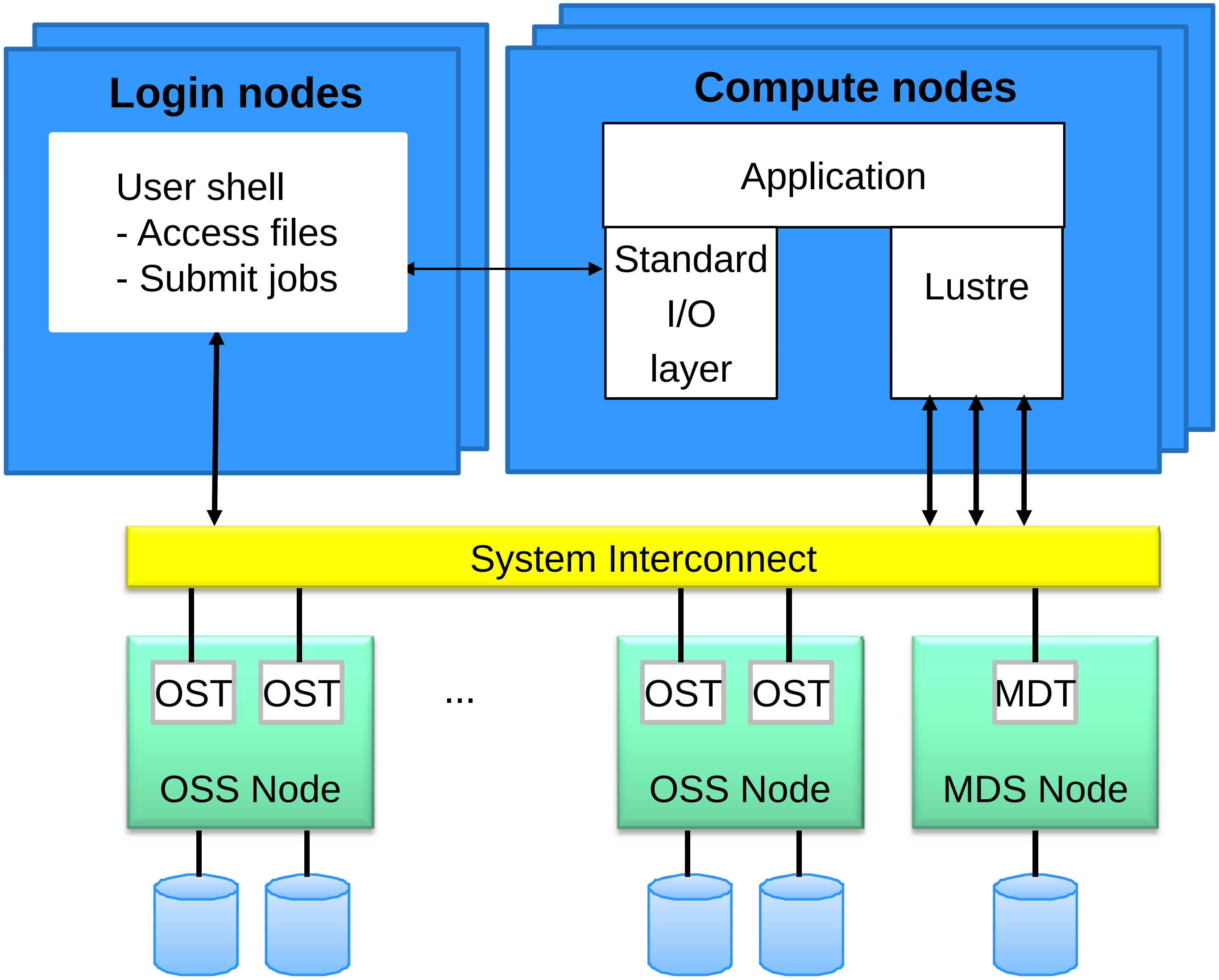
Lustre File System

Customer Documentation and Training

Cray XT/XE Systems I/O Support

- The compute nodes usually hand off I/O to the SIO or XIO (service I/O) nodes
- The `aprun` application launcher handles `stdin`, `stdout`, and `stderr`
  - Refer to the *Cray XT Programming Environment User's Guide* (S-2396), "I/O Support" in the "Catamount Programming Considerations" section

