XINNOR

All-flash Multinode High Availability for Lustre Disaggregated Implementations

Daniel Landau, Solution Architect landau.d@xinnor.io
Davide Villa, CRO davide.villa@xinnor.io

ABOUT XINNOR

- Founded in Haifa, Israel, in May 2022
- Background: 10+ years of experience with software RAID
- Mission: to offer the best data protection scheme for high performance storage
- Team: Around 50 people; >35 are accomplished mathematicians and industry talents from Global Storage OEMs
- >25 selling partners worldwide
- >100PB of end-customers data

Technology partners

Western Digital. Micron ScaleFlux KIOXIA SOLIDIGM. DapuStor

BeeGFS TUXERA LINBIT SUPERMICRO NVIDIA ATTO arm ingrasys

xiRAID, the fastest RAID for NVMe SSDs

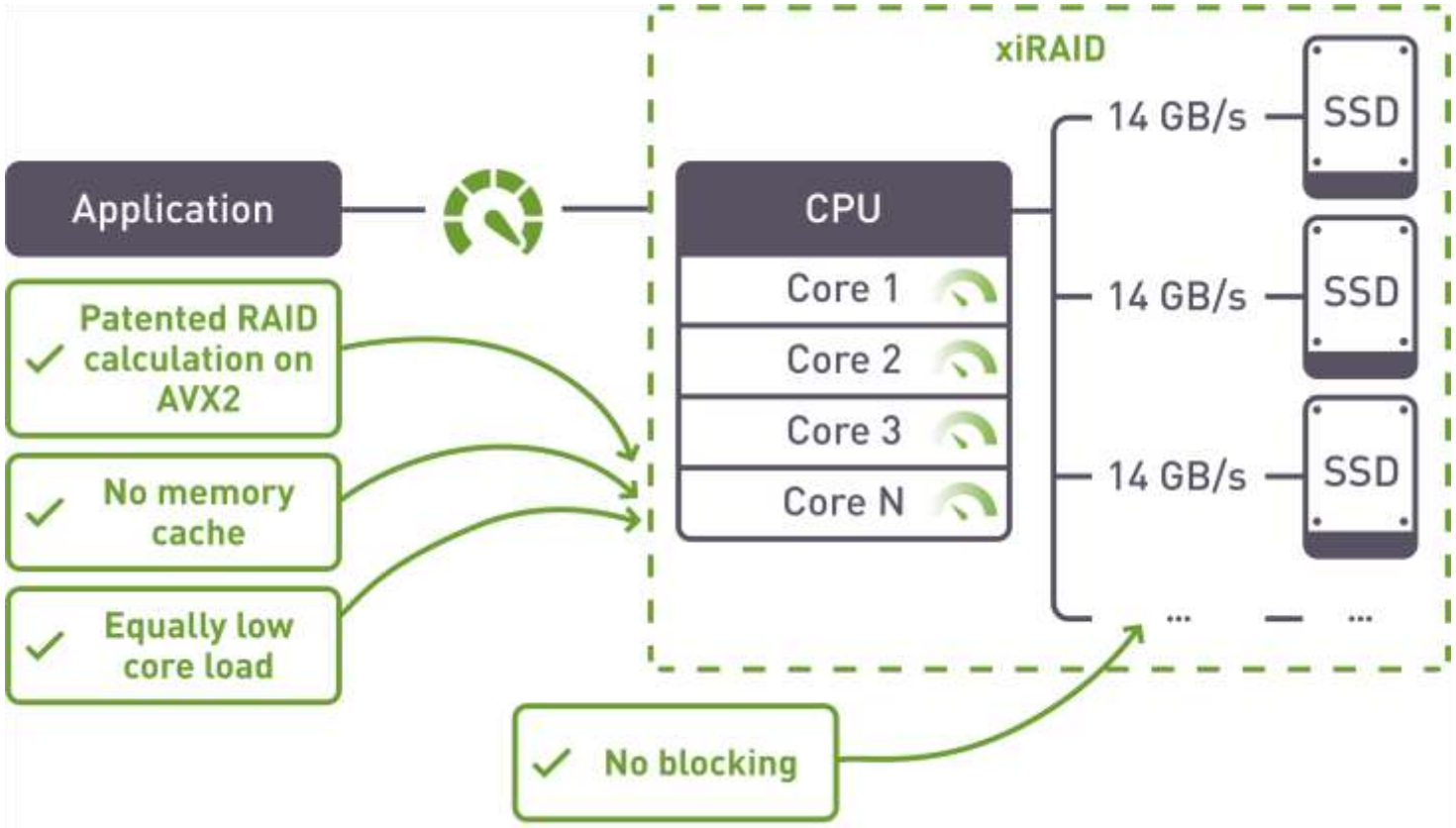
Superior performance in
normal operation

High performance in
degraded mode

No PCIe taxation
Minimal CPU load

Flexibility
RAID level 0, 1, 5, 6, 7.3, 10, 50, 60, 70 and N+M

Features rich
Including support for Pacemaker cluster



Other software data protection engines

Linux mdadm

- multiple RAID levels support
- good performance on seq read, low on write, random and degraded mode
- Pacemaker cluster support (according to the documentation)

