The Abdus Salam International Centre for Theoretical Physics

2494-24

Workshop on High Performance Computing (HPC) Architecture and Applications in the ICTP

14 - 25 October 2013

High availability Lustre FS implementation for genomic

Francesco De Giorgi
eXact Lab, Trieste

High availability Lustre FS
implementation for genomic

Francesco De Giorgi
francesco.degiorgi@exact-lab.it

ICTP, Trieste
October 24, 2013

eXact lab overview

eXperience on advanced computational technologies

- spin-off of the

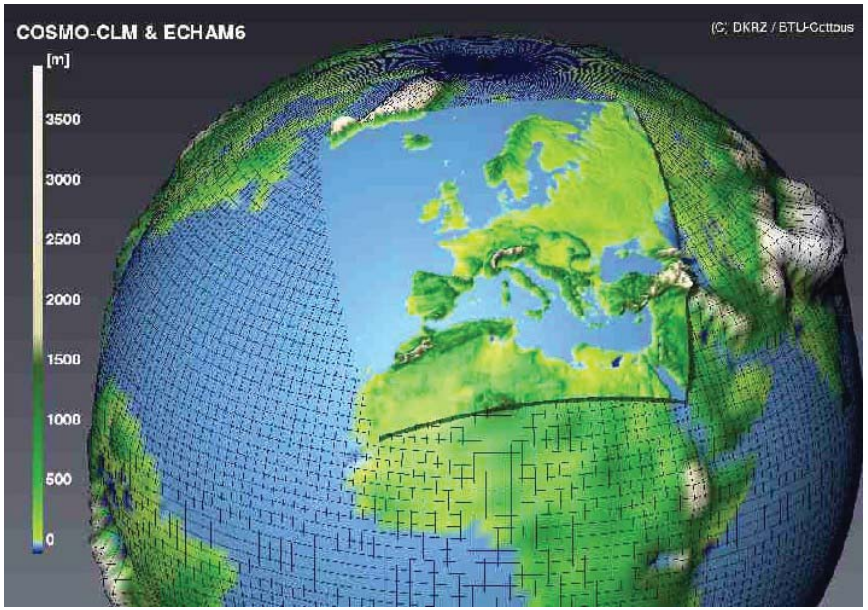
National Research Council of Italy

iom Istituto Officina dei Materiali

- leveraging academic expertise, the team provides solutions to the HPC market:
  - design and deployment of computational infrastructures
  - advanced training in HPC
  - advanced services for scientific/technical application

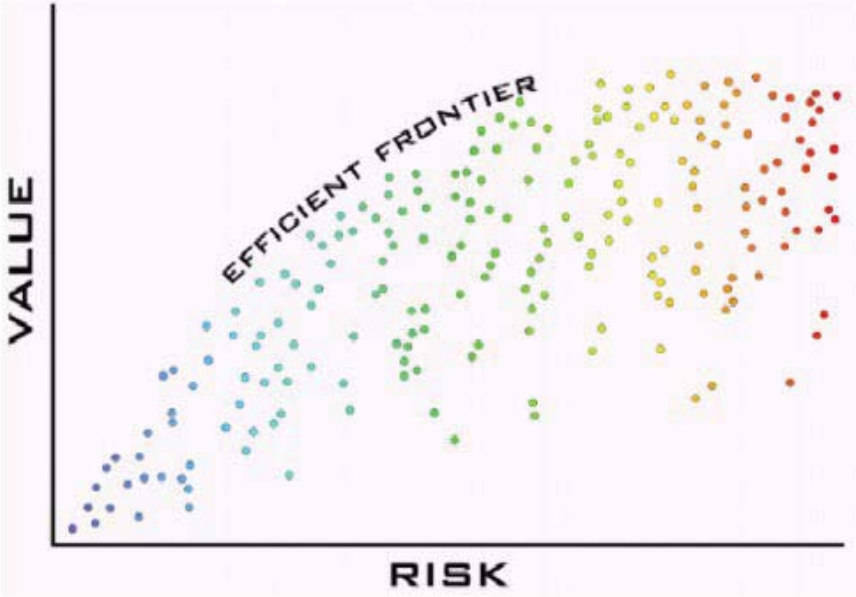
how HPC meets big data

- 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



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



how HPC meets big data

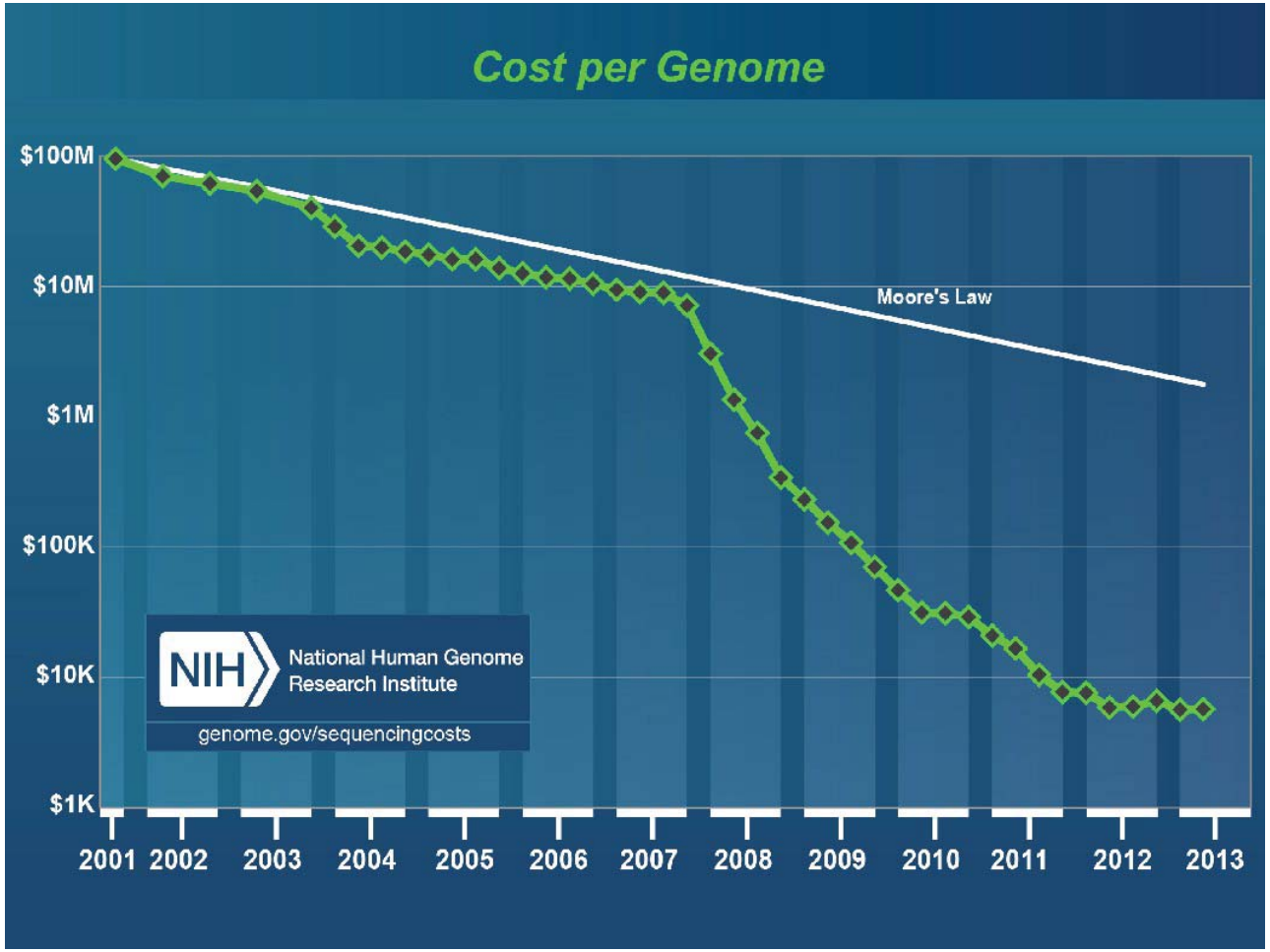
- HPDA: *High Performance Data Analysis*
  - tasks involving sufficient data volumes and algorithm complexity to require HPC resources
- Use cases
  - climate modeling
  - risk analysis
  - national security
  - life science

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

Life science: DNA sequencing

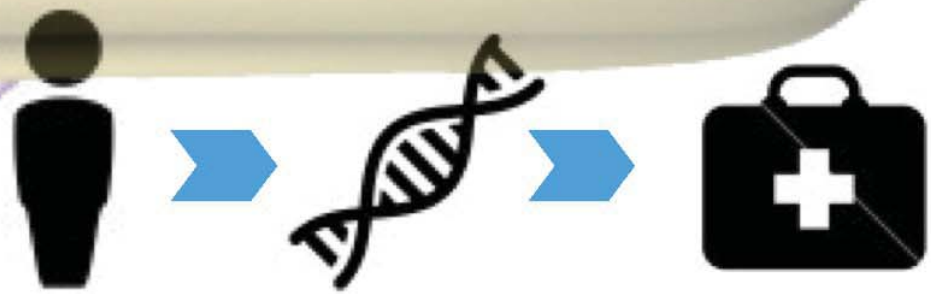
- High-throughput sequencing: a big data problem



DNA sequencing in HPC: a case study

eXact lab's deployment of a highly available and fault-tolerant HPC infrastructure for genomic research

- For the Center for Translational Genomic and Bioinformatics
  - in San Raffaele Hospital (Milan, Italy)
- Final goal is personalized medicine
  - *customization of healthcare by use of genetic information*



Customer's needs analysis

- **Huge amount of genomic data** from sequencer 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

