Lustre* HSM
Design Considerations Overview

Scaling capacity and performance without compromise using SGI® DMF™

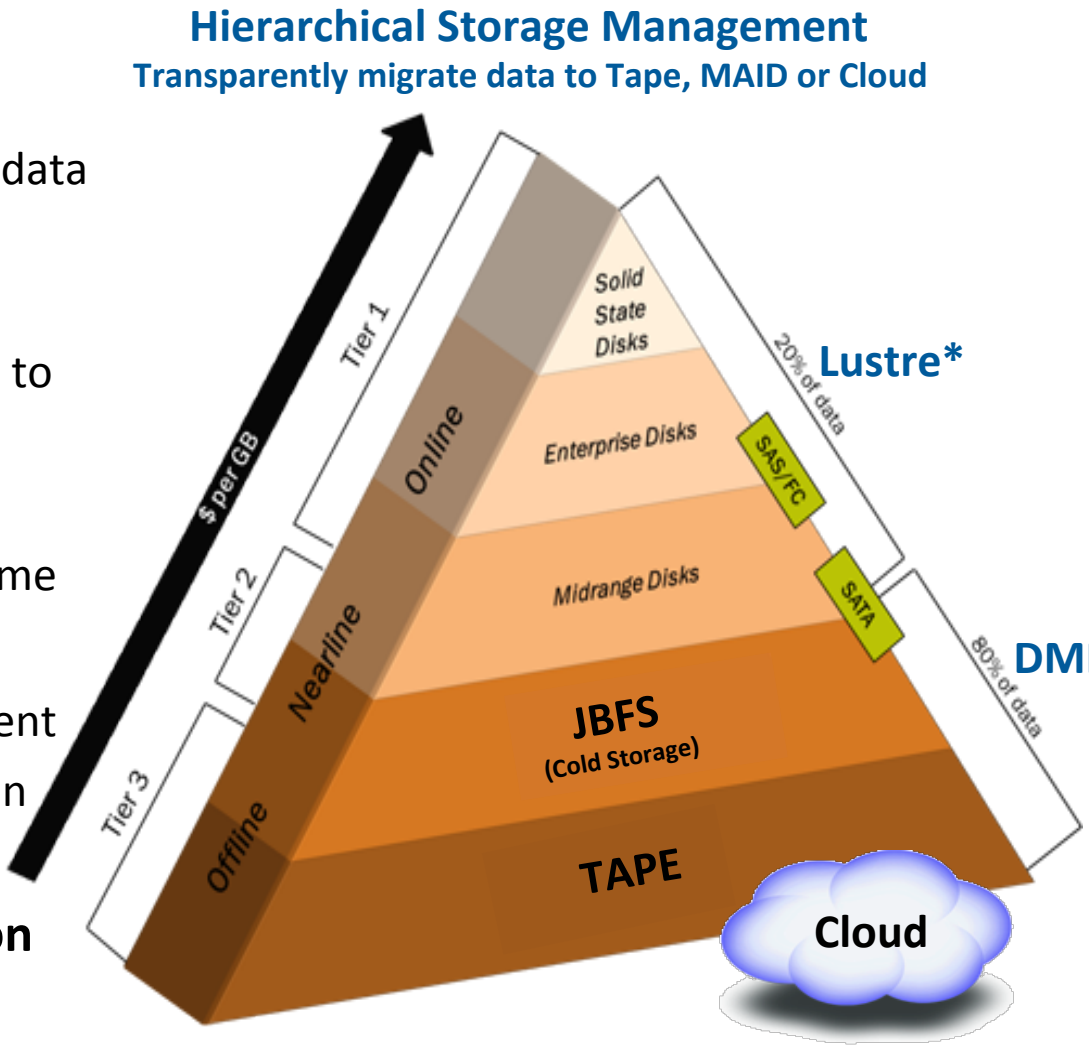
Capacity, Performance & Reliability

Robert Mollard

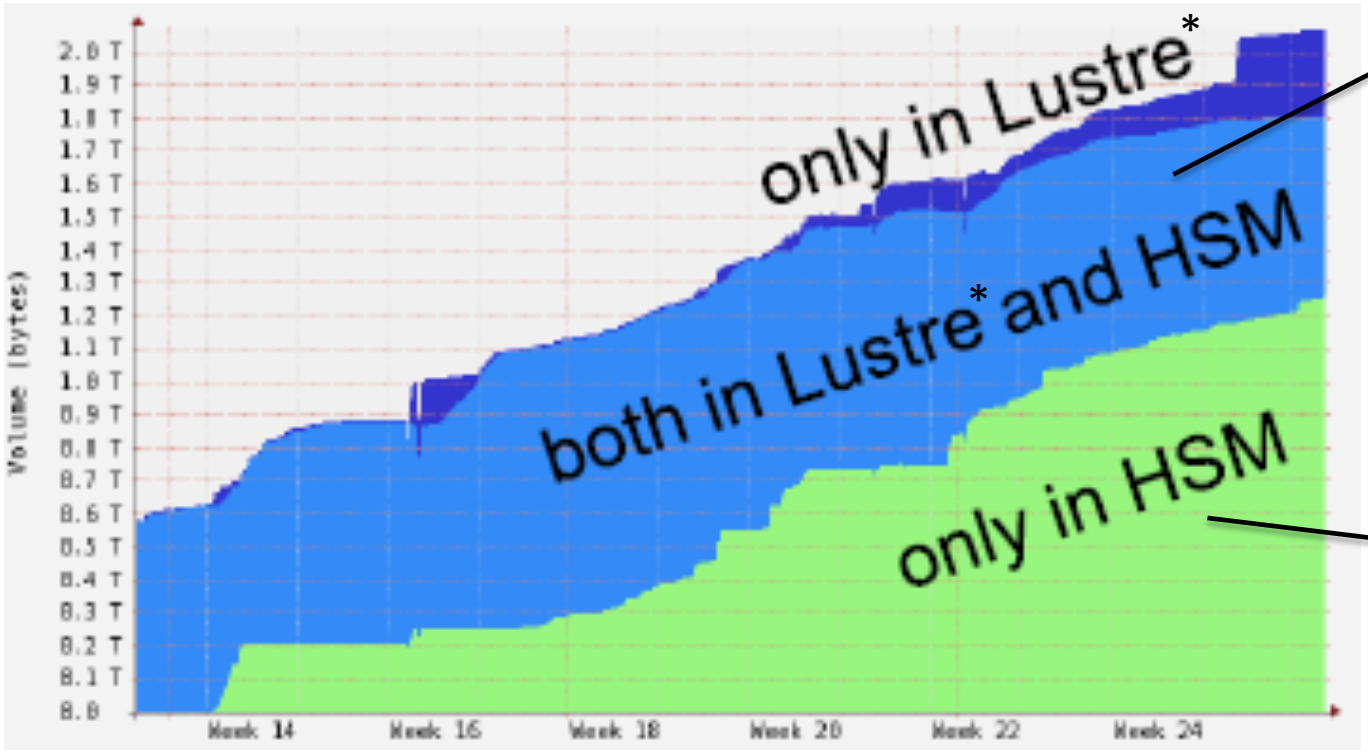
Senior Storage Specialist, APAC

HSM | Data Migration Facility (DMF)

- **Data life cycle management**
  - DMF manages the placement of data within multiple tiers of storage
- **Automated data migration**
  - From expensive, production disk to 2nd or 3rd tier storage
- **Transparent to user**
  - All data appears on line all the time
- **Key Benefits**
  - DMF reduces tier 1 disk investment
  - DMF reduces power consumption
  - DMF protects data long term
- **SGI® DMF™ 25 years in production**



Seamless Tiered Data Management



The most recent and active data is "live" in Lustre* and mirrored within DMF.
ALL DATA APPEARS ONLINE to users.

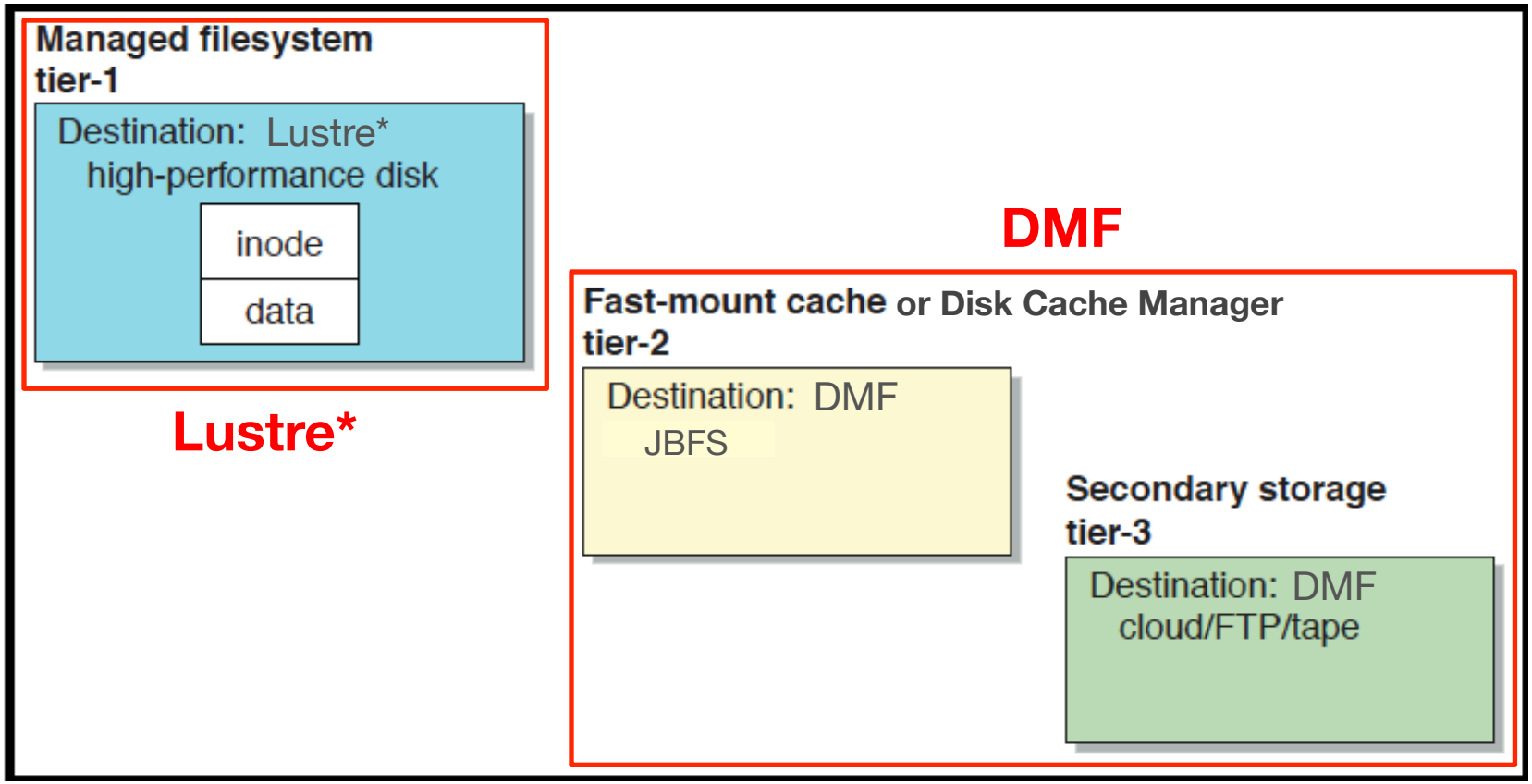
"Overflow" data is stored and protected within DMF on various cost-correct media types

