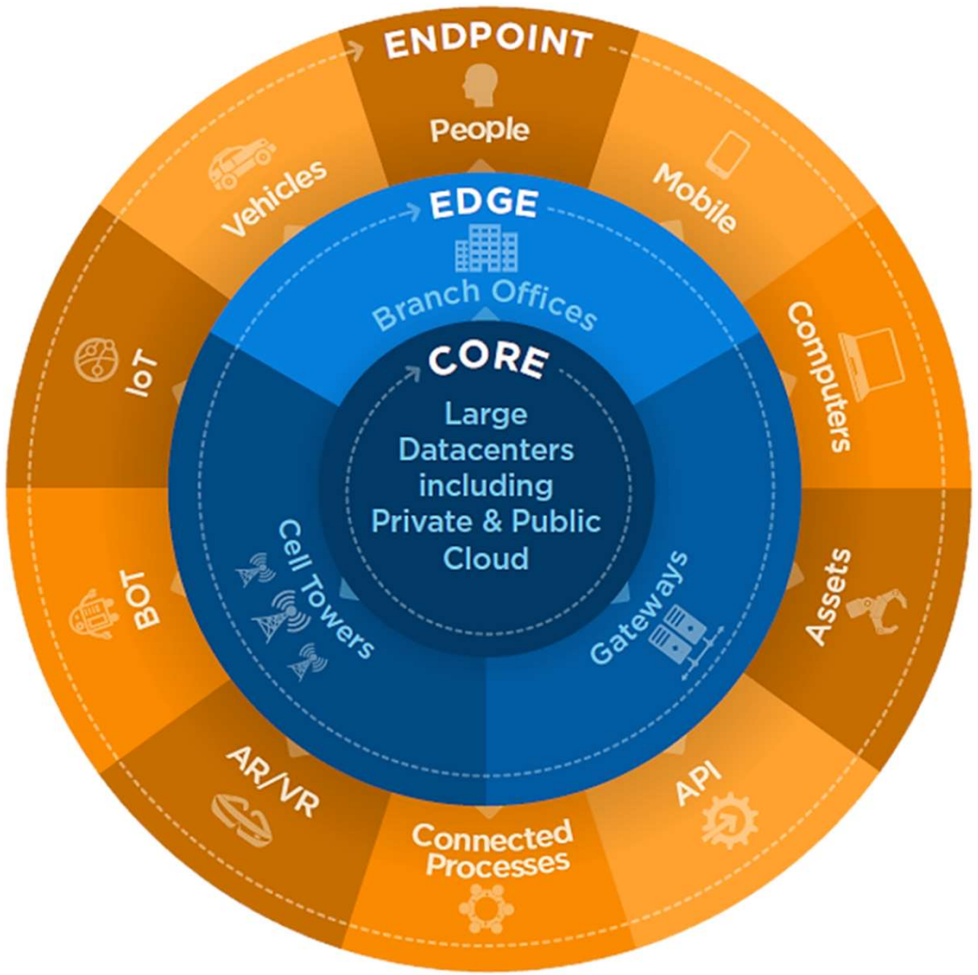

Hewlett Packard
Enterprise
STAYING AHEAD OF THE
DATA DELUGE
David Honey, HPC Solutions Architect
February 10th, 2021
intel

THE CONVERGENCE OF IOT, BIG DATA ANALYTICS AND ARTIFICIAL INTELLIGENCE CREATING A DATA DELUGE

THE INTERNET OF THINGS



The dramatic growth in cloud storage is affecting the mix and share of the Global StorageSphere installed base between core, edge, and endpoint.

The core is where the world's data is progressively being stored and will hold a 60% share of the StorageSphere installed base in 2024, up from 40% in 2019.
While the edge is growing nearly as fast as core, it will hold less than 10% of the overall Global StorageSphere installed base in 2024*

More than 37% of total data generated in 2020 (40ZB) will have significant business value! †

1 ZettaByte = 10^{21} = 1 Million Peta Bytes

*IDC Worldwide global storage forecast, 2020 - 2024

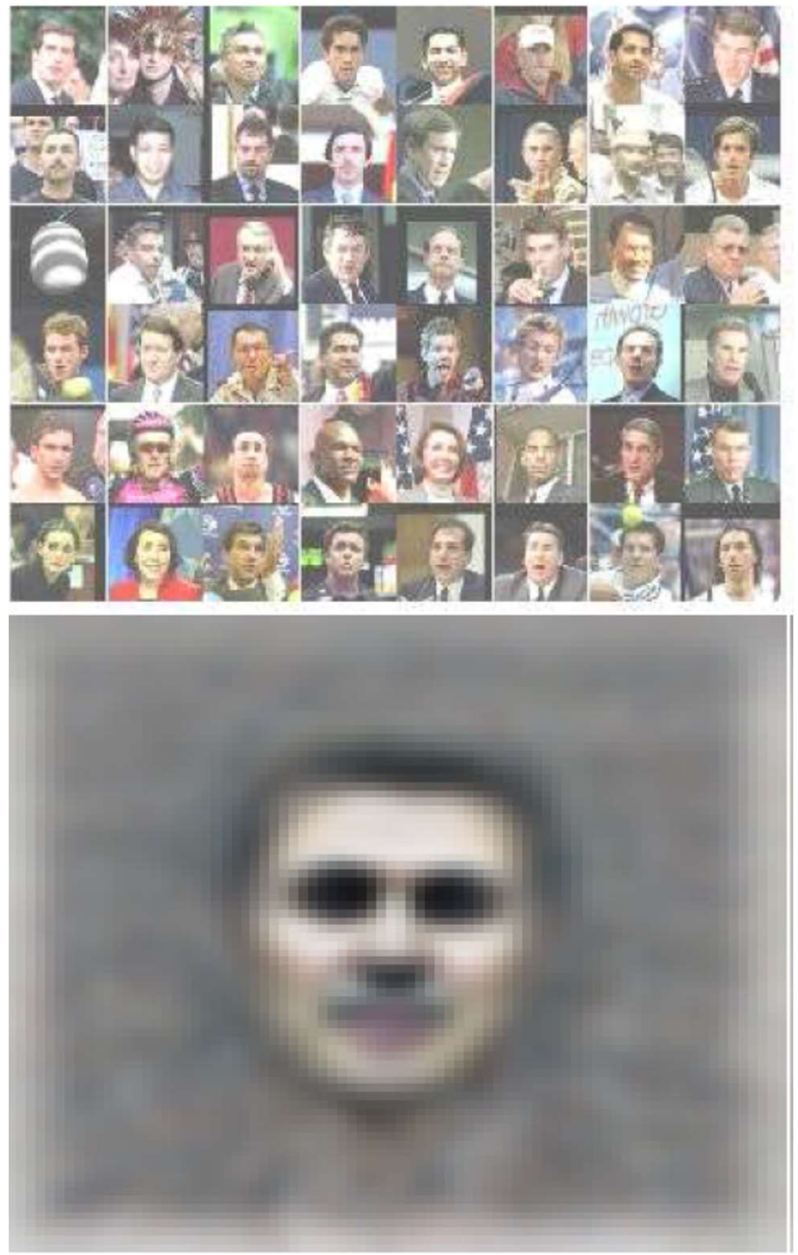
† IDC projection made in 2019

Hewlett Packard Enterprise | intel

THE DEEP LEARNING REVOLUTION

1. **Fast, cheap compute**
2. **Rich, public data collections**
3. **Capable neural networks**

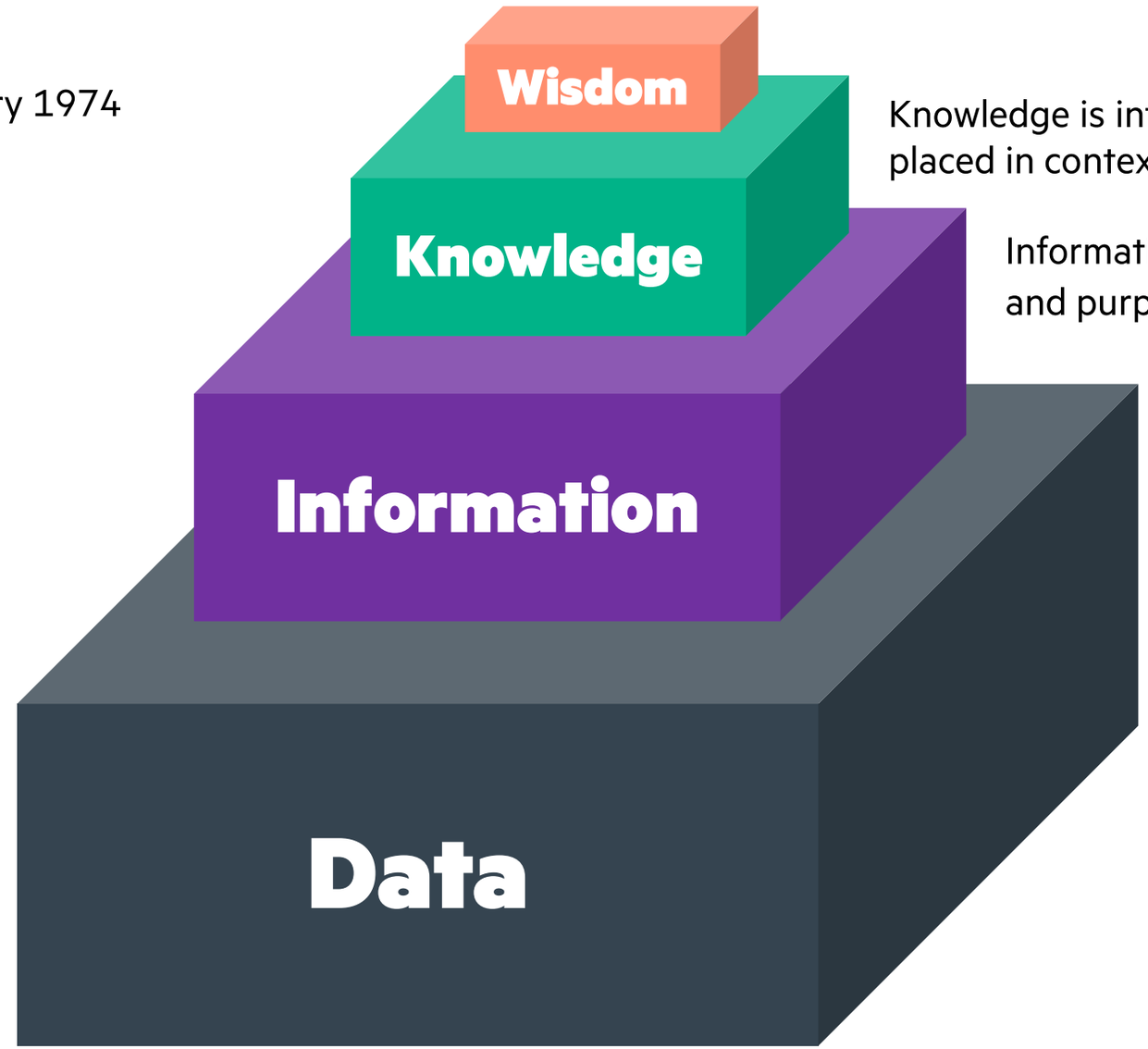
Millions of images are required to train a Neural Network (the more the better)



THE KNOWLEDGE HIERARCHY

Nicholas L. Henry 1974

Wisdom is knowledge that enables shared understanding. A pre-requisite for effective decision making.



