



**Hewlett Packard
Enterprise**

DMF 7 & Tier Zero: Scalable Architecture & Roadmap

Jan 2017



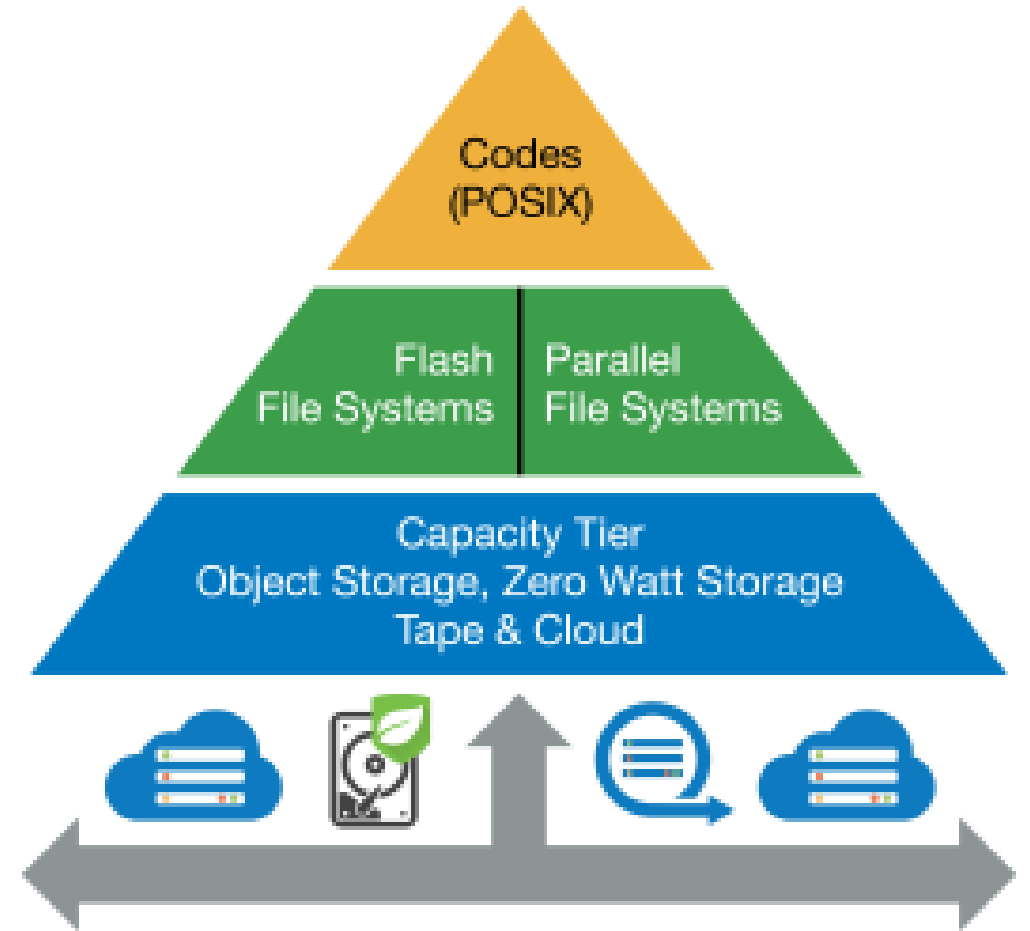
HPC (Compute/Storage) Roadmap

Forward-Looking Statements

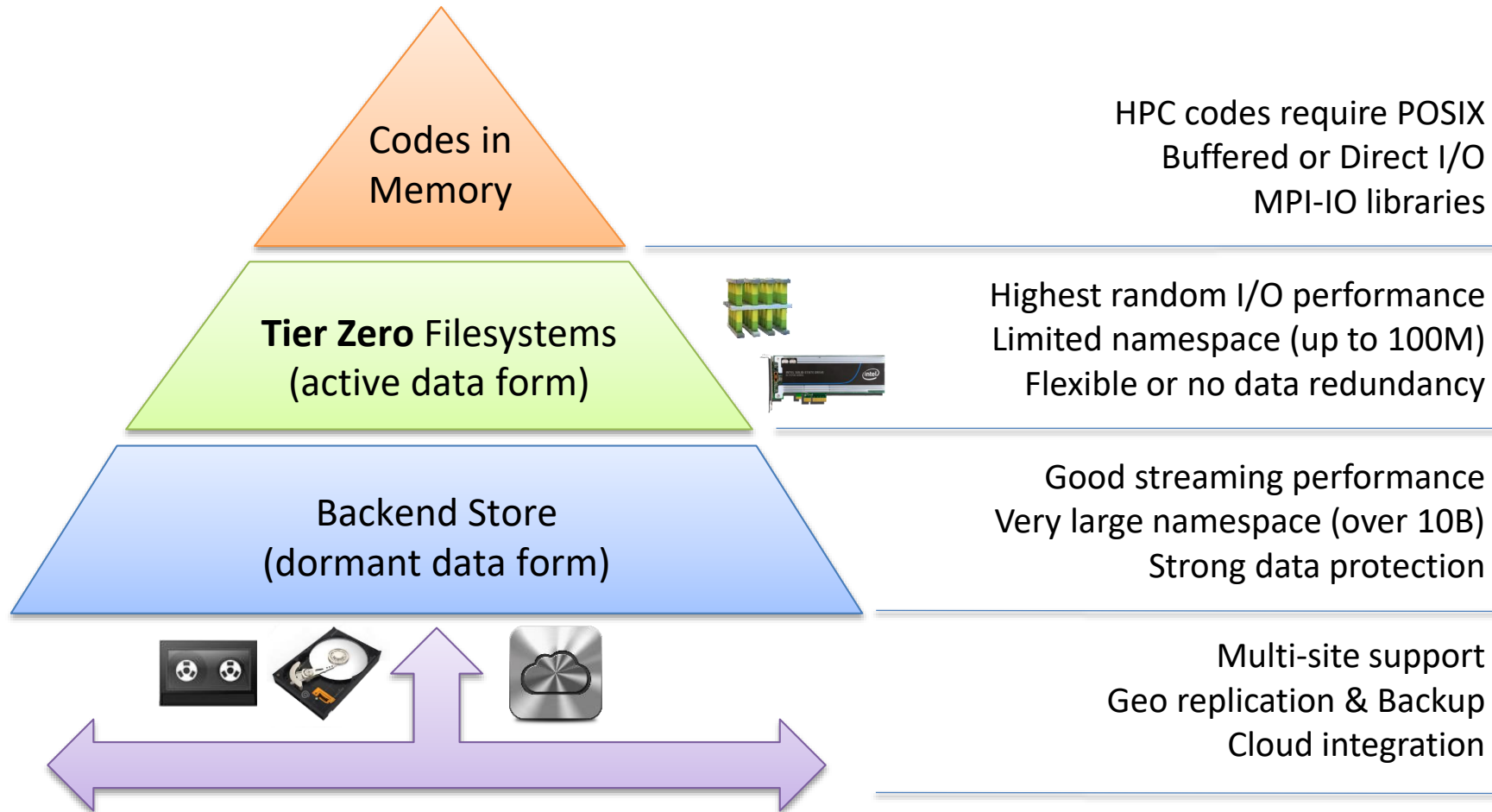
This document contains forward looking statements regarding future operations, product development, product capabilities and availability dates. This information is subject to substantial uncertainties and is subject to change at any time without prior notification. Statements contained in this document 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.

HPC Workflows | Solving Challenges **Inspiration**

- It's not just about FLOPs anymore
- It's about efficiently sharing active HPC data sets
- Understanding job I/O requirements
 - Accelerates performance
 - Improves successful completion
 - Enables co-scheduling to optimize utilization
- Placing data in proper tiers is key
 - Have data available in high-performance tier when job is starting
 - Store results in high resiliency tier when ready
 - Remove data from high-performance tier when done



HPC Workflows | Data Tiering Model **Active & Dormant Forms**





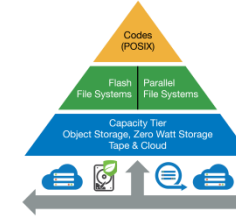
**Hewlett Packard
Enterprise**

DMF 7: Data Management Fabric

Scalable Distributed Architecture



Data Management Fabric | DMF 7 Platform Foundation



Lower Costs

DMF-based data management approach can yield up to **80% decrease in the cost of raw capacity** when compared to scaling out capacity using Tier 1 disks.



Flexible Infrastructure

DMF offers the most flexible storage infrastructure in the industry
RAID, Tape, Cloud & Object, ZWS



Limitless

DMF offers industry-leading virtually limitless storage & I/O capabilities
Total volume of data is only limited by physical media capacity.



Seamless Integration

DMF runs on Linux and integrates with the file systems already in use (XFS, CXFS and Lustre)
No application or workflow changes are required.

Data Management Fabric | DMF 7 Concepts & Vision

1 Integrate with High Performance Flash Filesystem

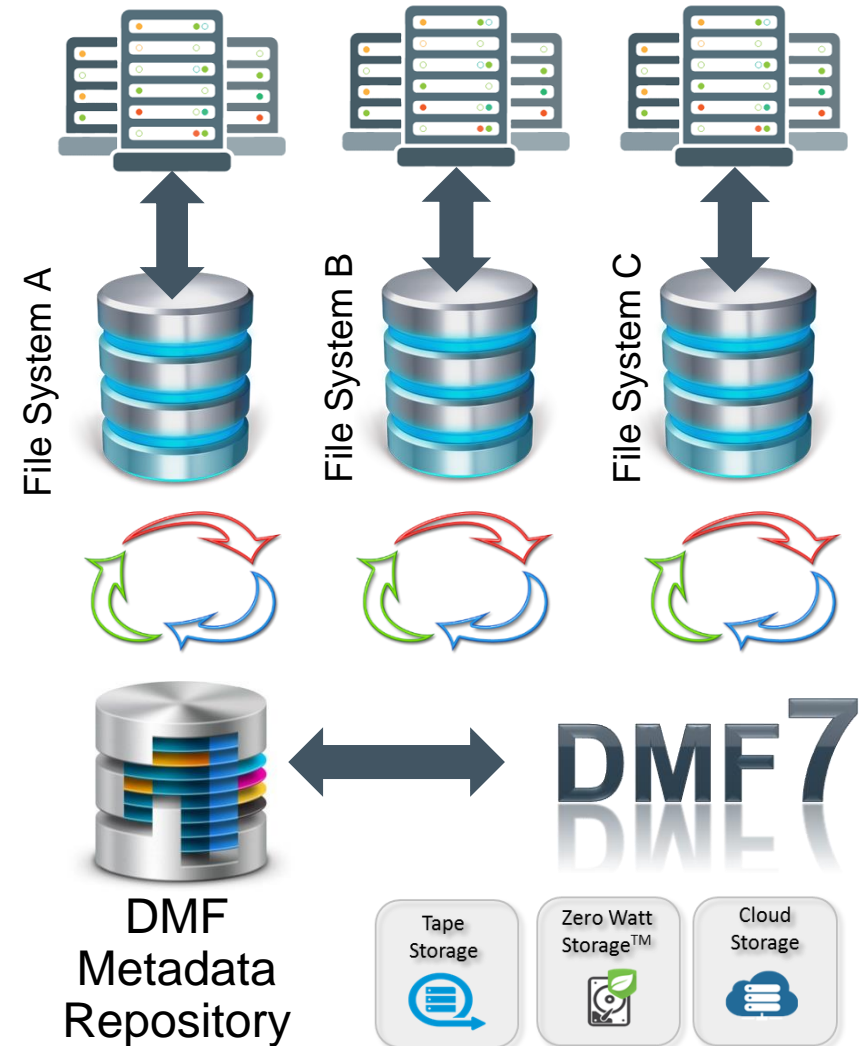
2 Offload Metadata Management to a Scalable Framework

