



IO on Pleiades: Tips and Techniques

**Webinar Presentation, Jan. 11, 2012
NASA Advanced Supercomputing Division**

Outline



Getting files in and out of NAS

Overview of Lustre hardware

Tweaking lustre for performance

Working with files on Lou's tape system

Parallel IO modules

When to use parallel IO

Questions

Upcoming attractions



Getting files in and out of NAS

Preliminary: set up SSH pass through to make moving files in and out of NAS much easier

http://www.nas.nasa.gov/hecc/support/kb/Setting-Up-SSH-Passthrough_232.html

Useful file transfer approaches: **sup**, **bbftp**, **bbscp**

http://www.nas.nasa.gov/hecc/assets/pdf/training/hecc_webinar_newuser_11-9-11.pdf

```
bridge3: bbftp -V -p 14 -s -u sheistan -e 'put file' lou.nas.nasa.gov
15825105635 bytes send in 30.3 secs (5.1e+05 Kbytes/sec or 3980 Mbits/s)
bridge3:
```

```
pfe4: bbftp -V -p 14 -s -u sheistan -e 'put file' lou.nas.nasa.gov
15825105635 bytes send in 140 secs (1.1e+05 Kbytes/sec or 861 Mbits/s)
pfe4:
```

```
pfe4: bbftp -V -p 14 -s -u sheistan -e 'put files*' lou.nas.nasa.gov
Doesn't do what you expect...
```

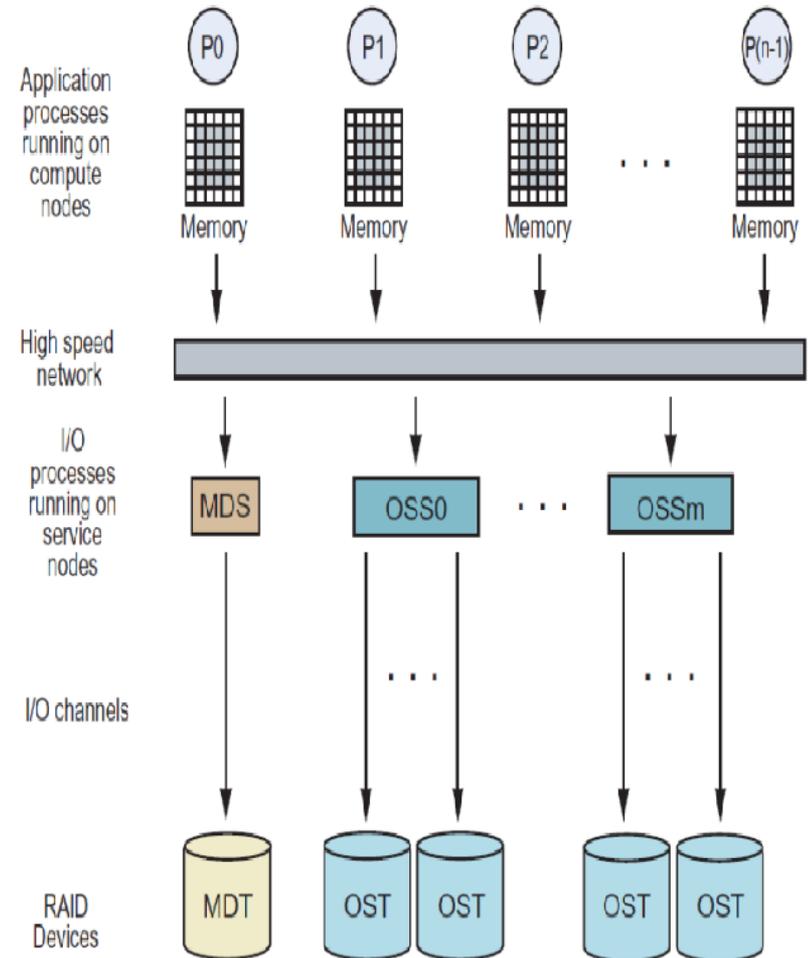
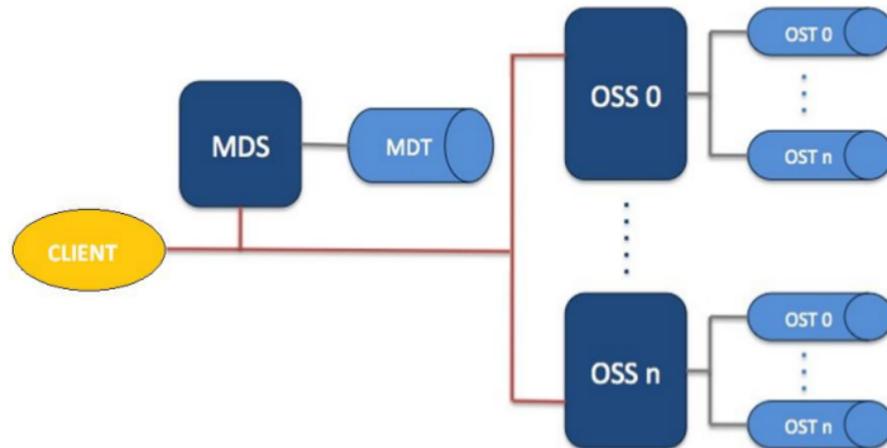
```
bridge3: bbscp -V -p 14 files* lou.nas.nasa.gov:
19236157440 bytes send in 108 secs (1.73e+05 Kbytes/sec or 1350 Mbits/s)
bridge3:
```