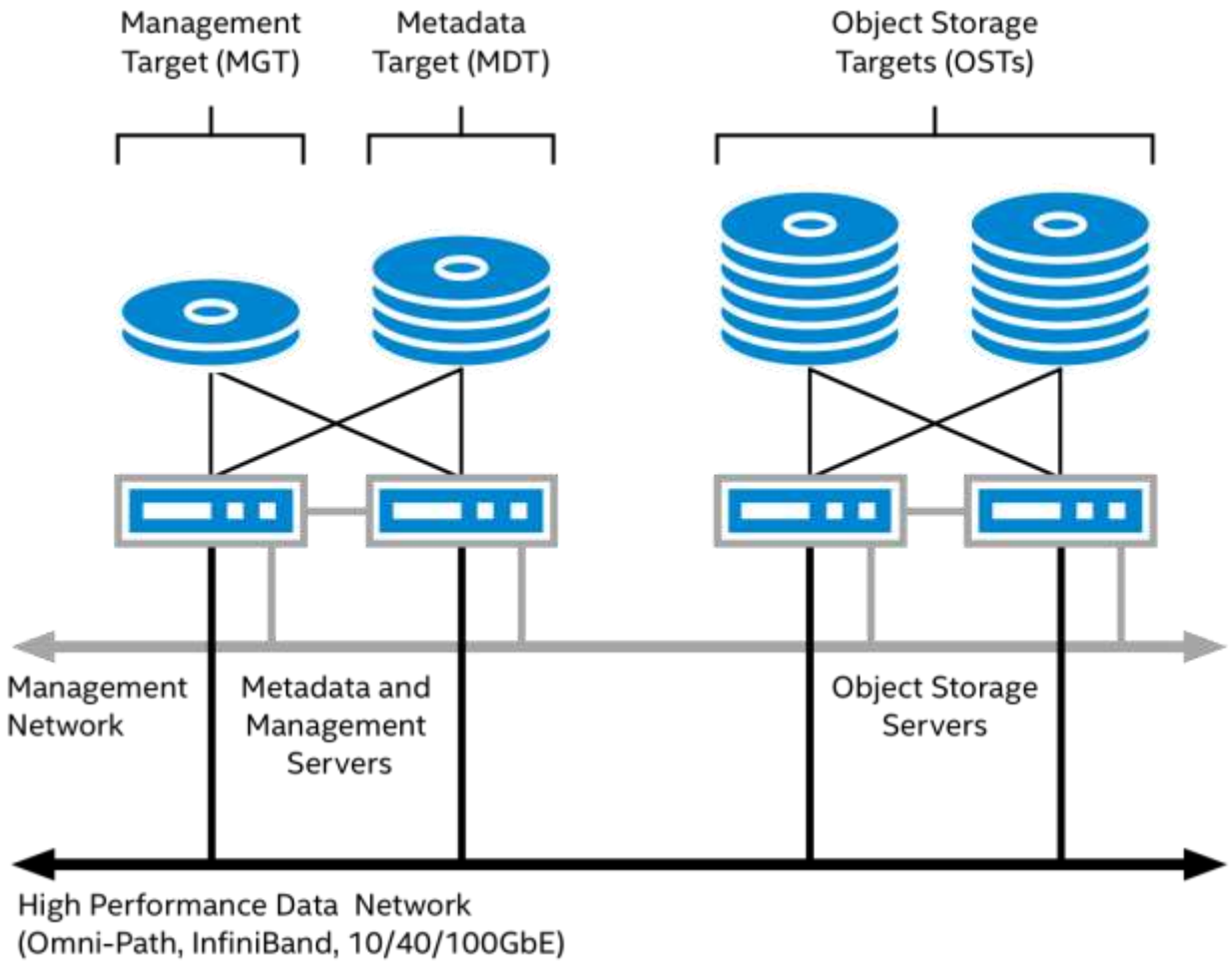
ZFS

- multiple redundancy levels supported for vdevs (raidz, raidz2, raidz)
- great flexibility
- medium performance for sequential workloads (waiting for the Direct IO support)
- Poor performance for random workloads
- Pacemaker cluster support

