

The status of Lustre and other open-source file systems at HPE

Torben Kling Petersen, PhD

Distinguished Technologist,
Global Lead HPC Storage Architect, HPC & AI BU

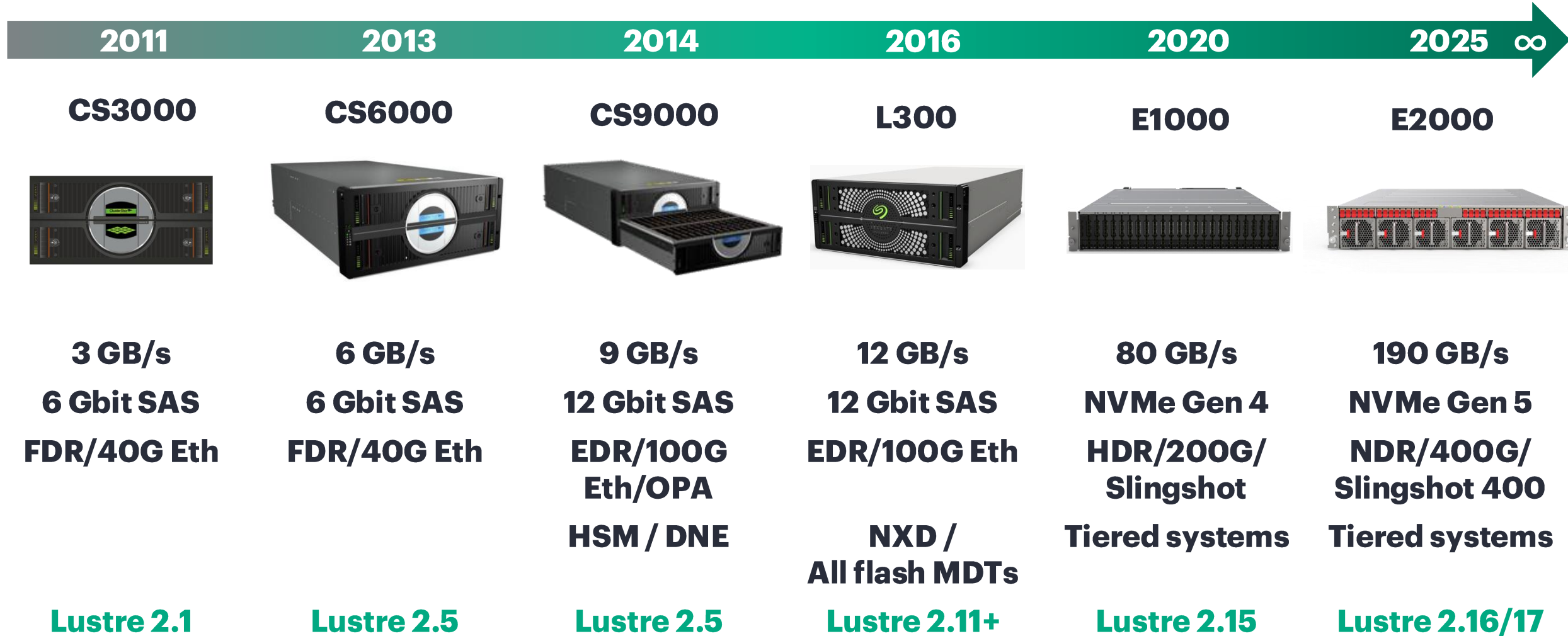
Andreas Müller

HPC Storage Technology Architect
EMEA Supercomputing Team

Agenda

-
- 01** General preamble
 - 02** How has the ever increasing cost of SSDs and RAM change things?
 - 03** Does the “perceived” lack of multi-tenancy features in Lustre hold us back?
 - 04** What about advanced data management and archiving in modern file systems?
-

