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

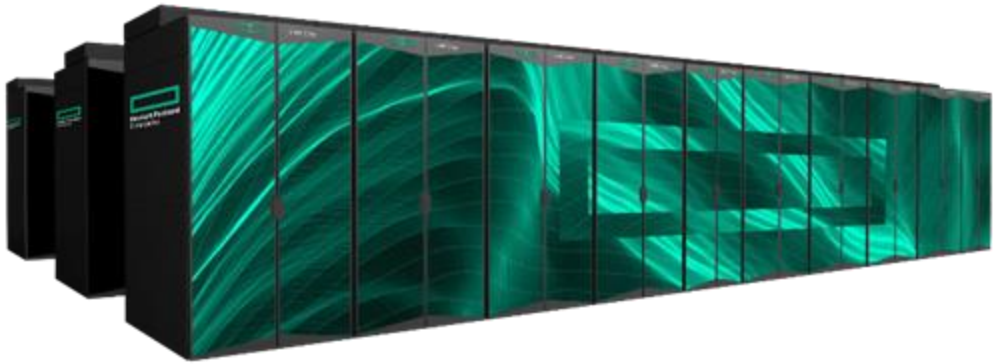
Agenda – Discussion points

- Open-Source at HPE
- How is Lustre holding up against the AI hype solutions?
- Does the “perceived” lack of multi-tenancy features in Lustre hold us back?
- Is performance no longer the key feature customers are looking for?
- What about advanced data management and archiving in modern file systems?

Providing options for leadership supercomputing sites

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

Supercomputer



Q1 - 2026

lustre

daos

Cray ClusterStor Storage Systems E1000
HPE Cray Supercomputing Storage Systems C500/E2000

DAOS storage layer running on HPE ProLiant DL servers

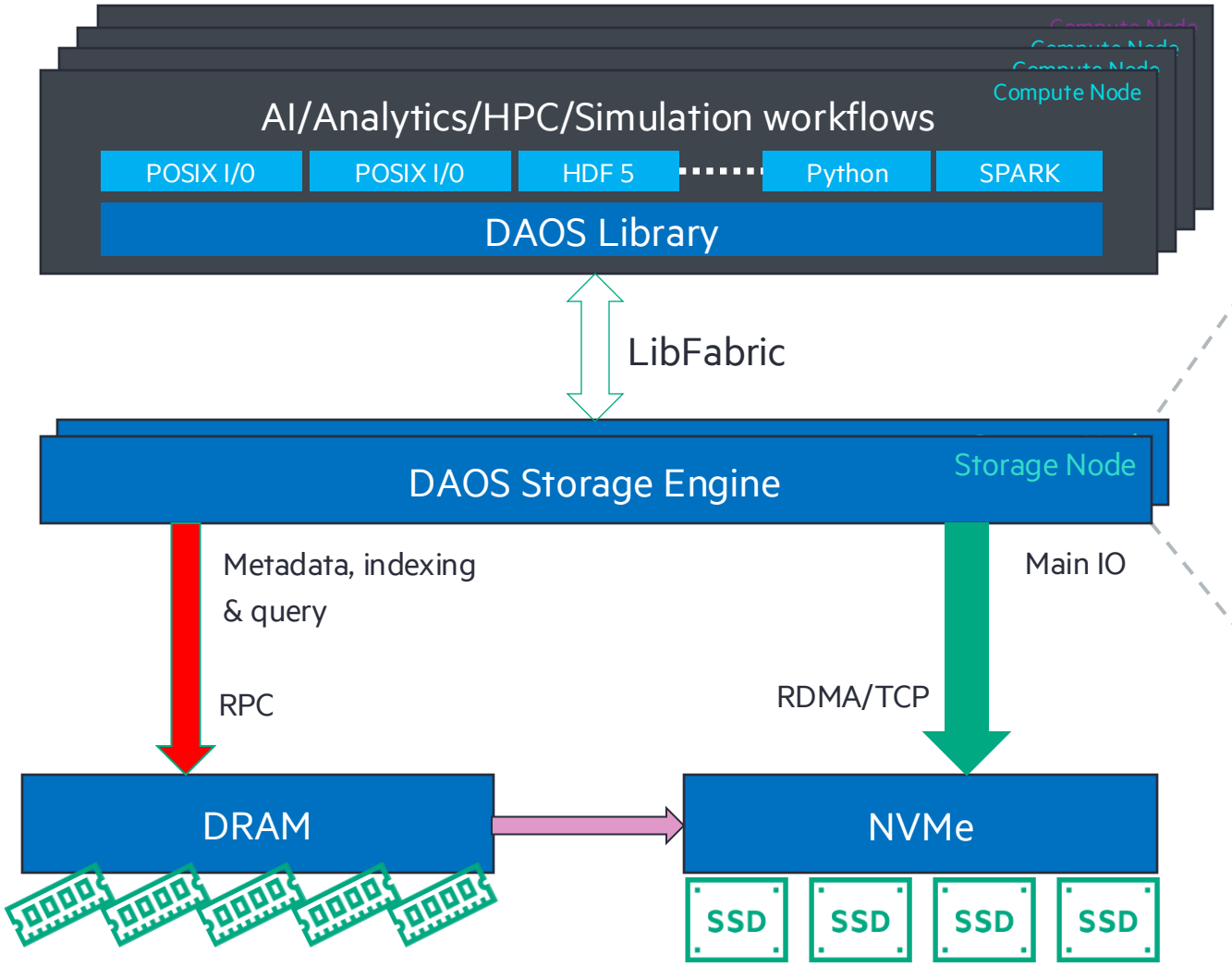
Lustre-based, high performance
SSD/HDD storage systems

