Massively Scalable Filesystems

Lustre Parallel Filesystem

Wil Mayers
Alces Software Ltd.

Agenda

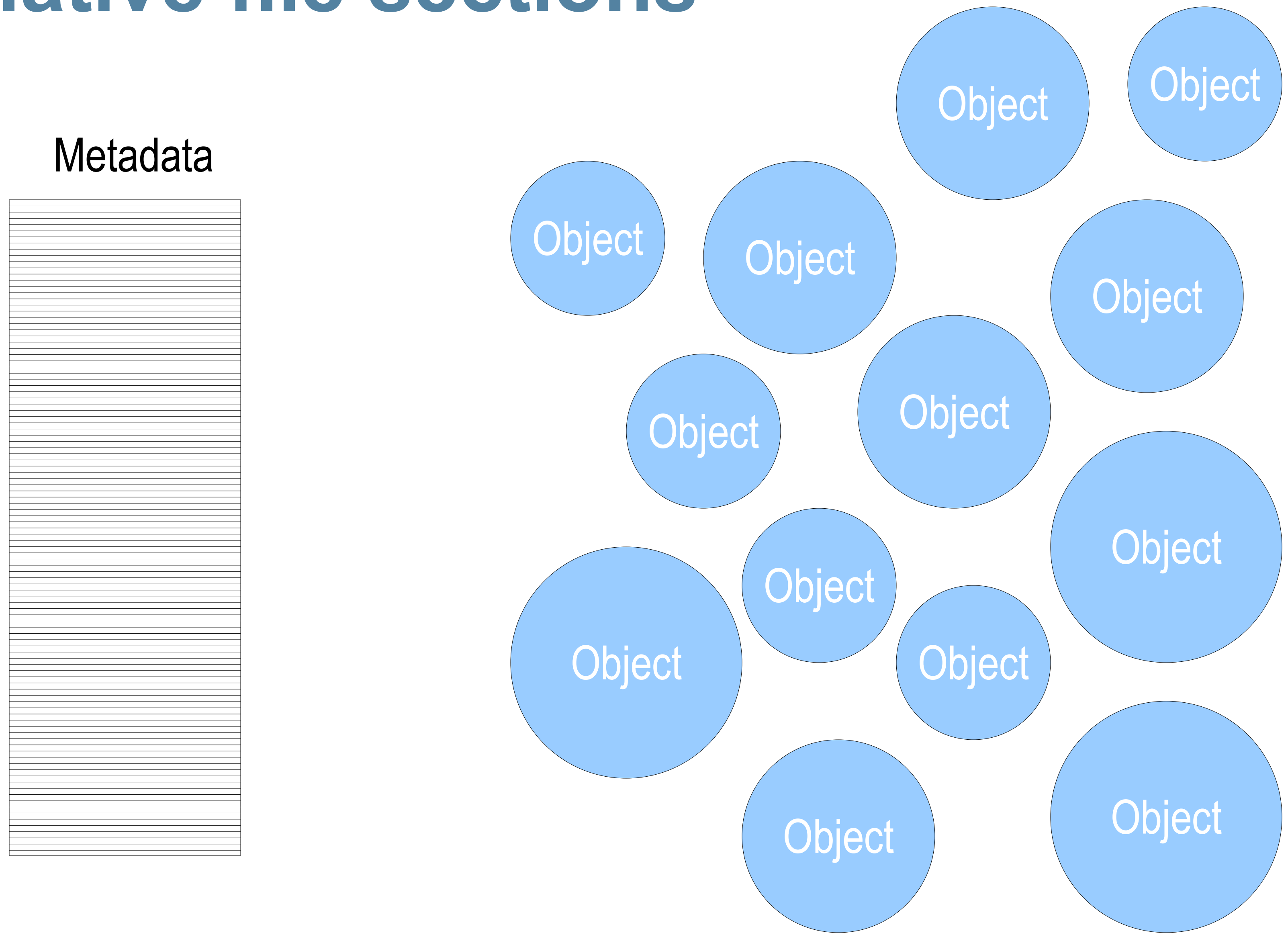
- Lustre Architecture
- Lustre at Sussex University
- Administrating Lustre
- Using and Troubleshooting Lustre

Lustre Architecture

Thinking about your files

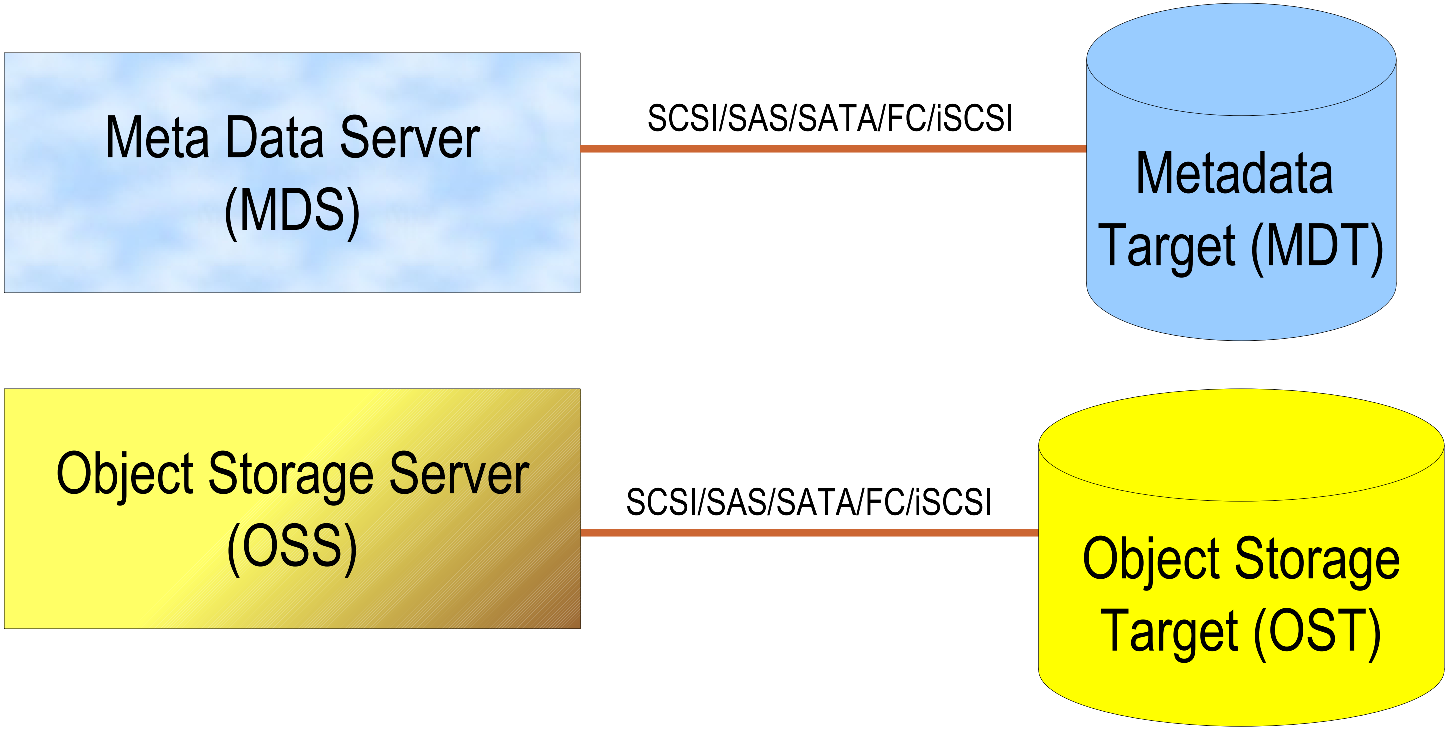
- Lustre provides an object storage service
- Files have metadata and data objects which are stored separately for maximum performance
- File metadata includes all file attributes
  - File name, size, permissions, ownership
  - Pointers to data objects
  - Metadata is useless without its file objects
- File objects hold the actual contents of the file
  - Objects may be any size, shape, compression level
  - Objects may be stored on any server
  - Objects are useless without their metadata

