eXact

Lustre failover experience

Lustre Administrators

and Developers Workshop

Paris

September 25, 2012

TOC

- Who we are
- Our Lustre experience: the environment
- Deployment
- Benchmarks
- What's next

Who we are

Company for technology transfer

eXact

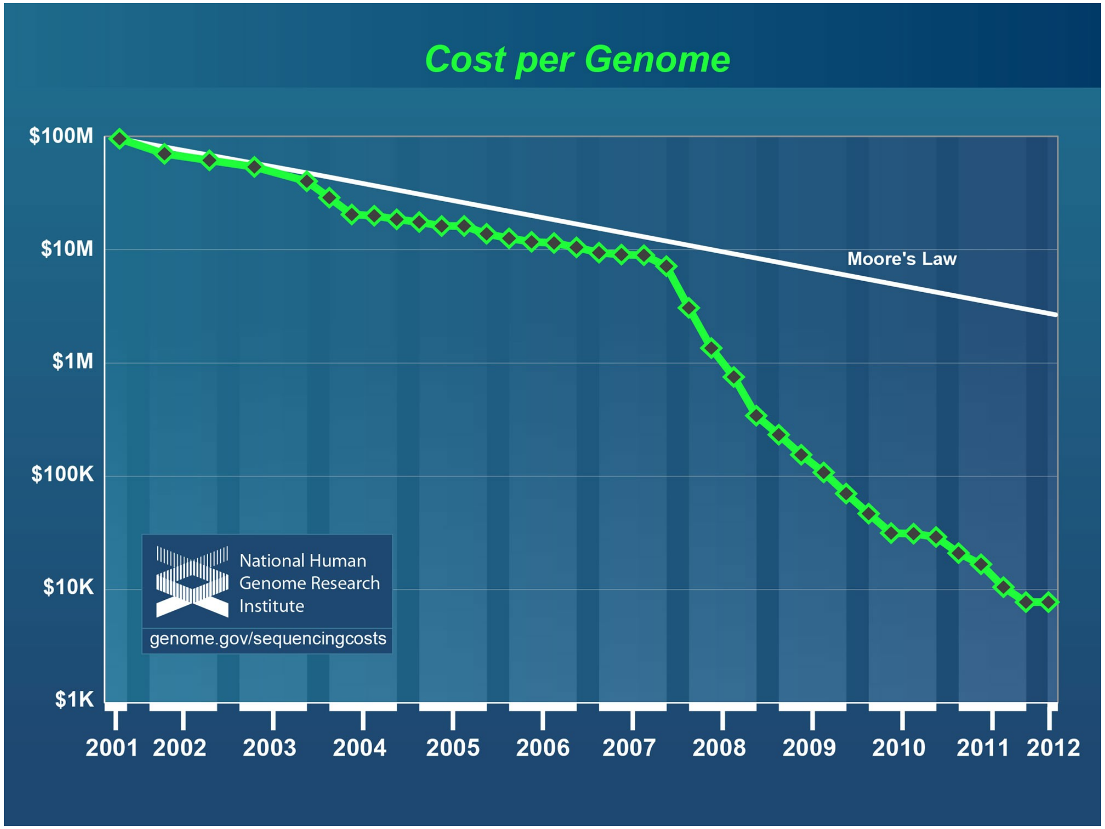
- HPC services
    - Cluster deployment
    - Storage solution
- Training
    - Sys admin and user oriented programs
- On-demand HPC

Environment

- Primary research institute in Italy
    - medical research
- Translational Genomic and Bioinformatics
    - personalized medicine: customization of healthcare by use of genetic information

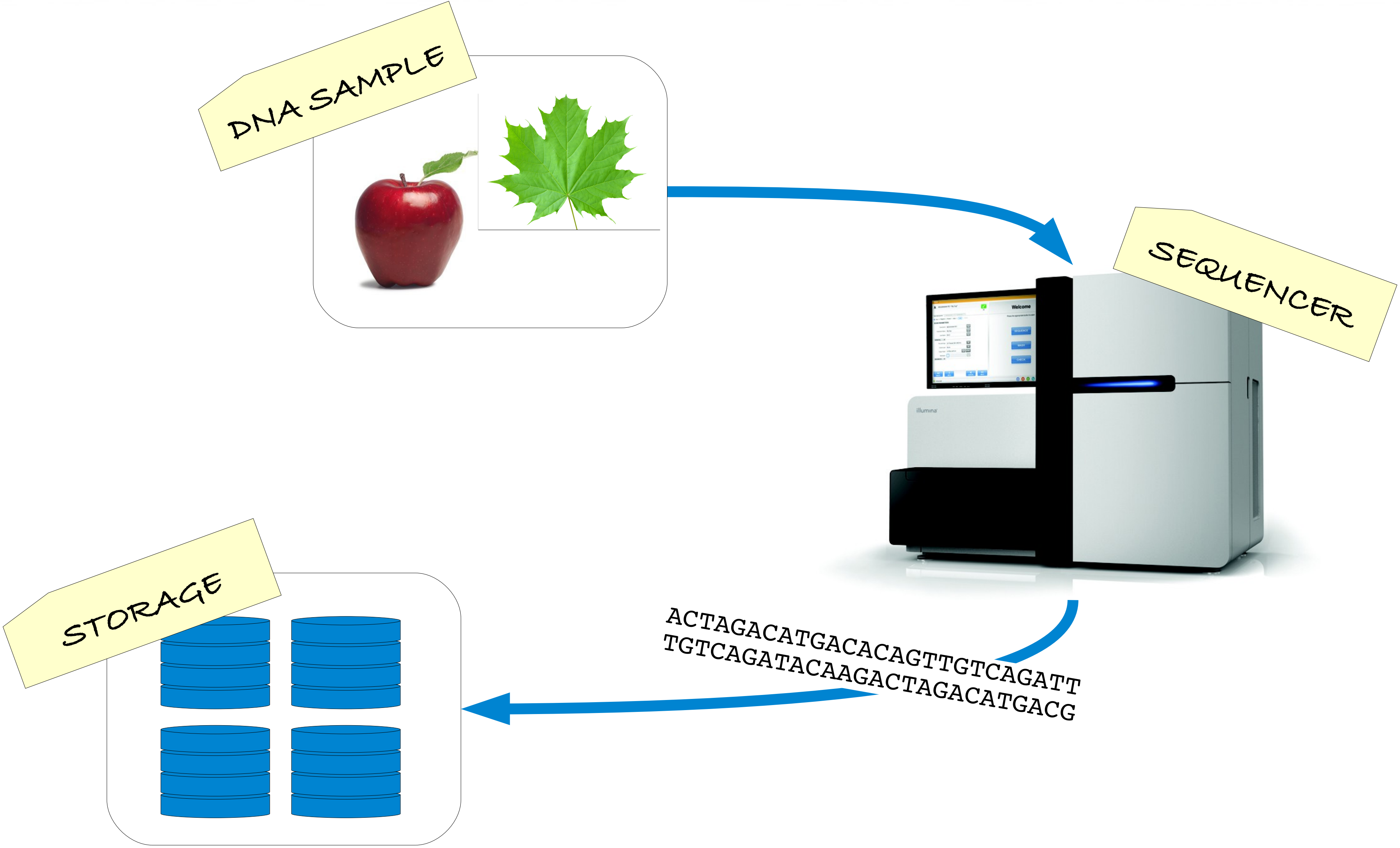
DNA sequencing

- High-throughput DNA sequencing
- The $1000 genome meme



Next Generation Sequencing is a big data problem!

