Choose Lustre

Stephen Simms

Manager, High Performance File Systems

Indiana University

https://www.flickr.com/photos/kubina/521327318

Lustre is scalable – 55 PB at LLNL

http://www.flickr.com/photos/shadowstorm/404158384/

https://www.flickr.com/photos/davesag/18735941

Lustre is fast – 1 TB/s at ORNL

http://www.flickr.com/photos/dvd5/163647219/

https://www.flickr.com/photos/johngarghan/5688711425

Lustre can support thousands of clients

https://www.flickr.com/photos/gogrimm/3525598147

Lustre is Open

Lustre is Open source software under GPLv2

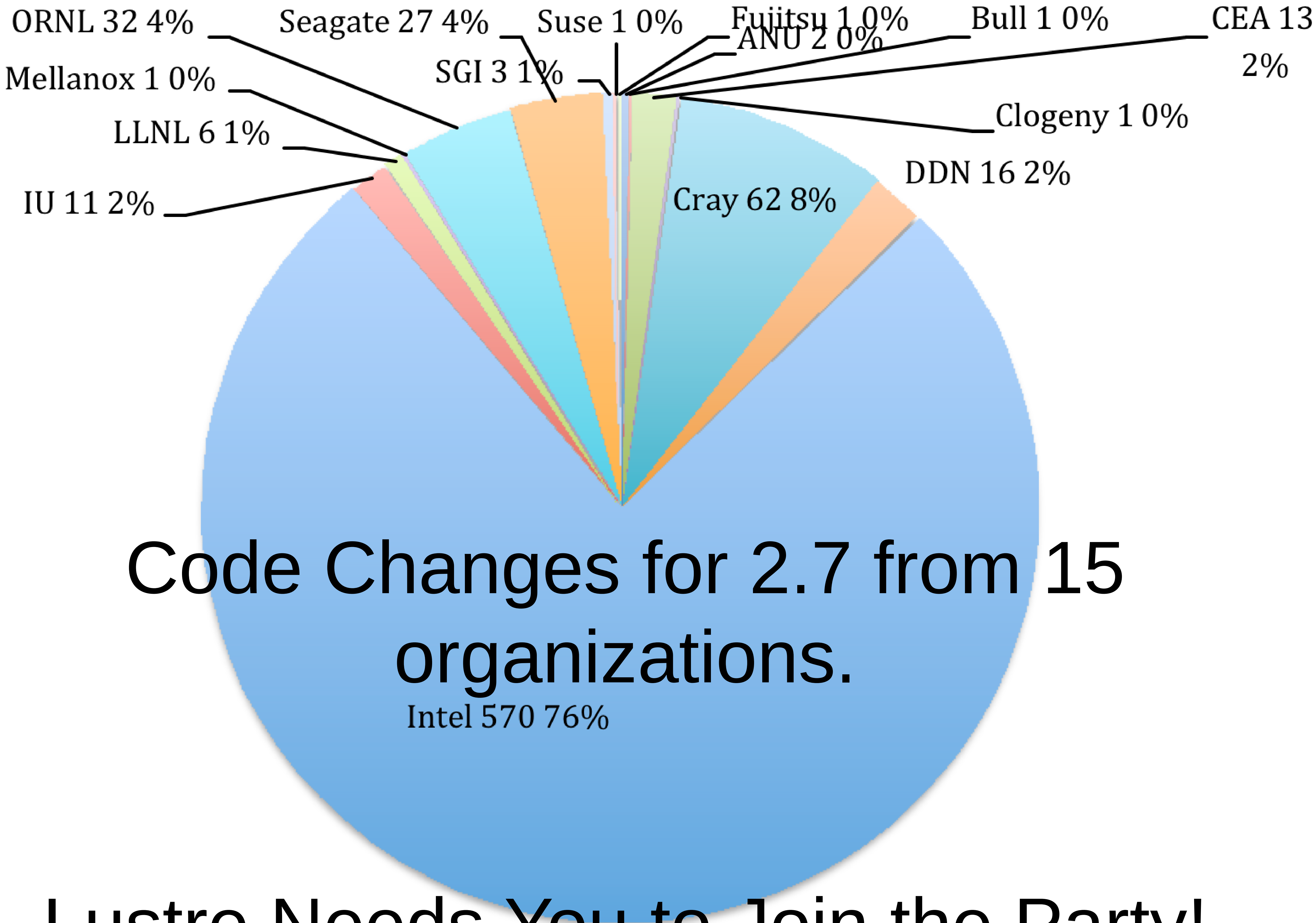
That means it’s
“Free Like Beer”
right?

Actually, more like a free puppy…

It Takes Lots of Work to Maintain Lustre

- Bug Fixes
- Rigorous Testing
- Feature Development
- Maintaining Documentation
- Tree Hosting
- Code Reviews by Peers

Many Hands Make Lighter Work



Code Changes for 2.7 from 15 organizations.