Tiered Data Management

HSM perspective: **regular** file

User perspective: **online** file

Before migrating

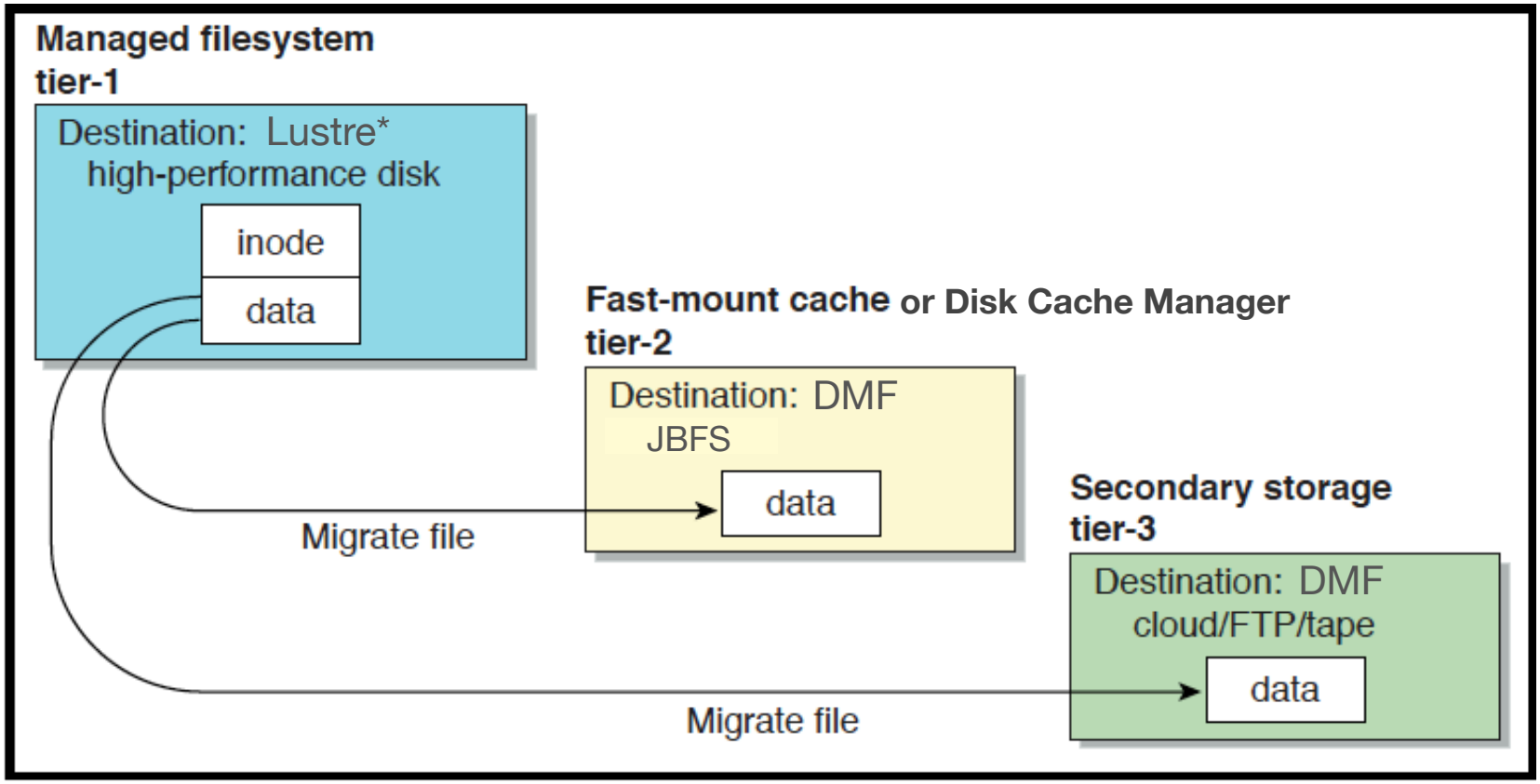


Tiered Data Management

HSM perspective: **dual-state** file

User perspective: **online** file

After migrating

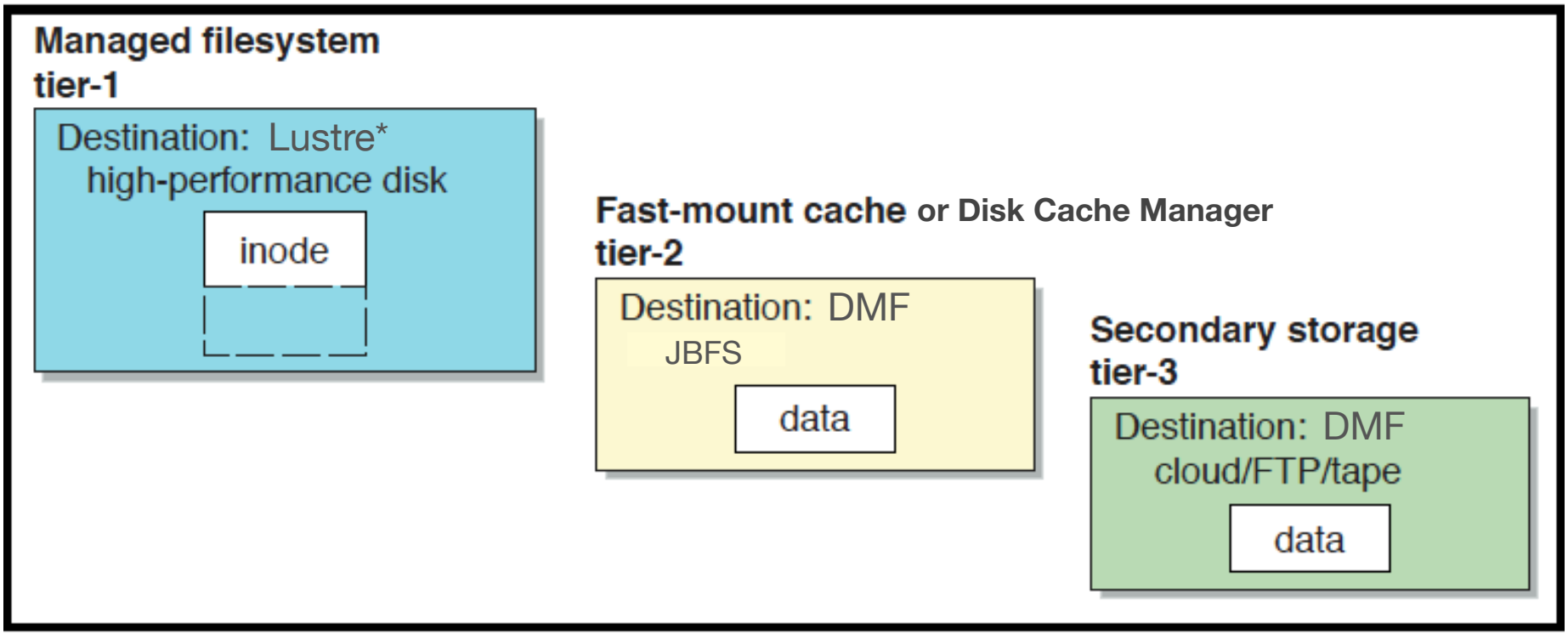


Tiered Data Management

HSM perspective: **offline** file

User perspective: **online** file

After freeing space

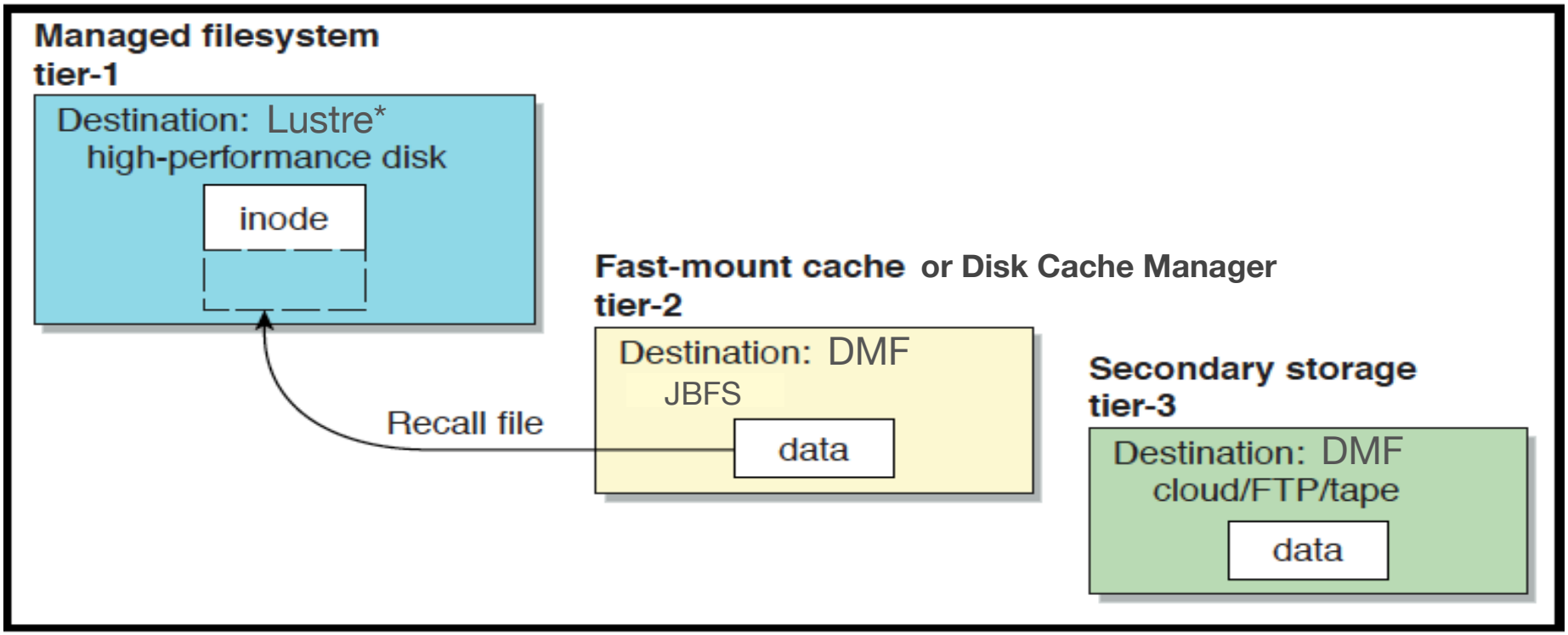


Tiered Data Management