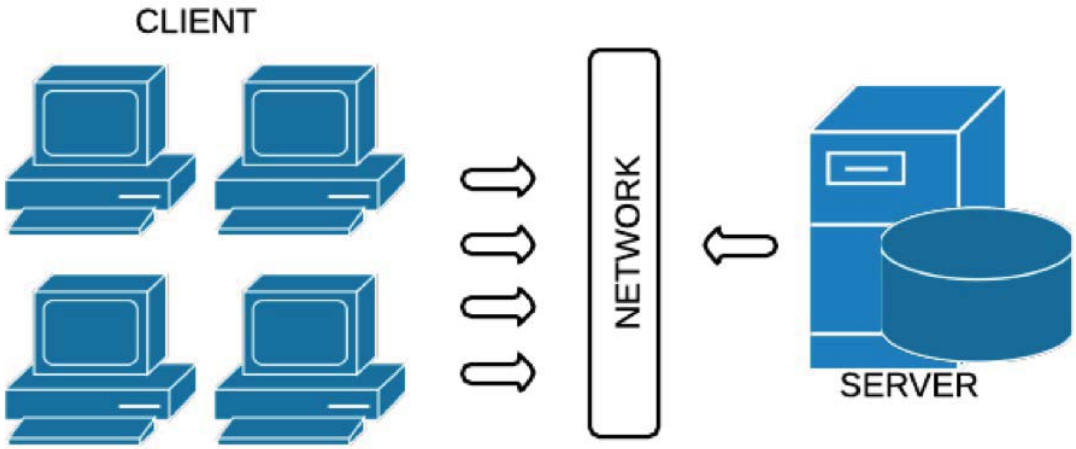
Customer's needs analysis

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

- **Lustre File System**
  - parallel and distributed
  - high throughput, low latency
  - scalable, redundant
  - high availability features

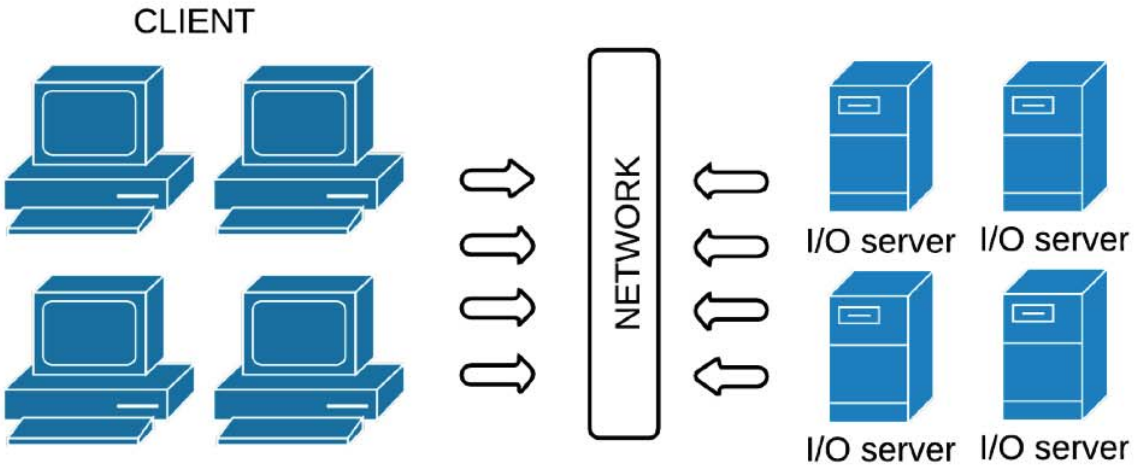
short intro to Lustre

- why a parallel and distributed file system?
- isn't NFS enough to address the requirements?
  - server is a single point of failure (unless data replication)
  - not scalable
  - limited by network



short intro to Lustre

- a parallel FS distributes data on many servers
  - parallel I/O
  - low access latency
  - high throughput
  - scalable



short intro to Lustre

- a parallel FS distributes data on many servers
  - **pros**
    - scalability
    - performance
    - data integrity (thanks to redundancy and failover mechanism)
  - **cons**
    - needs a metadata server
    - not so easy to configure and manage as infrastructure grows

Lustre file system

- part of Intel HPC strategy
  - road to exascale computing (by 2018)
  - 74% of CPUs in the TOP500 are Intel
- 6 on 10 in the TOP10 use Lustre
  - 60 in the TOP100
- used in heterogeneous environment
  - oil and gas, life science, meterology, finance, …
  - from a few Terabyte to tens of Petabyte

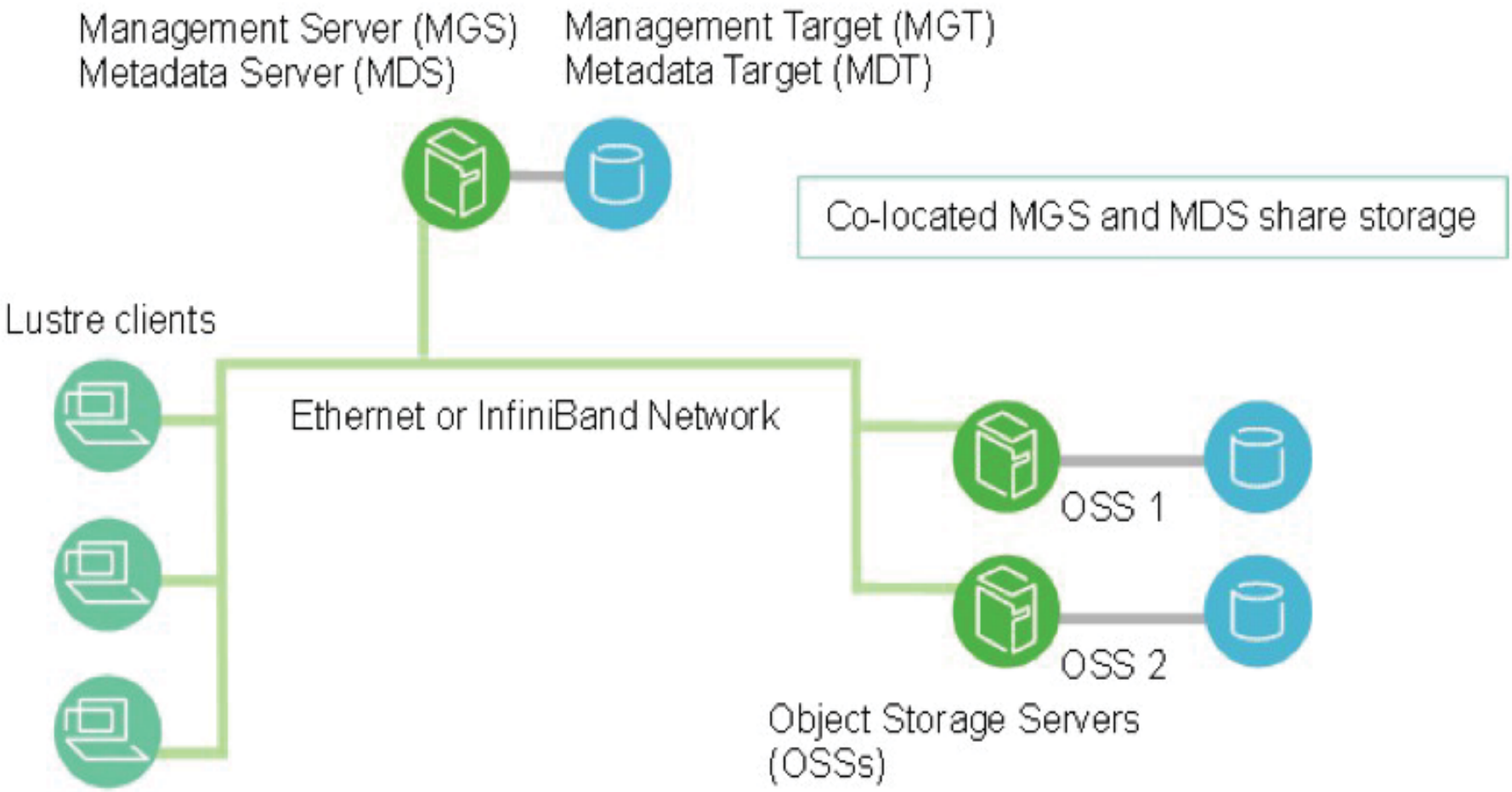
Lustre: servers and targets

- 3 main components
  - **MDS**, *MetaData Server*
    - store metadata in the MDT, *MetaData Target*
  - **OSS**, *Object Storage Server*
    - store data files on one or more objects, each of them exists on a OST, *Object Storage Target*
  - **Lustre client**, computational nodes, desktops, an entity able to mount a Lustre file system
    - each client sees a single, coherent, synchronized namespace

