High Availabilty Lustre FS implementation for Genomic

Data day

ICTP, Trieste

September 5, 2013

Francesco De Giorgi

what is eXact lab

***eX**perience on **a**dvanced **c**omputational **t**echnologies*

a private company led by IT experts with a strong background in physics and computer science, provides ***solutions*** in the HPC market

- services
  - cluster deployment
  - storage solution
- training
  - sys admin and user oriented programs

how HPC meets big data

- HPDA: *High Performance Data Analysis*
  - tasks involving sufficient data volumes and algorithm complexity to require HPC resources
- Use cases
  - climate modeling
  - risk analysis
  - national security
  - life science

use case: DNA sequencing

- In 2006 the XPRIZE Foundation offered $10 million to the first team that

  "... could sequence 100 whole human genomes at a cost of $10000 or less per genome, in 30 days or less ..."

- from September 5, 2013 to October 5

  ➔ **HURRY UP!!**

use case: DNA sequencing

- In 2006 the XPRIZE Foundation offered $10 million to the first team that

  “... could sequence 100 whole human genomes at a cost of $10000 or less per genome, in 30 days or less ...”

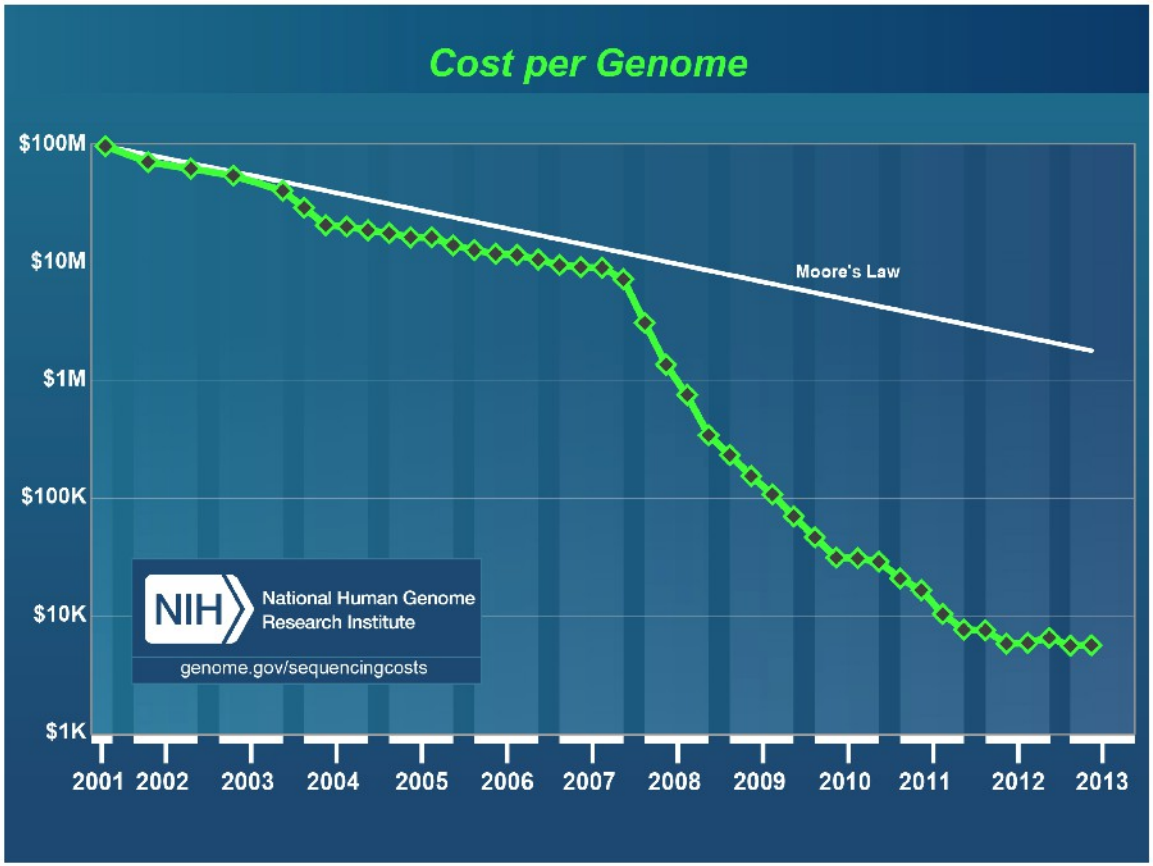
- ~~from September 5, 2013 to October 5~~

➔ XPRIZE canceled the price on August 22

"... genome sequencing technology is plummeting in cost and increasing in speed independent of our competition. Today, companies can do this for less than $5,000 per genome, in a few days or less ..."

DNA sequencing

- High-throughput DNA sequencing

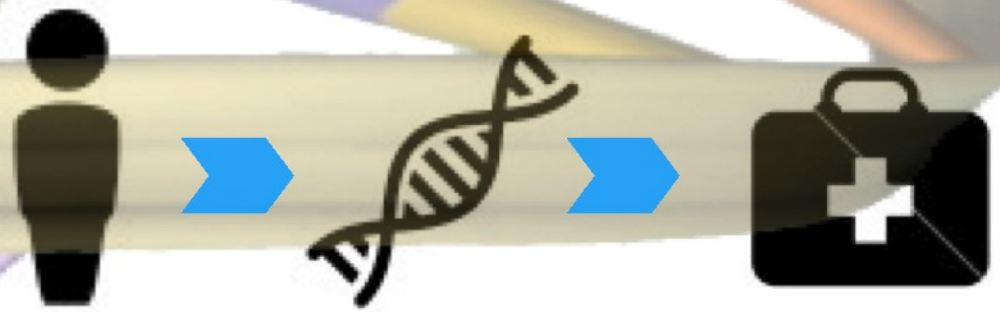


tremendous amount of data that need to be processed

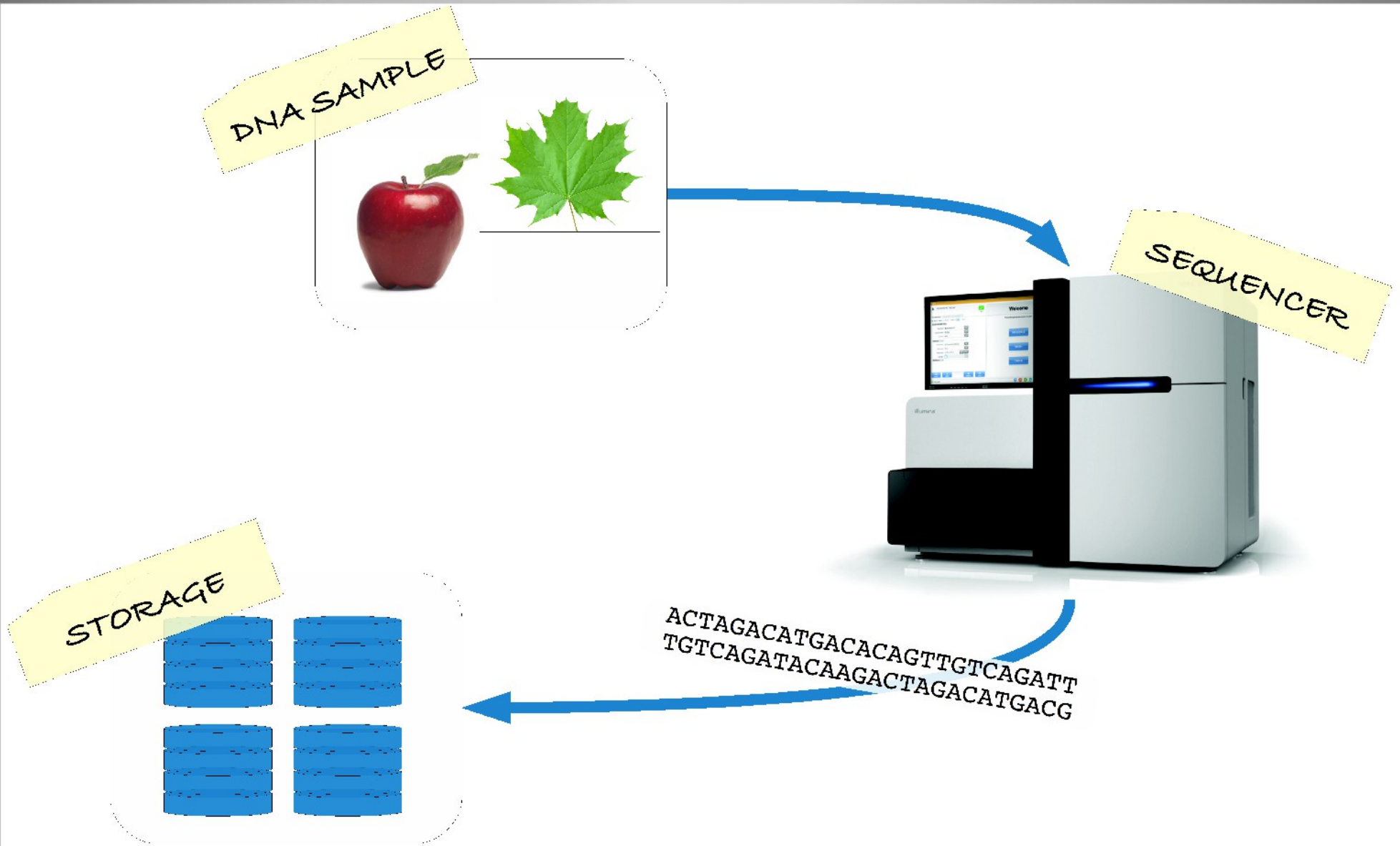
Next Generation Sequencing is a big data problem!