Lustre Needs You to Join the Party!

https://www.flickr.com/photos/alvi2047/3688993279/

Lustre is moving forward

- Hiccup when Lustre moved from Oracle
  - Lustre 2.0 – Fall 2010
  - Lustre 2.1 – Fall 2011
- Since then Lustre has accelerated
  - 2 major releases a year
  - Spring / Fall
- Users following Lustre releases
  - Majority using 2.5 in production

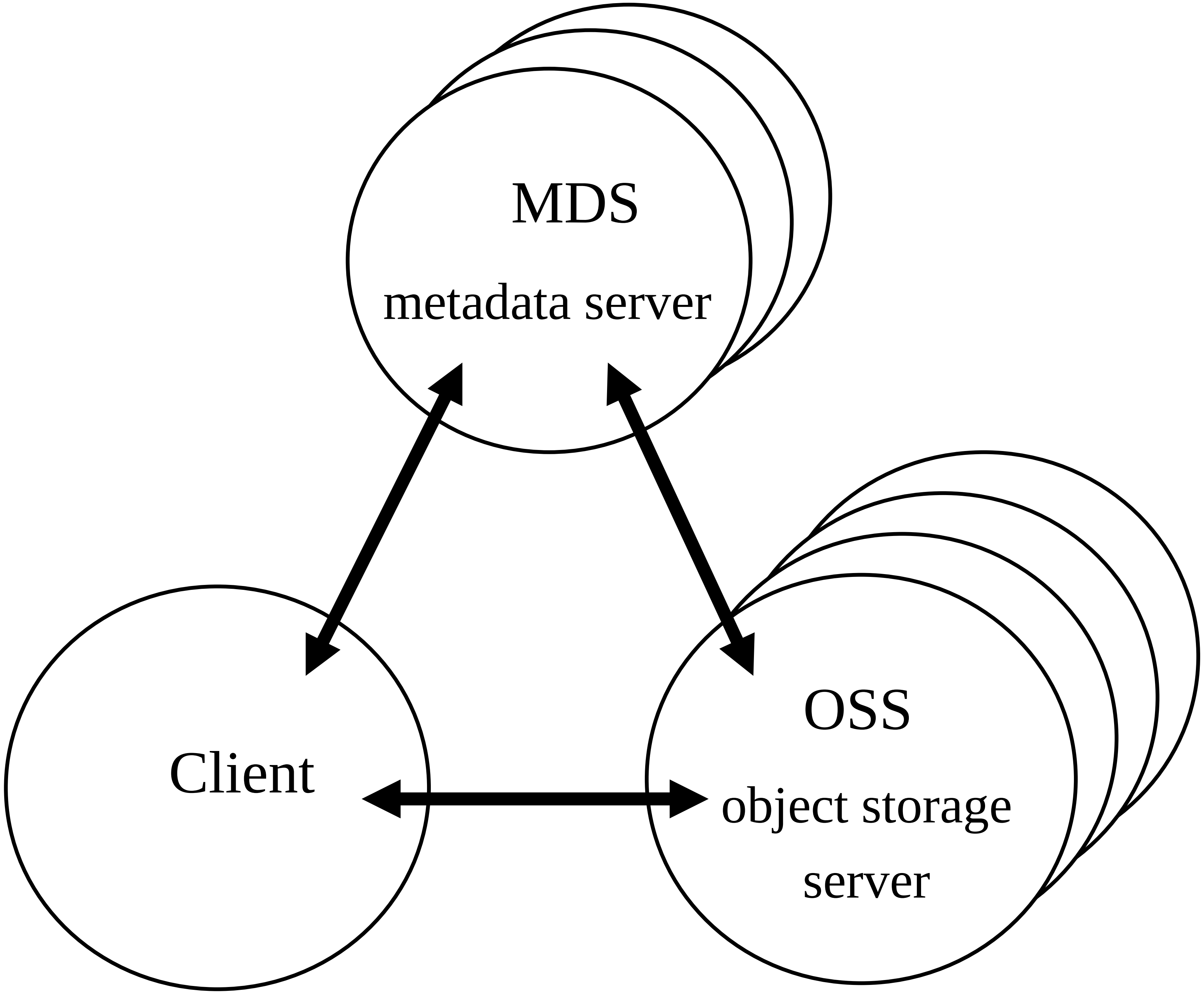
Alright, alright, stop the marketing

Lustre = Linux +Cluster

Lustre is a *parallel* distributed file system

- High performance filesystem used by >60 of the top 100 supercomputers in the world
- POSIX compliant – behaves like other file systems

