# The Status of Lustre at HPE

## OpenZFS, performance enhancements and other updates.

**Torben Kling Petersen**, PhD

**Distinguished Technologist**
**Lead HPC Storage Architect - EMEA & APAC**

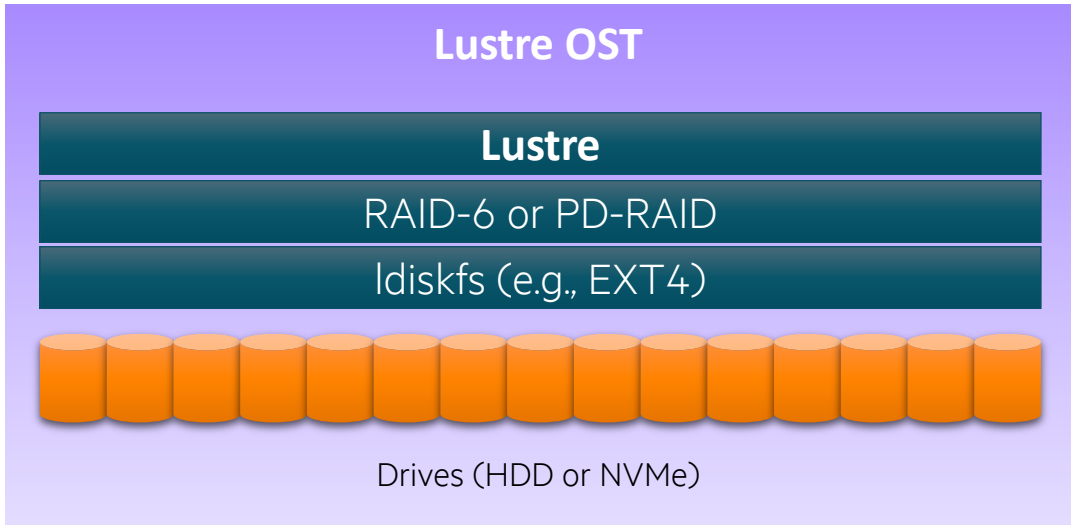
# Agenda ??

- OpenZFS etc …
- Performance enhancements
  - Lustre as a function of network protocols
- Data Management Framework
  - Update on current status

# OpenZFS

# Data Path Options

## Lustre Disk File System

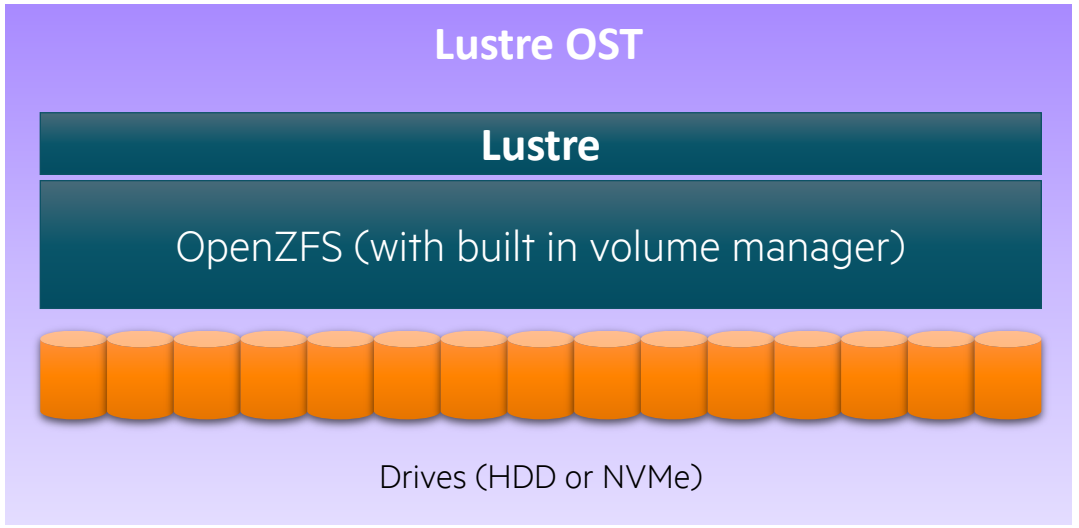



**ldiskfs benefits**

Better performance/IOPS
"Simple" trouble shooting
Proven functionality
Support for T10-PI & DPS

