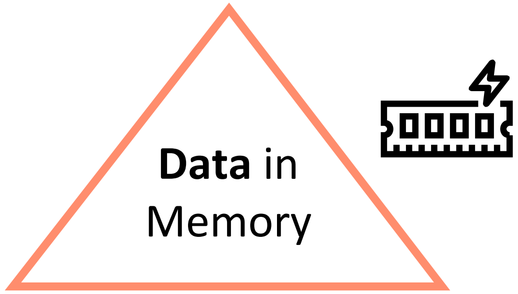
Data Movement & Tiering with DMF 7

Kirill Malkin
Director of Engineering

April 2019

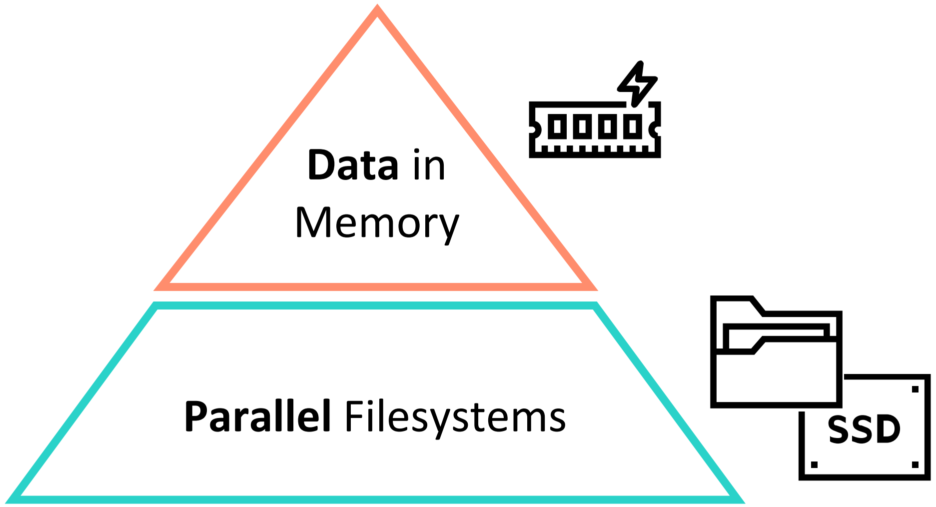
Why Move or Tier Data?

- We wish we could keep everything in DRAM, but…
  - It's volatile
  - It's expensive



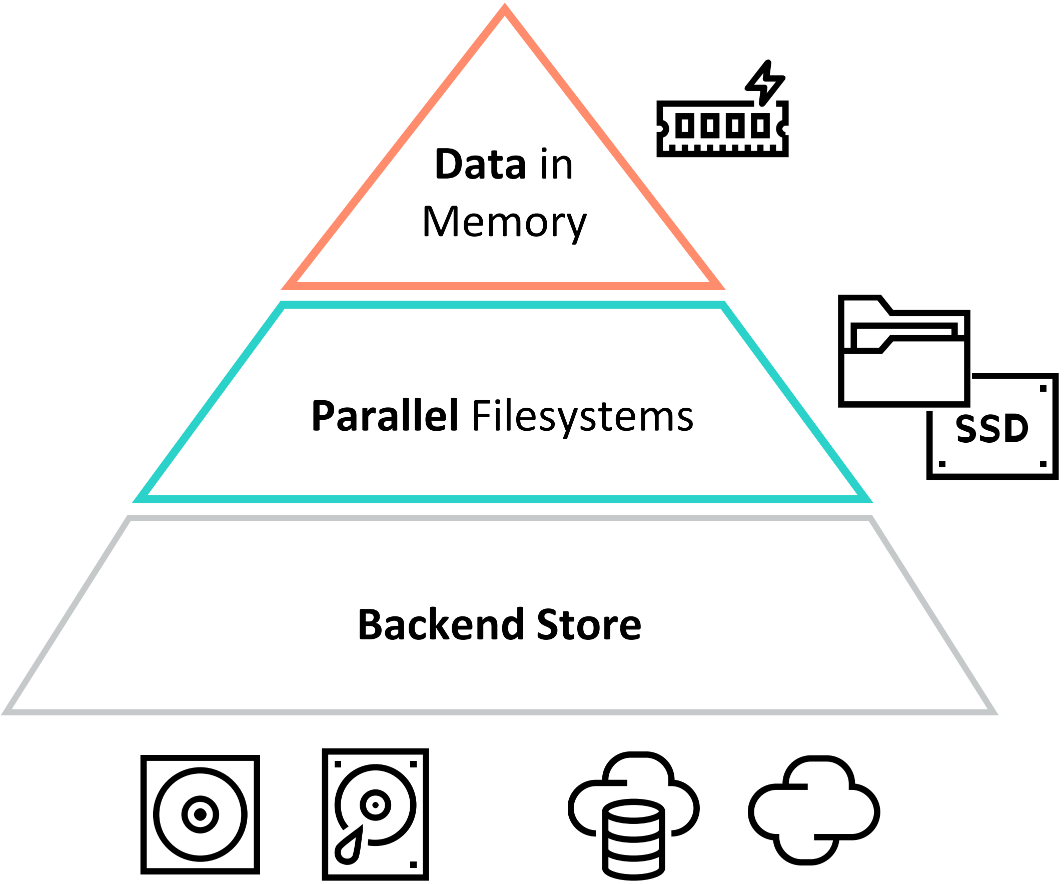
Why Move or Tier Data?

- We wish we could keep everything in DRAM, but…
  - It's volatile
  - It's expensive
- So we need to move data to and from non-volatile medium
  - Solid State or Magnetic
  - Make copies: snapshot, backup, archive