Relative file sections

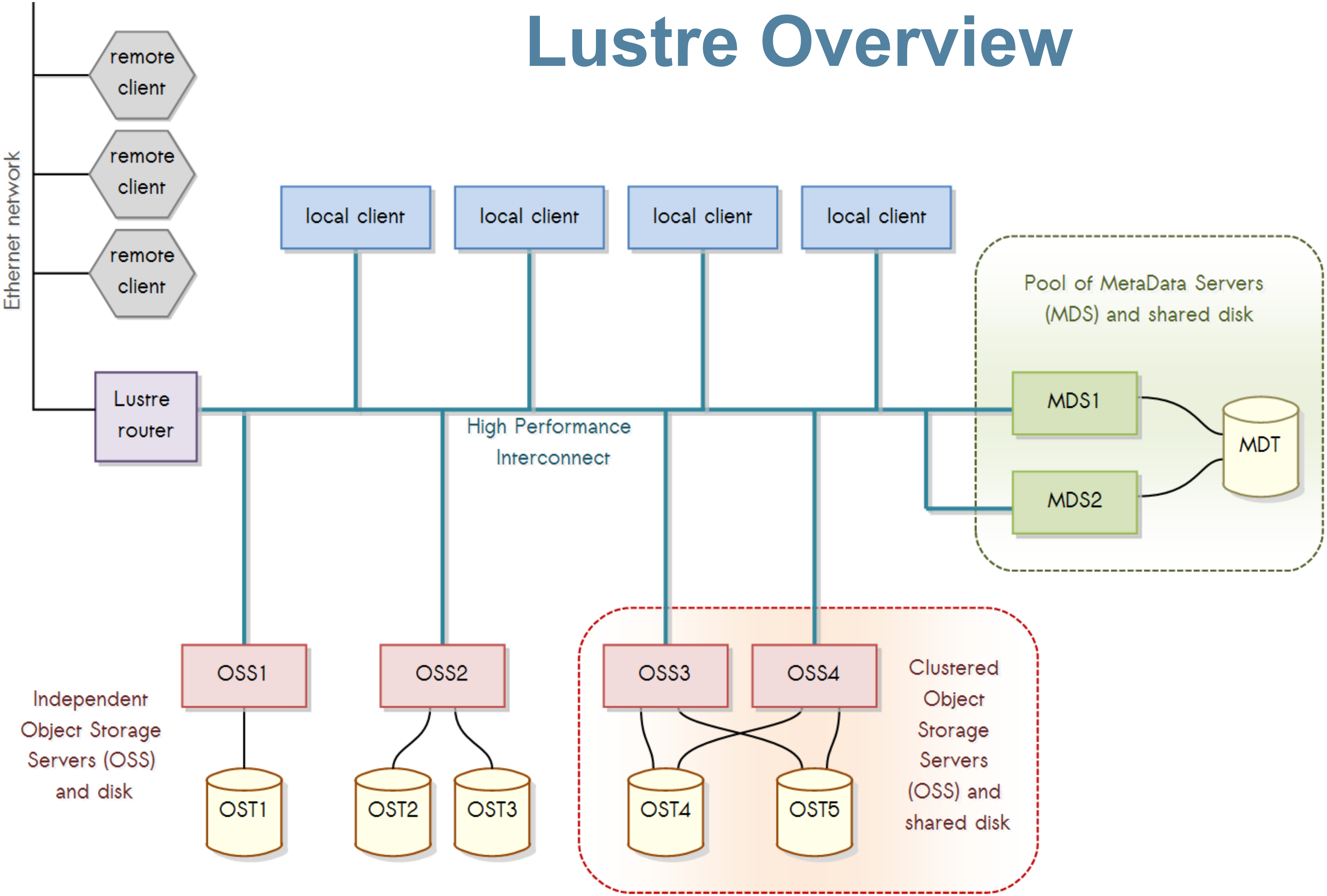


Lustre Servers

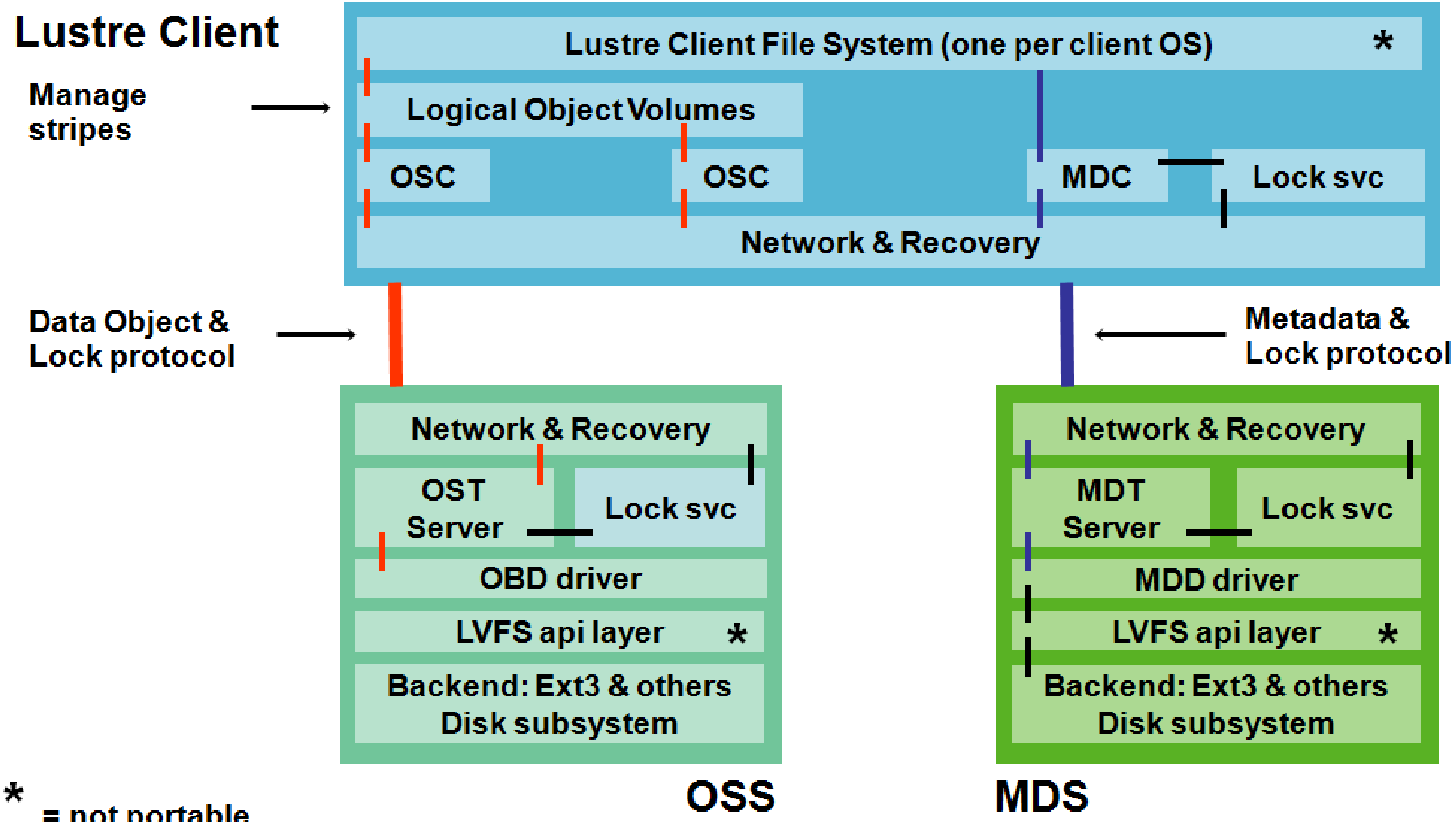
- Separate servers for meta data and object data



Lustre Overview


Ethernet network
remote client
remote client
remote client
Lustre router
local client
local client
local client
local client
Pool of MetaData Servers (MDS) and shared disk
MDS1
MDS2
MDT
High Performance Interconnect
Independent Object Storage Servers (OSS) and disk
OSS1
OSS2
OSS3
OSS4
OST1
OST2
OST3
OST4
OST5
Clustered Object Storage Servers (OSS) and shared disk