Fully supported to mix LDISKFS and OpenZFS in the same Lustre namespace

## Zettabyte File System




**ZFS benefits**

Copy on Write
Compression
Snapshot capability
Built in data integrity

HPE activity in the OpenZFS community

- Number of Patches: 12
- Number of Reviews: 10
- Number of Issues Caught , Regressions and worked with community to solve them: 10+

# Performance considerations – All Flash Arrays

- 20+ clients, Stonewalling IOR, GridRAID of OpenZFS (2x 12 NVMe based OSTs)
- Lustre 2.15 Clients and Server etc ….

| | IO (GB/s) | HDR IB LDISKFS | HW RoCE 200 GbE LDISKFS | HW RoCE 200 GbE OpenZFS - dRAID2 | HW RoCE 200 GbE lz4 compression |
|---|---|---|---|---|---|
| DIO 64PPN (GB/s) | Write | 56.6 | 59.3 | 52.2 | 45.7 |
| | Read | 85.4 | 85.4 | 79.2 | 65.0 |
| BIO 64PPN (GB/s) | Write | 61.3 | 63.2 | 55.2 | 48.9 |
| | Read | 83.4 | 83.6 | 76.6 | 54.2 |
| | | | | | |
| IOR Buffered IO 4K Random IOPS | Write | 88 652 | 85 236 | 26 777 | 25 190 |
| | Re-Write | 56 623 | 54 558 | 6 779 | 4 848 |
| | Read | 1 284 750 | 1 232 360 | 20 812 | 17 304 |

# Performance – OpenZFS on HDD based OSTs

- 20+ clients, Stonewalling IOR FFP, GridRAID (OST size: 53 HDDs) or OpenZFS dRAID2 (53[16d:2p:2s])
- Lustre 2.15 Clients and Server etc ….

| SSU-D# (GB/s) | IO (GB/s) | HDR IB LDISKFS | HW RoCE 200 GbE LDISKFS | HW RoCE 200 GbE OpenZFS - dRAID2 | HW RoCE 200 GbE lz4 compression |
|---|---|---|---|---|---|
| D1 - DIO 64PPN | Write | 18.3 | 18.3 | 9.3 | 19.6 |
| | Read | 19.1 | 18.8 | 19.6 | 38.9 |
| D1 - BIO 64PPN | Write | 17.8 | 18.2 | 8.7 | 18.7 |
| | Read | 15.4 | 15.7 | 18.3 | 33.3 |
| D2 - DIO 64PPN | Write | 35.6 | 36.1 | 15.1 | 26.9 |
| | Read | 34.8 | 36.3 | 31.2 | 42.9 |
| D2 - BIO 64PPN | Write | 34.1 | 34.8 | 23.7 | 42.0 |
| | Read | 29.8 | 30.6 | 28.4 | 44.5 |

# Performance update

# Performance Update – All Flash Arrays

- Samsung PM1733
- 20+ clients, Stonewalling IOR, GridRAID-12, Lustre 2.15 etc ….