eXact lab case study

- HPC services in a primary research institute
  - medical research
- Translational Genomic and Bioinformatics
  - personalized medicine: customization of healthcare by use of genetic information



Customer needs analysis



Customer needs analysis

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

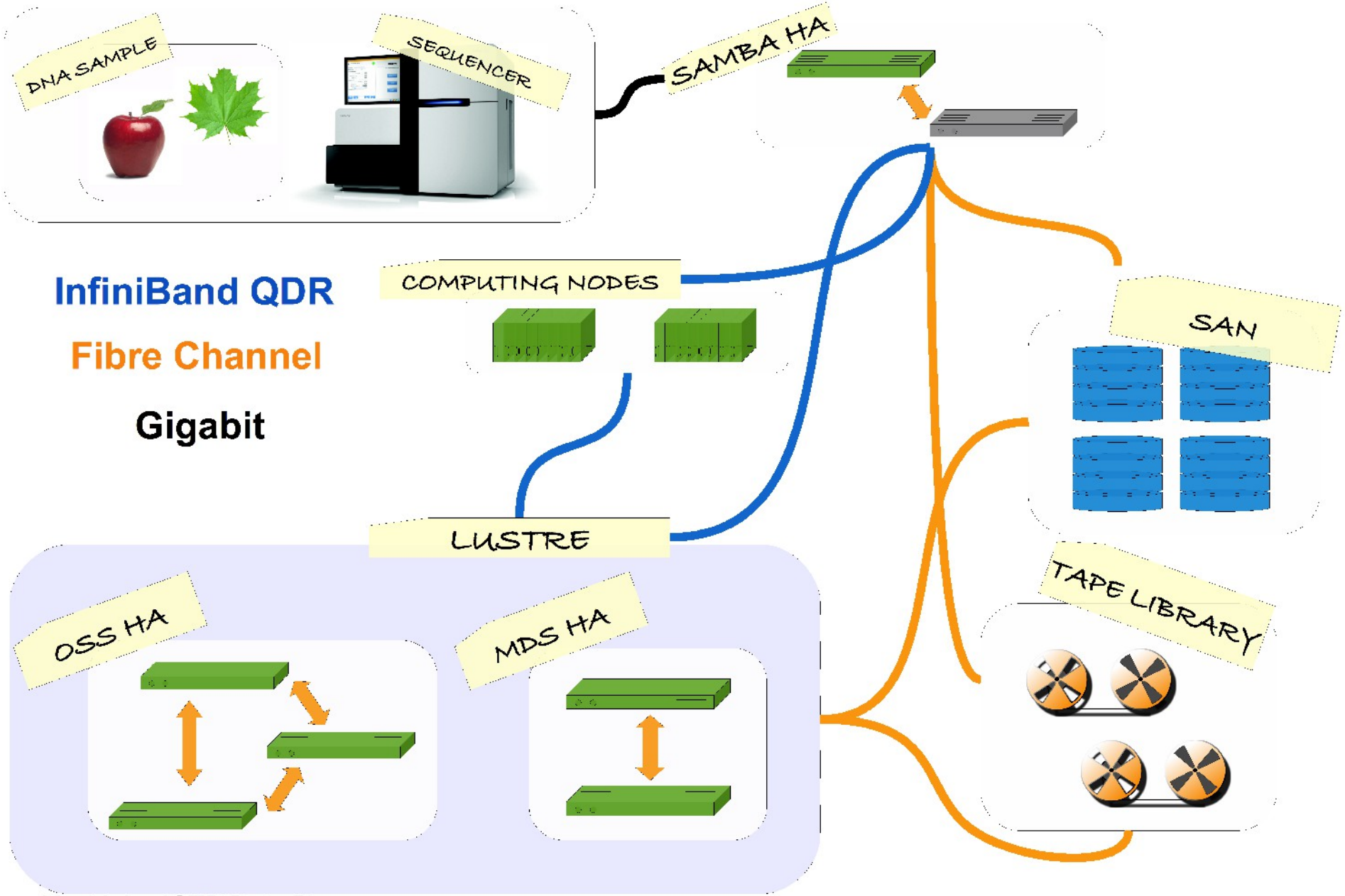
➔ Data from the sequencer need to be served to the computational infrastructure

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

Customer needs analysis

- Need for a **fast, high performance, highly scalable file system**, with robust failover and recovery mechanisms
- **Lustre File System**
  - parallel and distributed
  - high availability features