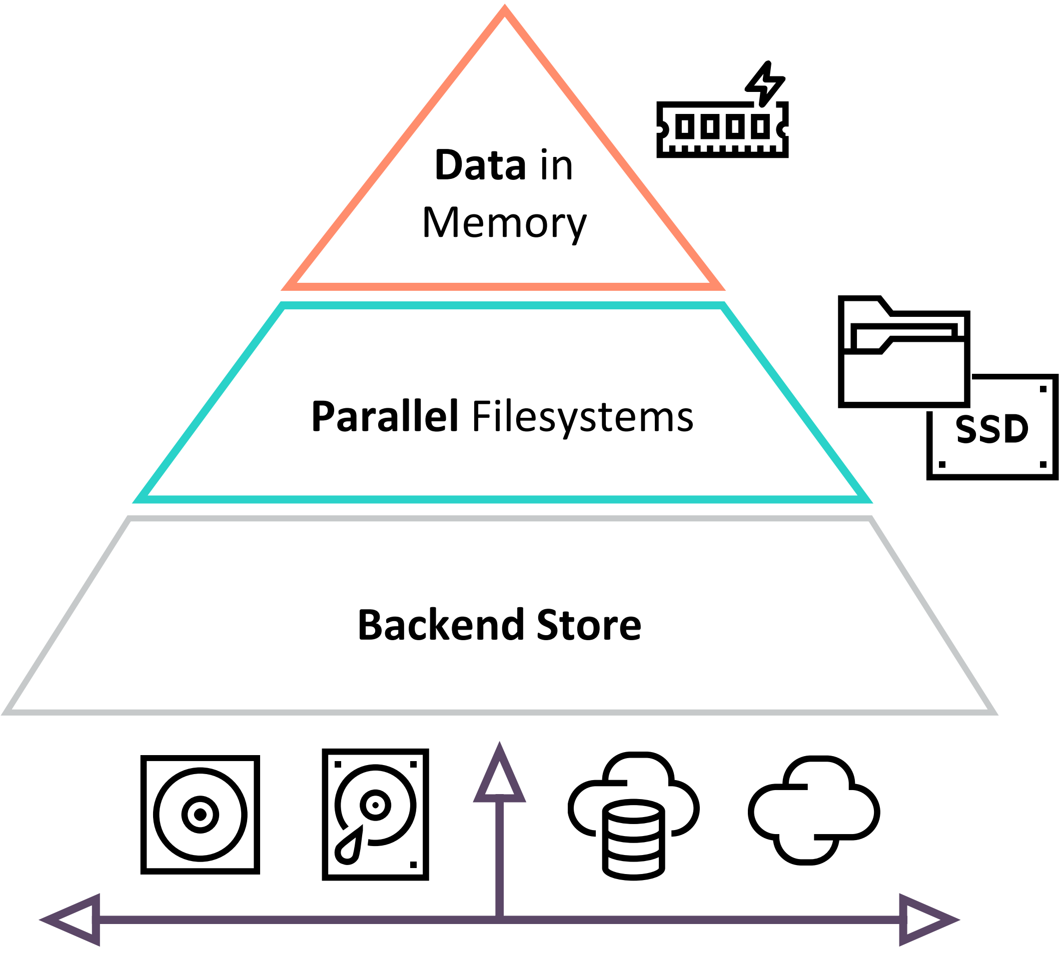
Why Move or Tier Data?

- We wish we could keep everything in DRAM, but…
  - It's volatile
  - It's expensive
- So we need to move data to and from non-volatile medium
  - Solid State or Magnetic
  - Make copies: snapshot, backup, archive
- When we move data to less expensive medium it's called tiering
  - Solid State to Hard Disk to Cloud to Tape



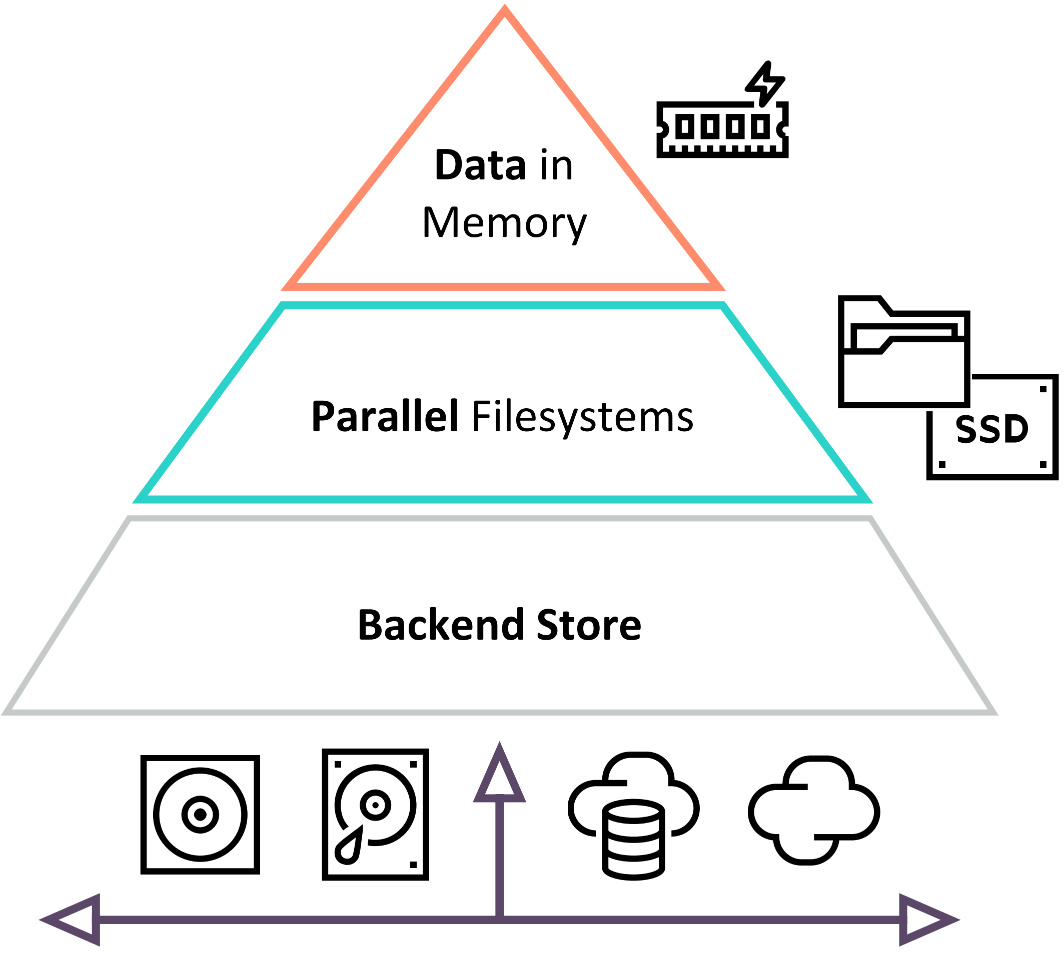
Why Move or Tier Data?

- We wish we could keep everything in DRAM, but…
  - It's volatile
  - It's expensive
- So we need to move data to and from non-volatile medium
  - Solid State or Magnetic
  - Make copies: snapshot, backup, archive
- When we move data to less expensive medium it's called tiering
  - Solid State to Hard Disk to Cloud to Tape
- We also move data because of locality
  - Different compute system
  - Another data center or another organization
  - Computing at Edge



Why Move or Tier Data?