Lustre: The Players



MDS

metadata server

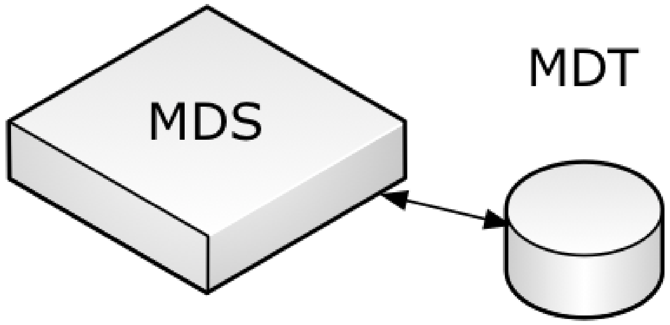
Client

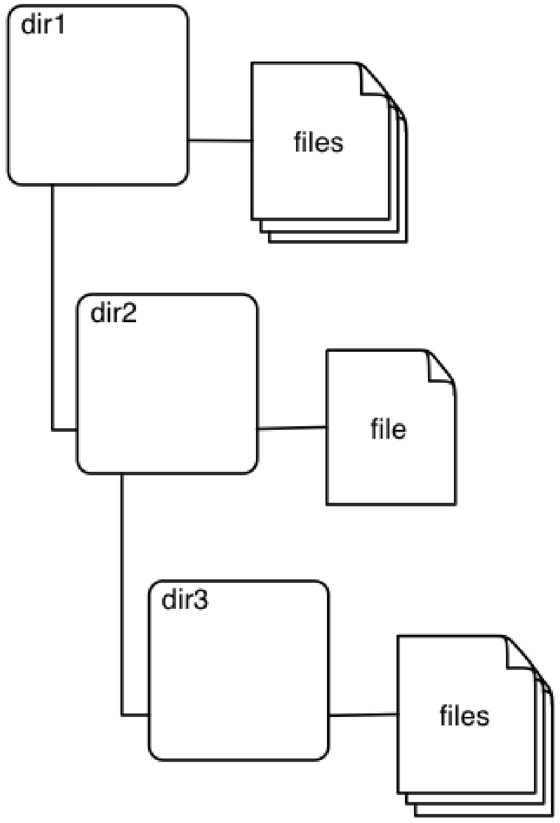
OSS

object storage server

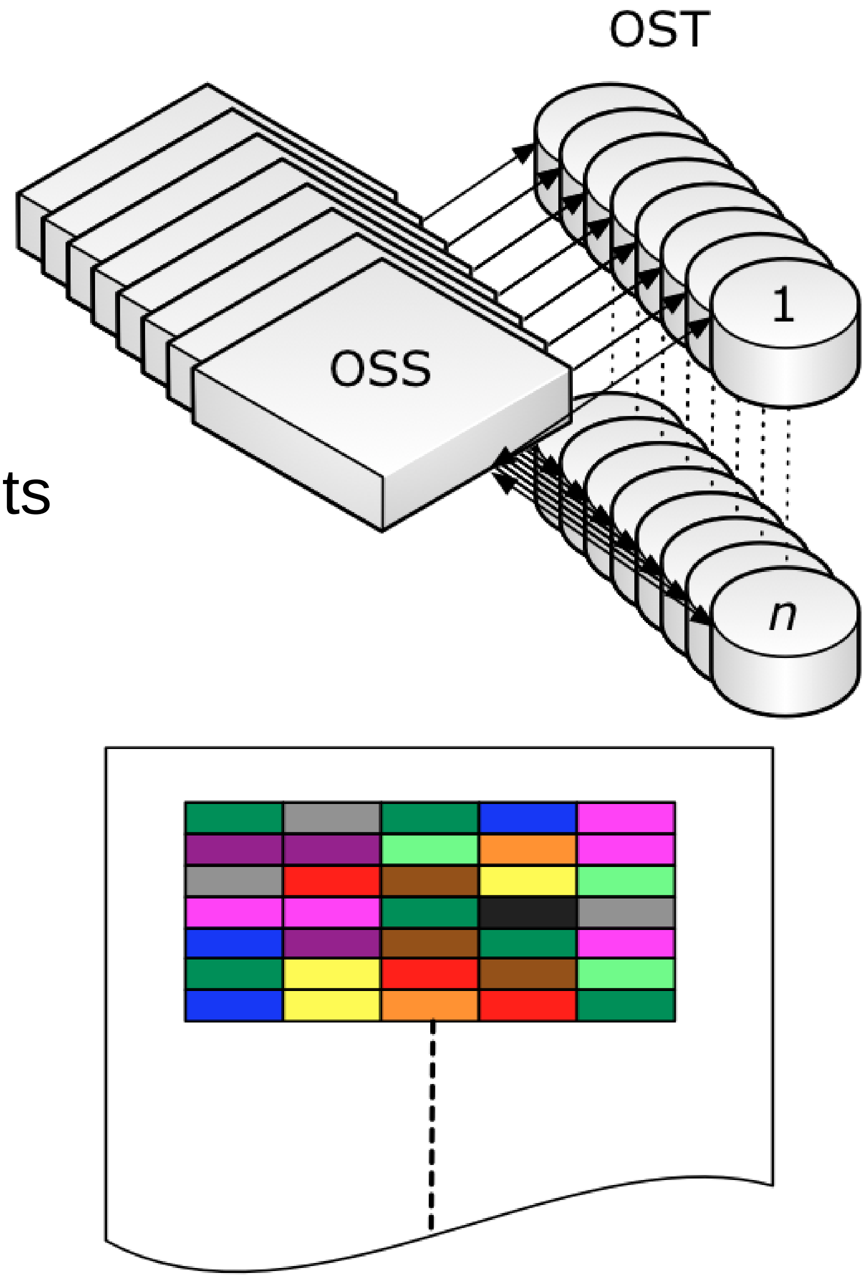
Lustre Architecture - MDS

- **Metadata Server (MDS)**
  - – Node(s) that manage namespace, file creation and layout, and locking.
    - Directory operations
    - File open/close
    - File status
    - File creation
    - Map of file object location
  - – Relatively expensive serial atomic transactions to maintain consistency
- **Metadata Target (MDT)**
  - – Block device that stores metadata