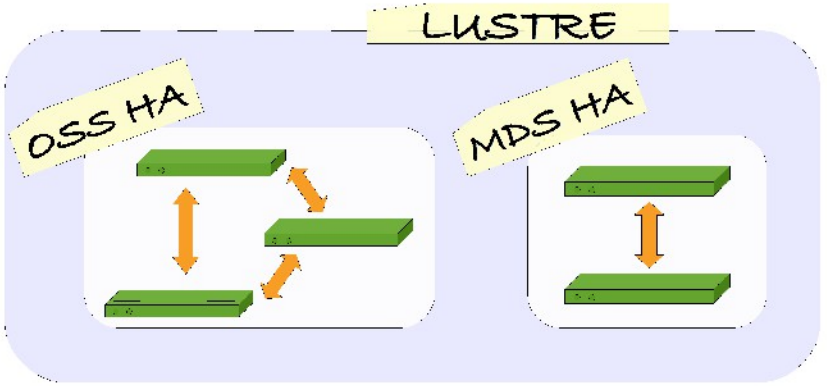
Infrastructure



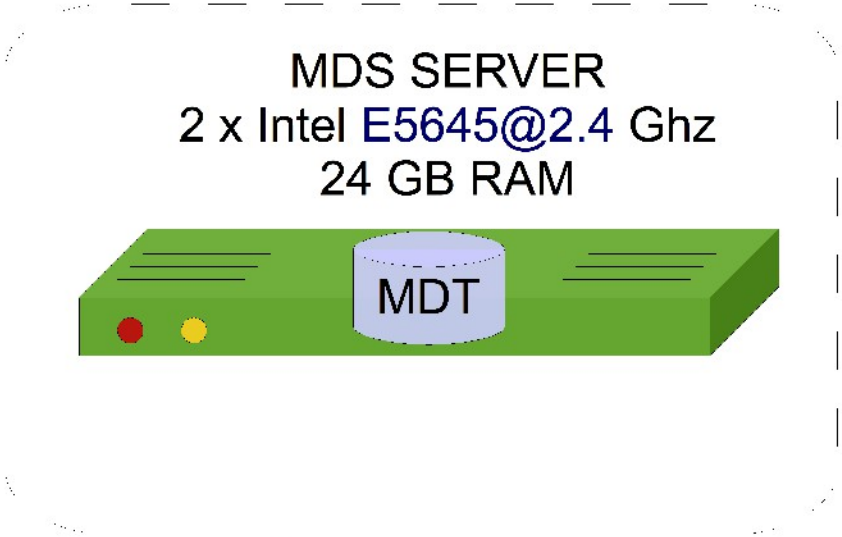
Lustre filesystem

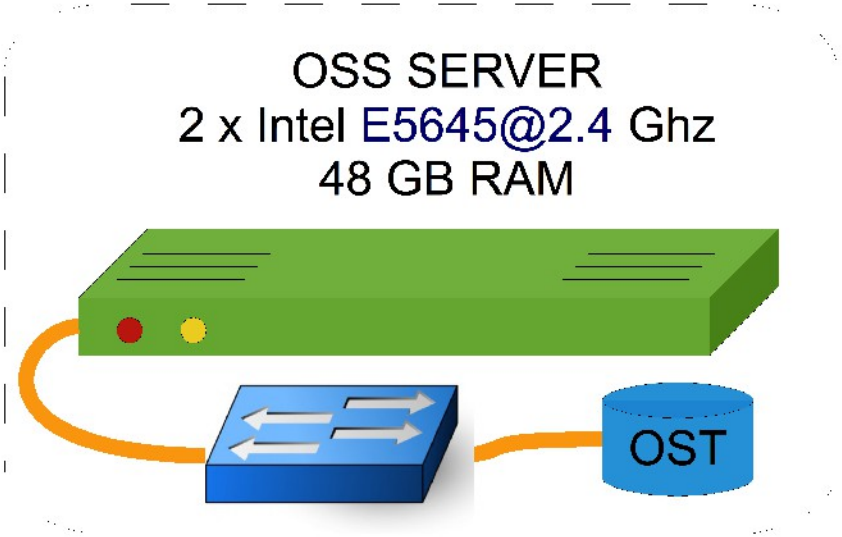
2 Lustre filesystems

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

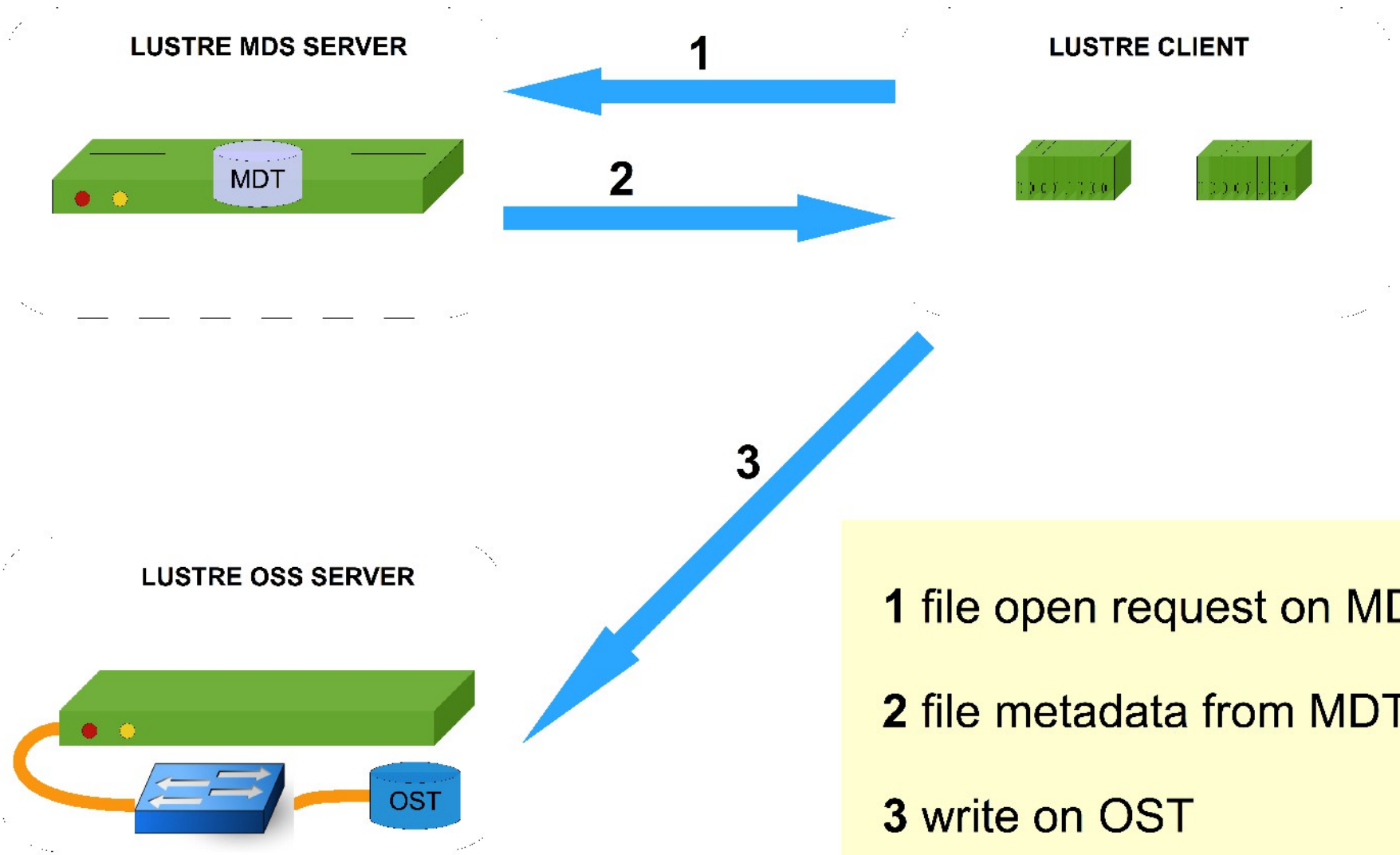


always available!