- We wish we could keep everything in DRAM, but…
  - It’s volatile
  - It’s expensive
- So we need to move data to and from non-volatile medium
  - Solid State or Magnetic
  - Make copies: snapshot, backup, archive
- When we move data to less expensive medium it’s called tiering
  - Solid State to Hard Disk to Cloud to Tape
- We also move data because of locality
  - Different compute system
  - Another data center or another organization
  - Computing at Edge
- **Doing this well requires Data Management**



Pushing Compute to Edge

Industry Trend

- Moving compute closer to where data is
  - Transferring large amounts of data produced by IoT devices is too expensive
  - Decision making must occur in real time, transfers take too long
  - AI, Big Data and HPC processing gravitates to mini-datacenters at Edge

Pushing Compute to Edge
Industry Trend

– Moving compute closer to where data is
  – Transferring large amounts of data produced by IoT devices is too expensive
  – Decision making must occur in real time, transfers take too long
  – AI, Big Data and HPC processing gravitates to mini-datacenters at Edge

– What happens to data & results produced at Edge?
  – Valuable IoT data as well as results should be preserved
  – Typically this means copying to one or more locations
  – Key to organizing and optimizing this process is distributed metadata
  – Workflow managers and users query metadata & schedule data movement in dormant form

Pushing Compute to Edge
Industry Trend

- Moving compute closer to where data is
  - Transferring large amounts of data produced by IoT devices is too expensive
  - Decision making must occur in real time, transfers take too long
  - AI, Big Data and HPC processing gravitates to mini-datacenters at Edge
- What happens to data & results produced at Edge?
  - Valuable IoT data as well as results should be preserved
  - Typically this means copying to one or more locations
  - Key to organizing and optimizing this process is distributed metadata
  - Workflow managers and users query metadata & schedule data movement in dormant form
- POSIX is still dominant access method in HPC
  - Typically, not 100% compliant – consistency optimized for performance
  - Significant dependency of codes on POSIX semantics
- Non-HPC applications store data differently
  - Buckets of objects in cloud, or S3-API
  - Emergence of Data Lakes

Pushing Compute to Edge
Industry Trend

- Moving compute closer to where data is
  - Transferring large amounts of data produced by IoT devices is too expensive
  - Decision making must occur in real time, transfers take too long
  - AI, Big Data and HPC processing gravitates to mini-datacenters at Edge
- What happens to data & results produced at Edge?
  - Valuable IoT data as well as results should be preserved
  - Typically this means copying to one or more locations
  - Key to organizing and optimizing this process is distributed metadata
  - Workflow managers and users query metadata & schedule data movement in dormant form
- POSIX is still dominant access method in HPC
  - Typically, not 100% compliant – consistency optimized for performance
  - Significant dependency of codes on POSIX semantics
- Non-HPC applications store data differently
  - Buckets of objects in cloud, or S3-API
  - Emergence of Data Lakes
- Moving POSIX data
  - Better done in dormant form where it is immutable
  - Once local to data center, data can be staged as POSIX and computed on

Data Management

What Challenges does it Solve?

Too much data

- 1PB or more of unstructured file data
- Need simple & cost-effective storage or backup solution

Too many files

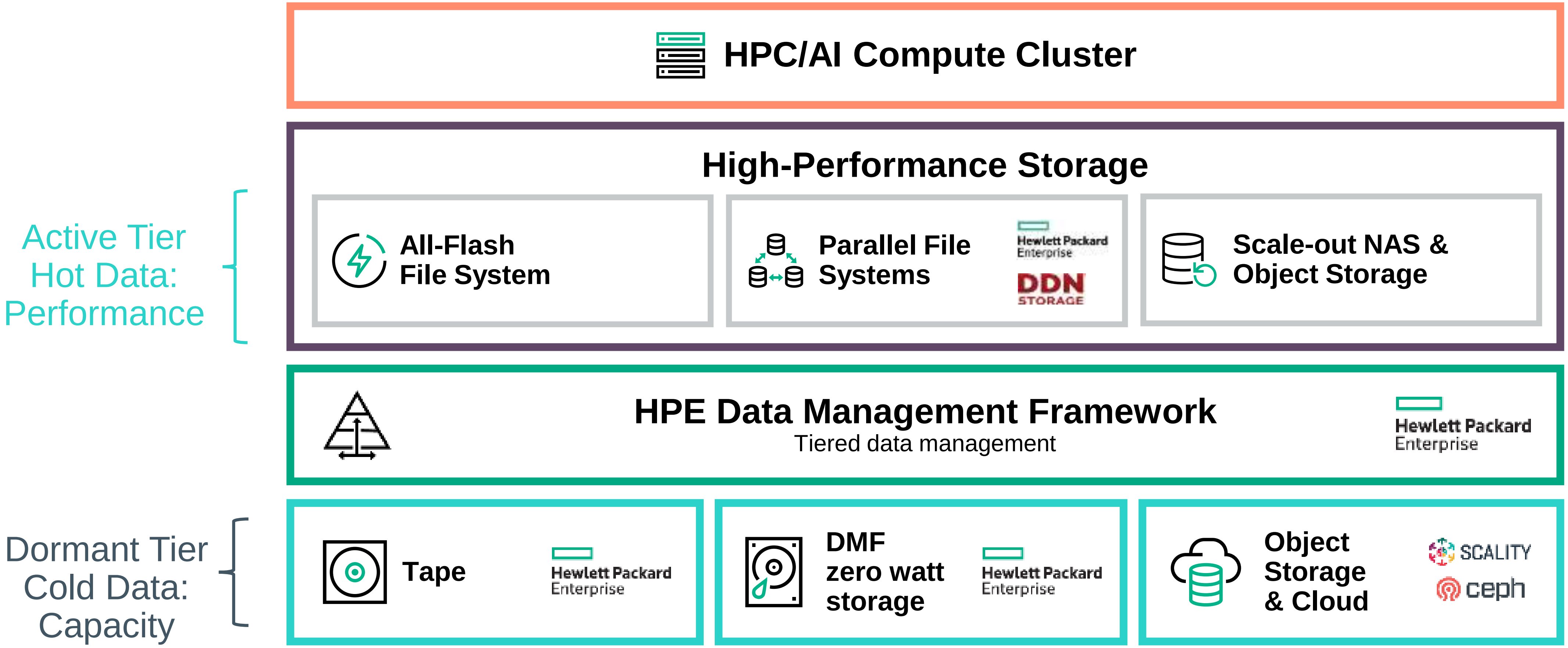
- Billions of files that require periodic movement
- Locate & construct datasets based on workflow

Need for Speed

- Workflow requires high bandwidth or I/O rate
- HPC, data analytics or AI clusters need faster storage

Introducing Data Management Framework

Active & Dormant Data Forms



Data Management Framework

Technology Highlights

- What is DMF?
  - Data Management hub encompassing entire data life cycle
  - Policy-driven POSIX data movement & tiering solution for tape, disk and cloud
  - Scalable metadata capture & search engine based on Big Data technology
  - Scalable parallel data transfer engine optimized for backend