**GridRAID**

| | IO (GB/s) | IB (HDR) Neo 6.4-010.86, 2.15 B5 Client | KFI 6.4-010.75, SHS 2.0.2 RC3 | TCP/IP (SS) 6.4-010.75 SS 2.0.2 RC3 | KFI to o2ib Routing Server 6.4-010.75 Slingshot Switch* |
|---|---|---|---|---|---|
| DIO 64PPN | Write | 57.8 | 57.9 | 20.4 | 56.3 |
| | Read | 85.4 | 82.2 | 71.6 | 72.7 |
| BIO 64PPN | Write | 60.9 | 63.2 | 20.9 | 59.2 |
| | Read | 83.1 | 83.6 | 58.1 | 69.9 |
| | | | | | |
| IOR Buffered IO 4K Random IOPS | Write | 84,843 | 86,015 | 47,980 | 84,587 |
| | Re-Write | 54,760 | 50,009 | 66,542 | 52,983 |
| | Read | 1,228,842 | 807,673 | 892,924 | 626,794 |

* Storage on KFI, 20 clients on HDR IB, 4 LNET routers

# Performance Update E1000 – HDD based OSTs

- 20+ clients, Stonewalling IOR, GridRAID-53, Lustre 2.15 etc ….

**GridRAID**

| SSU-D# | IO (GB/s) | IB (HDR) Neo 6.4-010.86 | KFI SHS 2.0.2 RC3 6.4-010.75 | TCP/IP SS 2.0.2 6.4-010.75 | KFI to o2ib Routing |
|---|---|---|---|---|---|
| D1 - DIO 64PPN | Write | 18.2 | 20.8 | 18.2 | 21.1 |
| | Read | 18.6 | 19.5 | 18.6 | 17.5 |
| D1 - BIO 64PPN | Write | 17.7 | 19.8 | 18.7 | 20.1 |
| | Read | 14.5 | 15.6 | 15.0 | 12.7 |
| D2 - DIO 64PPN | Write | 36.0 | 39.3 | 19.8 | 38.1 |
| | Read | 34.1 | 37.3 | 37.8 | 34.9 |
| D2 - BIO 64PPN | Write | 34.2 | 37.7 | 20.0 | 38.0 |
| | Read | 30.5 | 31.4 | 31.7 | 27.0 |

- 20+ clients, Stonewalling IOR, GridRAID-53, Lustre 2.15 etc ….

# Single MDT RAID-10

MDT0
0K files
Non-DOM
Unique Directory
Files Only

| MDtest (Single MDT) | IB (HDR) Neo 6.4-010.86 | KFI SHS 2.0.2 RC3 6.4-010.75 | TCP/IP SS 2.0.2 6.4-010.75 | KFI to o2ib Routing |
|---|---|---|---|---|
| File Creates per Second | 151,498 | 128,755 | 130,818 | 93,779 |
| File Stats per Second | 692,449 | 524,554 | 567,823 | 339,517 |
| File Reads per Second | 278,693 | 246,305 | 204,109 | 121,274 |
| File Removes per Second | 140,252 | 135,207 | 123,293 | 111,930 |