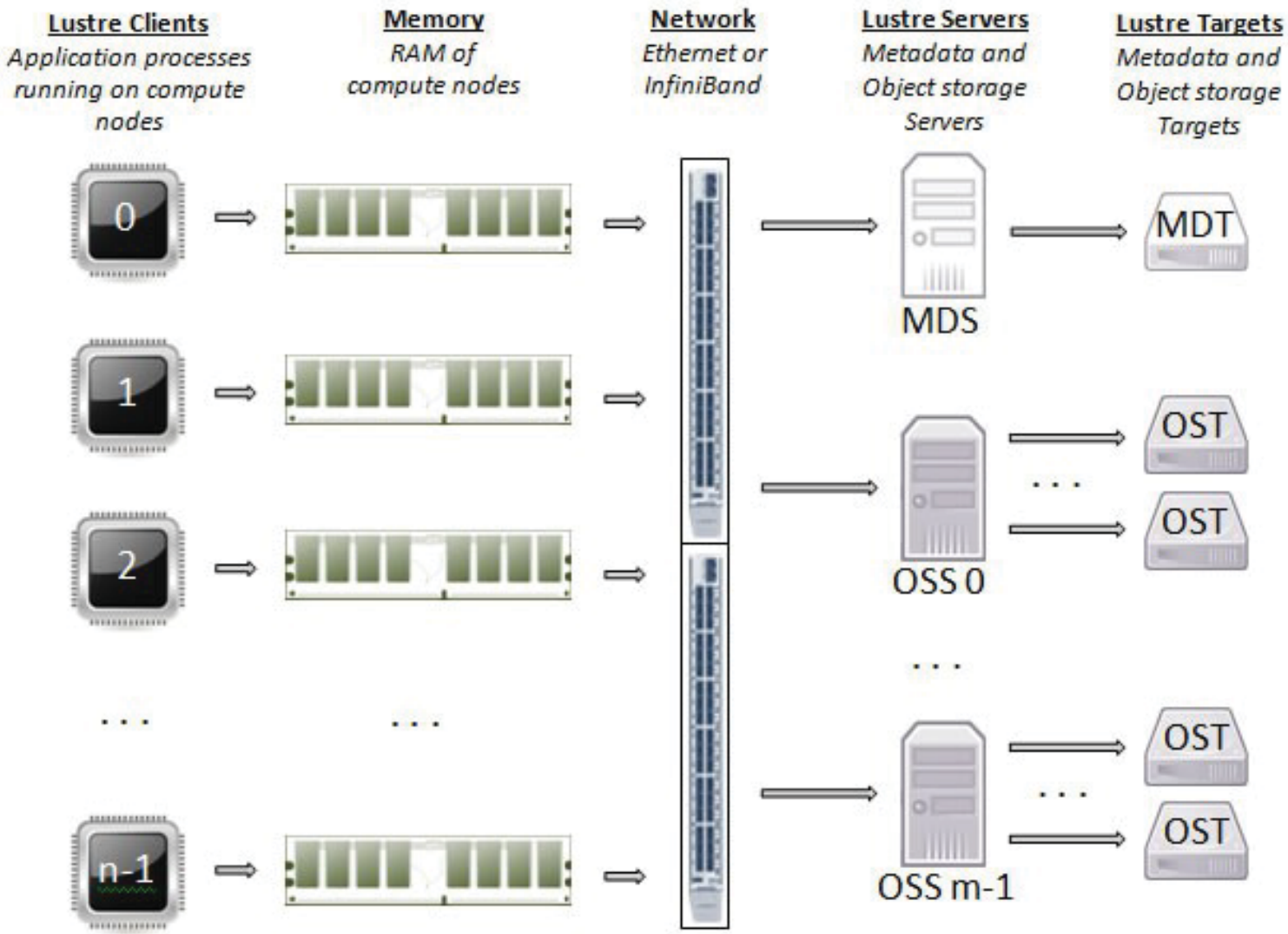
Lustre: network

- communication between servers and client is managed by Lustre networking (LNET)
- Lustre supports many high performance, low latency networks
  - InfiniBand: OpenFabric OFED (o2ib)
  - Myrinet: MX
  - any network carrying TCP traffic (GigE, 10GigE, IPoIB)
- permits RDMA, when supported by underlying networks (InfiniBand, Myrinet MX)

Basic architecture

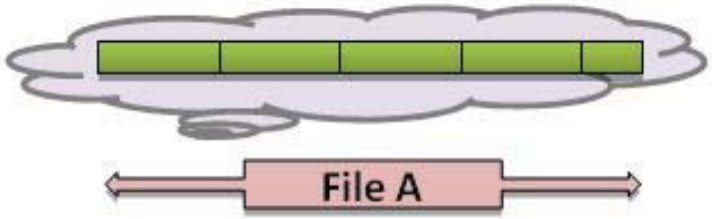


Lustre components

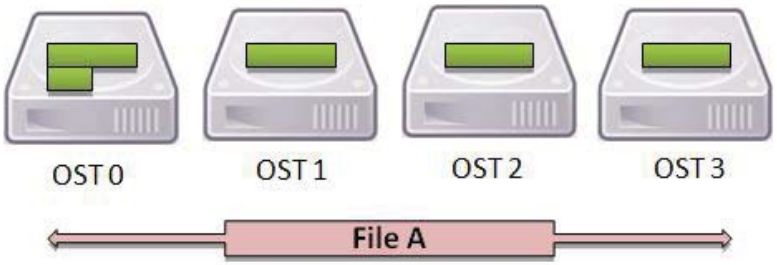


Lustre file striping

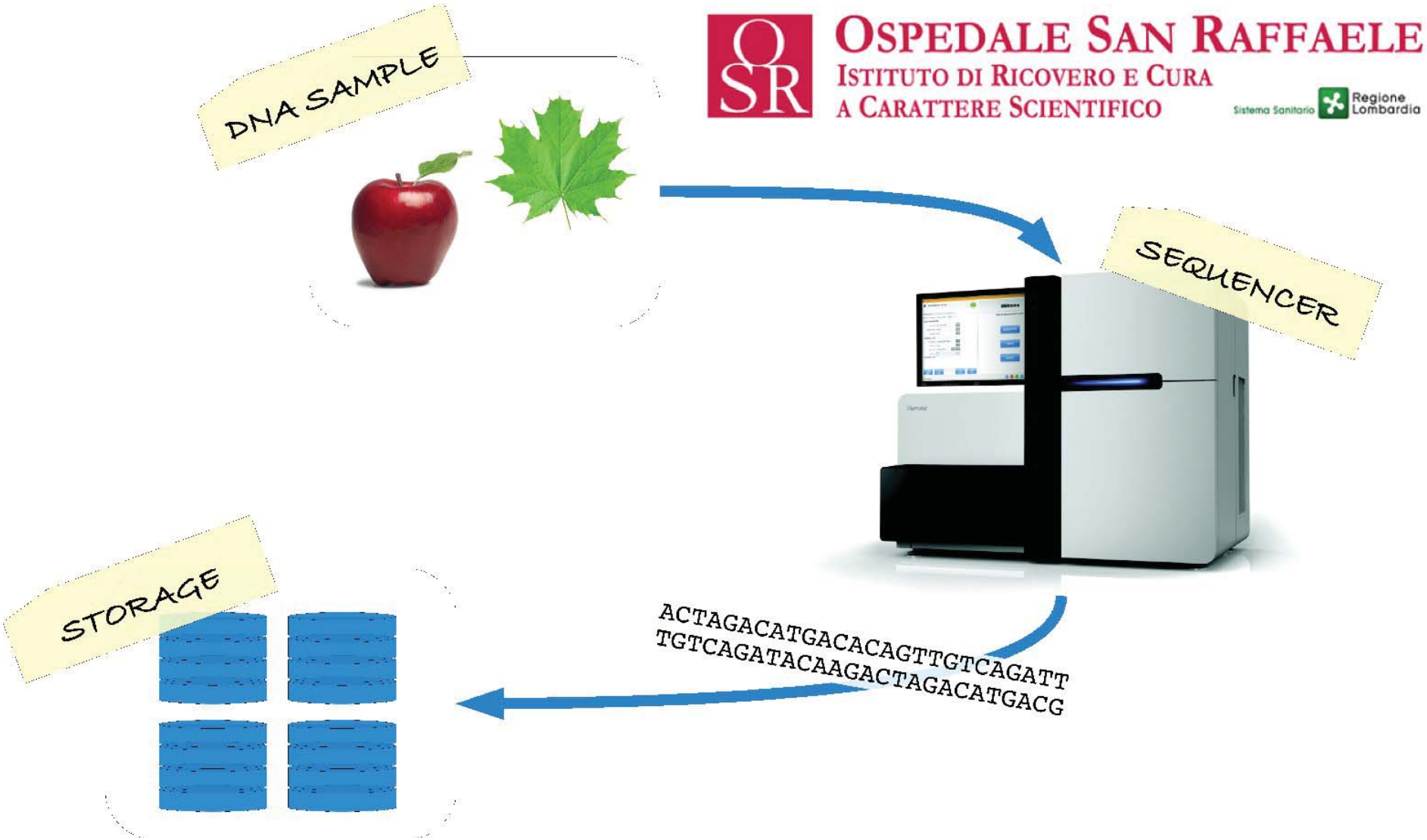
- Lustre can *stripe* a single file on multiple OSS
  - normally a file is a byte sequence...



