Managing trillions of objects
Supercomputing Asia 2018
David Honey
March 2018
Improved Data Management

➢ When do I need it?
  ▪ When the volume of data stored exceeds what can be managed by available resources
  ▪ When the number of objects exceeds what current tools are capable of handling
  ▪ When too much effort is needed to make data accessible
  ▪ When manual tasks are repeated over and over
  ▪ When delays getting access to data erodes the benefit of storing it
  ▪ When staff can’t cope with the range of tools needed to manage data
  These are especially true where the value of data increases over time

➢ Why do I need it?
  ▪ ‘Data is the new Intellectual Property’*
  ▪ To stay ahead of the competition
  ▪ Successful companies extract business value from data
  ▪ To maximise return on investment
  ▪ To deliver new capabilities based on new technology
  ▪ To reduce the cost of compliance
  ▪ Because future AI applications will extract more value from data, but only if they can get at it

* From Moneyball by Michael Lewis
It’s not just about FLOPS
- It’s about efficiently sharing active data sets

Understanding job IO requirements
- Enables co-scheduling to optimise utilisation
- Improves successful completion
- Accelerates performance

Having data in the right tier is key
- High performance tier at job initiation
- High resilience tier when results are ready
- Purge from high performance tier when done
You know it’s time for change when;

- Users spend more time searching for data than doing research
  - Walking the directory tree takes days
  - Terminals disconnect before resultants are returned

- Users lose data
  - If there’s no way to recover from ‘operator error’

- Regular operations never finish
  - Data isn’t backed up
  - Quick search indexes aren’t up to date

- Maintenance eats further into production time
  - File System checking (fsck) runs for 6+ hours

*Machine learning workflows will compound these issues*
Performance vs Capacity Tiers | Active & Dormant data

- **Active**
  - HPC codes require POSIX
  - Buffered or Direct I/O
  - MPI-IO libraries
  - Highest random I/O performance
  - Partial namespace
  - Flexible or no data redundancy

- **Idle**
  - Good streaming performance
  - Very large namespace
  - Strong data protection
  - Multi-site support
  - Geo replication & Backup
  - Cloud integration

- **Dormant**
## Physical Storage Tiers | Performance bands

<table>
<thead>
<tr>
<th>Tier</th>
<th>Role</th>
<th>Objects</th>
<th>Technology</th>
<th>Architecture</th>
<th>Data residency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>Computation</td>
<td>10,000’s</td>
<td>Local DIMMs</td>
<td>Dedicated to node, highly transient</td>
<td>Seconds, minutes Overwritten continuously</td>
</tr>
<tr>
<td>Tier 0 Burst buffer</td>
<td>Very fast POSIX file system</td>
<td>100,000’s</td>
<td>Shared NVMe drives</td>
<td>Dynamically populated, job specific data on shared storage</td>
<td>Hours/days/weeks Overwritten at job end</td>
</tr>
<tr>
<td>Tier 1 Scratch</td>
<td>High Performance POSIX file system</td>
<td>10,000,000’s</td>
<td>Parallel File System on RAIDed SSD, HDD hybrid</td>
<td>Transient data on shared storage, dedicated to cluster</td>
<td>Flush regularly</td>
</tr>
<tr>
<td>Project Storage</td>
<td>Fast object store for infrequently accessed data</td>
<td>100,000,000’s</td>
<td>Large HDD storage with parity</td>
<td>Global FS</td>
<td>Flush at project end</td>
</tr>
<tr>
<td>Tier 2 Cold Storage</td>
<td>High performance archive cache</td>
<td>1,000,000,000’s</td>
<td>Zero Watt Storage Power down HDD</td>
<td>Nearly online</td>
<td>Years</td>
</tr>
<tr>
<td>Tier 3 Archive</td>
<td>Inactive data</td>
<td>1,000,000,000,000’s</td>
<td>Object or Tape</td>
<td>Nearly online</td>
<td>Years, decades</td>
</tr>
</tbody>
</table>

* Indicative Only
HPE DMF automates data movement between all Tiers

Storage Technologies | Relative Costs

- High IOPS and Throughput
  - SSD Flash
  - Block-based RAID solutions
  - File & Object-based Network Attached Solutions
  - PFS

- Performance / Cost
  - NAS

- Longevity and Protection
  - MAID
  - Zero-Watt Disk Solutions
  - Automated Libraries