3 Capture Filesystem Events in Realtime Changelog to avoid scans

4 Provision Namespaces On Demand by HPC Scheduler in addition to Static Filesystems

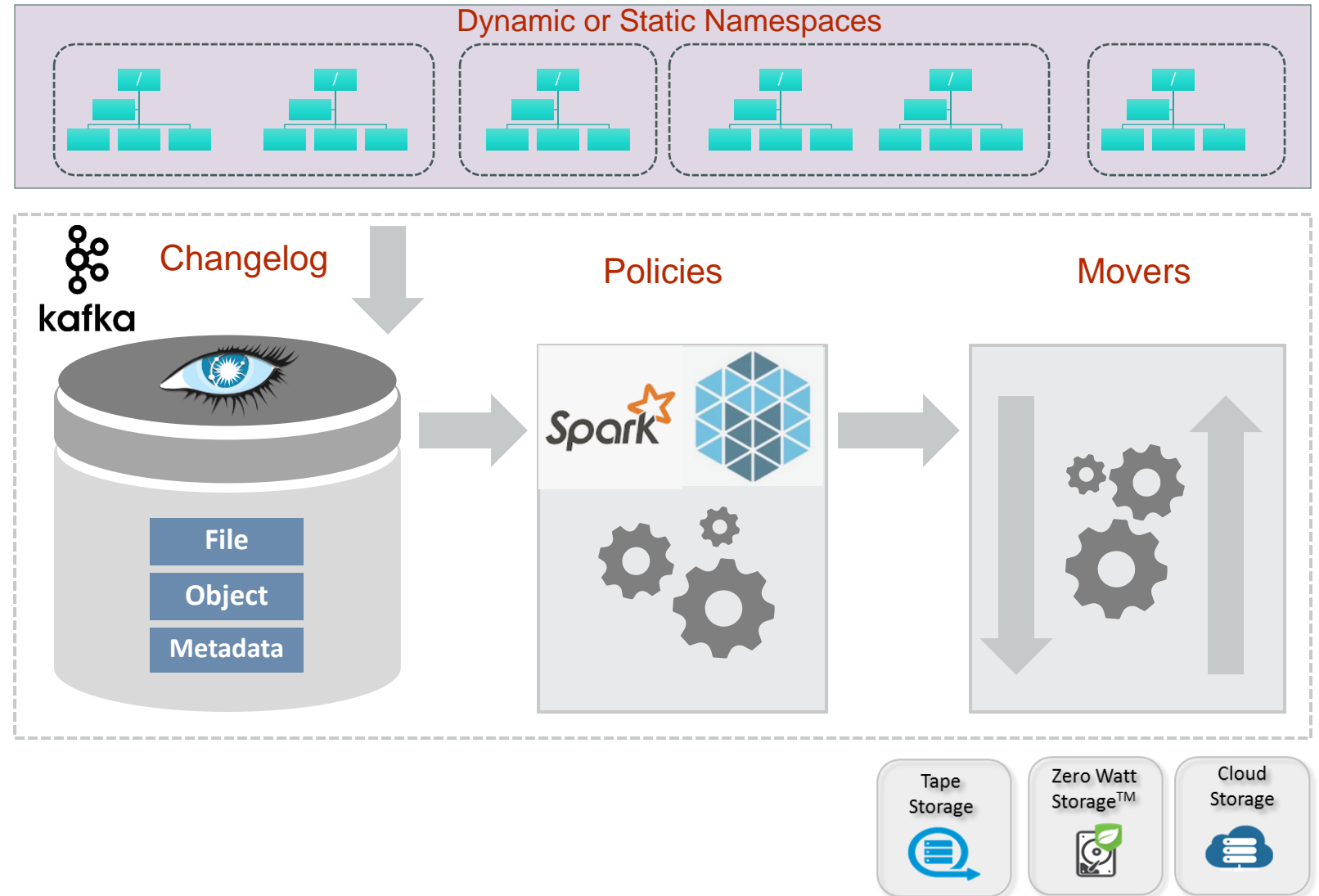
5 Optimize Data Transfer by Chunking and Parallel Data Movers

6 Keep Transparency. Replace Traditional Backup



Data Management Fabric | DMF 7 Architecture

- Kafka for Changelog processing
- Cassandra for Scalable Metadata
- Mesos for Task Scheduling
- Spark for Query Engine
- It's *SMACK!* Because it is Big Data.



Data Management Fabric | DMF 7 File System Integration

Key Takeaways

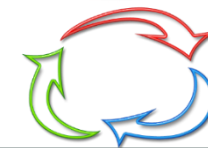
Native Integration Approach

DMF 7 includes tightly-coupled integration with supported file system environments.

- **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 the XFS/CXFS File System:** DMF 7 will provide seamless support of CXFS – and updated versions of the CXFS file system will include additional features supporting native integration that will remove the requirement 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.



- [C]XFS
- Lustre
- GPFS



DMF7

Direct file system integration and log processing with scalable metadata management

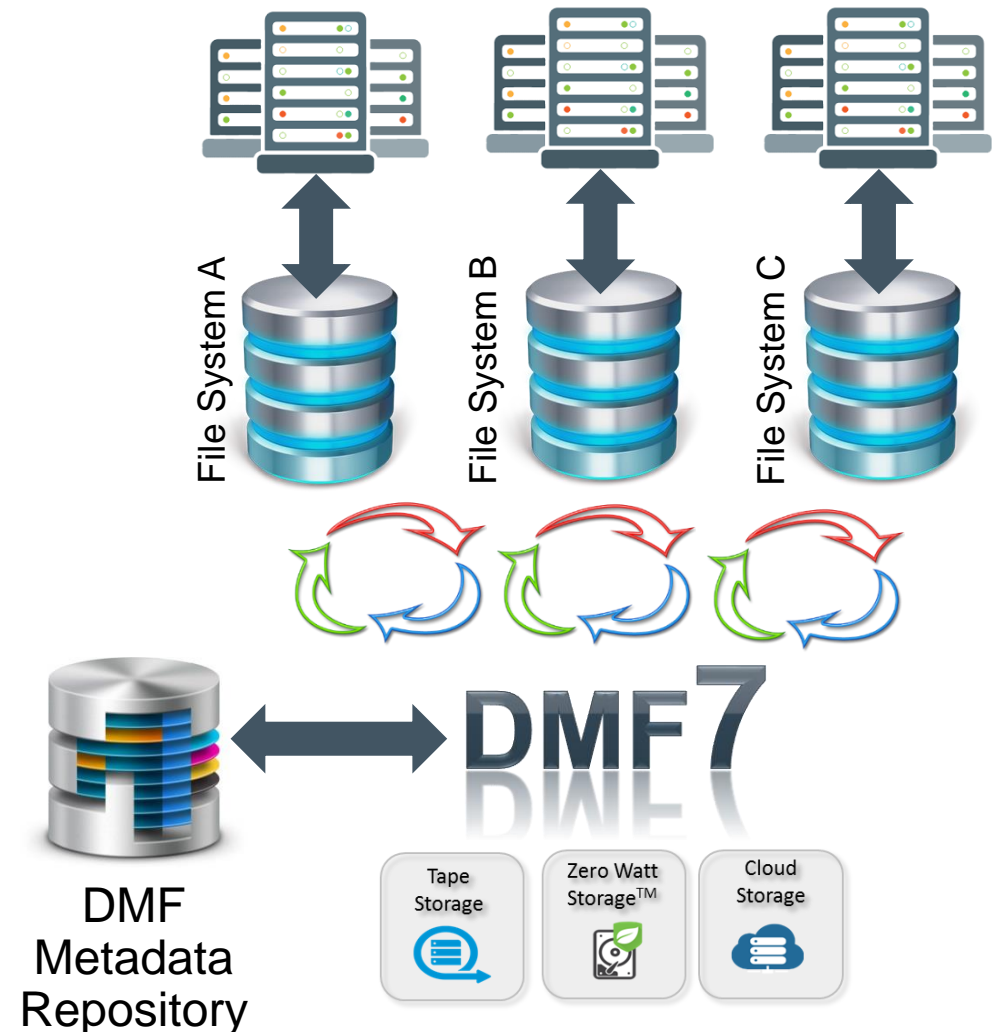
Data Management Fabric | DMF 7 Authoritative Metadata Repository

Key Takeaways

Authoritative Metadata Repository

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

- **Reduced Reliance on inodes:** With metadata and file system history stored in the DMF metadata database, there is reduced dependence on individual file system inodes. In the event of file system corruption or accidental data deletion, both the file system structure and contents can be recovered using DMF and its metadata information.
- **Extensible Metadata Information:** The DMF metadata repository is extensible to enable the storage 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.



Data Management Fabric | DMF 7 Job Scheduler Integration

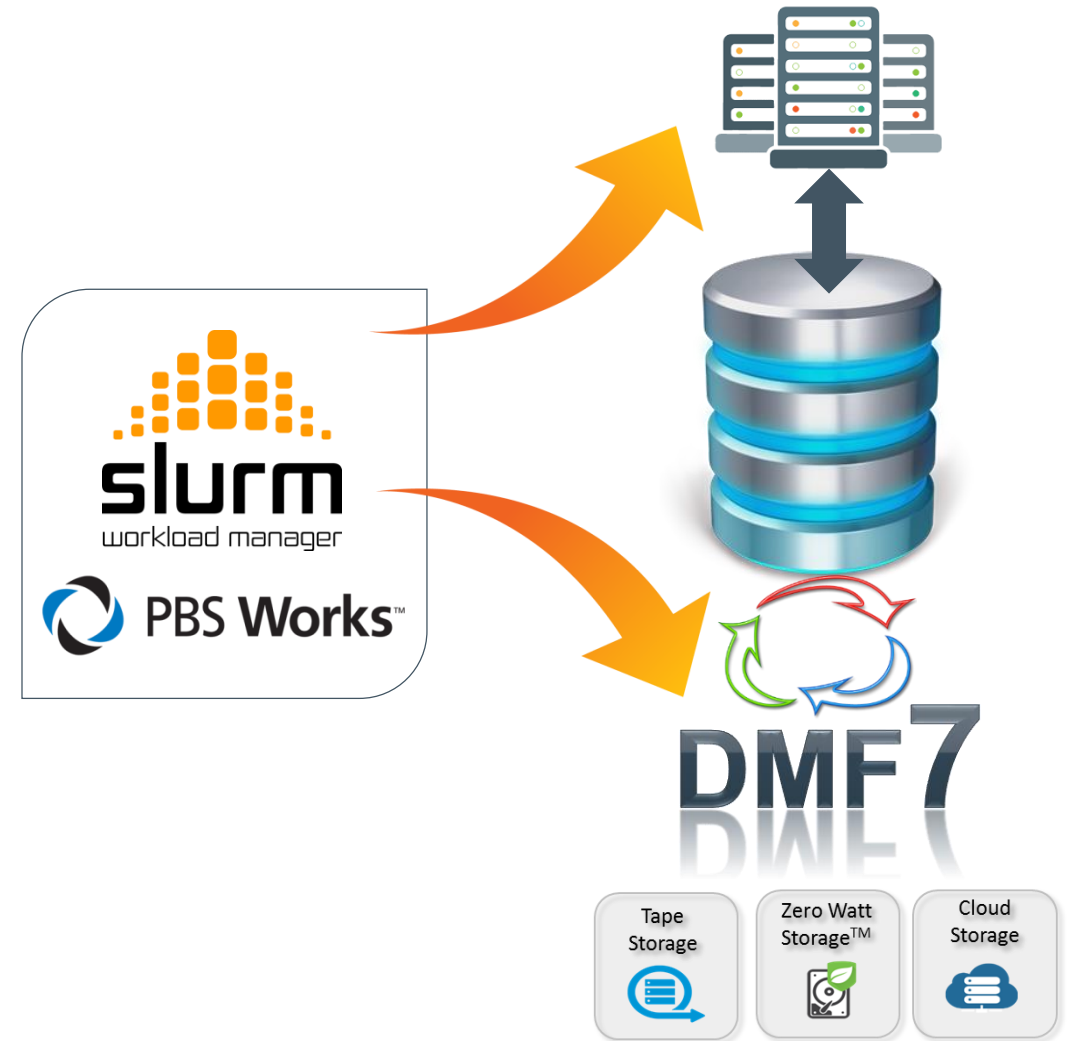
HPC use case

Key Takeaways

Job Scheduler Integration

DMF 7 is API-enabled to allow integration 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.