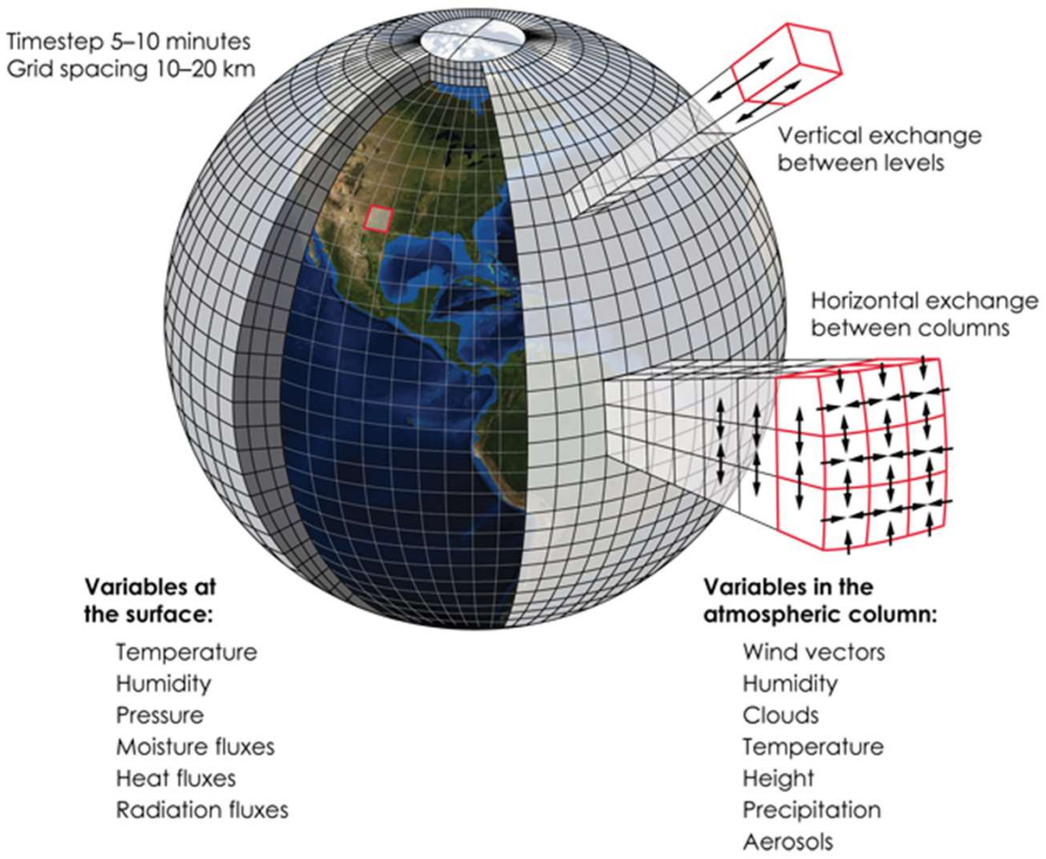
Knowledge is information placed in context

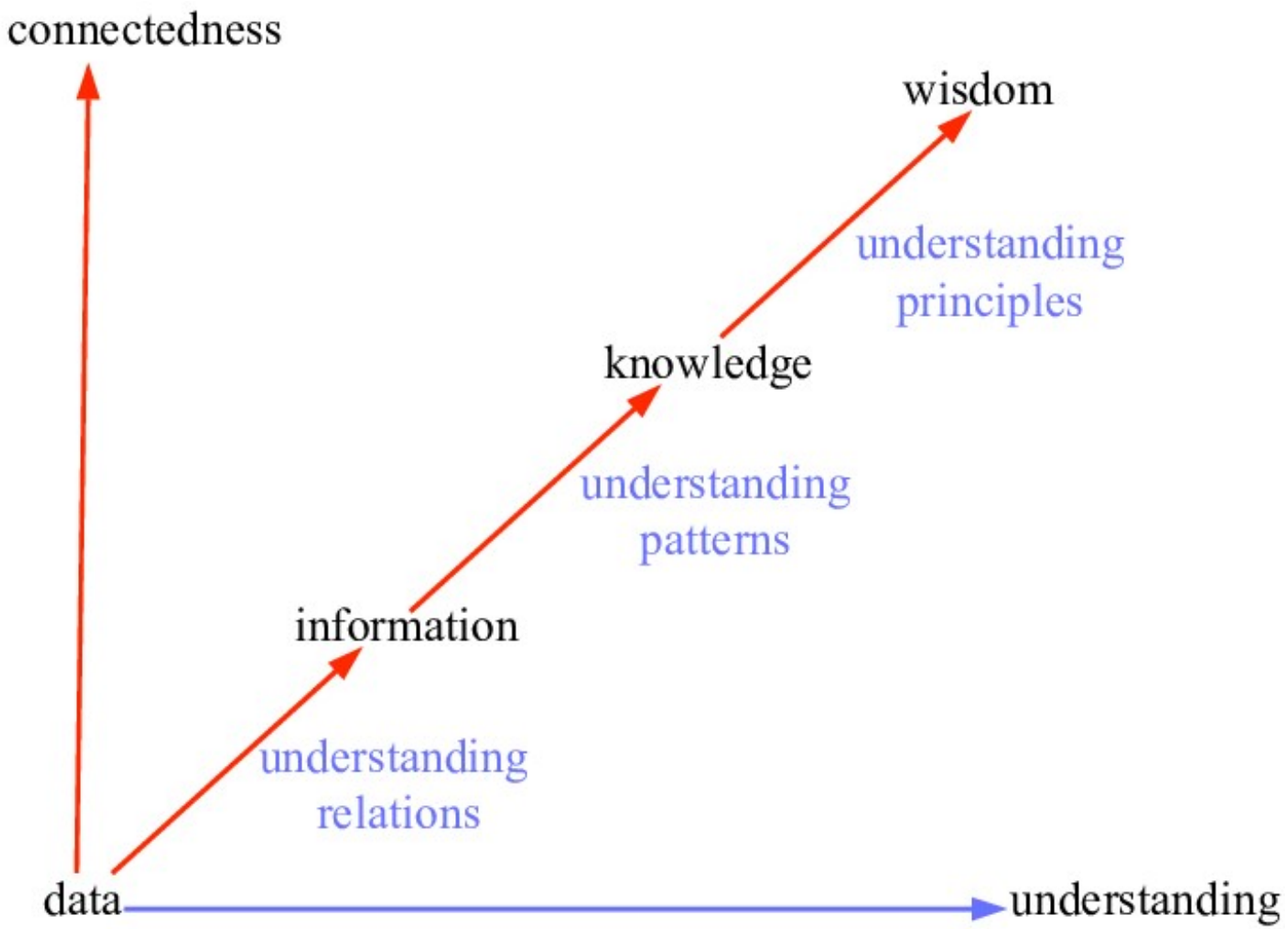
Information is data with relevance and purpose