Tip: Do an MD5sum on the files before and after transferring to verify copy

```
bridge3:heistand% md5sum tullis_data.tar
e5e484d718f3e3ad8366f2ab4beedf46  tullis_data.tar
bridge3:
```

The Joys of Lustre

- **Metadata Server (MDS)** makes metadata stored in the **MDT(Metadata Target)** available to Lustre clients.
- Each MDS manages the names and directories in the Lustre filesystem and provides network request handling for the MDT.
- **Object Storage Server(OSS)** provides file service, and network request handling for one or more local **OSTs**.
- **Object Storage Target (OST)** stores file data (chunks of files).





The Joys of Lustre (cont)

To check your quota on the /nobackup filesystem

```
pfe: lfs quota -u nas_username /nobackup/nas_username
```

Setting the stripe on an empty directory

```
pfe: lfs setstripe -c 12 -s 4M -i -1 directory
```

```
pfe:
```

This will spread the data created in the directory across 12 OSTs, in chunks of 4M on each, starting with a random OST.

(Best not to pick an OST to start on)

To correct the stripe of a whole directory:

```
pfe: mkdir good_dir
```

```
pfe: lfs setstripe -c 16 -s 4M good_dir
```

```
pfe: cp -rp bad_dir/* good_dir/
```

```
pfe: rm -rf bad_dir
```

```
pfe:
```

Guidelines for choosing the values:

Amount of data in a fortran binary write() call

Number of processors writing at once

Overall size of files, e.g. stripe files << 1G across just 1 stripe

Lustre tips:

Limit 'ls -l' operations on directories with large numbers of files

Delay between file inquiries inside codes and scripts.

Open files read only unless you really need to write to the file.



Moving files to/from Lou (archive)

Tip: Use the bridge{1,2,3,4} front end nodes as they have faster connections

Don't compress files before archiving

- They get compressed automatically when written to tape
- Doing compression on Lou will bog the machine down for everyone

Useful DMF commands on lou:

dm`ls` `'dmls -l'` will look a lot like a normal `'ls -l'` command

But will also have file status info:

- `REG` not managed by DMF
- `MIG` migrating
- `DUL` dual-state
- `OFL` offline
- `UNM` unmigrating

dm`get` `dmget [-a] list_of_files`

- Get a number of files from tape. List multiple files
- For better efficiency.

dm`put`: `dmput -r list_of_files`

- Puts online files back onto tape.

Note: currently no quotas on lou. This may change so try to plan ahead on your usage.

Combining small files into tarballs will help reduce quota issues related to the number of files.

```
bridge3: tar -cf - dir_or_files_to_tar | ssh -q lou 'cd someplace ; cat > filename.tar'
```



Parallel IO Libraries

Libraries installed as modules on Pleiades:

- HDF5: `hdf5/1.8.7/gcc/mpt`
- NetCDF: `netcdf/4.1.3/gcc/mpt`
- MPI-IO:
 - `mpi-sgi/mpt.2.04.10789`
 - `mpi-mvapich2/1.6/intel`
 - `mpi-intel/4.0.2.003`

This module requires the environment variables:

```
I_MPI_EXTRA_FILESYSTEM=on  
I_MPI_EXTRA_FILESYSTEM_LIST=lustre
```

Useful tutorials on parallel IO in it's many forms:

<http://www.osc.edu/supercomputing/training/pario/>

<http://www.youtube.com/user/HPCUniversity/search?query=parallel+IO>

Things to look for when the data is wrong:

- Wrong file offset in an MPI-IO call
- Calling a routine that really needs to be called by every rank, but either it's not called by everyone, or called with a communicator that isn't defined everywhere. e.g. not `MPI_COMM_WORLD`



When to use parallel IO

You may want to do parallel IO if:

- Your main output files are or need to be in HDF or NetCDF format
- The post processing step needs a single or just a few really really large files
- You are writing check point/restart files and may change process count on a restart
- You plan on scaling to large ($> \sim 1000$) processes and will be output large amounts of data
- The data you are writing is too large to easily route it to a single process for IO



Slides Prepared by Steve Heistand

pdf and recording of this webinar will be available shortly at:
<http://www.nas.nasa.gov/hecc/support/training.html>

Next webinar

“How Can I Speed up my Interactive Connection to NAS?”
tentatively scheduled on Feb. 8, 2012

Suggestions for future webinar topics are welcomed