How I/O works in Lustre

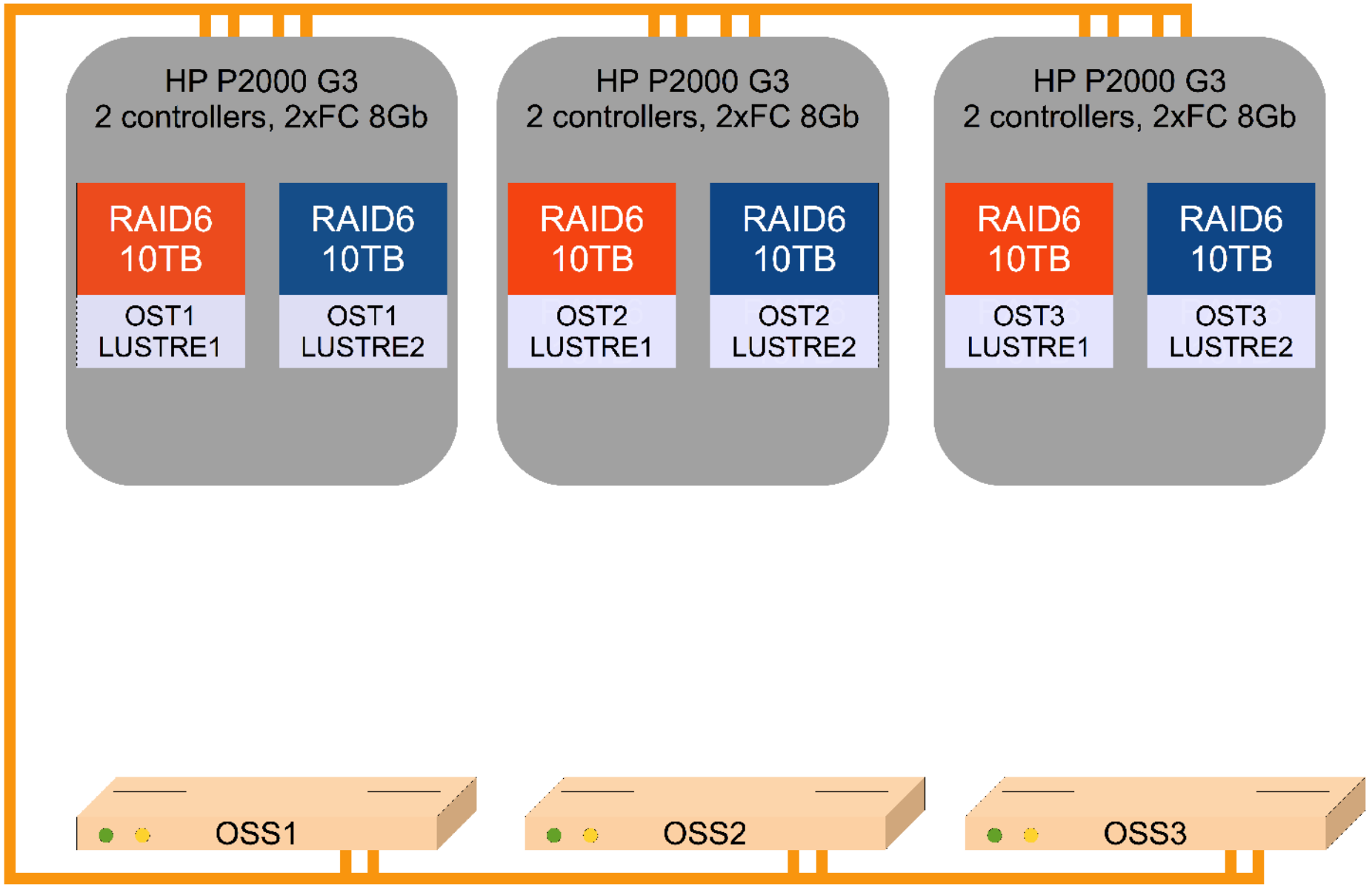


1 file open request on MDT

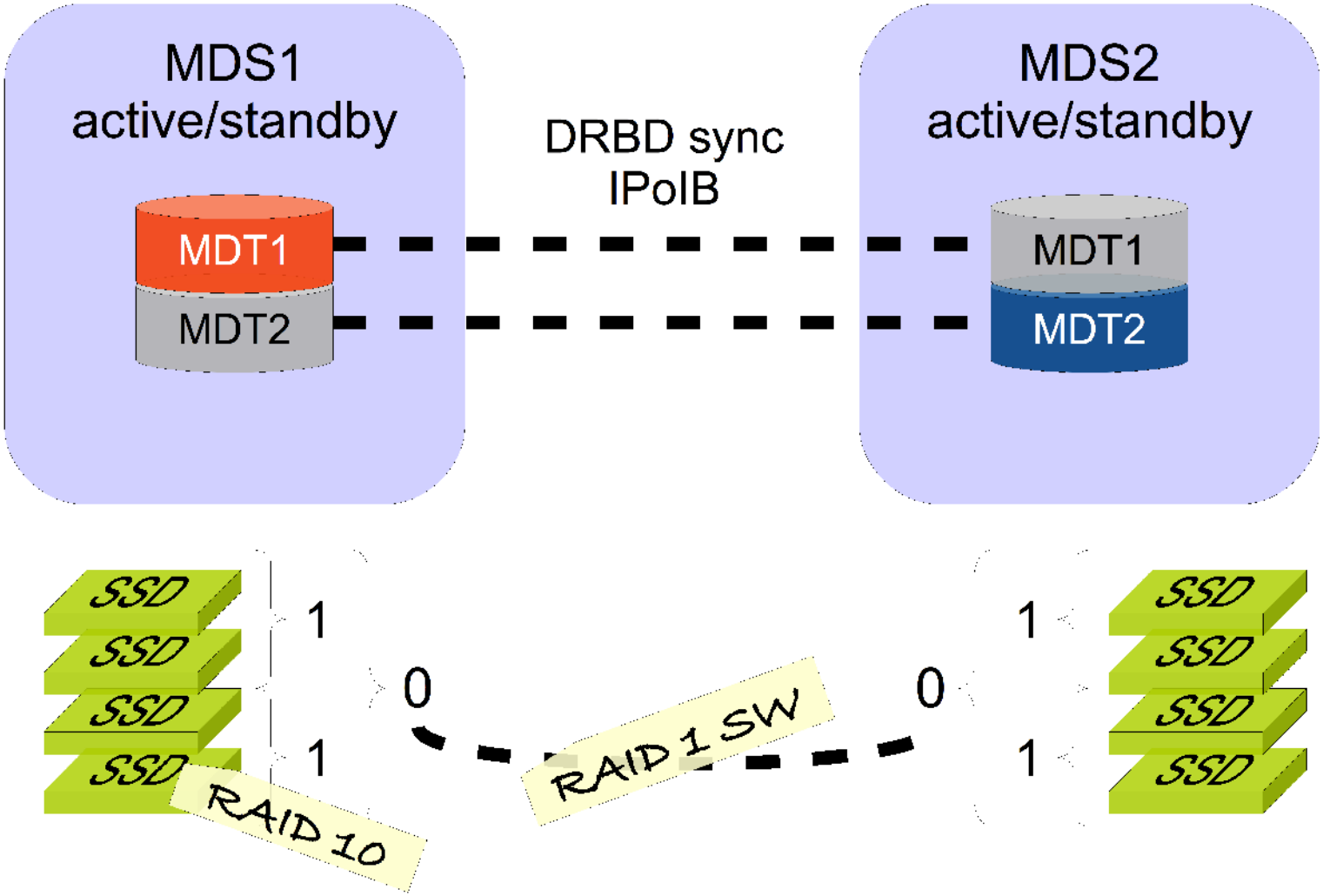
2 file metadata from MDT

3 write on OST

Object Storage Targets on SAN



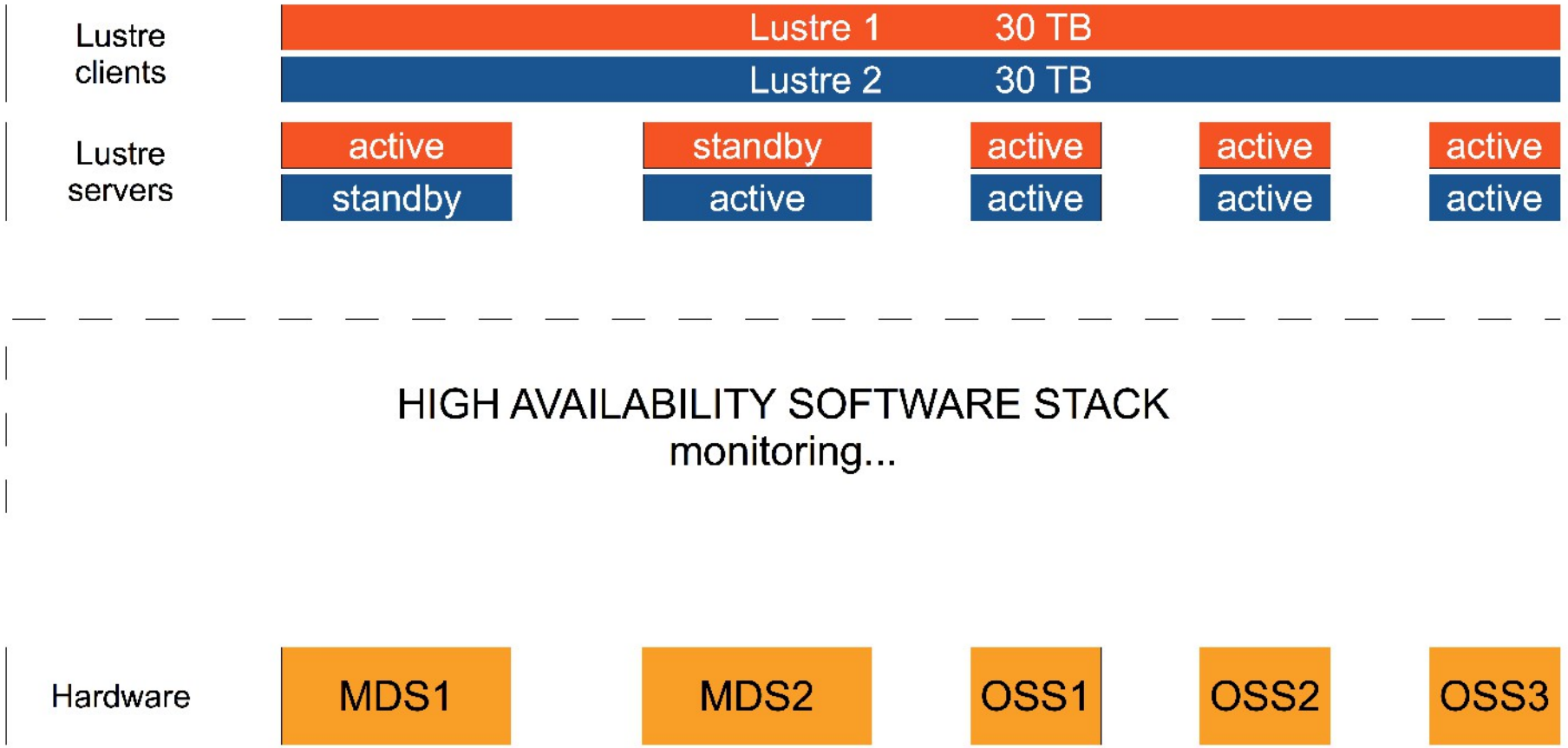
MetaData Targets: locally on MDSs