Lustre Clustering

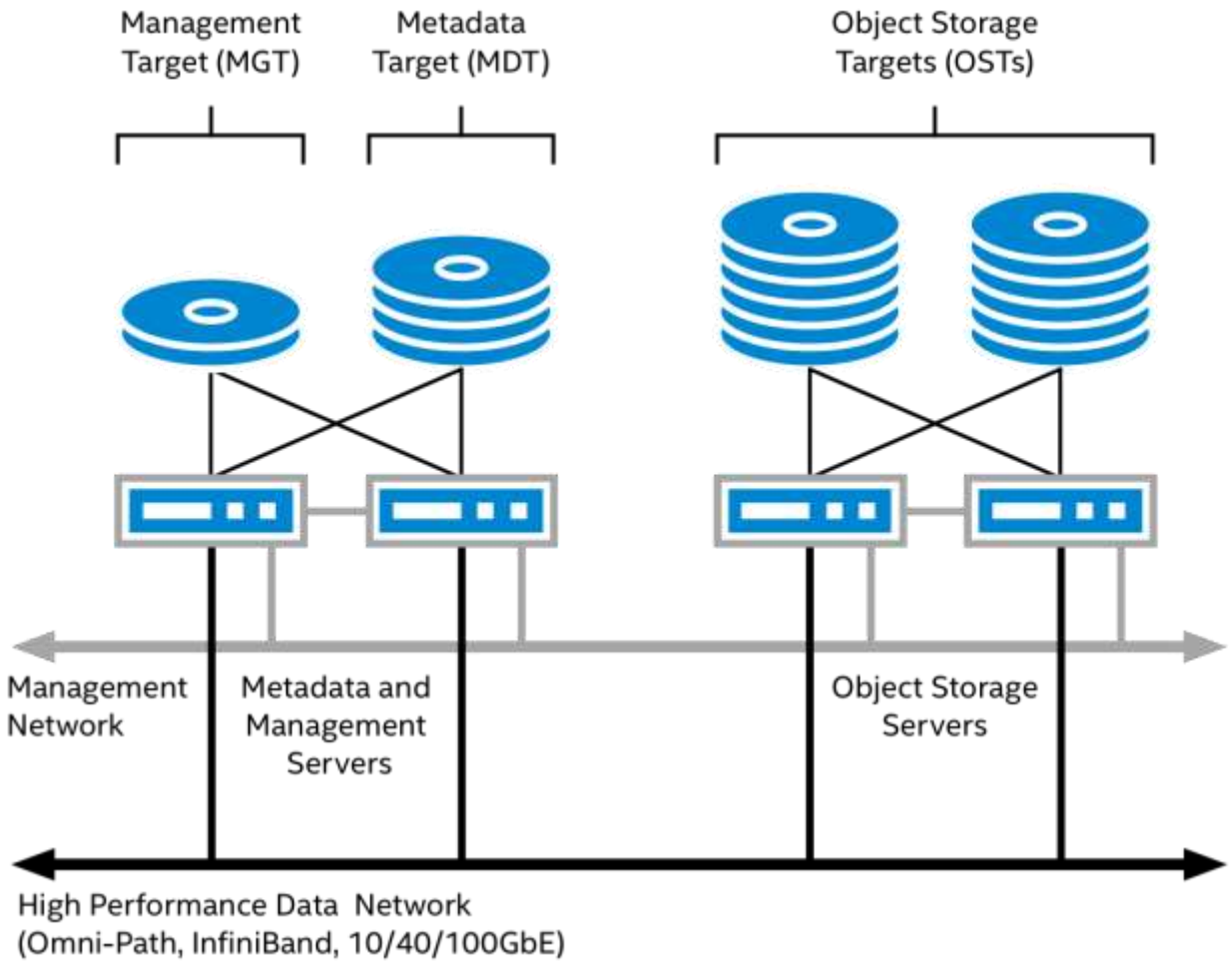
HA Cluster expectations

- Service continuity in case of a server failure (must have)
- No or minor performance degradation after a failover (should have)



Lustre HA: classical approach

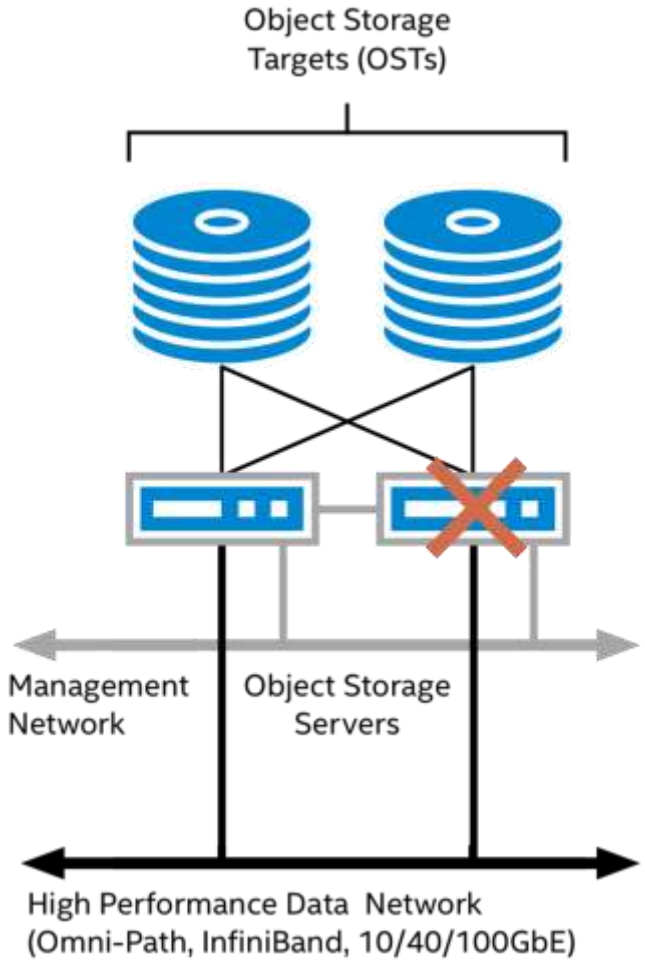
- A-P or A-A failover pairs
- Two nodes Pacemaker HA cluster as building blocks
- Shared storage:
  - Shared LUNs from a disc array with SAN access
  - A shared set of drives, combined into a virtual device using some data protection engine locally on Lustre servers



