HPE
Data Mobility in the HPC World

Torben Kling Petersen, PhD

Distinguished Technologist

Data Mobility – Setting the stage

What it ISN'T …

- A well-defined product
- A deliverable with a plan of record
- Something being actively developed

What it is …

- Data Mobility is a concept with a goal to create a reference architecture for data movement, data insight and long-term data curation
- A mapping of software components available internally or externally
- Customer information gathering

Data Mobility – a definition ..

Movement

- Policy based data migration between file systems, sites or systems
- Scalable architectur to handleExaBytes of data and billions of files

Transformation

- Data modicifations as part of workflows
- Changing file formats and packaging
- Compression, containerisation, de-duplication

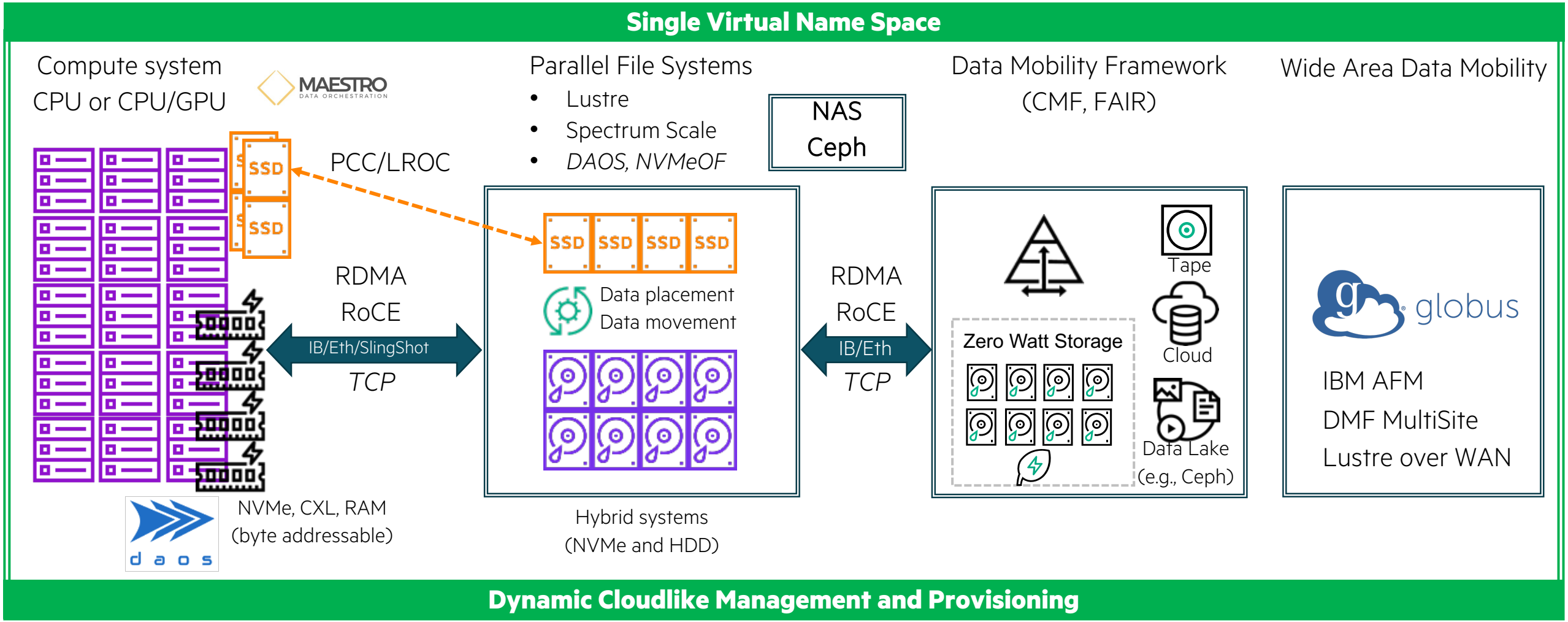
Tiering

- Data movement within a single name space (e.g. CDS/DLM/etc.)
- Just in time data availability from ANY tier
- Support for complex workflows including data gathering from any data storage system