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

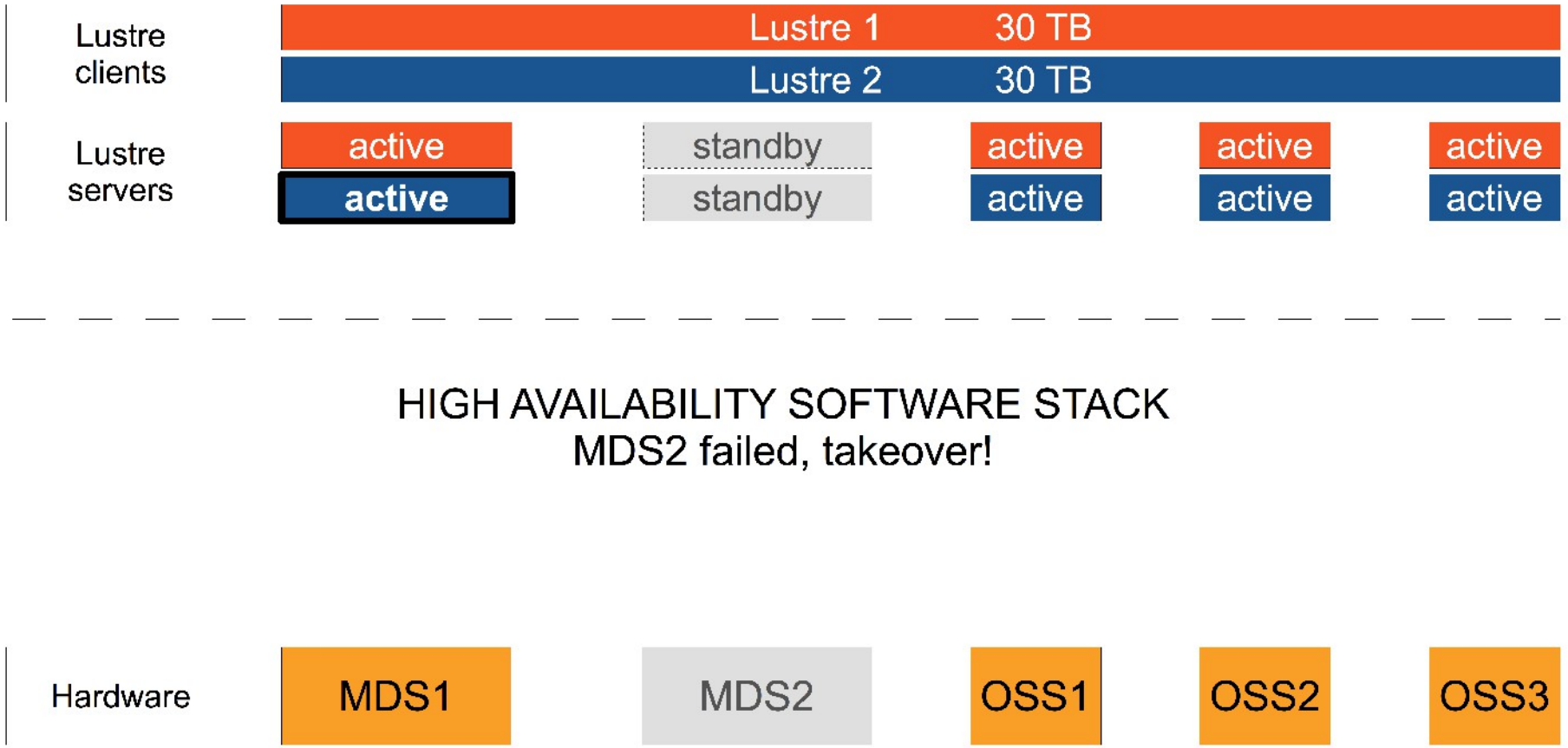
Lustre high availability

- In production, no failures



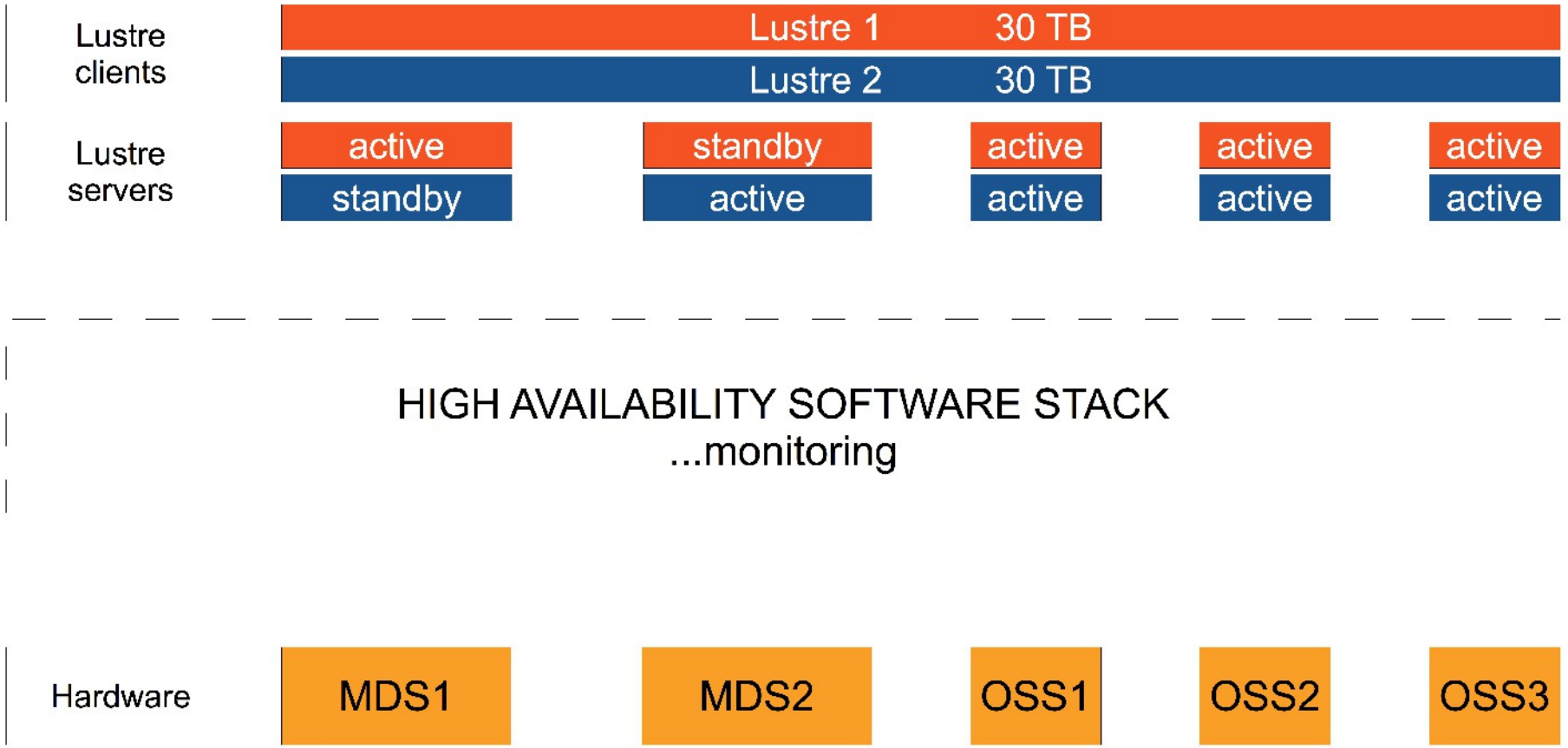
MDS2 failure

- MDS1 will take over the service of MDS2



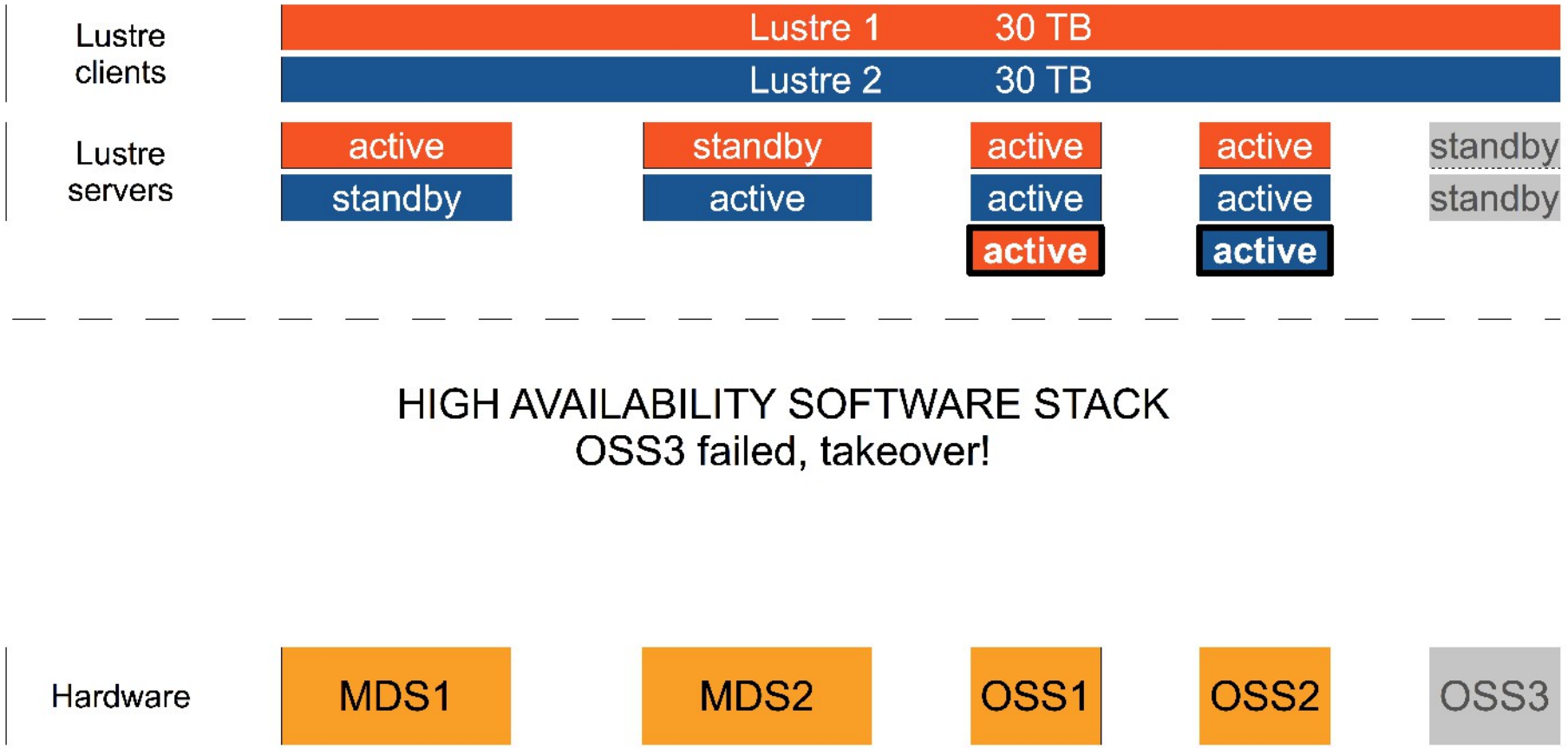