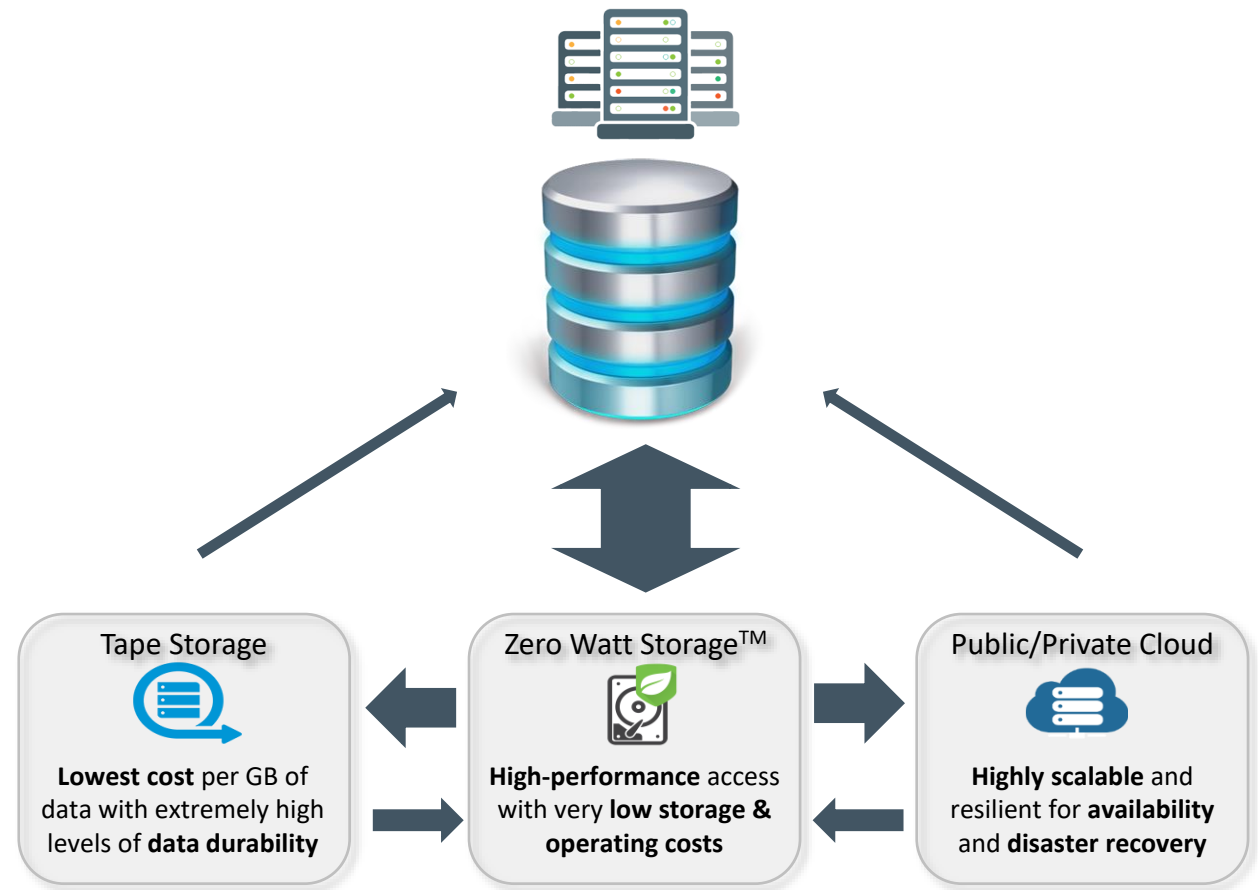
Data Management Fabric | DMF 7 Leveraging Zero Watt Backend

Key Takeaways

Zero Watt Storage as Data Cache

DMF 7 can maintain and track multiple copies of backend objects. It also supports creating and deleting the copies based on policy or request. This functionality replaces and augments DiskMSP and FMC use cases

- **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



Data Management Fabric | DMF 7 Continuous Backup & Versioning

Key Takeaways

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.

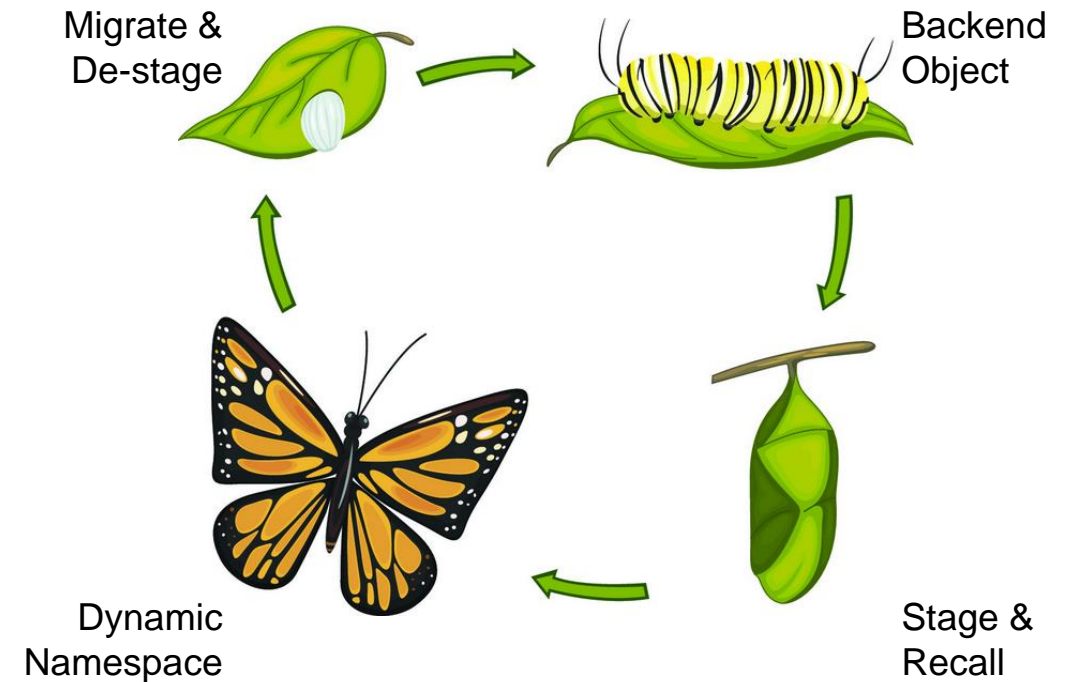
- **High-Quality Backup:** DMF will automatically make backup copies of files and metadata on a rolling basis using policy definitions that can be tailored by administrators
- **File Versioning:** DMF maintains both metadata information and file data from prior versions of files so that administrators 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.

Automated tiering and replication of new and modified files supports easy file restoration and “point in time” data access based on versions

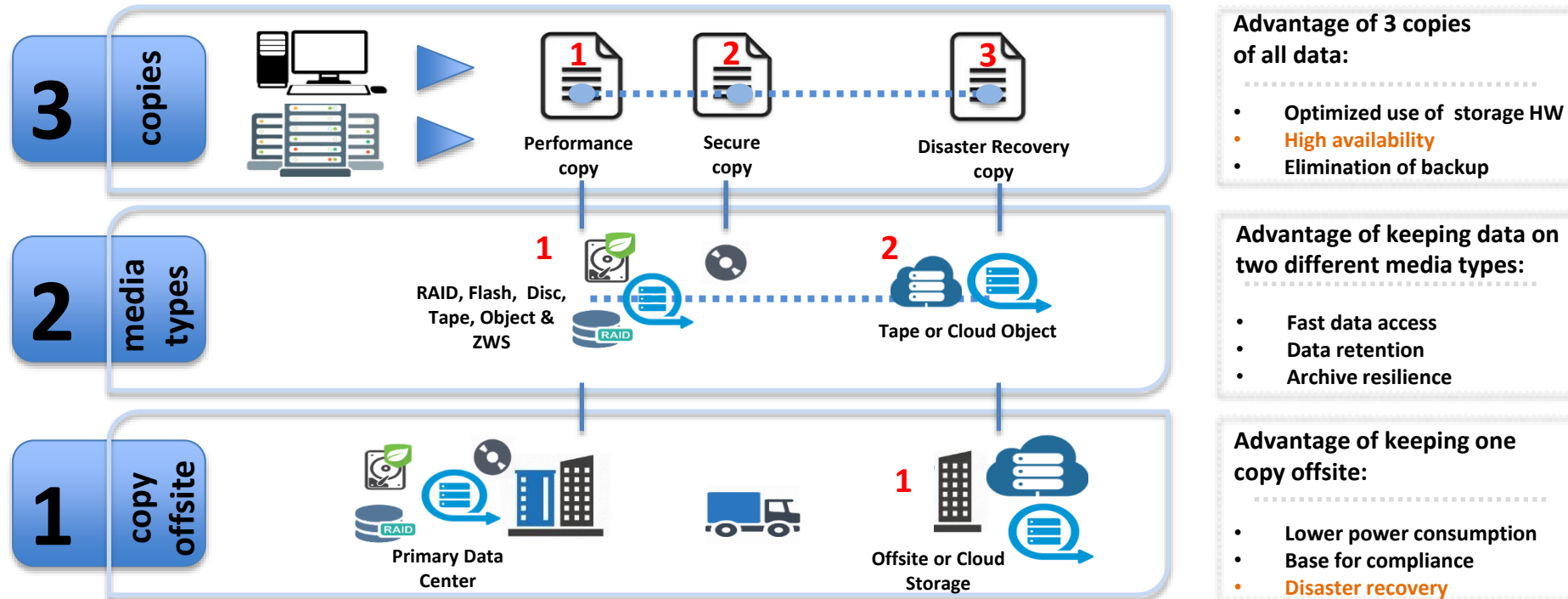


Data Management Fabric | DMF 7 Data Metamorphosis

- Store bulk of data in backend storage, on premises or in cloud
- Capture file metadata and rich metadata in a distributed database
 - Handling HPC metadata is a Big Data problem!
- Automatically stage, recall, migrate and/or de-stage objects to high-performance namespaces just-in-time, based on workflow
 - Avoid machine time charges to wait for I/O!
- Data in backend storage can be directly accessible by as read-only data source via plugins
 - S3 objects can be made available for non-HPC Big Data clusters
 - Backend storage is easy to replicate across distance. Metadata will replicate with the distributed database



Data Management Fabric | DMF Data Protection Strategy





**Hewlett Packard
Enterprise**

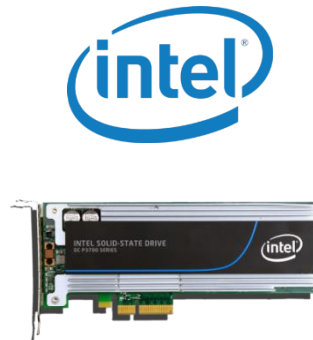
Tier Zero Namespaces

Towards 500 clients and beyond



Data Management Fabric | UV300 Accelerated All-Flash File System

- Single scale-up SMP system delivers
 - Compute:
 - Up to 480 Intel Xeon™ cores (32 sockets)
 - Memory:
 - Up to 24TB DRAM
 - Non-volatile Memory:
 - Up to 128TB FLASH
 - Up to 30 Million IOPs
 - Up to 200GB/s
- All in one rack
- No switches, no arrays, no problems