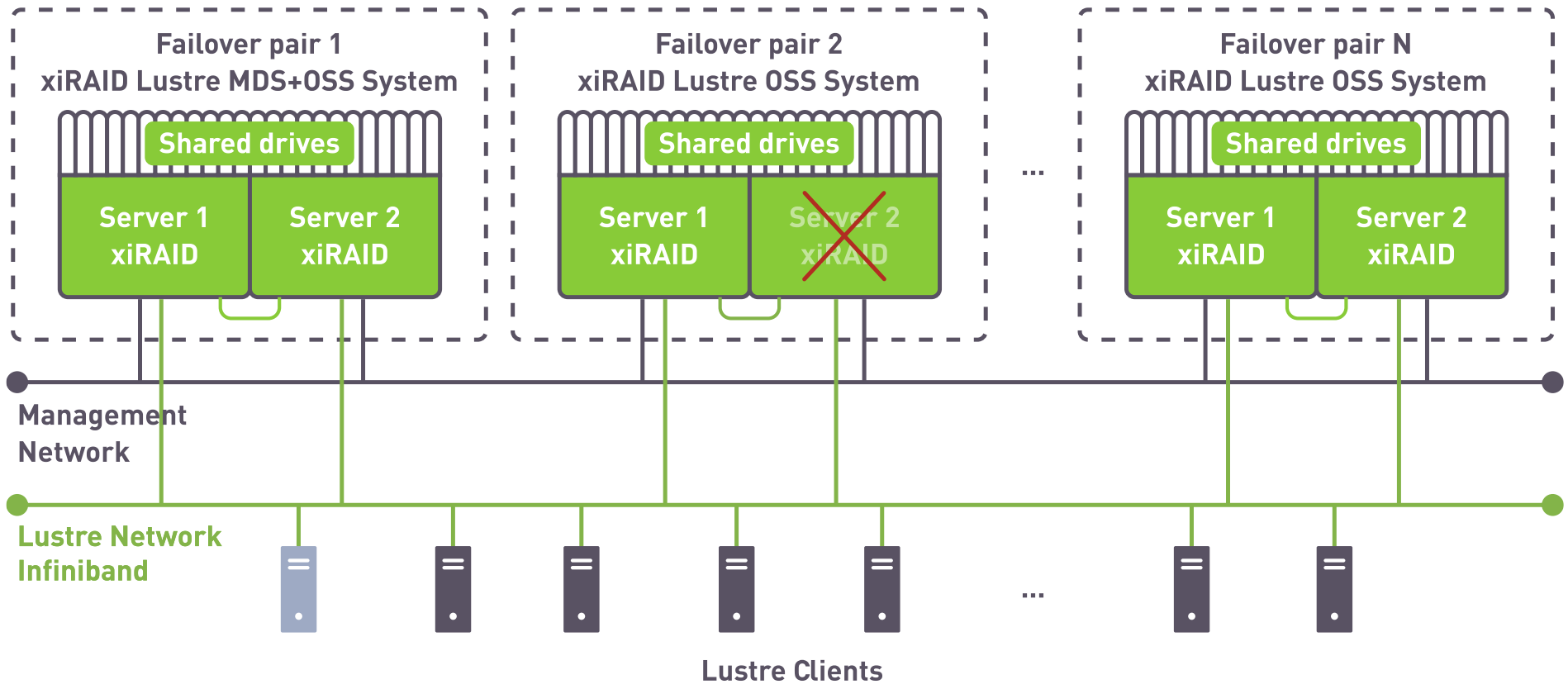
Failover pair performance

- In case of a single server failure, the remaining server will work under increased workload
- It may cause performance degradation up to 50%
- To avoid performance degradation, each server should be originally sized to be able to serve two times more intensive workload. Each server should have:
  - up to twice LNET connections throughput
  - up to twice drives connections throughput
  - up to 2 times amount of RAM
  - up to 2 times amount of CPU
  - etc...

=> Higher the system price.



Performance impact to the overall filesystem



Performance degradation at one of a single failover pair usually affects the performance of all the filesystem or a pool (depending on pools and striping configuration).

NVMe shared access options

Dual-ported drives:

- NVMe drives are available as 1x4 (single port) or 2x2 (dual port) mode;
- In dual ported mode, a single NVMe drive can be connected to different hosts;
- Unlike SAS drives, it’s not possible to get the full performance from a single port, if a drive is connected in dual-port mode

EBOFs

- Just a bunch of NVMe drives, connected to Ethernet
- Supports NVMe-oF protocol (RDMA and TCP/IP)
- Requires additional configuration on the initiator side (connection/multipath)
- Usually has two IO modules for redundancy. In this case, it requires dual-ported drives.

Storage Bridge Bay NVMe Systems

SBB systems: architecture overview

- Two servers and multiple dual-ported NVMe drives in a single box
- Each server has access to all the drives
- Some models have internal network connection between servers
- All servers have IPMI for fencing
- Available from multiple vendors:

Celestica™ SUPERMICRO® VIKING Enterprise Solutions A Sanmina Division

SBB systems: peculiarities

Problem: a single drive is connected by 2 PCIe lines to each server only – no way to achieve the full NVMe performance from a single server.

Solution: split each NVMe in two namespaces and use first namespaces in the RAID groups running by default at the first server and second namespaces in the different RAID groups running by default at second server.

- It will double the number of Lustre OSDs.
- Does not work in case of failover.

SBB systems: advantages and disadvantages

Advantages:

- Simple solution
- Small datacenter footprint
- Drives directly connected to servers

Disadvantages:

- Specialized hardware
- Additional NVMe configuration required to get full NVMe performance
- Storage layer performance degradation in failover state can’t be overcome
- Servers are co-located and share some components