- Brief history of the product

Data Migration Facility
Cray + SGI

DMF 1.0-2.5
Cray

DMF 2.6-3.11
SGI

DMF 4.0-6.9
SGI

→

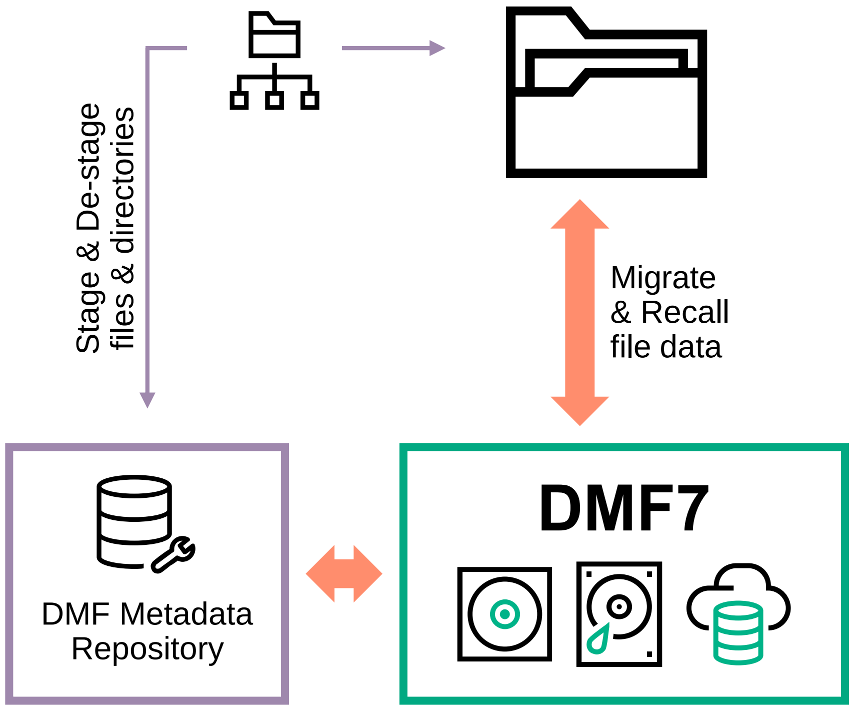
Data Management Framework
HPE

DMF Suite (7.1)
HPE

File System Management

What Can DMF 7 Do?

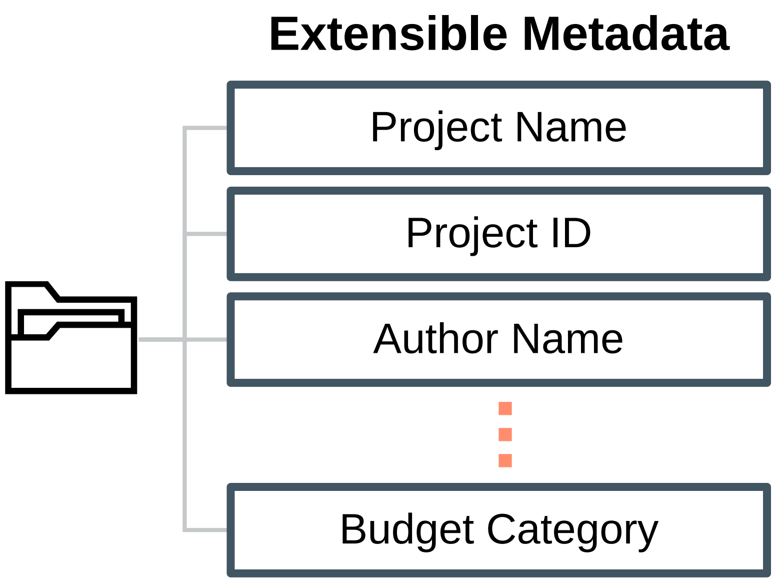
- Maintain namespace reflection with file and directory metadata that can be queried independently of filesystem or data
- Transparently migrate & recall files on Lustre, HPE XFS and other parallel filesystems to and from versioned backend store
- De-stage & stage files and directories, including all metadata, among managed namespaces
- Recover files, directories and entire file systems, replacing backups
- Store files in a “dormant” form without file system representation
- Construct and manage datasets based on file & directory metadata, including extended attributes
- Stage datasets just-in-time on demand via API or HPC job scheduler
- Tier, copy or move datasets according to policy or workflow