HSM perspective: **unmigrating** file

User perspective: **online** file

Recalling file data from cache

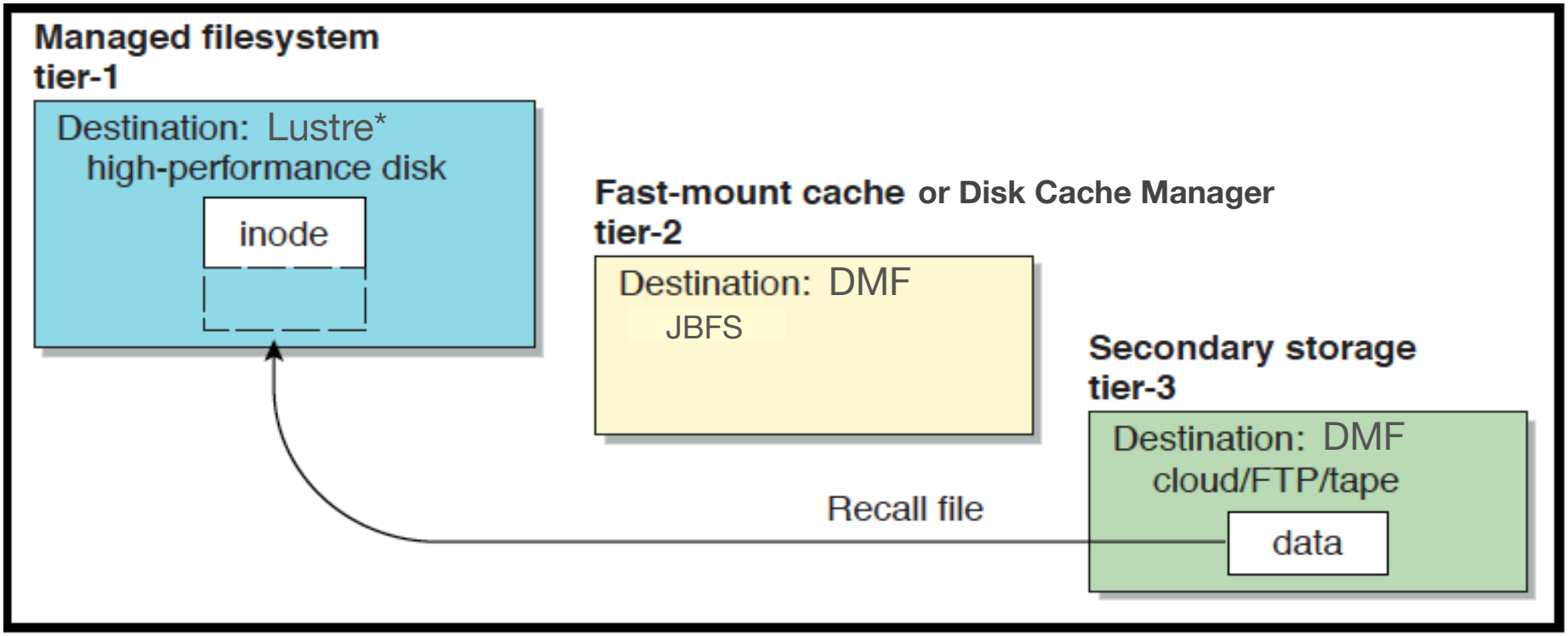


Tiered Data Management

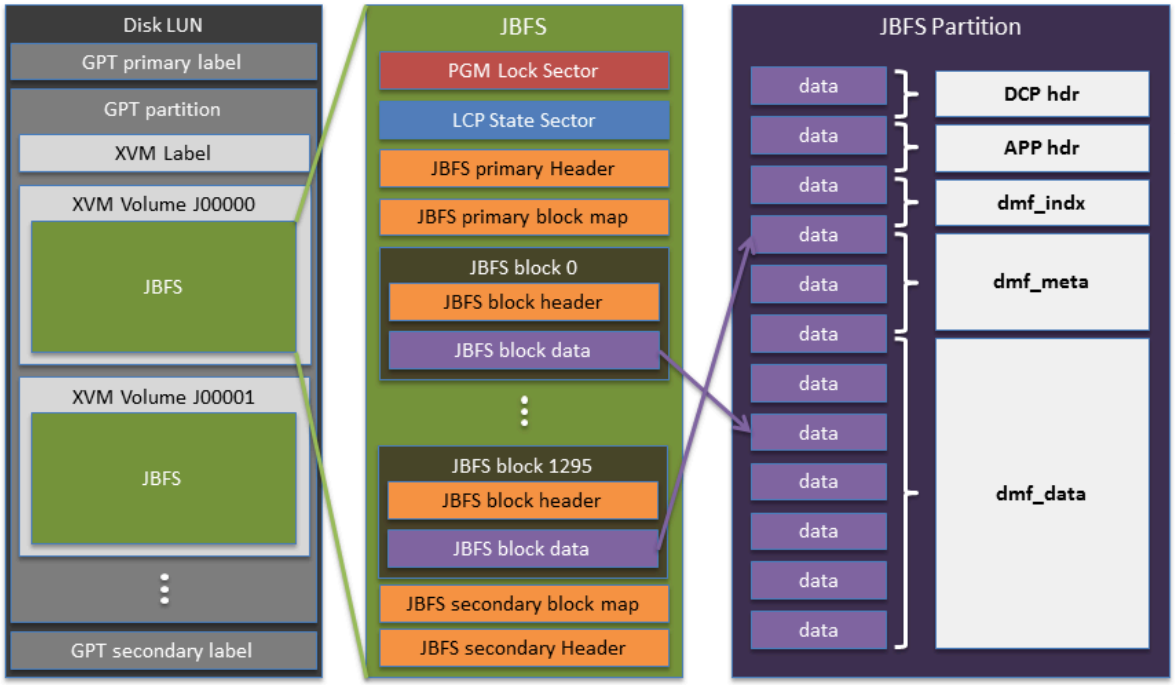
HSM perspective: **unmigrating** file

User perspective: **online** file

Recalling file data from cache

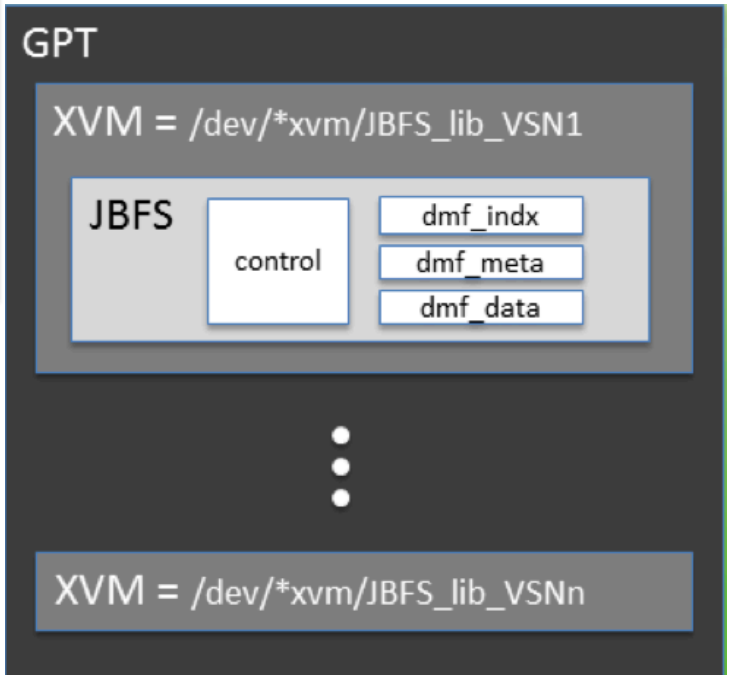


JBFS | Disk Structure



XVM volume name must be "JBFS_{lib}_{PCL}"

- JBFS – Fixed
- {lib} – OpenVault Library Name
- {PCL} – unique 6-character value [0-9A-Z]