Customer needs analysis



DNA SAMPLE

SEQUENCER

ACTAGACATGACACAGTTGTCAGATT
TGTCAGATACAAGACTAGACATGACG

STORAGE

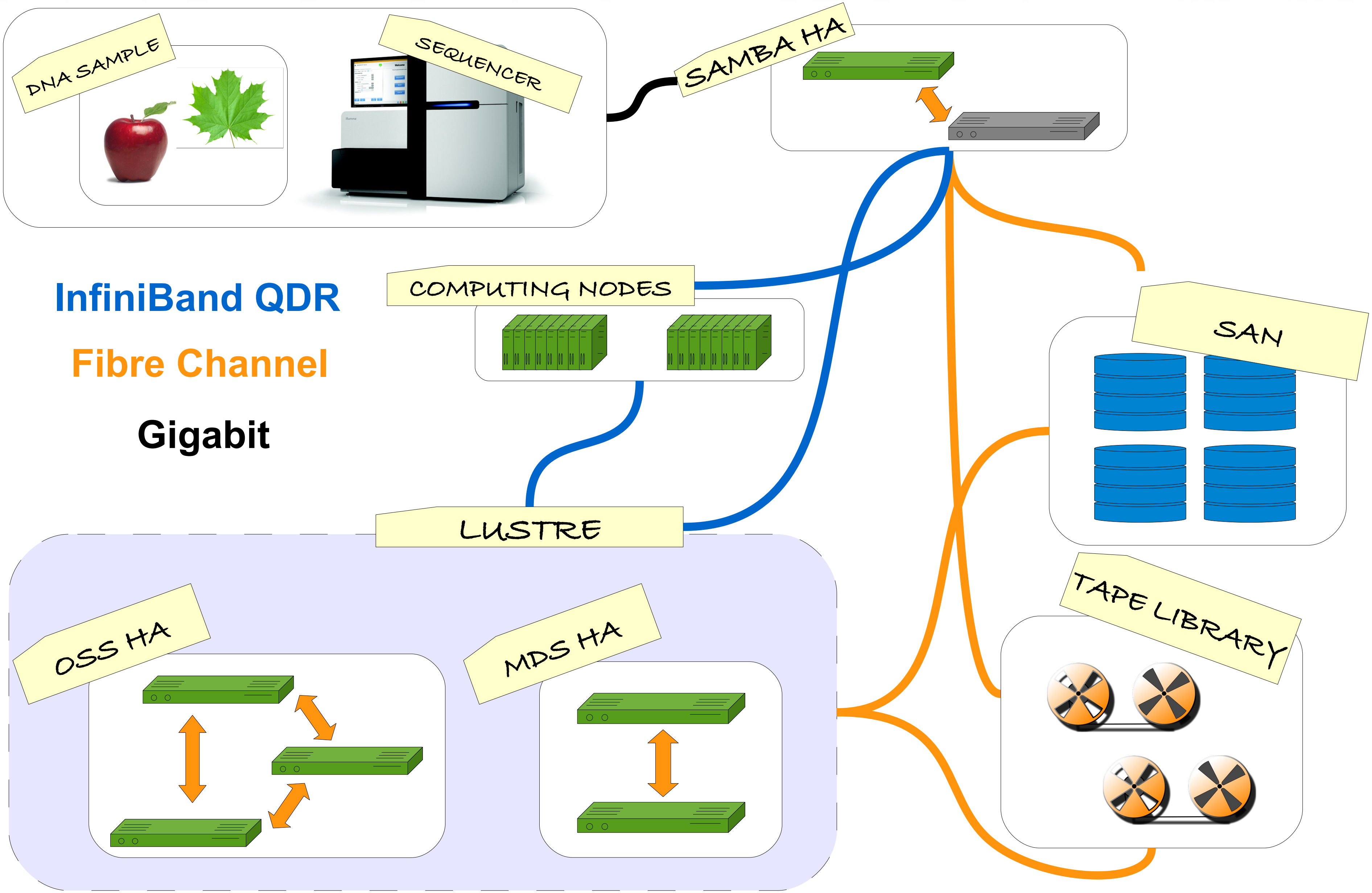
Customer needs analysis

- Lot of genomic data from Illumina Hi-Seq 2000
  - To backup (~20k € per run)
  - To post-process
  - Always available

Data from the sequencer need to be served to the computational infrastructure

Need for a fast, high performance, highly scalable file system, with robust failover and recovery mechanisms