Archiving

- HSM
- Incremental backups
- Multiple backends including off-site locations

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 Key Services

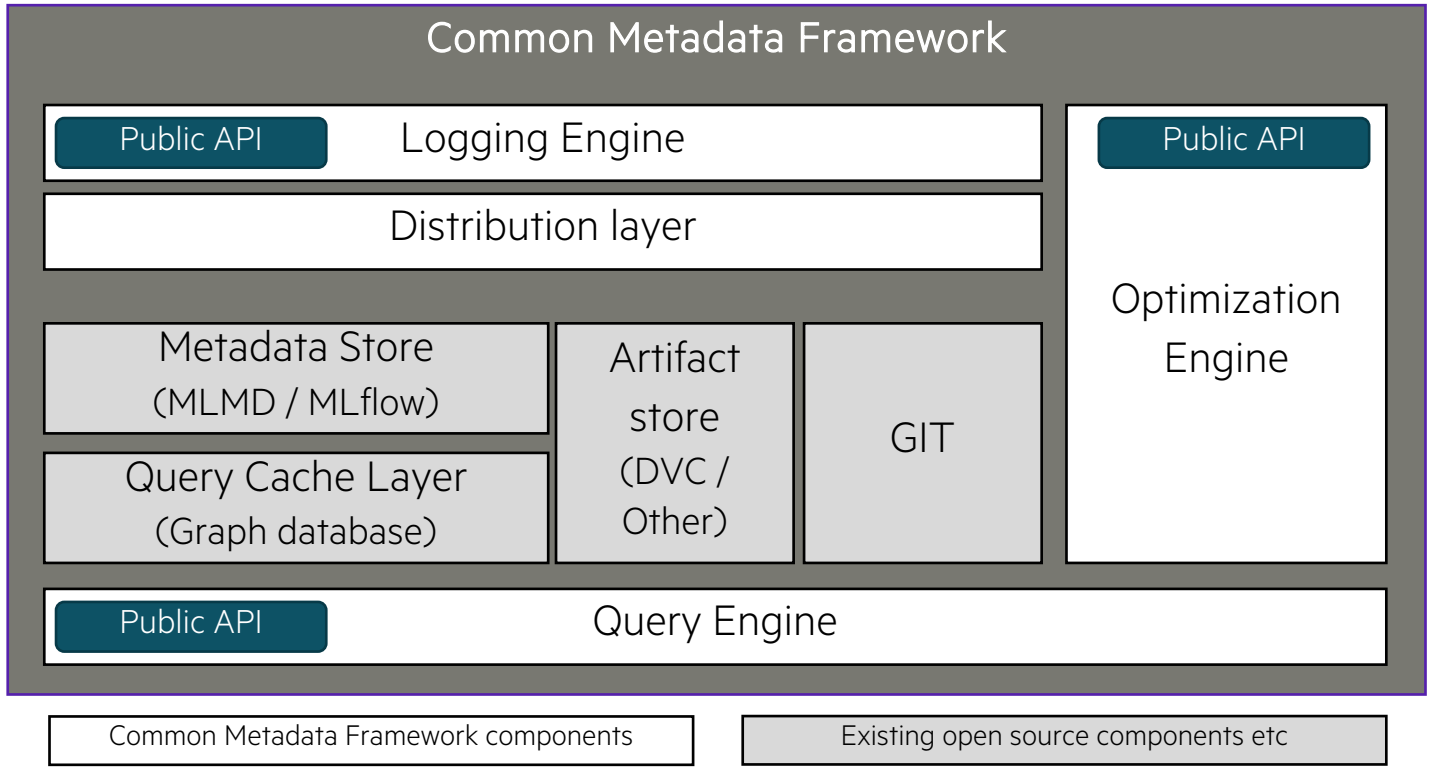
Key Components

- Metadata
  - Centralized
  - Extended (manually, machine assisted, AI …..)
  - Comprehensive – map and consolidate all monitored name spaces
  - User (including end user) and API accessible
  - Federated with locality index
- Data movement
  - Includes Tiering, copying, and archiving
  - Dynamic and scalable
  - Parallel where possible
  - Controlled (guaranteed movement of file) and using checksums
  - Interruptible with automatic clean-up
  - Secure (possibly encrypted data in flight)
- Monitoring
  - Done, planned and inflight
  - Statistics
  - Consolidated reporting
- Security
  - Auditing
  - Encrypting
  - Secure delete

Common Metadata Framework (CMF) architecture

- Tracking Data, Metadata, Code
  - Also tracking versioned data
- GIT like experience for metadata
- Global View of all metadata
- Extensible and modular

- CMF components:
  - Metadata library/query engine
  - CMF Local Client – synchronizing metadata
  - CMF central server
    – single instance or distributed
  - Optimized central repository for code, metadata and data