Cray I/O Architecture



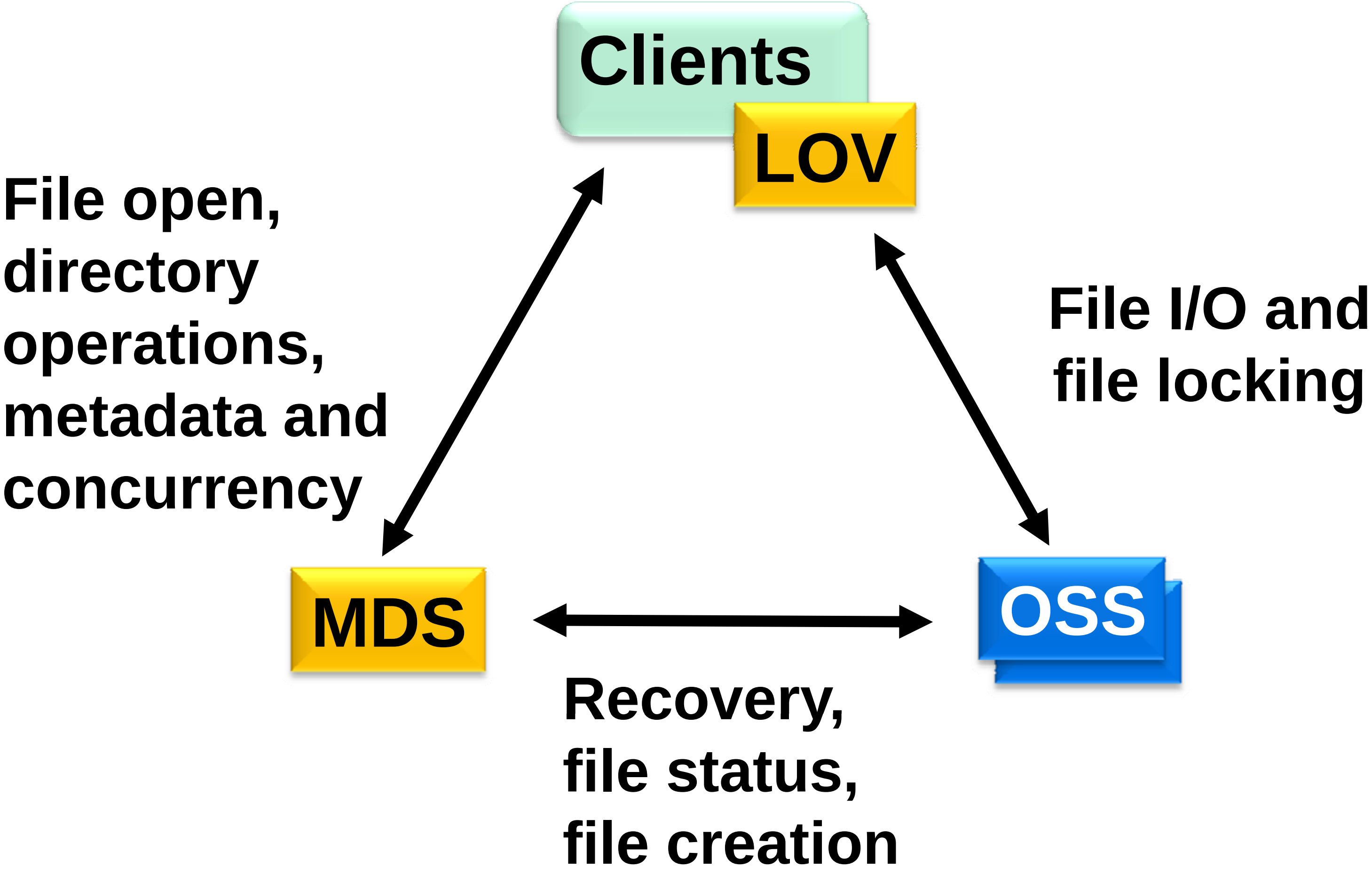
Lustre Fundamentals

- One MDS and one or more OSTs comprise a single Lustre file system
- If you want to create another Lustre file system, you must configure it on separate disk devices

Lustre Terminology

- MDS – metadata server
  - The server node
- MDT – metadata target
  - The softbackend storage
    - The backend storage is an ldiskfs (ext3) file system
    - On Cray systems, this is a RAID volume
    - LUNs are formatted according to the underlying vendor device
- OSS – object storage server
  - The server node, supports multiple OSTs
- OST – object storage target
    - The backend storage is an ldiskfs (ext3) file system
    - The multi-block allocator (MBA) is used for performance

