

# Lustre Basics

## Category: Lustre on Pleiades

A Lustre filesystem is a high-performance, shared filesystem (managed with the Lustre software) for Linux clusters. It is highly scalable and can support many thousands of client nodes, petabytes of storage and hundreds of gigabytes per second of I/O throughput. On Pleiades, the Lustre filesystems are named "/nobackupp."

### Main Lustre components:

- Metadata Server (MDS)

1 or 2 per filesystem; service nodes that manage all metadata operations such as assigning and tracking the names and storage locations of directories and files on the OSTs.

- Metadata Target (MDT)

1 per filesystem; a storage device where the metadata (name, ownership, permissions and file type) are stored.

- Object Storage Server (OSS)

1 or multiple per filesystem; service nodes that run the Lustre software stack, provide the actual I/O service and network request handling for the OSTs, and coordinate file locking with the MDS. Each OSS can serve up to ~15 OSTs. The aggregate bandwidth of a Lustre filesystem can approach the sum of bandwidths provided by the OSSes.

- Object Storage Target (OST)

multiple per filesystem; storage devices where the data in user files are stored. Under Linux 2.6 (current OS on Pleiades), each OST can be up to 8TB in size. Under SLES 11, each OST can be up to 16 GB in size. The capacity of a Lustre filesystem is the sum of the sizes of all OSTs.

- Lustre Clients

commonly in the thousands per filesystem; compute nodes that mount the Lustre filesystem, and access/use data in the filesystem.

### File Striping

A user file can be divided into multiple chunks and stored across a subset of the OSTs. The chunks are distributed among the OSTs in a round-robin fashion to ensure load balancing.

Benefits of striping:

- allows one to have a file size larger than the size of an OST
- allows one or more clients to read/write different parts of the same file at the same time and provide higher I/O bandwidth to the file since the bandwidth is aggregated over the multiple OSTs

Drawbacks of striping:

- higher risk of file damage due to hardware malfunction
- increased overhead due to network operations and server contention

There are default stripe configurations for each Lustre filesystem. However, users can set the following stripe parameters for their own directories or files to get optimum I/O performance:

#### 1. stripe\_size

the size of the chunk in bytes; specify with k, m, or g to use units of KB, MB, or GB, respectively; the size must be an even multiple of 65,536 bytes; default is 4MB for all Pleiades Lustre filesystems; one can specify 0 to use the default size.

#### 2. stripe\_count

the number of OSTs to stripe across; default is 1 for most of Pleiades Lustre filesystems (/nobackupp[10-60]); one can specify 0 to use the default count; one can specify -1 to use all OSTs in the filesystem.

#### 3. stripe\_offset

The index of the OST where the first stripe is to be placed; default is -1 which results in random selection; using a non-default value is NOT recommended.

Use the **lfs setstripe** command for setting the stripe parameters.

```
pfe20% lfs setstripe -s stripe_size -c stripe_count -o
stripe_offset dir|filename
```

For example, to create a directory called dir1 with a stripe\_size of 4MB and a stripe\_count of 8, do

```
pfe20% mkdir dir1
pfe20% lfs setstripe -s 4m -c 8 dir1
```

Also keep in mind that:

- When a file or directory is created, it will inherit the parent directory's stripe settings.

- The stripe settings of an *existing file* can not be changed. If you want to change the settings of a file, you can create a new file with the desired settings and copy the existing file to the newly created file.

## Useful Commands for Lustre

- To list all the OSTs for the filesystem

```
pfe20% lfs osts
```

- To list space usage per OST and MDT in human readable format for all Lustre filesystems or for a specific one, for example, /nobackupp1:

```
pfe20% lfs df -h  
pfe20% lfs df -h /nobackupp1
```

- To list inode usage for all filesystems or a specific one, for example, /nobackupp1:

```
pfe20% df -i  
pfe20% df -i /nobackupp1
```

- To create a new (empty) file or set directory default with specified stripe parameters

```
pfe20% lfs setstripe -s stripe_size -c stripe_count -o  
stripe_offset dir|filename
```

- To list the striping information for a given file or directory

```
pfe20% lfs getstripe dir|filename
```

- To display disk usage and limits on your /nobackup directory (for example, /nobackupp1):

```
pfe20% lfs quota -u username /nobackupp1
```

or

```
pfe20% lfs quota -u username /nobackup/username
```

To display usage on each OST, add the -v option:

```
pfe20% lfs quota -v -u username /nobackup/username
```

See the **lfs man page** for more options and information.

Article ID: 224

Last updated: 17 Dec, 2012

Computing at NAS -> Best Practices -> Lustre on Pleiades -> Lustre Basics

<http://www.nas.nasa.gov/hecc/support/kb/entry/224/?ajax=1>