Lustre high availability

- In production, no failures



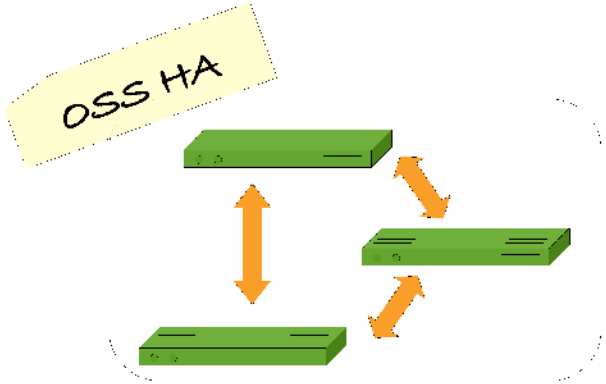
OSS3 failure

- OSS1 and OSS2 will take over its service

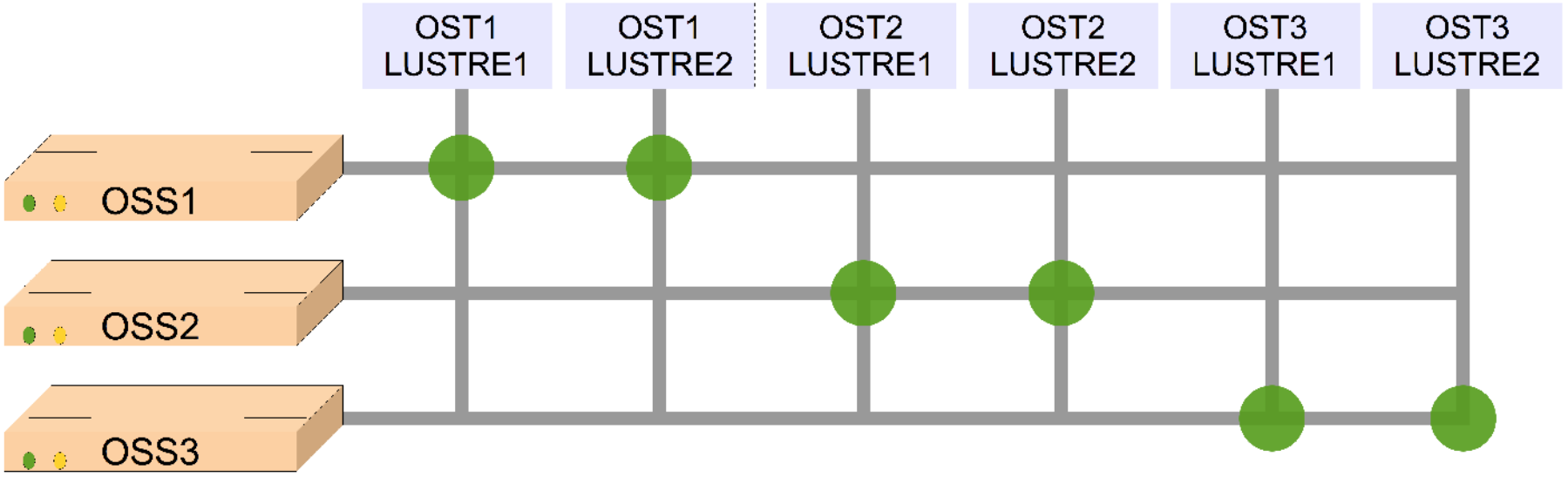


High availability on OSSs

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