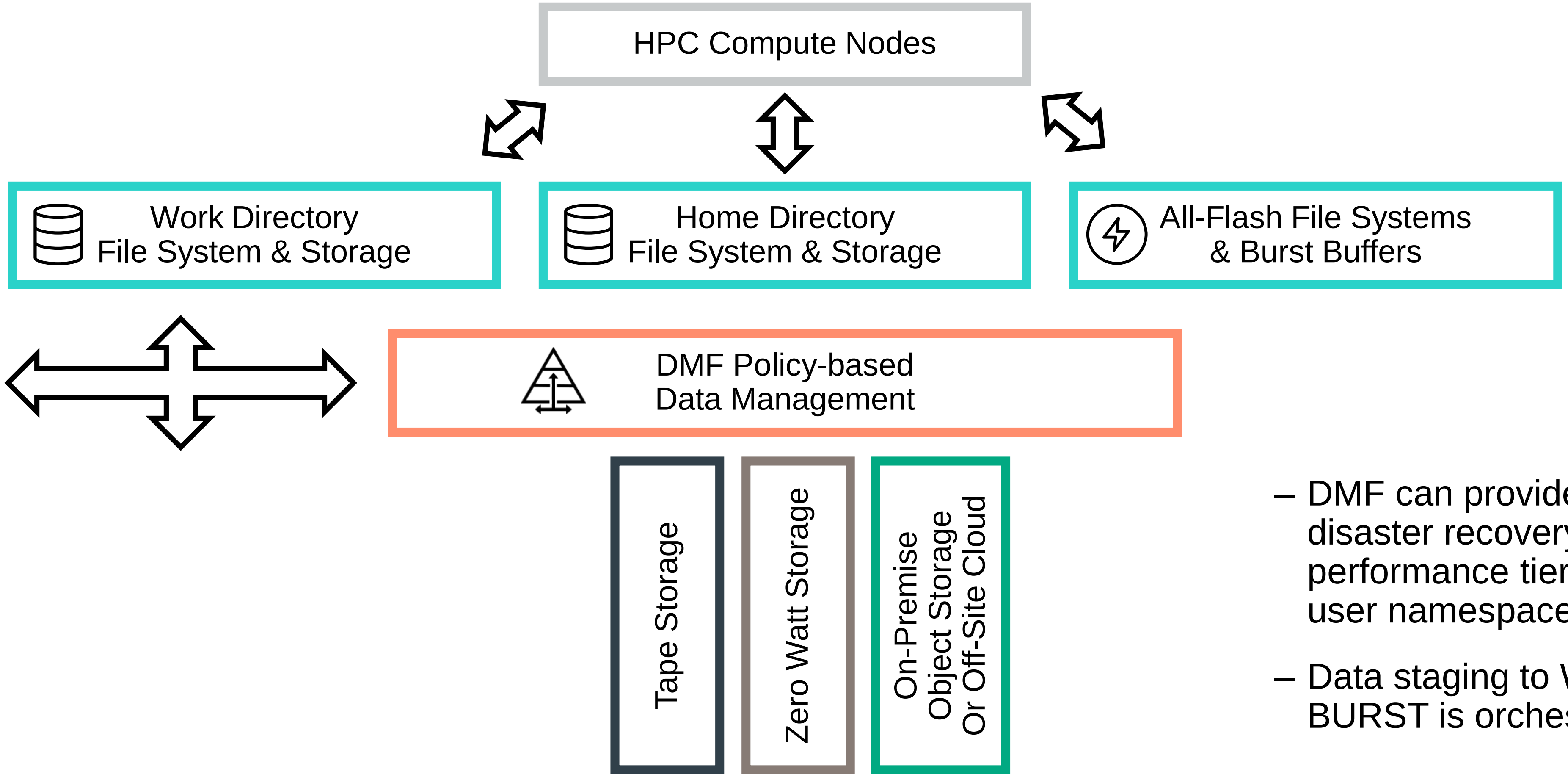
Extensible Metadata Support

Data management flexibility and precision with extensible metadata

- DMF v7 is based on scalable metadata repository
- Repository functions as a long term data store for information about file system structure, attributes, contents and evolution over time
- Metadata repository supports POSIX extended attributes on files and directories, e.g. project name, project ID, etc.
- Queries can be run against metadata including extended attributes for precise and flexible selection of files, e.g. data set creation
- Additionally, policies can be run against the results of metadata queries for data movement, archiving, etc.



Vertical & Horizontal Data Movement

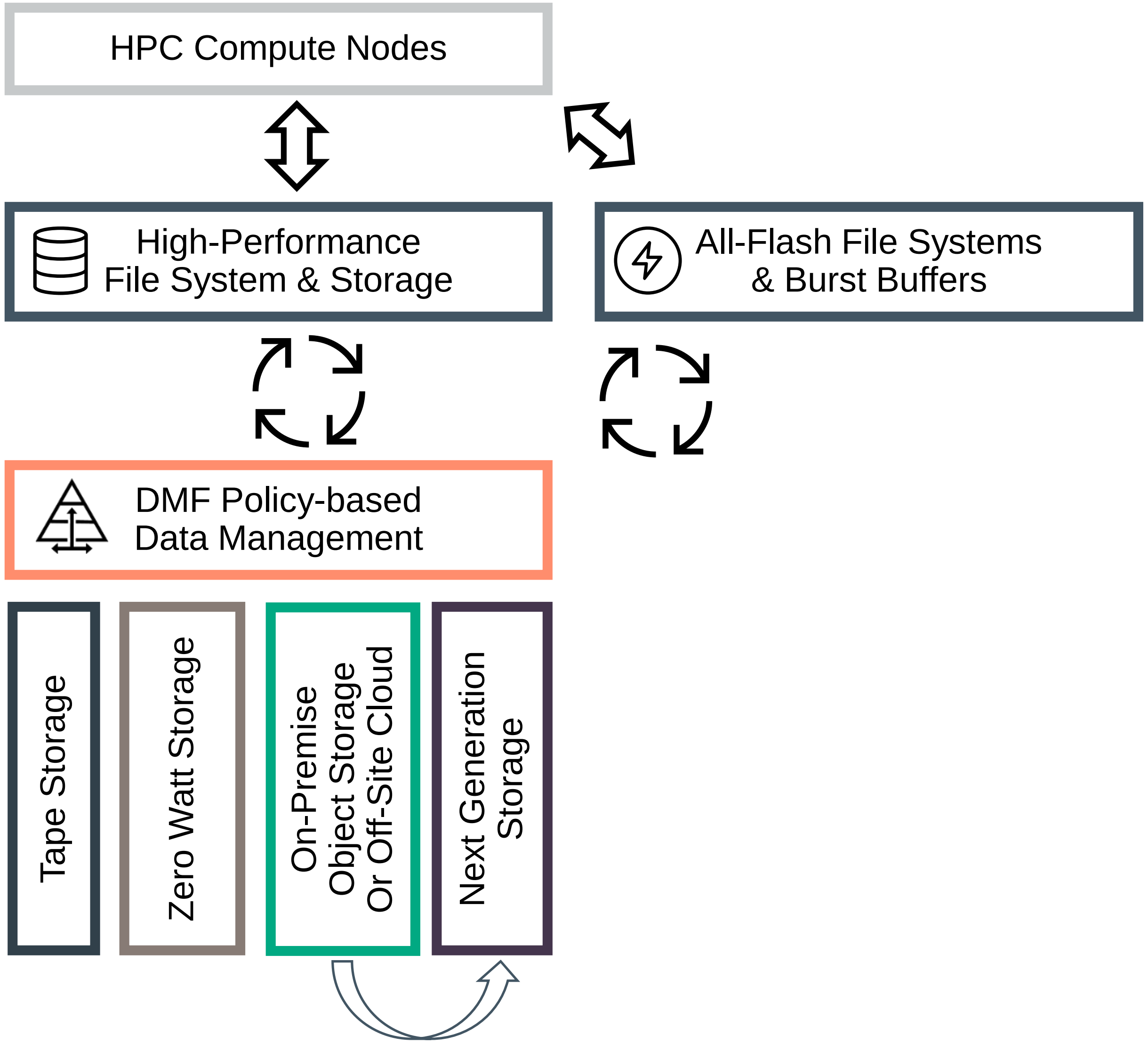


- DMF can provide backup, disaster recovery and high-performance tiered storage for user namespaces
- Data staging to WORK and BURST is orchestrated by DMF