Data are discrete, objective, facts that have no meaning in isolation

KNOWLEDGE HIERARCHY IN EARTH SCIENCES

Weather, Climate, Remote Sensing, Geospatial, Seismic

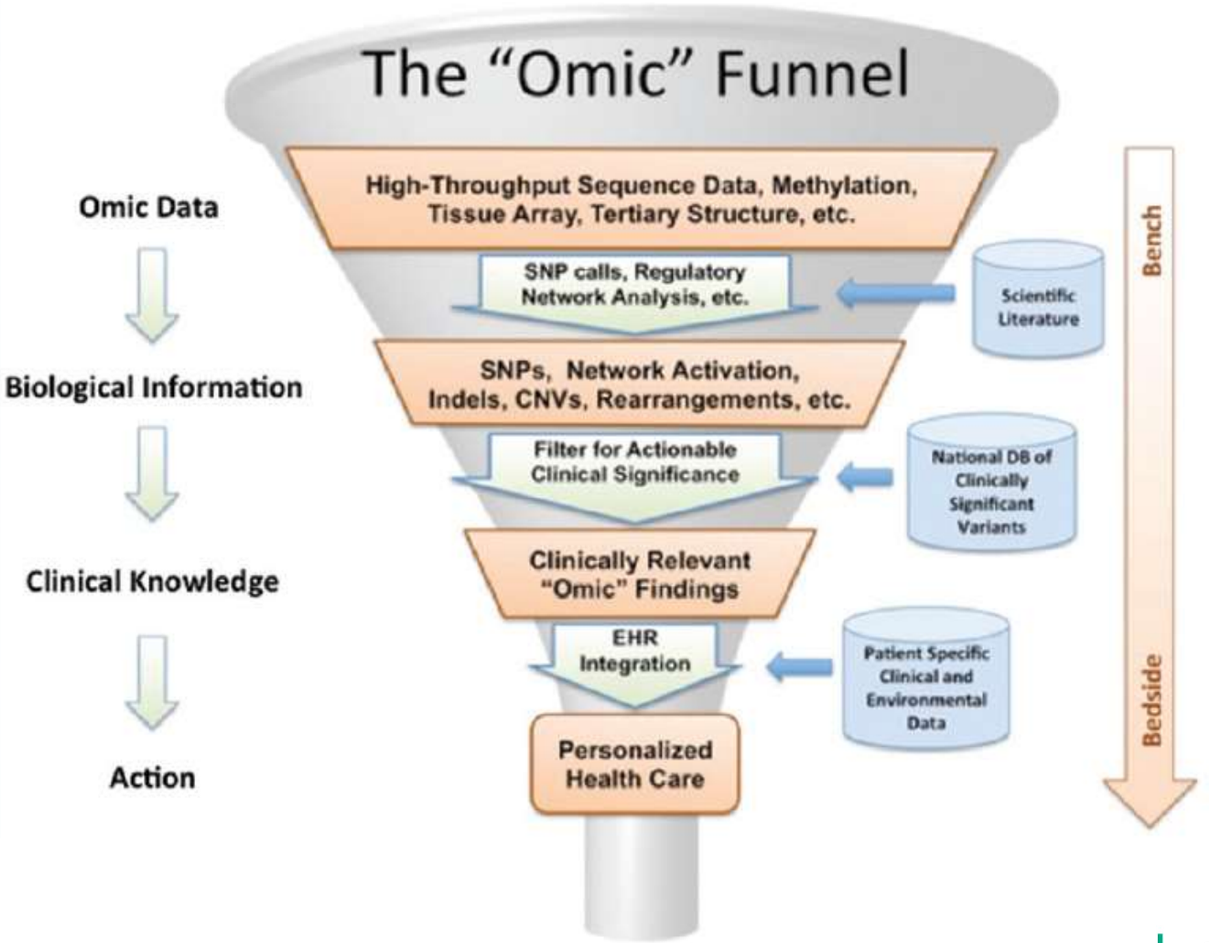




Hewlett Packard Enterprise | intel

IN LIFE SCIENCES

Genomics - Transcriptomics - Proteomics - Metabolomics

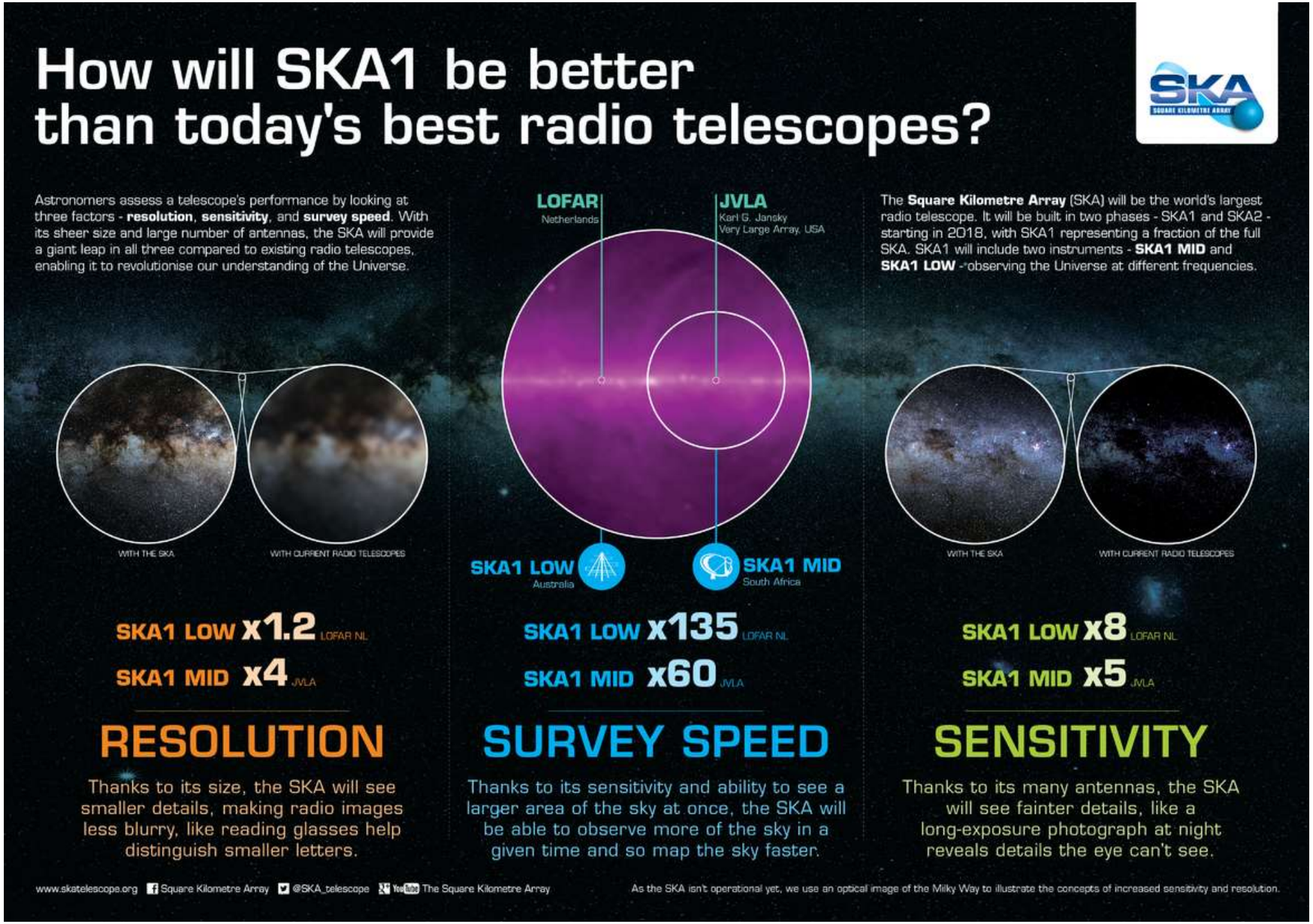


10TB per day per sequencer

IN RADIO ASTRONOMY

SKA an extreme example of a high volume instrument

ASKAP generates 2.5GBps



Hewlett Packard Enterprise | intel

INTRODUCING HPE DATA MANAGEMENT FRAMEWORK VERSION 7

What Challenges does it Solve?

Too much data

- 1PB or more of unstructured data
  - Simple, cost-effective storage & data protection solution

Too many files

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

Need for Speed

- Workflow demands high bandwidth or I/O rate
  - Fast storage for HPC, data analytics or AI clusters

DATA MANAGEMENT WITH DMF7

Benefits

Tiered Storage

- Scalable storage tiering and backup
  - Support for high-latency media such as tape and cloud

Metadata Search

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

Flash Scratch

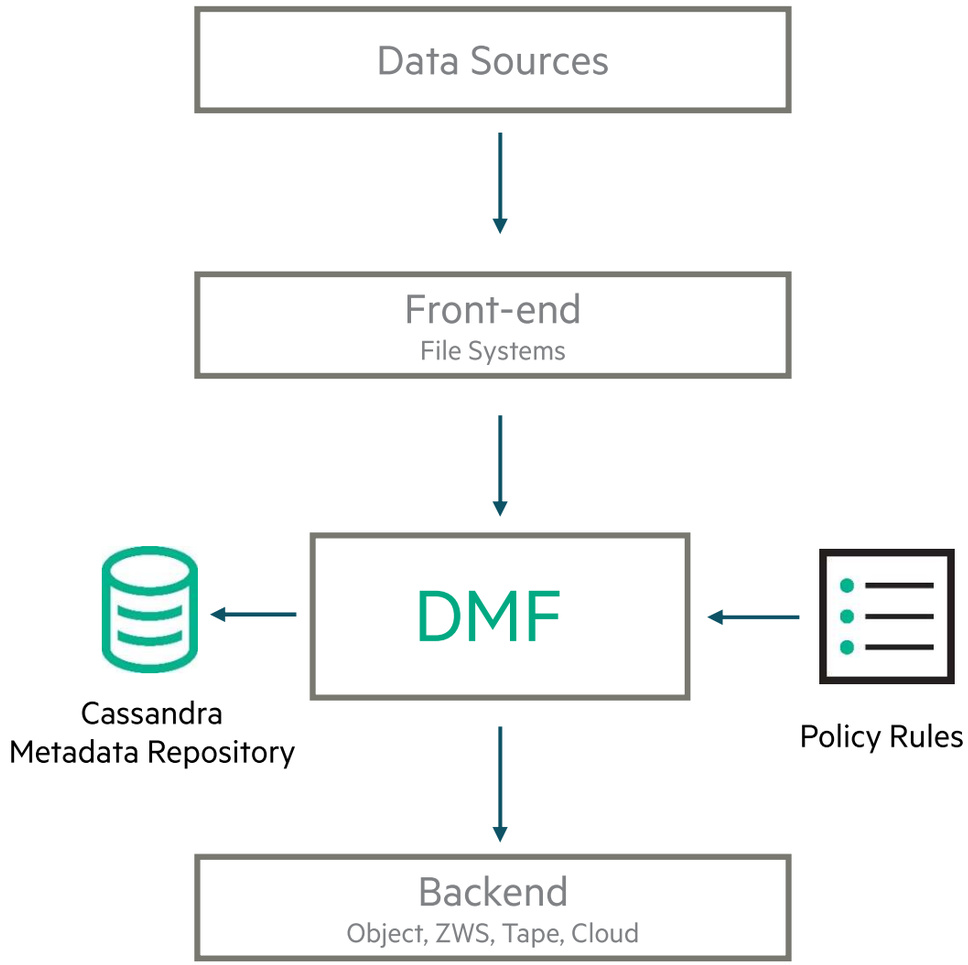
- Right-sized, flash-based, “burst buffer” namespaces
  - Delivers high throughput and millions of IOPS

DATA MANAGEMENT FRAMEWORK

Conceptually

DMF 7 is all of these

- Scalable Hierarchical Storage Management System
- Data Manager supporting HPC workflows
- Posix-to-Object Bridge
- Data Transfer Engine optimized for throughput
- Space or workflow driven Data Policy Engine