Infrastructure



InfiniBand QDR

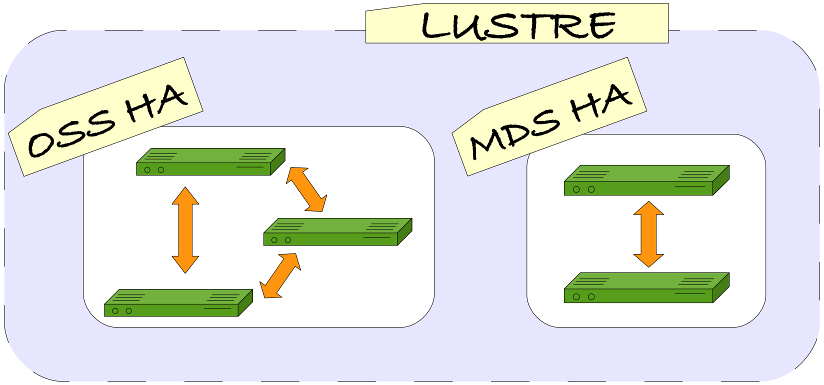
Fibre Channel

Gigabit

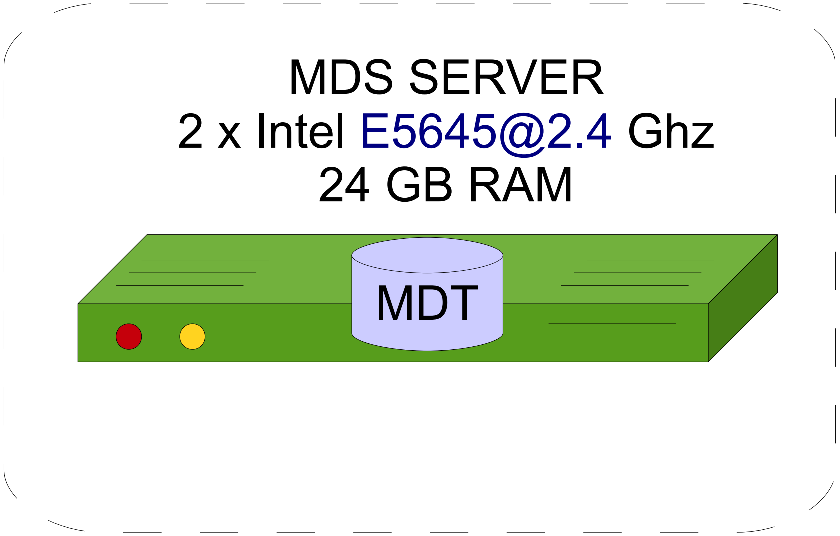
Lustre filesystem

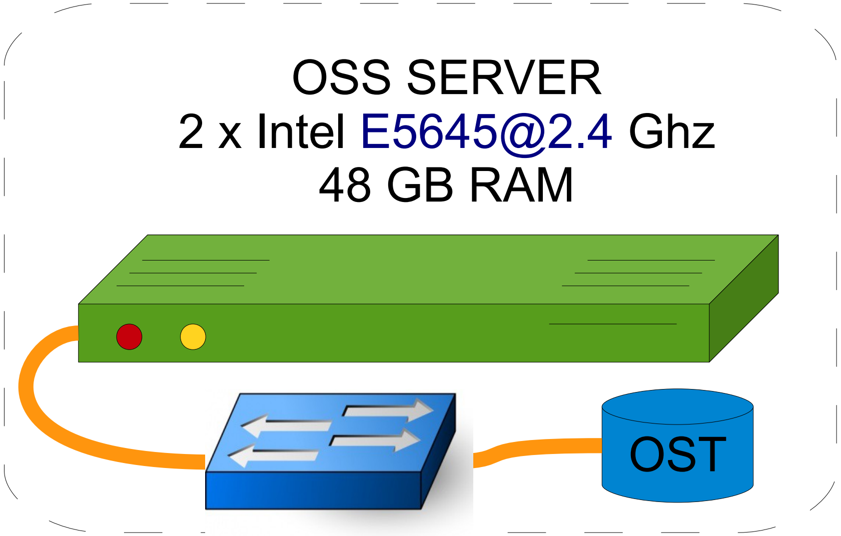
2 Lustre filesystems

- 2 MDSs, 3 OSSs
- ~50 clients
- 60 terabytes from SAN

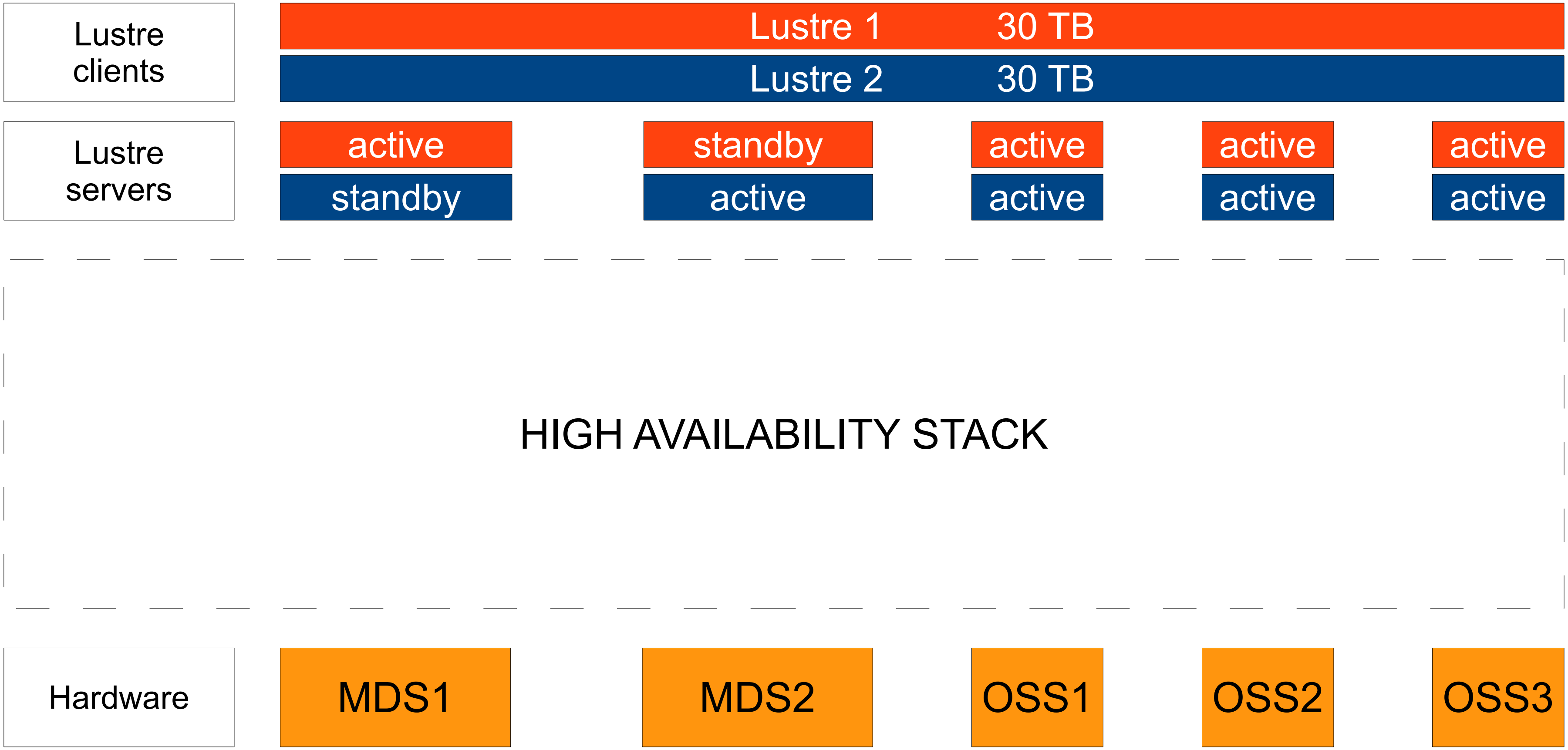


always available!