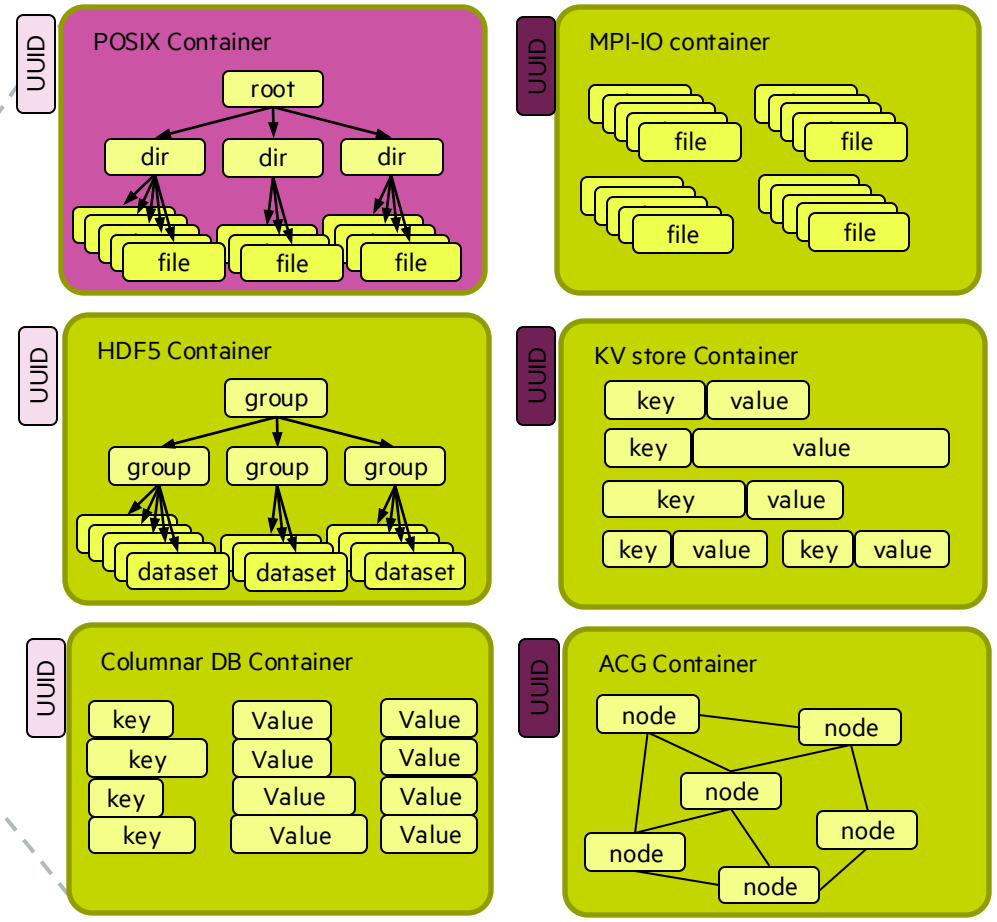
DAOS-based, extreme IOPS layer
All Flash Storage Solution

Why DAOS ??