Node Interaction


Clients

LOV

MDS

OSS

File open, directory operations, metadata and concurrency

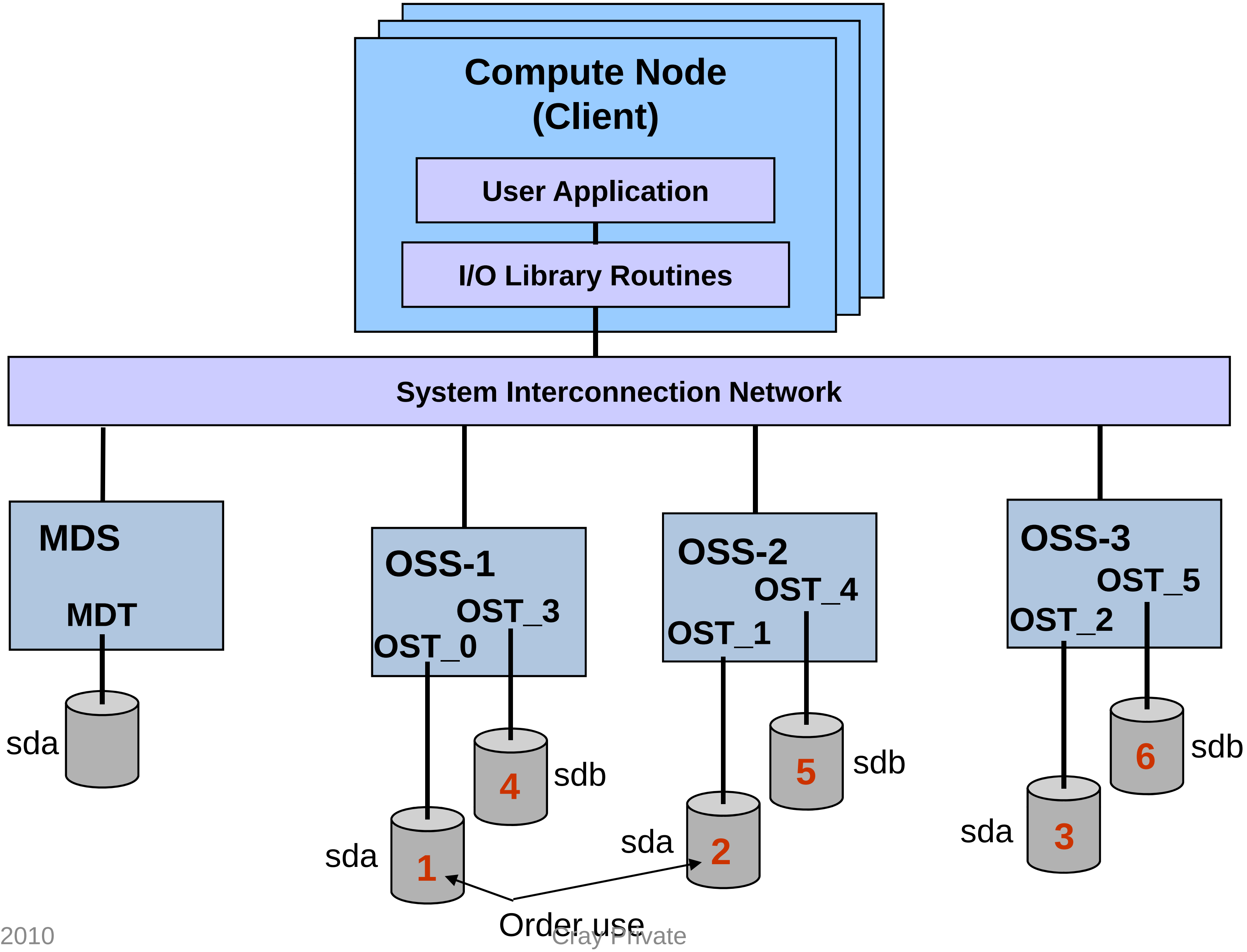
File I/O and file locking

Recovery, file status, file creation

Lustre Terminology

- **Stripe width - the number of OSTs to write the file across**
  - **Cray recommends a default stripe width of one to four**
    - **Recommends against striping across all OSTs**
  - **Can be set either at the file or directory level**
    - **When you stripe at the directory level, new files inherit the stripe width of the parent directory**

Lustre Diagram



Using the df Command

- **Use the standard `df` command to locate the mount point for a Lustre file system**

```
% df -t lustre
Filesystem        1K-blocks       Used Available Use%   Mounted on
8@ptl:/nid00008_mds/client
                 5644895560 4509828872 848322232  85%   /lus/nid00008
36@ptl:/nid00036_mds/client
                 1128979112  447543988 624086236  42%   /lus/nid00036
%
```