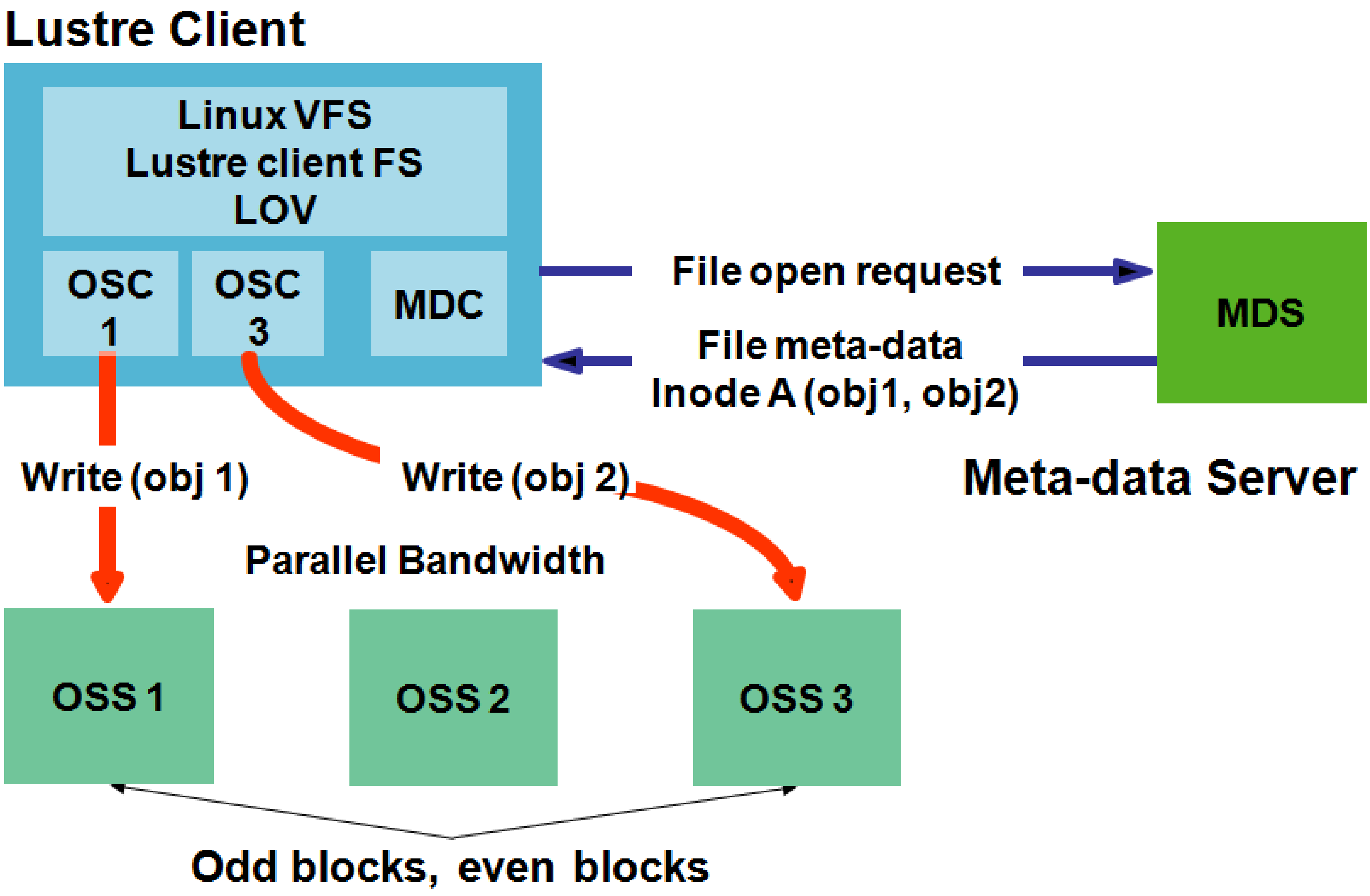
Lustre Architecture



* = not portable

observe reuse of many modules

Lustre file transactions



Stand-alone servers

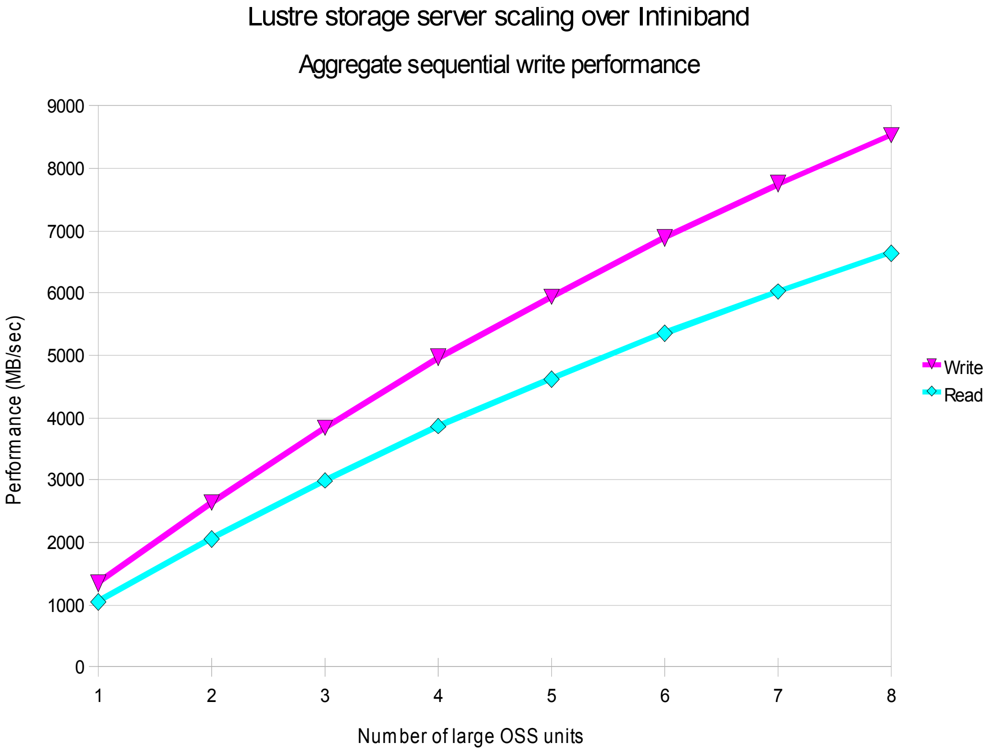
- Basic server with attached storage
- Need reasonably powerful servers
  - Dual 2.4Ghz quad-core CPUs; 5520 chipset preferred
  - At least 12GB RAM
  - Redundant power supplies
  - Network/interconnect card
- Plenty of bandwidth to storage
  - No more than 24 disks per SAS card
  - Can use SATA or SAS drives
  - Remember boot drives, parity and spare disks

High availability servers

- Lustre does not provide file level redundancy
- High availability achieved using two servers
  - > Both servers must be identical
  - > Both servers must have dedicated boot drives
  - > Lustre storage must be dual attached to both servers
    - – Requires SAS disk drives or disk array
    - – HW RAID cards in servers CANNOT be used
  - > Metadata servers: active/passive mode only
  - > Object storage servers: active/active possible

OSS unit scaling

- Capacity scales linearly as OSS are added
- Performance scales near-linearly as OSS added
- Interconnect bandwidth can limit scaling



Lustre Network Performance

Results per OSS (dual xeon/x86_64 server):

- GIGE: 118 MB/sec, 20k RPC/sec
- Trunked 4 x gigE: 400MB/sec, 30k RPC/
- Myrinet: 200 MB/sec, 35k RPC/sec
- 10gigE: 600-800MB/sec, 45k RPC/sec
- Infiniband: 700-1800MB/sec, 60k RPC/sec