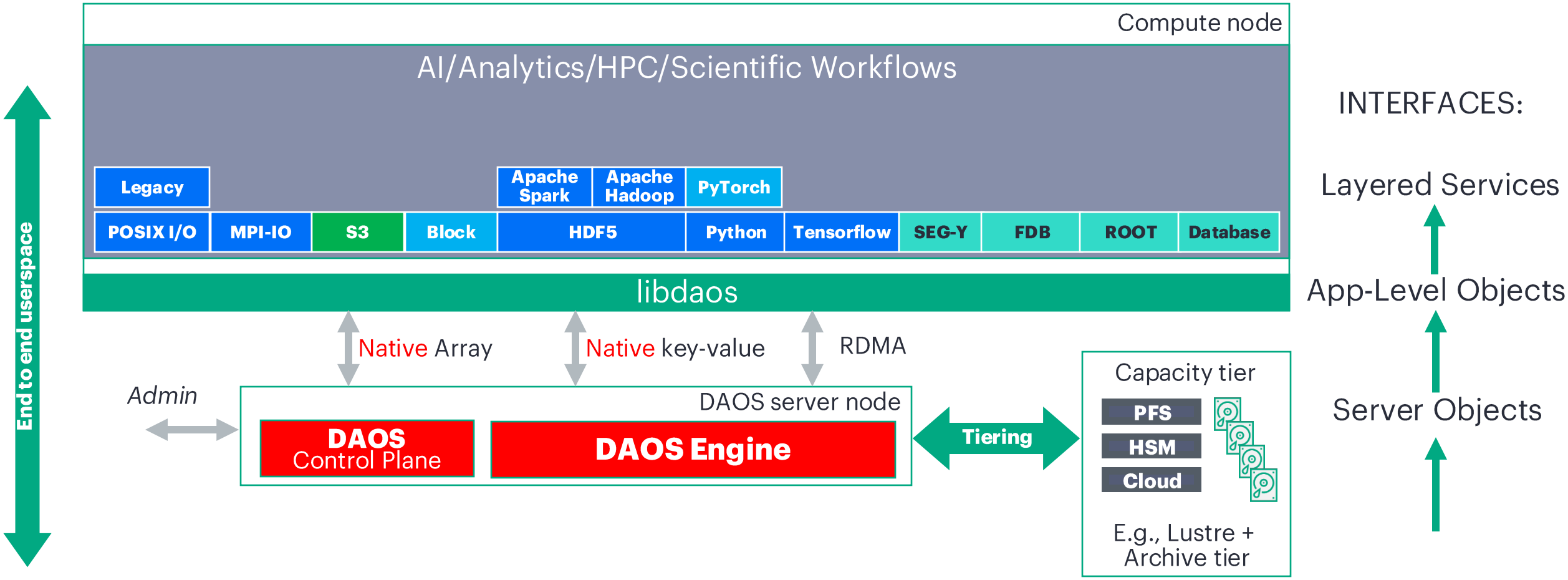
- Augment Lustre where extreme IOPS are required
- All code is in User Space
  - Avoiding difficult kernel developments etc
  - Includes the client which does not require complicated compilation for specific client OS releases
  - Runs on simple single node clusters such as HPE ProLiant systems
- No complicated HA design with twin tailed storage/fail-over
  - All erasure coding is over the network from the client
  - Allows for per file RAID settings supporting everything from replication to n+m RAID schemas
- Write once letting the client handle erasure coding
  - Contrary to other systems such as VAST or Weka that requires the data to be written twice
- Designed by members of the original Lustre development team
  - Experience in creating high performance file systems is key
- 100% Open Source

DAOS Exascale Storage Architecture (DRAM based)





Distributed Asynchronous Object Storage & its Interfaces



How is Lustre holding up against the AI hype solutions?