MDS SERVER
2 x Intel E5645@2.4 Ghz
24 GB RAM


OSS SERVER
2 x Intel E5645@2.4 Ghz
48 GB RAM

Lustre high availability



Lustre clients

Lustre 1 30 TB

Lustre 2 30 TB

Lustre servers

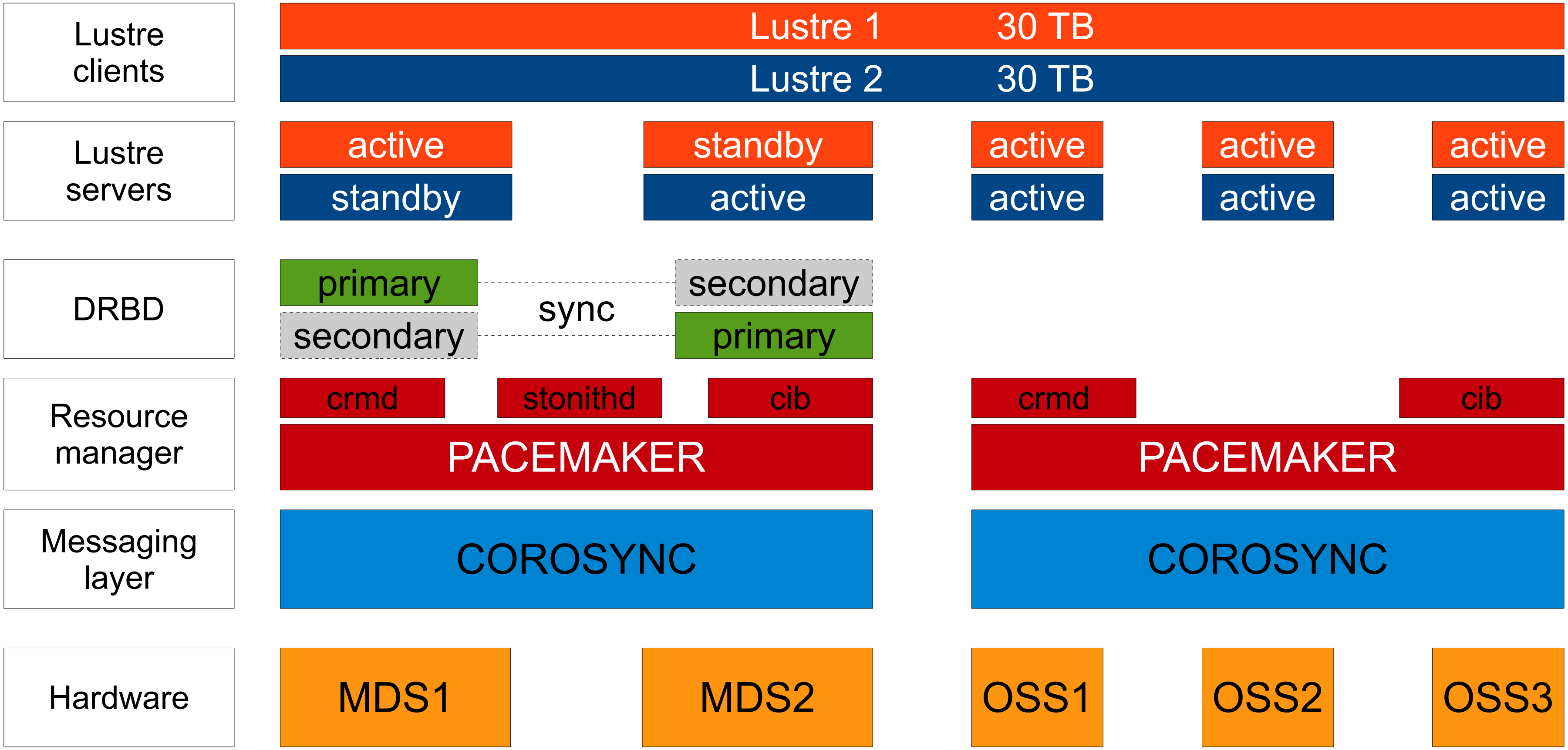
active	standby	active	active	active
standby	active	active	active	active