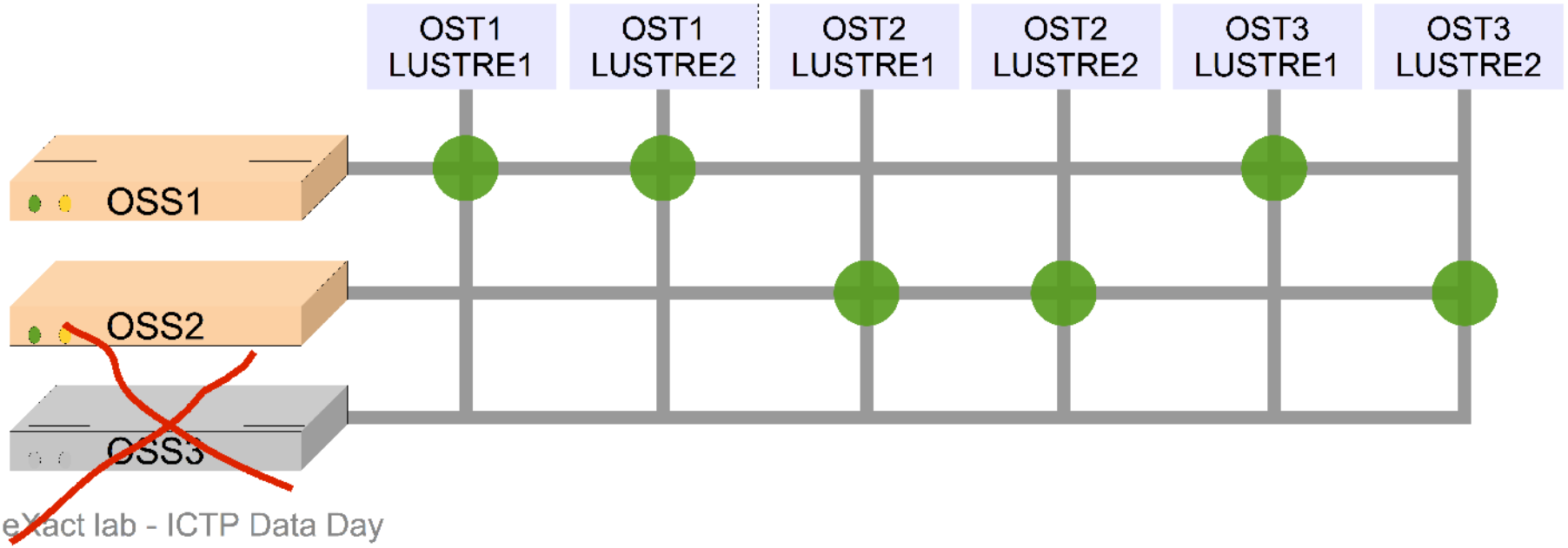
- In production, each OSS server mounts 2 OSTs



*both the links!

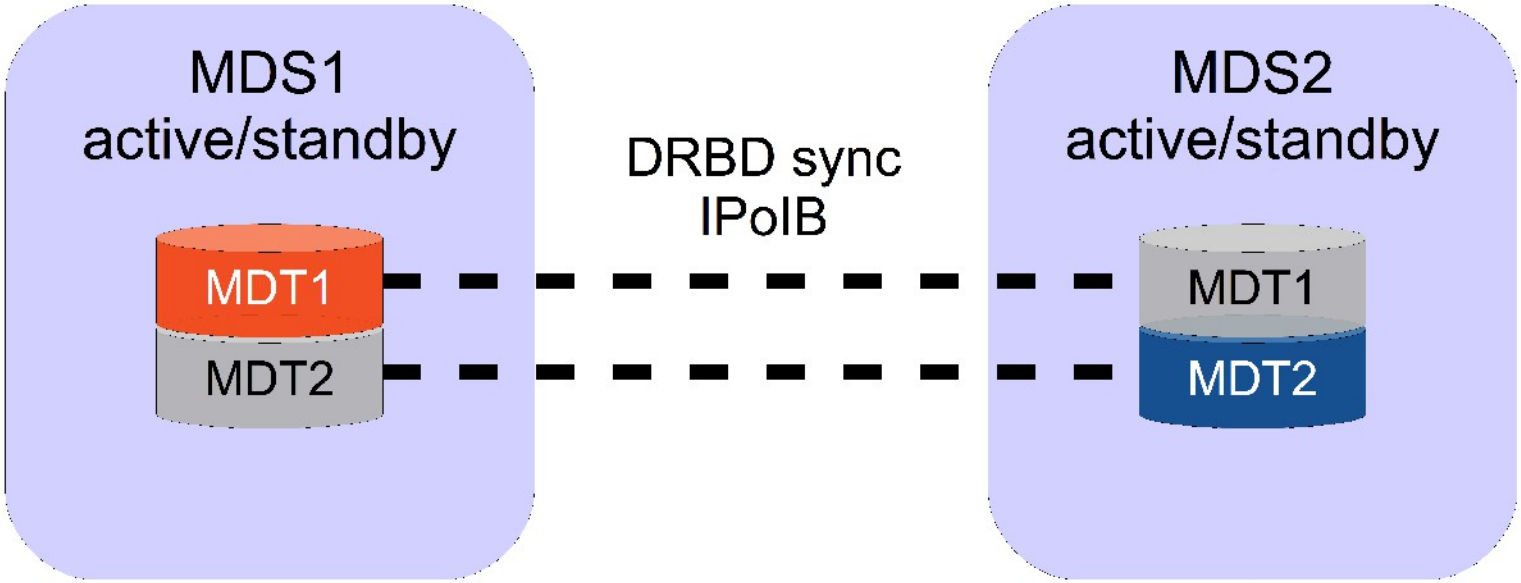
High availability on OSSs

- If OSS3 fails
  - the HA software will acknowledge the failure
- OSS2, OSS1 receive a new OST each