- Data Volume
  - Tape
Data Management | Automatic Cost Optimised Data Placement

Tiered Virtual Storage Pool

Tier 1 Disk
$$/TB

Tier 2 Disk
$/TB

Tier 3
¢/TB

Primary Storage
Fast Nearline
Slow Nearline

Used Space
Data Management | Automatic Cost Optimised Data Placement

Administrator-Defined Free Space Thresholds

- Tier 1 Disk: $$/TB
- Tier 2 Disk: $$/TB
- Tier 3: $$/TB

Primary Storage
- Fast Nearline
- Slow Nearline

Used Space
Data Management | Automatic Cost Optimised Data Placement

New data automatically migrated to lower tiers

Primary Storage

Tier 1 Disk
$$/TB

Tier 2 Disk
$/TB

Tier 3
¢/TB

Fast Nearline

Free Space Thresholds

20%

80%

Slow Nearline

= Used Space
When Free Space reaches 20%

- Tier 1 Disk: $$/TB
- Tier 2 Disk: $$/TB
- Tier 3: ¢/TB

Free Space Thresholds:
- Primary Storage: 20%
- Fast Nearline: 80%
- Slow Nearline: 20%
Data Management | Automatic Cost Optimised Data Placement

Dual state files automatically offlineed

Tier 1 Disk
$$/TB

Tier 2 Disk
$/TB

Tier 3
¢/TB

Primary Storage
Fast Nearline
Slow Nearline

20%
80%
Free Space Thresholds
Data Management | Automatic Cost Optimised Data Placement

And space is reclaimed for new data

- Tier 1 Disk: $$/TB
- Tier 2 Disk: $/TB
- Tier 3: ¢/TB

Primary Storage | Fast Nearline | Slow Nearline
Data Management | Automatic Cost Optimised Data Placement

Repeat

Tier 1 Disk
$$/TB

Tier 2 Disk
$/TB

Tier 3
¢/TB

Primary Storage
Fast Nearline
Slow Nearline

Free Space Thresholds
20%
80%
Data Management | Automatic Cost Optimised Data Placement

Countless times

<table>
<thead>
<tr>
<th>Tier 1 Disk</th>
<th>Tier 2 Disk</th>
<th>Tier 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$$/TB</td>
<td>$$/TB</td>
<td>¢/TB</td>
</tr>
</tbody>
</table>

Primary Storage | Fast Nearline | Slow Nearline

20% 80%
Free Space Thresholds
Data automatically recalled on access

Tier 1 Disk
$$/TB

Tier 2 Disk
$/TB

Tier 3
¢/TB

20%

80%

Free Space Thresholds

All files appear online in original location
Data Management | Automatic Cost Optimised Data Placement

Costs Optimised over Time and TCO Dramatically Reduced

Tier 1 Disk
$$/TB

Tier 2 Disk
$/TB

Tier 3
¢/TB

All files appear online in original location
HPE is bringing the experience from SGI in managing some of the world’s largest data environments to the growing needs of today’s AI market.

- **Data life cycle management**
  - DMF manages data placement on heterogeneous tiers of storage

- **Key Benefits**
  - Open, published interfaces
  - Protects data long term with minimal overhead
    - Metadata only backups
    - Automatic CRC checking
    - Hardens archives against wholesale corruption events; hardware failures,
    - Write once - read seldom up to 63 copies with immutability
  - Simplifies management of large data collections
    - Immensely scalable
    - Policy driven
    - Daily audit with exception reporting
    - Transparent, automated storage technology refresh
  - Reduces RAID disk investment and power consumption
  - Open, published interfaces

- **DMF™ 28 years in production**
- **Active User Community (DMFUG) based in Australia**

---

**Testimony | Data Management Facility**

**Proven in production use for nearly 3 decades**

---

**HPE is bringing the experience from SGI in managing some of the world’s largest data environments to the growing needs of today’s AI market.**
Next Generation
Data Management Framework - DMF
**DMF | Design Objectives**

1. Integrate with High Performance flash
2. Scalable Metadata – 100s of Trillions of files
3. Capture T1 filesystem events in Real-time Changelog to avoid scans
4. Provision Namespaces On Demand by job scheduler in addition to Static Filesystems
5. Optimize Data Transfer by Chunking and Parallel Data Movers
6. Keep Transparency. Replace Traditional Backup
DMF | Architected for Big Data

- Native Changelog consumers for HPC file systems
- High Performance Distributed Metadata db
- Distributed in memory Scheduler
- Distributed Query Engine

Dynamic or Static Namespaces in T0 or T1

Filesystem metadata mirrored in real-time to a scalable database

Refer to the DMF 7 Getting Started Guide here:
Authoritative Metadata Repository

DMF 7 is based on a scalable metadata repository that is the long term data store for information about file system structure, attributes, contents and evolution over time.

- **Reduced Reliance on inodes**: In the event of file system corruption or accidental data deletion, both the file system structure and contents can be recovered using DMF objects and metadata information.

- **Extensible Metadata Information**: The DMF metadata repository is extensible to enable the storing of additional file metadata that can include information on projects, job runs, researcher/lab details and data retention requirements.

