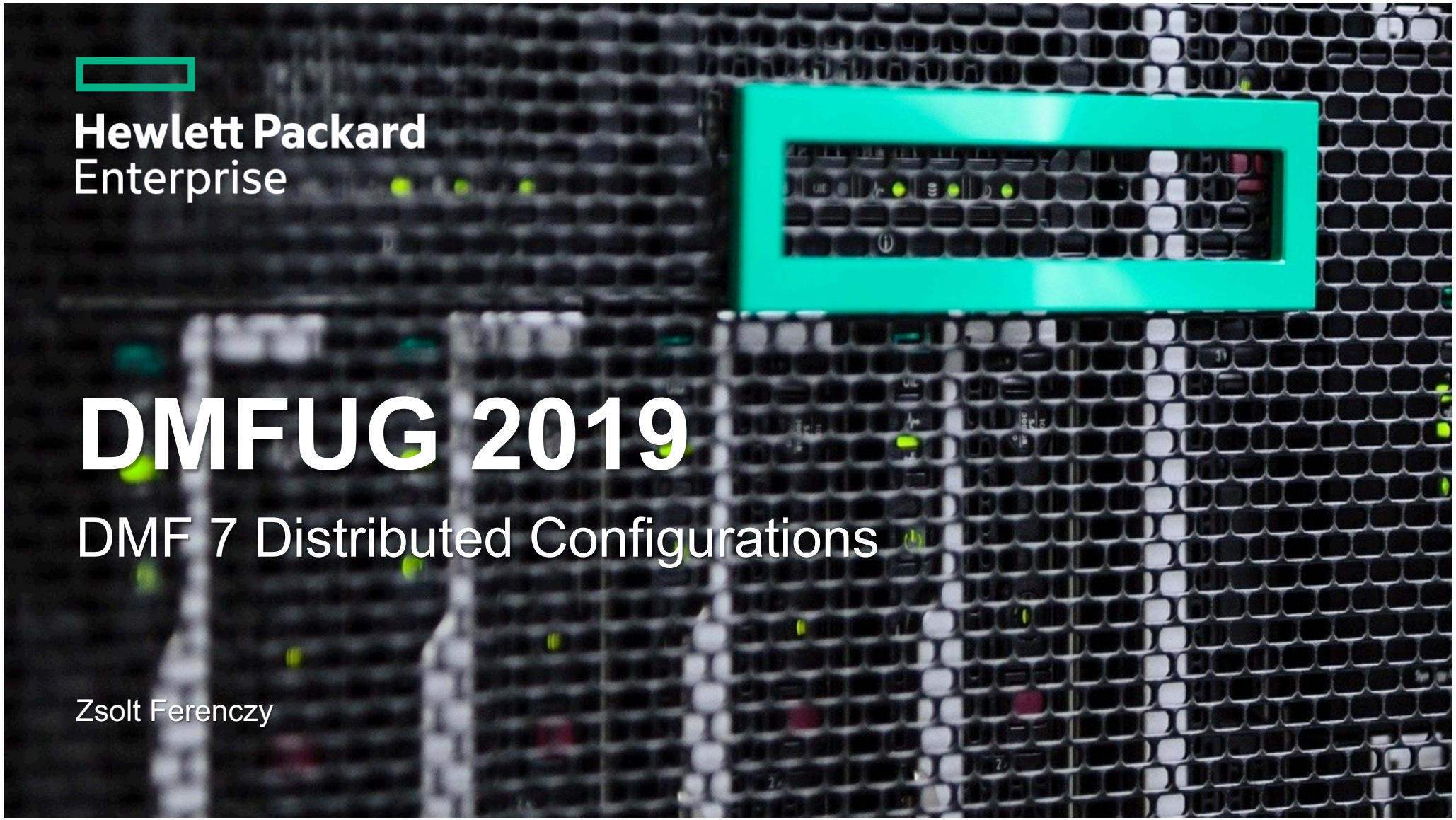

Hewlett Packard
Enterprise
DMFUG 2019
DMF 7 Distributed Configurations
Zsolt Ferenczy

Confidentiality Notice

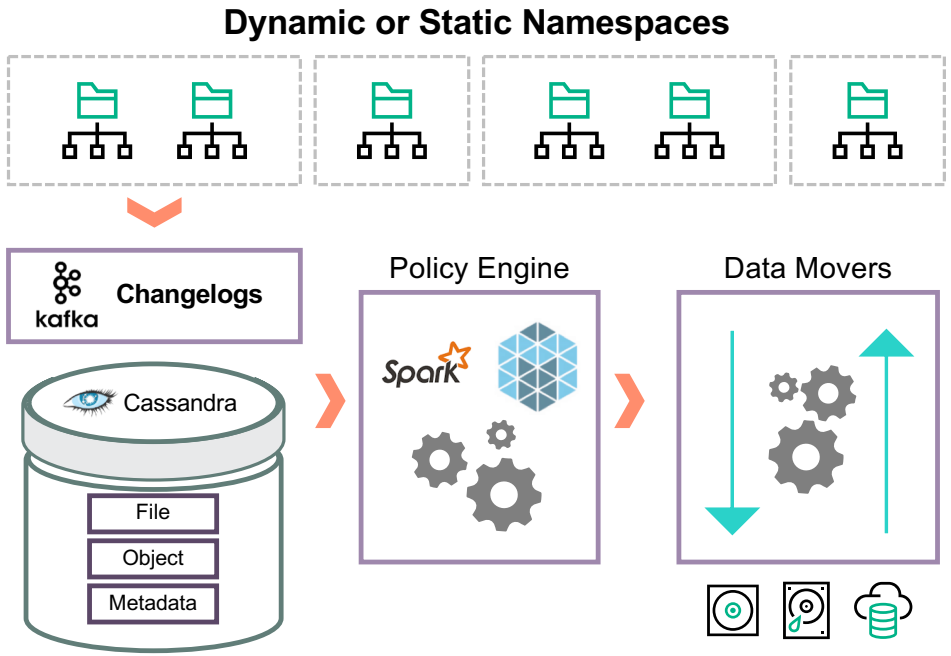
- **The information contained in this presentation** is proprietary to Hewlett Packard Enterprise (HPE) Company and is offered in confidence, subject to the terms and conditions of a Confidential Disclosure Agreement