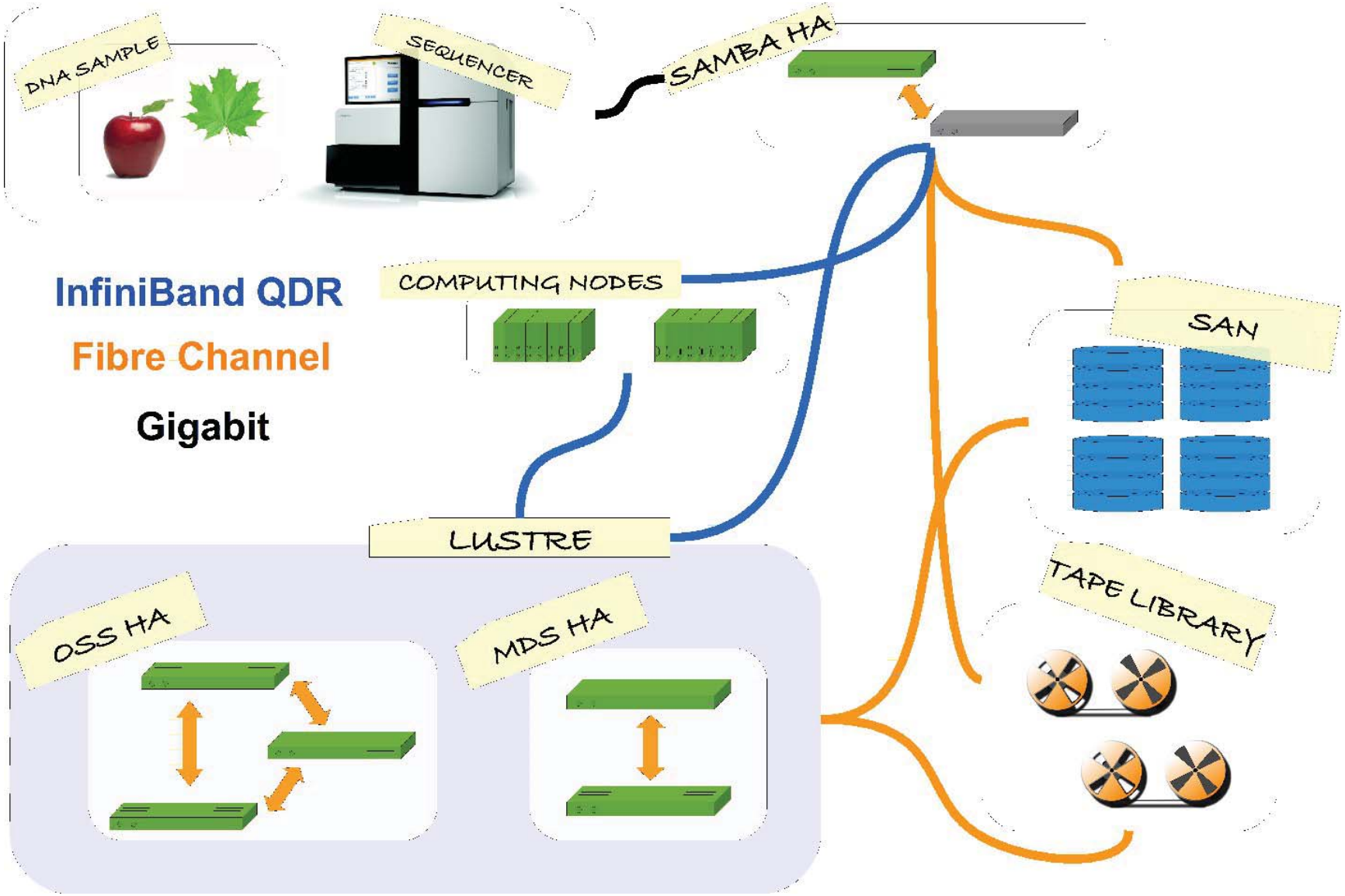
  - ...but Lustre can divide it into chunks



back to case study: the workflow



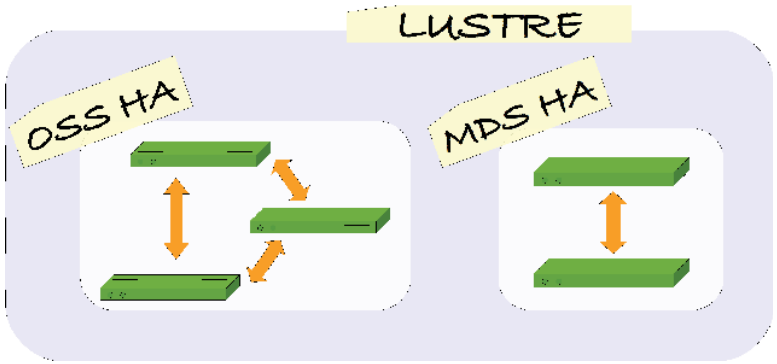
infrastructure overview



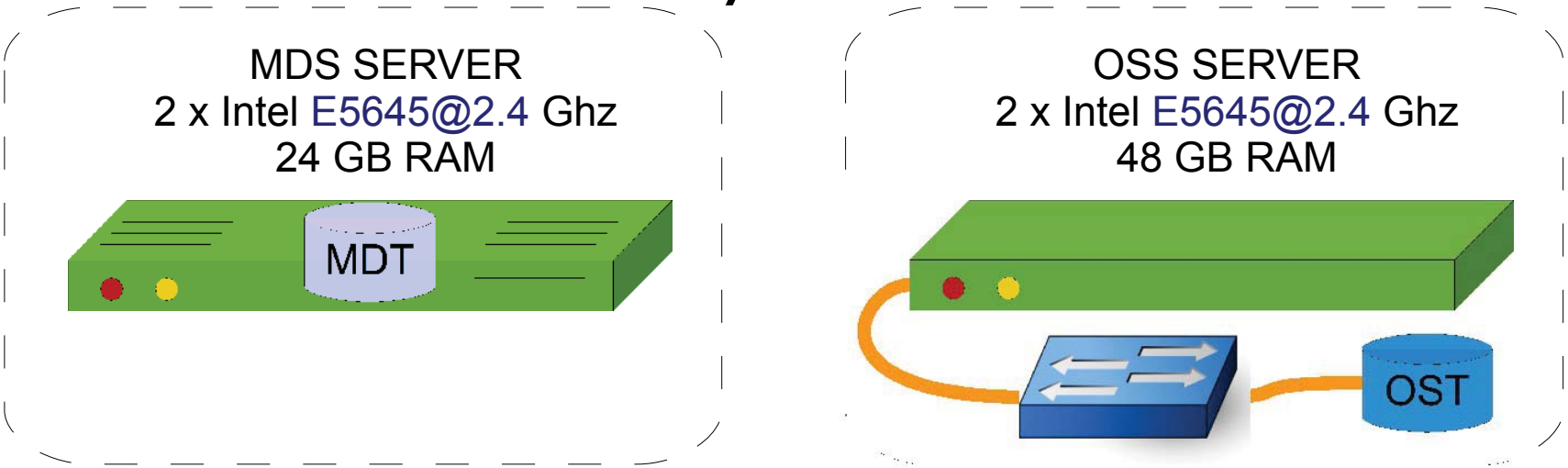
Lustre filesystem

2 Lustre filesystems

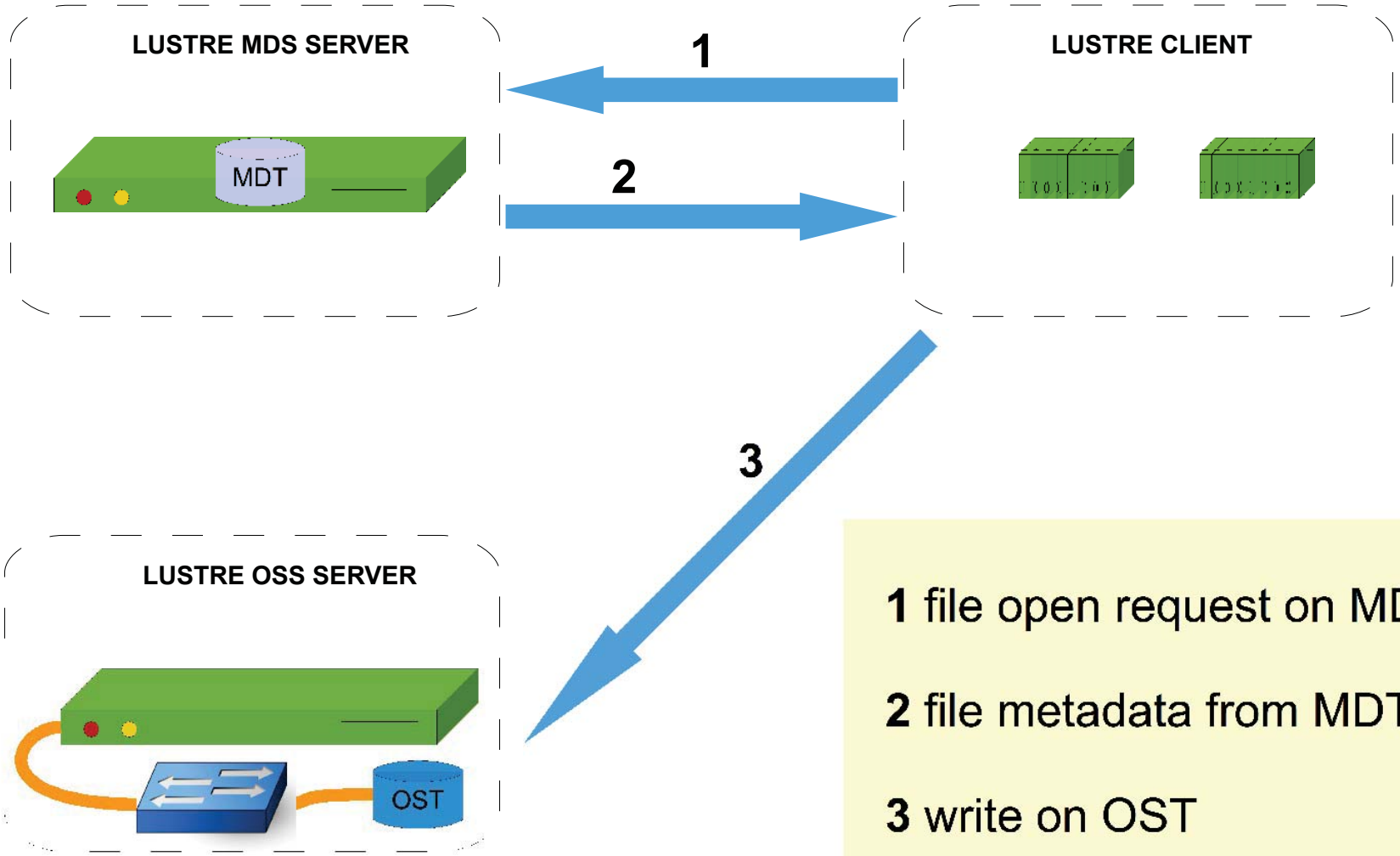
- 2 MDSs, 3 OSSs
- ~50 clients
- 60 terabytes from SAN