High availability on MDSs

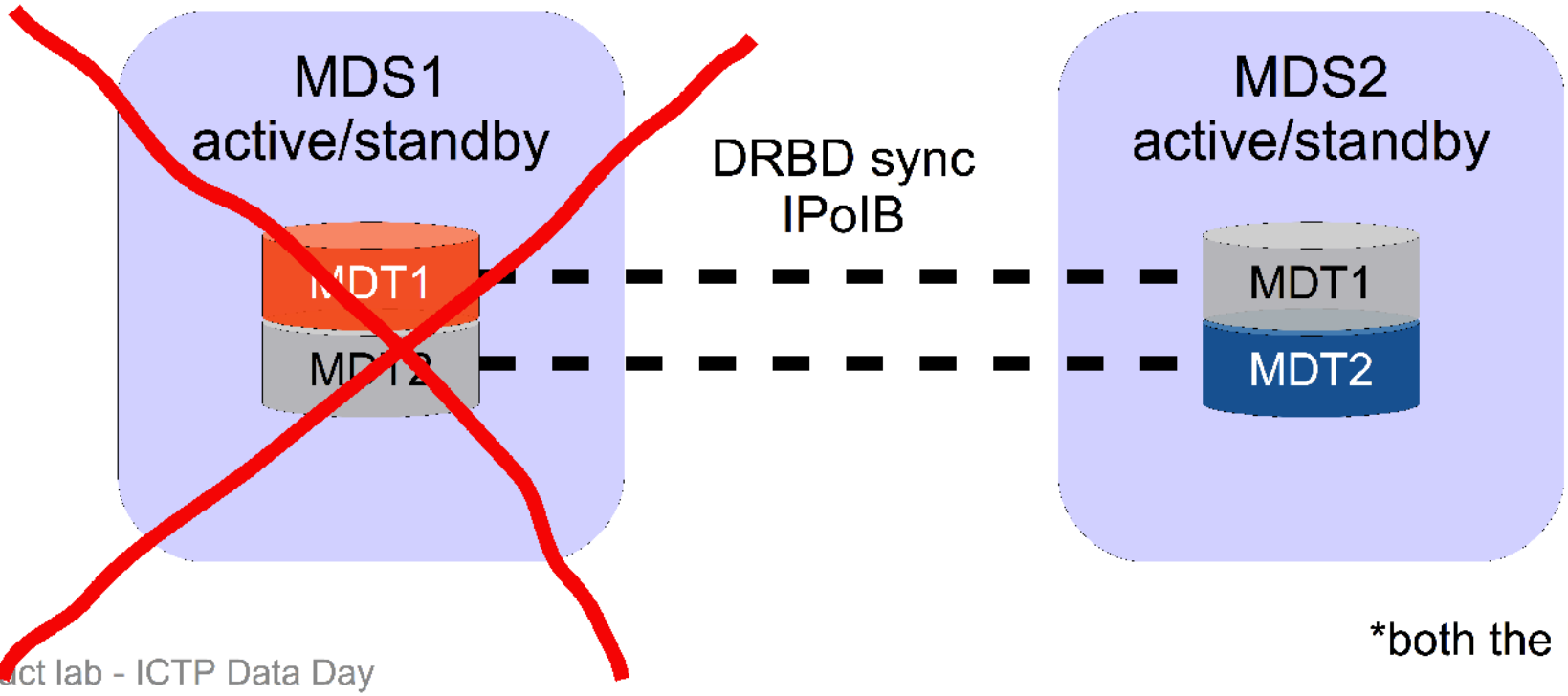
- In normal condition
  - MDTs are replicated between MDSs
  - only one replica is active for Lustre client



*both the links!

MDS1 failure

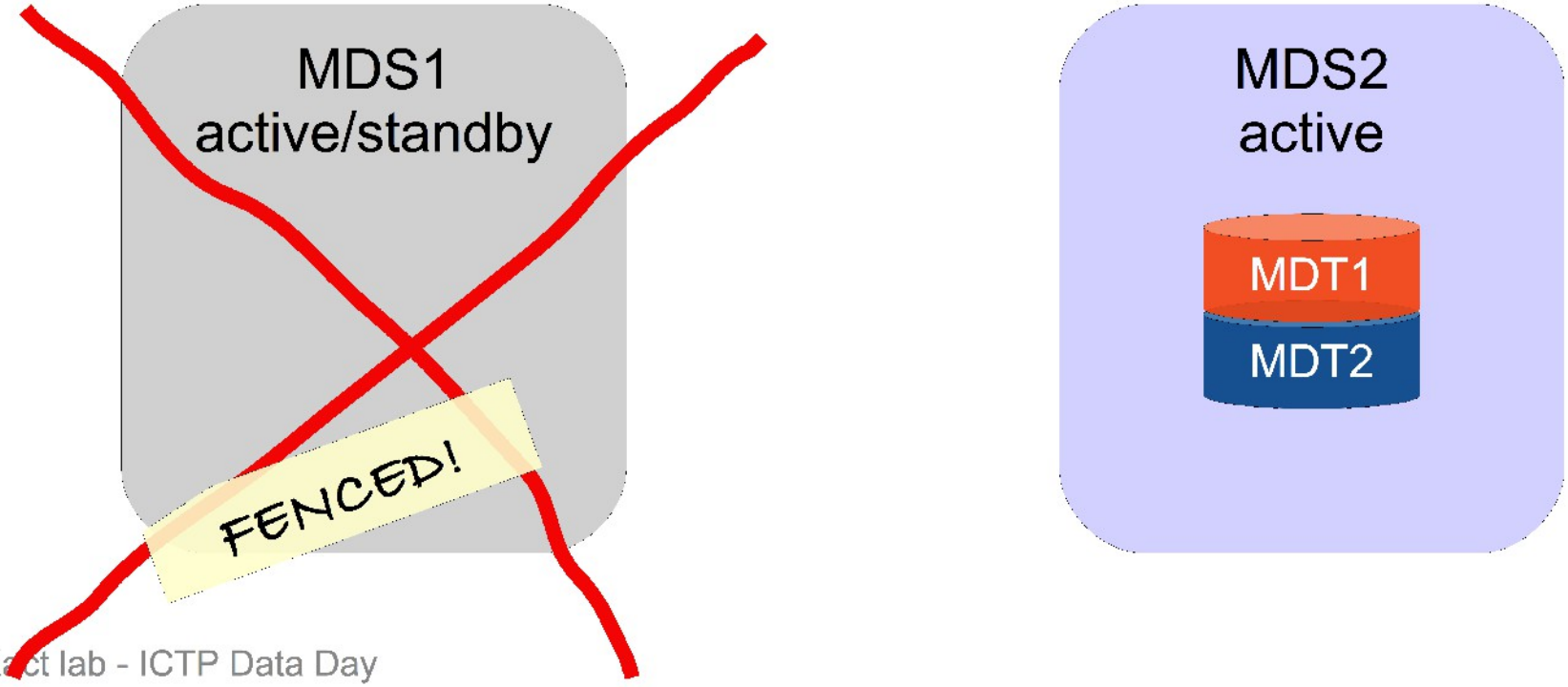
- data integrity **MUST** be ensured
- MDS1 *irresponsive* ≠ *isn't accessing my data*
- MDS1 must be powered off by MDS2



*both the links!

MDS1 failure

- STONITH aka ***S**hoot **T**he **O**ther **N**ode **I**n **T**he **H**ead!*
  - MDS1 fails
  - MDS2 takes over its services
    - MDS2 forces MDS1 to power off



The way to ensure MDT integrity

High availability tests

- Unplug → *failover*
  - Power*
  - InfiniBand*
  - Fibre Channel (OSS)*
  - InfiniBand + Fibre Channel
- Replug → *failback*

➔ Completely transparent for clients!

High availability tests

- Unplug → *failover*
  - Power*
  - InfiniBand*
  - Fibre Channel (OSS)*
  - InfiniBand + Fibre Channel
- Replug → *failback*

➔ Completely transparent for customers!

High availability tests

- Unplug → *failover*
  - Power*
  - InfiniBand*
  - Fibre Channel (OSS)*
  - InfiniBand + Fibre Channel
- Replug → *failback*

➔ Completely transparent for doctors!

High availability tests

- Unplug → *failover*
  - Power*
  - InfiniBand*
  - Fibre Channel (OSS)*
  - InfiniBand + Fibre Channel
- Replug → *failback*

➔ Completely transparent for patients!

Q&A

High Availabilty Lustre
FS implementation
for Genomic

info@exact-lab.it www.exact-lab.it