- **HPE makes no warranties regarding the accuracy of this information**. This document contains forward looking statements regarding future operations, product development, product capabilities and availability dates. This information is subject to substantial uncertainties and is subject to change at any time without prior notification. Statements contained in this document concerning these matters only reflect Hewlett-Packard Enterprise's predictions and / or expectations as of the date of this document and actual results and future plans of Hewlett-Packard Enterprise may differ significantly as a result of, among other things, changes in product strategy resulting from technological, internal corporate, market and other changes. This is not a commitment to deliver any material, code or functionality and should not be relied upon in making purchasing decisions.

Software Architecture

Modern open source architecture

- 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



Available Now

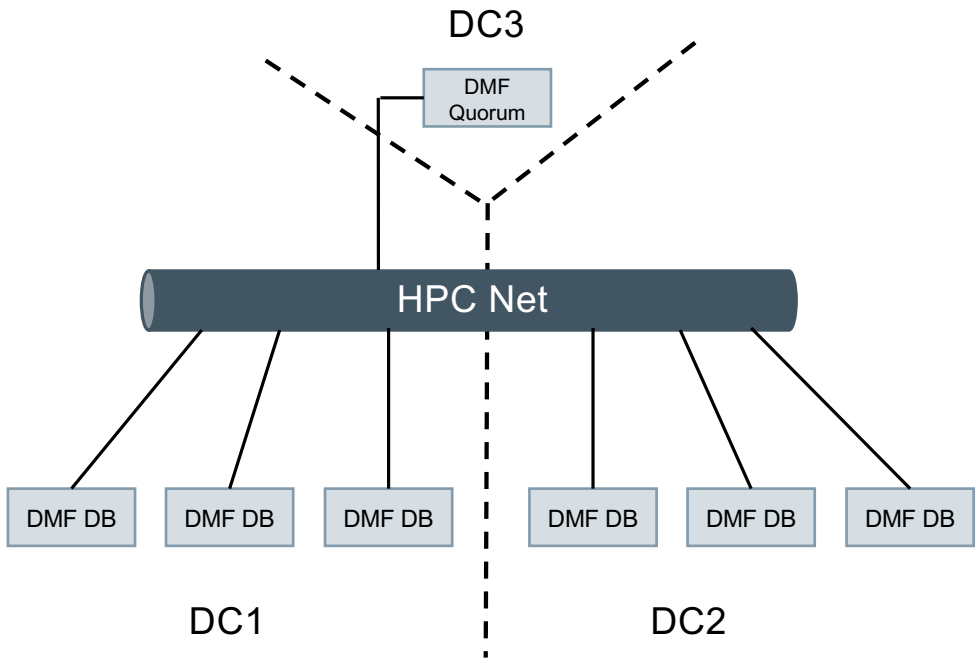
- Multi Data Center / Independent Failure Domains
- Independent Storage Islands

Future

- Full Federation
- Multi-Site Replication

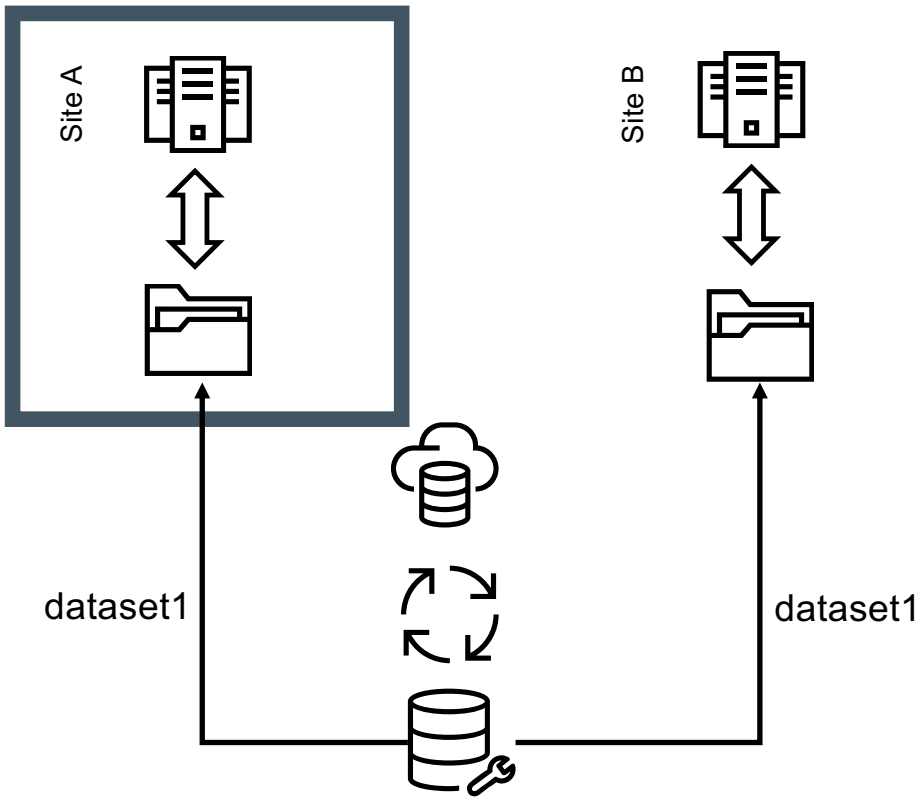
Data Management Fabric | **DMF 7** Multi Data Center