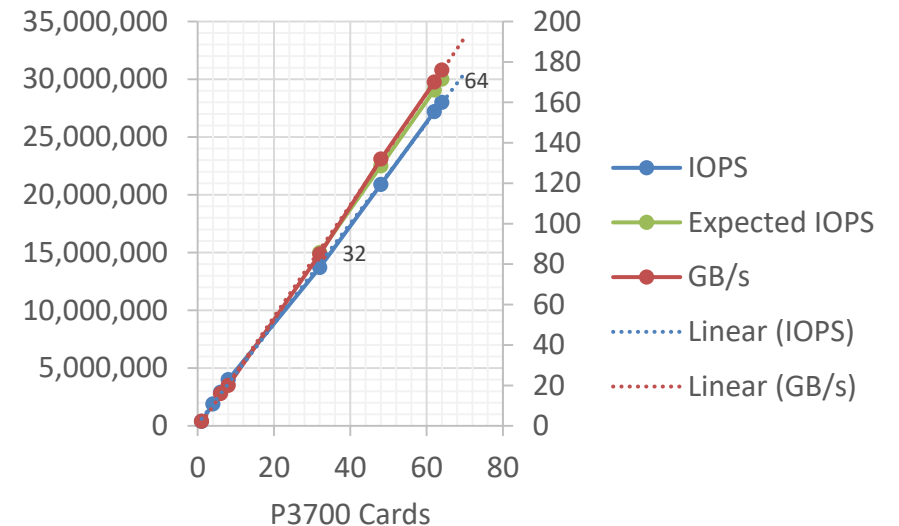


Intel P3700
400K IOPS
2.4GB/s



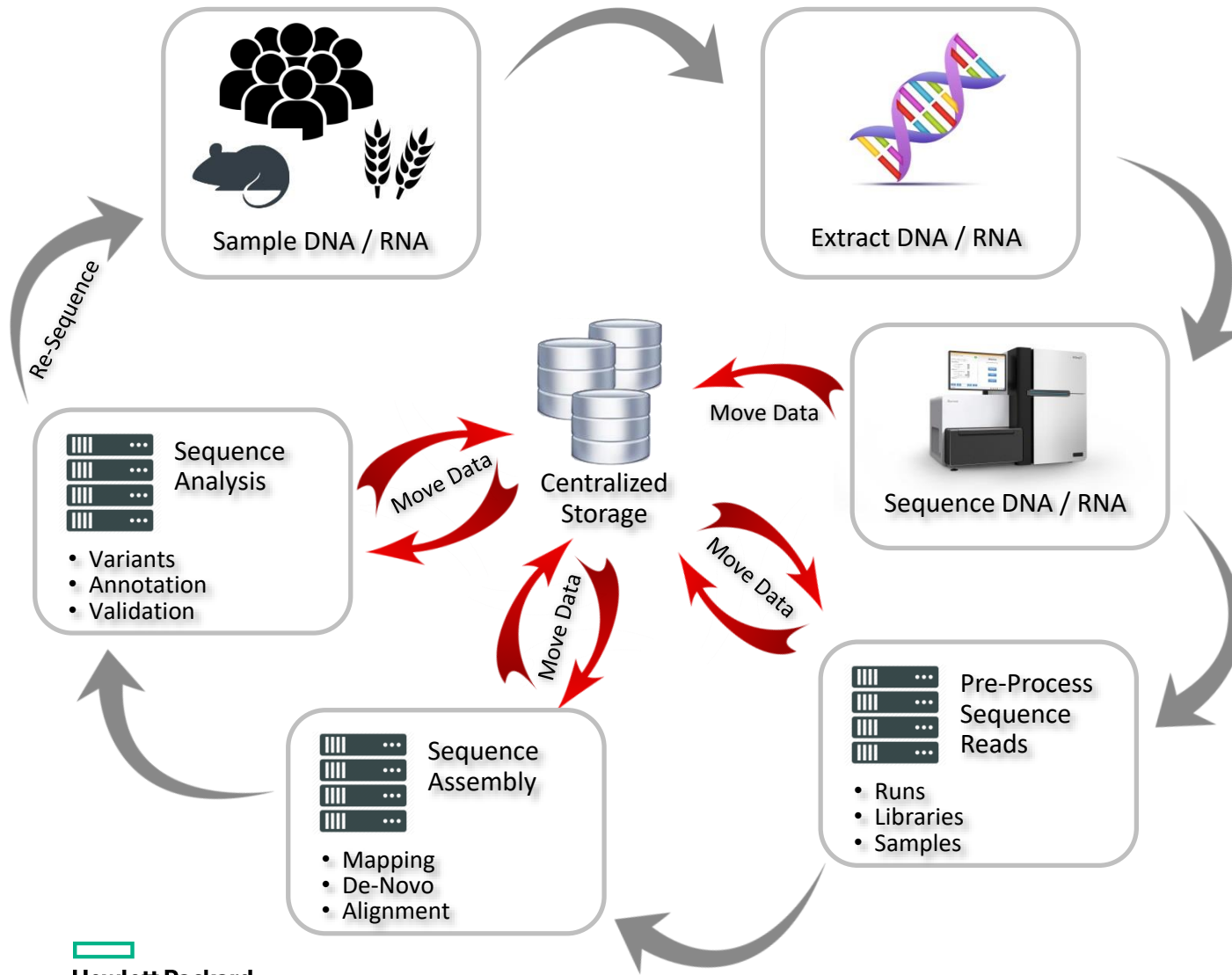
32 Socket SGI UV 300 System
with 64 Intel NVMe Flash Cards

- Up to 30M IOPs
- Up to 200 GB/s



* Plan of Intent

Data Management Fabric | Compute & Storage for Genomics



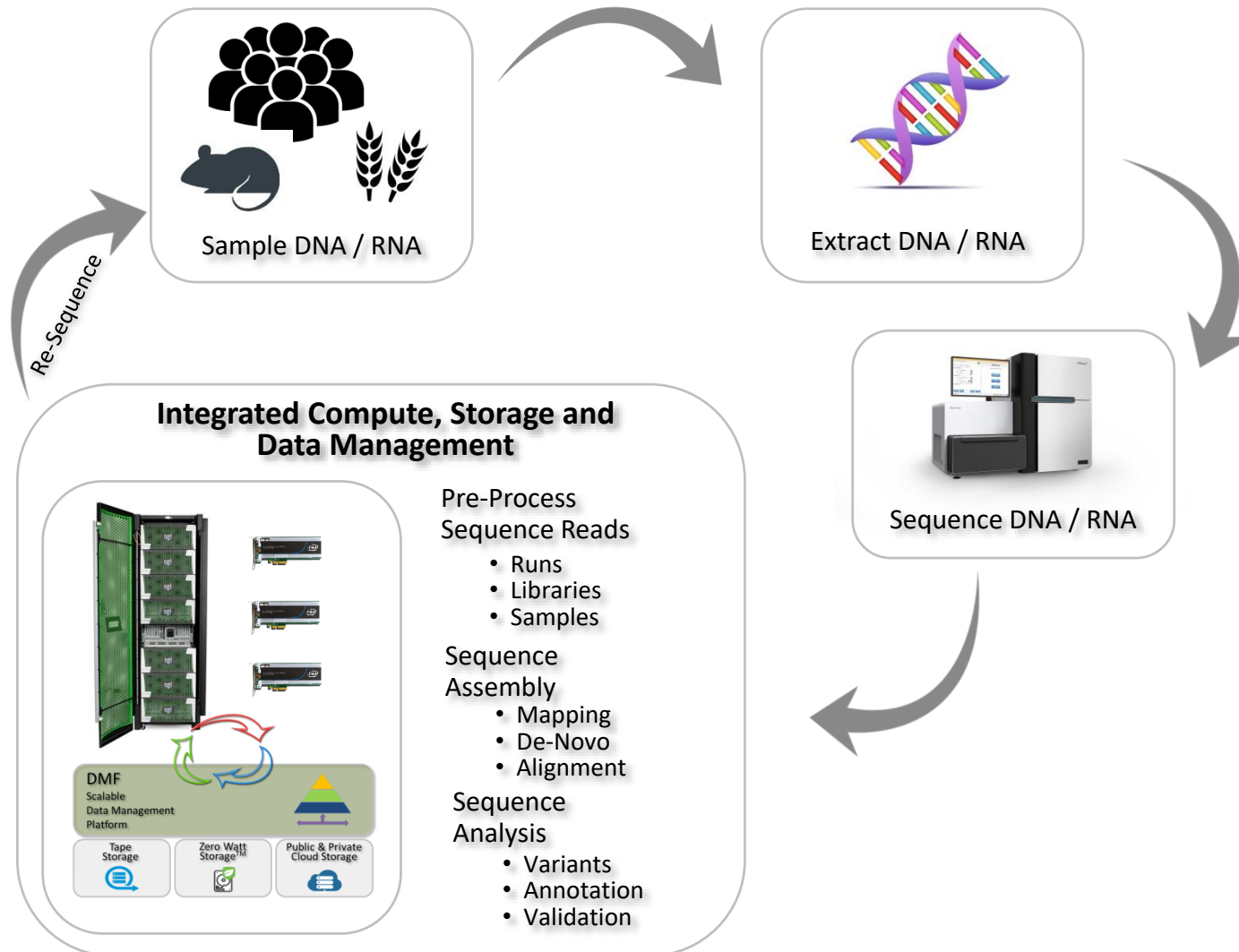
Traditional Model

Overall Approach

Centralized data store model with distributed data processing.

- **Large Scale Data Movement:** Large data sets of up to 500GB per genome can drive significant load on data fabrics and storage systems. Genomics data is often comprised of large numbers of small files related to sequence data which can put pressure on file system metadata management.
- **Workflow Performance Impacts:** Repeated data movement operations related of multi-step workflows can impact the overall speed at which work is completed.

Data Management Fabric | Converged Solution for Genomics



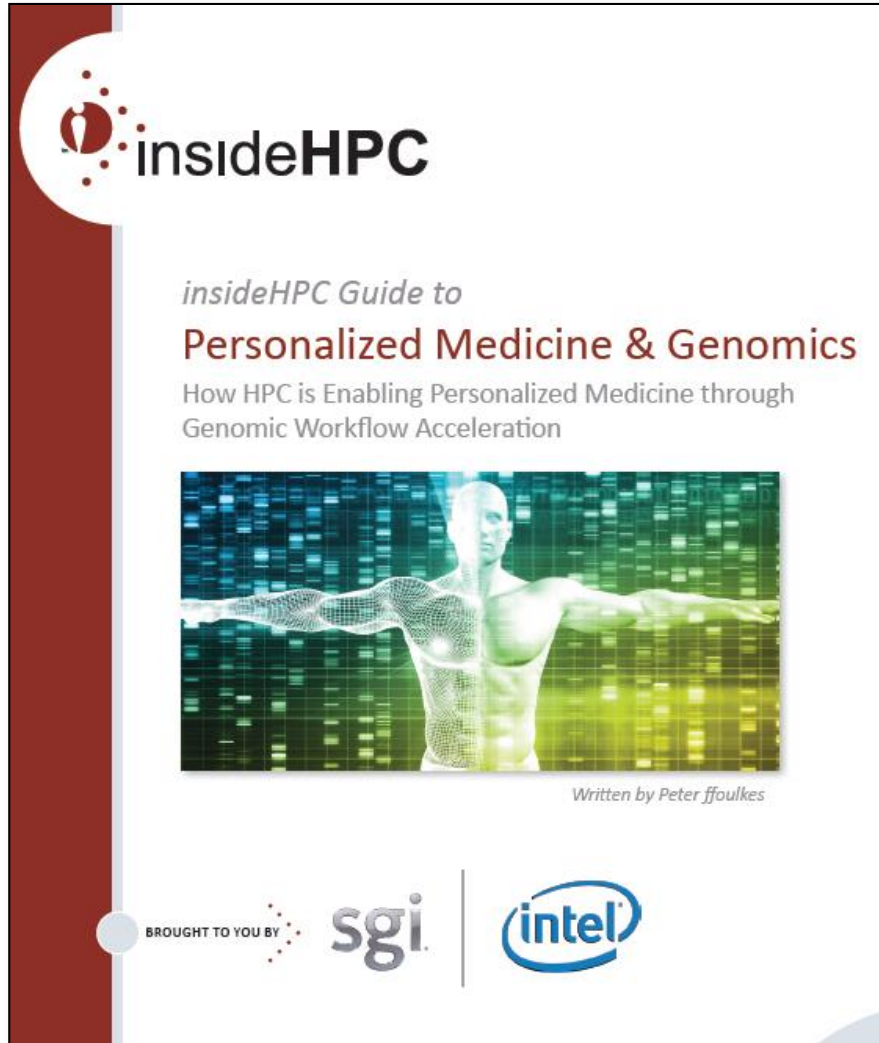
Accelerated Model

Overall Approach

Converged compute-and-storage model leveraging and all-flash performance tier backed by advanced large-scale data tiering and protection.

- **In-Place Data processing:** Data sets land on a high-performance flash-based storage tier – and multi-phase processing occurs in-place without the need for data movement to a remotely connected storage system. As processing completes, data is transparently moved to a cost-optimized storage and data protection layer to free up flash storage for new sequence operations.
- **Workflow Performance Acceleration:** Repeated data movement operations related of multi-step workflows can impact the overall speed at which work is completed.