HIGH AVAILABILITY STACK

Hardware

MDS1	MDS2	OSS1	OSS2	OSS3

Lustre high availability



Lustre clients

Lustre 1 30 TB

Lustre 2 30 TB

Lustre servers

active standby active active active

standby active active active active

DRBD

primary secondary

secondary primary

sync

Resource manager

crmd stonithd cib

PACEMAKER

crmd cib

PACEMAKER

Messaging layer

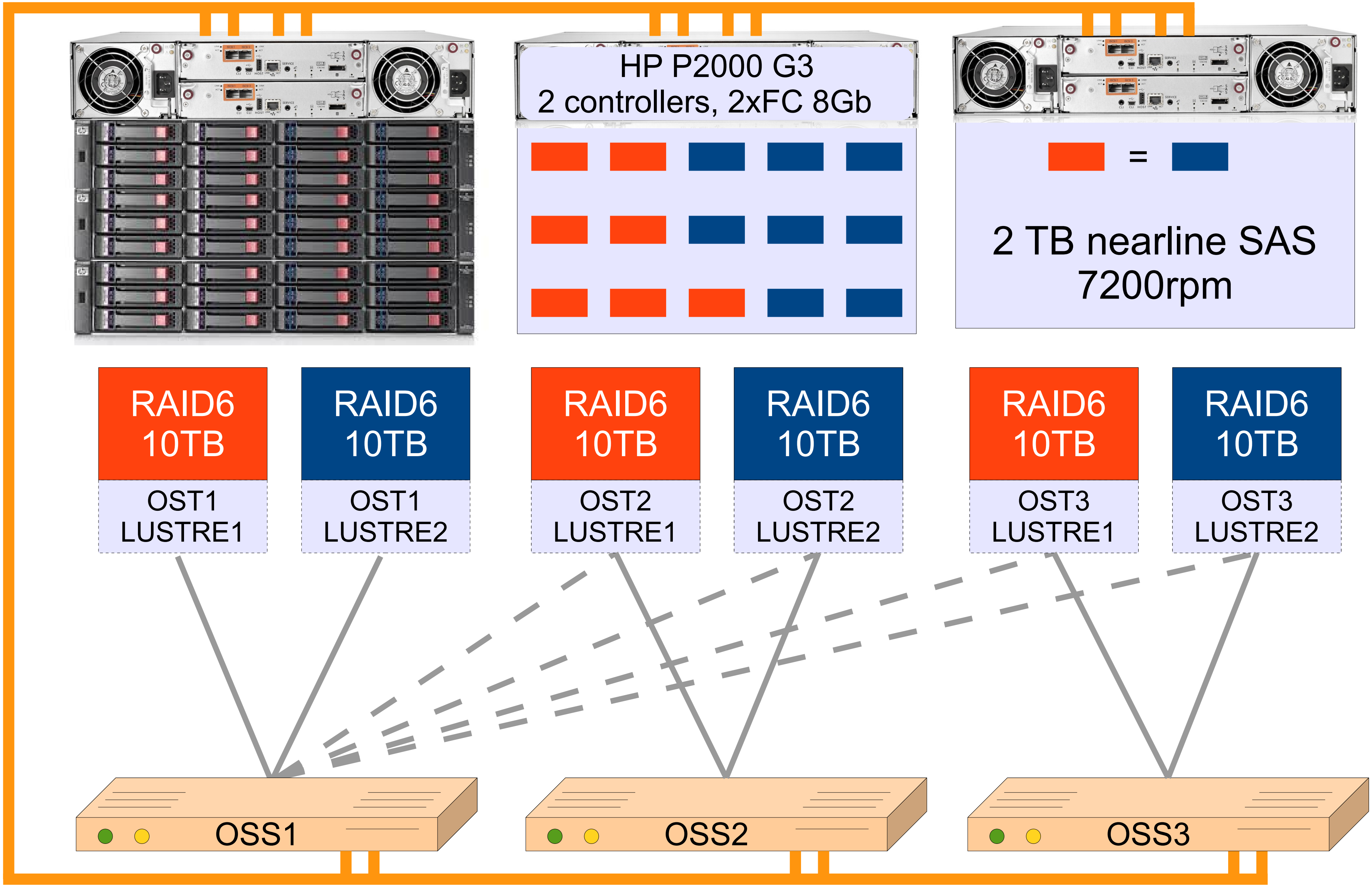
COROSYNC

COROSYNC

Hardware

MDS1 MDS2 OSS1 OSS2 OSS3

SAN provisioning for Lustre OSSs



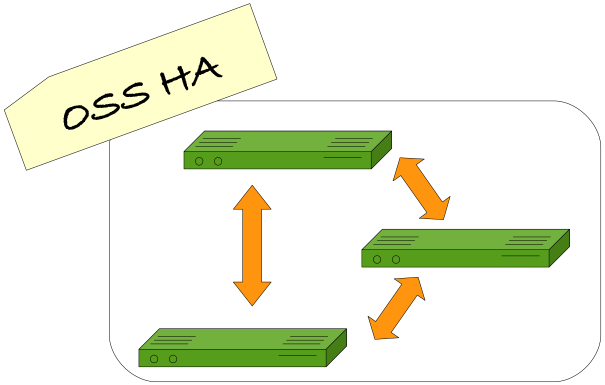
HP P2000 G3
2 controllers, 2xFC 8Gb

🟥 = 🟦

2 TB nearline SAS
7200rpm

High availability on OSSs

- Failures
  - Power*
  - Fibre channel*
  - InfiniBand*

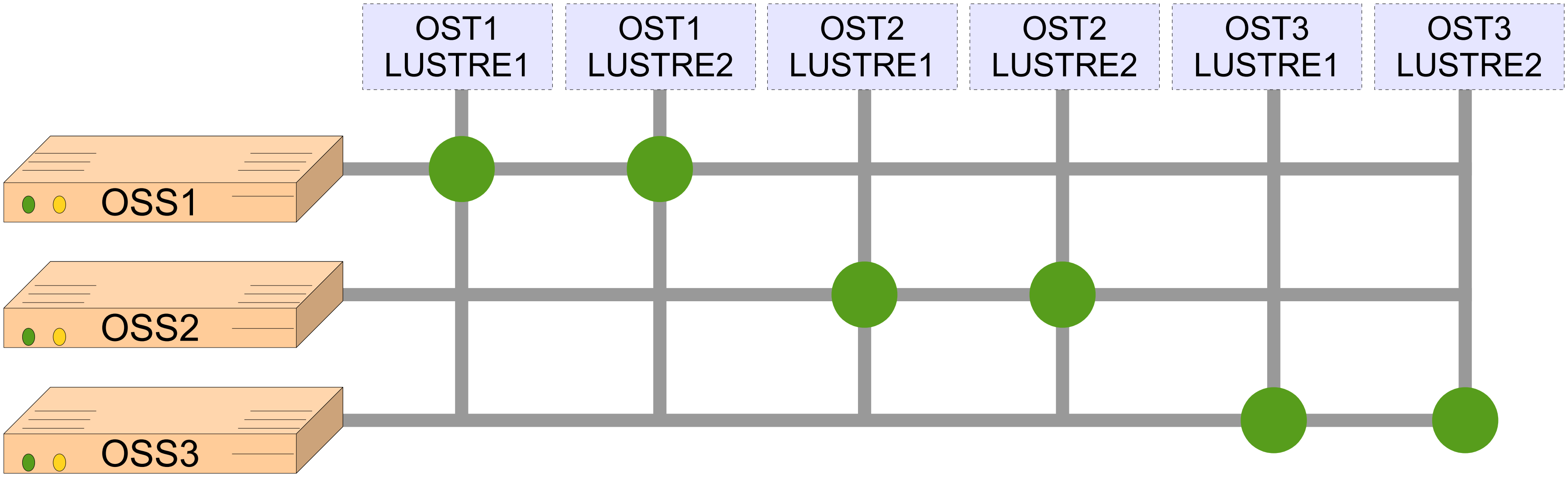


- Weights distribution, scoring mechanism
  - Each OST has a score with respect to each OSS
  - A OSS mounts an OST when that OST has the highest score on that OSS

*both the links!