Memory cache significantly helps performance

- OSS read cache
- Disk drive and controller cache
- Client side cache

Please note: These are example performance figures and may not actually be true

Typical Lustre bottlenecks

Storage volumes

- Disk drive data bandwidth and IOPS
- Disk controllers saturation

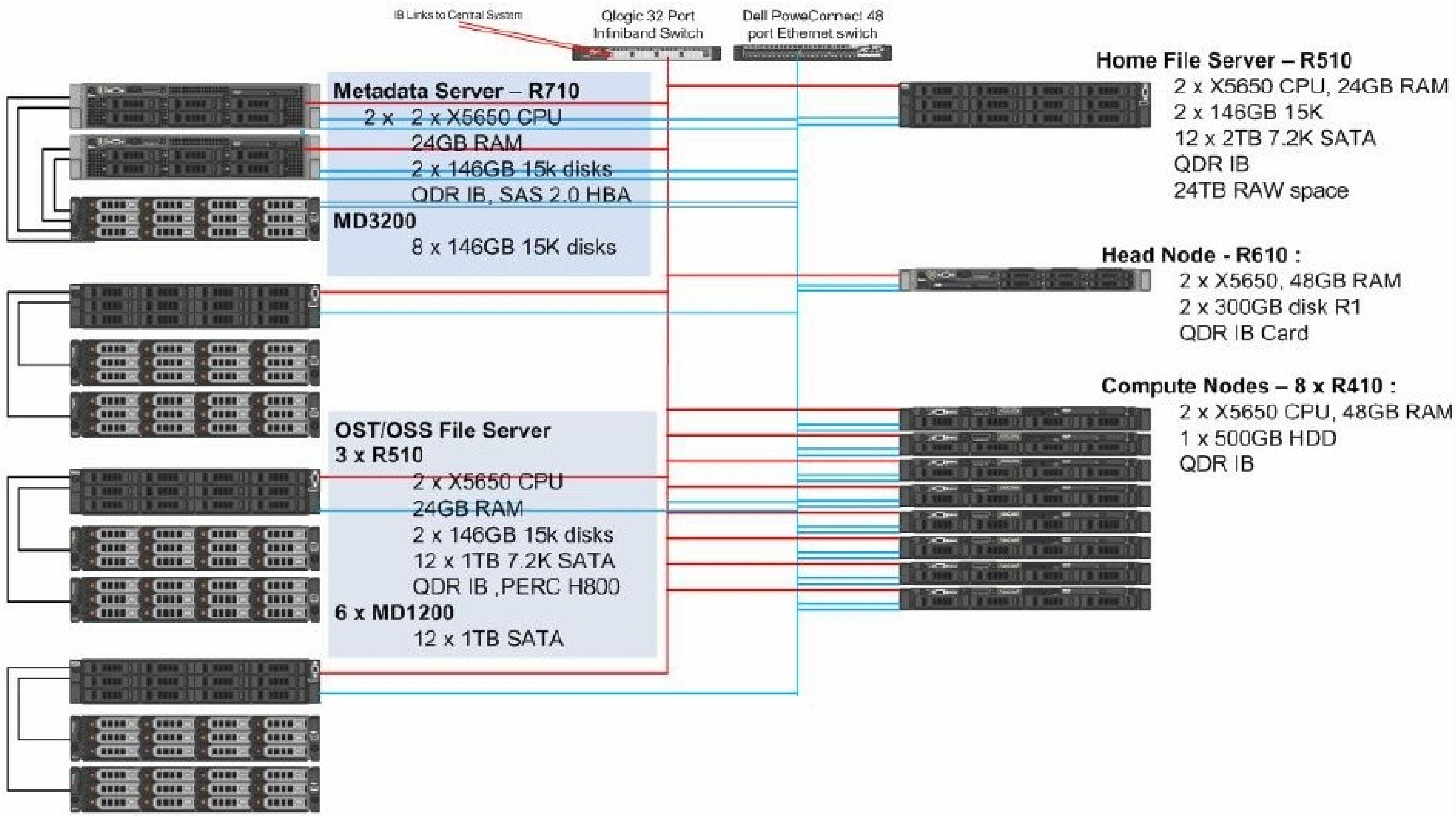
Interconnect performance

- Ethernet collisions
- Infiniband oversubcription
- Interaction with other traffic

Per-server thread limits

- OSS thread limitations (~2GB/s)
- MDS thread limitations (~15,000 RPC/s)
- Client limitations (~1GB/s)

Sussex Uni Lustre Solution



IB Links to Central System

Qlogic 32 Port Infiniband Switch

Dell PoweConnect 48 port Ethernet switch

Metadata Server – R710
2 x 2 x X5650 CPU
24GB RAM
2 x 146GB 15k disks
QDR IB, SAS 2.0 HBA

MD3200
8 x 146GB 15K disks

OST/OSS File Server
3 x R510
2 x X5650 CPU
24GB RAM
2 x 146GB 15k disks
12 x 1TB 7.2K SATA
QDR IB ,PERC H800

6 x MD1200
12 x 1TB SATA

Home File Server – R510
2 x X5650 CPU, 24GB RAM
2 x 146GB 15K
12 x 2TB 7.2K SATA
QDR IB
24TB RAW space

Head Node - R610 :
2 x X5650, 48GB RAM
2 x 300GB disk R1
QDR IB Card

Compute Nodes – 8 x R410 :
2 x X5650 CPU, 48GB RAM
1 x 500GB HDD
QDR IB

Solution Diagram		Customer:	Sussex University
		Date:	28/06/2010
Project	ATLAS HPC	Dell SC	Nick Jefferson

Created with ALF/PD
Dell Confidential Document

Lustre filesystem solution

Metadata server (MDS) pair

- High availability MDS pair
- 2 x Dell R610 1U servers
  - Dual 2.4Ghz processors
  - 12GB RAM
  - RAID1 system disks
  - Qlogic 7340 HCA
- Shared metadata target (MDT)
  - Dell PowerVault MD3200 SAS array
  - Multi-path dual controller connection
  - 8 x 300GB SAS disks (300M files)

Lustre filesystem solution

Object storage servers (OSS)

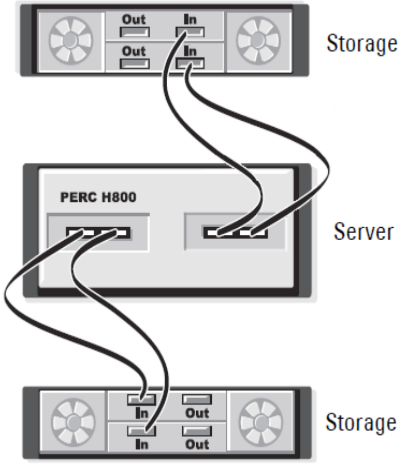
- Three identical standalone OSS machines
- Dell R510 2U servers
  - – Dual 2.4Ghz processors, 24GB RAM
  - – RAID1 system disks
  - – Qlogic 7340 HCA
- Three OST volumes
  - > ost1 = PERC H700 + 12 x 2TB SATA drives
  - > ost2 = PERC H800 + 12 x 2TB SAS drives
  - > ost3 = PERC H800 + 12 x 2TB SAS drives
- Two connected MD1200 arrays per OSS

Physical cabling