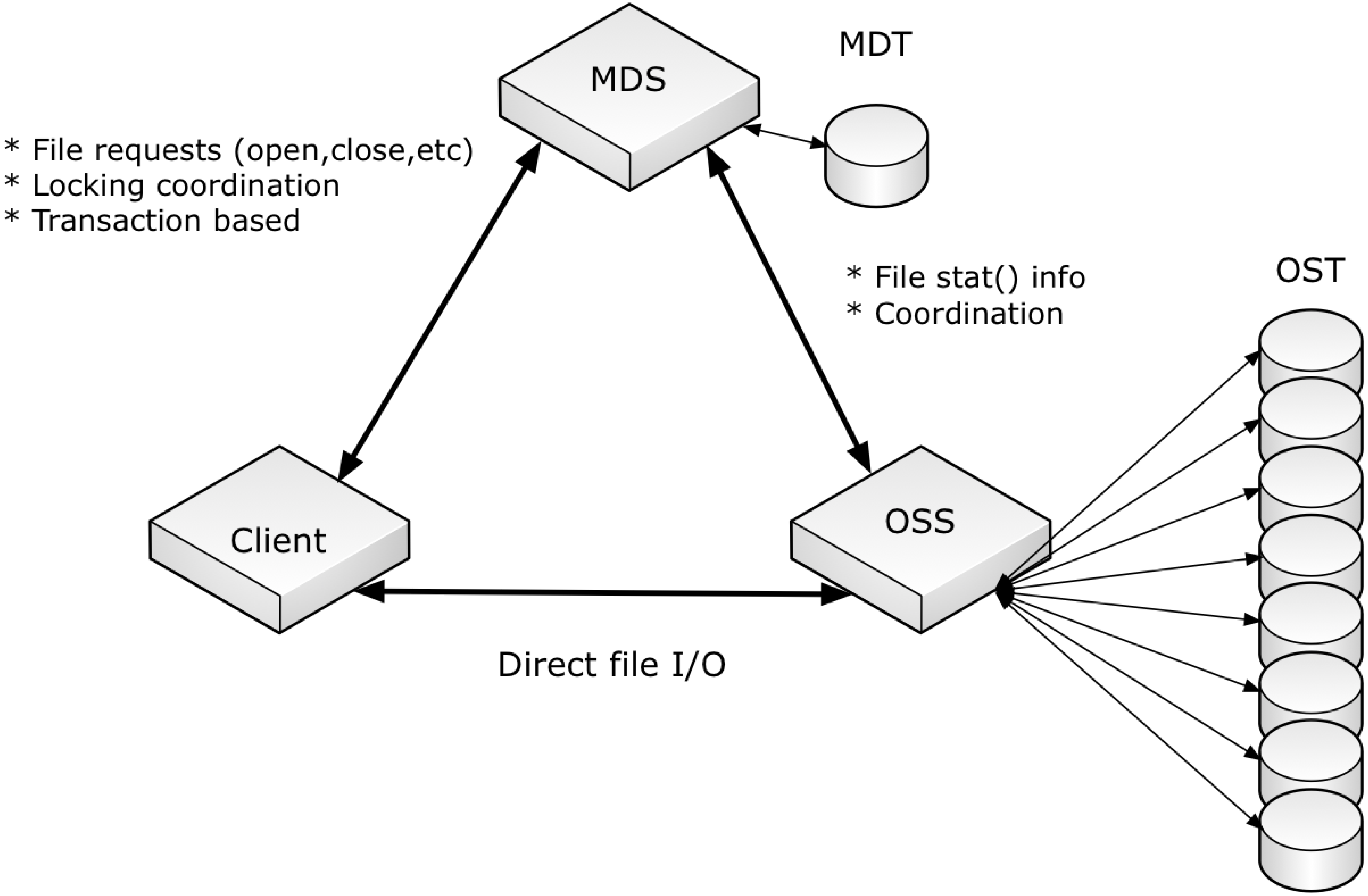


Lustre Architecture - OSS

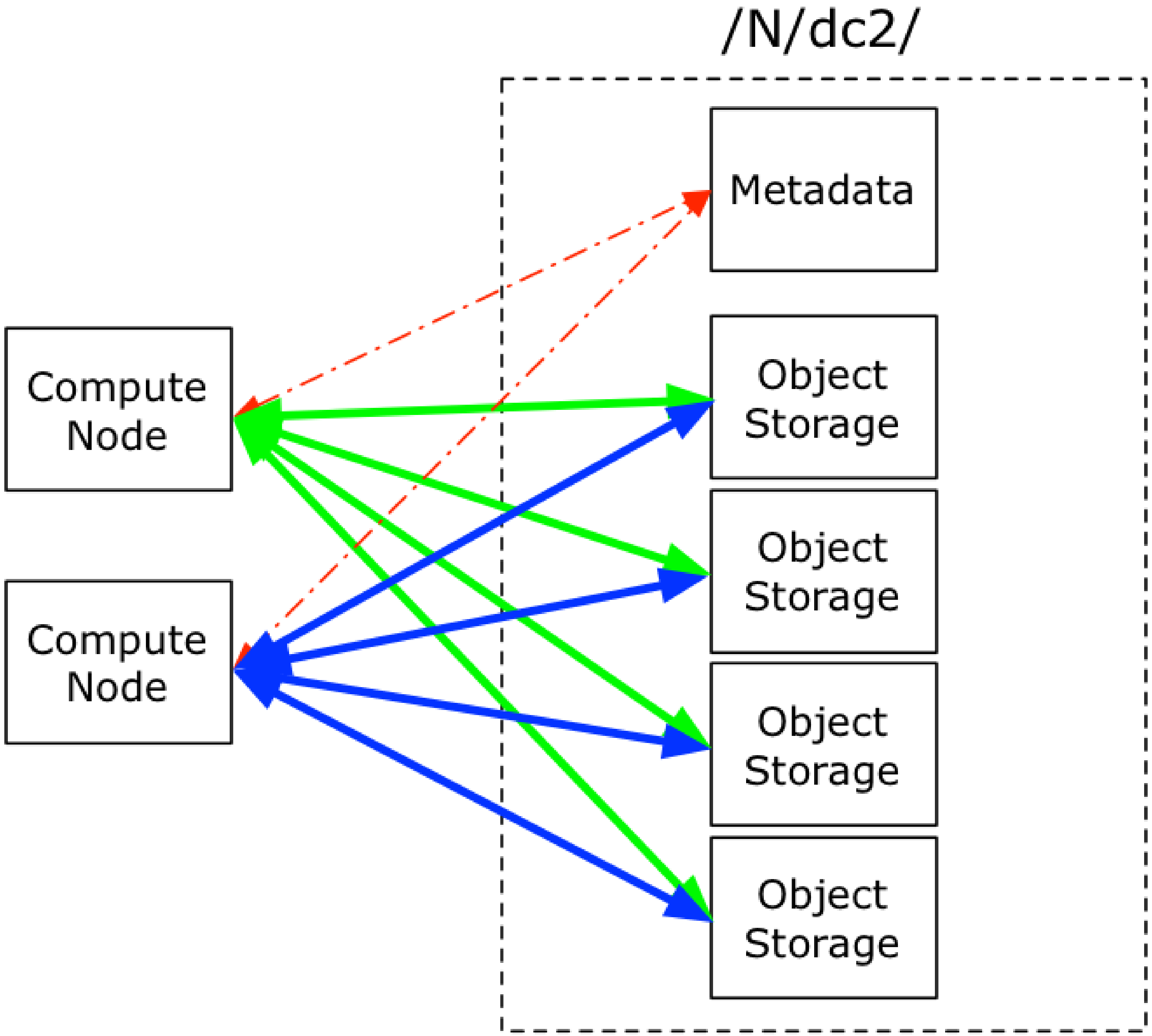


- **Object Storage Server (OSS)**
  - Multiple nodes that manage network requests for file objects on disk.
- **Object Storage Target (OST)**
  - Block device that stores file objects