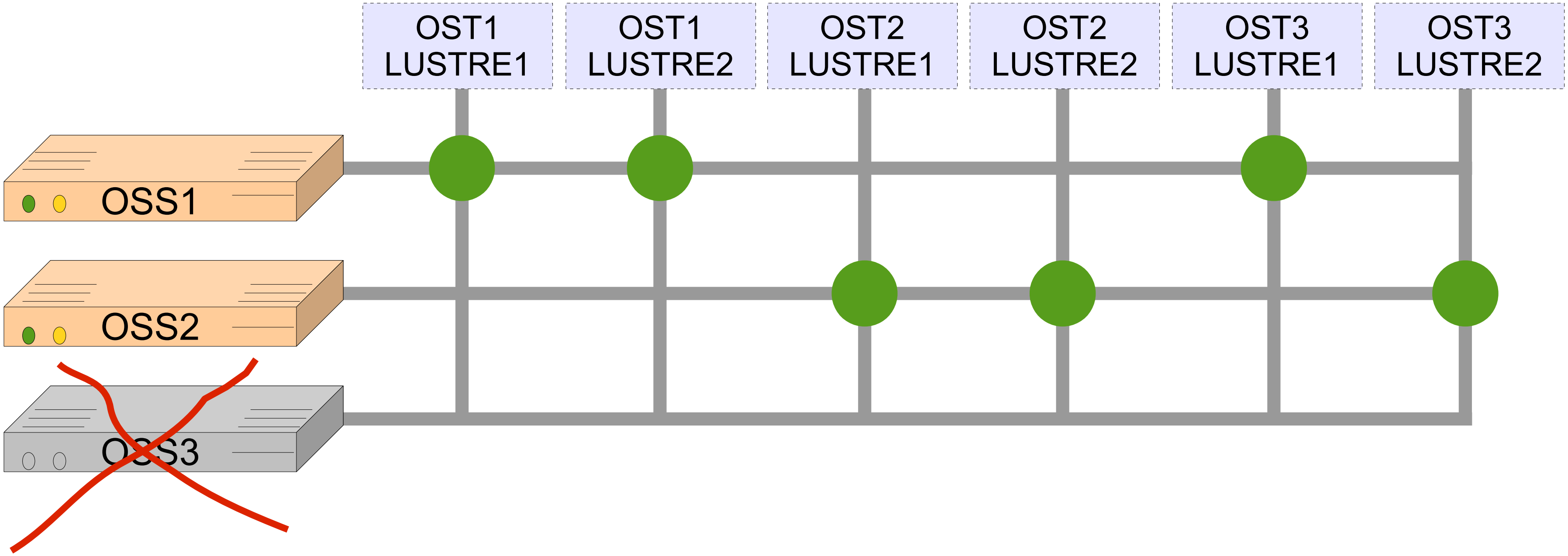
High availability on OSSs

	OST1 LUSTRE1	OST1 LUSTRE2	OST2 LUSTRE1	OST2 LUSTRE2	OST3 LUSTRE1	OST3 LUSTRE2
OSS1	1000	1000	600	800	800	600
OSS2	800	600	1000	1000	600	800
OSS3	600	800	800	600	1000	1000

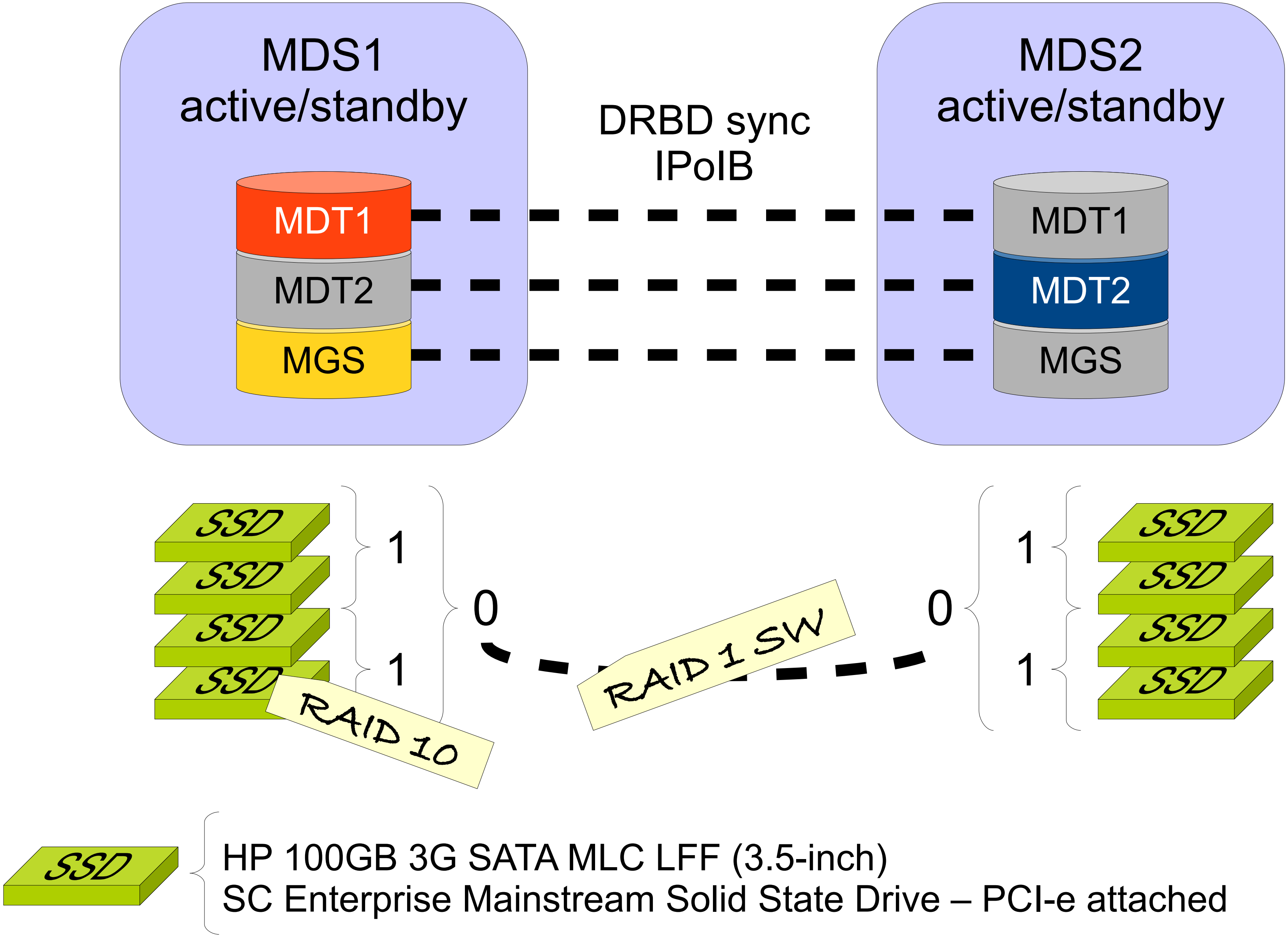


High availability on OSSs

- OSS3 fails
    - OST3LUSTRE1, OST3LUSTRE2 → -INF score
- OSS2, OSS1 receive a new OST



