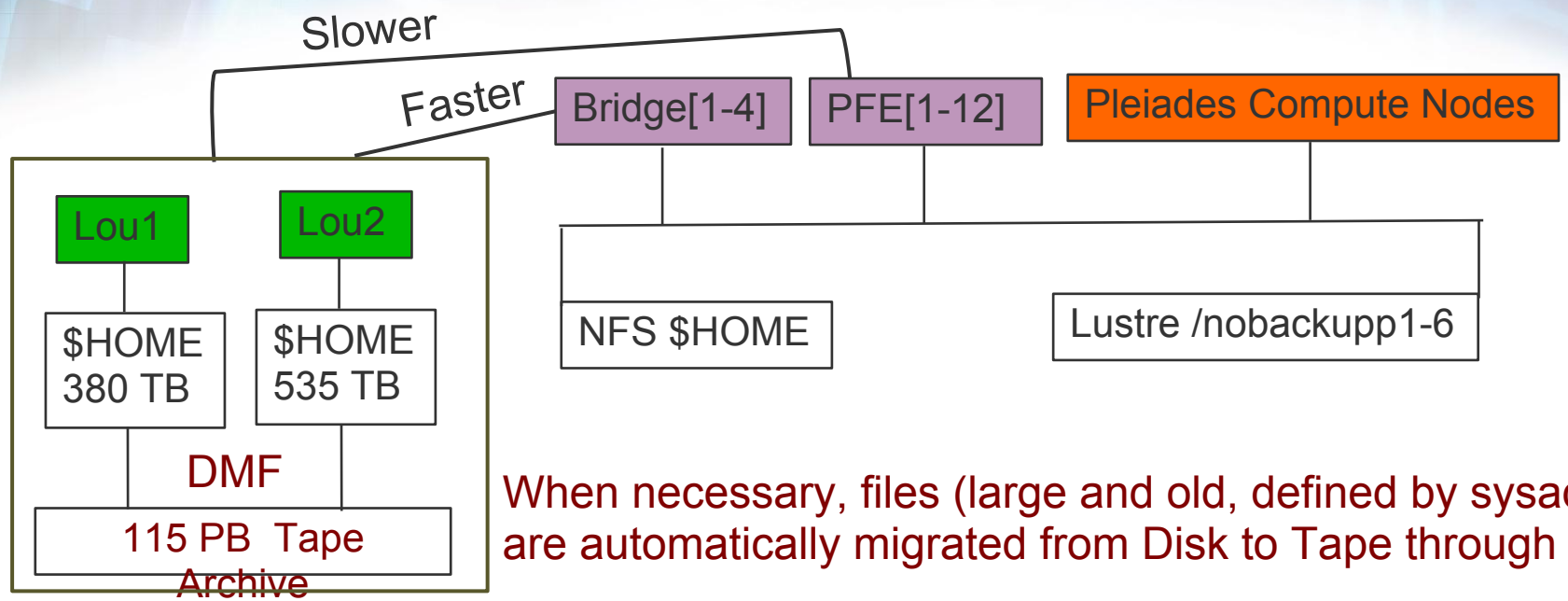


	HOME	Lustre
Directory Name	/u/username	/nobackup/username
Disk Quota	soft/hard: 8/10 GB	soft/hard: 500 GB/1TB
Inode Quota	None	soft/hard: 75,000/100,000
Quota Command	*quota -v*	*lfs quota –u username /nobackup/username*
Backup Policy	Daily	None
Default Stripe Count	N/A	1
Default Stripe Size	N/A	4 MB

Managing Data for Active Use (cont'd)

- Striping on Lustre Filesystems
  - Creating a new directory with a larger stripe count *pfe% mkdir new_dir*
    pfe% lfs setstripe -c 16 -s 4M new_dir
  - Changing the stripe count of an existing directory *pfe% cp –rp old_dir/* new_dir*
    pfe% rm –rf old_dir
    pfe% mv new_dir old_dir
- tar and mtar
  - Creating or extracting a tar file *pfe% tar cvf my.tar * (or tar cvf my.tar .)*
    pfe% tar xvf my.tar
  - Using mtar when creating or extracting a tar file on Lustre (stripe count based on the sum of all component files) *pfe% pwd*
    pfe% /nobackup/username
    *pfe% mtar cvf my.tar **
    pfe% mtar xvf my.tar
- Sharing Files
  - chmod command *pfe% chmod –R g+rx mydir*
  - setfacl & getfacl command *pfe% setfacl –R –m u:jbrown:rx mydir*
    pfe% getfacl mydir

Managing Archive



When necessary, files (large and old, defined by sysadm) are automatically migrated from Disk to Tape through DMF

- Which Lou should I use?

lou1% mylou
Your Mass Storage is lou2
Store files there in your home …

Lou1/Lou2	HOME / Tape
Disk/Tape Quota	None
Inode Quota	soft/hard: 250,000/300,000
Quota Command	*quota -v*

Combining small files into tar file will help reduce issues with inode quota

Managing Archive (cont'd)