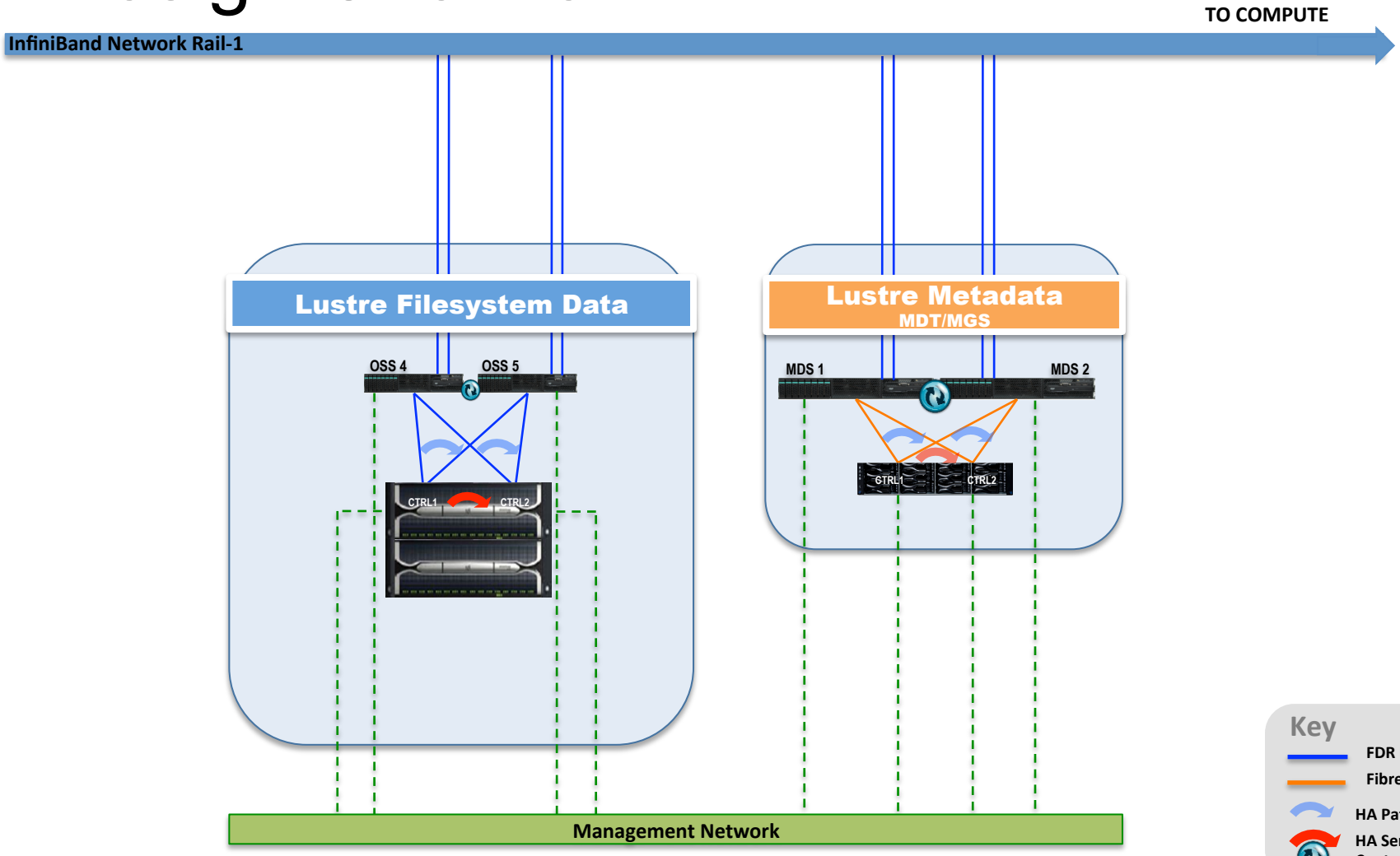


Preparing a disk device for use with JBFS consists of three basic steps:

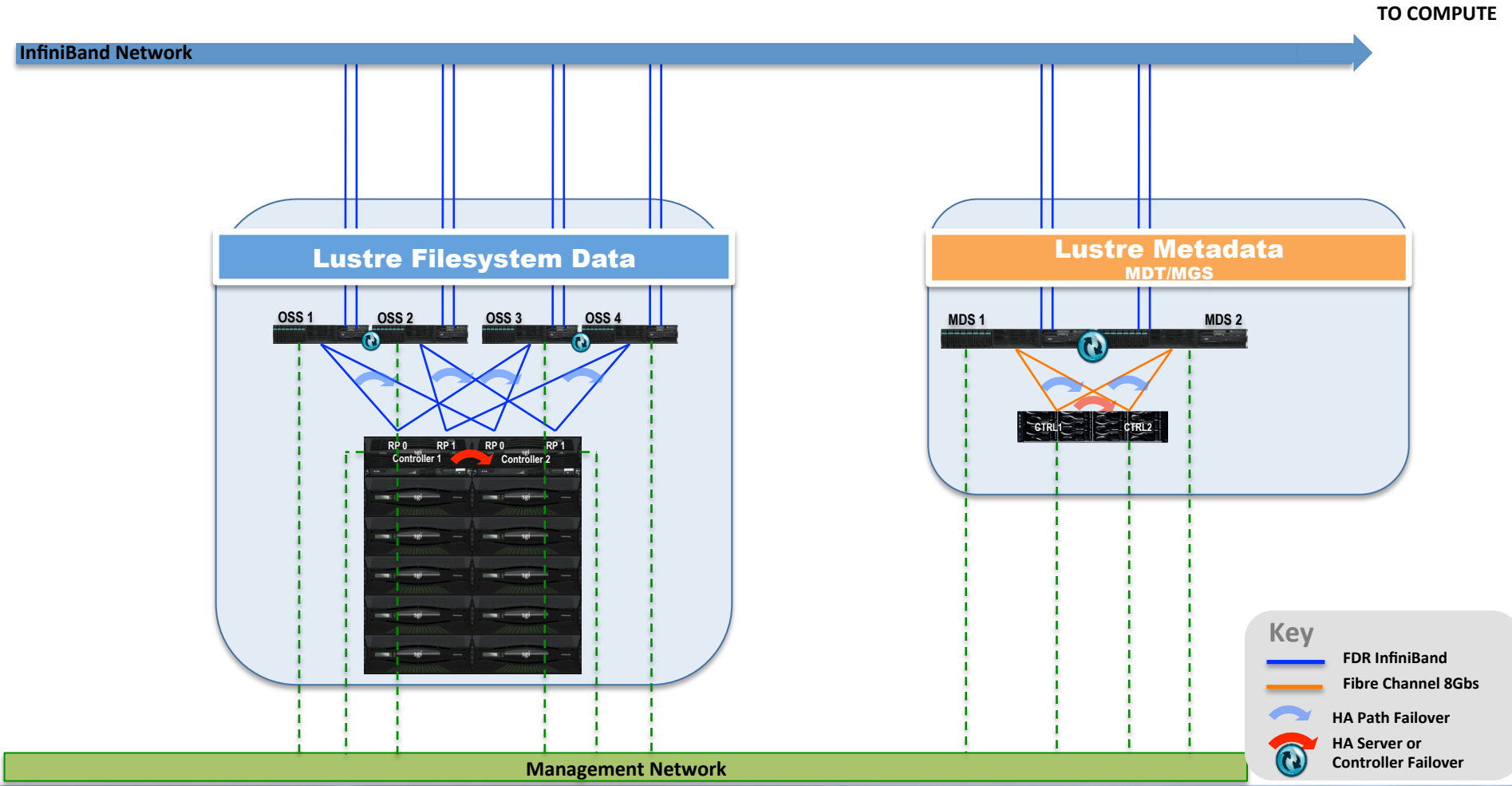
1. Apply the GPT labels
2. Apply the XVM labels
3. Apply the JBFS format

Lustre FS
Design Overview

Lustre FS | Small Building Block Design Overview



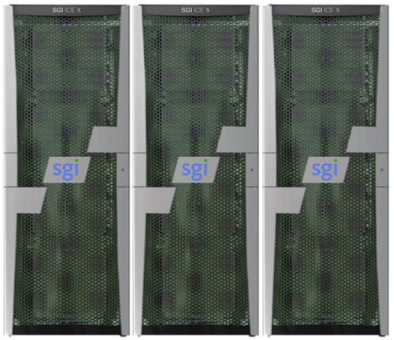
Lustre FS | Large Building Block Design Overview