Brief history of ClusterStor



HPE powers 6 of the world's Top 10 most powerful supercomputers

All powered by HPE Cray Supercomputing Storage Systems

Aggregated:
1.59 EB usable capacity
19 TB/s throughput

First Cray E2000
Storage in Top 10

Lawrence Livermore
National Laboratory

#2



OAK RIDGE
National Laboratory

#3



Argonne
NATIONAL LABORATORY

#4



eni

#6



eni

#8



CSCS

#10



of 100% DLC GPUs

44,544

37,888

63,744

13,920

13,888

10,752

of 100% DLC CPUs

44,544

9,472

21,248

13,920

3,472

10,752

Interconnect

HPE Slingshot

HPE Slingshot

HPE Slingshot

HPE Slingshot

HPE Slingshot

HPE Slingshot

ClusterStor
Storage Systems

400 PB
>2 TB/s

690 PB
11 TB/s

200 PB
1.8 TB/s

38 PB
1.3 TB/s

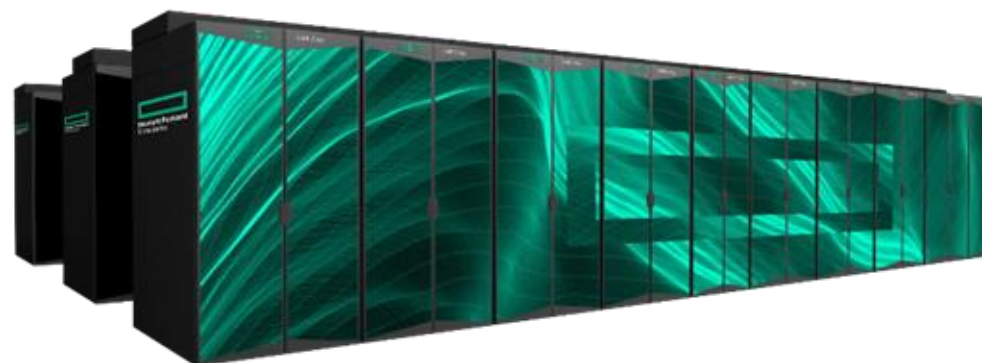
125 PB
1.5 TB/s

116 PB
1.4 TB/s

For the most demanding leadership supercomputing sites*

Two-tier, open-source storage architecture analogous to storage system of Aurora

Supercomputer



Throughput storage



lustre

Lustre on NVMe + HDD tiered solutions
HPE Cray Supercomputing Storage Systems **E2000**

Lustre-based, high performance
SSD/HDD storage systems

IOPS storage



daos

DAOS storage layer running on HPE ProLiant DL servers
HPE Cray Supercomputing Storage Systems **K3000**

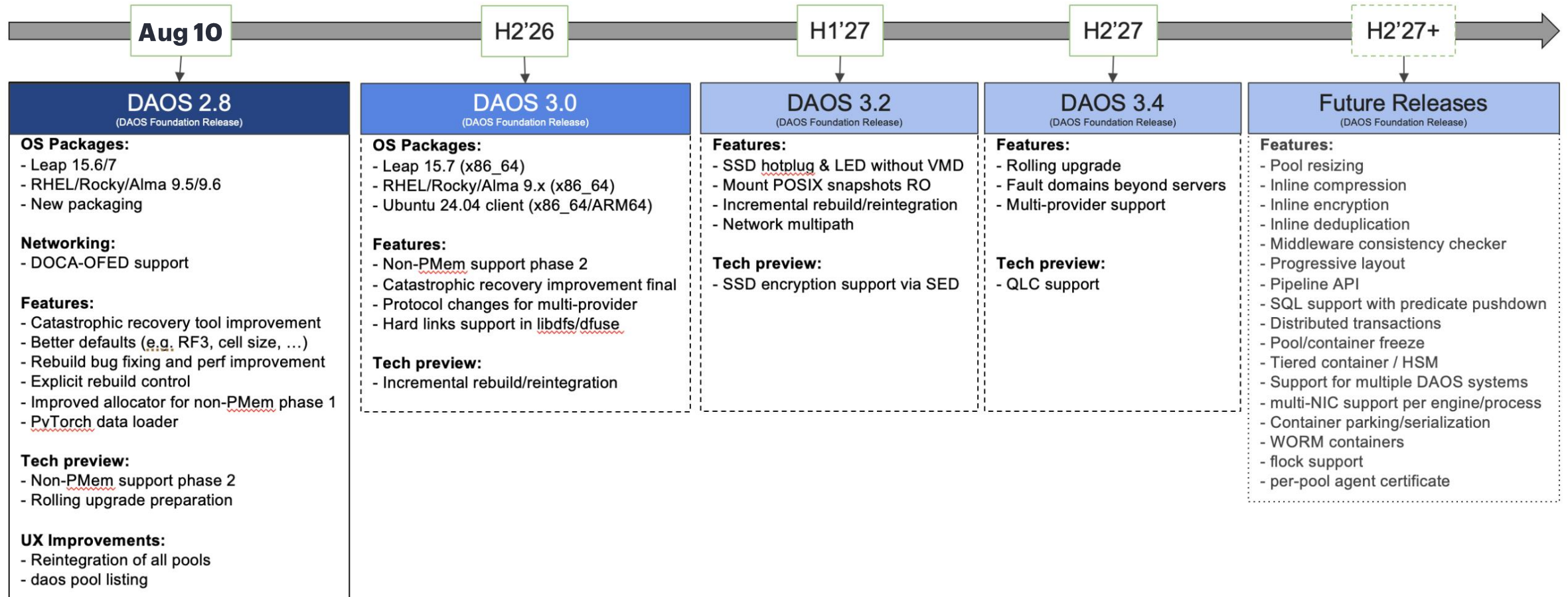
DAOS-based, extreme IOPS layer
All Flash Storage Solution

*For 95%+ of sites the Lustre on NVMe/HDD will deliver sufficient performance for at least 5 years

DAOS roadmap

Color coding schema:

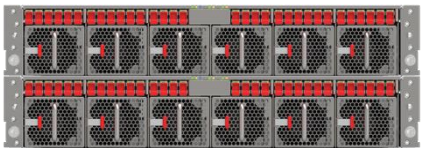
- Committed (or released) release/features
- In-planning release/features
- Future possible release/features