<u>always available!</u>



How I/O works in Lustre

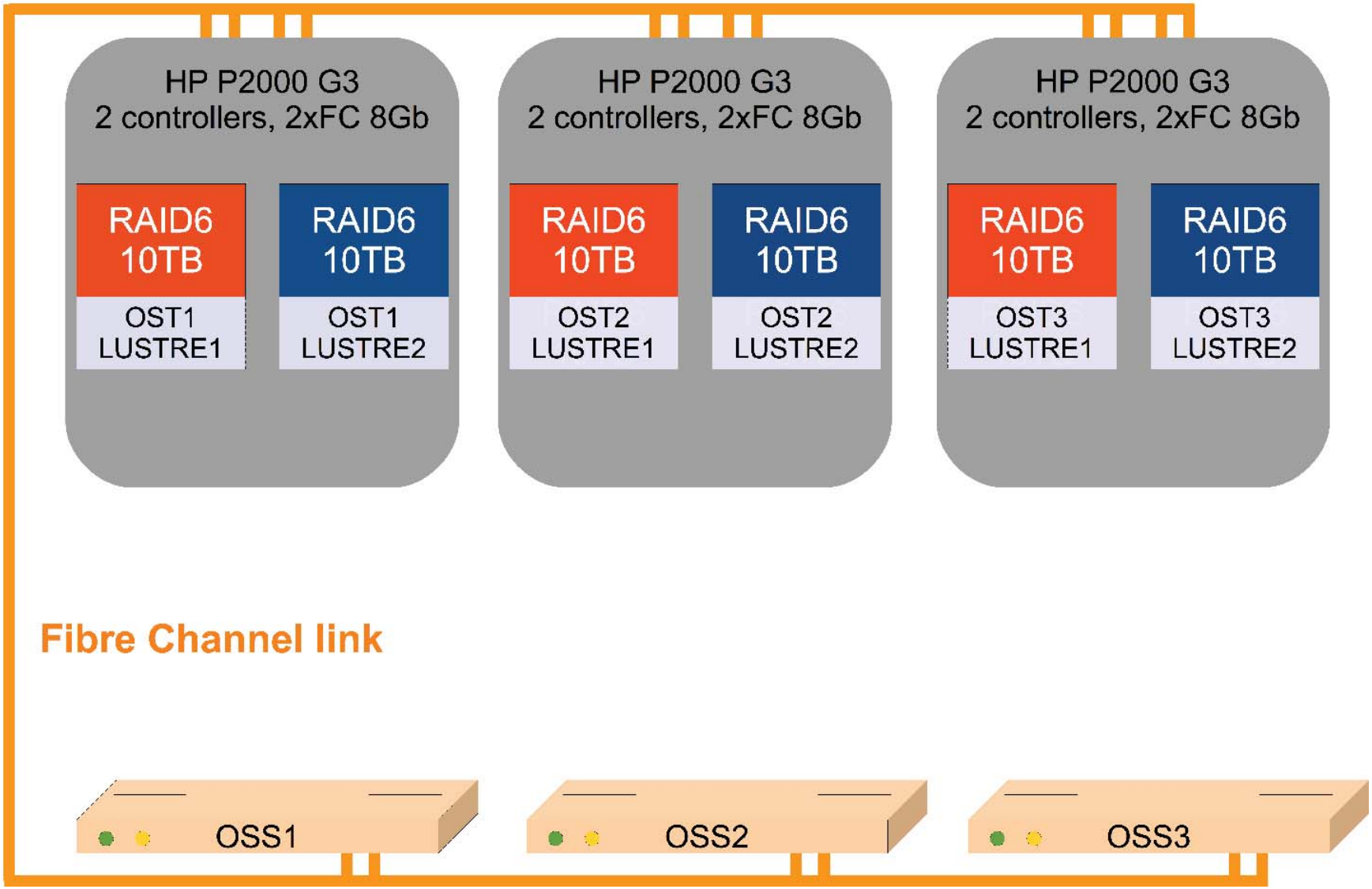


1 file open request on MDT

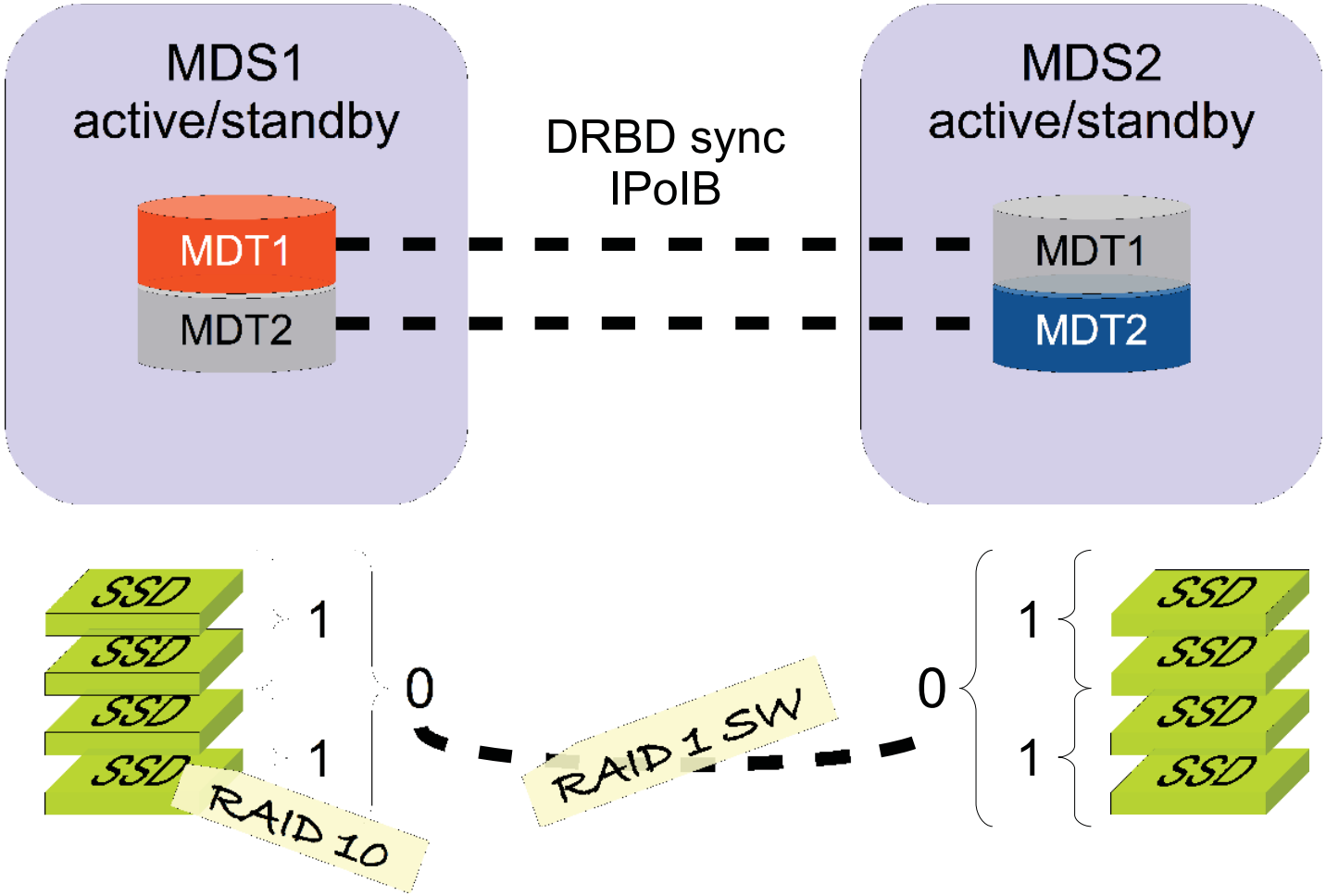
2 file metadata from MDT

3 write on OST

Lustre components: OSTs and OSSs



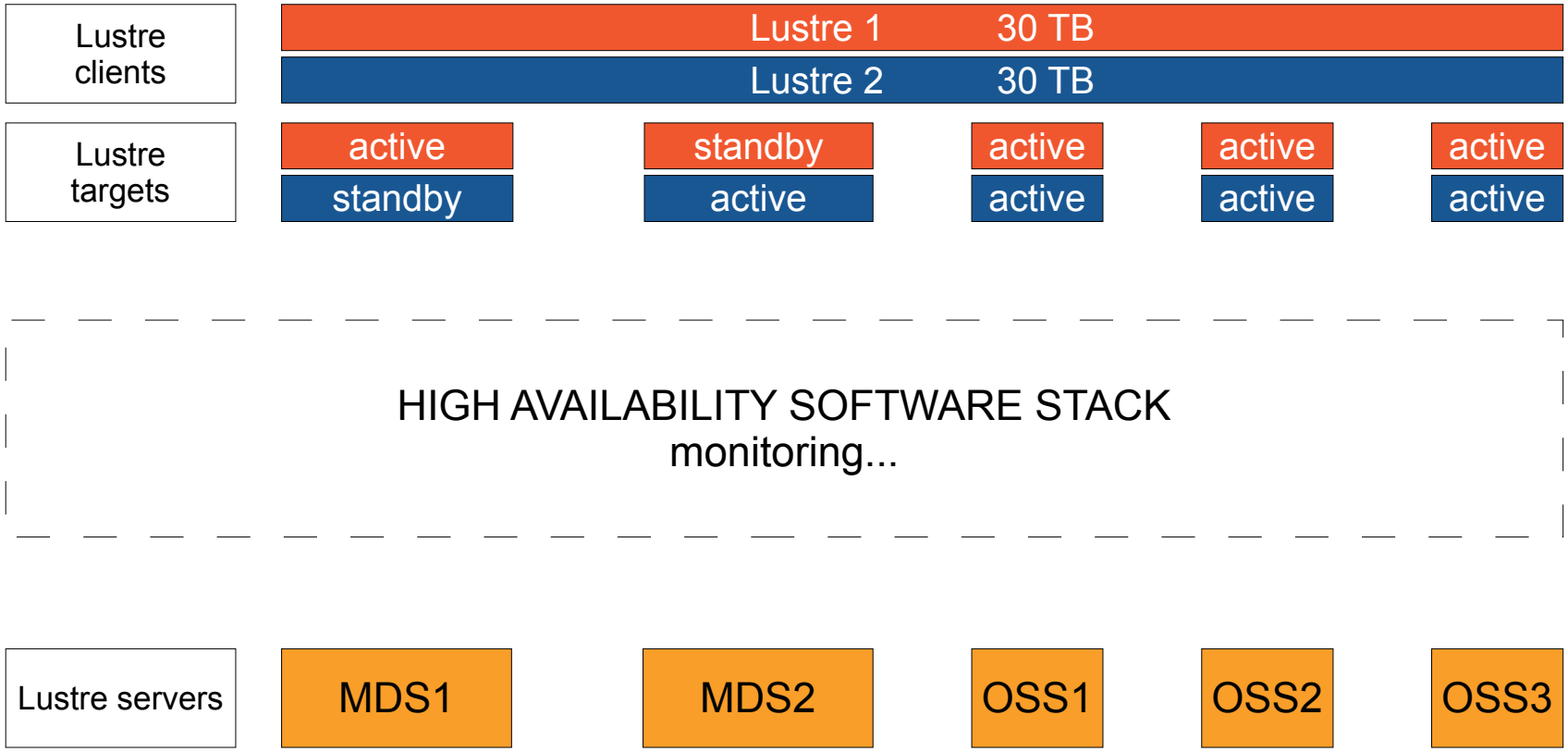
Lustre components: MDSs and MDTs



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

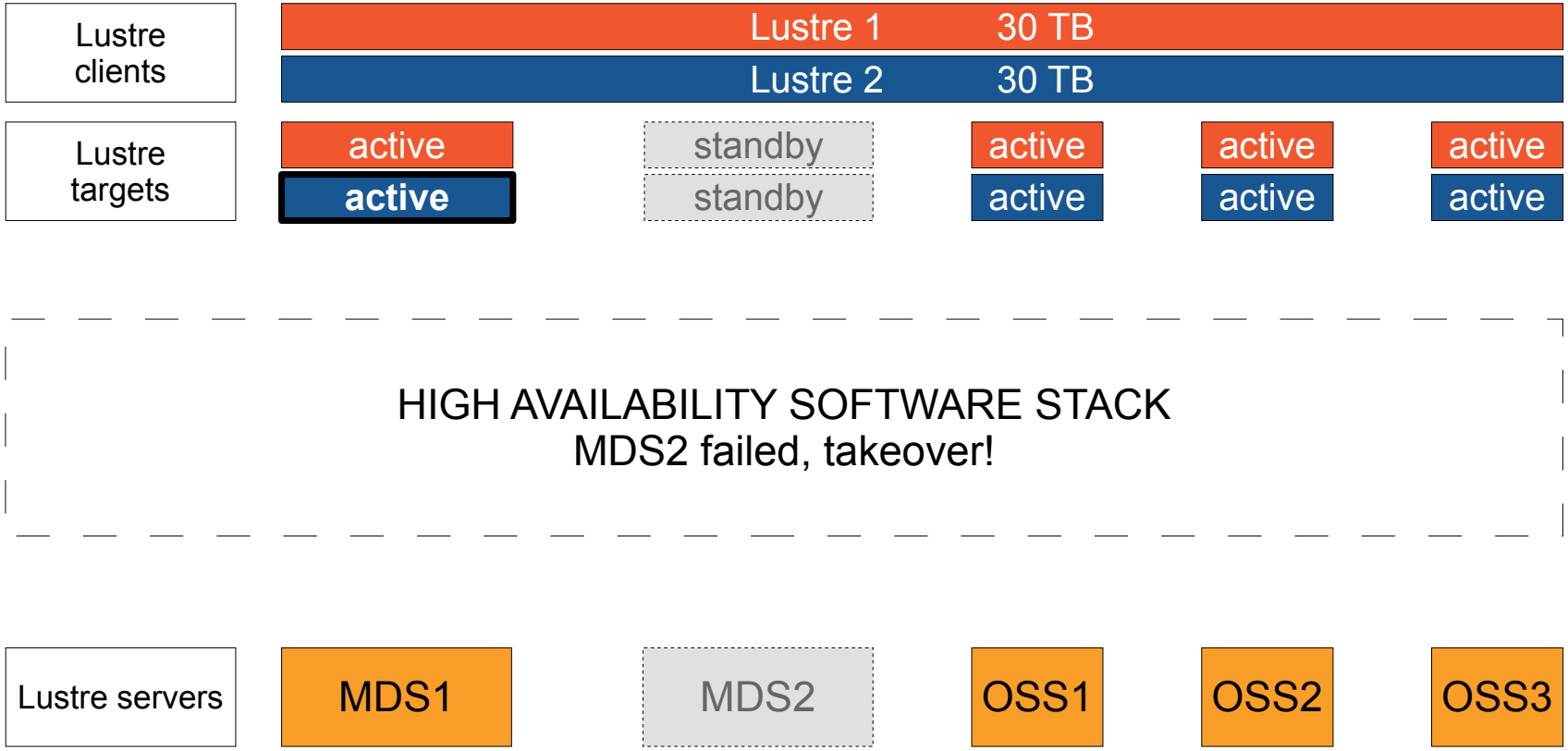
Lustre high availability

- In production, no failures