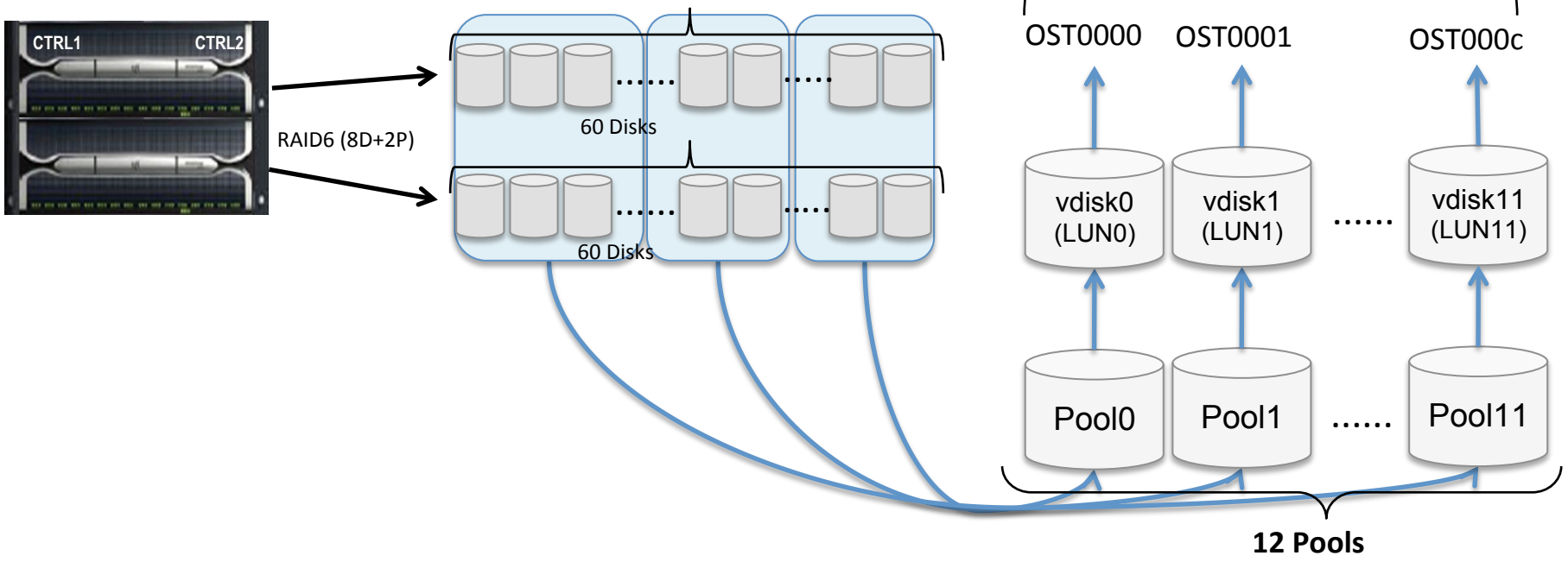


OST | Storage Configuration

Small Building Block

- One LUN (OST) per RAID set

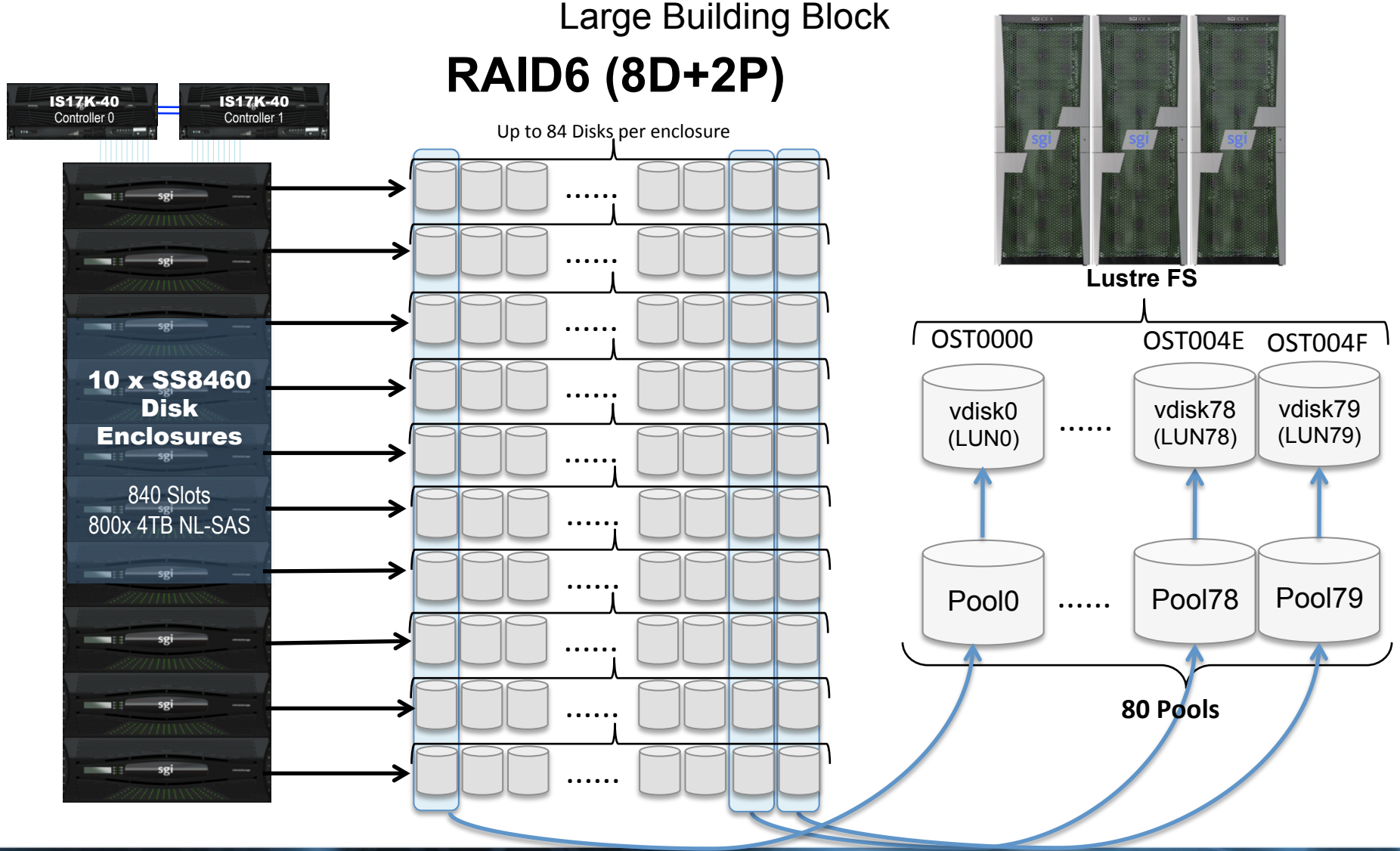




OST | Storage Configuration

Large Building Block

RAID6 (8D+2P)



Lustre HSM

Overview of Lustre HSM

Features:

- Migrate data to and from external storage (HSM)
- Free disk space when needed
- Bring back data on cache-miss
- Policy management (migration, purge, soft rm,...)
- Import from existing backend
- Disaster recovery (restore Lustre filesystem from backend)

Supported HSM Actions

Archive

- Archiving a file means pre-copying a file from Lustre to an external HSM.
- A Copy Tool (“copytool”) reads file content and copies it to an external HSM.
- Once it has been copied, a file is then ready to be released.

Release

- Remove all file data objects.
- Synchronous action which does not involve the copytool nor coordinator.

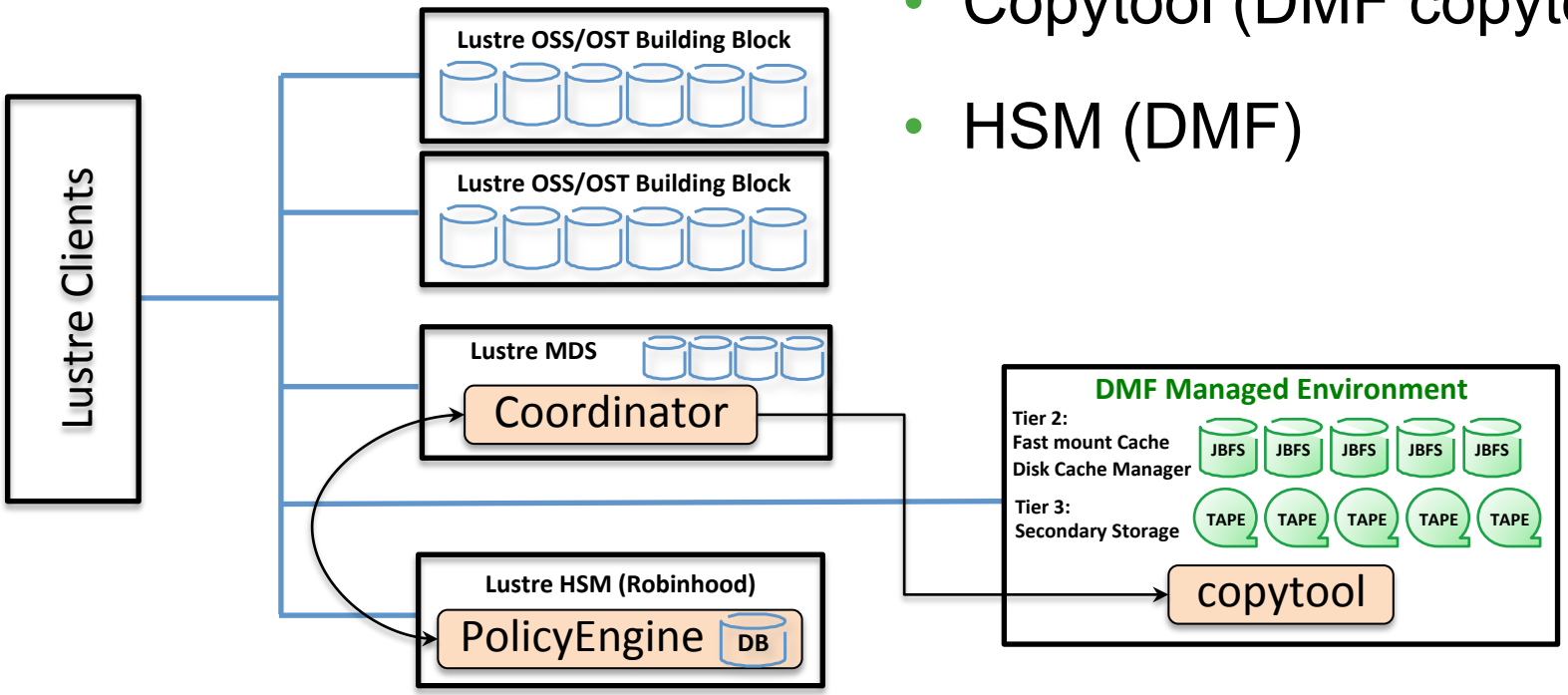
Restore

- All file accesses are blocked until the file is fully restored.
- Copytool will write file data back from an external HSM to Lustre.
- File data accesses are unblocked when the restore is finished.