Hewlett Packard Enterprise

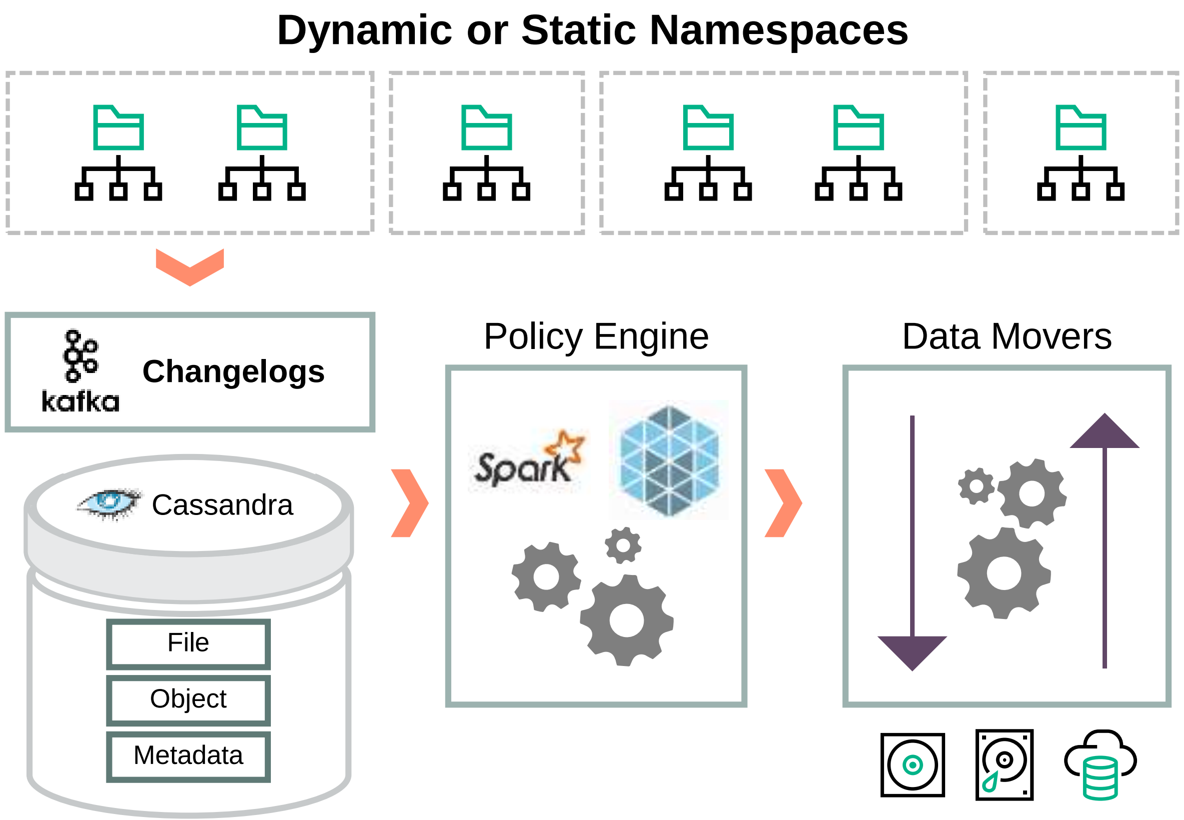
Transition to New Technology

- Manage introduction of new storage technologies over time without disruption
  - Seamlessly manage migration, validation and consolidation of massive data sets
  - Perform migration over period of weeks or months with no impact to user data access
  - Stage managed data to burst buffers or all-flash filesystems



DMF 7 Software Architecture
Designed for Scalability

State-of-art open source components
- Kafka for Changelog processing
- Cassandra for Scalable Metadata
- Mesos for Task Scheduling
- Spark for Query Engine
- Zookeeper for Configuration
- Containerized Components
- Dedicated Components per Filesystem
- Component Level HA

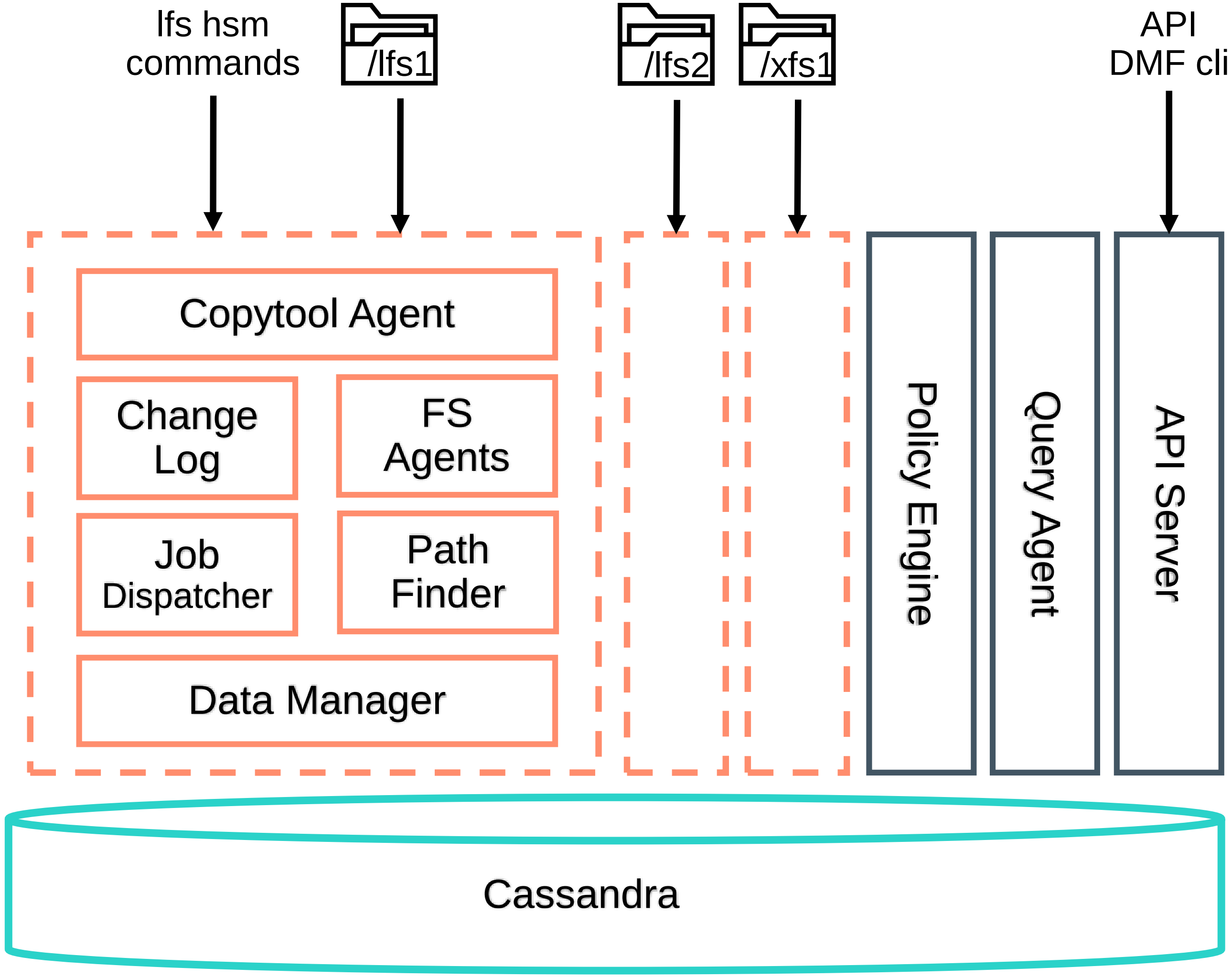


Hewlett Packard Enterprise

Solution Scaling & Extensibility

Unified Scalable Front-End

- DMF 7 has a unified scalable front end for Lustre, HPE XFS and other filesystems (e.g. GPFS)
- Same Query and Policy engine for the all filesystem types
- Same DMF CLI for all filesystem types
- Lustre lfs hsm commands are supported along with the native DMF CLI