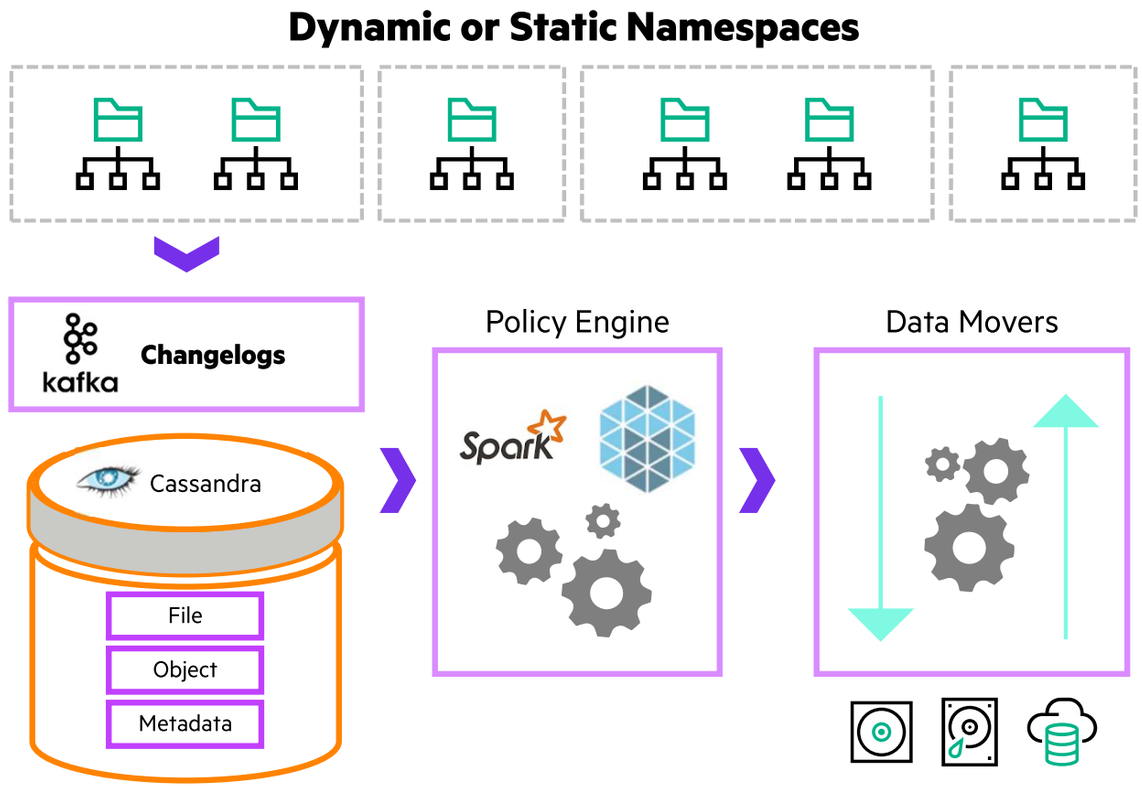
OPEN SCALABLE SOFTWARE TECHNOLOGIES

Modern open-source architecture

- Kafka for Changelog processing
- Cassandra for Scalable Metadata
- Mesos for Task Scheduling
- Spark Query Engine

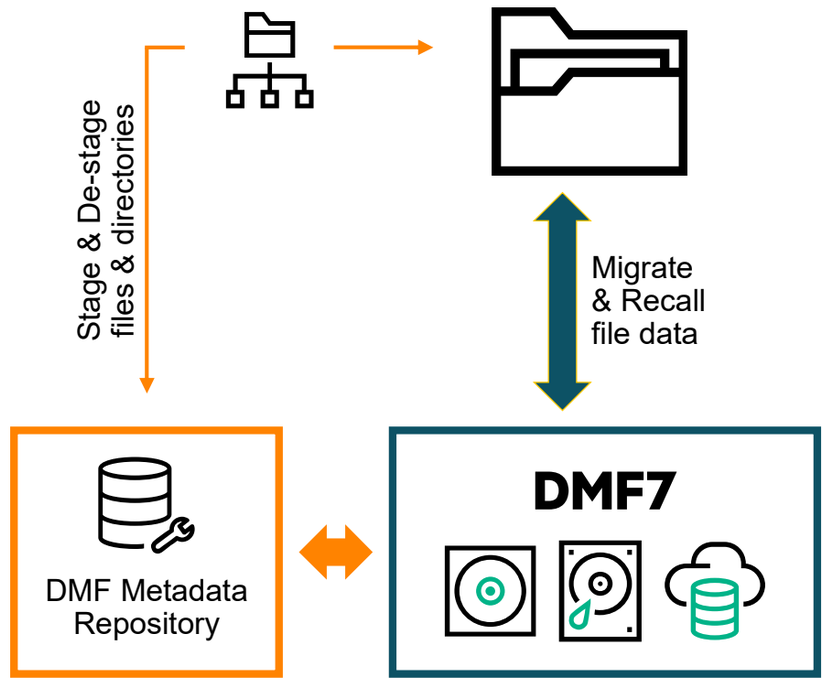
– Scalable for capacity and performance

– Flexible new ways to manage data with the ability to create/delete/recover namespaces



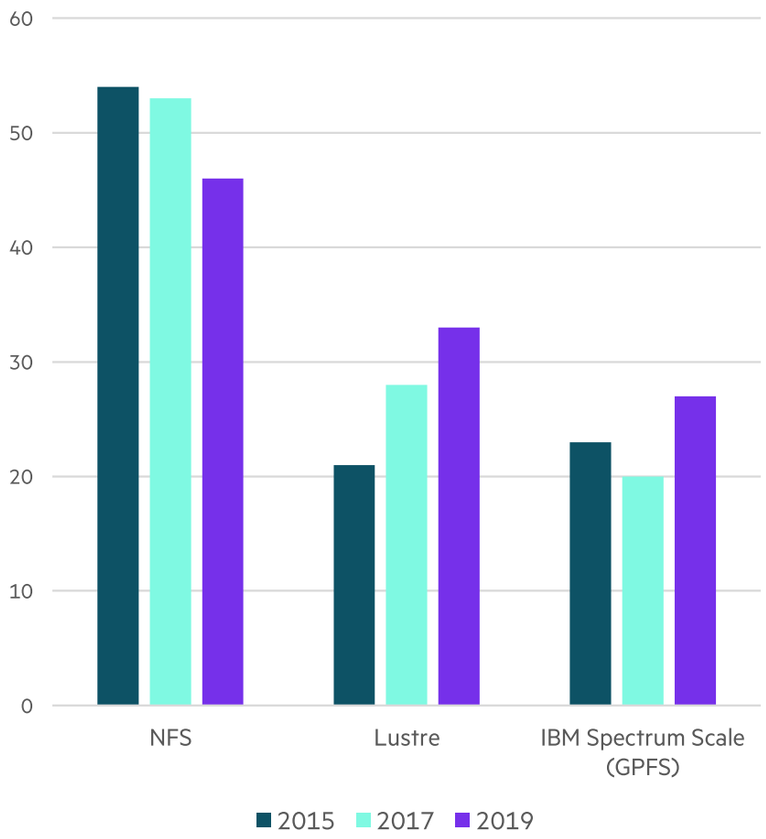
FILE SYSTEM MANAGEMENT
What Can DMF 7 Do?

- 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, GPFS 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



THE RISE OF LUSTRE IN HPC

HPC file system trends: Top 3



- While **NFS** remains the most widely adopted file system, it has **dropped from being utilized at 54% of the sites down to 46%.** NFS was one of the first file systems that could handle the initial scale of early HPC systems. Its first mover status, coupled with the fact that it's still adequate in smaller scale HPC systems today, is reflected in its continued, **albeit shrinking**, wide adoption.

HPE offering: Scale out NFS clusters or Qumulo

- **Lustre** utilization has grown from 21% to 32.5%, and Lustre's open source approach has enabled it to **mature its feature set across a wide number of areas including performance, resiliency, reliability, and scalability.** This maturity has also given it stability and the capability of scaling to meet the demands of petascale and emerging exascale configurations.

HPE offering: Cray ClusterStor E1000

- **GPFS/Spectrum Scale** adoption has grown at a slightly lower rate, from 23% to 26.8%. This modest growth in adoption can be attributed to a **slower feature advancement pace due to its proprietary nature** along with a **market shift away from IBM's dominance in the HPC sector** and a **change in the pricing model** away from user-license-based to capacity-based.

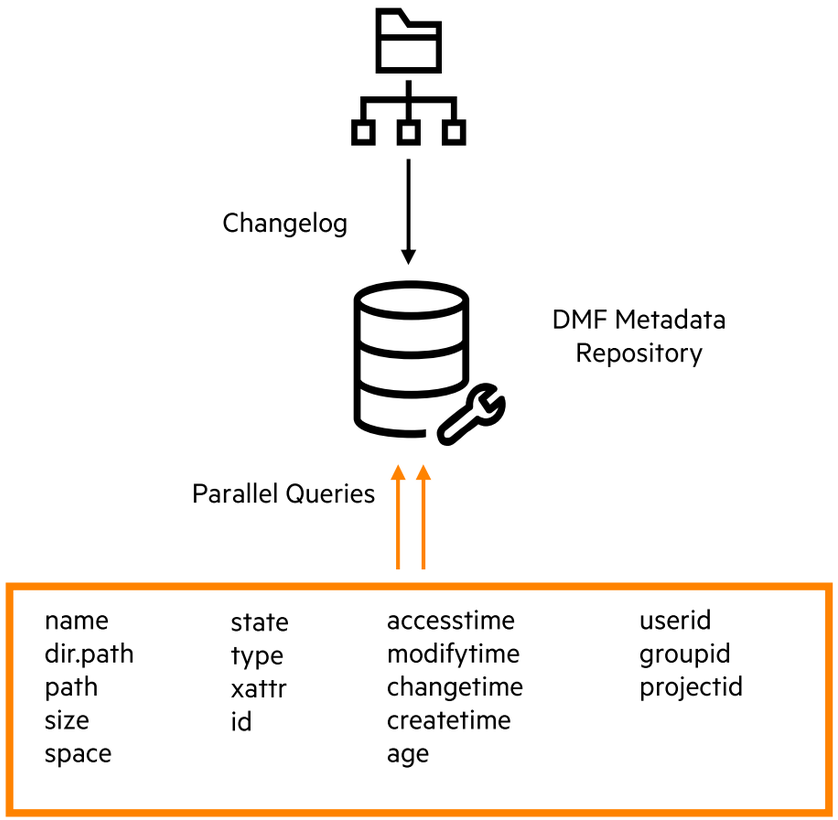
HPE offering: SpectrumScale with Erasure Coding

Hewlett Packard Enterprise | intel

Source: Hyperion Research, Special Study: Shifts Are Occurring in the File System Landscape, June 2020

FILE SYSTEM MANAGEMENT
Reflection of Managed Namespaces

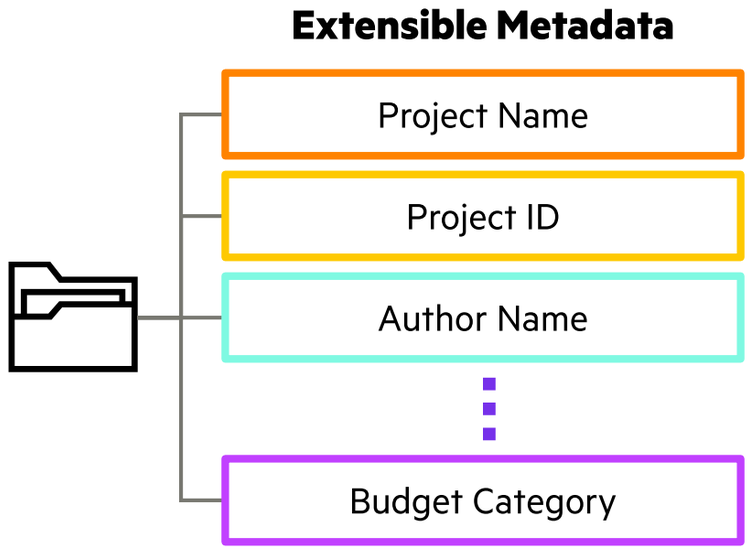
- DMF 7 captures file-system metadata changes via the changelog stream and stores them as a set of tables in a highly scalable database.
- Metadata Protection
  - DMF migration preserves the namespace along side the data.
- Parallel Metadata Queries
  - Using big data tools, such as Apache Spark, enables parallel metadata queries on extremely large data sets.
- Filesystem Offload
  - Allows for indexed metadata searches, data versioning, and policy execution without putting additional strain on the system.