Lustre Commands

- `lfs` is a Lustre utility that can:
  - Provide file system configuration information

    ```
    lfs df
    ```

  - Create a file or directory with a specific striping pattern

    ```
    lfs setstripe
    ```

  - Display file striping patterns

    ```
    lfs getstripe [directory | file name]
    ```

  - Find file locations

    ```
    lfs find [directory | file name]
    ```

    - For example, to find directories or files on a particular OST

    ```
    lfs find –r –obd ost5_UUID /work/rns
    ```

  - Display quota information

    ```
    lfs quota -u|g <name> file system
    ```

Client View of File System Space

- **To view the individual OSTs**

```
% lfs df
UUID                  1K-blocks       Used Available  Use%  Mounted on
nid00008_mds_UUID    1003524776   57906856 945617920     5  /lus/nid00008[MDT:0]
ost0_UUID            1128979112 1094326220  34652892    96  /lus/nid00008[OST:0]
ost1_UUID            1128979112 1076393372  52585740    95  /lus/nid00008[OST:1]
ost2_UUID            1128979112  894139784 234839328    79  /lus/nid00008[OST:2]
ost3_UUID            1128979112 1006132924 122846188    89  /lus/nid00008[OST:3]
ost4_UUID            1128979112  725581028 403398084    64  /lus/nid00008[OST:4]
filesystem summary: 5644895560 4796573328 848322232    84  /lus/nid00008

UUID                  1K-blocks       Used Available  Use%  Mounted on
nid00036_mds_UUID    1003524776   57871880 945652896     5  /lus/nid00036[MDT:0]
ost0_UUID            1128979112  504892876 624086236    44  /lus/nid00036[OST:0]
filesystem summary: 1128979112  504892876 624086236    44  /lus/nid00036
%
```

File Striping and Inheritance

- Lustre distributes files across all OSTs
- The default stripe width is set in the configuration file
- Users can create files and directories with various striping characteristics
  - New files inherit the striping of the parent directory
  - Striping across more OSTs generally leads to higher peak performance on large files, but may not be best for small files
  - CANNOT change the stripe pattern on an existing file
  - CAN change the stripe pattern on a directory
- Improper striping, such as in the following list, may result in inefficient use of your Lustre file system:
  - Writing a very large file to a single OST
  - Creating a directory where files do not circle through the OSTs
  - Striping a small file across many OSTs

`lfs` Command

- **Use the `lfs` command to manage striping characteristics**
  - **To define striping for a file or directory:**

```
lfs setstripe [--size s] [--offset o] [--count c]
  [--pool p] <dir|filename>
```

`--size|-s`	**stripe-size, 0 means use the default**
`--offset|-o`	**starting ost, -1 means use the default (round robin)**
`--count|-c`	**stripe count, 0 means use the default**
`--pool|-p`	**name of OST pool**

Defaults are defined in the Lustre configuration file

  - **To view striping for a file or directory:**

```
lfs getstripe <file-name|dir-name>
```