Data Management | **DMF** Tape Storage Integration

- DMF is certified with libraries from HPE, as well as Spectra Logic, IBM and Oracle (StorageTek)
  - Streams to tape drive at native rates, even for small files
  - Block ID positioning for fast seek
- Support for latest LTO-8 and Enterprise-class drive technology
- Advanced feature support for accelerated retrieval and automated library management
  - Supports Data Integrity Verification (DIV) and Logical Block Protection (LBP) available with Oracle T10k and IBM LTO7 drives

Zero Watt Storage | **High-Density Storage for DMF**

Performance-Oriented & Power-Managed

Hardware-with-software solution optimized for use with the HPE DMF data management platform

Cost-optimized = Total cost of storage competitive with Tape and lower than Cloud

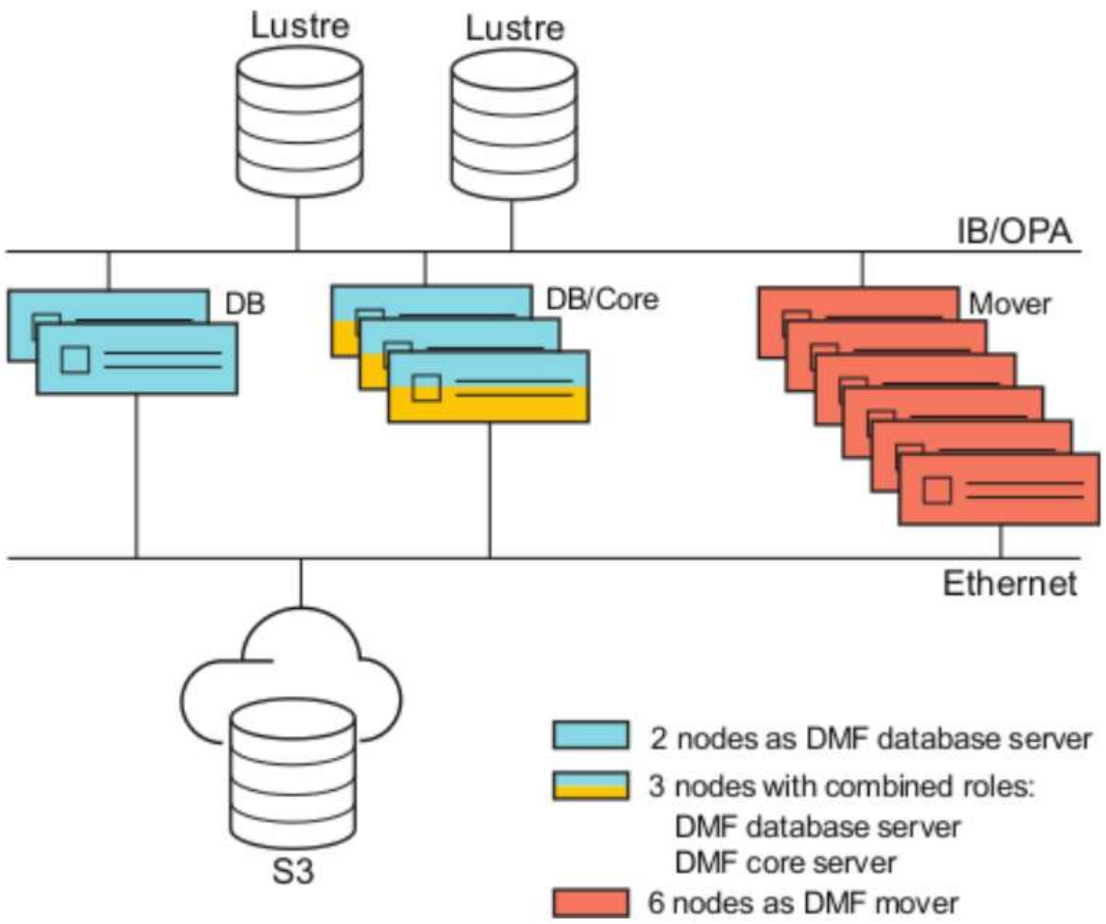
High Performance disk based tier provides instant access to first byte and high throughput streaming

On-premise storage tier which can be used as a capacity storage tier as well as a fast mount cache or “relief” to RAID arrays