EXTENSIBLE METADATA SUPPORT

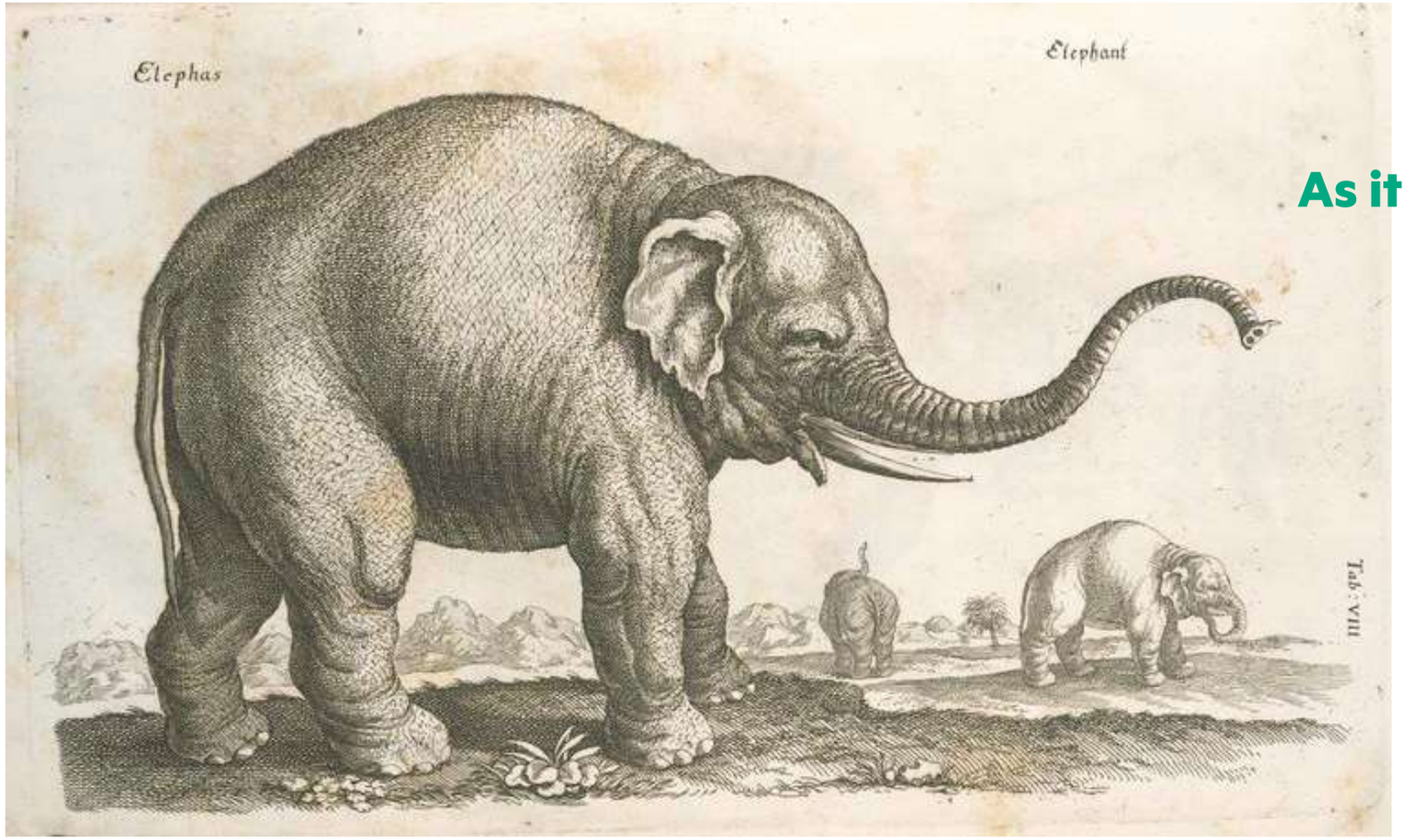
Data management flexibility and precision with extensible metadata

- DMF v7 is based on scalable metadata repository
- Repository functions as 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.