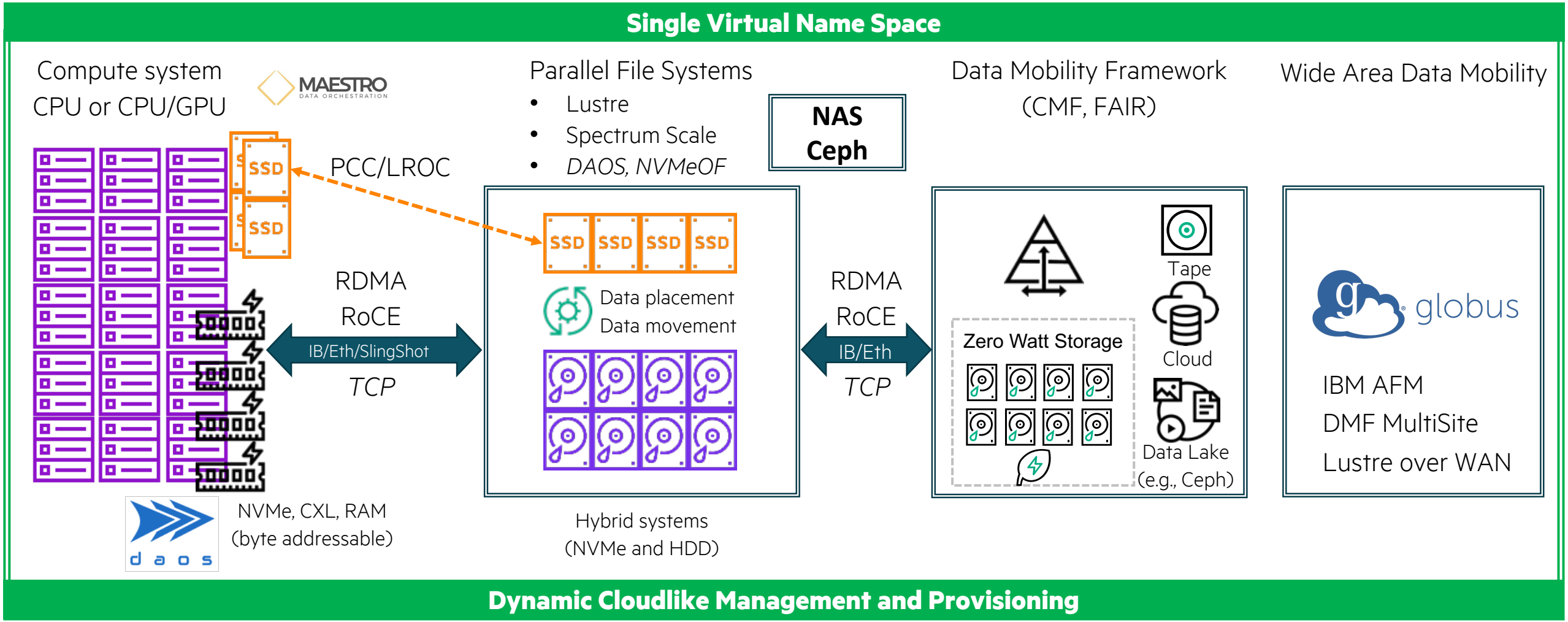
MDT0
0K files
Non-DOM
Unique Directory
Directory+Files

| Single MDT | IB (HDR) Neo 6.4-010.86 | KFI SHS 2.0.2 RC3 6.4-010.75 | TCP/IP SS 2.0.2 6.4-010.75 | KFI to o2ib Routing |
|---|---|---|---|---|
| Directory Creates/sec | 107,436 | 86,704 | 98,210 | 73,931 |
| Directory Stats/sec | 428,791 | 372,082 | 365,379 | 305,304 |
| Directory Removes/sec | 166,191 | 172,449 | 173,496 | 147,126 |
| File Creates/sec | 157,753 | 120,276 | 119,888 | 85,217 |
| File Stats/sec | 707,471 | 561,325 | 570,214 | 349,901 |
| File Reads/sec | 247,182 | 233,629 | 205,334 | 120,455 |
| File Removes/sec | 144,569 | 135,103 | 124,497 | 113,482 |

# Data Migration Framework

## Short update

# POSSIBLE TIERED STORAGE SOLUTIONS (ON PREM OR OFF …)




# Existing (partial) solutions




## Across the core data mobility

Komprise – Analyze, mobilize and monetize file and object data. Via Subscription, managed through a global file system, the Komprise Cloud File System.

## Across the cloud data mobility

Aparavi - Identify, Classify, optimize and move unstructured data. Cloud-based user experience.

Spectra Vail – Multi-cloud data management, object-based global data store.