Parallel storage rack

	layout	power (watt)	weight (kg)	SAS C0P0	SAS C0P1	SAS C1P0	SAS C1P1	eth0	eth1	eth2	eth3
				SAS cabling targets				**LAN cabling targets**			
42											
41											
40											
39											
38											
37											
36											
35											
34											
33											
32											
31											
30											
29											
28											
27											
26											
25											
24											
23											
22	R610 MDS server 1	385	18	MDTC0P0	MDTC1P0			Cluster LAN	IPMI LAN	-	heartbeat LAN
21	R610 MDS server 2	385	18	MDTC0P1	MDTC1P1			Cluster LAN	IPMI LAN	-	heartbeat LAN
20 19	MD3200 array (MDT)	150	25								
18 17	R510 OSS1 (O1)	420	30	O1A1C0	O1A1C1	O1A2C0	O1A2C1	Cluster LAN	-		
16 15	MD1200 array 1 (O1A1)	150	22								
14 13	MD1200 array 2 (O1A2)	150	22								
12 11	R510 OSS2 (O2)	420	30	O2A1C0	O2A1C1	O2A2C0	O2A2C1	Cluster LAN	-		
10 9	MD1200 array 1 (O2A1)	150	22								
8 7	MD1200 array 1 (O2A2)	150	22								
6 5	R510 OSS3 (O3)	420	30	O3A1C0	O3A1C1	O3A2C0	O3A2C1	Cluster LAN	-		
4 3	MD1200 array 1 (O3A1)	150	22								
2 1	MD1200 array 1 (O3A2)	150	22								

Rack weight		150
Total power/weight	3,080	433
Total heat (BTU/hr)	10,509	



Administrating Lustre

Mounting Lustre

Start-up and shut-down procedures

- Lustre now integrated with *mountconf*
- MDS and OSS server control
- Client mounts
- Start-up order
  - OSS machines first
  - MDS machines second
  - Clients last
- Shut-down order
  - Unmount clients first
  - Shutdown MDS second
  - Shutdown OSS last

Authentication and user control

Integration with Cluster services

- IPoIB addressing for Lustre
- DNS and host name integration
- User authentication
  - LDAP integration for users
  - Critical for Lustre to work properly
- Logging in to the systems
  - Standard users cannot log in
  - Admin login as privileged user
  - Call-home service for support

Managing system security

Passwords and data access

- Lustre uses standard Linux authentication
- POSIX access restrictions apply
- User permissions from LDAP
- *norootsquash* by default
- Passwords should be changed regularly
  - Remember to update *heartbeat* fail-over configuration for MDS servers when changing passwords