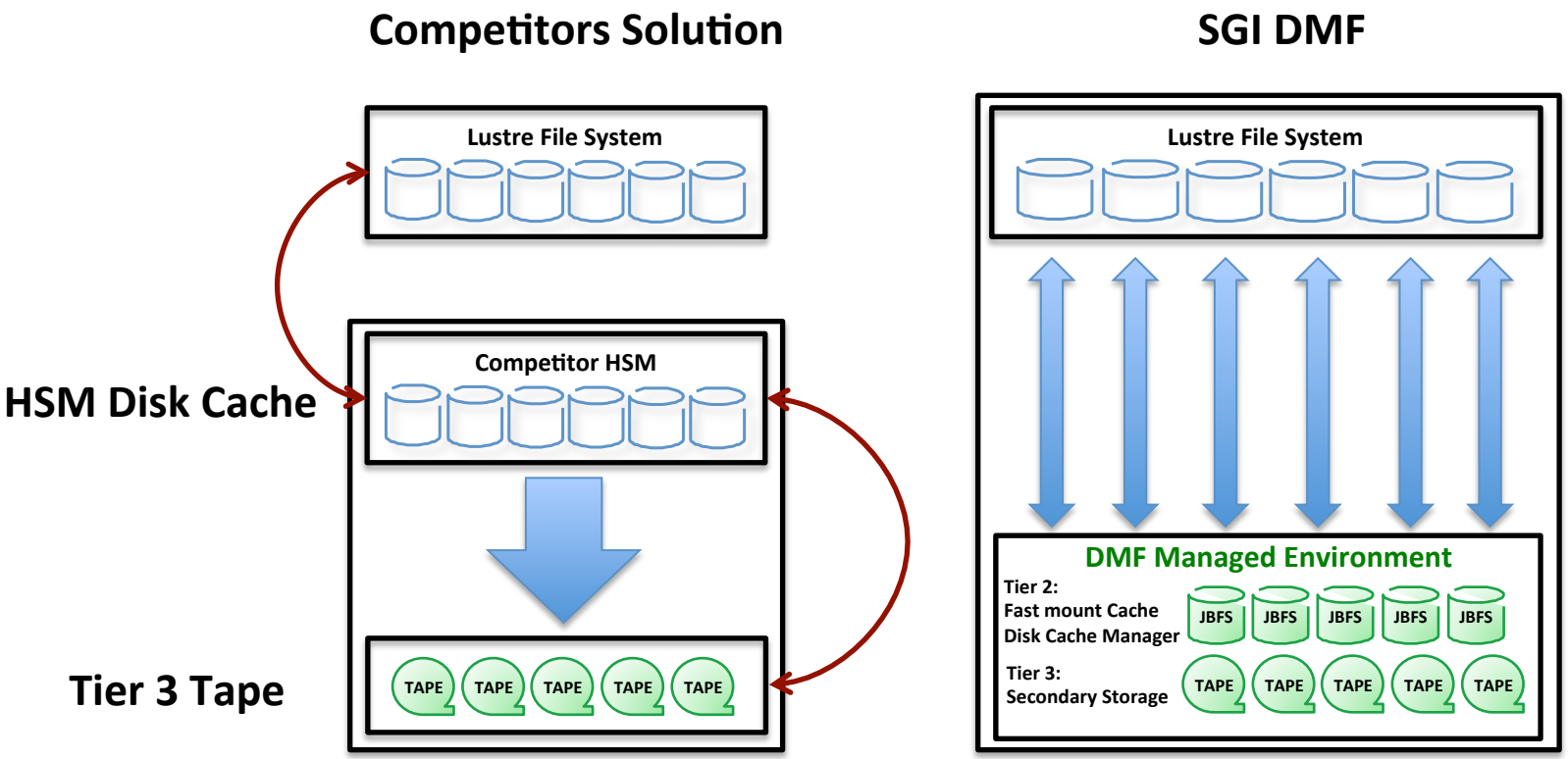
Overview of Lustre HSM

Components

- Coordinator
- Policy Engine (robinhood)
- Copytool (DMF copytool)
- HSM (DMF)



Lustre HSM with DMF direct archiving comparison



Robinhood Policies

Robinhood manages 3 types of policies

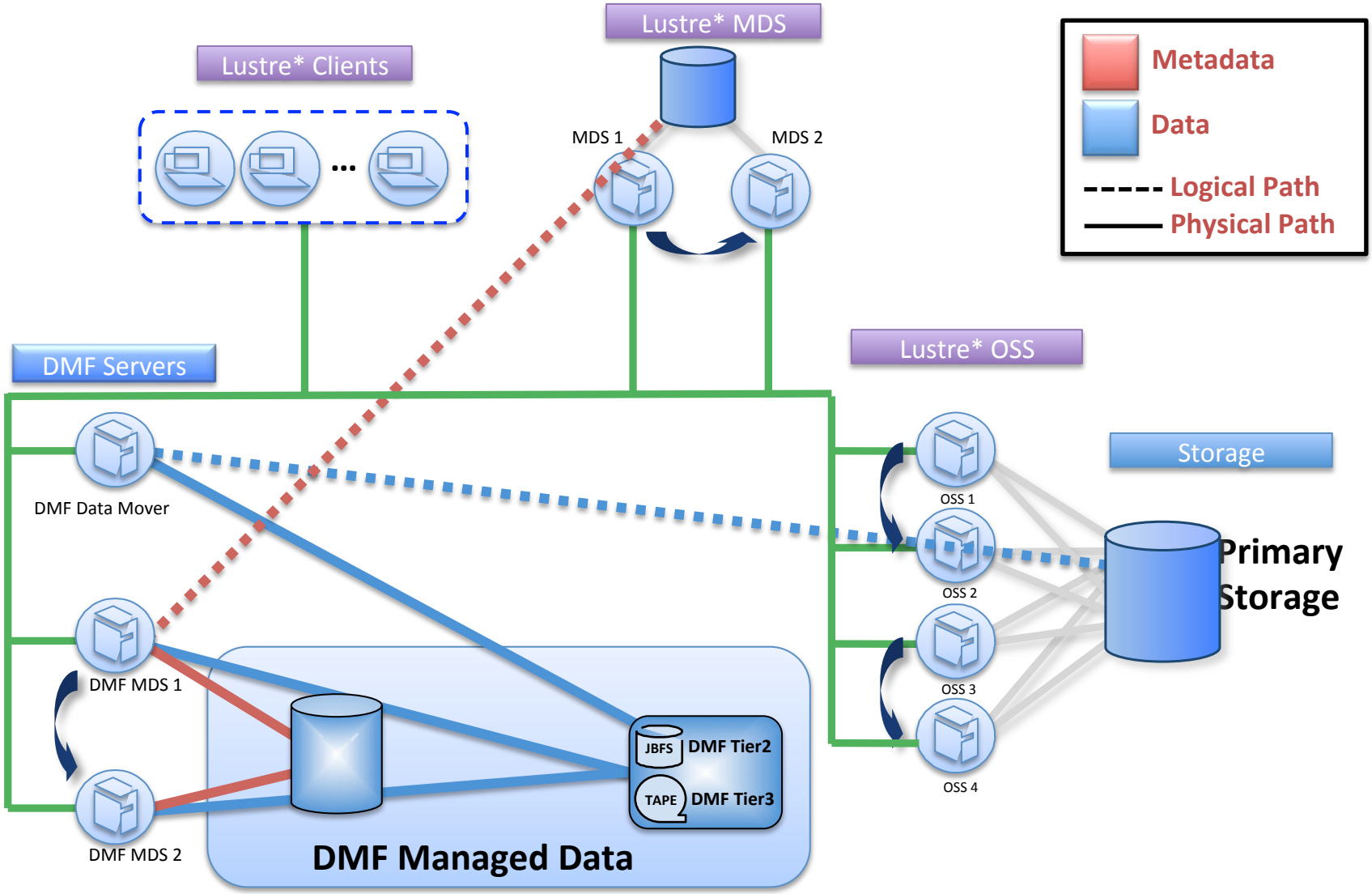
- Migration policy
- Purge policy
- Removal policy

Policies

- File class definitions, associated to policies
- Based on file attributes (path, size, owner, age, xattrs, ...)
- Rules can be combined with boolean operators
- LRU-based migration/purge policies (Least Recently Used)
- Entries can be white-listed

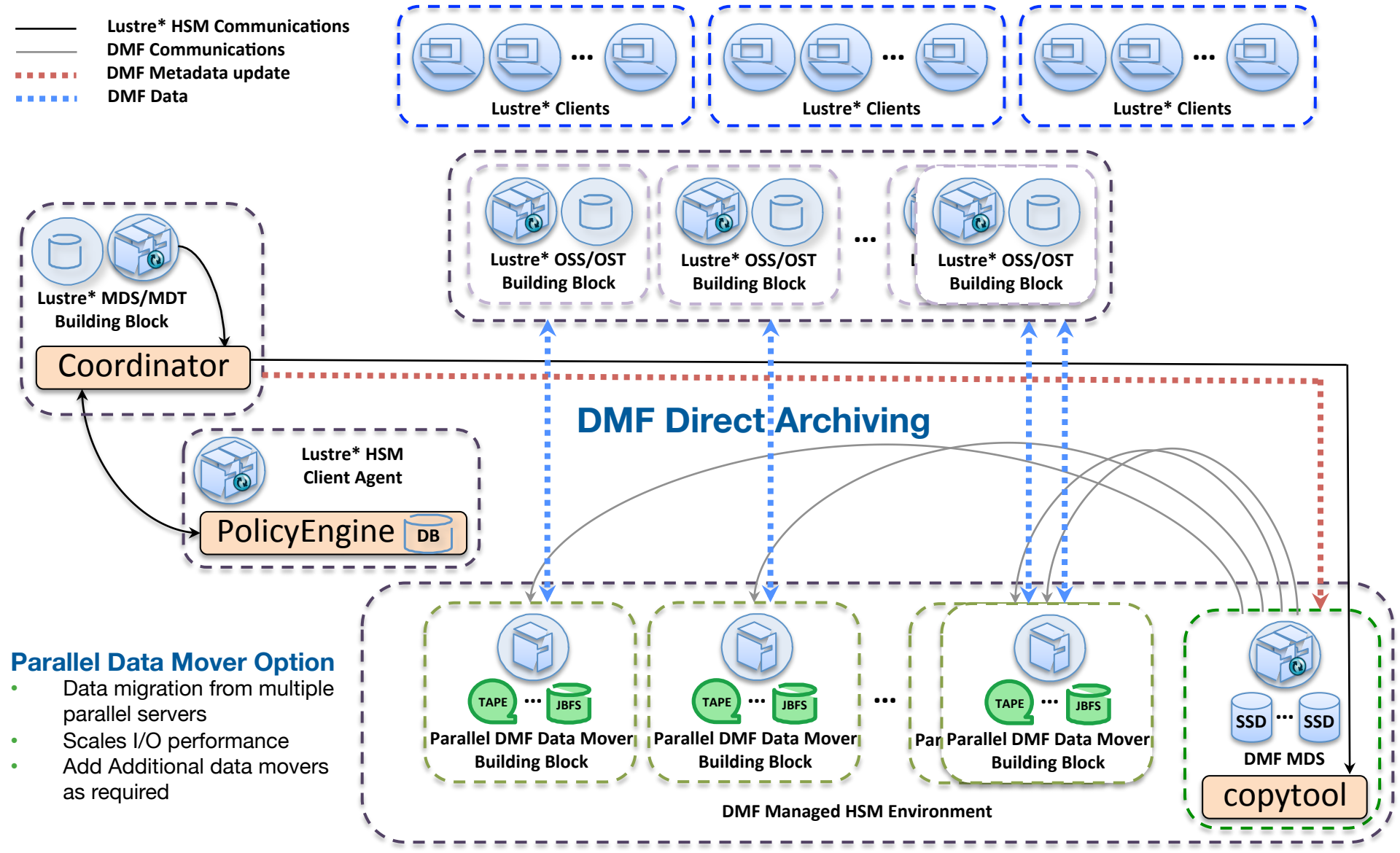
Why consider Lustre HSM?