Data Management Fabric | Converged Solution for Genomics



insideHPC

insideHPC Guide to
Personalized Medicine & Genomics
How HPC is Enabling Personalized Medicine through
Genomic Workflow Acceleration

Written by Peter ffoulkes

BROUGHT TO YOU BY **sgi** | **intel**

Earlham Institute

formerly The Genome Analysis Centre (TGAC)



Earlham Institute's projects span the breadth of life sciences as well as technology development, data infrastructure, and knowledge exchange with a focus on genomics and bioinformatics which is applied to plants, animals, and microbes. By partnering closely with HPC technology leaders such as SGI, Earlham Institute boasts world-class compute and storage infrastructure that allows scientists to undertake some of the most challenging data-intensive research in the fields of genomics and biosciences.

abbvie

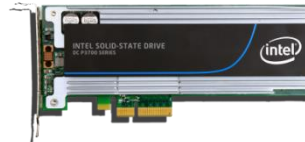
formerly Stemcentrx

abbvie | MEDICAL INFORMATION

Stemcentrx is a San Francisco-based startup that was founded in 2008 and was recently acquired by AbbVie. They are pioneering new approaches to treating cancer. By processing an individual's DNA sequences, the company is investigating many of the largest and most lethal cancers to identify cancer stem cells and deliver drug treatments that eliminate the cells that are responsible for tumors.

Data Management Fabric | **Standard Server** All-Flash File System

- 2-Socket Servers Running SGI R-Pool with 8-16 NVMe devices
 - Non-volatile Memory:
 - 16-32TB Flash per Server
 - Up to 6.5 Million IOPs
 - Up to 38 GB/s
- Scalable using SGI all-flash file system for dynamic file system management



Intel P3700
400K IOPS
2.4GB/s

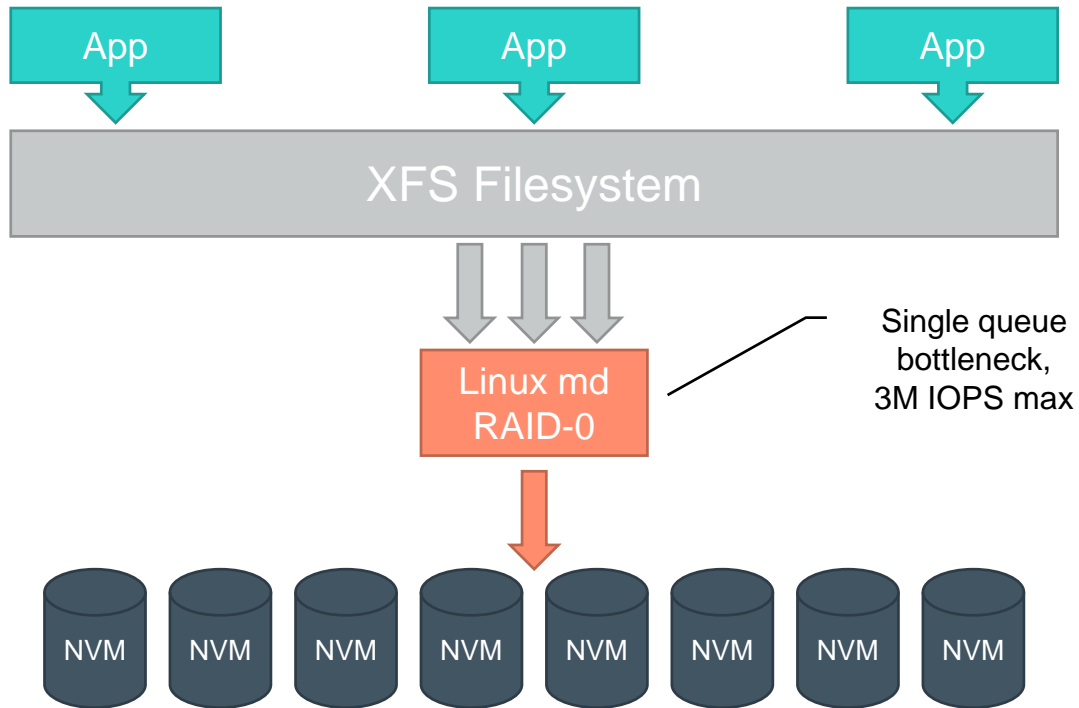


* Plan of Intent

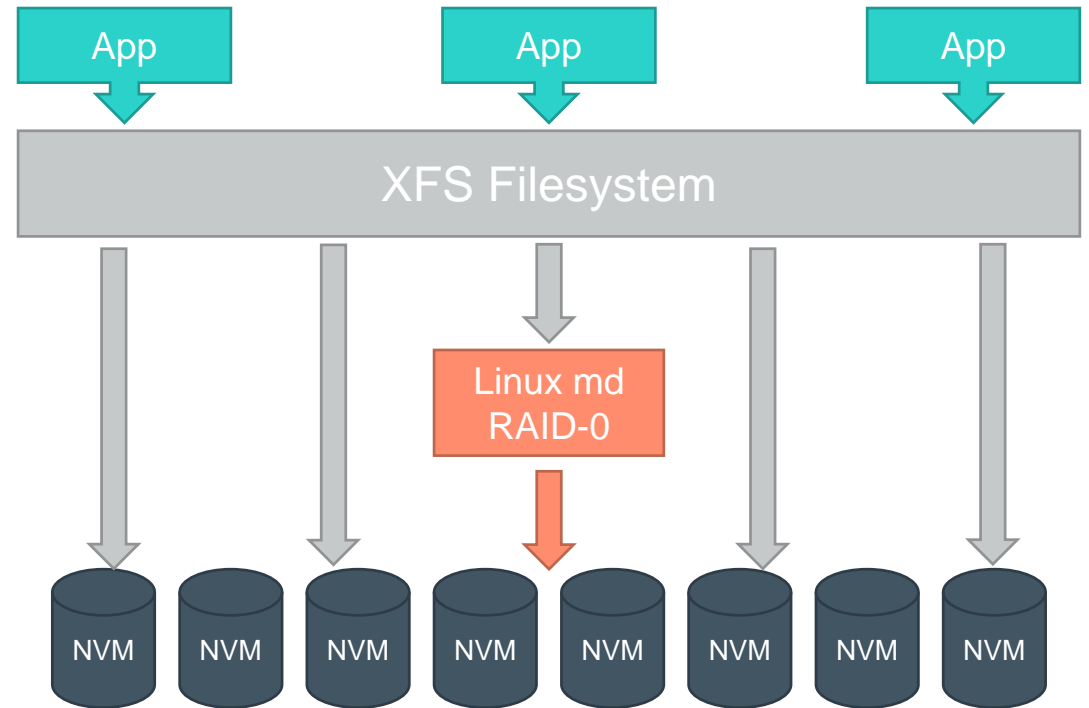
Data Management Fabric | Tier Zero Performance

As shown at ISC15

XFS + MDRAID: 3M read IOPS max



SGI eXFS: 13M read IOPS (32 cards)



Direct parallel access to NVMe queues

Data Management Fabric | Tier Zero R-Pool Architecture

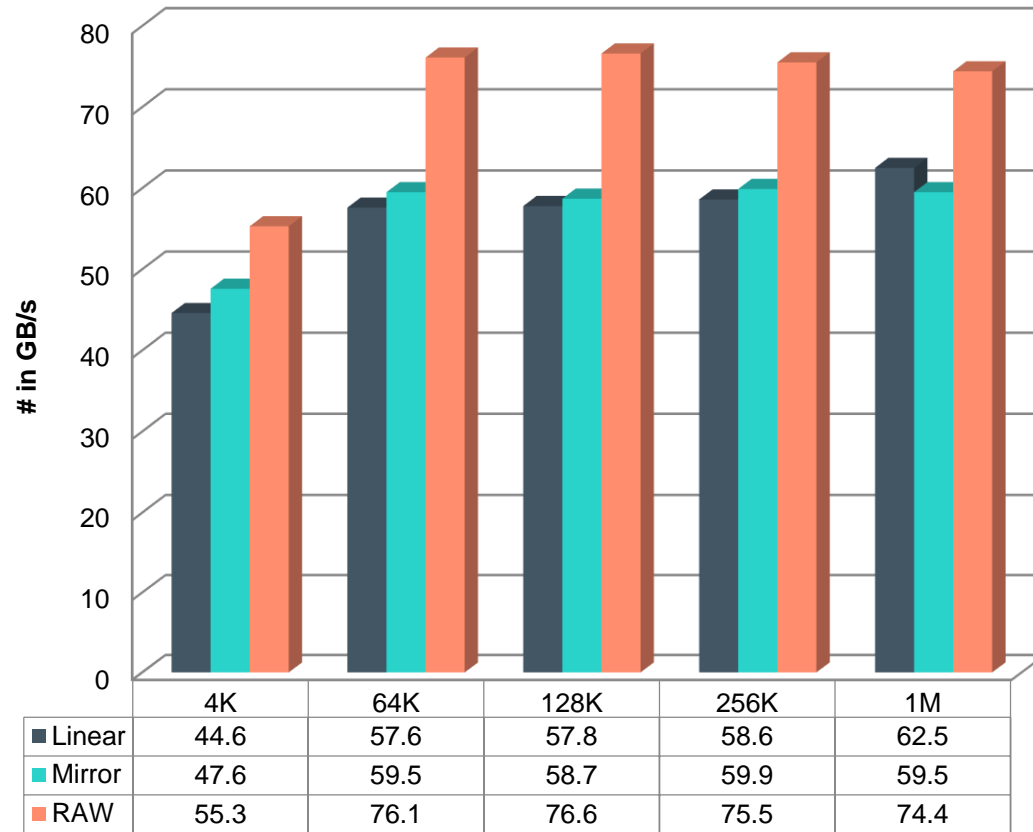
NUMA aware Hardware Abstraction & Storage Virtualization

- Storage devices
 - NVMe-capable Solid State Devices
 - Includes external PCIe arrays
- Extent
 - Storage device capacity is broken into fixed size extents, typically 1GB
 - Extents are broken into fixed size chunks
- Chunk
 - Segment of an extent that can be defined as data or parity
 - Selectable size from 1M to 8M based on expected random overwrites
- Sheet
 - Collection of extents with chunks of same size organized in stripes
- Virtual Volumes
 - Volumes are broken into fixed size stripes
 - Stripes are mapped when they are first written
 - Only mapped stripes point to physical storage
 - Volume stripes can be unmapped via commands from OS
 - Used for filesystem (XFS, CXFS, ext3)
- Writes are direct-in-place and do not require remap or copy-on-write
- Flexible Protection
 - All actively written stripes are mirrored
 - Parity is calculated for inactive stripes and mirror stripe is released

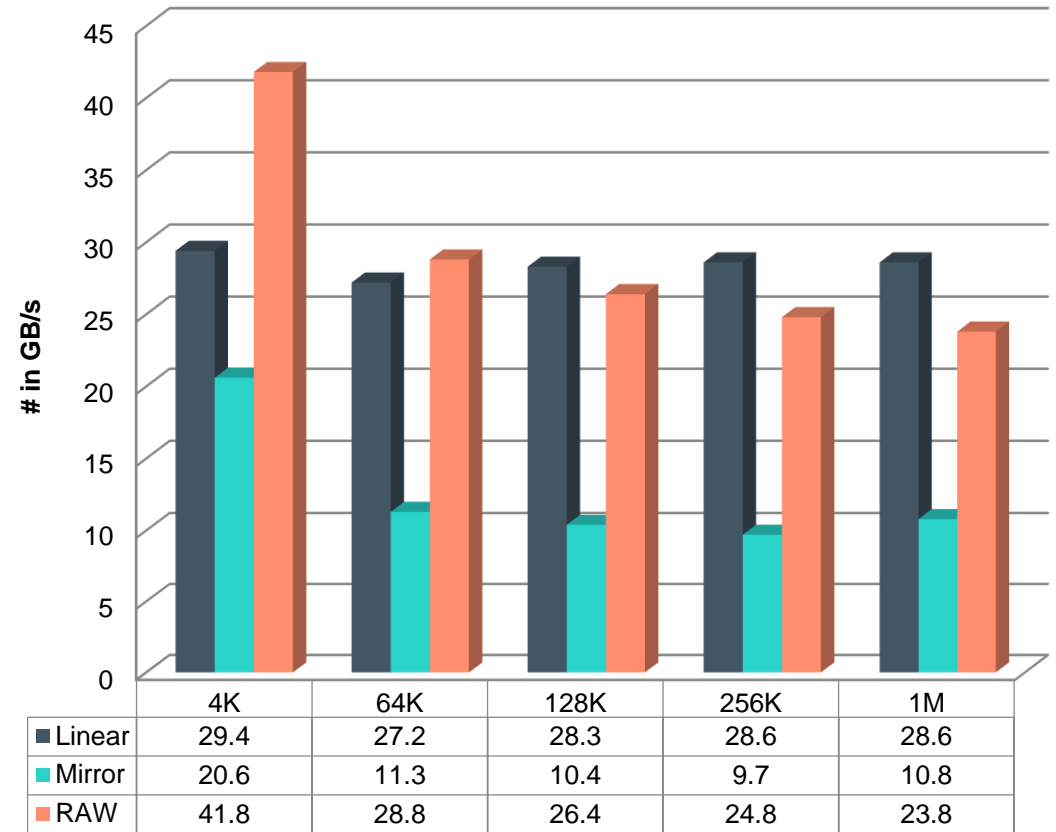
R-Pool Layout (8D:128K) | **FIO - 32 IODEPTH - 64 JOBS**