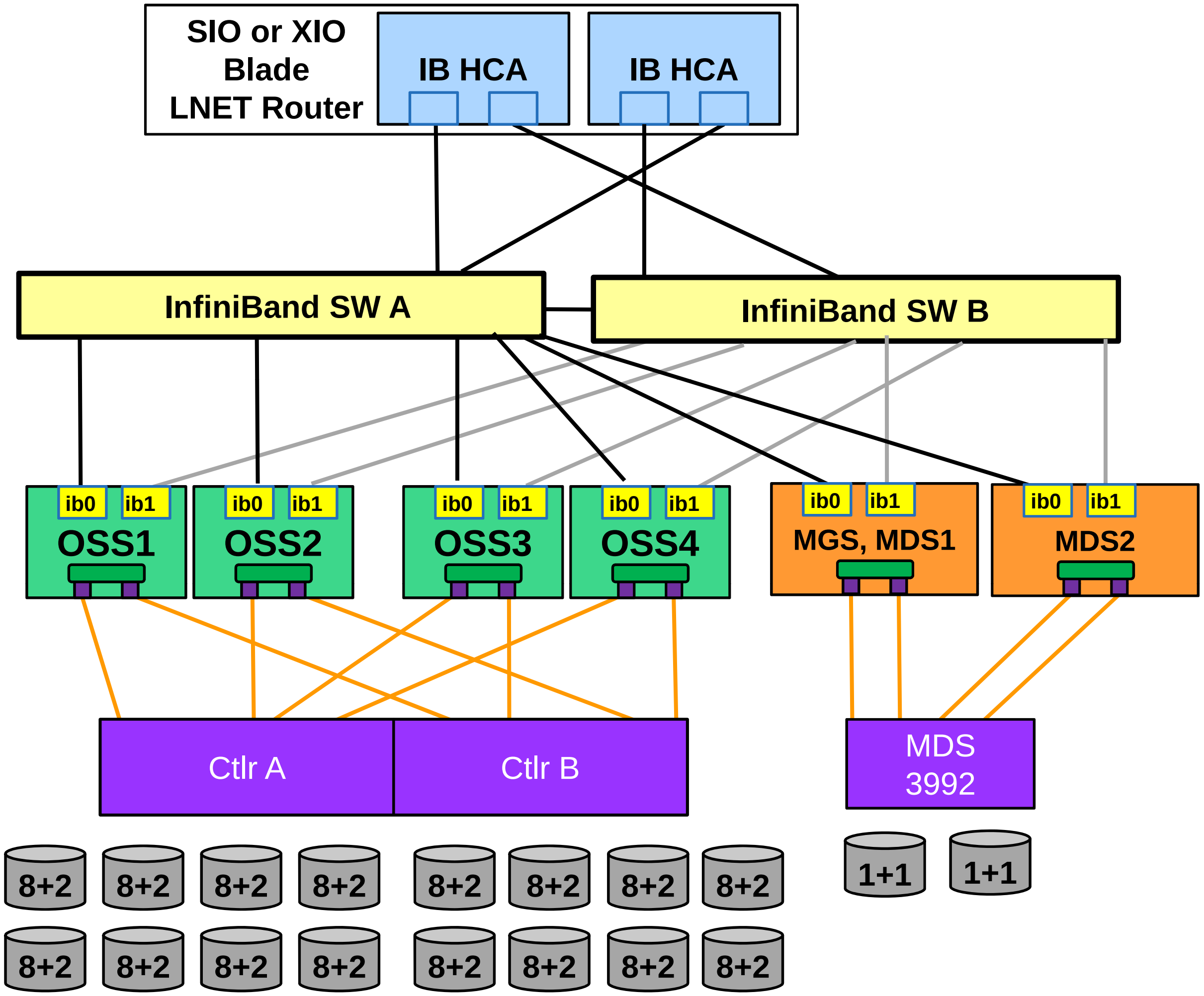
lfs Command Example

```
/rns> mkdir rick
nid00008/rns> lfs getstripe rick
OBDS:
0: ost0_UUID ACTIVE
1: ost1_UUID ACTIVE
2: ost2_UUID ACTIVE
3: ost3_UUID ACTIVE
4: ost4_UUID ACTIVE
rick
(Default) stripe_count: 2 stripe_size: 1048576 stripe_offset: 0
nid00008/rns>
nid00004:/work # cd rick
nid00004:/work/rick # touch file_one
 rns/rick> lfs getstripe file_one
OBDS
0: ost0_UUID ACTIVE
1: ost1_UUID ACTIVE
2: ost2_UUID ACTIVE
3: ost3_UUID ACTIVE
4: ost4_UUID ACTIVE
file_one
        obdidx           objid          objid            group
             3        22189877      0x1529735                0
             4        23084122      0x1603c5a                0
```

lfs Command Example

```
nid00008/rns> lfs setstripe rick –s 3
nid00008/rns> cd rick
rns/rick> touch file_two
rns/rick> touch file_three
rns/rick> lfs getstripe *
OBDS:
0: ost0_UUID ACTIVE
  [clip …]
4: ost4_UUID ACTIVE
file_one
        obdidx           objid          objid            group
             3        22189877      0x1529735                0
             4        23084122      0x1603c5a                0
file_three
        obdidx           objid          objid            group
             3        22189897      0x1529749                0
             4        23084142      0x1603c6e                0
             0        21685921      0x14ae6a1                0
file_two
        obdidx           objid          objid            group
             3        22189895      0x1529747                0
             4        23084140      0x1603c6c                0
             0        21685919      0x14ae69f                0
```

Basic Elements of an esFS Configuration


SIO or XIO
Blade
LNET Router
IB HCA
IB HCA
InfiniBand SW A
InfiniBand SW B
ib0
ib1
OSS1
ib0
ib1
OSS2
ib0
ib1
OSS3
ib0
ib1
OSS4
ib0
ib1
MGS, MDS1
ib0
ib1
MDS2
Ctlr A
Ctlr B
MDS
3992
8+2
8+2
8+2
8+2
8+2
8+2
8+2
8+2
1+1
1+1
8+2
8+2
8+2
8+2
8+2
8+2
8+2
8+2