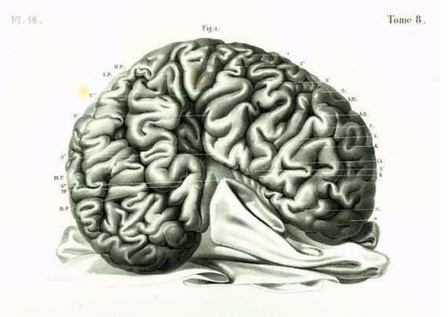
DISCOVERY, PROVENANCE AND REUSE

Metadata is key to these research activities



Data

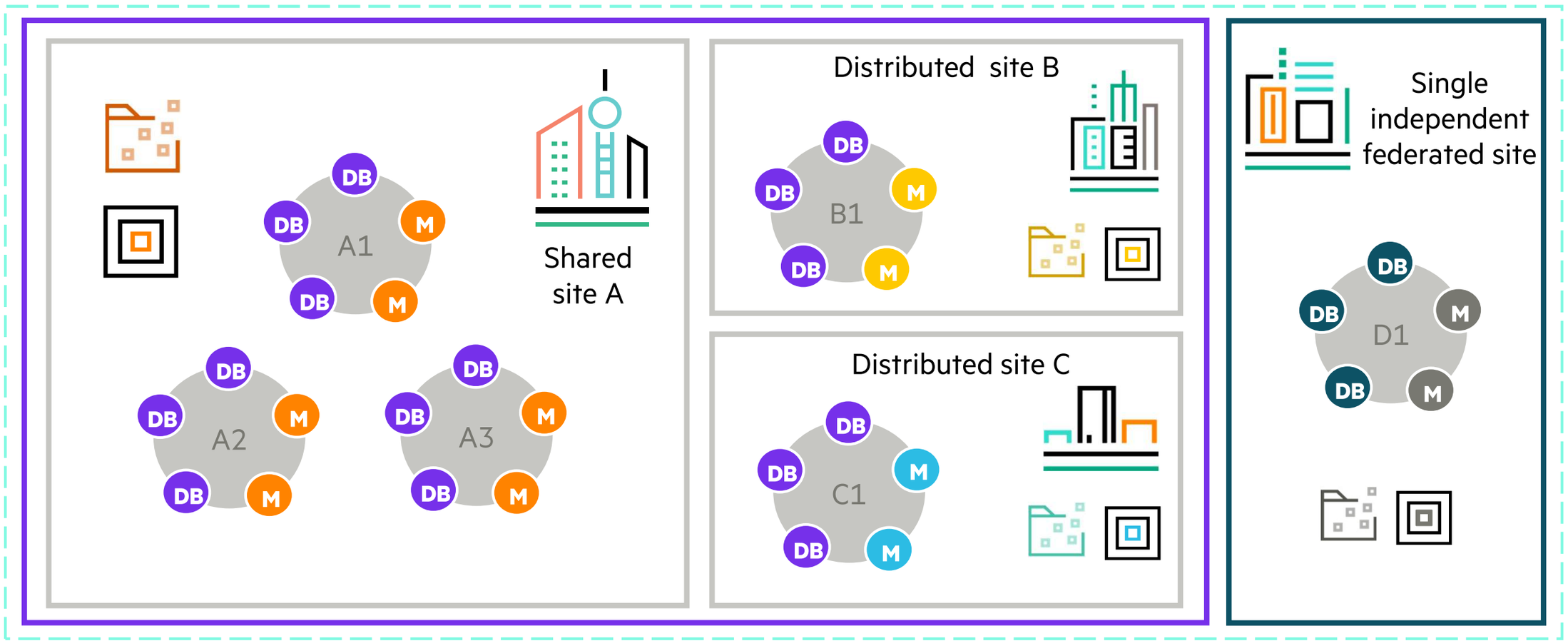
As it is to scalable data management



Metadata

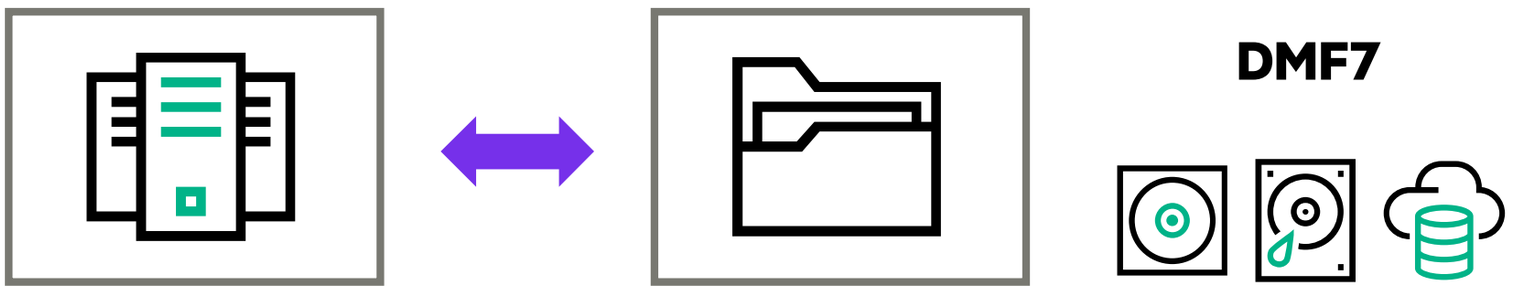
Hewlett Packard Enterprise | intel

FEDERATED DMF



INTEGRATED BACKUPS AND FILE VERSIONING

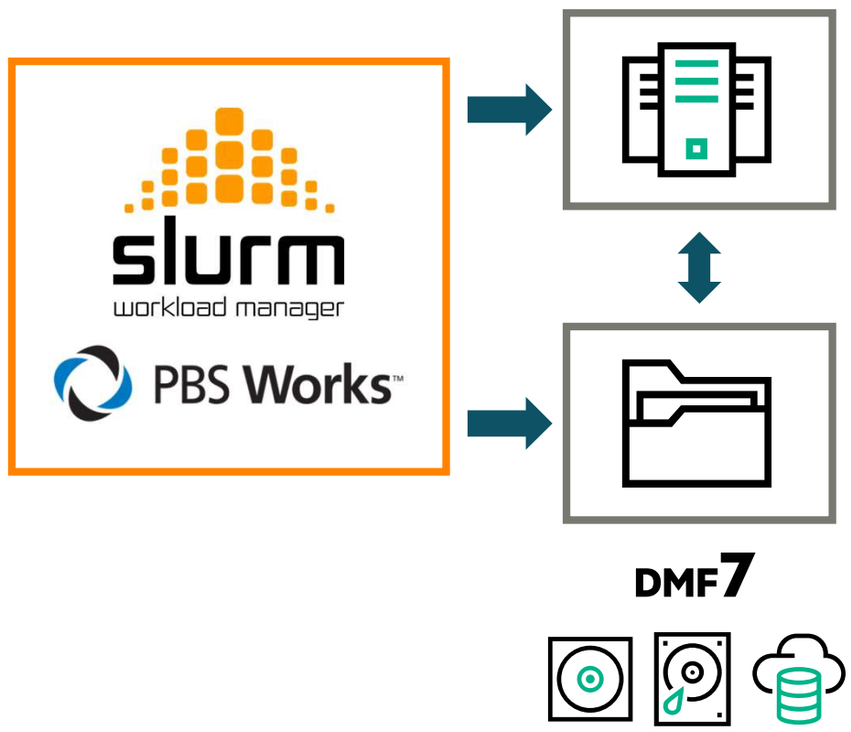
- Large volumes of data can be backed up via automated migrations without impacting users
  - Policies trigger migrations of all modified files & directories at specific intervals or after periods of inactivity
  - Filesystem namespace is periodically captured into backup datasets
- Users can view past versions of backup data sets and individual files and restore the desired version
  - Complete history of the evolution and contents of file systems maintained by DMF v7
- Replication of results for specific job runs, or for validating the correct operation of modified system codes, is enabled via Point in Time restoration of file systems from backup datasets



JOB SCHEDULER INTEGRATION

DMF v7 jobs can be scheduled via standard HPC job schedulers:

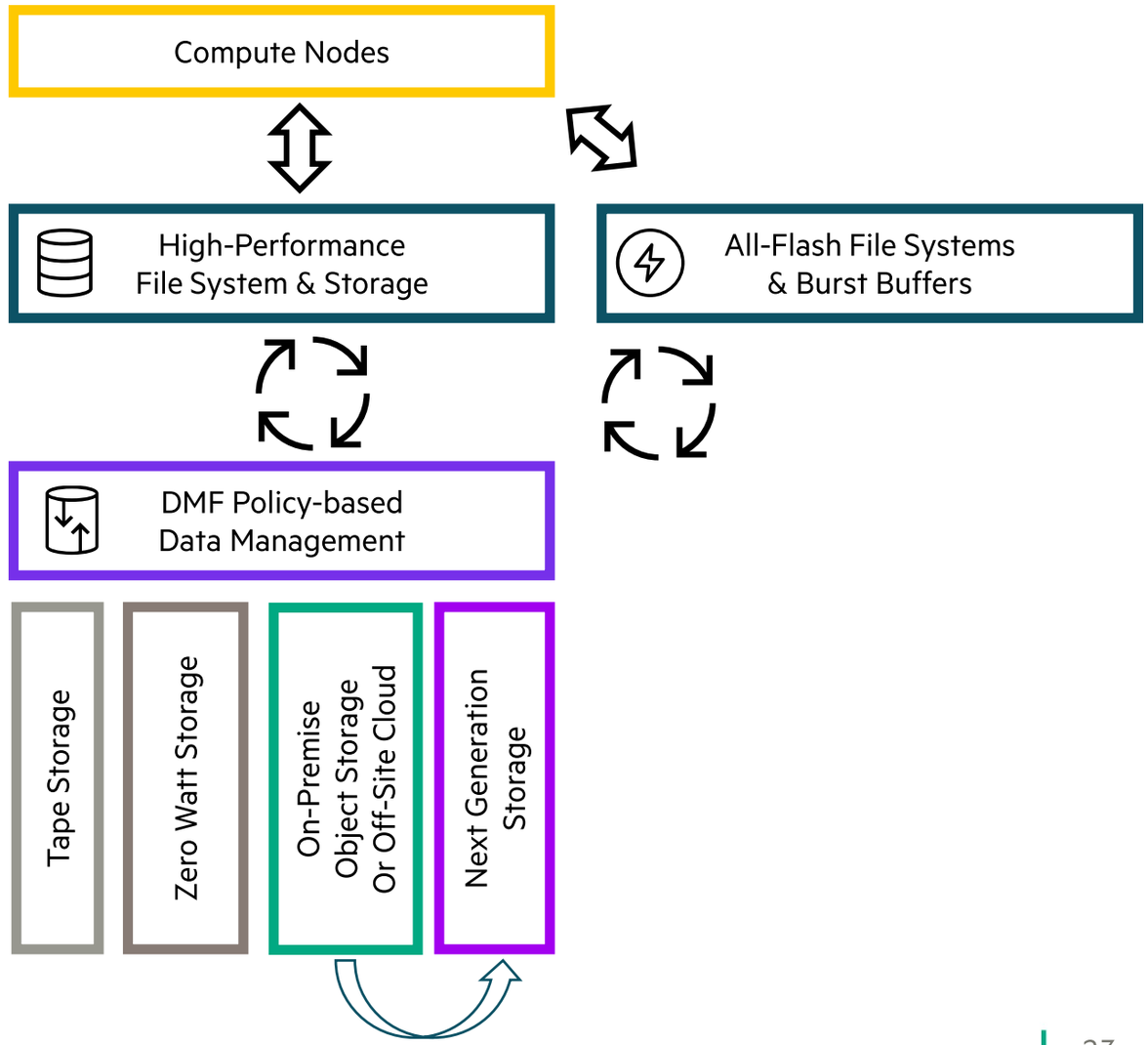
- DMF v7 is API-enabled for integration with job schedulers, e.g. Slurm, PBS Works
- Data set “labels” can be defined by an administrator to “name” a data set
  - Simplifies job management and reproducibility of results in the future.
- Job scheduler definitions can include information to place data sets on the fastest tier of storage in advance of job initiation
- After a job is done, its data set can be de-staged and migrated to a designated tier by policy



FUTURE TECHNOLOGY INTEGRATION

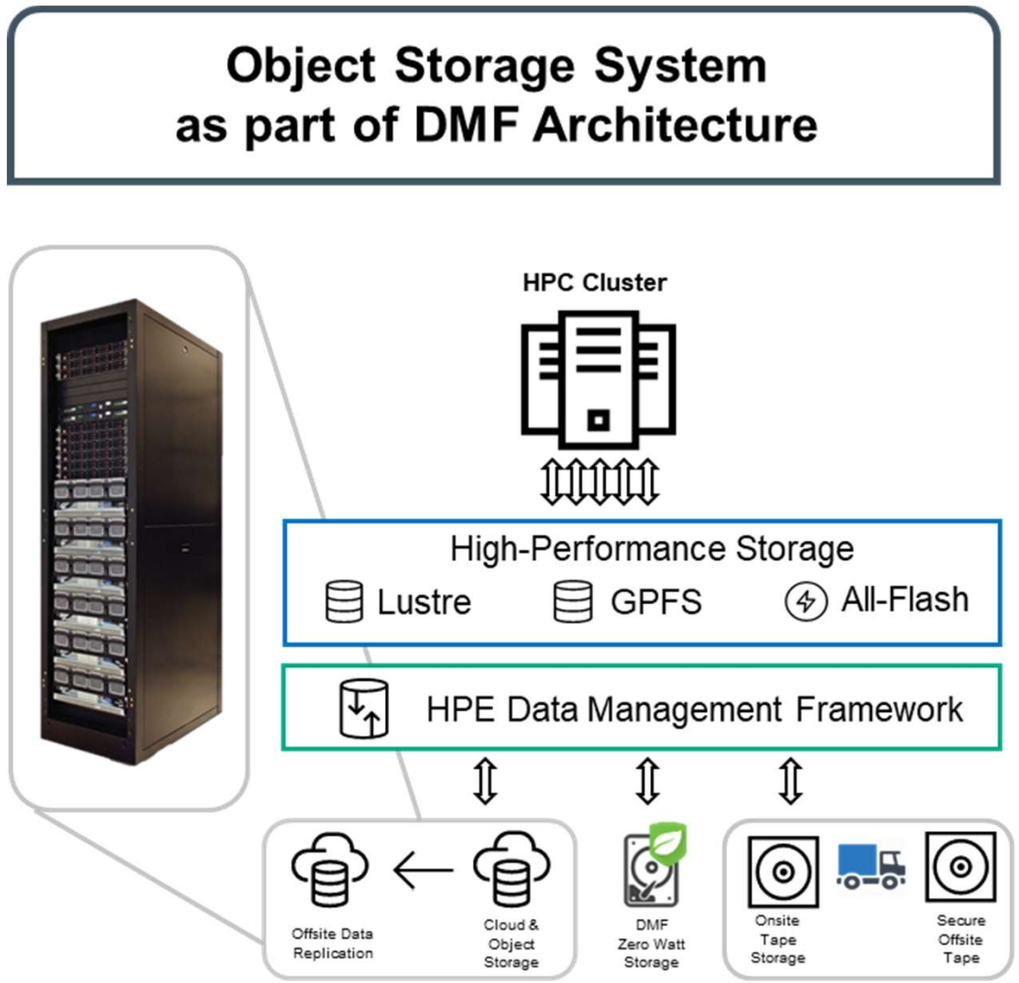
Manage introduction of new storage technologies over time without disruption

- Seamlessly manage migration, validation and consolidation of massive data sets
- Perform the migration over a period of weeks or months with no impact to user data access
- Stage managed data to burst buffers or all-flash filesystems



DMF OBJECT STORAGE SUPPORT

- **Standards-based Integration:**
  - Use of S3 interface enables compatibility with Scality, CEPH, Amazon S3, DDN WOS, HGST Active Archive, NetApp StorageGrid, DELL|EMC ECS, and open source alternatives
- **Scalability & Throughput:**
  - Scalable DMF connections to object storage environment
  - DMF Parallel Data Mover architecture with high availability, balancing and failover
- **Flexibility:**
  - Ability to blend object storage with alternative storage options including Zero Watt Storage (performance) or tape (off-site disaster recovery)