Simplest Lustre System

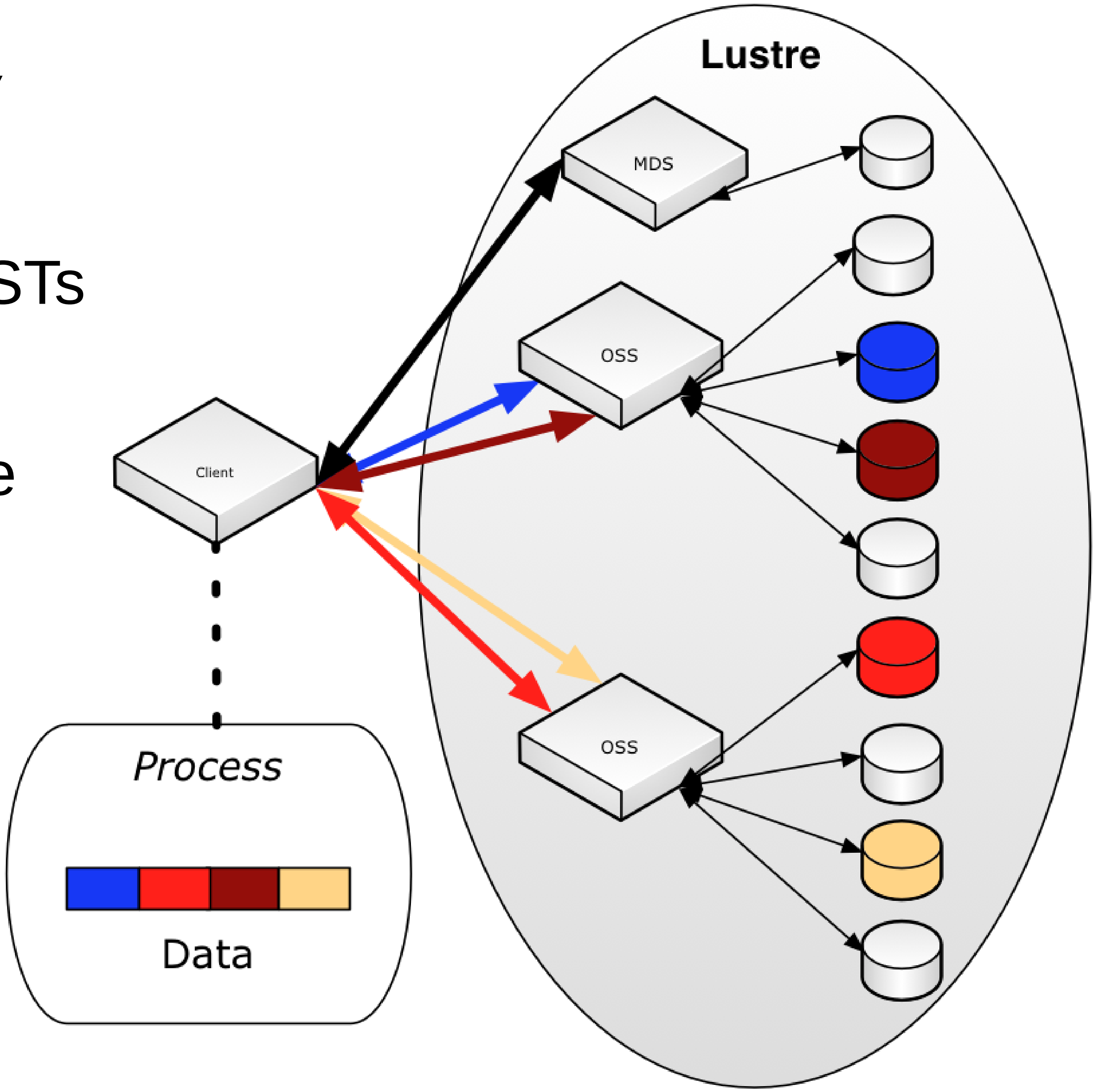


Lustre Parallel I/O

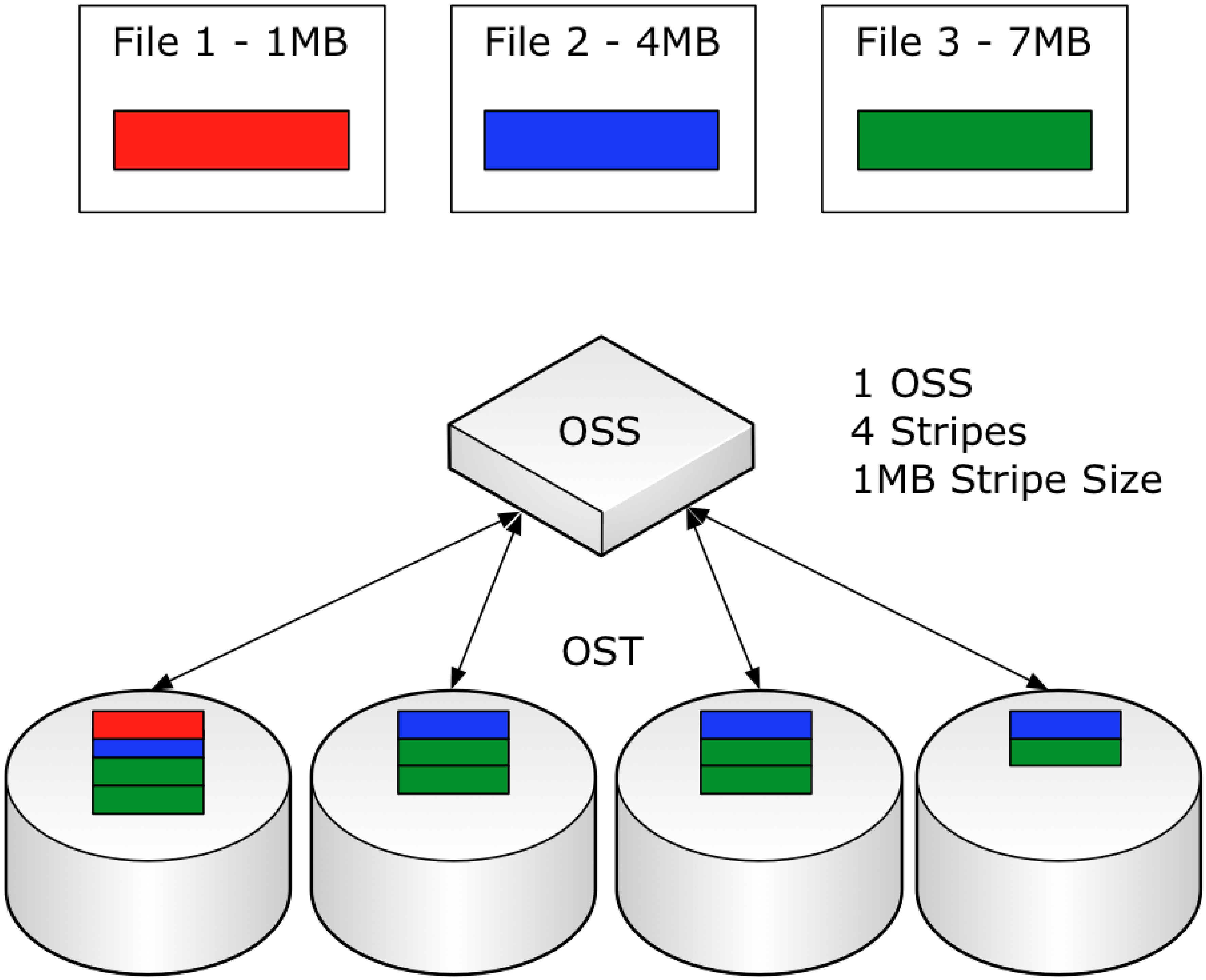


Striping Data

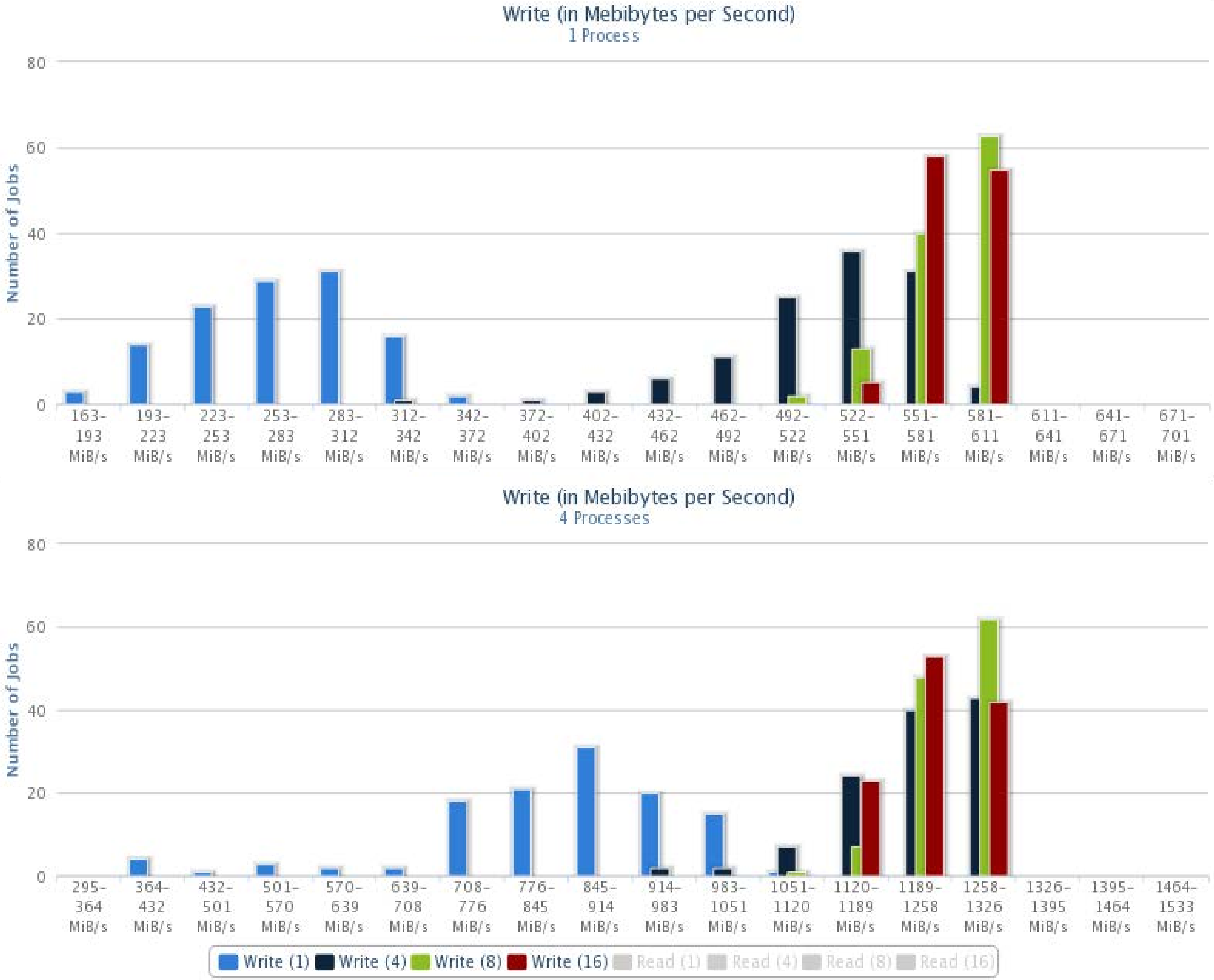
- Lustre allows you to control *how* data is written, if you want
  - *Stripe* data across multiple OSTs
    - can stripe files OR directories
  - Can increase I/O performance with reading and writing
  - With DNE2 metadata can be *Striped* across multiple MDTs
- Striping analogous to RAID 0
- Default striping set by sysadmin