Lustre vs DAOS performance comparison

Comparative sizing:

- 2x E2000 SSU-Fs:
 - 64 E3.S drives @ 15.36 TB
 - 8x SlingShot 400 NICs
 - GridRAID (8+2)
 - Usable capacity: 693 TB
- 4x K3000 nodes
 - 64 E3.S drives @ 15.36 TB
 - 8x SlingShot 400 NICs
 - EC 2+1
 - Usable capacity: 655 TB



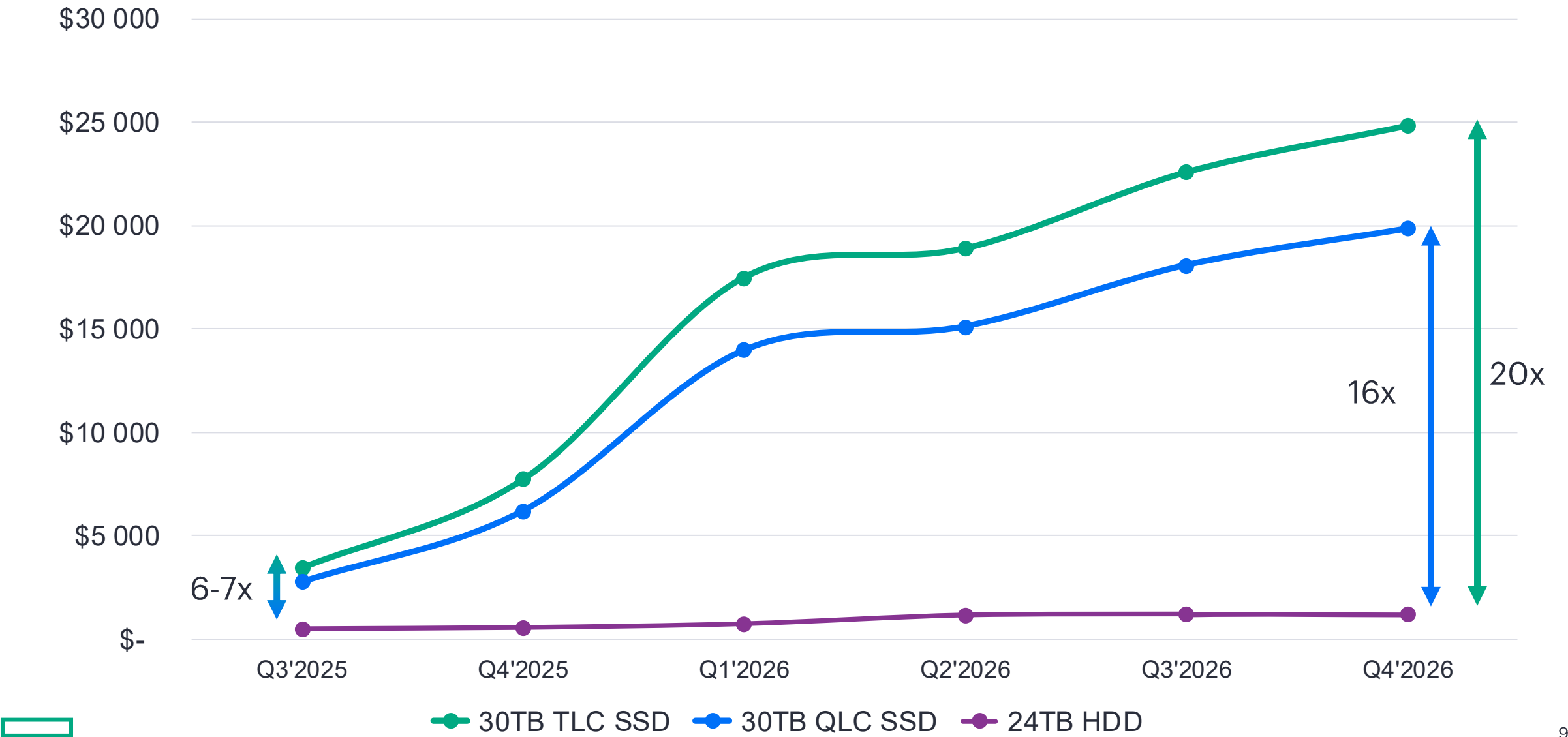
IOR (GB/s)	E2000	K3000
Write	261	248
Read	377	372
IOPS (M)		
Write	0.4	8
Read	4.8	16

Setting the Stage

Flash and RAM prices are changing the world

Isn't it time to re-consider HDDs for both
AI/ML and HPC storage??