DMF Direct Archiving | Data Flow



Lustre* HSM | Communication & Data Flow

* = Some names and brands may be claimed as the property of others



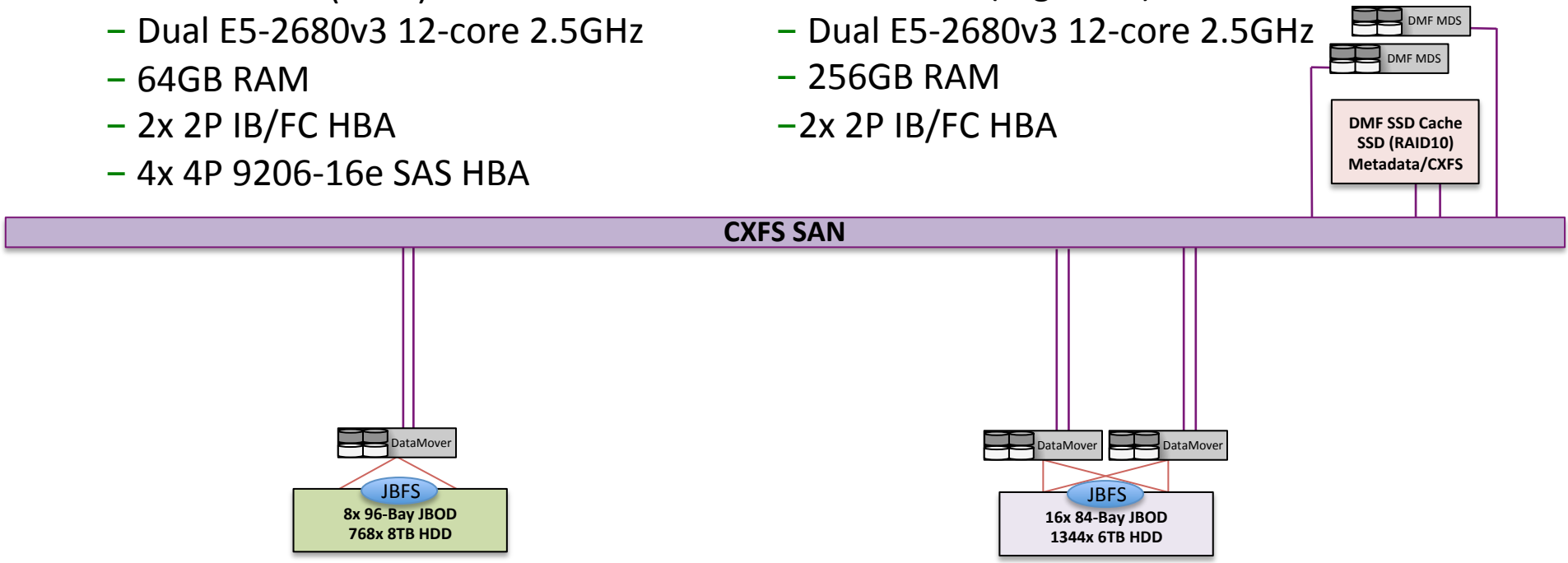
Parallel Data Mover Option

- Data migration from multiple parallel servers
- Scales I/O performance
- Add Additional data movers as required

JBFS | Tier 2

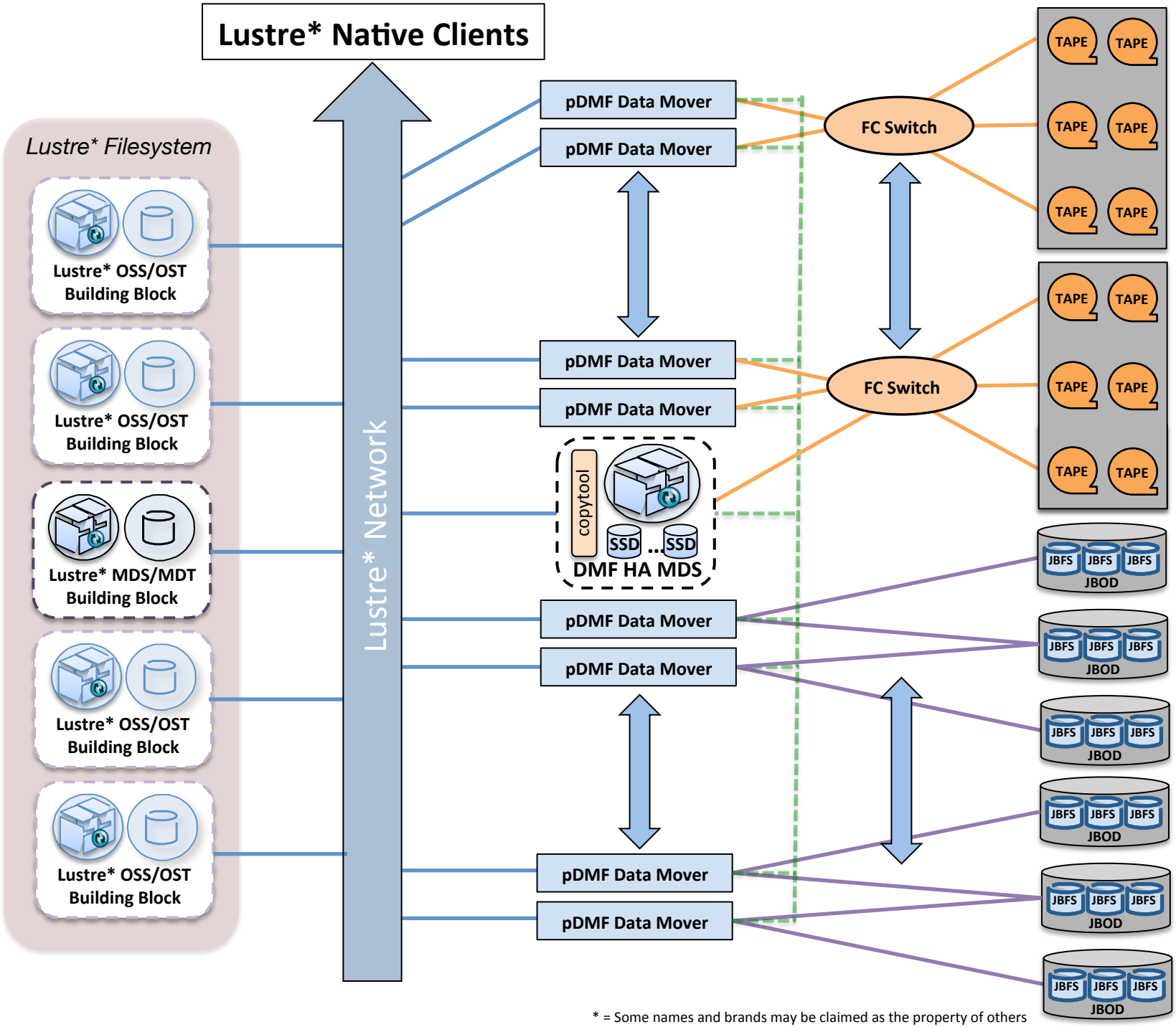
Fast Mount Cache // Disk Cache Manager

- DMF DataMover (Peak) 2U
  - Dual E5-2680v3 12-core 2.5GHz
  - 64GB RAM
  - 2x 2P IB/FC HBA
  - 4x 4P 9206-16e SAS HBA
- DMF/CXFS MDS (Highland)
  - Dual E5-2680v3 12-core 2.5GHz
  - 256GB RAM
  - 2x 2P IB/FC HBA



8x JBOD Direct connect
1x JBOD Daisy Chain (performance limited)

16x JBOD Direct connect
2x JBOD Daisy Chain


Lustre* Native Clients
Lustre* Filesystem
Lustre* OSS/OST Building Block
Lustre* OSS/OST Building Block
Lustre* MDS/MDT Building Block
Lustre* OSS/OST Building Block
Lustre* OSS/OST Building Block
Lustre* Network
pDMF Data Mover
pDMF Data Mover
pDMF Data Mover
pDMF Data Mover
copytool
SSD
SSD
DMF HA MDS
pDMF Data Mover
pDMF Data Mover
pDMF Data Mover
pDMF Data Mover
FC Switch
FC Switch
TAPE
JBFS
JBOD
* = Some names and brands may be claimed as the property of others

Key Points

- New Lustre* and DMF features allow cost effective scalability without compromising performance
- SGI DMF provides a high performance parallel HSM for Lustre* with direct archiving to tier 2/3 storage targets
- SGI DMF – JBFS delivers a tier 2 fast mount cache or disk cache manager with built in power management capabilities$
- The Result:
  - Cost effective capacity, reduced TCO (low cost/power storage tiers)
  - Proven long-term data protection (DMF – 25 years in production)
  - Improved operational procedures (simplified access to data)
  - Scalable performance within archive tiers (parallel DMF)