- **Enhanced Privacy and Protection**: File systems not in active use may be completely de-staged and removed from active access.
Filesystem searches can be done through the database with much better scale, speed, and flexibility.

dmf find fs:xyz --query "file.size > 1 mb"
dmf find fs:xyz --query "file.state = REG"
dmf find fs:xyz --query "dir.path = '/foo/bar'"
dmf find fs:xyz --query "file.xattr['user']['tag'] = 'foo'"

Queries can be compound using SQL-like syntax;

dmf find fs:xyz --query "dir.path = '/foo/bar' and file.state = OFL and file.size > 1 mb"

Queries invoke data operations for example ‘put’ triggers policy based migration.

dmf put --query "dir.path = '/foo/bar'" --fsid:xyz

Queries with data operations can also assign datasets names;

dmf put --query "dir.path = '/foo/bar'" --fsid:xyz --set foo
dmf destage --query "dir.path = '/foo/bar' and file.state = REG and file.size > 1 mb" --fsid:xyz --set bar

Once a dataset name is assigned, users can operate on the data set;

dmf find obj --query "object.tags contains 'foo'"
dmf stage --set bar /somewhere/new
DMF | File System Integration

Native Integration
DMF 7 is tightly-coupled integration with HPC file systems

- **Accelerated Awareness**: Combined HSM API and change log approach provide near real time DMF awareness of file system events such as file creation/deletion, file modification and file access by users and applications.

- **Continued Support for HPE Clustered Extent File System**: DMF 7 will provide seamless support of HPE’s shareable file system including additional features such as supporting native integration without the need for periodic file system scanning.

- **No Requirement for the Robinhood Engine (Lustre)**: DMF’s ability to perform native Lustre change log processing and metadata management eliminates the requirement for the Robinhood component and separate tracking database.

- **Roadmap to Native Spectrum Scale (GPFS) Integration**: Planned future releases of DMF 7 will provide native integration with Spectrum Scale using the existing GPFS DMAPI application interface.
Continuous Backups and File Versioning
DMF 7 policies may be configured with policies to store copies of both new and updated files to the capacity tier at specific intervals or after certain periods of file change inactivity.

• **High-Quality Backup:** DMF will automatically make backup copies of files and metadata on a rolling basis using policy definitions that can be tailored by administrators.

• **File Versioning:** DMF maintains both metadata information and file data from prior versions of files so that administrators have a complete history of the evolution and contents of file systems.

• **“Point in Time” Restoration:** Administrators can re-stage file systems – or portions of file systems – using a point-in-time designation for use in replication of results for specific job runs or for validating the correct operation of modified system codes.
Job Scheduler Integration

DMF 7 integrates with job schedulers for operations such as data pre-staging to a high-performance flash tier in advance of job execution.

- **Data Pre-Staging or Recall Based on Metadata**: Job scheduler definitions can include information on required data sets that should be on the fastest tier of storage in advance of job initiation.

- **Data Set Definition**: Job administrators can define labeled "data sets" that are a collection of files/directories associated with a specific job type. This process can simplify job management and enable more easily reproducible results in the future.

- **Data Migration or De-Staging After Job Completion**: Data can be migrated or de-staged from high-performance storage based on automated policies – or job administrators can direct the system to migrate or de-stage data after job completion.
Data Management | **DMF Dynamic File Systems**

- **Job Scheduler**
  - PBS Professional
- **Compute Nodes**
  - Execution of Codes
- **High-Performance Storage**
  - Data & Permissions Populated
- **Capacity Object Storage Pool**
  - Encrypted Pool of All Data Objects
- **Metadata Repository**
  - Data Set 277

Decrypt and Populate Data

Hewlett Packard Enterprise
Data Management | **DMF Dynamic File Systems**

- **Job Scheduler**
  - PBS Professional
  - Job Complete

- **Compute Nodes**
  - Execution of Codes

- **High-Performance Storage**
  - Data & Permissions Populated

- **Capacity Object Storage Pool**
  - Encrypted Pool of All Data Objects

- **New Data Encrypted & Updated Object Versions**

- **Changes?**

- **Metadata Repository**
Global Metadata Namespace
DMF 7 metadata repository is based on a modern distributed database that can span multiple sites.

- **Multiple sites supported:** Distributed database allows multiple sites to participate in the metadata replication, provide view of distributed objects and support metadata queries.

- **Replication via cloud or object storage:** Once a copy of data is in the cloud and the object record is replicated, any DMF 7 enabled site can stage these data objects in a local filesystem.

- **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, data can be pre-fetched into ZWS prior to staging to the filesystem.
Zero Watt Storage as Data Cache

DMF 7 can track multiple copies of backend objects. It also supports creating and deleting the copies based on policy or request.