- Split a DMF cluster across different failure domains
- Separate racks in 1 DC
- Multi-DC on a campus
- Allows DMF to continue to function if one domain fails
- Requires high speed network
- Requires using a quorum node in a separate failure domain from the DB nodes
- Not applicable to WANs
- Not a DMF7 metadata Federation
- All metadata is replicated between DCs



Hewlett Packard Enterprise

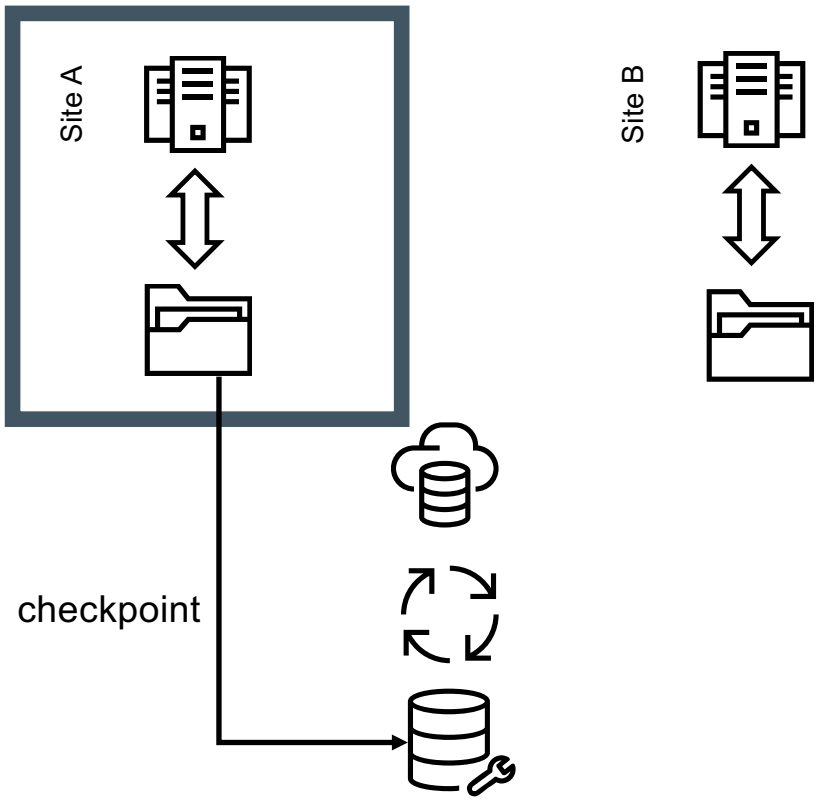
Data Management Fabric | **DMF 7** Production HPC Job Failover



- Stage Job data set into POSIX at both sites
- Production Job started at Site A

Hewlett Packard Enterprise

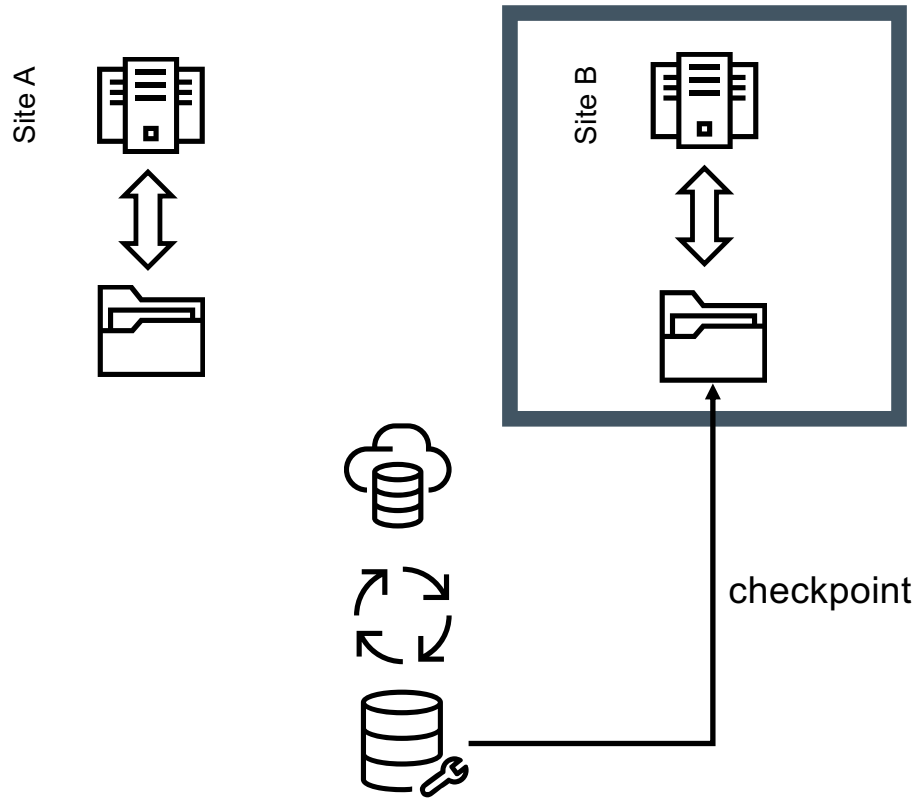
Data Management Fabric | **DMF 7** Production HPC Job Failover