$ = on supported hardware

Questions & Responses

Robert Mollard
Senior Storage Specialist, Asia Pacific
rmollard@sgi.com

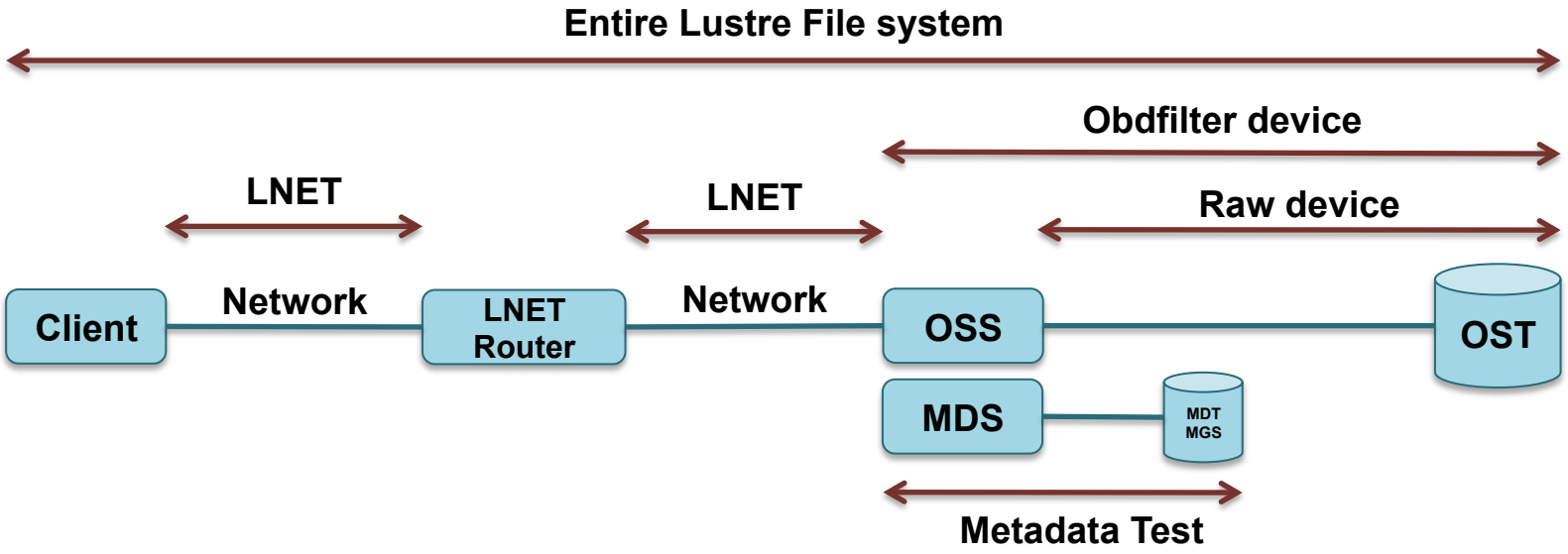
Extra Slides

Lustre Benchmarking

Lustre Benchmark Approach

There are a couple of layers for the Lustre benchmarking, in order to get maximum Lustre performance, the measurement and analysis of the following is crucial:

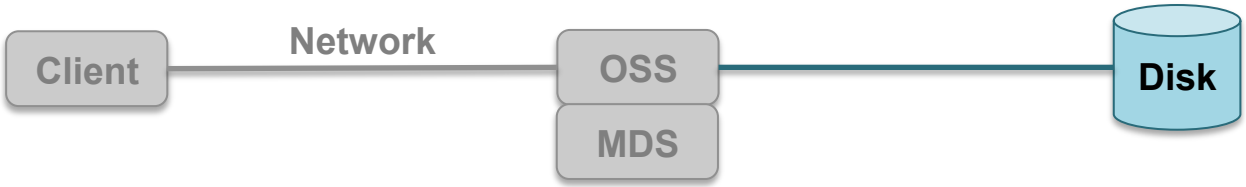
1. Raw device (server to disk device without the file system, sgp_dd-survey & xdd)
2. Obdfilter device (Lustre backend device, OSS <-> OST)
3. LNET (OSS <-> Client without disk)
4. IO throughput - Entire Lustre File system (OSS/OST <-> Client)
5. MetaData testing (mdtest & fdtree)



Raw device

Bare metal storage performance

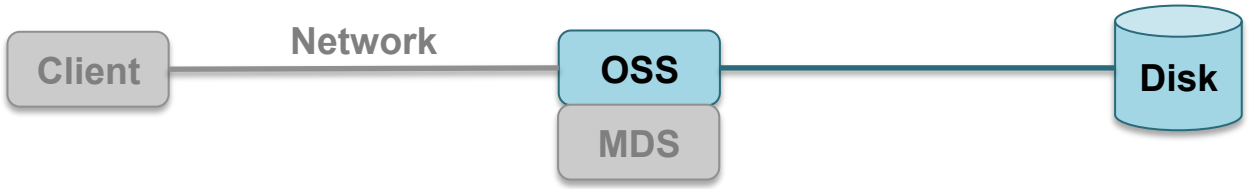
1. Make sure storage controller's peak performance
2. sgp_dd, xdd are major tools for measuring raw performance (standard dd to block device is not sufficient and can be misleading)
3. sgp_dd-survey is based on sgp-dd and is part of Lustre
4. xdd works on the multiple nodes simultaneously which is important to understand aggregate raw performance



Obdfilter device (Lustre backend device)

obdfilter-survey is major tool to test obdfilter performance

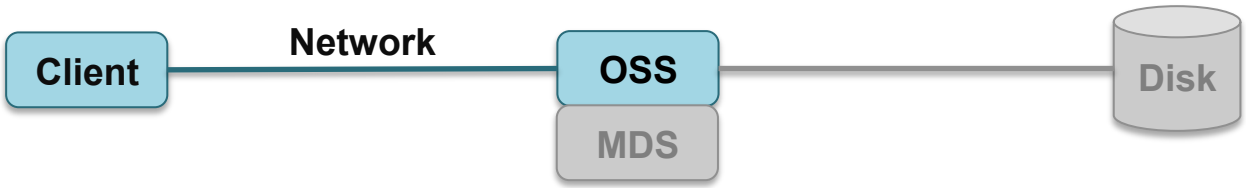
1. Benchmark runs on OSSs without Lustre clients
2. Results show Lustre backend performance
3. obdfilter-survey is a part of Lustre
4. obdfilter-survey works on multiple OSSs simultaneously



LNET(Lustre Network)

LNET Self-Test

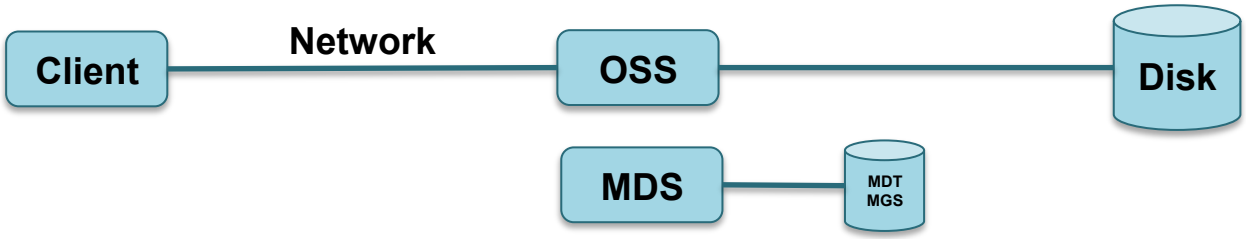
1. Benchmark on between Lustre clients and OSSs with Lustre protocols
2. Runs over LNET and LNDs(Infiniband, 10GbE, 1GbE, etc)
3. Lnet-selftest is part of Lustre. (kernel modules and user-land utilities)
4. Also useful for regression tests, performance testing and verification of Lustre Network layer



Entire Lustre file system

Once the clients mount the Lustre file system, then general I/O benchmark tools work well.

1. **I/O Throughput**
   - IOR
   - IOZONE
2. **Metadata**
   - mdtest
   - fdtree
   - etc..

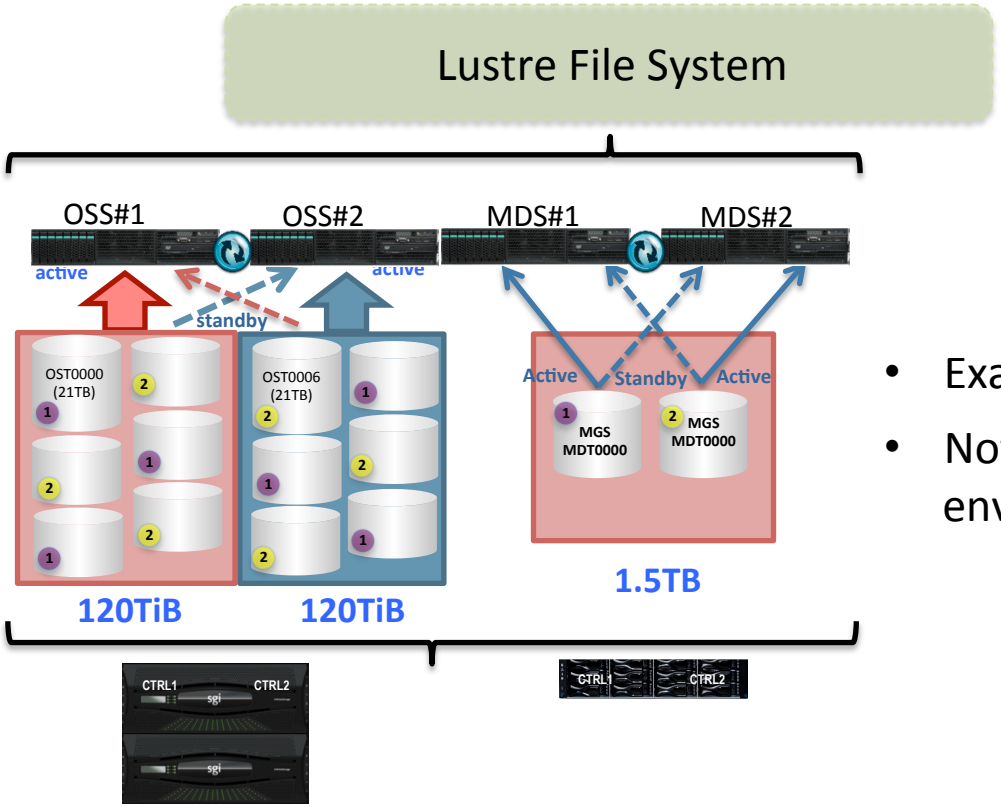


Extra Slides

Lustre Configuration Options

Lustre FS Configuration

Dual file system balanced setup



- Example of Dual file system serving.
- Not recommended for performance environments.

Key

- Active Path
- Standby Path
- 1 Scratch File System
- 2 Home File System

- Balanced OSTs to OSS to ensure performance and failover

CIFS Gateway for Lustre