Media price increase over time



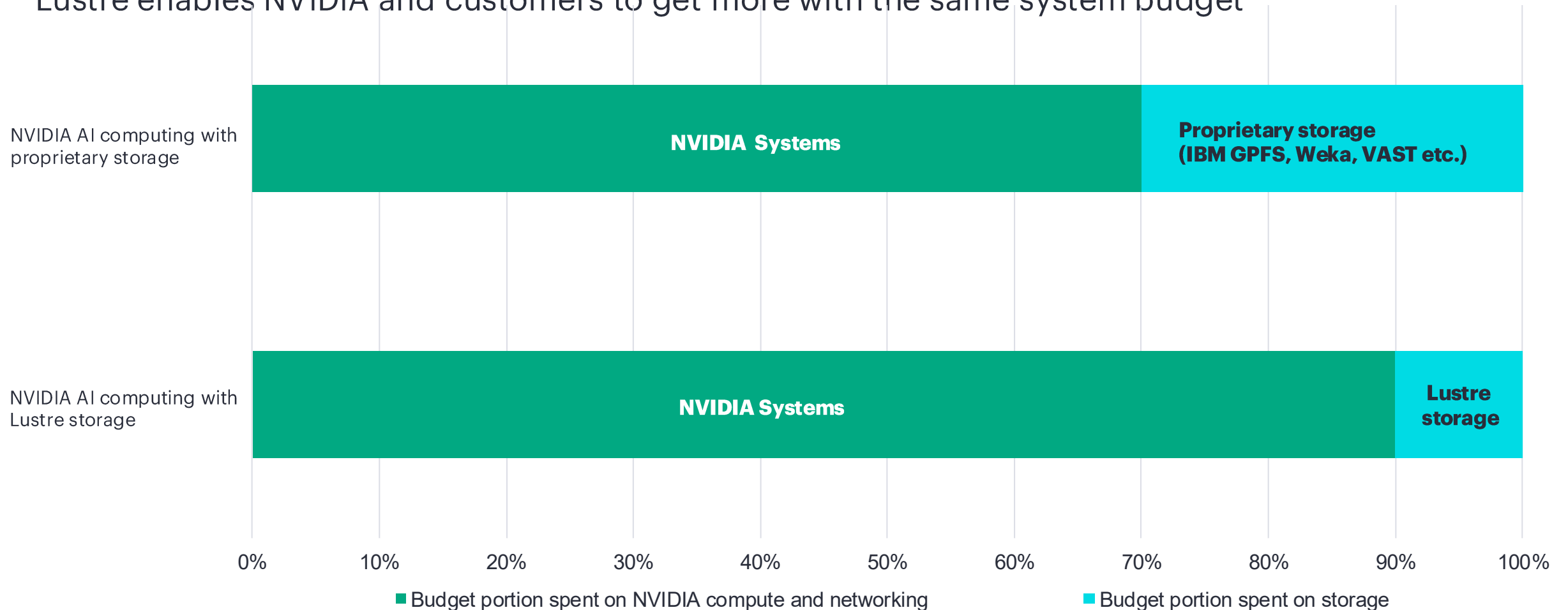
Impact of HDDs in the “big picture”

System comparison based on:
≥ 15 PB usable
≥ 500 GB/s Read
≥ 250 GB/s Write
3-year support



The “Lustre Effect” in large-scale NVIDIA system (PODs and SuperPODs)

Lustre enables NVIDIA and customers to get more with the same system budget



Given a fixed customer budget, customers can buy significantly more NVIDIA GPUs when deployed with cost-effective Lustre storage!

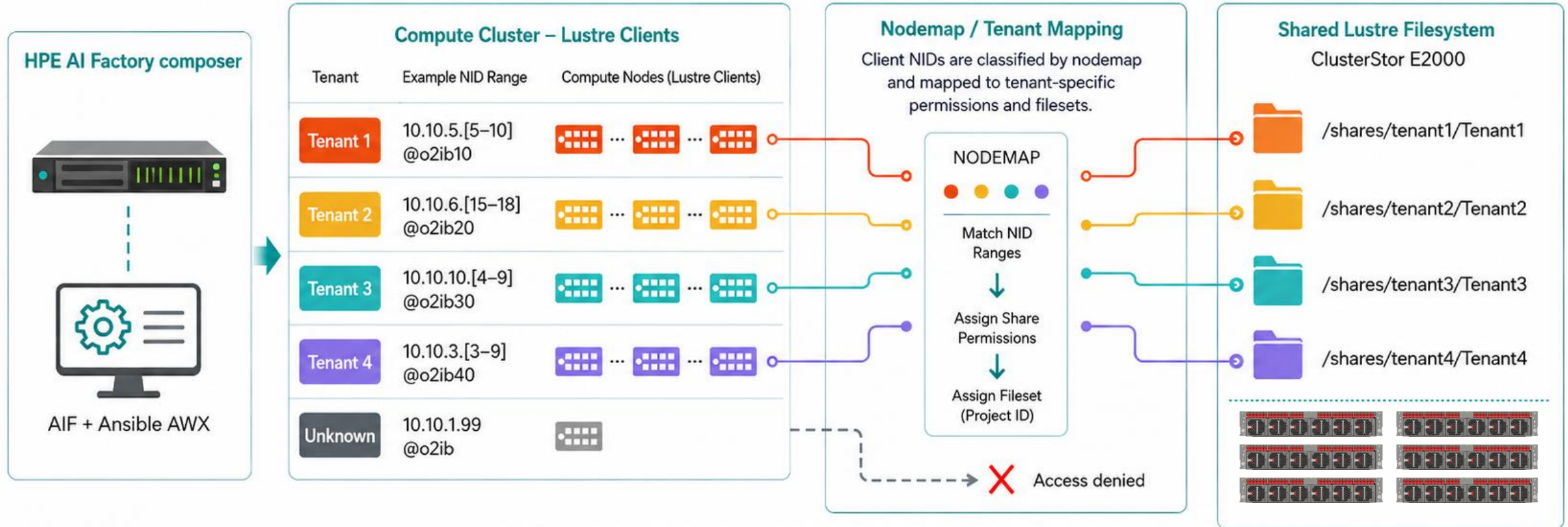
Does the “perceived” lack of multi-tenancy features in Lustre hold us back?