- Job intermediate output files and check point restart files are periodically migrated
- Interval may be set via the DMF policy engine
- HPC applications may also make use of the DMF API to trigger migrations at predetermined time steps
- DMF replicates data and metadata

Hewlett Packard Enterprise

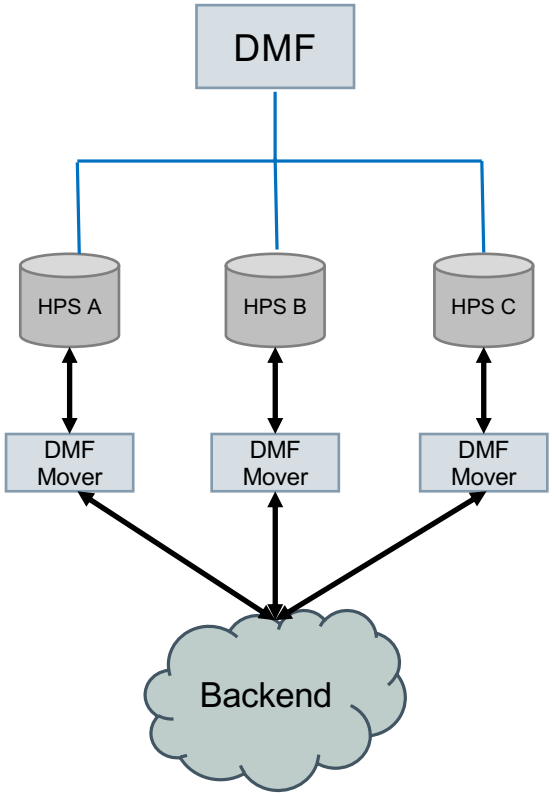
Data Management Fabric | **DMF 7** Production HPC Job Failover



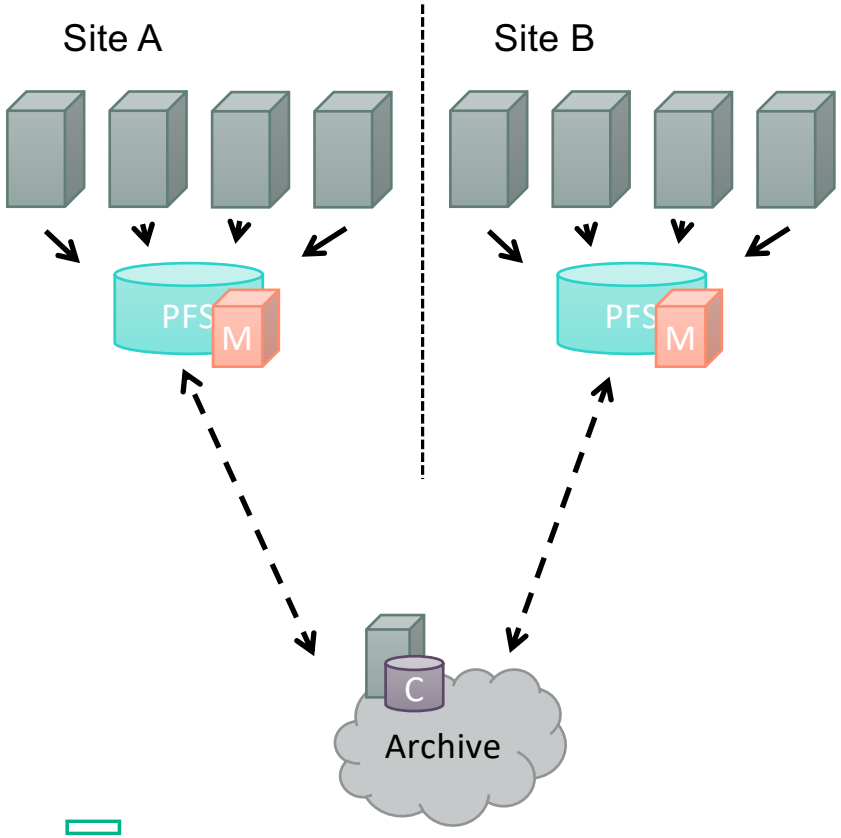
- In the event of failure of the primary site, the latest checkpoint is staged into POSIX at site B
- Job is restarted
- As DMF supports file versions, any time step that was generated during the run at site A can be staged

Data Management Fabric | **DMF 7** Islands

- Allows a single instance of DMF to manage independent filesystems from a common backend
- For Tape/ZWS, an instance of Library Server per Island is required
- Supports
  - Multiple non-clustered EXFS file servers
  - Multiple CXFS clusters
  - Traditional Active/Stand-by LXVM/EXFS HA



HPC Data Management | **Federated Multi-Site**



- Multi-site HPC filesystems are difficult to deploy
- Replication at the POSIX filesystem layer has too much overhead for the performance requirements of HPC application
- Replication at the block layer is incompatible with HPC parallel filesystems
- Replication above the filesystem, rsync/copy/etc, has high administration overhead and is non-scalable for large data sets