DMF TAPE STORAGE INTEGRATION

- DMF is certified with libraries from HPE, as well as Spectra Logic, IBM, Oracle, Quantum and Overland
  - Streams to tape drive at native rates, even for small files using packing
  - Block ID positioning for fast seek
- Support for latest LTO-8 and TS1160 Enterprise 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 LTO drives
  - Recommended Access Order (RAO) and SpectraLogic's TAOS
- RoCEv2 drive support coming

LTO-8 12TB
360MB/s

TS1160 20TB
400MB/s

ZERO WATT STORAGE

Software-defined DMF warm storage tier with optimised power utilisation

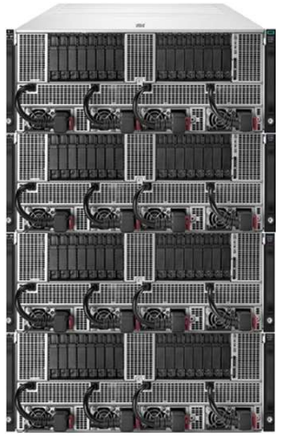
4RU
127 kg

HPE D8000 JBOD

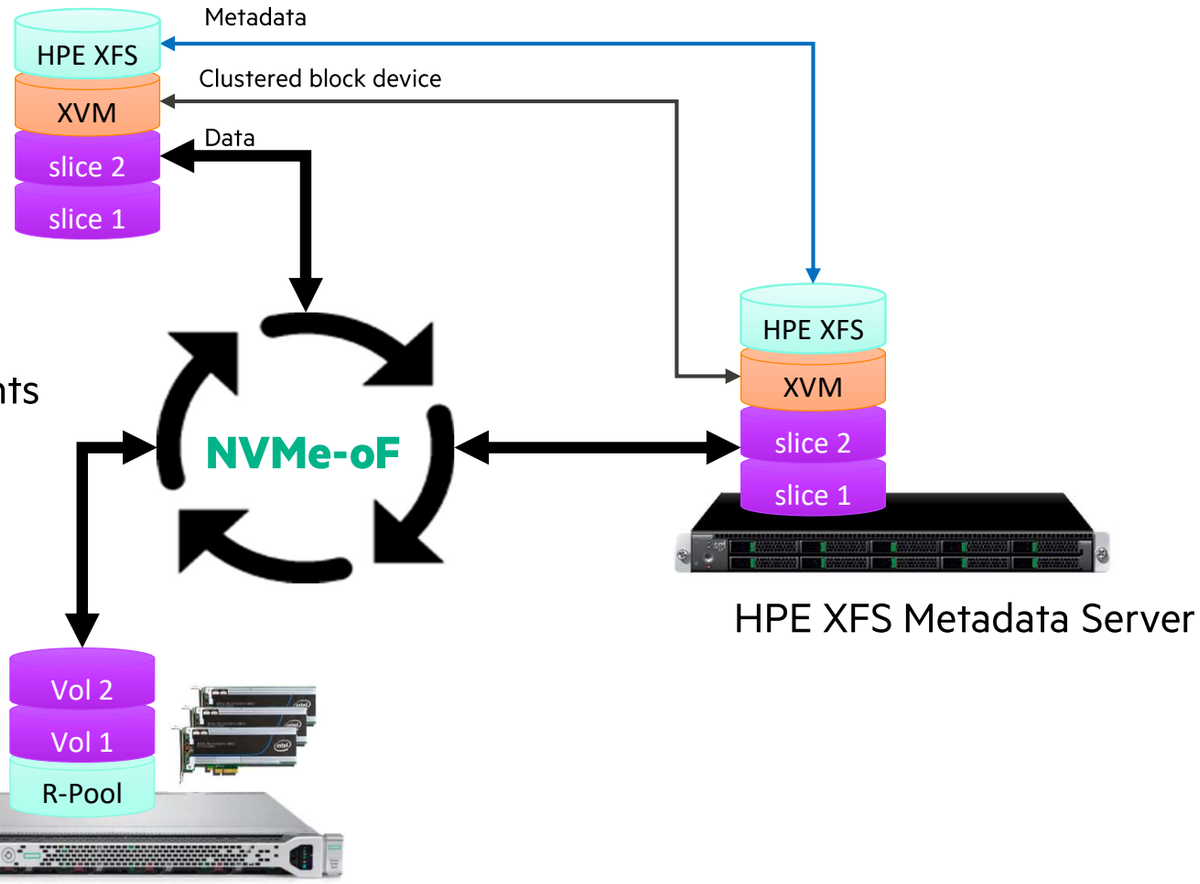
- 4U High Density Enclosure
- 106 3.5” Drives per Enclosure
- Up to 8 Enclosures (848 drives) per ZeroWatt Scalable Unit
- 1.5 PB DMF ‘Single Copy’ Usable per Enclosure
- 13.5 PB DMF ‘Single Copy’ Usable per Scalable Unit (max)
- **Individual Drive Spin Down**
- Every HDD is a vTape Drive

HPE TIERZERO & RPOOL

Burst buffer storage tier on all flash 1RU servers



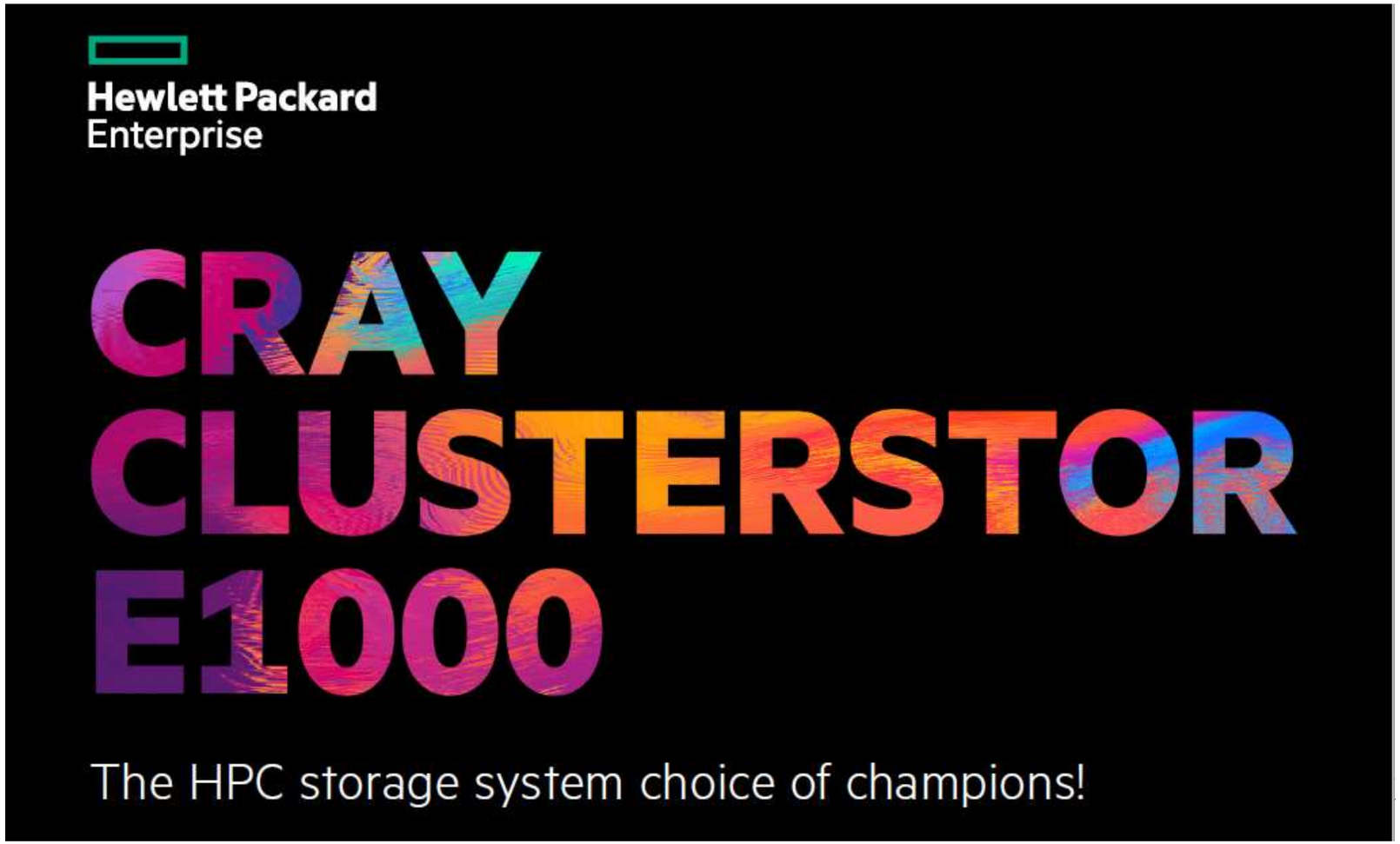
Compute Node clients



HPE XFS Metadata Server

NVMe server(s)

PARALLEL FILE SYSTEM APPLIANCE



https://h20195.www2.hpe.com/v2/Getdocument.aspx?docname=a50002176enw

One 19" rack has 10 PB of usable storage capacity[4]
and an aggregate throughput rate up to 1.4 TB/s[5]

To put that into perspective: One Cray ClusterStor E1000 rack could...