## Edge, Core, Cloud data mobility

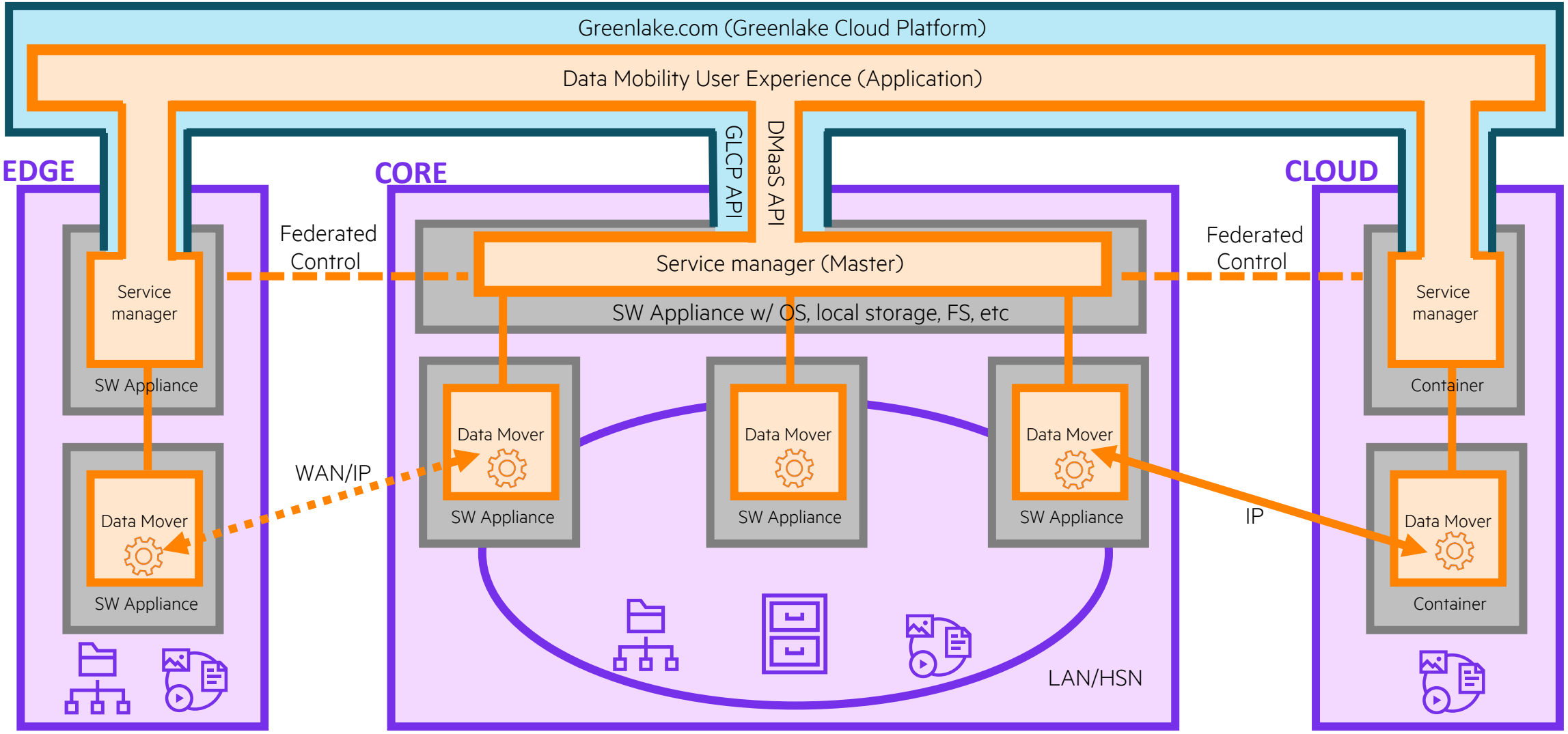
Cohesity – SW defined on virtual or physical, or as a Service in the cloud. Cloning for test/dev, snapshot integration with HPE arrays, NAS integration with SmartFiles.

Ctera – Global file system. Cloud-based SaaS distributed file storage solution incorporating unstructured data management. Store, access, share and protect files.