Short answer: Badly …

Why:

- In the beginning, it was all about read performance and IOPS
- Then it was all about Jupyter Notebooks
- Followed by the “need” for S3 and NFS
- Quotes like “AI is not HPC”
- Multi-tenancy became a big topic

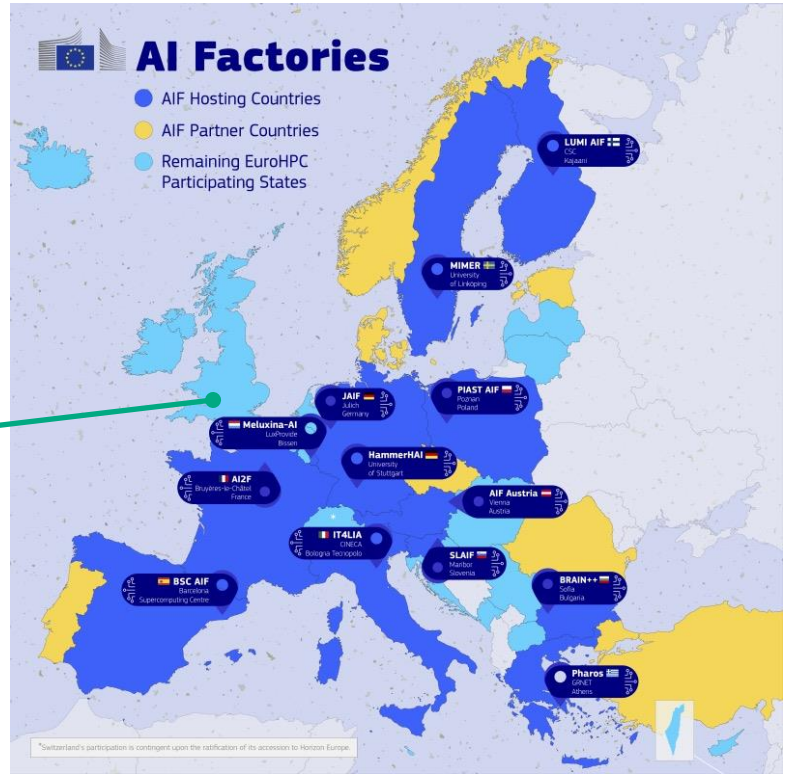
- The competition marketed the hell out of these ideas …
  - And added that every storage admin needed GUI tools to do the work
  - And only SSDs were good enough

Digging deeper …

“Lustre is too complex and modern AI/ML apps require S3 or NFS”

Not true, most of the current sites receiving “AI Factories” already use Lustre !

EuroHPC AI Factories



AI Factory Italy
Delivery FY26

Dutch AI, Netherlands
Delivery FY27

LumiNext, Finland
Delivery FY27

HammerHAI, Germany
Delivery FY26

GRNet AIFactory Greece
Delivery FY26/27

HeliosAI, Poland
Delivery FY26

MeluxinaAI, Lux (LUXProvide)
Delivery FY26

MIMER AIFactory Sweden
Delivery FY26

Spain AI, Spain
Delivery FY26

Is performance no longer the key feature customers are looking for?

Flash system Benchmarking Examples

Single E2000 SSU-F node

- 32 E3.S NVMe
- 2 OST/OSS
- IB NDR
- NPS=4
- 32 NDR clients
- ost_num_threads = 1536
- Direct or buffered I/O with stonewalling

Single SSU-F	IO (GB/s)	E1000 – IB (HDR) Neo 6.6/LDISKFS 1 OST/OSS	E2000 – IB (NDR) Neo 7.0/LDISKFS 2 OSTs/OSS	E2000 – IB (NDR) Neo 7.0/LDISKFS RAID10
Direct I/O	Write	62.6	117.9	88.6
	Read	85.5	185.1	126.1
Buffered I/O	Write	65.4	139.6	86.4
	Read	83.5	190.8	118.2

Type	Write	Read
Single node, single stream	8.6 GB/s	9.7 GB/s
Single node, multi stream	49.5 GB/s	49.3 GB/s
Single SSU-F with 3x NICs (DIO)	119.0 GB/s	266.0 GB/s

Example: GPT-3 LLM Checkpoint @ 60 GB/sec write speed in 85 seconds

IOPS do not matter, only throughput in GB/s as it is sequential write of very large files[1]

Epoch time: 3 hours I NVIDIA recommendation: Checkpoint every 9 minutes I GPT-3 checkpoint size: 5,120 GB in 64 ckpt files (80 GB files)

Duration for writing the checkpoint: 5,120/60 = 85 seconds

Cray ClusterStor E2000 (1 SSU-F)	HPE Solutions for Weka (7 HPE Alletra 4110)	Vast Data Universal Storage (9x9 Cluster)
• 4 CPUs • 64 NVMe SSD • HSN ports: 8	• 14 CPUs • 63 SSD • HSN ports: 28	• 36 CPUs • 9 Flash Enclosures • HSN ports: 72

[1] Source: VAST Data presentation, LLM Checkpointing – IO Calculations, S. Kartik, 2023

Scenario: Customer wants to cut down checkpoint time with 240 GB/s storage

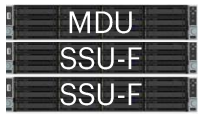
To go from **85 seconds wait/idle time** for GPU nodes every 9 minutes to **22 seconds**

Customer gets 730,000 more GPU hours out of the 1,024 GPU cluster per year!