Nvidia RA dictating new functionality and design

- Nvidia is still writing the rule book
 - NCP certification
 - Quantum-X ethernet (with Bluefield NICs)
 - Magnum IO
 - KV Cache & CMX
- Simplistic approach to dynamic setup
 - GUI based
 - Ansible based (or similar)
 - No Root interactions (e.g., no more `lctl` commands on mgs)
 - User self administration ??
- 100% Multi-tenancy
 - Monitoring per tenant
 - Authentication service per tenant
 - Tenant separation
 - VLANs, encryption, auditing, billing etc ...

Multi-tenancy & Isolation – AI Factory Design

One Lustre file system securely partitioned by Nodemap-based isolation



Single Filesystem logically and securely shared between tenants



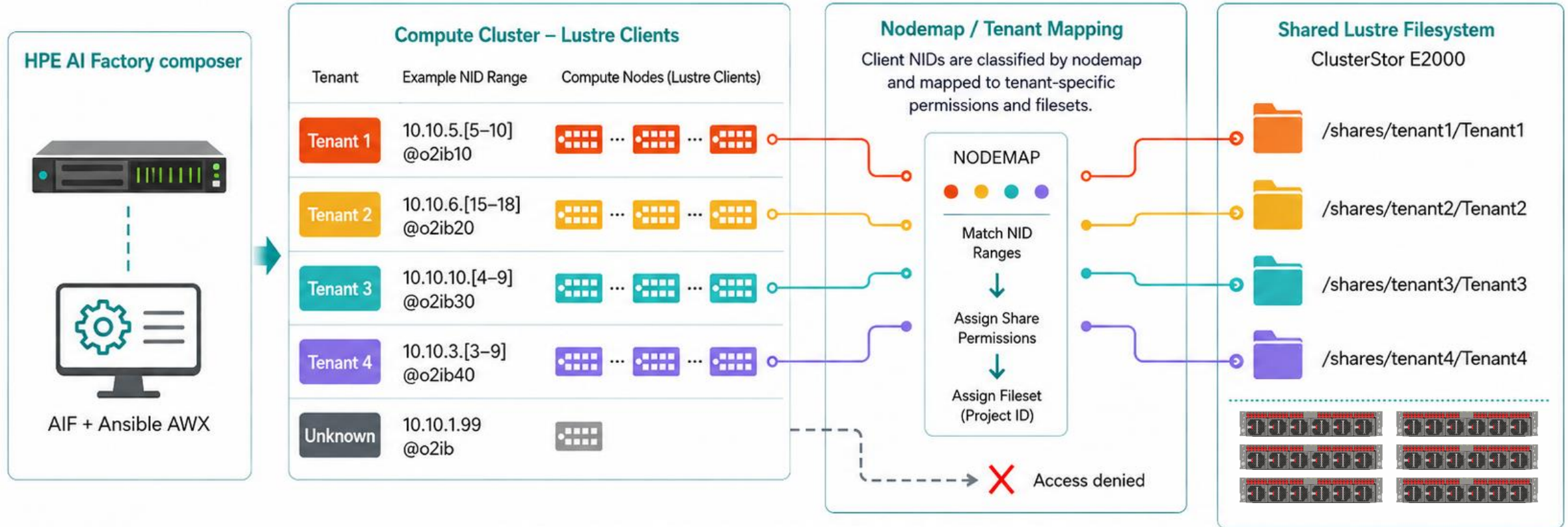
- Tenants operate securely and indcently within a single Lustre namespace
- Nodemap groups client nodes by their NIDs and enforces a designated fileset path (subdirectory)
- Tenant is assigned to a nodemap by defining UID/GID mapping rules



- Each tenant can receive a project ID with inherited POSIX attributes on subdirectories
- Optional QoS, quotas, and network isolation via VLANs or Pkeys
- Fast tenant setup and teardown managed by AIF Ansible AWX automation