High availability services

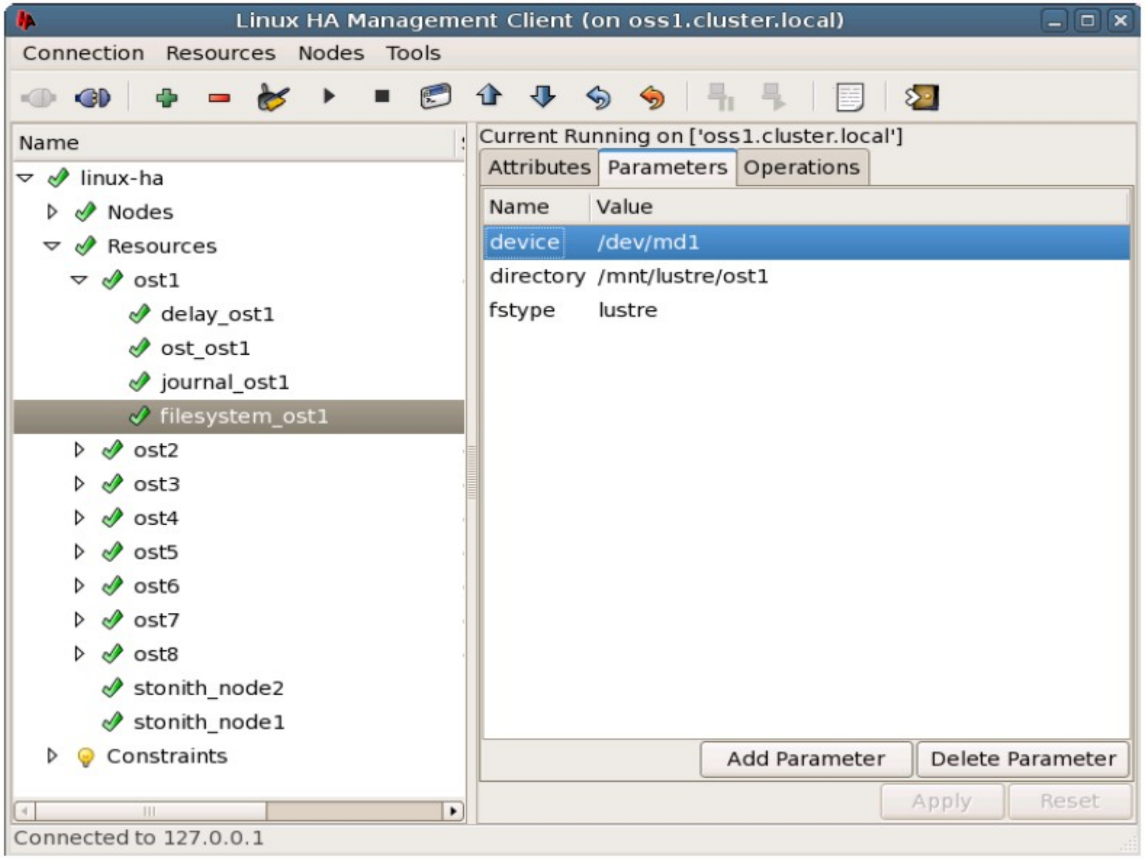
Configured for Metadata server on *mds1/mds2*

- Lustre MDS service is automatically started and stopped via *heartbeat*
- *Heartbeat* will fail to start if Lustre is mounted manually
- STONITH configured via Dell BMC
- Be extra careful when performing system administration to avoid accidental failover
- *heartbeat* service does not start automatically at boot; after failure, check servers over first before re-enabling
- MDS service is sticky

High availability services

Administration with graphical user interface

- Use the *heartbeat-gui* package for administration:



- linux-ha
  - Nodes
  - Resources
    - ost1
      - delay_ost1
      - ost_ost1
      - journal_ost1
      - filesystem_ost1
    - ost2
    - ost3
    - ost4
    - ost5
    - ost6
    - ost7
    - ost8
    - stonith_node2
    - stonith_node1
  - Constraints

Current Running on ['oss1.cluster.local']

Name	Value
device	/dev/md1
directory	/mnt/lustre/ost1
fstype	lustre

Connected to 127.0.0.1

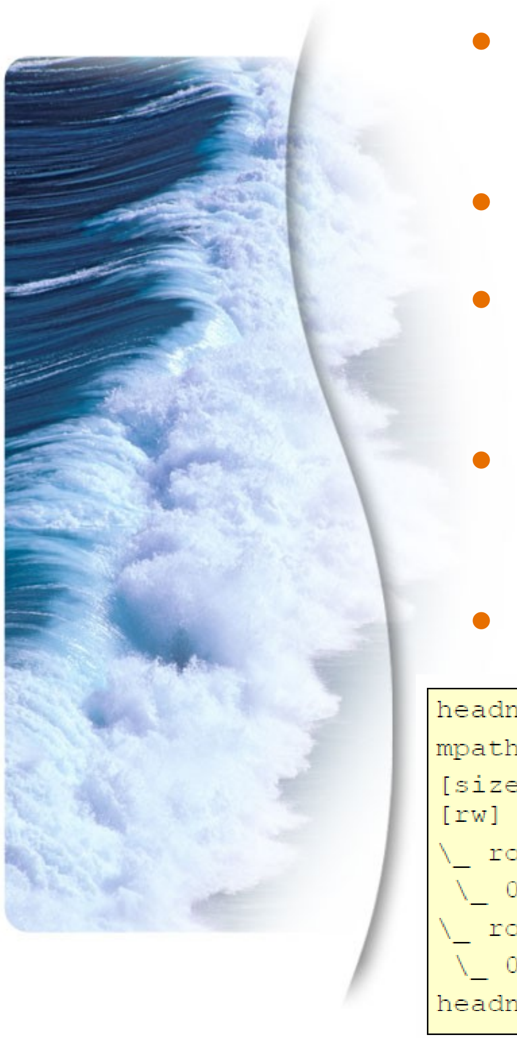
RAID volumes

Configuration, monitoring and administration

- Metadata servers
  - system disks monitored via *checkraid* command
  - MD3200 monitored via Dell storage manager
- Object storage servers
  - disks monitored via MegaCLI utility
  - *check-lustre-oss* command quickly confirms status of all RAID volumes
- All groups are RAID1/RAID6 protected
  - Disk failures reported by these commands
  - Replace failed disks as soon as possible
  - If in doubt, email support@alces-software.com