Epoch time: 3 hours I NVIDIA recommendation: Checkpoint every 9 minutes I GPT-3 checkpoint size: 5,120 GB in 64 ckpt files (80 GB files)

Duration for writing the checkpoint: 5,120/240 = 22 seconds

Cray ClusterStor E2000 (2 SSU-F)	HPE Solutions for Weka (14 HPE Alletra 4110)	Vast Data Universal Storage (36x36 Cluster)
• 6 CPUs • 96 NVMe SSD • HSN ports: 12	• 28 CPUs • 224 NVMe SSD • HSN ports: 56	• 144 CPUs • 36 Flash enclosures • HSN ports: 288



Just add 1 SSU-F.
Drive performance with at least
57% fewer CPUs and 65% fewer SSDs!

This is why Lustre should be de-facto standard for large GPU clusters on-premises and in the public cloud!

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

Including other notable “issues”

How Lustre is often described by the competition .....

Feature	VAST	Weka	Lustre
VLAN support	✓	✓	✗
DEAR	✓	✓	-
Encryption over the Wire	✓	✓	✗
Multi-Tenant Access control	✓	✓	✗
Multi-Tenant Isolation	✓	✓	✗
Multi-Tenancy over shared HW	✓	✓	✗
Multi-Tenant Performance	✓	✓	✗
Fine grained quotas	✓	✓	-
No requirement for custom clients	✓	✓	✗
Support for NFS/S3/SMB	✓	✓	✗
Advanced Data Management	✓	✓	✗
GUI based administration	✓	✓	✗

How Lustre is often described by the competition .....

Feature	VAST	Weka	Lustre
VLAN support	✓	✓	✓
DEAR	✓	✓	✓
Encryption over the Wire	✓	✓	✓
Multi-Tenant Access control	✓	✓	✓
Multi-Tenant Isolation	✓	✓	✓
Multi-Tenancy over shared HW	✓	✓	✓
Multi-Tenant Performance	✓	✓	✓
Fine grained quotas	✓	✓	✓
No requirement for custom clients	✓	✓	✗
Support for NFS/S3/SMB	✓	✓	✓
Advanced Data Management	✓	✓	✓
GUI based administration	✓	✓	✓

According Your Needs and Use Cases

NFS Connection Protocol	Single Mount BW (Read)
LEGACY NFS (TCP SINGLEPATH)	**~ 2 GB/s**
VAST NFS (RDMA SINGLEPATH)	**~ 8,5 GB/s**
NFS NCONNECT (TCP w UPSTREAM)	**~ 10 GB/s**
NFS MULTIPATH (TCP w MULTIPATH)	**~ 20 GB/s**
VAST NFS MULTIPATH (RDMA w MULTIPATH)	**~ 50 GB/s**
VAST NFS MULTIPATH w/ GPUDirectStorage	**~ 175 GB/s**

Complexity in setting up Multi-Tenancy ??

Lustre requires admins to be root on the mgs node:

```
# create admin nodemap
mgs# lctl nodemap_add admin
mgs# lctl nodemap_add_range --name admin --range <n00 NIDs>
mgs# lctl nodemap_add_range --name admin --range <n01 NIDs>
mgs# lctl nodemap_modify --name admin --property trusted --value 1
mgs# lctl nodemap_modify --name admin --property admin --value 1
mgs# lctl nodemap_modify --name admin --property deny_unknown --value 0

# define default nodemap properties
mgs# lctl nodemap_modify --name default --property trusted --value 0
mgs# lctl nodemap_modify --name default --property admin --value 0
mgs# lctl nodemap_modify --name default --property deny_unknown --value 1
mgs# lctl nodemap_set_fileset --name default --fileset /null

# Enable user, group, and project quotas
mgs# lctl conf_param <fsname>.quota.ost=ugp
mgs# lctl conf_param <fsname>.quota.mdt=ugp

# enable nodemaps
mgs# lctl nodemap_activate 1
```

Complexity in setting up Multi-Tenancy ??