Striping Example

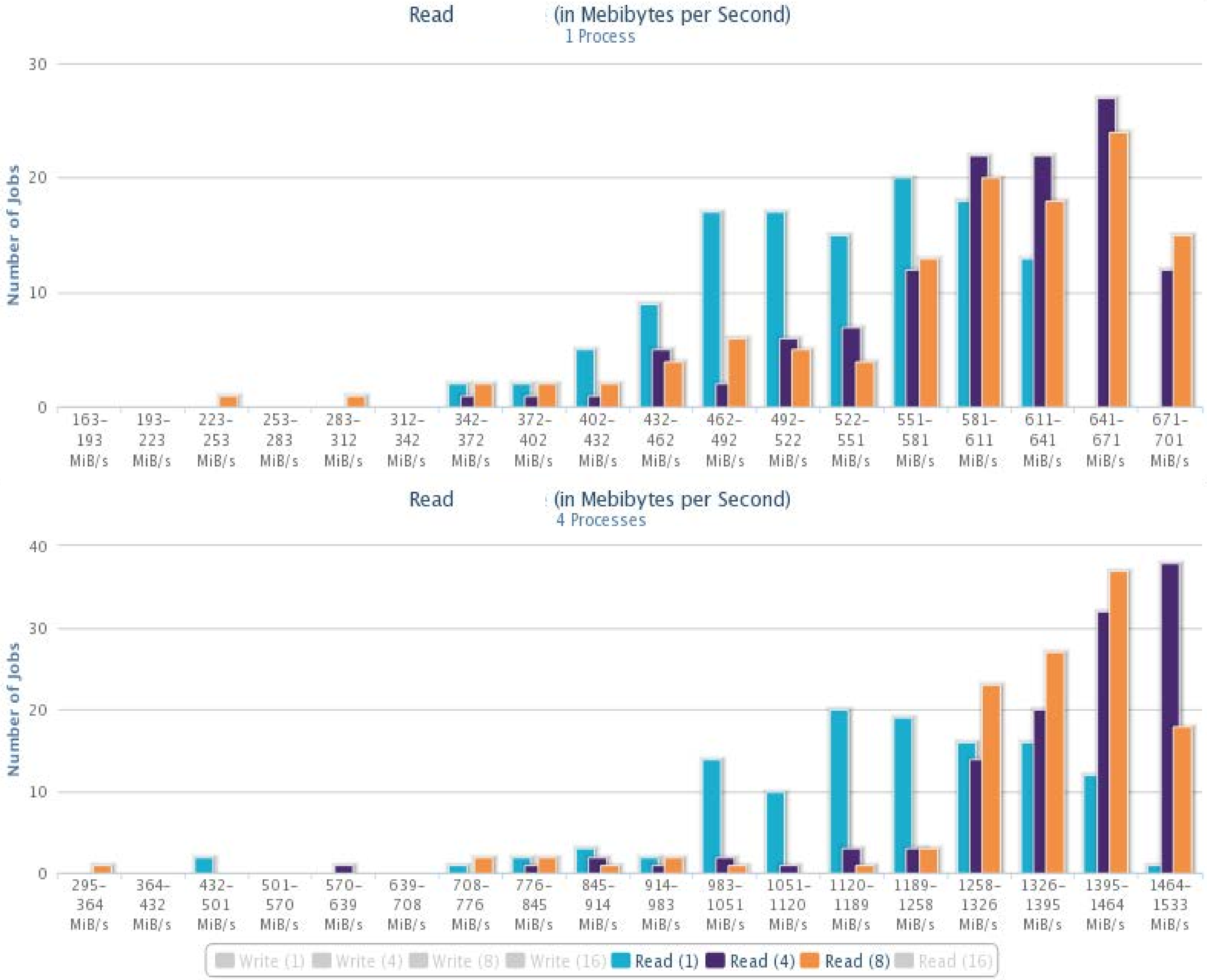


1 OSS
4 Stripes
1MB Stripe Size



Write (in Mebibytes per Second)
1 Process

Write (in Mebibytes per Second)
4 Processes


Read (in Mebibytes per Second)
1 Process
Number of Jobs
0
10
20
30
163-193 MiB/s
193-223 MiB/s
223-253 MiB/s
253-283 MiB/s
283-312 MiB/s
312-342 MiB/s
342-372 MiB/s
372-402 MiB/s
402-432 MiB/s
432-462 MiB/s
462-492 MiB/s
492-522 MiB/s
522-551 MiB/s
551-581 MiB/s
581-611 MiB/s
611-641 MiB/s
641-671 MiB/s
671-701 MiB/s
Read (in Mebibytes per Second)
4 Processes
Number of Jobs
0
10
20
30
40
295-364 MiB/s
364-432 MiB/s
432-501 MiB/s
501-570 MiB/s
570-639 MiB/s
639-708 MiB/s
708-776 MiB/s
776-845 MiB/s
845-914 MiB/s
914-983 MiB/s
983-1051 MiB/s
1051-1120 MiB/s
1120-1189 MiB/s
1189-1258 MiB/s
1258-1326 MiB/s
1326-1395 MiB/s
1395-1464 MiB/s
1464-1533 MiB/s
Write (1)
Write (4)
Write (8)
Write (16)
Read (1)
Read (4)
Read (8)
Read (16)

Lustre Striping

Advantages	Disadvantages
Bandwidth – file objects are striped across OSTs, so bandwidth is the aggregate I/O rate	**User Overhead** – Time and thought required to understand your I/O patterns and create stripe layout for directories and files
File Size – file objects striped across OST can have a total size larger than available space on any single OST	**System Overhead** – Additional stripes means more OST lookups to determine the size of the file (more time)

- Striping will not benefit *ALL* applications

Take Home Message

Choose Lustre!

It scales – size, speed, clients

It’s open, growing, and needs your help

It gives users powerful options

Tools available to help with installation

Filets, Chops, Removes household odors

Act now and no salesman will visit your home

チキンフィレ
サンド
季節限定サンド
CHOOSE
LUSTRE

Thank You for Your Time and Attention

Open Scalable File Systems, Inc.
3855 SW 153rd Drive
Beaverton, OR 97006
Ph: 503-619-0561
Fax: 503-644-6708
admin@opensfs.org

www.opensfs.org