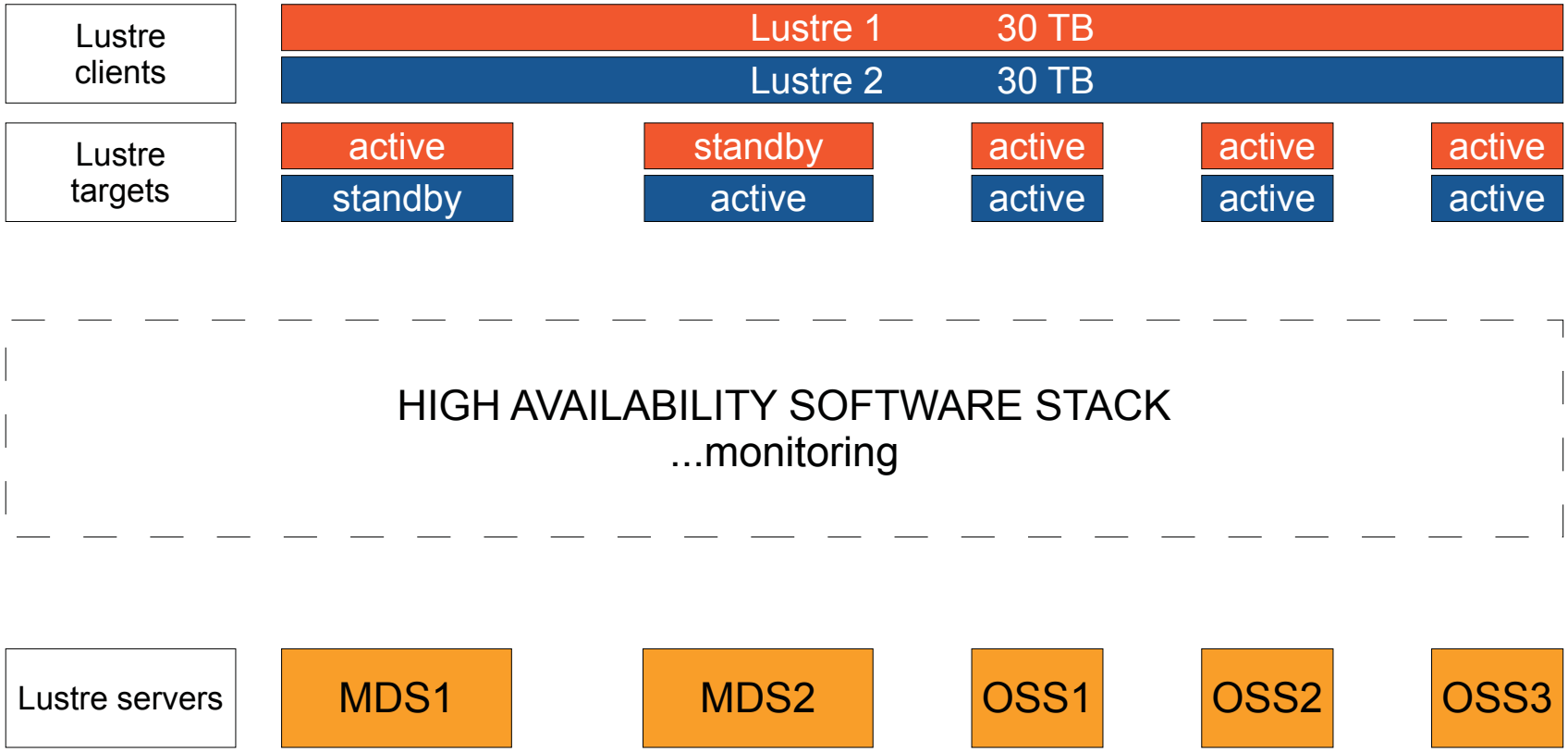
MDS2 failure

- MDS1 will take over the service of MDS2



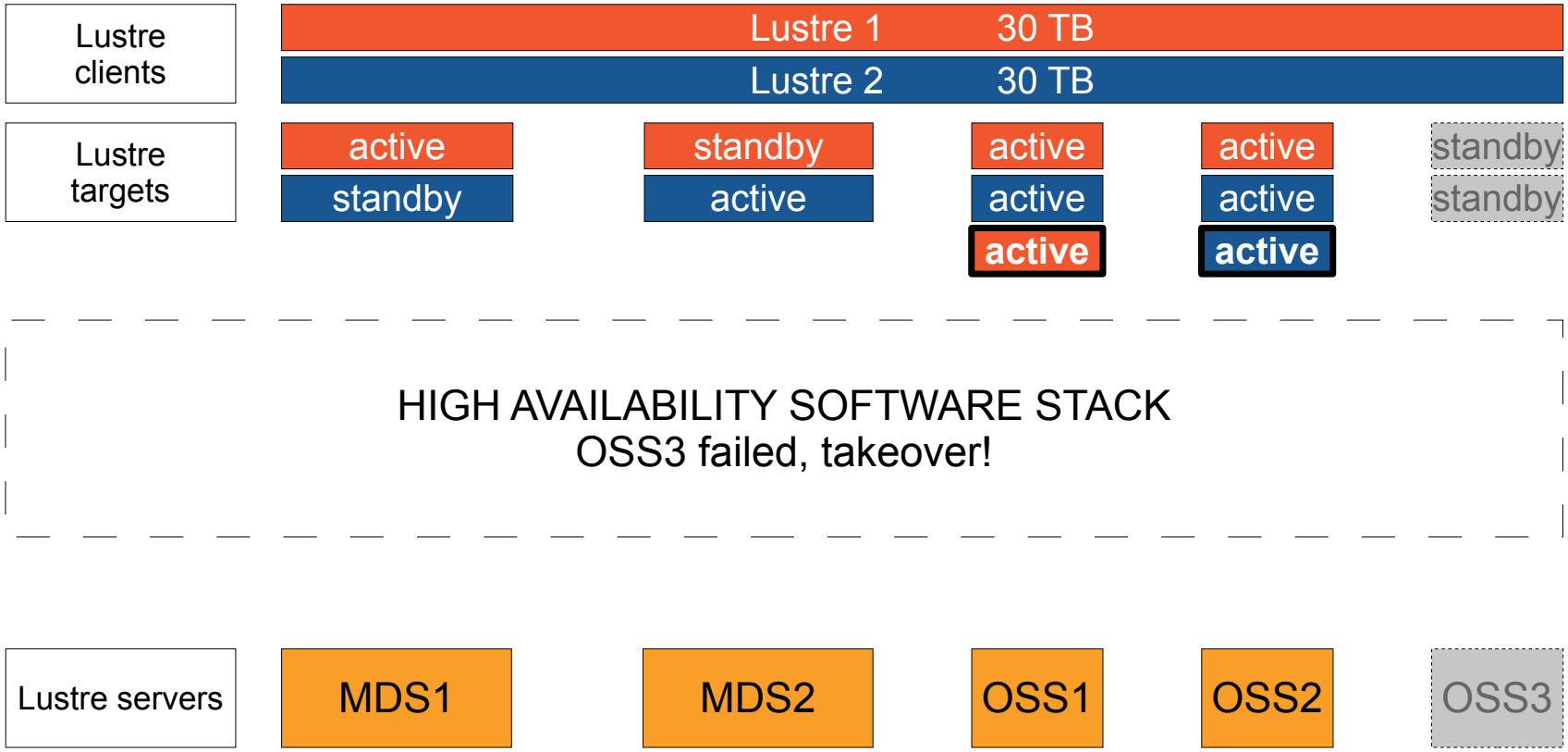
Failback

- MDS2 recovers its services



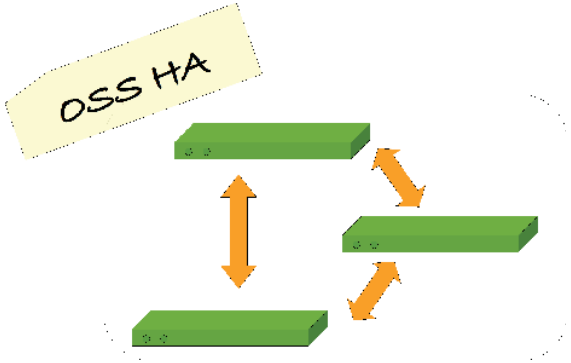
OSS3 failure

- OSS1 and OSS2 will take over its service

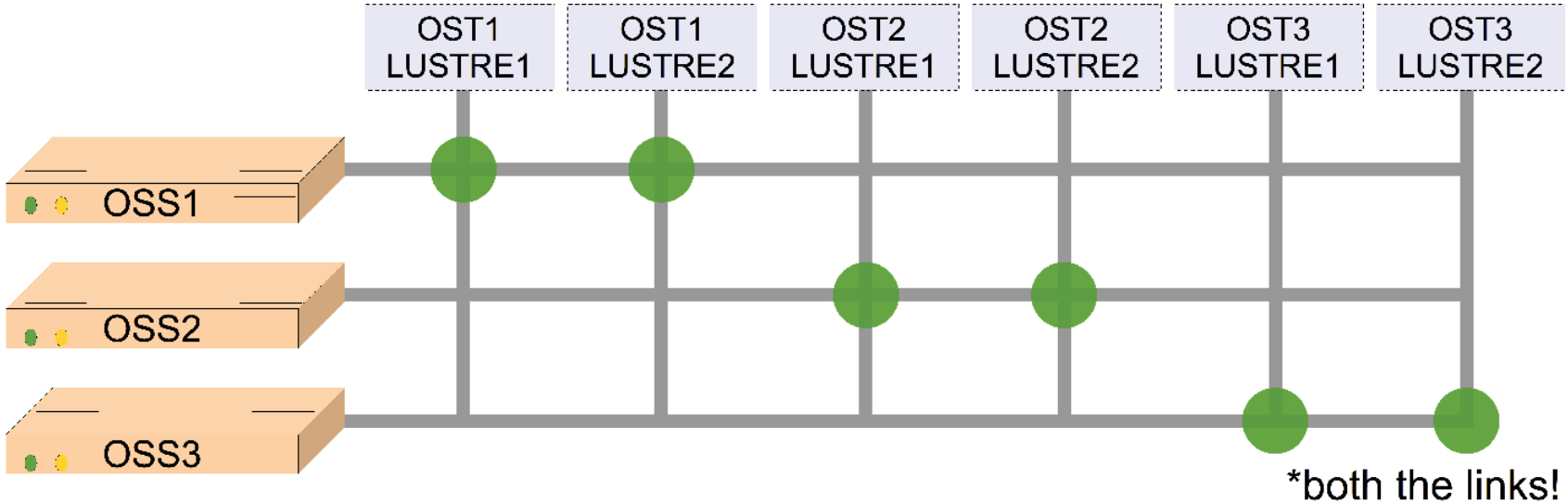


High availability on OSSs

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

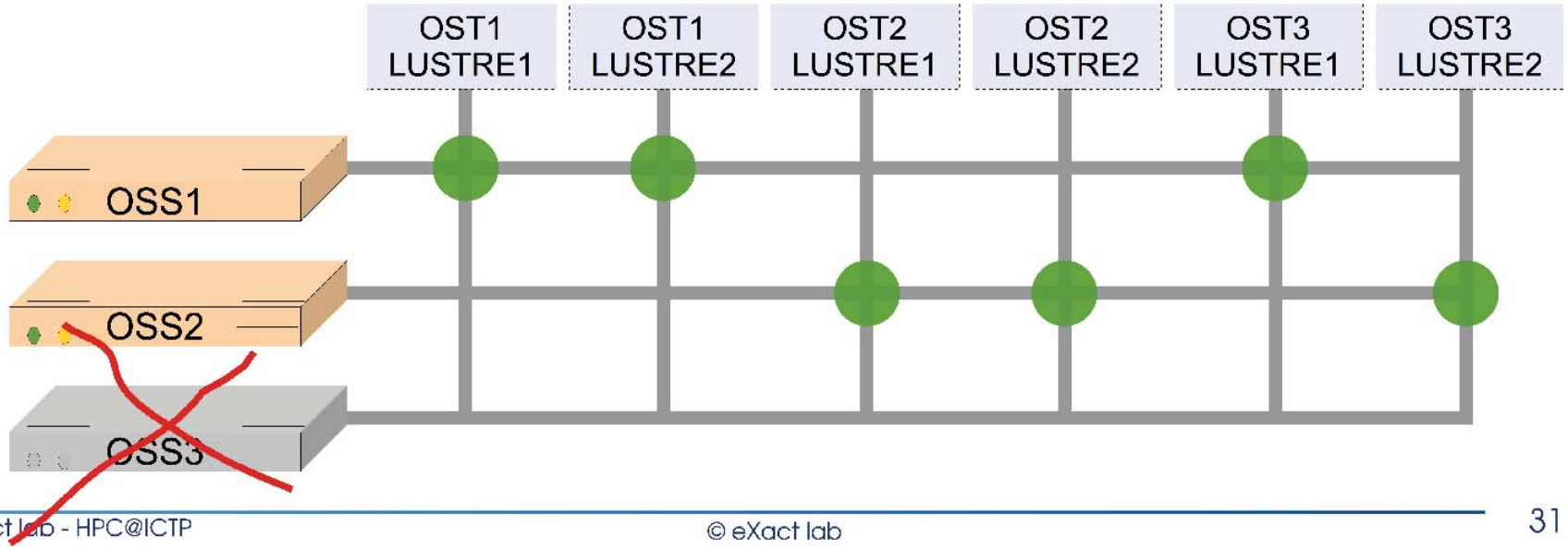