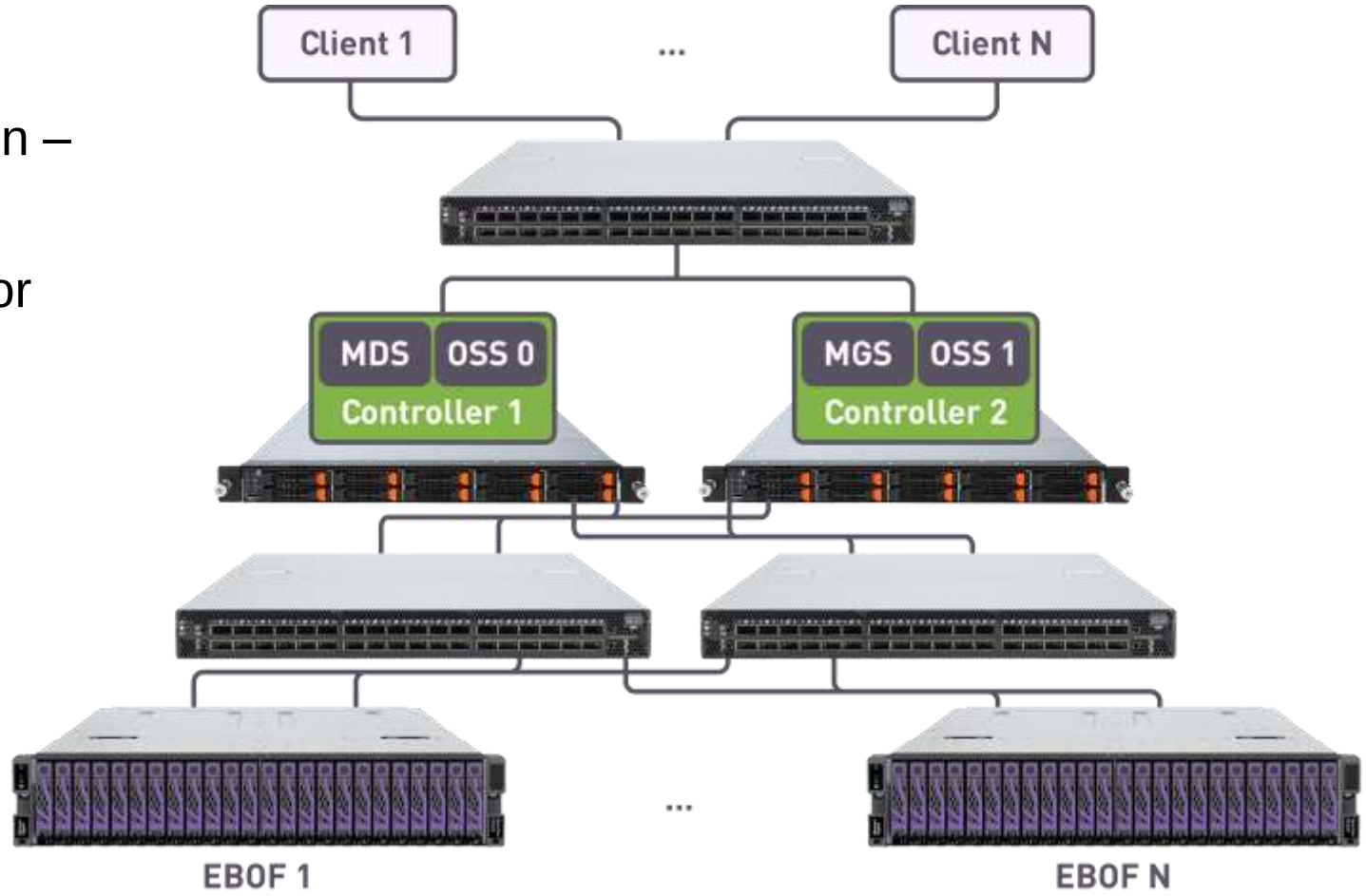
Dual-node clusters with EBOFs

EBOFs

- NVMeOF RDMA and TCP protocols by RoCE
- Usually, 24 drives max support
- Usually, dual port drives required
- 2 IO modules
- 3-6 network interfaces per IO module
- 100 or 200 GbE port speed

Dual-node cluster with EBOFs

- Generally available servers
- Network cards for EBOF connection – can be bottlenecks
- EBOFs can be connected directly or via switches
- In case of direct connection, pay attention to the NIC speeds at both sides
- If using just a single switch – it's a single point of failure



Dual-node cluster with EBOFs: advantages and disadvantages

Advantages

- Generally available hardware
- No additional NVMe configuration needed to get the full performance from the drives
- No hard limit in the number of NVMe drives used in the configuration
- Storage layer performance degradation can be avoided, if EBOF to server network is properly sized

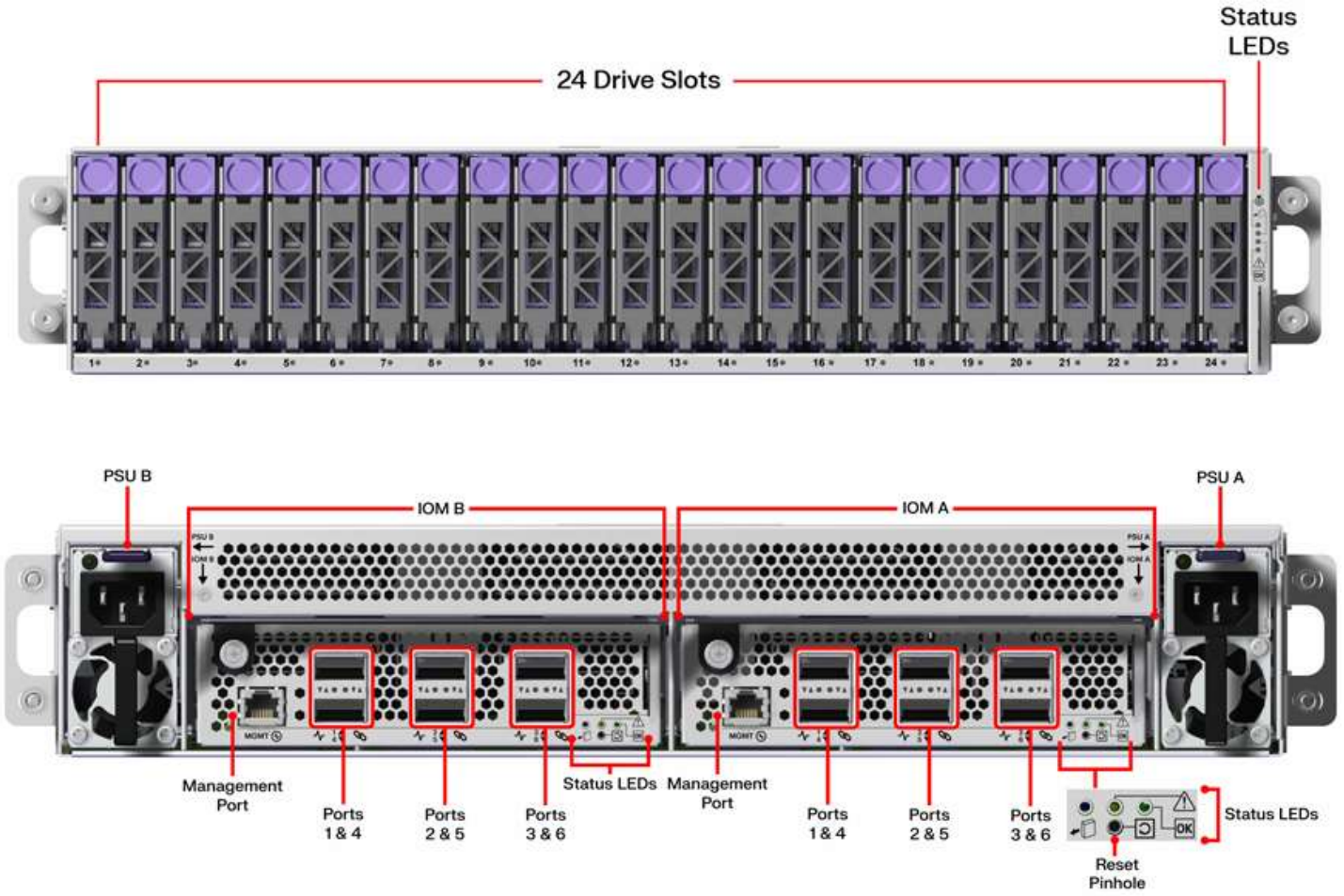
Disadvantages

- Redundant network for EBOFs to servers connection required
- Increased complexity vs SBB
- High datacenter footprint vs SBB
- To prevent performance degradation in degraded mode, each node must have redundant CPU power, memory and network throughput, which will not be used most of time