...download all movies on Netflix U.S. in

20
SECONDS

...complete one input/output operation for every person in Texas in

1
SECOND

...store as much data as a row of 4-drawer filing cabinets lined up side by side

172
TIMES AROUND THE EARTH

Hewlett Packard Enterprise | intel

DATA MANAGEMENT FRAMEWORK

DMF Use Cases

Bottomless archive & backup

Control Storage Costs

- Organizations are faced with tens of petabytes of research data that they need to store & protect
- DMF can cost-effectively and transparently store & protect massive amounts of data by copying & tiering to lower cost storage, such as object storage and High-Latency Media (HLM)

Holistic data management

Manage complex datasets

- Modern workflows require moving data between compute clusters, training/inference platforms and active archive, introducing multiple namespaces
- DMF can manage multiple filesystems using policies to coordinate data transfers between multiple filesystems, object storage and HSM backend

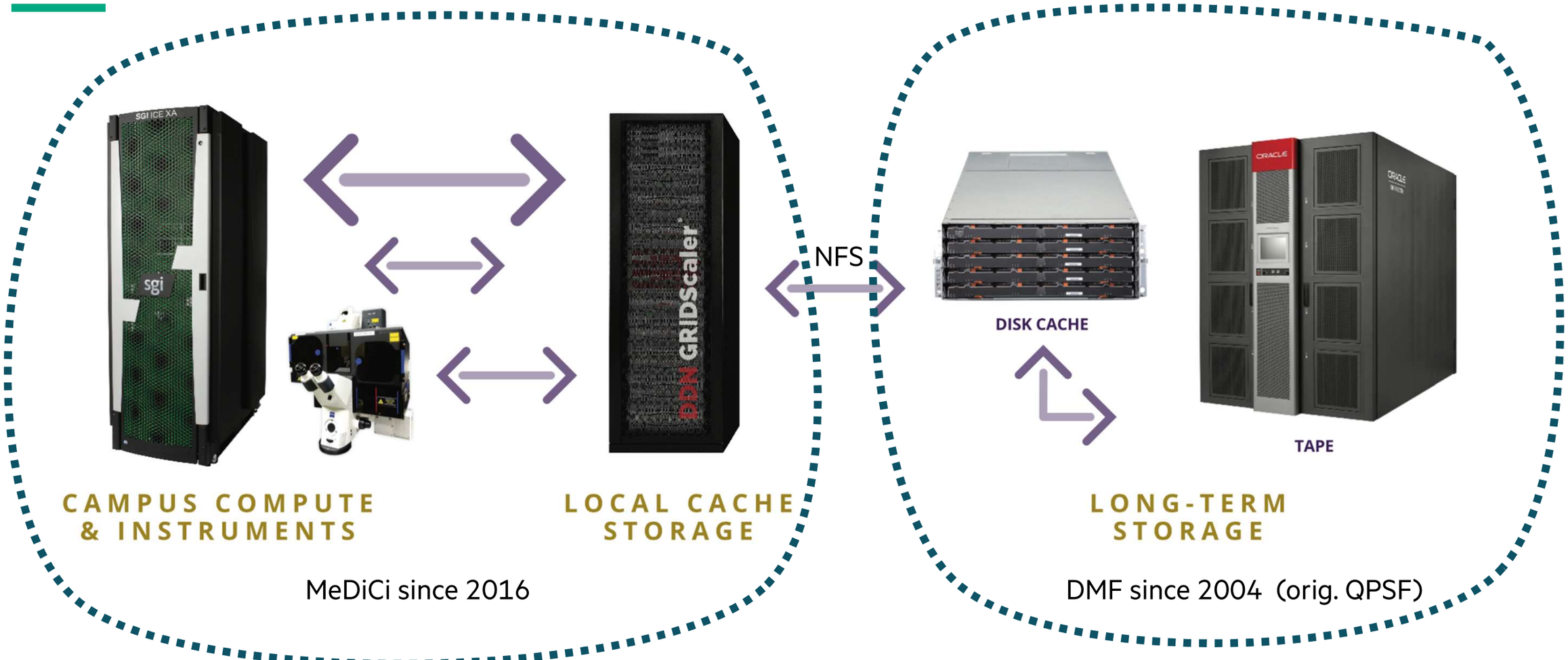
Challenging workloads & burst buffers

Accelerate workflows

- HPC users typically request 5-10% of total scratch capacity as high performance flash storage for bandwidth- or IOPS-intensive workloads
- DMF can stage files into a flash-based dynamic namespace created on demand by HPC scheduler or workflow manager, and then de-stage results into backend and release the namespace

CASE STUDY:

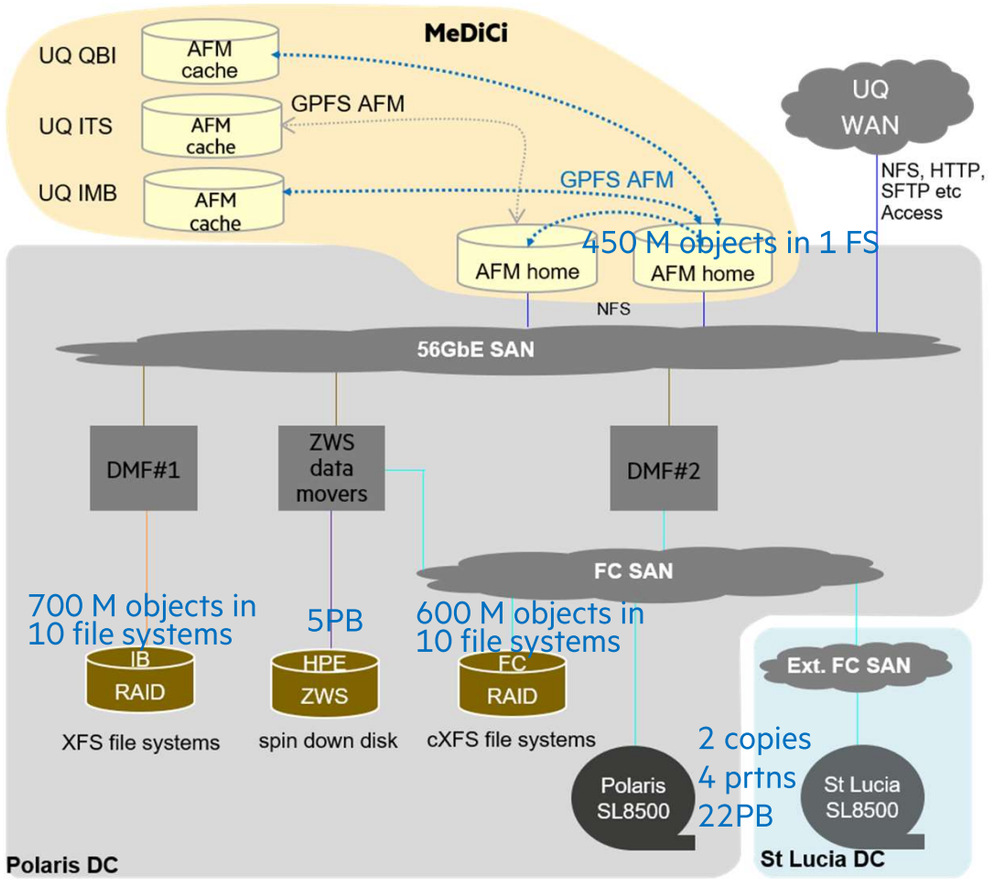
UQ's Research Computing Centre (RCC)

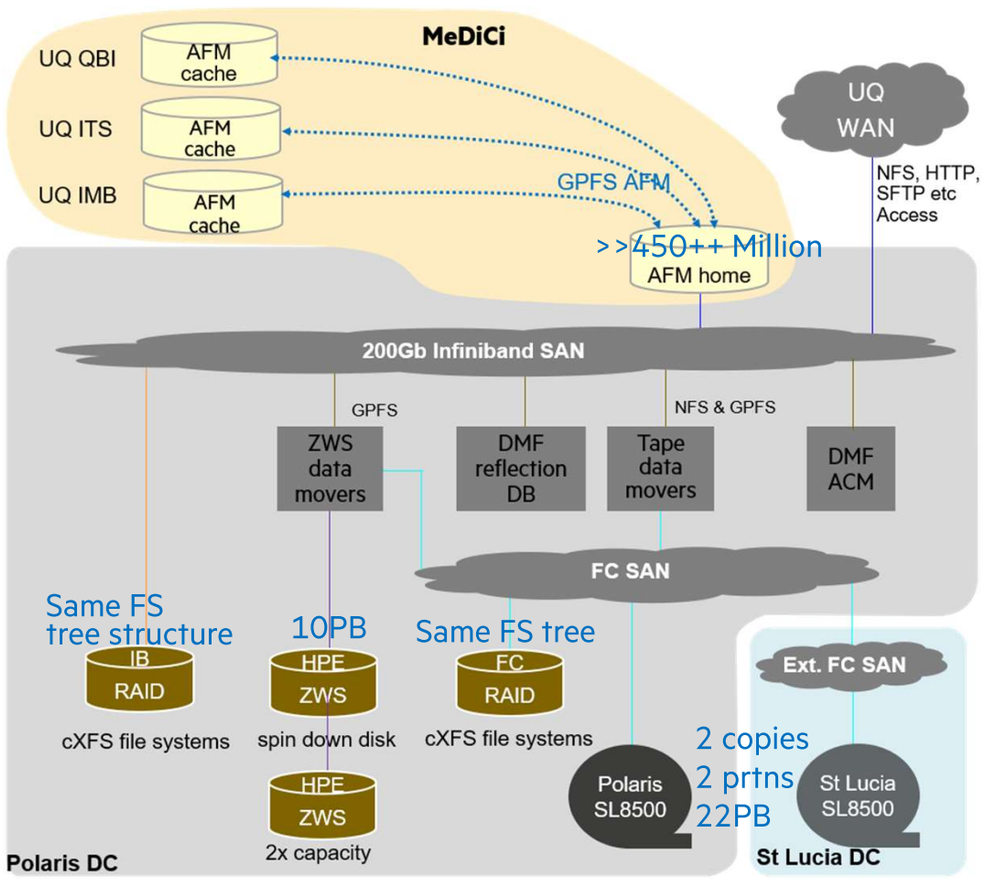


Hewlett Packard Enterprise | intel

CASE STUDY:
UQ RCC DMF6 to DMF7 upgrade

Metadata migration with data in place.





Before
- AFM home GPFS cluster nodes are NFS clients of DMF managed file system

After
- DMF data mover nodes are clients of AFM home GPFS cluster

Hewlett Packard Enterprise | intel

CASE STUDY:
UQ RCC DMF6 to DMF7 upgrade

Outcomes

- Adhered closely to the plan schedule that was based on timing estimates obtained during dry runs
- 1.3 Billion files and directories accessible immediately following cut-over including soft deleted files (not a byte was lost)
- Single namespace
- Increased power and flexibility to query 20 file system's metadata and HSM repository objects
- 8PB of online data and 22PB of nearline (tape) data accessible immediately following cut-over
- Data recovered (user deleted files) days after cut-over
- Faster servers and networks, new firmware, latest OS and DMF software
- Risk mitigation due to fully reversible upgrade process - only metadata was migrated (data-in-place upgrade)
- Fast secondary tier (ZWS) capacity doubled from 5PB to 10PB
- Direct data transfer over native GPFS protocol will ultimately replace NFS hop to and from long term storage
- Increased scale will allow researchers to grow data collections with enterprise grade data protection and disaster recovery

Questions?

THANK YOU

david.honey@hpe.com

Hewlett Packard Enterprise | intel