Hewlett Packard Enterprise

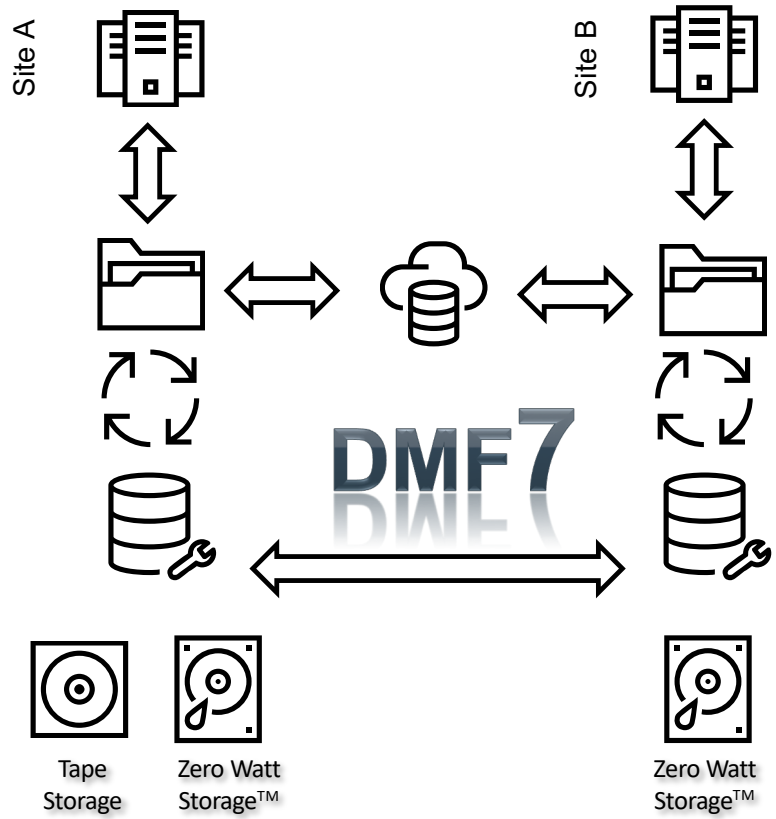
Data Management Fabric | **DMF 7** Global Namespace Solution

Key Takeaways

Global Metadata Namespace

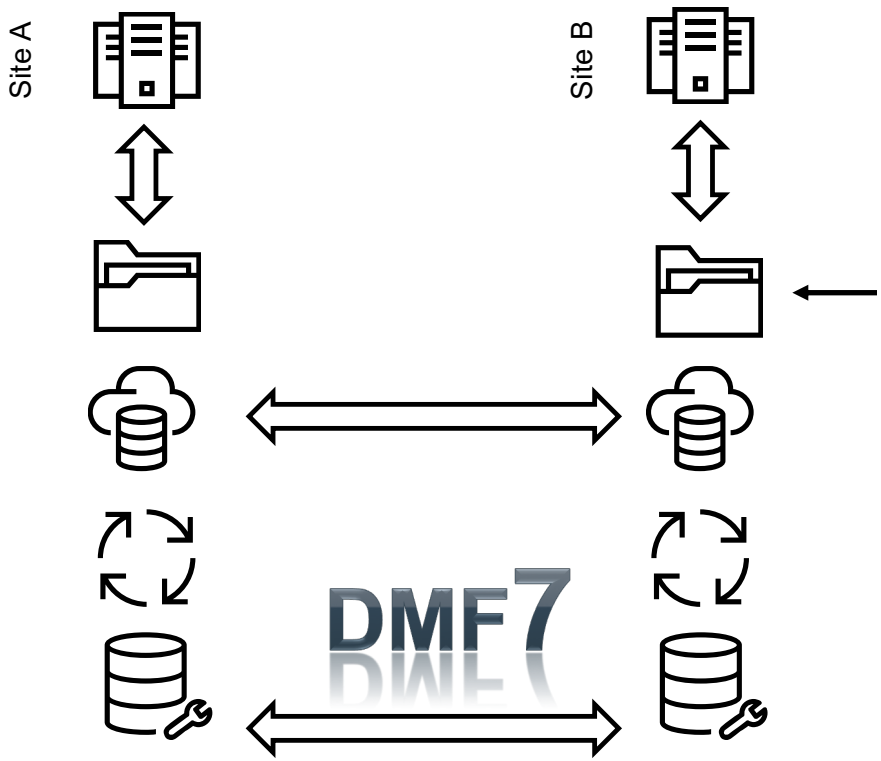
DMF 7 metadata repository is based on a modern distributed database that can span multiple sites. This enables replication of metadata records across all sites, running local metadata queries and staging local filesystems.

- **Replication via cloud or object storage:** Once a copy of data is in the cloud and the object record was replicated, any DMF 7 enabled site can stage these data objects in a local filesystem
- **Multiple sites supported:** Distributed key-value Cassandra database used in DMF 7 allows multiple sites to participate in the metadata replication, provide view of all data objects and support metadata queries
- **Optimize access by caching in ZWS:** The local copies can be cached in the Zero Watt Storage backend and replicated to the cloud later. Similarly, the data can be pre-fetched in ZWS prior to staging to the filesystem