In ClusterStor, we will allow admins (on the admin nodes) to set this up with a single command:

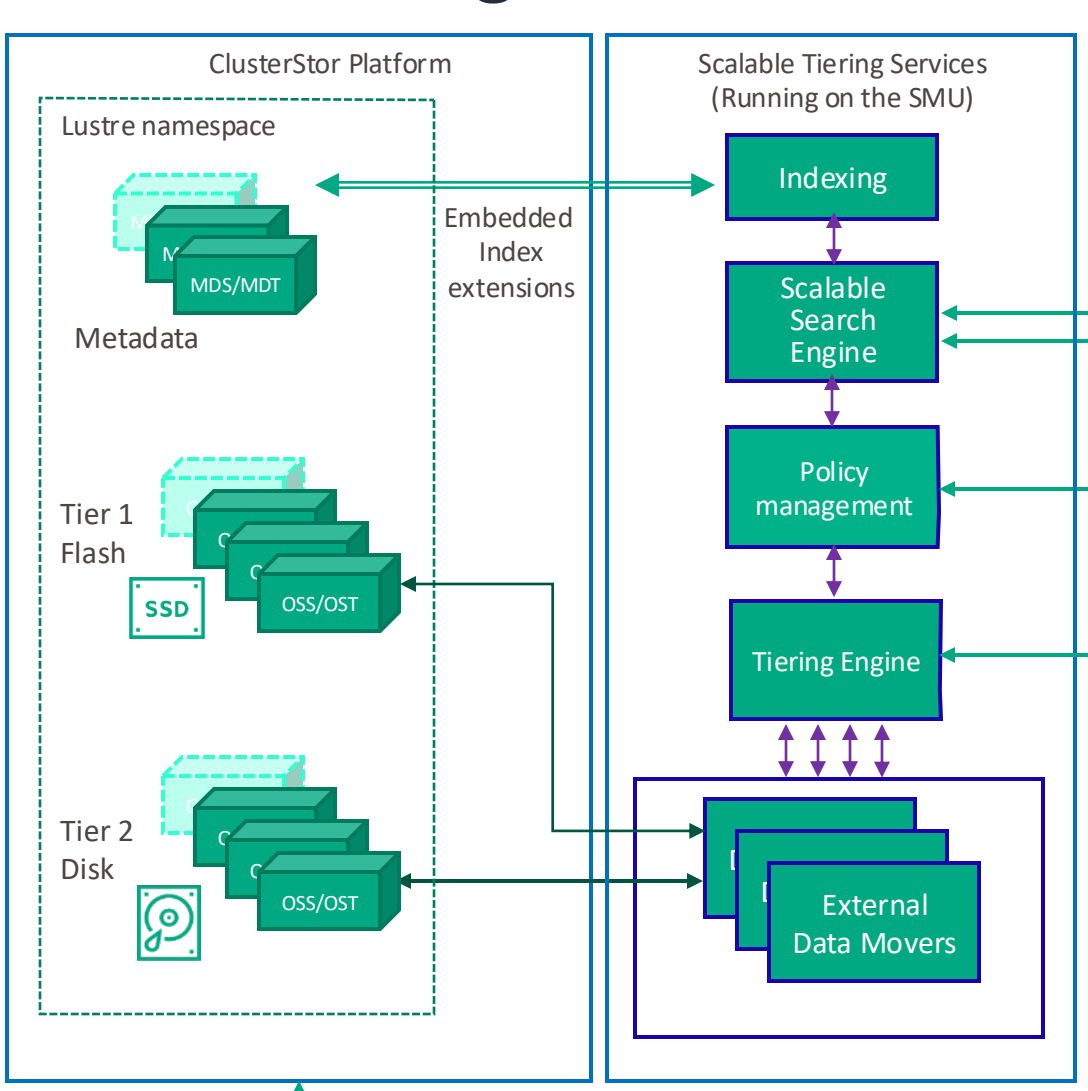
```
cscli lustre mt manage enable [-h]
```