- In production, each OSS server mounts 2 OSTs



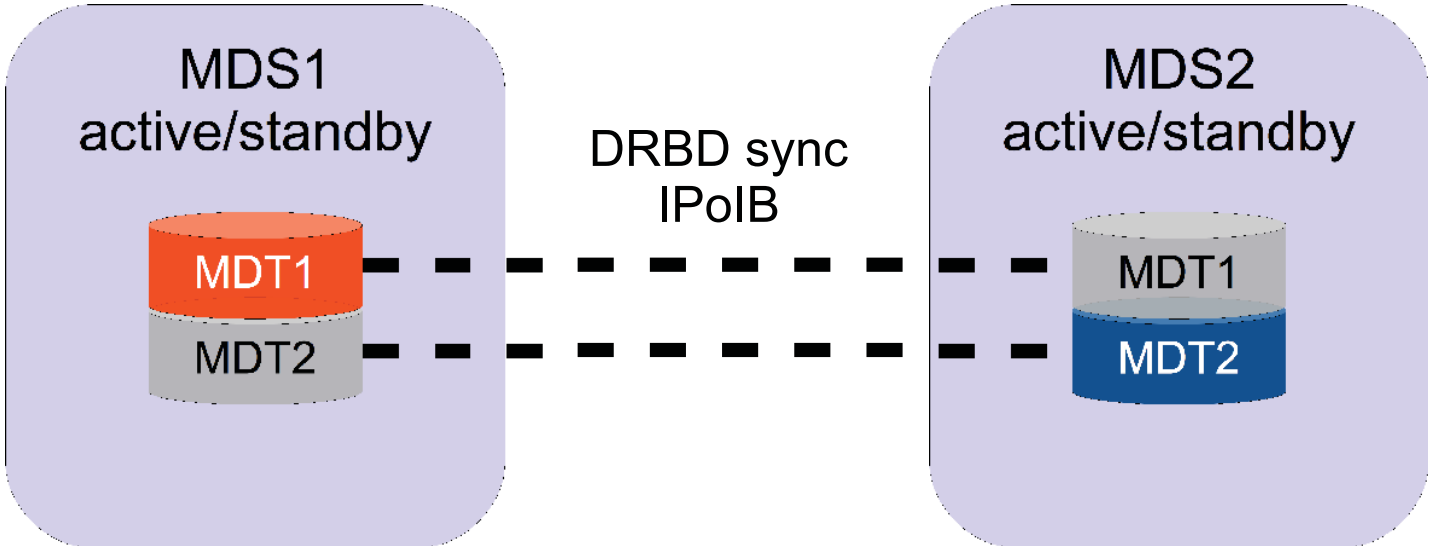
High availability on OSSs

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



High availability on MDSs

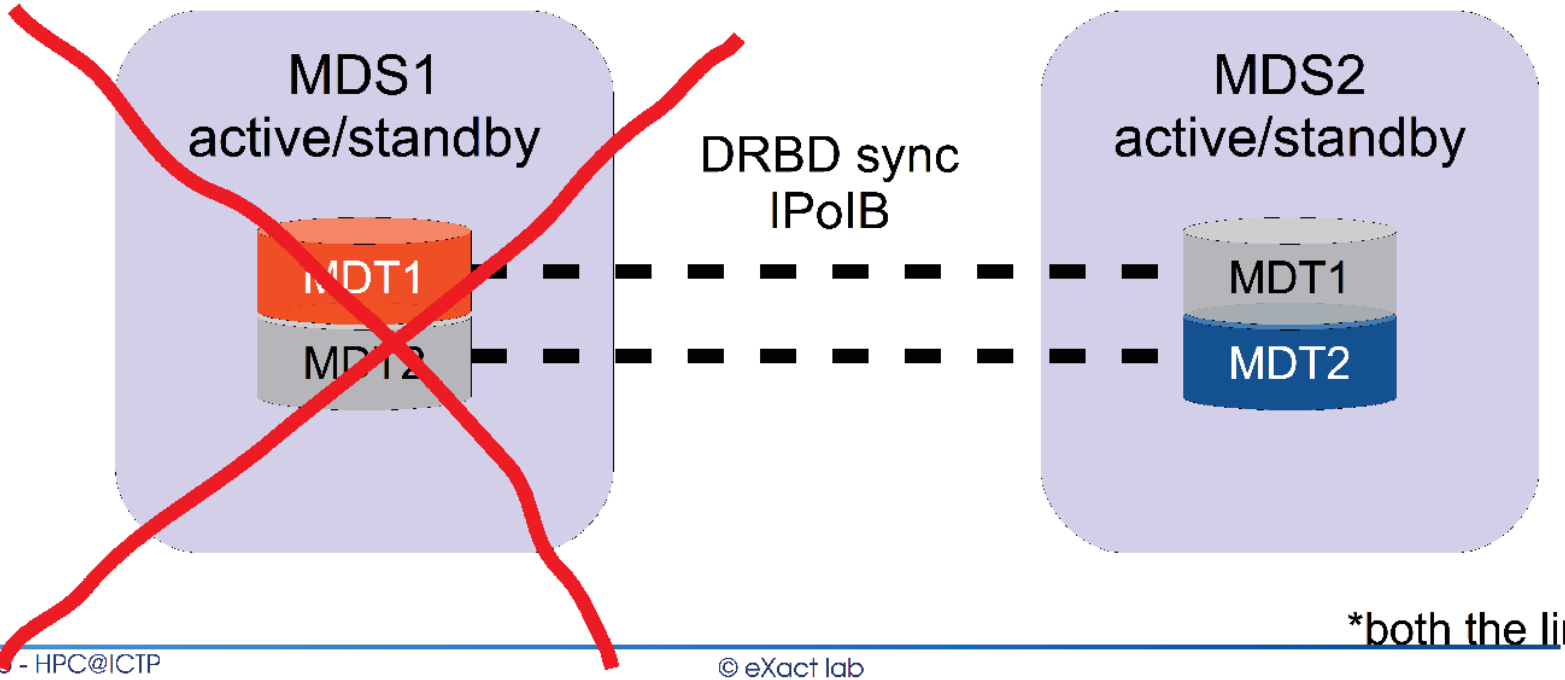
- In normal condition
  - MDTs are replicated between MDSs
  - only one replica is active for Lustre client



*both the links!

MDS1 failure

- data integrity **MUST** be ensured
- MDS1 *irresponsive* ≠ *isn't accessing my data*