26 Intel DC P3700 800GB NVMe Cards, 16 socket MC990 X

RANDOM READ - BANDWIDTH



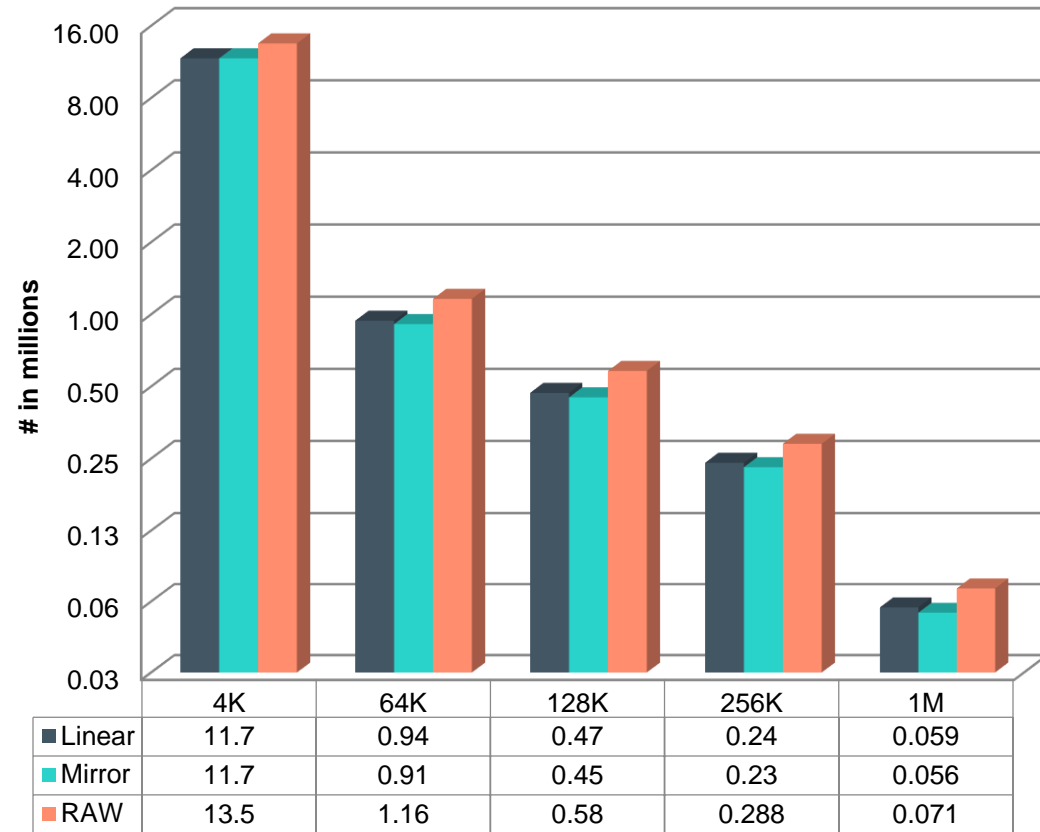
RANDOM WRITE - BANDWIDTH



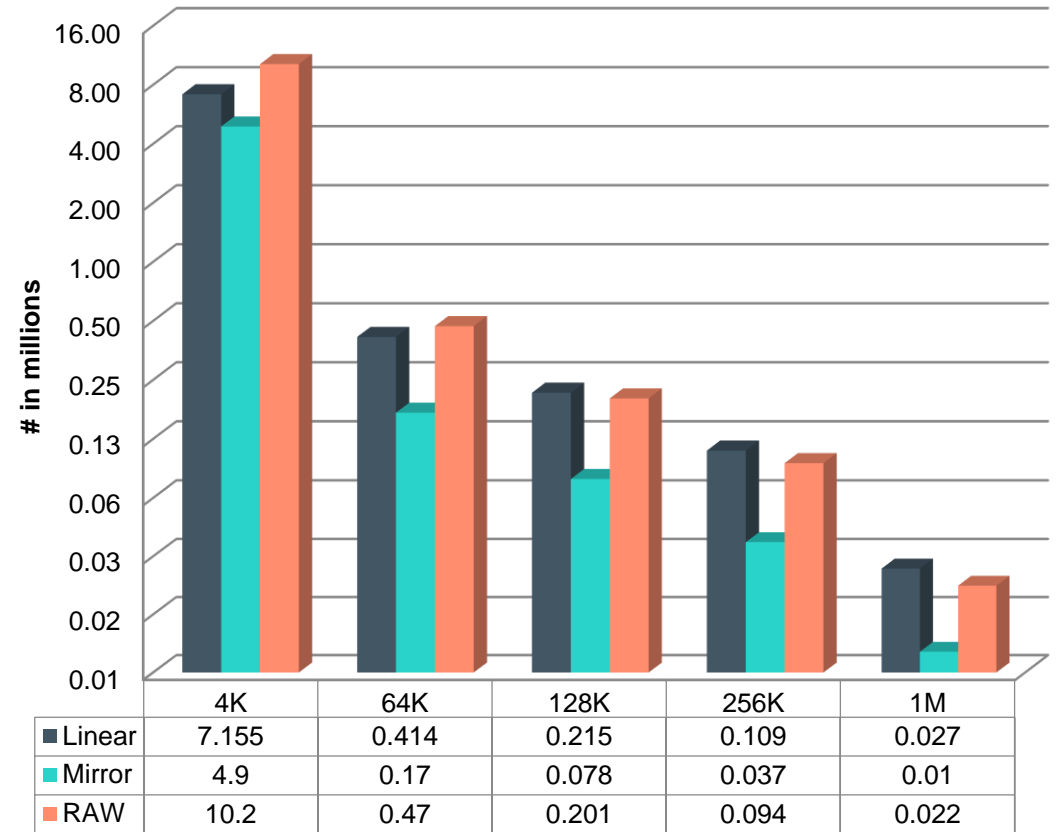
R-Pool Layout (8D:128K) | **FIO - 32 IODEPTH - 64 JOBS**