- **Fast migration via Zero Watt**: Migrate and release space on the managed filesystem faster by creating one or two copies in ZWS then creating copies on tape and/or cloud in the background.

- **Pre-stage or recall via Zero Watt**: Bring data from the tape and/or cloud into ZWS first, then quickly stage or recall into the managed filesystem.

- **Fast recall from tape**: Keep one copy of data written to tape in ZWS. When the data is recalled, it will be copied from ZWS, so the tape does not need to be mounted.

- **No-fee recall from cloud**: Keep one copy of data written to cloud in ZWS. When the data is recalled, it will be quickly copied from ZWS without incurring cloud data retrieval fees.
Software-based DMF warm tier storage option with minimized power utilization paired with the HPE D6020 JBOD

High Performance
Power Optimized
Extended Drive Lifespan

- 12 Gbps SAS JBOD performance provides ‘fast’ hard disk tier to stream data to active ‘hot’ storage
- Each drive is individually managed by DMF to track data activity and data layout
- Drives can be spun down when not in use to minimize power consumption and increase drive lifespan
- HPE D6020 5U 70 bay JBOD first to be qualified
Scale-out NFS v3, v4 and SMB File Serving
Tier Zero enables high-bandwidth file serving with multiple NAS heads for massive scalability.

- **Multiple NAS Heads:** Each serves up Tier Zero Namespaces via NAS protocols, so compute nodes can see all files in the NAS space. Clients use standard NFS/SMB drivers supplied by all OS vendors.

- **Growth in two dimensions:** Storage size and I/O performance. Grow by adding Tier Zero nodes with flash to increase capacity or grow by adding NAS servers and I/O ports to increase performance.

- **High availability:** Designed so there is no single point of failure. Scales to 6,000+ clients and supports seamless failover and load balancing.
SOLUTION ATTRIBUTES

• All nodes have full POSIX access to the flash tier and parallel file system

• Flash configuration provides burst buffer capabilities and persistent shareable POSIX file system functionality in a single layer

• For expanded tiered data management capabilities, DMF can tier data from/to this layer into object & cloud storage, Zero Watt buffer storage or tape in order to deliver virtually infinite capacity as well as integrated backup, archive and disaster recovery capabilities
Tier Zero | Linear Performance Scaling Dynamic Namespaces

- 7.5 Million IOPS 25 GB/s
- 15 Million IOPS 50 GB/s
- 22.5 Million IOPS 75 GB/s
- 30 Million IOPS 100 GB/s
- 37.5 Million IOPS 125 GB/s

Storage Capacities:
- 40 TB
- 80 TB
- 140 TB
- 180 TB
- 220 TB

Scaling Multipliers:
- X2
- X3
- X4
- X5

Hewlett Packard Enterprise
**DMF | Vision**

**Scale-Out HPC**
- CFD
- FEA
- CEM
- Chemistry
- Physics
- Scalable applications
- Diverse mixed workload throughput

**Scale-UP HPC**
- I/O intensive jobs
  - Local scratch
  - Out-of-core
  - Per-job private file systems
  - Dynamic shared scratch
  - Tier Zero

**Science Cloud**
- IaaS
- HPCaaS
- Virtual laboratories
- The long tail of HPC
- Industry Outreach
- Portals / collaboration

**Memory Centric Computing**

**Exascale**
- Metadata
  - Version control
  - Scheduler integration
  - Workflow and automation
  - Prestaging of data
  - Query and discovery
  - Analytics and reporting
  - Fabric volume management
  - File system orchestration

**Parallel File Systems**
- High bandwidth shared storage
- Streaming workloads

**Shared flash**
- High performance shared storage
- IOPs workloads

**Bulk disk / MAID**
- Fast recall
- Spin down

**Object Storage**
- Geographic replication
- Multi-site locality

**Tape**
- Secure
- Reliable
- Encrypted
- On-site and off-site

**Hewlett Packard Enterprise**
HPE Scalable Storage with Lustre ZFS

High Performance Storage Solution

Meets Demanding I/O requirements
Performance measured for a Gen 9 building block:
• Up to 17 GiB/s Read/15 GiB/s Writes with EDR
• Up to 16 GiB/s Reads andWrites with OPA
• Up to 21 GiB/s Reads and 15 GiB/s Writes with all SSD

Density Optimized Design For Scale
• Dense Storage Design Translates to Lower $/GB
• Linear performance and capacity scaling

Innovative Software Features

Leading Edge Yet Enterprise Ready Solution
• ZFS RAID provides Snapshot, Compression & Error Correction
• HPE Lustre Monitoring Platform

Services and support

Installation and support services
• Factory tested and validated, deployment services for installation
• 24/7 Support services
Thank you