## Edge to core data mobility

Globus – SaaS, non-profit service. Enabled via the cloud, secure transfers for research data. Focus on collaboration across sites and institutions. Supports user definable workflows “Flows”

# Data Mobility as a Service concept



# System components

- User Application
  - Connected to all service managers via an API for the service
- Service Managers
  - Highly available, support continuous operations
  - Federated between Core and edge, core/edge to cloud
- Data Movers
  - Internal/external data movement could utilize compute nodes via SLURM or PBS Pro based controlled jobs
    - Alternative is using dedicated nodes for greater control.
  - Dedicated data movers for:
    - Indirect path on premises to accommodate network/security limitations
    - Edge-to-core for managing long latency, poor transmission quality
    - Edge/core to cloud for cloud bursting performance

# Required end points

- File
  - HSM enabled parallel file system
    - Lustre, Spectrum Scale
    - Includes tiering between Flash and HDD based components
  - Other POSIX compliant file systems
    - DAOS, CortX, BeeGFS, Ceph, etc.
  - NFS
  - SMB/CIFS
- Object
  - Vendor solutions
    - Scality, Ceph
  - Cloud Service Providers
    - Amazon, Microsoft, Google
- Device
  - Linear/Tape
    - LTO, IBM TS

# Reporting requirements

- Service
  - files moved
    - Success and Failure stats
  - bytes moved
  - jobs run
  - data rate
  - User count
- Storage
  - System utilization
  - Utilization trends
- Data
  - Characterization by size, age, owner
  - Copies
  - Storage space utilized
  - Grouping
  - Compliance hold
  - Classified

Data Mobility Service

https://128.162.236.238/dashboard?site_id=DEF456

BlueJay DMF HPE JIRA Cray JIRA HPE MFA SignUp

# Data Mobility Service

## Dashboard

Demo Site

### System Status

**Filesystems**

1
Total

**Movers**

5
Total

4
Healthy

1
UnHealthy

**Jobs**

5
Active

4
Pending

5
Completed

4
Failed

### Managed Data

**Analyzed:**
5 Filesystems


### Alerts

**Network Interface eth1 on group sjc-array2003 failed**
Hardware
2021-11-02 15:50

**File share rtp-prod-share-dev has reached 75% of the hard quota limit**
Capacity
2021-11-02 15:50

Preview

Metadata Summary

Putty R-Desktops Desktop

Data Mobility Service ... Data Mobility Service ... Data Mobility Service ... Data Mobility Service ... Data Mobility Service ... Data Mobility Service ...

A/C

11:48 AM
Wednesday
2/8/2023

# Open questions …

- Do we need “intelligent” tools or is brute force good enough ??
- Are Lustre and/or GPFS running out of steam in the next 5-7 years ??
  - If so, how do we handle the many EB of data and trillions of files ??
- Migrating data to new (and probably) larger file systems ?
  - On day 1, opportunistically or not at all ??
- Data migration tools ??
  - rsync (msrsync, Lustre rsync), PCP, Pftool, Shift-C, Mutils, psync, dsync, UFTP, BBCP etc ??
- Archiving futures?
  - ”Tape is dead”  (or is it ??)
  - Cloud based cold storage ??
  - Disk based systems (zero watt implementations) ??
- Where do we go from here ??

# Summary ??

- Lustre is alive and healthy at HPE …
- Most of the development efforts are focused on fixes and specific enhancements
  - mostly performance related but functionality is also important
- Thorough testing or every release is key
  - More than 3,000 tests in our repository
  - > 500 standard tests are run on any major release
    - Many Lustre features tested during repeated FOFB
  - In depth performance sweeps are performed to find regressions or other issues
- The entire ecosystem (not just Lustre) is equally important
  - Includes data management, archiving, security, etc.
  - ANY feature or function can (and will …) break at scale

- Despite rumours to the contrary, there's still a lot of life left in Lustre ….

# While claiming the honor …

- The cast of many:

John Fragalla, Bill Loewe, Sergey Shlepakov, Kris Woolsey, Michael Moore, Petros Koutoupis, Brent Petit, Dipak Ghosh, Andreas Müller, Mark Wiertala, Cory Spitz, ……

Apologies to the ones I forgot …..


tkp@hpe.com