26 Intel DC P3700 800GB NVMe Cards, 16 socket MC990 X

RANDOM READ - IOPS

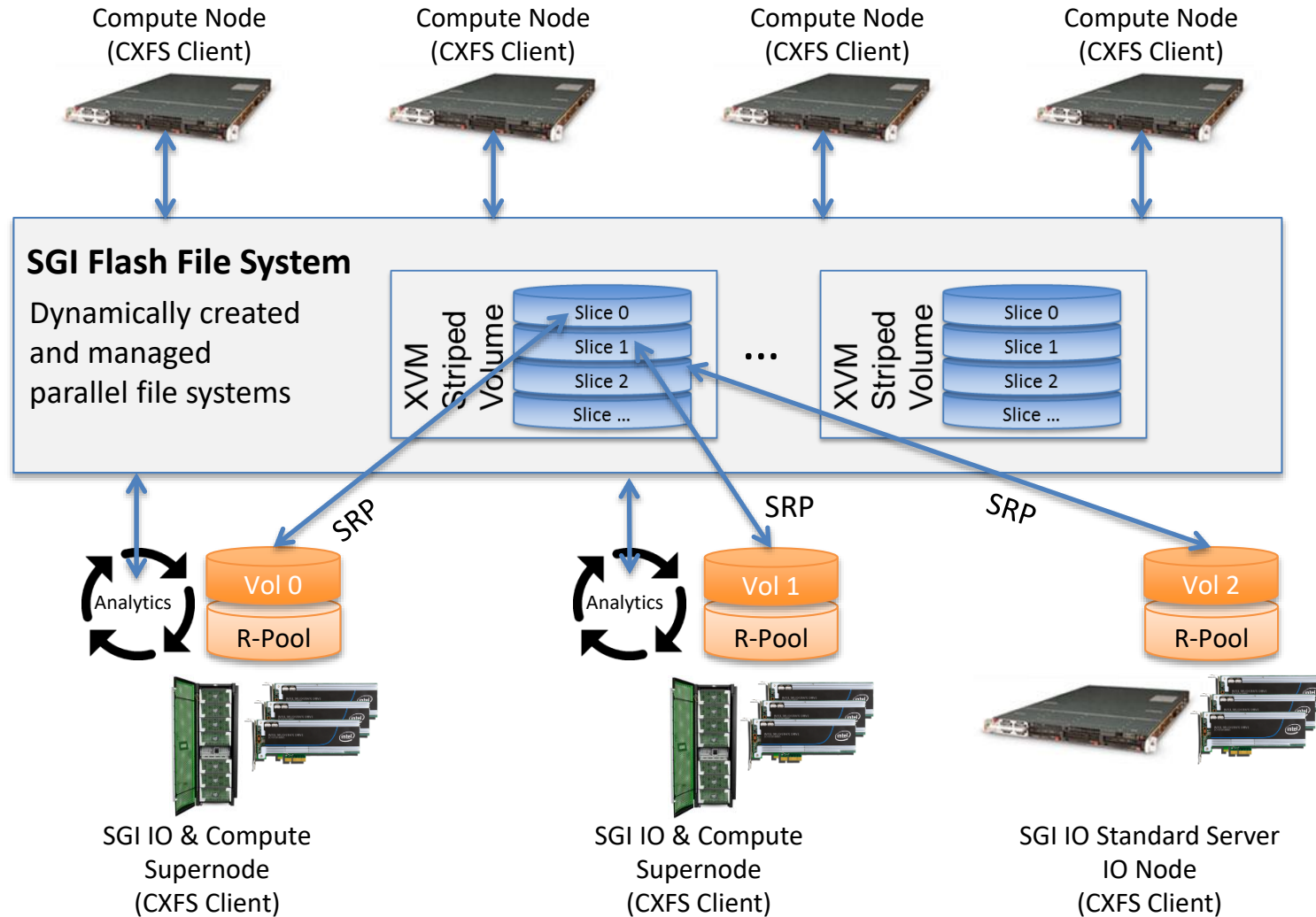


RANDOM WRITE - IOPS



Data Management Fabric | Accelerated Data All-Flash File System

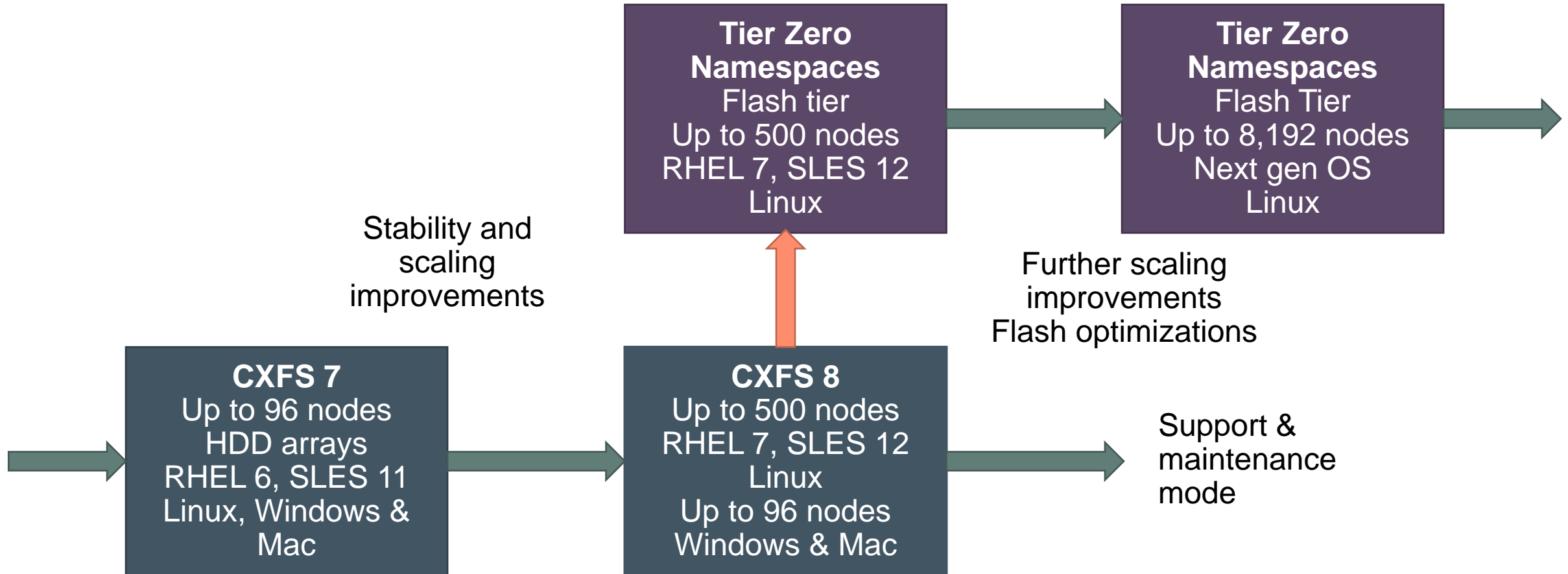
SGI UV-based IO & Compute Supernodes serve dual roles as high-performance block storage devices and scale-up compute nodes with full data access. Alternately, standard 2-socket servers can be used as I/O-centric systems.



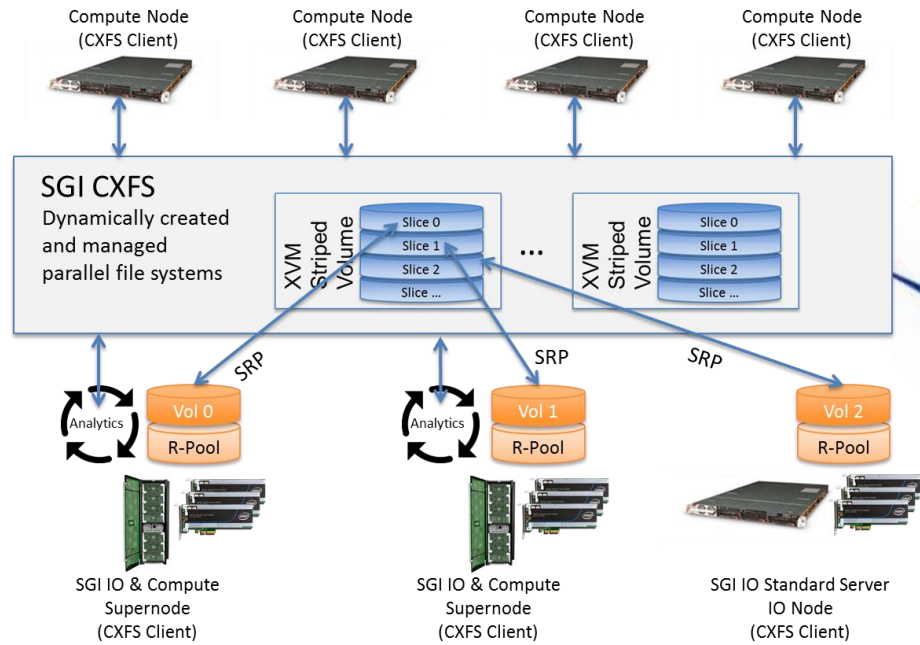
* Plan of Intent

Data Management Fabric | Accelerated Data Towards Flash Tier

CXFS & Tier Zero Namespaces



Data Management Fabric | Accelerated Data All-Flash File System



DMF performs dynamic file system creation, monitoring and data management based on user policies – or through job scheduler integration. Data is staged, recalled, migrated and de-staged for use by both cluster compute nodes and UV-based analytics nodes (where applicable).

Tape Libraries (Hardware) Low Cost & High Durability	Zero Watt Storage Low Cost & High Performance	Cloud / Object Storage High Scalability & Geo-Distribution

* Plan of Intent

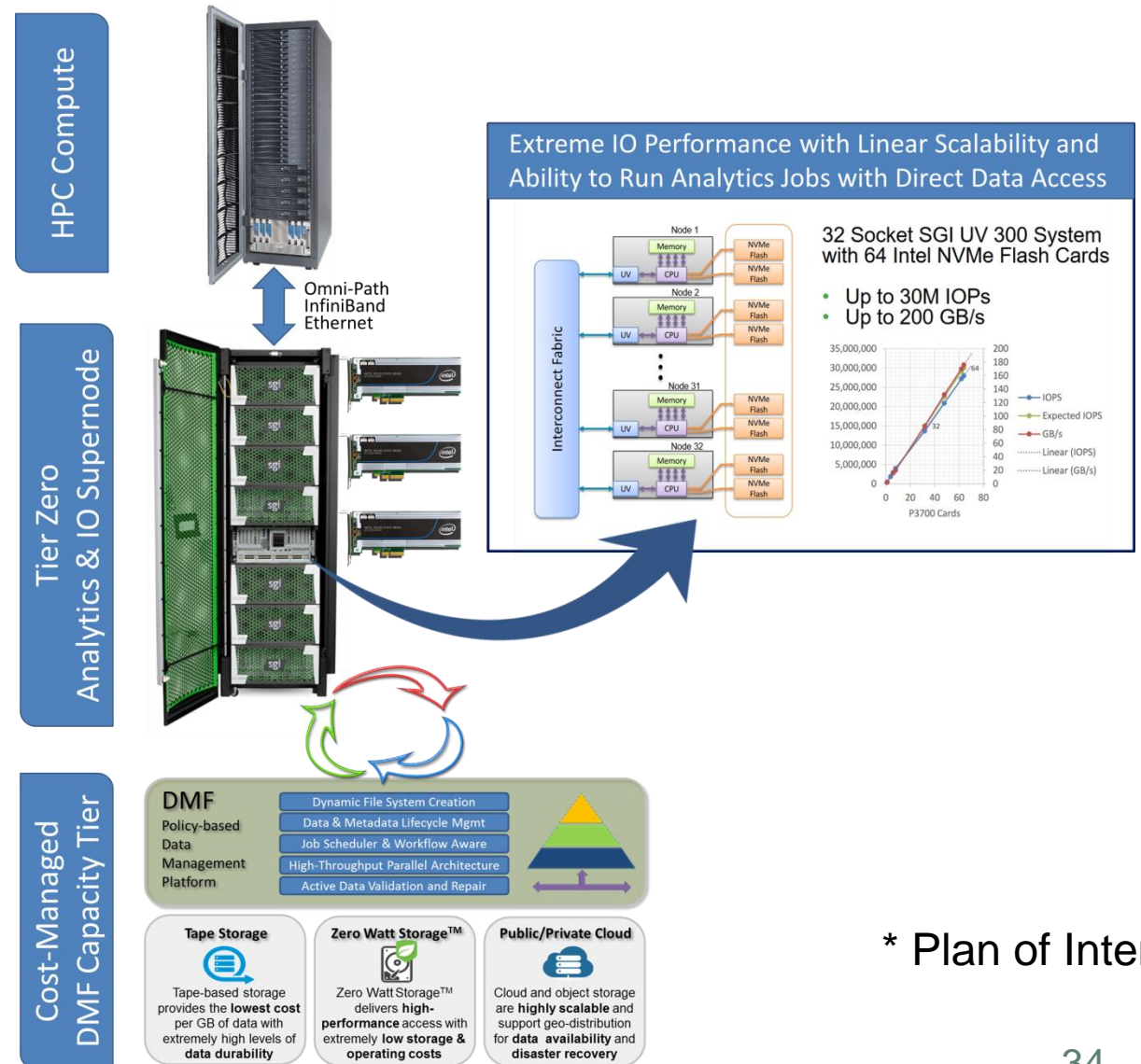
Data Management Fabric | All-Flash Storage Layer Tier Zero

Tier Zero Storage Key Concepts

Overall Approach

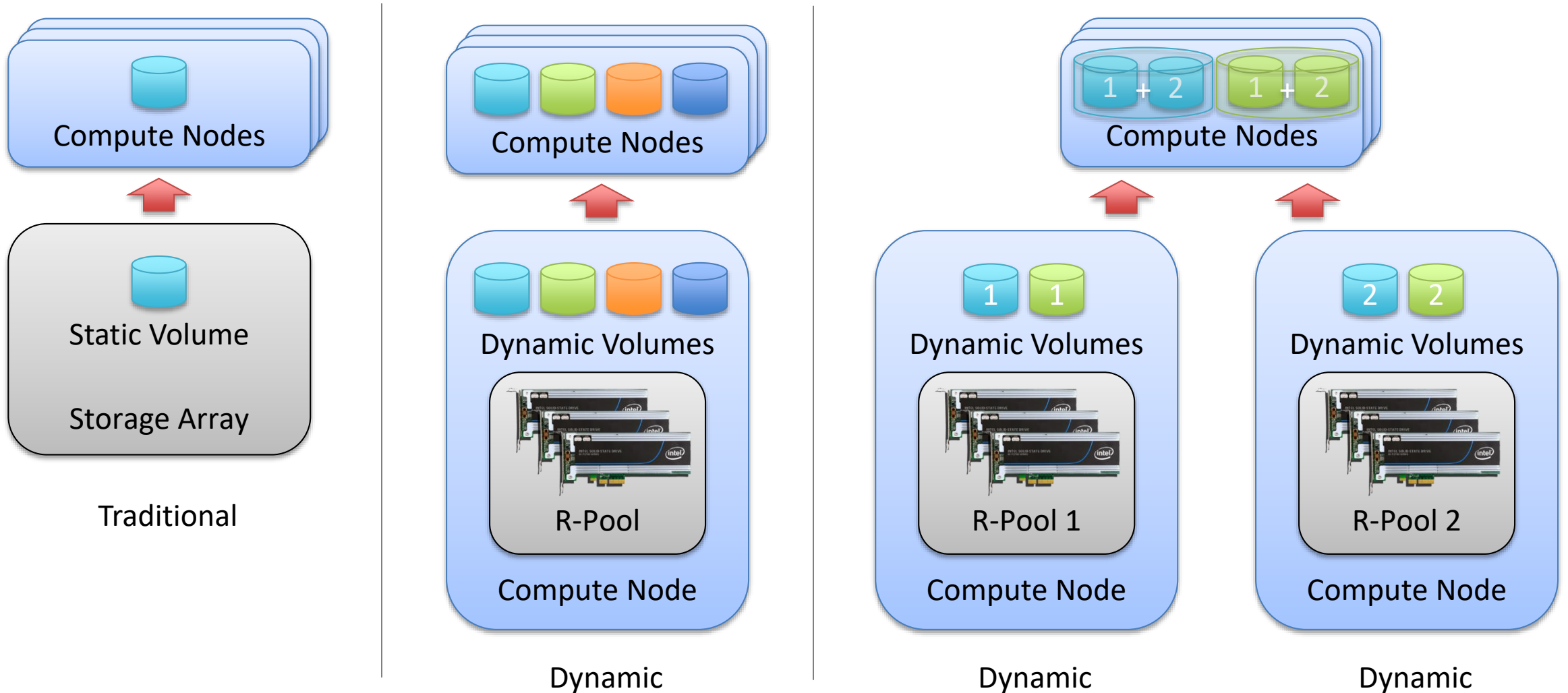
Flash-based Converged Compute & Storage Architecture for Workflow Acceleration

- **All-Flash Storage Tier Integrated Within Compute:** Supports both standard 2-socket storage nodes or SGI UV scale-up systems as IO supernodes for managing sets of high-performance flash storage
- **On-Demand Data Tiering under Job Scheduler Control:** All-flash tier is sized to manage active data associated with in-flight compute jobs. Integration with standard HPC job schedulers (SLURM, PBS Professional, etc) for data staging and de-staging orchestration to move data from high-performance tier to capacity tier managed by SGI DMF.