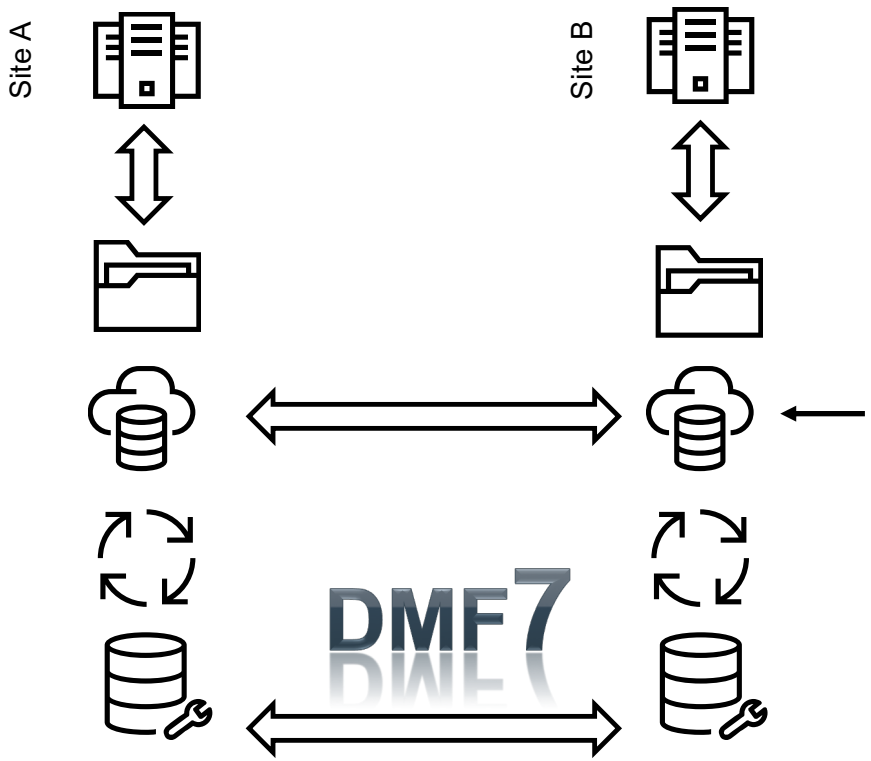
Hewlett Packard Enterprise

Data Management Fabric | **DMF 7** Global Namespace Solution



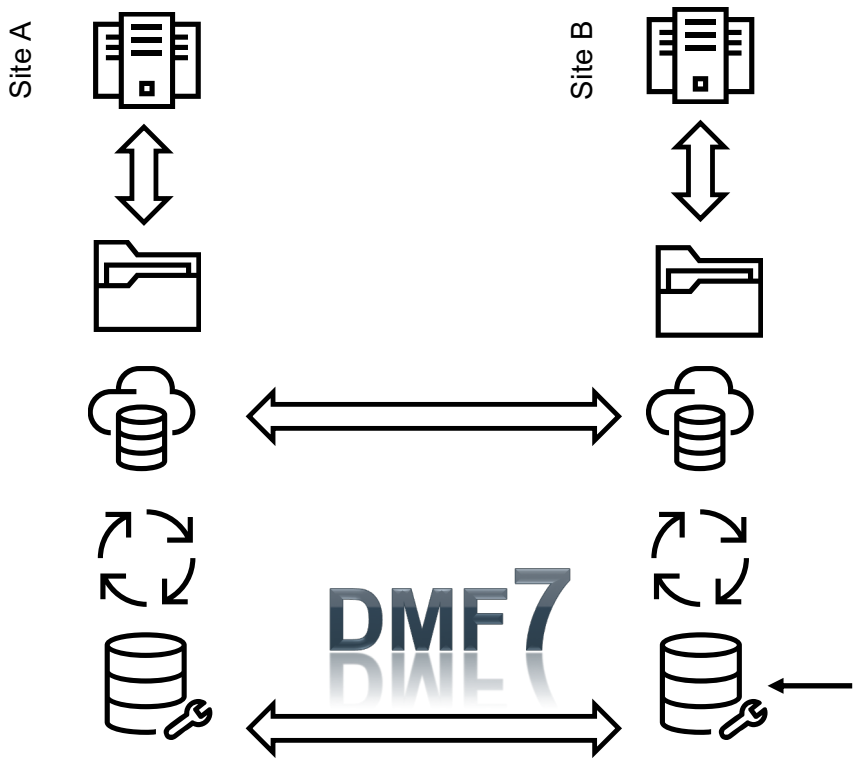
- POSIX Filesystem layer at each site is 100% independent
- Full performance of the hardware
- No added latency for remote access or synchronization
- No external applications for synchronization