Managing multi-path disk devices

Metadata server shared storage array targets



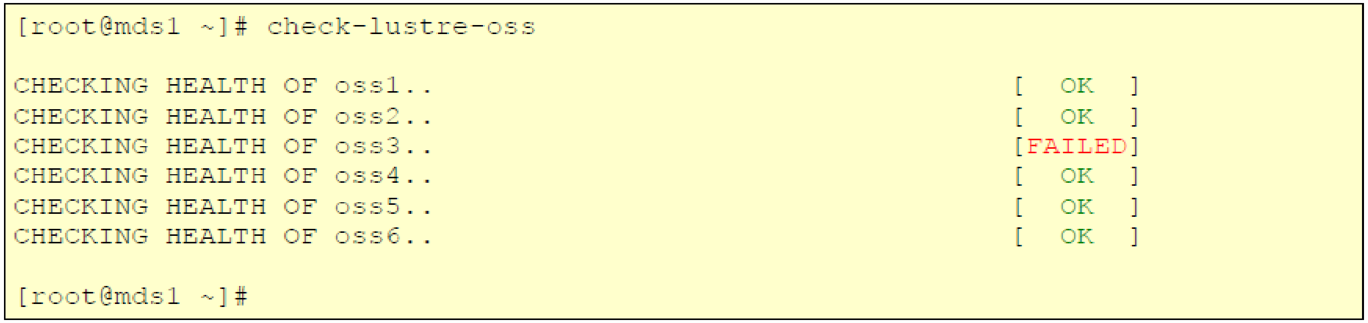
- Dell MD3200 supports Linux multi-path drivers
- One controller owns volume
- Second controller presents *ghost* device, but does not allow I/O
- On failure of the primary controller, secondary controller takes over
- Use *multipath* command to view status:

```
headnode:/ # multipath -ll
mpath1 (36842b2b00018fcf421000026623e722f62) dm-0 DELL,MD32xx
[size=1.4T][features=3 queue_if_no_path pg_init_retries 50][hwhandler=1 rdac]
[rw]
\_ round-robin 0 [prio=100][active]
 \_ 0:0:1:0  sdc 8:32  [active][ready]
\_ round-robin 0 [prio=0][enabled]
 \_ 0:0:0:0  sda 8:0   [active][ghost]
headnode:/ #
```

RAID volumes

Configuration, monitoring and administration

- Object storage servers
  - *check-lustre-oss* command confirms status of Lustre OST filesystems:



```
[root@mds1 ~]# check-lustre-oss

CHECKING HEALTH OF oss1..                                    [  OK  ]
CHECKING HEALTH OF oss2..                                    [  OK  ]
CHECKING HEALTH OF oss3..                                    [FAILED]
CHECKING HEALTH OF oss4..                                    [  OK  ]
CHECKING HEALTH OF oss5..                                    [  OK  ]
CHECKING HEALTH OF oss6..                                    [  OK  ]

[root@mds1 ~]#
```

  - Application monitors for change from known-good configuration
  - Will detect disk volume and controller changes
  - Will notice if server system changes significantly
  - Reset configuration after server maintenance

Using and Troubleshooting Lustre

Lustre filesystem usage

How to use a Lustre filesystem

- Lustre is mounted on nodes and headnode system like a normal POSIX filesystem
- Save files, read them, set permissions as normal
- Mount information added into /etc/fstab just like a normal network filesystem
- Can be used without special knowledge
  - Administrators can determine default storage policies
  - When unavailable, Lustre mount will hang like a hard-mounted NFS server
  - Best practice is not to mount in root fs

Lustre file striping

Where files are stored

- Lustre has a default file storage policy
  - New files are **not** striped over OSTs
  - New blocks allocated on the best OST
- Default policy can be modified for individual files, directories or entire tree
  - Target OST pool (by name)
  - Stripe count (how many OSTs to use)
  - Stripe size (how much data per OST)
- Existing files are unaffected
  - Support script available to redistribute filesystem contents

Lustre file striping

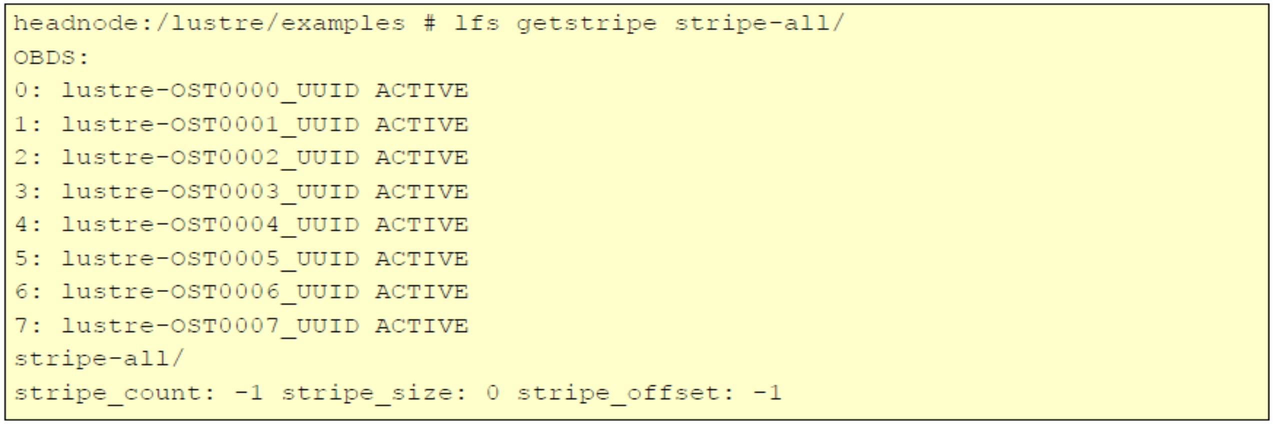
Why stripe?

- Striping large files can lead to massive performance increases
- 300MB/sec on unstriped file
  - single OST
  - 12 disks in RAID6
  - one QDR IB adapter
- 4GB/sec on same filesystem for striped files
  - nine OSTs
  - 108 disks in nine separate RAID6 groups
  - three QDR IB adapters
- Benefit from OSS cache, less contention, more controllers

Lustre file striping

Where files are stored

- Use the *lfs getstripe* and *lfs setstripe* commands to influence policies:



```
headnode:/lustre/examples # lfs getstripe stripe-all/
OBDS:
0: lustre-OST0000_UUID ACTIVE
1: lustre-OST0001_UUID ACTIVE
2: lustre-OST0002_UUID ACTIVE
3: lustre-OST0003_UUID ACTIVE
4: lustre-OST0004_UUID ACTIVE
5: lustre-OST0005_UUID ACTIVE
6: lustre-OST0006_UUID ACTIVE
7: lustre-OST0007_UUID ACTIVE
stripe-all/
stripe_count: -1 stripe_size: 0 stripe_offset: -1
```

- Forced striped settings can have unexpected results
- File availability considerations
- Extra stripe information stored on MDS