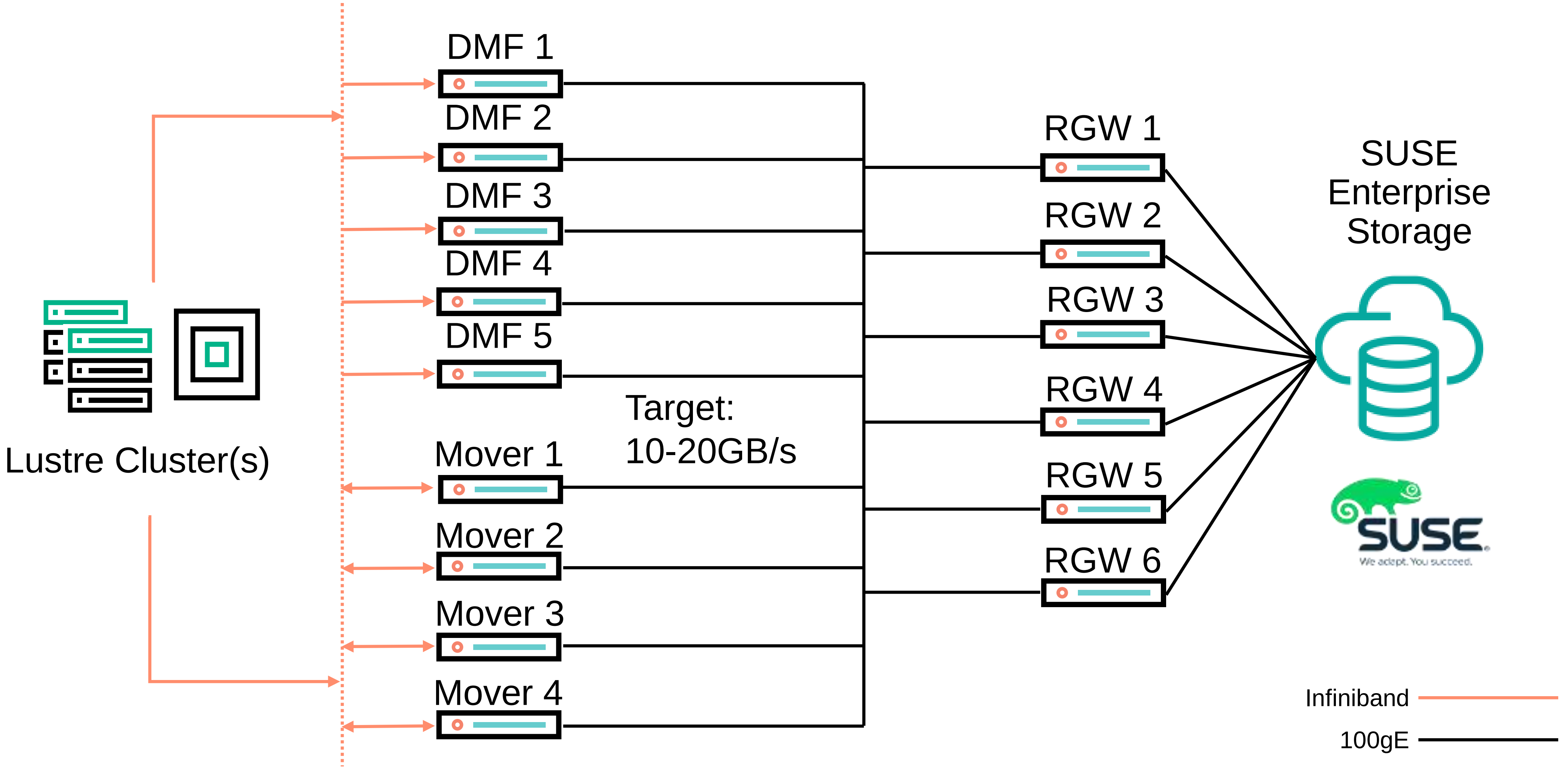
Hewlett Packard Enterprise

Solution Scaling & High Availability

– DMF can scale by adding nodes that perform the required roles
  – Add nodes that are DMF database servers to scale metadata capability
  – Add nodes that are DMF movers to scale data migration capability
– Example of large Lustre configuration
  – Five nodes each provide the DMF database server. Three of those nodes also act as the DMF core server
  – Six nodes are the DMF movers

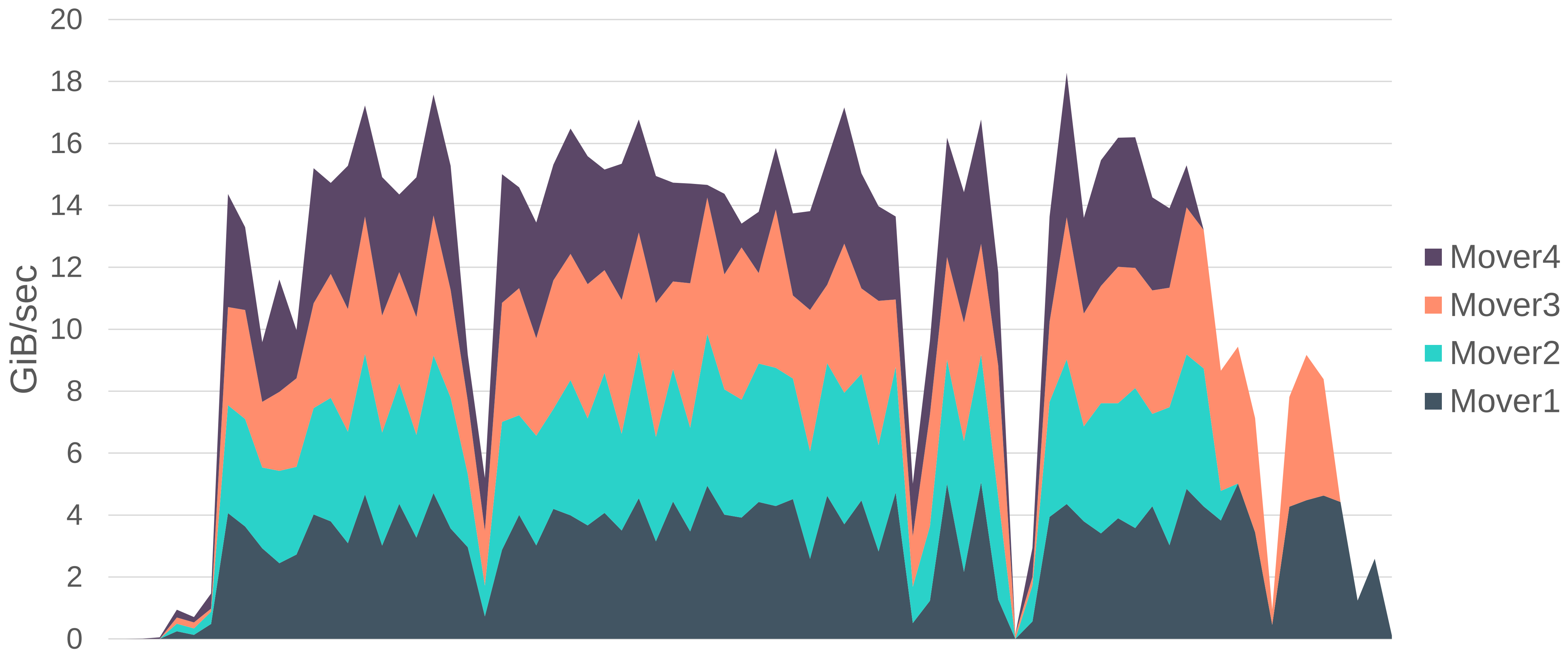


Data Management | **DMF 7** Design of Lustre Tiering to SES



DMF7 S3 Mover Performance

Four DL360 Mover Nodes | Lustre to SES



Hewlett Packard Enterprise

Summary | DMF 7

Tiered Storage

- Scalable storage tiering and backup
- High-latency media such as tape and cloud

Metadata Search

- Locate, select and move large groups of files
- Standard and user-assigned attributes

Flash Scratch

- Right-sized, flash-based, “burst buffer” namespaces
- High throughput and millions of IOPS

Hewlett Packard Enterprise

Thank you

kirill.malkin@hpe.com