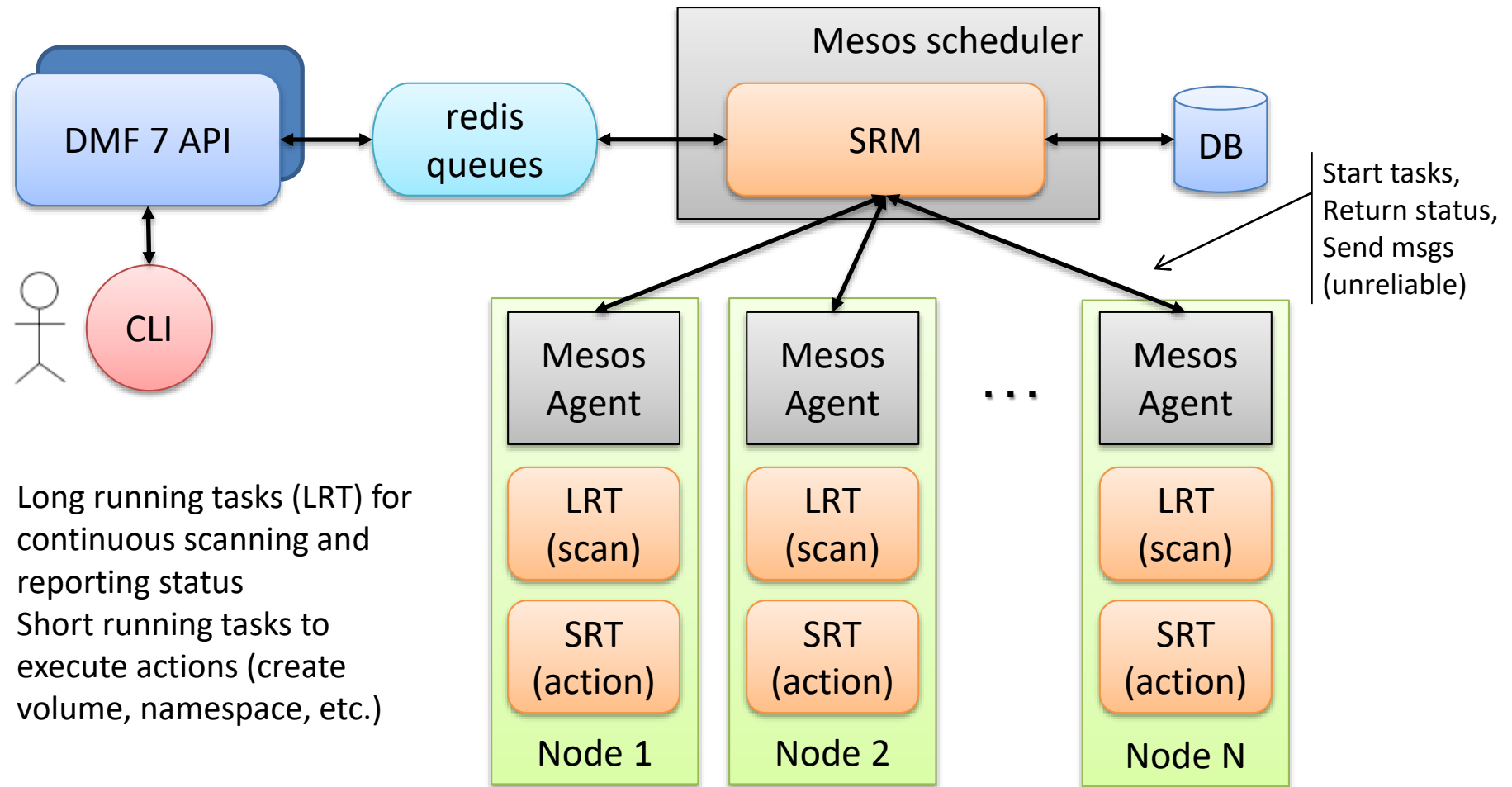


* Plan of Intent

Data Management Fabric | Shared Dynamic Namespaces

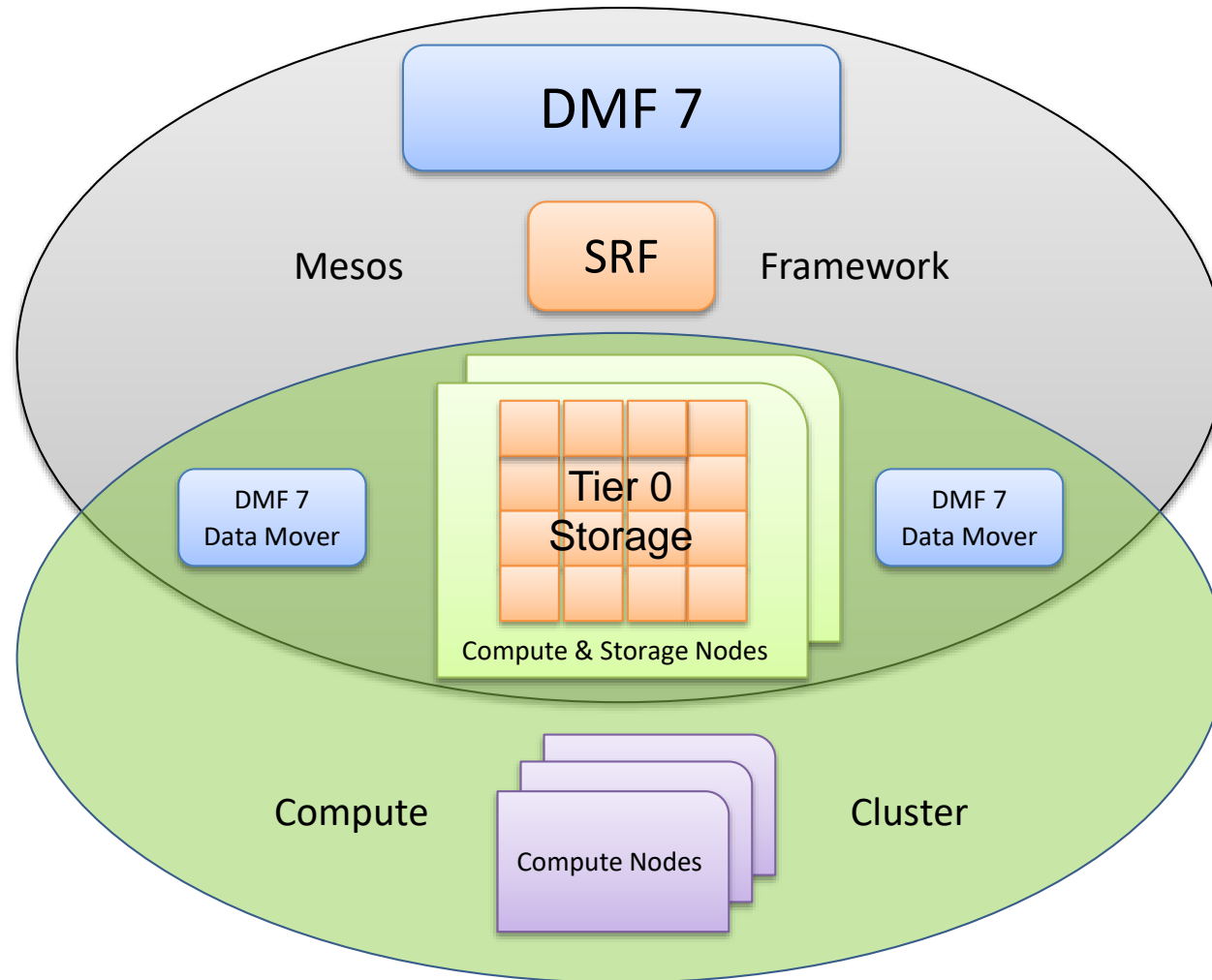


Data Management Fabric | Storage Resource Manager (Mesos)

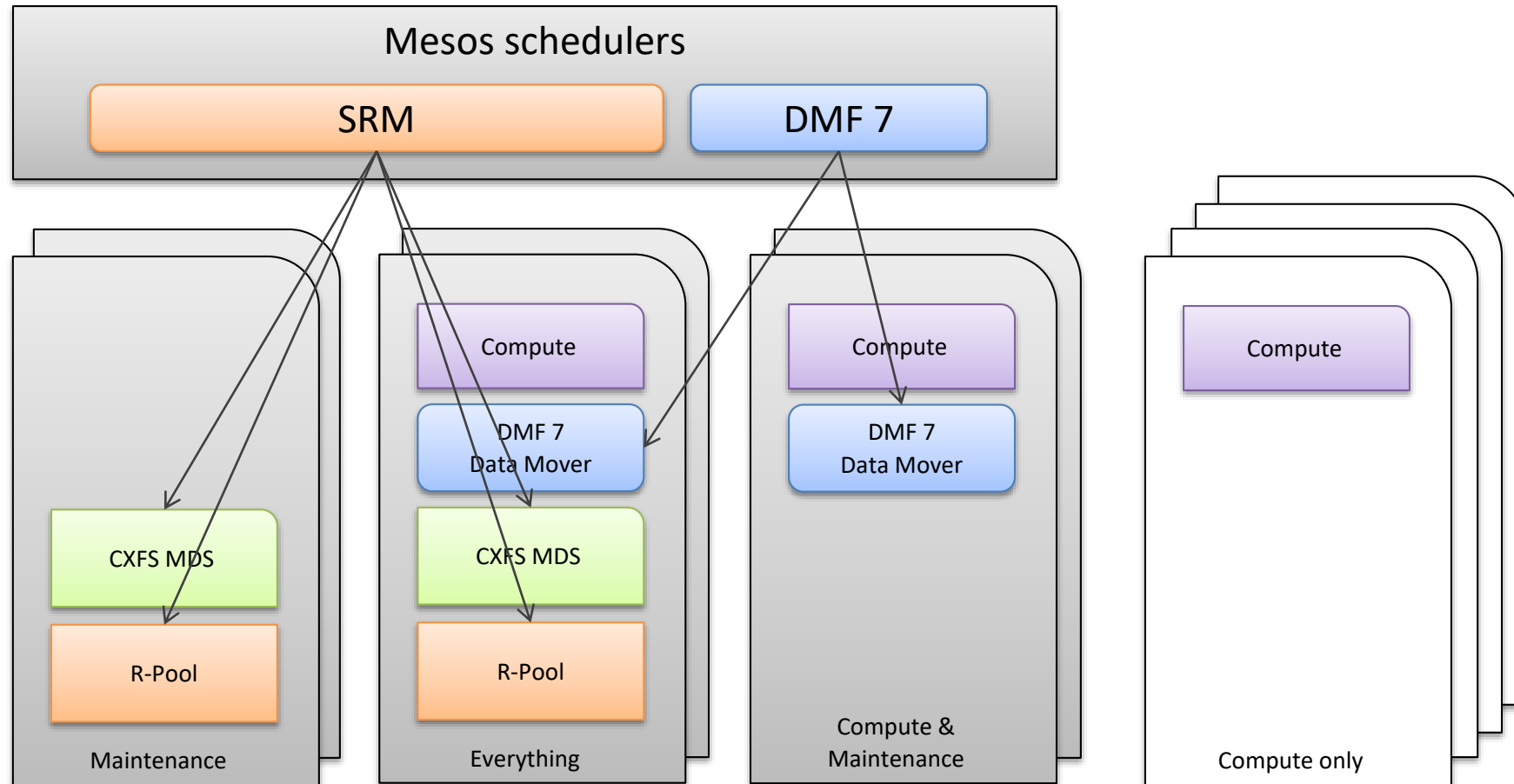


- Long running tasks (LRT) for continuous scanning and reporting status
- Short running tasks to execute actions (create volume, namespace, etc.)

Data Management Fabric | **Compute Cluster** with DMF 7



Data Management Fabric | Compute Cluster Node Configurations





Hewlett Packard
Enterprise

Thank you