Lustre filesystem administration

Querying filesystem space

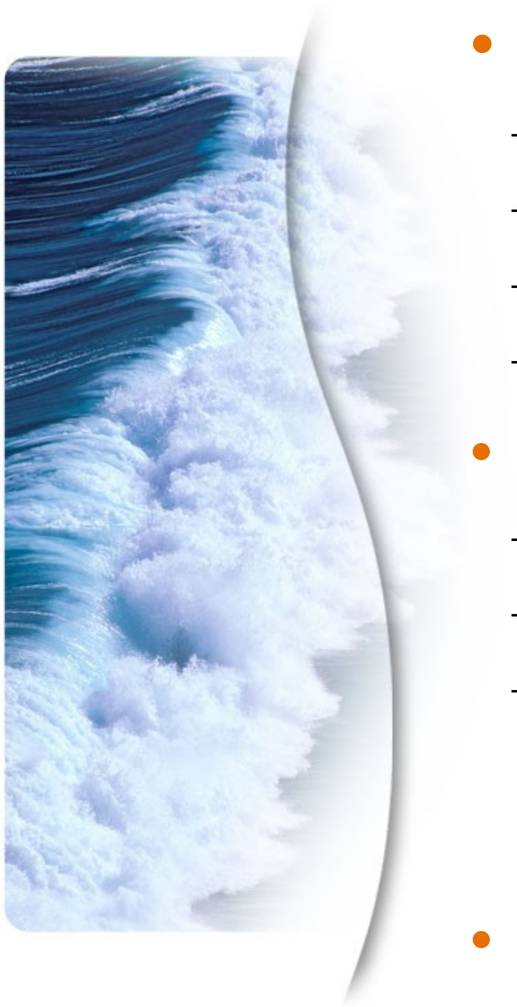
- The standard *df -h* command reports the total available space on the filesystem
- Filesizes are estimates (*glimpse* method)
- The *lfs df* command reports usage per Lustre server:

```
# lfs df
UUID 1K-blocks  Used Available  Use% Mounted on
mds-lustre-0_UUID 9174328  1020024  8154304  11% /mnt/lustre[MDT:0]
ost-lustre-0_UUID 94181368 56330708  37850660  59% /mnt/lustre[OST:0]
ost-lustre-1_UUID 94181368 56385748  37795620  59% /mnt/lustre[OST:1]
ost-lustre-2_UUID 94181368 54352012  39829356  57% /mnt/lustre[OST:2]
filesystem summary:282544104167068468 39829356  57% /mnt/lustre

# lfs df -i
UUID  Inodes  IUsed  IFree  IUse% Mounted on
mds-lustre-0_UUID 2211572 41924 2169648 1% /mnt/lustre[MDT:0]
ost-lustre-0_UUID 737280 12183 725097 1% /mnt/lustre[OST:0]
ost-lustre-1_UUID 737280 12232 725048 1% /mnt/lustre[OST:1]
ost-lustre-2_UUID 737280 12214 725066 1% /mnt/lustre[OST:2]
filesystem summary: 2211572  41924  2169648 1%  /mnt/lustre[OST:2]
```

Lustre troubleshooting

What to do when things go wrong



- Filesystem availability requires
  - MDS to be up and working
  - All OSS to be up and working
  - Infiniband networks to be working
  - Ethernet networks to be working
- Problem components cause filesystem hangs
  - Similar to hard-mounted NFS filesytem
  - Jobs will (should) wait on blocked I/O
  - No loss of data will occur during a reboot of
    - Any single Lustre server
    - Client servers
- I/O will resume when filesystem is available

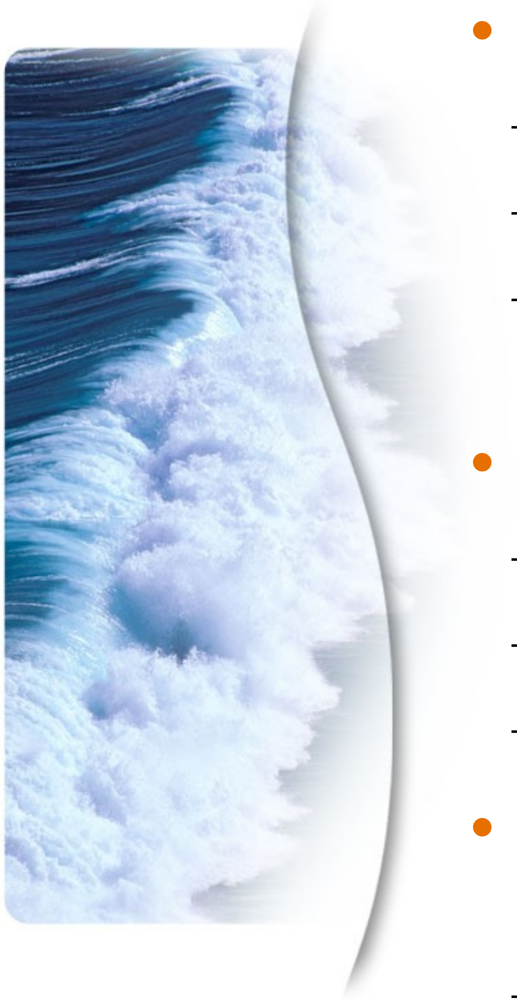
Lustre troubleshooting

Default recovery actions

- On failover, all clients automatically search for *failnode*
- Dead/slow/broken clients are *evicted* after a timeout period and I/Os are rolled back
  - Can be caused by interconnect failures
  - Prevents disruption of other clients
- On OSS reboot, clients cache I/O requests until server is available again
- On MDS reboot, filesystem enters recovery period while clients reconnect
  - 5 minute timeout while all clients reconnect
  - Any missing clients are evicted after timeout

Lustre troubleshooting

Lustre log files



- Lustre logs to kernel and system log files
  - Lots of status and debugging information
  - Primary method of diagnosing problems
  - Lustre bugs (LBUG) are rare and should be investigated further
- Most issues are not caused by Lustre
  - Interconnect problems
  - Name service unavailable (NIS/LDAP/AD)
  - Storage volume problems
- Contact support for assistance diagnosing problems
  - email *support@alces-software.com*

Lustre troubleshooting

Checking a Lustre filesystem

- Lustre uses *ext4* backing filesystems
- Storage volumes on MDS and OSS machines can be checked in parallel
- Lots of storage = long check process
- It is almost **never** necessary to use *lfsck* command
  - Checks Lustre filesystem integrity
  - Easy to remove important data by mistake
- Contact support for assistance diagnosing problems
  - email *support@alces-software.com*

Lustre troubleshooting

Steps to troubleshooting hanging Lustre filesystem

1. Is the problem affecting just one node?
   - Try from another filesystem client; if it works, try restarting the affected node
2. Are the MDS and OSS servers all running?
   - Check Lustre is mounted
   - Use the *check-lustre-oss* command to query status
3. Is the Lustre interconnect up?
   - Use the *ping* command to confirm
4. Is the local name service working?
   - Use *ypcat* or *getent* commands to confirm
5. Check the MDS and OSS logs for Lustre errors
6. Contact support – *support@alces-software.com*

Questions?