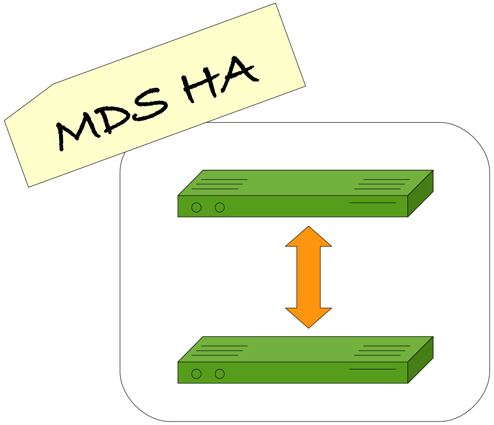
Metadata target

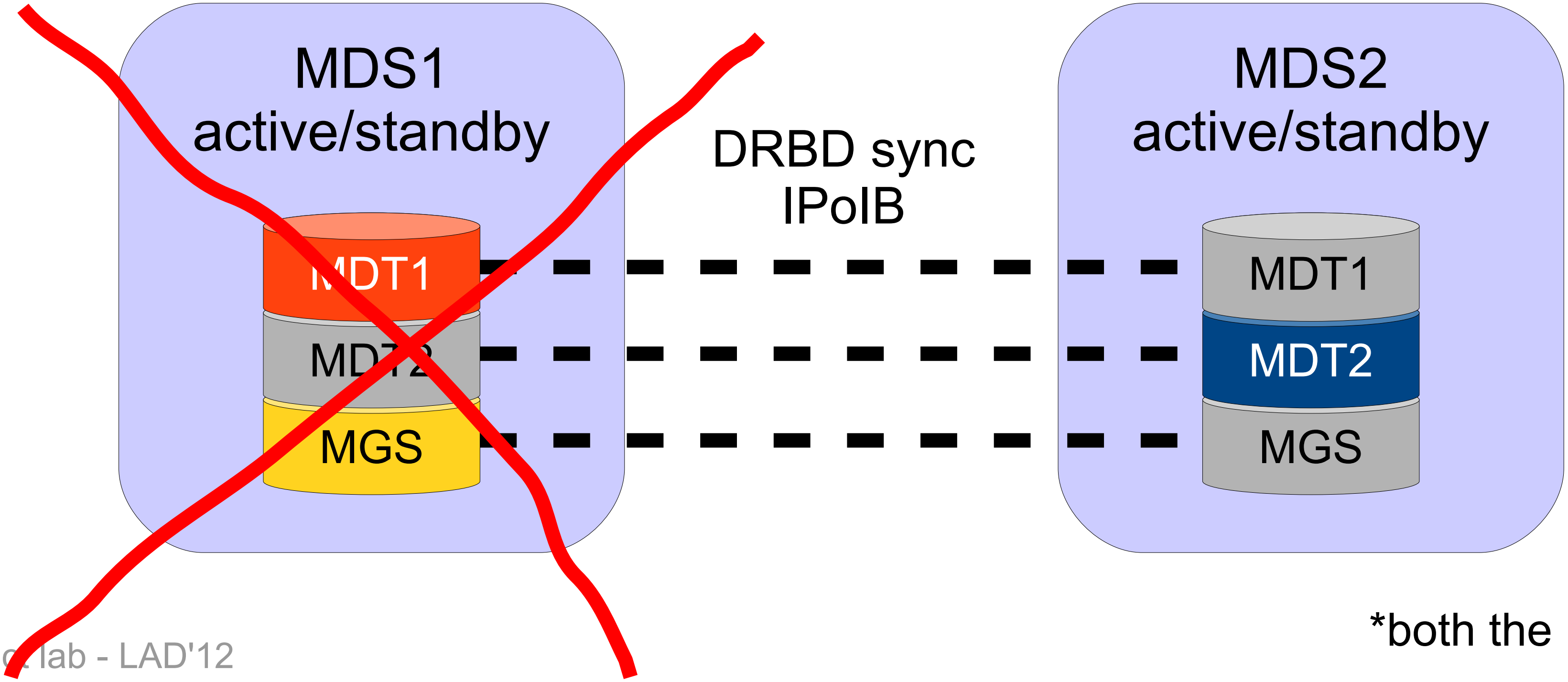


HP 100GB 3G SATA MLC LFF (3.5-inch)
SC Enterprise Mainstream Solid State Drive – PCI-e attached

High availability on MDSs

- Failures
  - Power*
  - InfiniBand*





MDS1
active/standby

MDT1

MDT2

MGS

DRBD sync
IPoIB

MDS2
active/standby

MDT1

MDT2

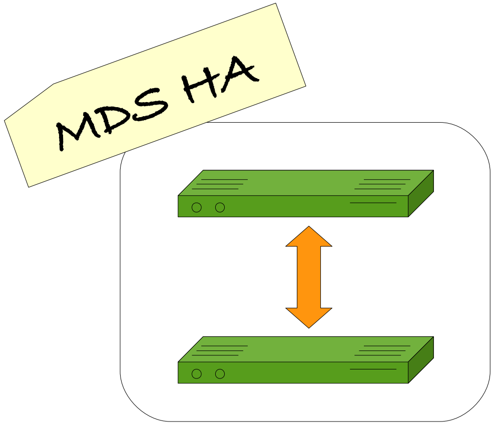
MGS

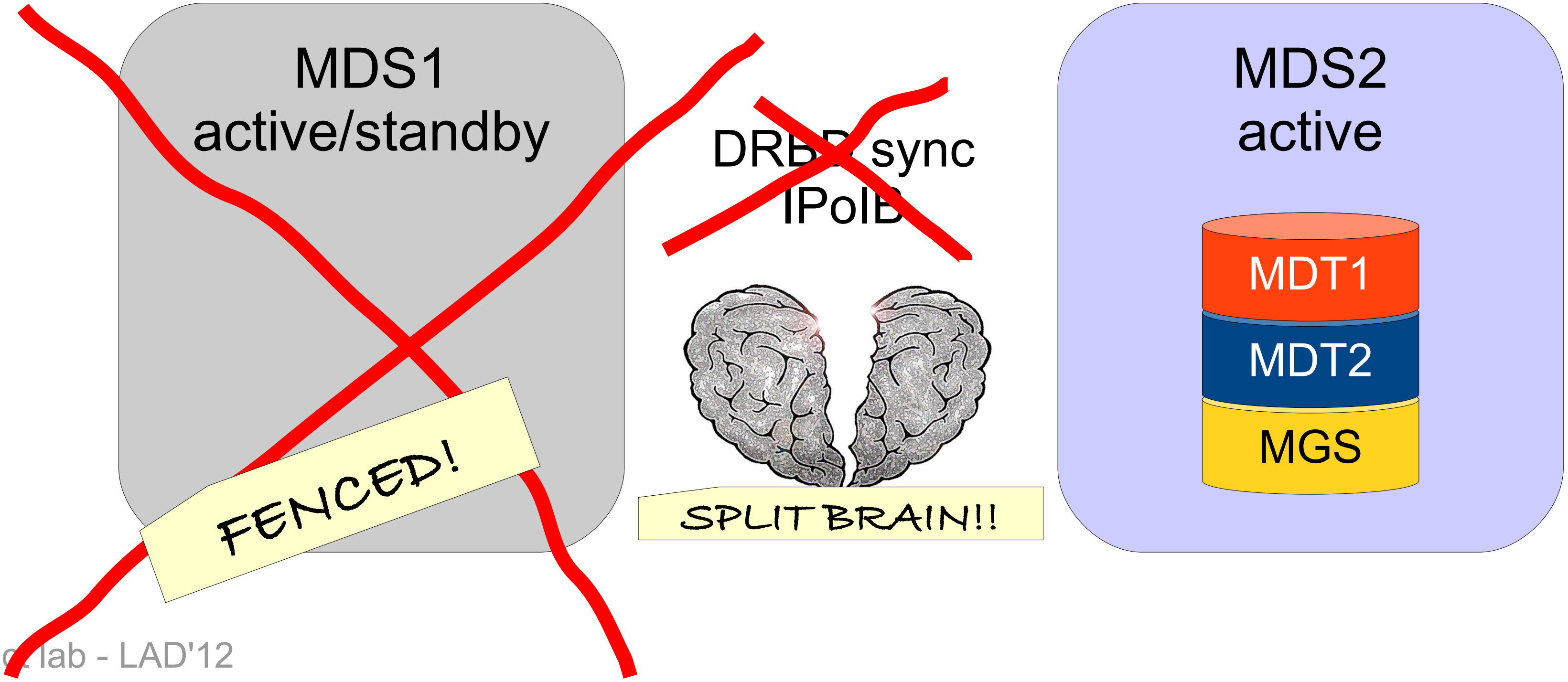
*both the links!