Data Management Fabric | **DMF 7** Global Namespace Solution



- Metadata is replicated
- Global view for the entire namespace from either site

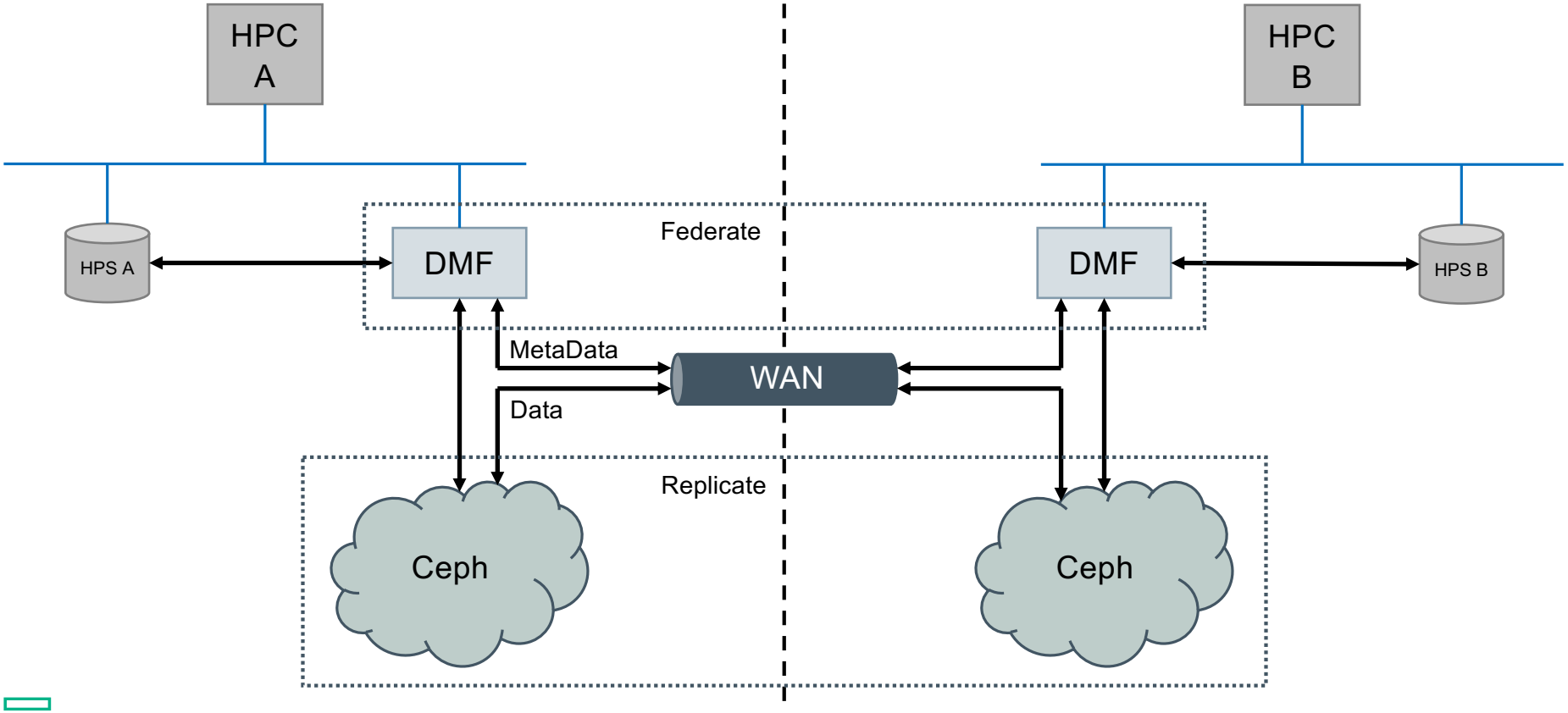
Data Management Fabric | **DMF 7** Global Namespace Solution



- Dormant data is replicated
- Access to full data set from either site
- Dormant data promoted to local POSIX namespace for job execution