Multi-tenancy & Isolation – AI Factory Design

One Lustre file system securely partitioned by Nodemap-based isolation



Single comprehensive control plane

- All multi-tenancy actions are managed by the Neo DP-API (e.g., Redfish/Swordfish) supporting both internal and external token-based access
- ALL monitoring (hardware, software, services etc) and most of the control plane for other Lustre functionality are handled by the DP-API
- ClusterStor use a very similar design to the SlingShot and Cray SuperComputer management planes (all can be controlled by a single instance)

https://support.hpe.com/hpesc/public/docDisplay?docId=sd00002916en_us&page=index.html

ⓘ Allocated IP ranges: 10.230.7.97, 10.230.7.99, 10.230.7.[101-112]

IP Range Start

IP Range End

10.230.7.97

10.230.7.126



Save

Apply

IP Routing ⓘ



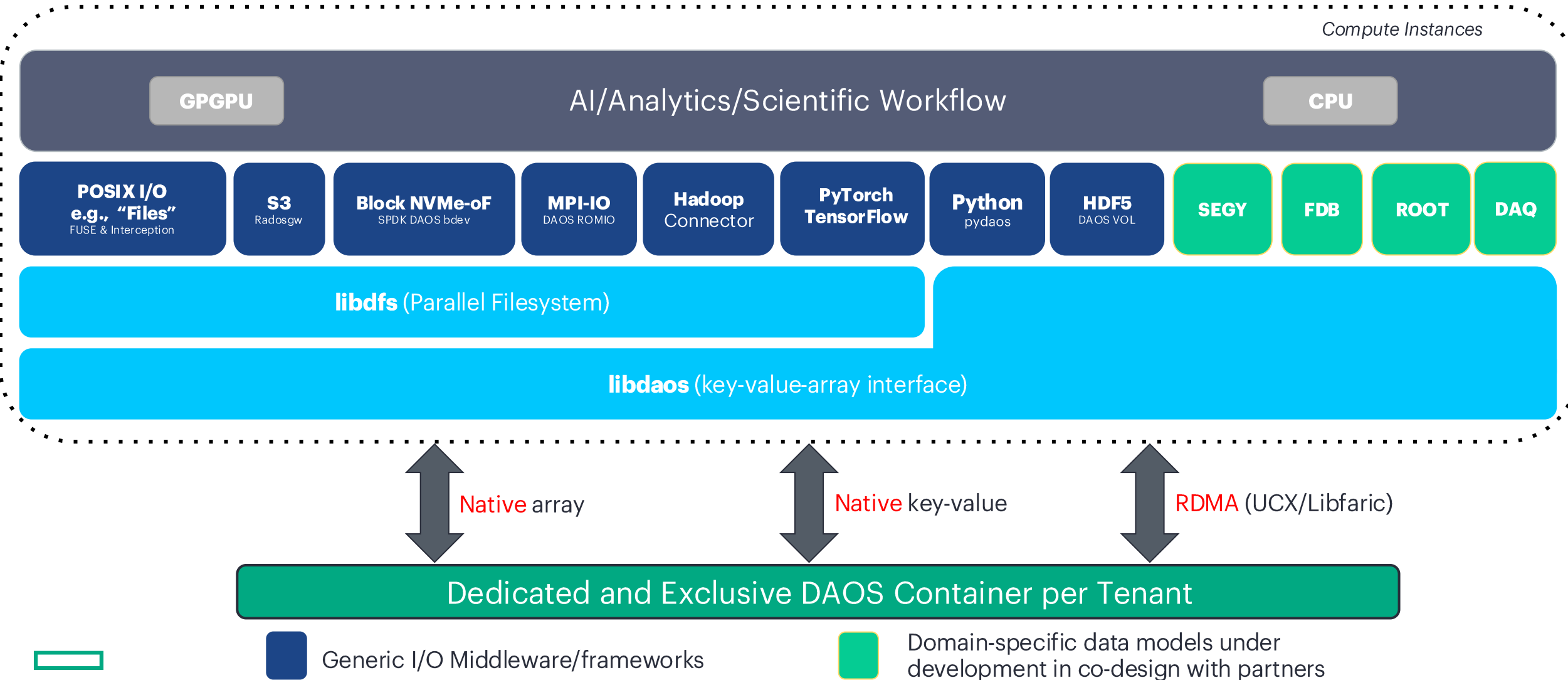
Choose file to upload (.txt)