- clients must not be able to reach MDS1!

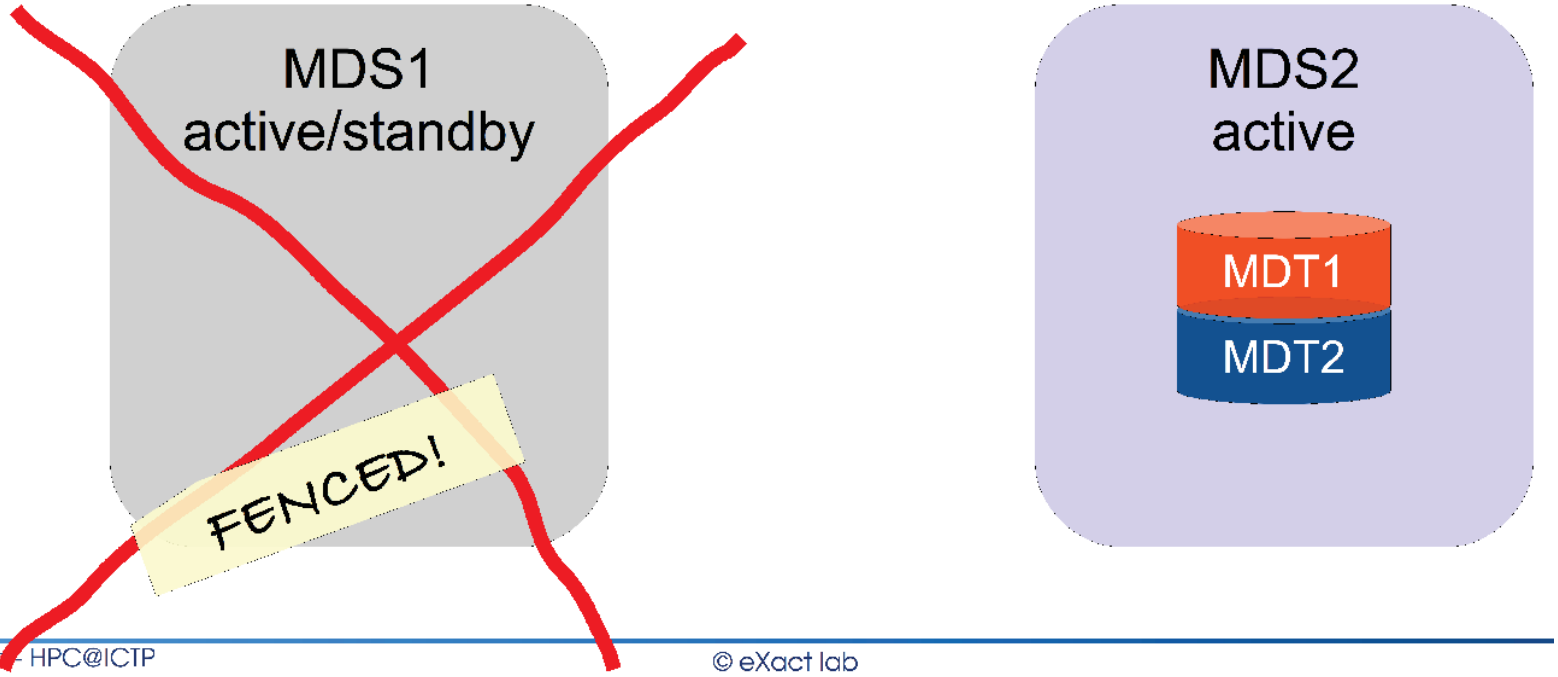


*both the links!

how to ensure data integrity

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 powers off MDS1 !**



High availability tests

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

➔ Completely transparent for clients!

Q&A

High Availabilty Lustre
FS implementation
for Genomic

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