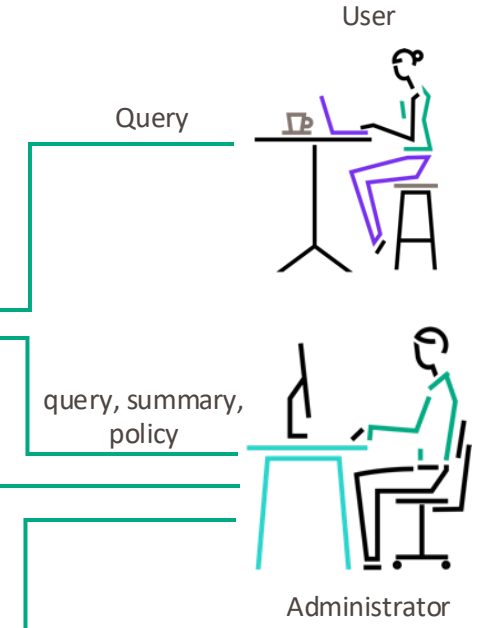
Based on a pre-defined (but editable) json file:

```
"SMT": {
    "Enabled": true,
    "Nodemaps": [
        {
            // details of admin node map, i.e. n00 and n01 nids
            "Name": "admin",
            "NidRange": "<n00-n01 nids>"
            "Trusted": 1,
            "Admin": 1,
            "DenyUnknown": 0
        },
        {
            // details of default node map
            "Name": "default",
            "Trusted": 0,
            "Admin": 0,
            "DenyUnknown": 1
        }
    ]
    "ProjectQuotas": {
        "<fsname>.quota.ost": "ugp",
        "<fsname>.quota.mdt": "ugp",
    },
    "NodeMapActive": 1,
    "OverSubscriptionEnabled": false
}
```

Advanced data management and archiving in modern file systems?

Total Tiering Solution



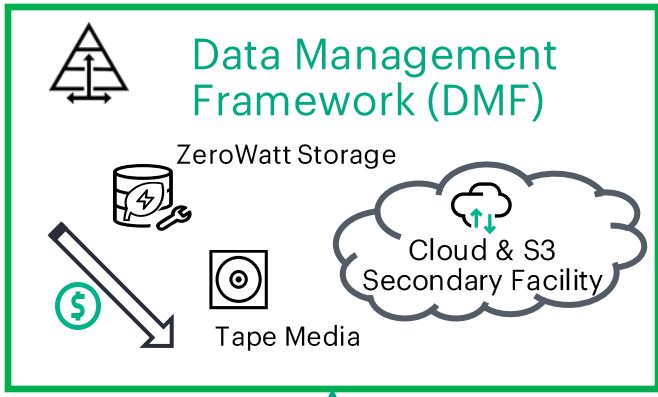


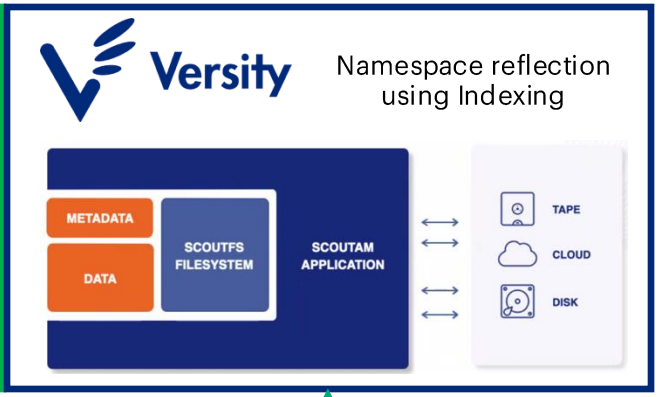
Tiering Services

"Within the file system" tiering and file search index for active data

Enables advanced data management

- Embedded FS Index
- Optimized search
- Customizable policies
- Data movement via policy or CLI
- Scale out data movers
- Data purging
- OST Rebalancing/Drain OSTs





Bottom line, how do we actually compare .....

Feature	VAST	Weka	Lustre	DAOS
VLAN support	✓	✓	✓	✓
Multi-Tenancy Support	✓	✓	✓	✓
End to End Encryption	✓	✓	✓	✓
Support for NFS/S3/SMB	✓	✓	✓	✓
Advanced Data Management	✓	✓	✓	✓
Atomic writes	✗	✗	✓	✓
Custom Clients	✓	✓	✓	✓
HDD support	✗	✗	✓	✗
Open-Source	✗	✗	✓	✓
Scalability (Cap/Perf)	++	+	+++	++
Cost at scale	€€€	€€€€	€	€€
ILOM stack	✗	✗	✓	(✓)
Marketing budgets	€€€€	€€€	-	-

Thank You

For listening to a madman's ramblings

tkp@hpe.com
andreas.mueller2@hpe.com