Save

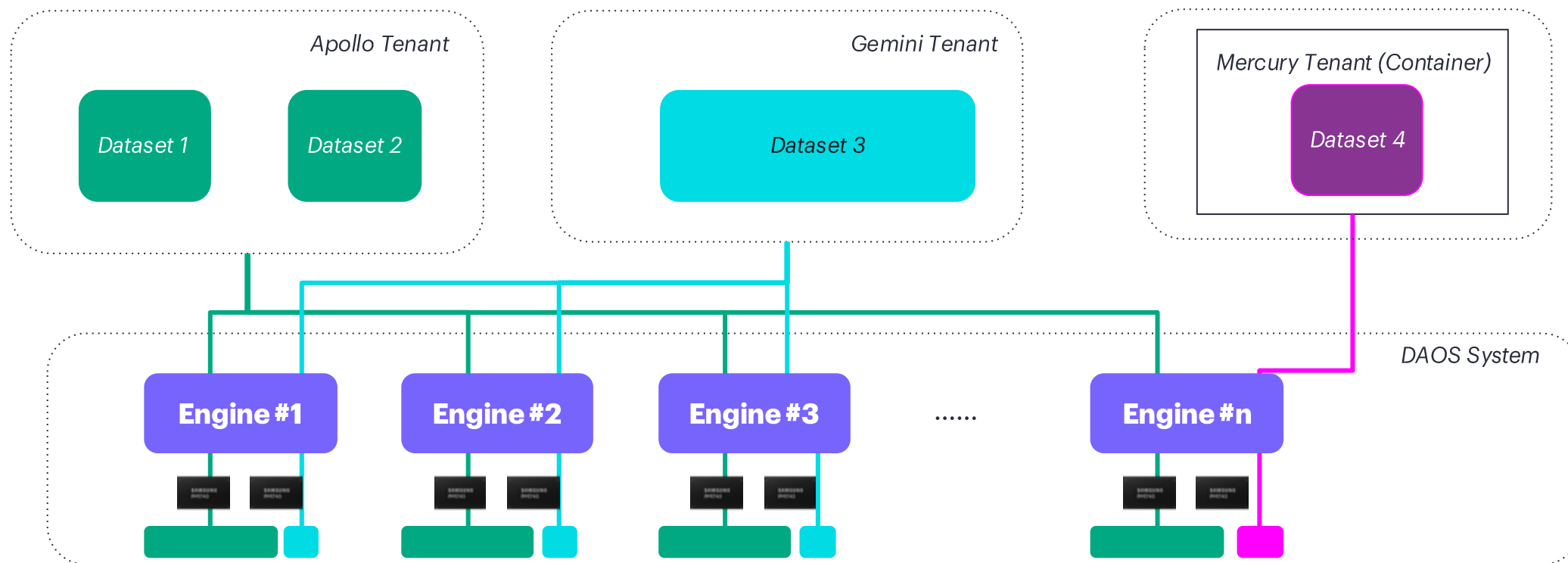
Apply

Multi-tenancy in DAOS

Erasure coding and “I/O Personality” runs on Client



Storage Pooling - Multi-tenancy



Pool 1 	Apollo Tenant	100PB	20TB/s	200M IOPS
Pool 2 	Gemini Tenant	10PB	2TB/s	20M IOPS
Pool 3 	Mercury Tenant	30TB	80GB/s	2M IOPS

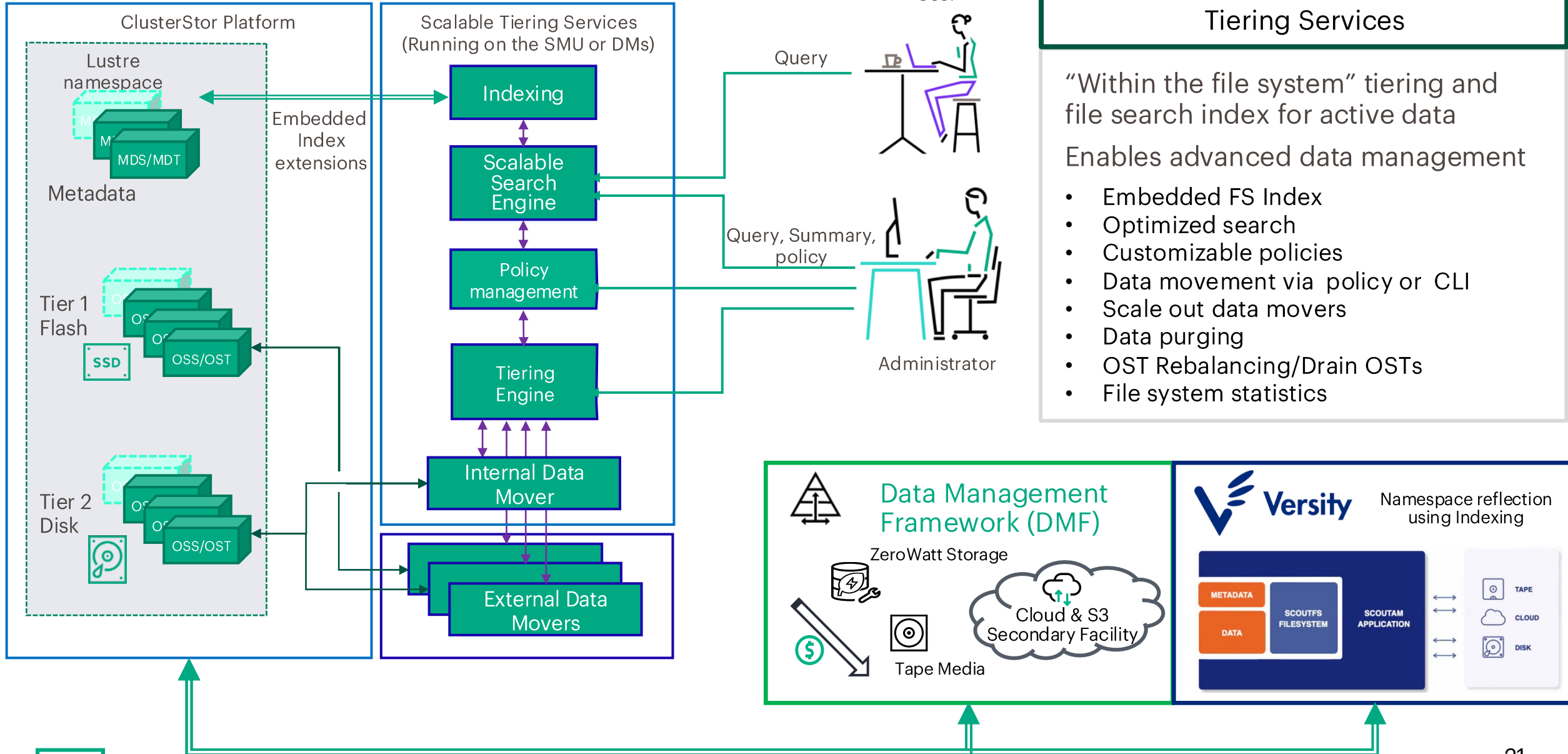
**What about advanced data management
and archiving in modern file systems?**