Metadata integrity

High availability on MDSs

- STONITH!
  - MDS1 failovers
  - MDS2 takeovers




MDS1
active/standby

FENCED!

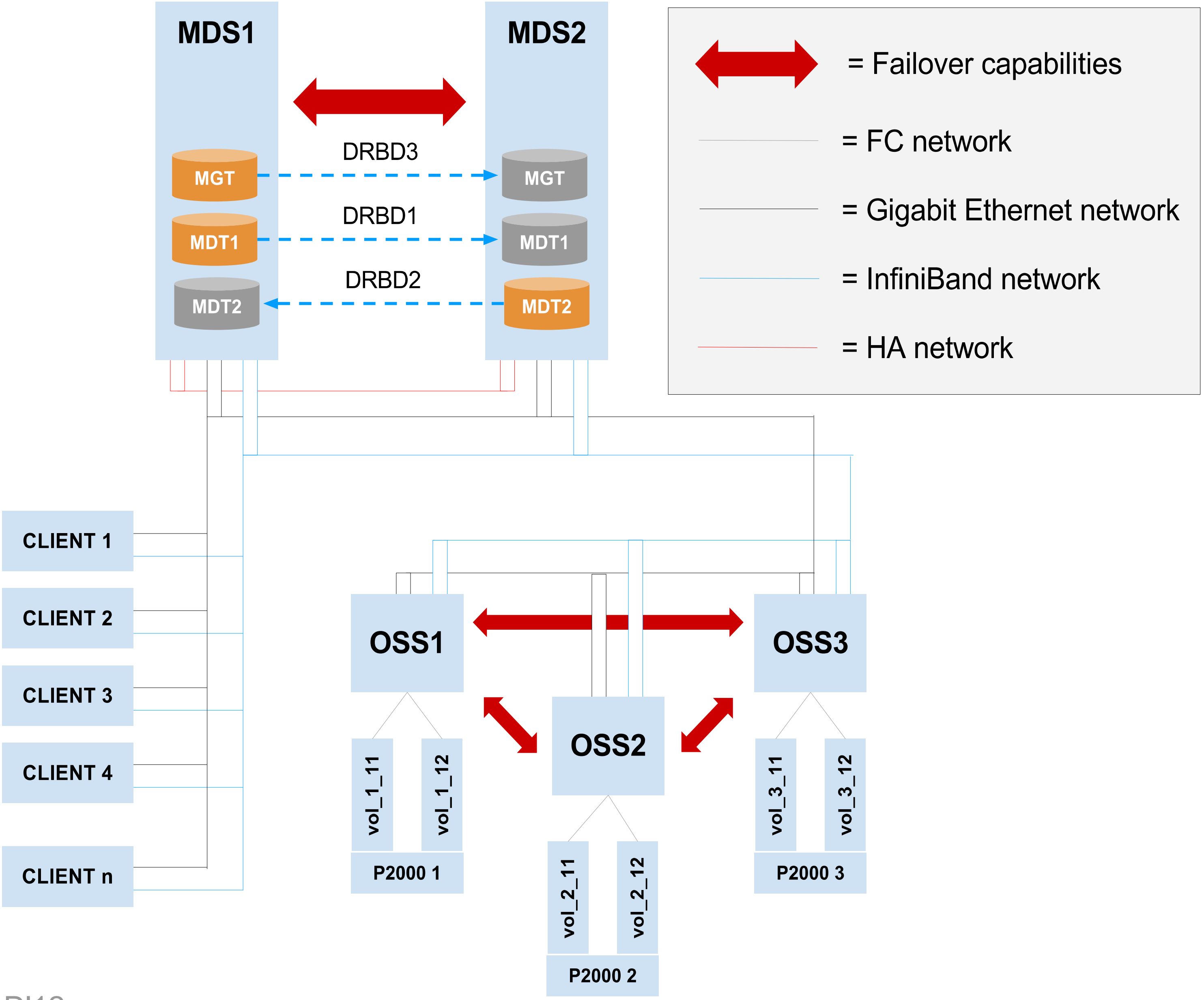
DRBD sync
IPoIB

SPLIT BRAIN!!

MDS2
active

MDT1
MDT2
MGS

Lustre network



Lustre HA: SW stack

- CentOS 6.2
- Lustre 2.1.1
    - Lustre-kernel RPM on I/O server
    - Lustre patchless client RPM modules on clients
    - Lustre iokit
    - Shine
- Pacemaker/corosync

High availability tests

- Unplug → failover
  - Power
  - InfiniBand
  - Fibre Channel (on OSS)
  - InfiniBand + Fibre Channel
- Replug → failback
  - Automatic on OSSs
  - Manual on MDSs
    - No automatic split-brain resolution!

DOWNTIME = ~120s

Performance on OSTs

	XDD 1 thread		Sgpdd-survey 16 threads	
	READ	WRITE	READ	WRITE
2 TB SINGLE DISK	150	150	/	/
RAID6 (7 DISKS)	400	400	330	590

Results in MB/s

Performance on MDTs

	FILE CREATION		OPERATION ON DIRECTORIES	
	16 threads	64 threads	16 threads	64 threads
MDT on SSD	1000	2600	60000	68000
MDT on HD	800	1200	30000	30000

Results in operations per second

What's next

- Upgrade to 2.1.3
  - (almost) no downtime thanks to HA
- Monitoring the HA software stack
  - DRBD on MDTs

eXact

www.exact-lab.it

info@exact-lab.it