- **Transfer Data from Pleiades bridge nodes (faster) or pfe's (slower) to Lou**

bridge1% scp my.tar lou2:
bridge1% bbftp –e 'setnbstream 2; put my.tar' –E 'bbftpd –s –m 2' lou2
bridge1% shiftc my.tar lou2: (use 2 streams if using bbftp underneath)

- **Transfer Data from Pleiades to Lou in a PBS Job**

In your PBS script, you are on a compute node such as r130i0n0, which has no direct connection to lou. Need to ssh to a bridge node (or a pfe) to do the file transfer. Add the following in your PBS script:

ssh –q bridge1 "ssh –q lou2 mkdir –p $SAVDIR"
ssh –q bridge1 "scp –q $RUNDIR/* lou2:$SAVDIR" (or bbftp, or shiftc)

- **Validate Integrity of Data Transferred (using md5sum or shiftc --verify)**

bridge1% md5sum my.tar
afb1d69c6c0c293f05f9f6405d481c7f my.tar
lou2% md5sum myfile
afb1d69c6c0c293f05f9f6405d481c7f my.tar

bridge1% shiftc --verify my.tar lou2:

Managing Archive (cont'd)

	DMF Commands (on Lou only)
dmls	Same as *ls* + list file migration status
dmget	Retrieve migrated files from Tape to Disk (tape robot fetch tape cartridge, retrieve file, put away the tape)
dmput	Cause file to migrate from Disk to Tape
dmfind	Find files

	File Migration Status
REG	File is on Disk Only
DUL	File is on both Disk and Tape
OFL	File is on Tape Only
MIG	File is migrating from Disk to Tape
UNM	File is un-migrating from Tape to Disk

```
% dmls –l
    -rw-r----- 1 user group    14713 Mar 1 17:02 (REG) file1
  -rw-r----- 1 user group    75645 Mar 1 17:02 (REG) file2
  -rw-r----- 1 user group  8014713 Mar 1 17:02 (OFL) file3
  -rw-r----- 1 user group 71175645 Mar 1 17:02 (OFL) file4
```

```
% dmget file3 file4; echo Done
    [pause, then]
    Done
% dmput –r –w file1 file2

% dmget `dmfind . –state ofl –print`
% dmput `dmfind . –state reg –print`
```

A single dmget with multiple files can be more efficient than multiple dmget of single file if those files are on the same tape

-r releases the disk copy after the tape copy is made
(So the file will be in OFL mode, not DUL mode)
-w wait for the migration to complete before returning to shell prompt

Find all offline files in the current directory and then retrieve them
Find all online files in the current directory and offline them

References in NAS Knowledge Base

http://www.nas.nasa.gov/hecc/support/kb/

Knowledge Base | News | Downloads | Ask a Question

Search: Advanced search

76 »

- Knowledge Base
  - New User Orientation
  - FAQ
  - Updates on Issues
  - Policies
  - Troubleshooting
  - Tips & Tricks
  - The HEC Environment
  - Computing at NAS
    - Computing Hardware
      - Pleiades
        - Pleiades: Introduct
        - Pleiades Hardware
        - Pleiades Configurat
        - Harpertown Proces

HECC Home / Support Home / KB Home / Computing at NAS / Computing Hardware / Pleiades / Pleiades Hardware Overview

Pleiades Hardware Overview

Pleiades, the seventh most powerful supercomputer in the world, represents NASA's state-of-the-art technology for meeting the agency's supercomputing requirements, enabling NASA scientists and engineers to conduct modeling and simulation for NASA missions. This distributed-memory SGI ICE cluster is connected with InfiniBand in a dual-plan hypercube technology.

This system contains the following types of Intel Xeon processors: X5670 (Westmere), X5570 (Nehalem), E5472 (Harpertown) and NVIDIA M2090 GPU. Pleiades is named after the astronomical open star cluster of the same name.

System Architecture

- Manufacturer - SGI
- 185 racks (11,776 nodes)

Article ID: 76

Last updated: 24 Oct, 2011

Print

Export to PDF

Views: 1431

Posted: 21 Jul, 2010
by Dunbar J.

Updated: 24 Oct, 2011
by Beyk D.

References in NAS Knowledge Base

Type in the following article ID in the search field to get to the page
or
use the URL format http://www.nas.nasa.gov/hecc/support/kb/entry/76

140: File Transfer Overview
142: Remote File Transfer Commands
300: File Transfers Using Shift
143: Outbound File Transfer Examples
144: Inbound File Transfer through SFEs Examples

64: Description of Various Filesystems
224: Lustre Basics
226: Lustre Best Practices
156: Quota Policy on Disk Space and Files
291: Using mtar to Create or Extract Tar Files on Lustre
279: Using Access Control Lists for File Sharing

148: Archiving Data Overview
244: Mass Storage Systems: Lou1 and Lou2
125: Streamlining File Transfers from Pleiades Compute Nodes to Lou
243: Checking File Integrity
250: Data Migration Facility DMF Commands

Questions?