EBOFs redundancy

Multiple EBOFs: EBOF failure protection

- Most of modern EBOFs have 2 IO modules
- But usually, they have common NVMe drives backplane
- Some probability of a complete EBOF failure exists
- The probability of failure grows with the number of EBOFs in a cluster

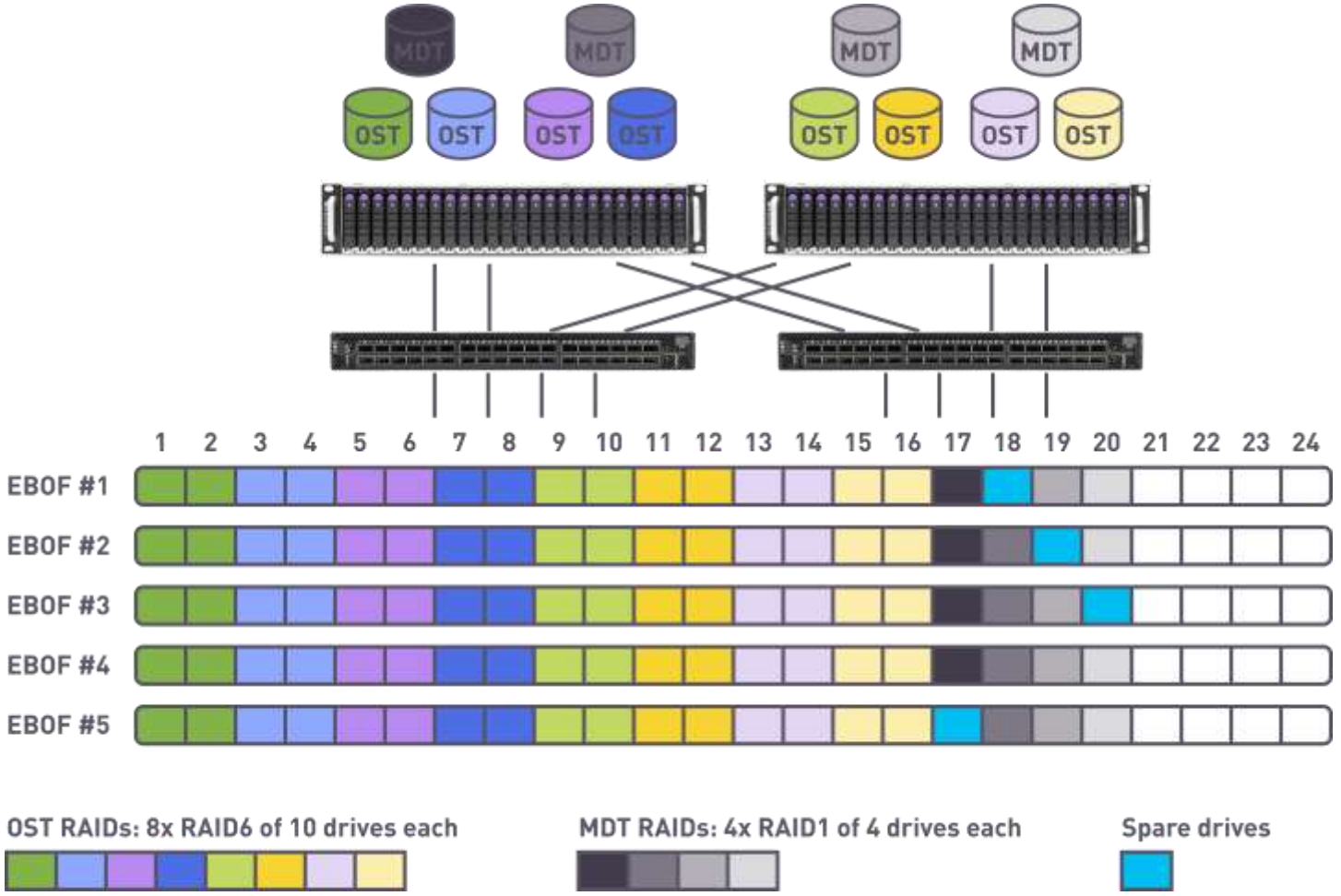


Multiple EBOFs: EBOF failure protection

Build the RAID group using from each EBOF no more than number of parity drives:

- RAID1 and 5 – no more than 1 drive
- RAID6 – no more than 2 drives
- RAID7.3 – no more than 3 drives

In case of an EBOF failure, the RAID groups will still work in degraded mode.