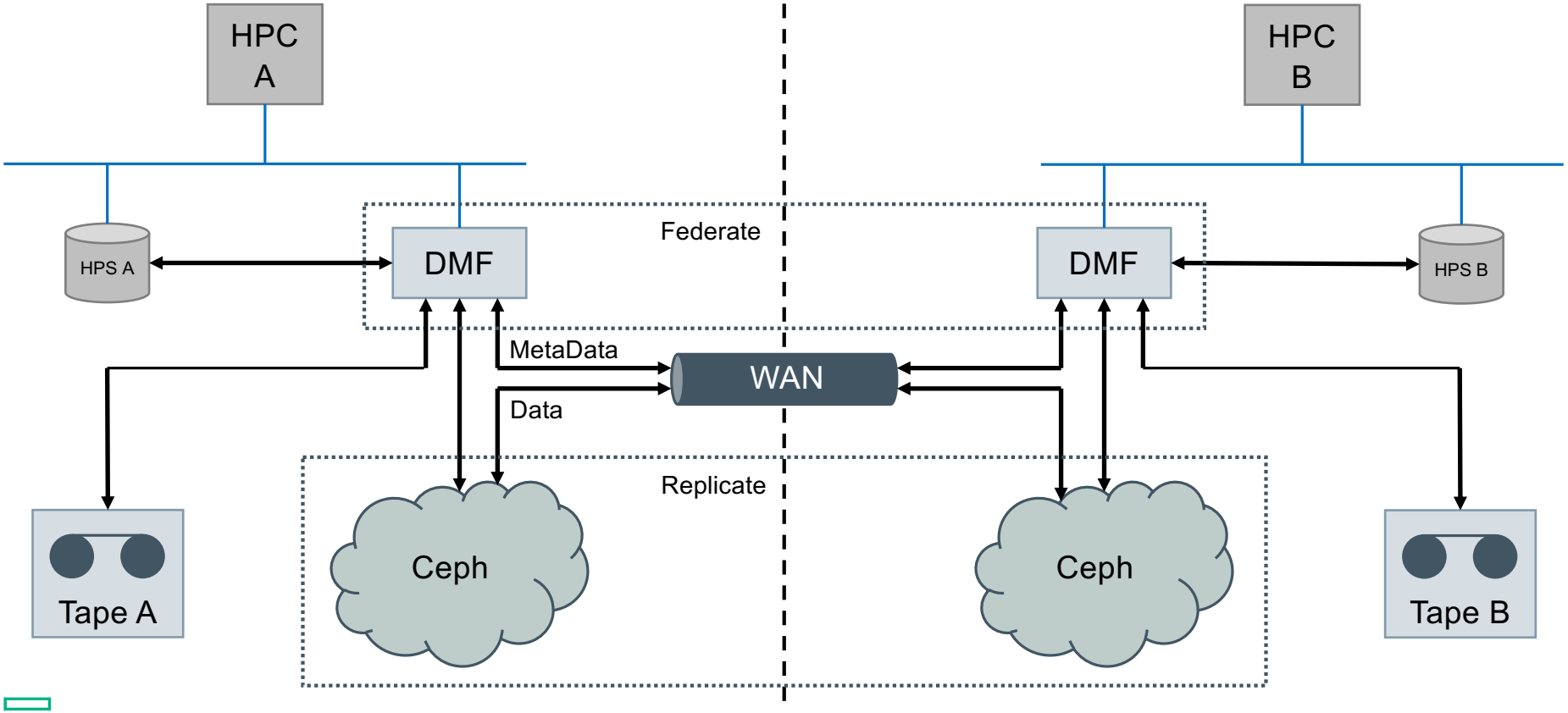
Hewlett Packard Enterprise

Data Management Fabric | **DMF 7** Multi-site High Level Design



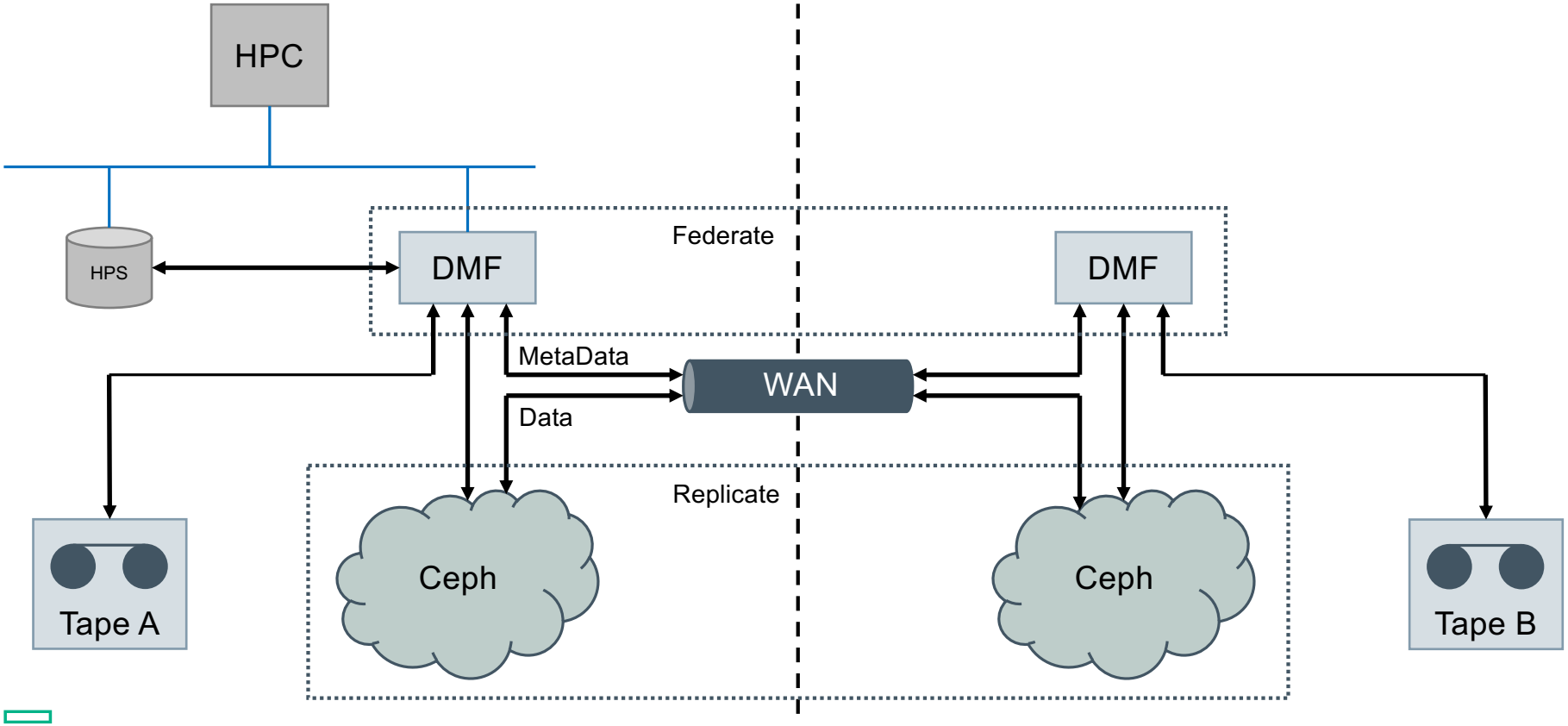
Hewlett Packard Enterprise

Data Management Fabric | **DMF 7** Multi-site High Level Design



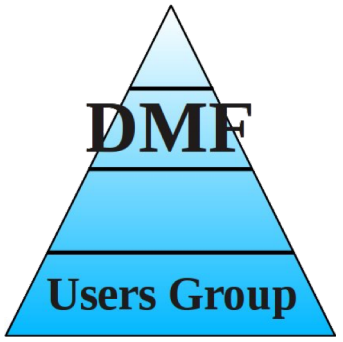
Hewlett Packard Enterprise

Data Management Fabric | **DMF 7** Multi-site High Level Design



Hewlett Packard Enterprise

Hewlett Packard
Enterprise



Thank you