by Alex Woodie | September 8, 2026



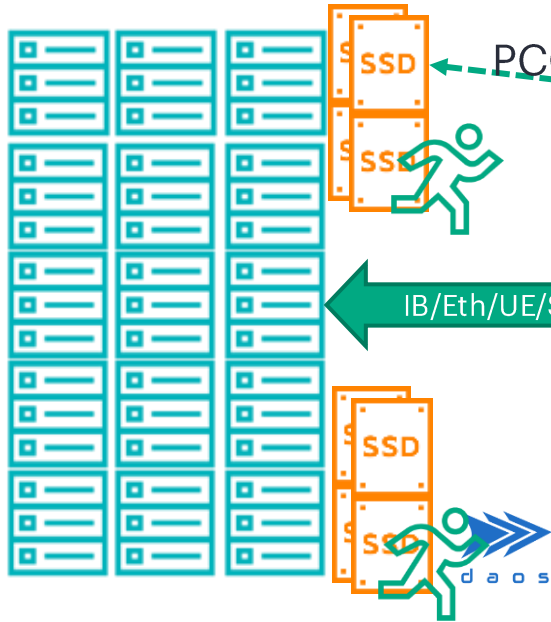
Total Tiering Solution



Data Mobility Framework - Concept

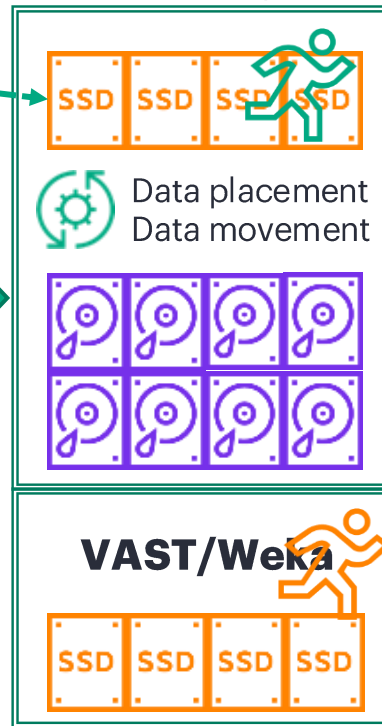
Single Virtual Name Space

Compute system
CPU or CPU/GPU

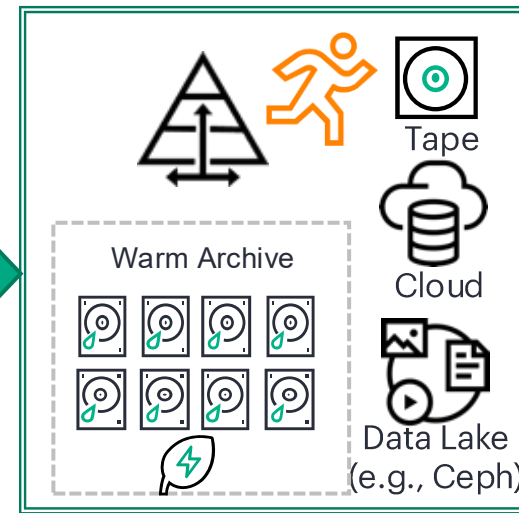


Parallel File Systems

- Lustre
- IBM Storage Scale



Data Mobility
Framework
(CMF, FAIR, iRods)



Wide Area Data
Mobility



Unified Metadata architecture (POSIX + Object)
Policy based, scalable Data Mobility agents and engines

Data Mobility Framework



Thank You

For listening to a madman's ramblings

tkp@hpe.com

andreas.mueller2@hpe.com