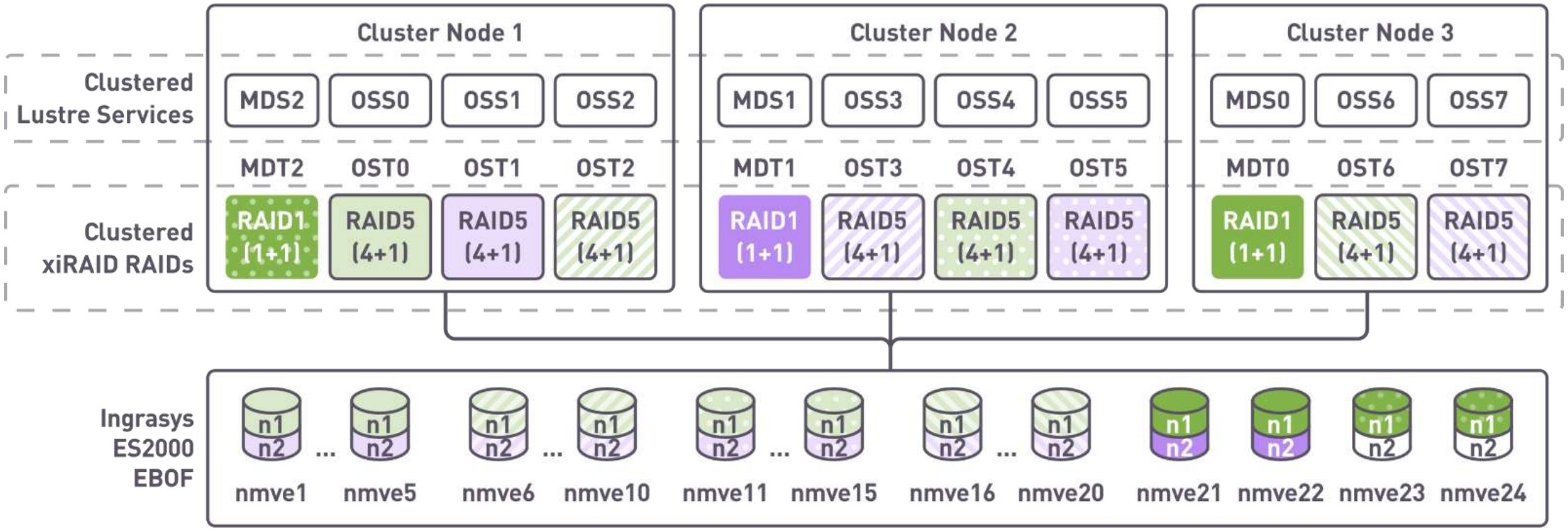


Multinode clusters

Dual-node is not the only HA option

- Typical Lustre clustering configurations = failover pairs (dual node clusters)
- Pacemake+Corosync cluster limit = 16 or 32 cluster nodes
- Using EBOFs, NVMe drives can be shared among multiple nodes
- Lustre (2.15.6) itself has no problems with listing multiple service nodes

Multi-node cluster test environment



mkfs.lustre --servicenode

```
# mkfs.lustre --mdt --fsname=lustre0 --index=2 --servicenode=192.168.45.100@o2ib \
--servicenode=192.168.45.101@o2ib --servicenode=192.168.45.102@o2ib \
--mgsnode=192.168.45.100@o2ib --mgsnode=192.168.45.101@o2ib \
--mgsnode=192.168.45.102@o2ib /dev/xi_r_mdt2
```

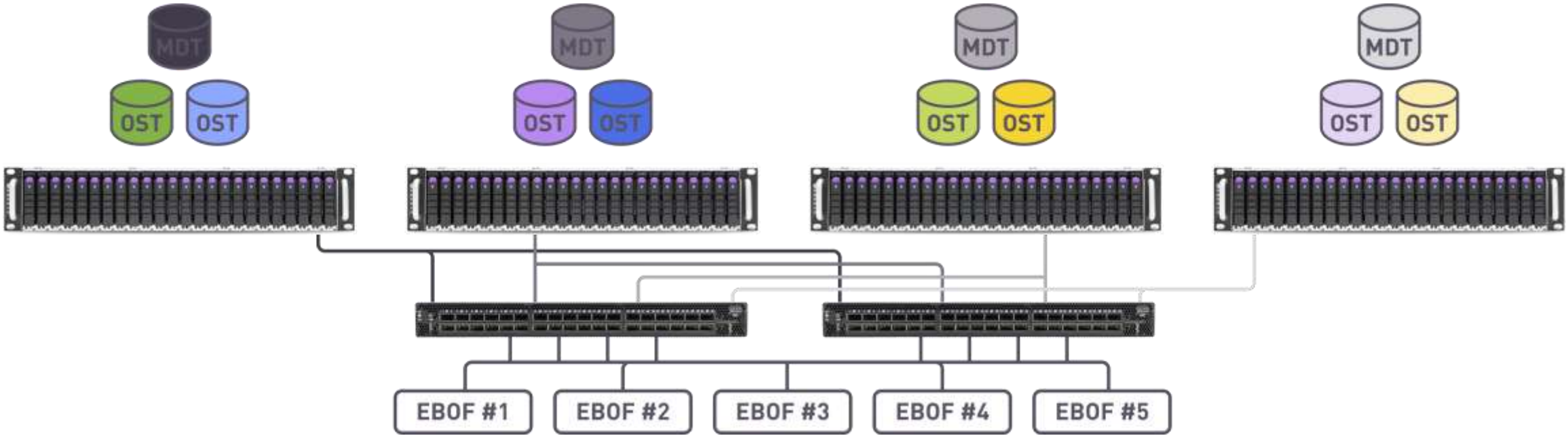
```
# mount -t lustre
192.168.100.100@o2ib:192.168.100.101@o2ib:192.168.100.102@o2ib:/lustre0 /l
```

Multinode cluster with EBOFs

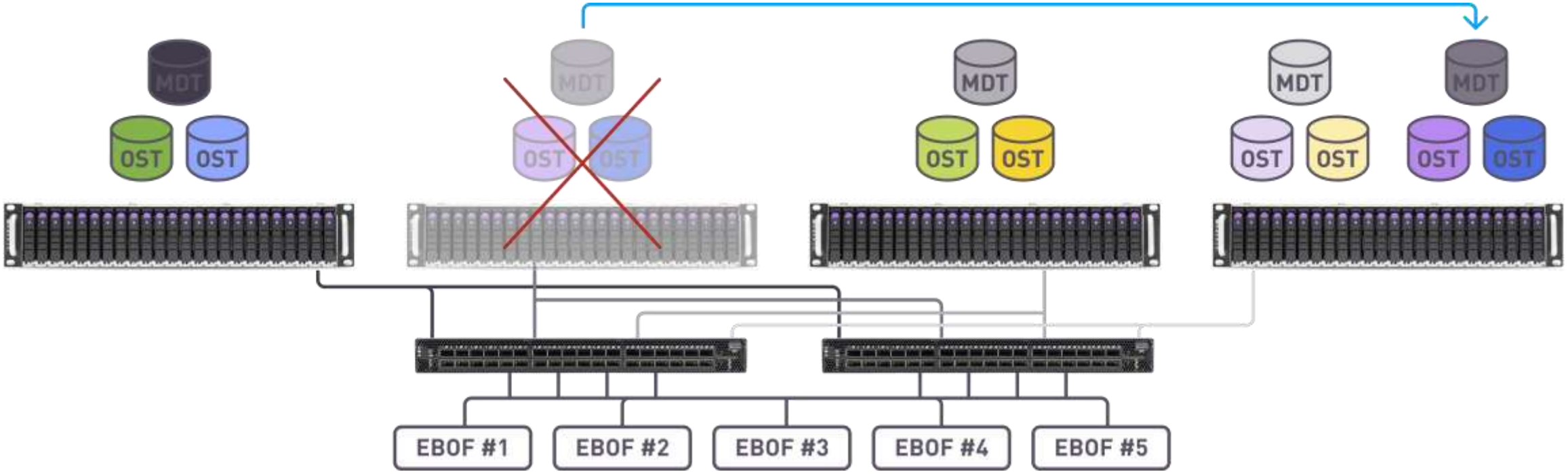
Configuration options:

- Non-dedicated redundant nodes
- Dedicated redundant nodes

Multinode cluster with a non-dedicated redundant node



Multinode cluster with a non-dedicated redundant node (failover)



Multinode cluster with a non-dedicated redundant node

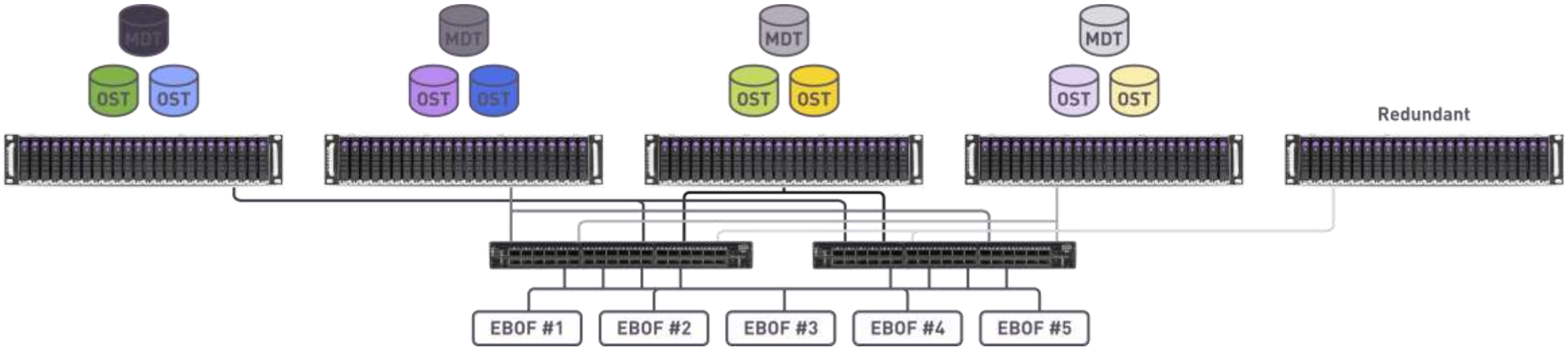
Advantages:

- Generally available hardware
- No additional NVMe configuration needed to get the full performance from the drives
- No hard limit in the number of NVMe drives used in the configuration
- Storage layer performance degradation can be avoided, if EBOF to server network is sized for that

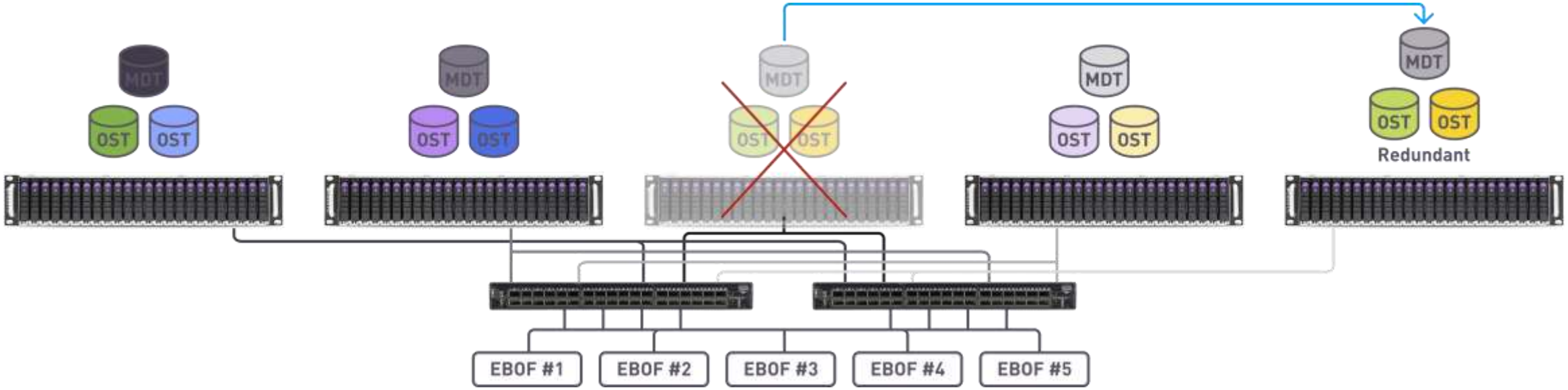
Disadvantages:

- Required redundant network for EBOFs to servers connection
- Increased complexity vs SBB
- High datacenter footprint vs SBB
- To prevent performance degradation in degraded mode, each node must have redundant CPU power, memory and network throughput, which will not be used most of time

Multinode cluster with a dedicated redundant node



Multinode cluster with a dedicated redundant node (failover)



Multinode cluster with a dedicated redundant node

Advantages:

- Generally available hardware
- No additional NVMe configuration needed to get the full performance from the drives
- No hard limit in the number of NVMe drives used in the configuration
- Each server is sized exactly to its workload
- Low redundancy level
- **No performance degradation in case of a server failure**

Disadvantages:

- Required redundant network for EBOFs to servers connection
- Increased complexity vs SBB
- High datacenter footprint vs SBB

Prove it yourself:
xinnor.io