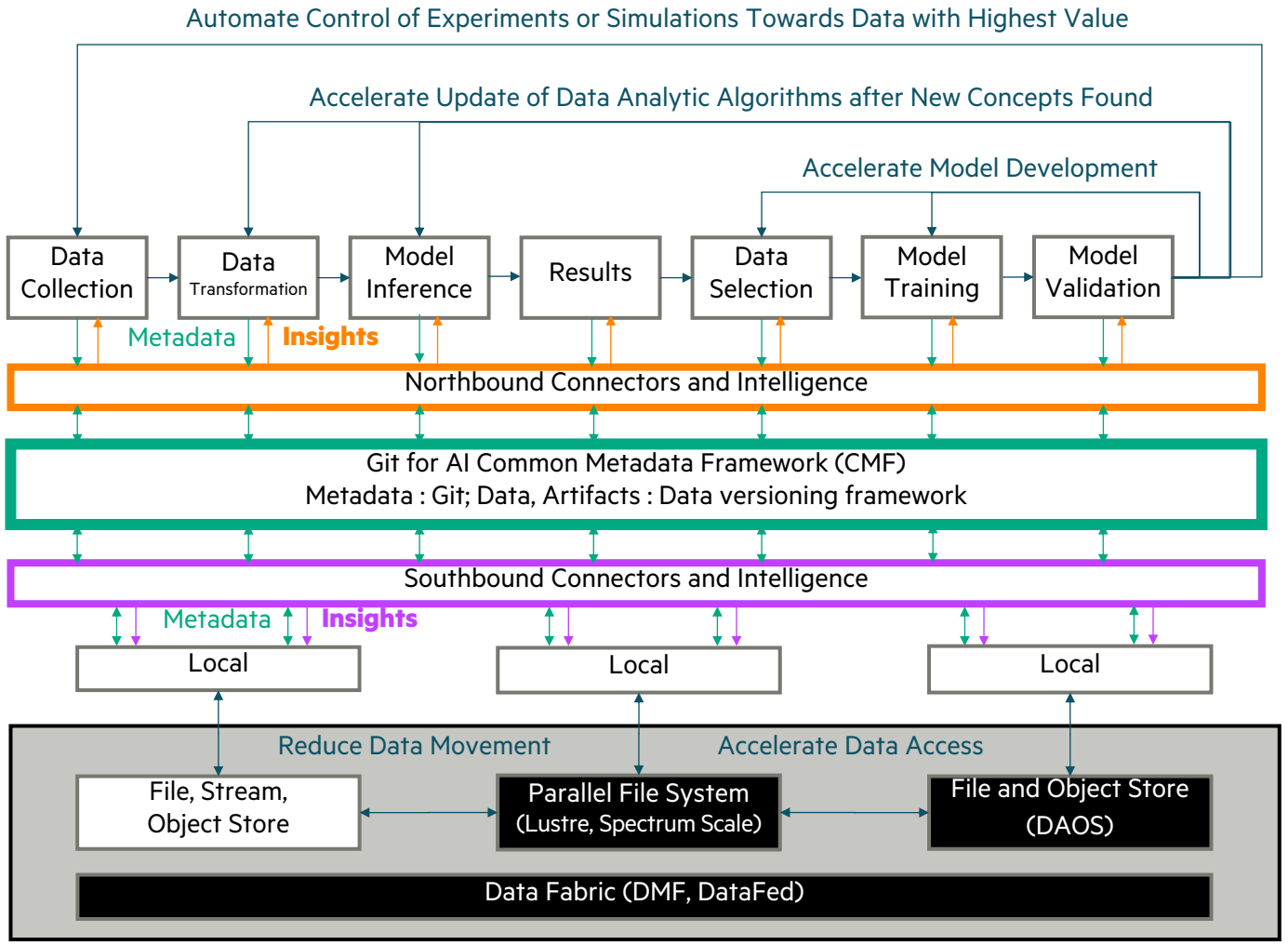
https://github.com/HewlettPackard/cmf

AI and ML workflows

- Model optimization intelligence
  - Parameter recommendation
  - Model recommendation
- Data centric intelligence
  - Data selection, gradation, caching
  - Data search, processing, augmentation
- Computational steering
  - Integration with HPC simulation environment
  - Looking for collaboration opportunities
- Experiment steering

- Does this data flow and data movement make sense?

Data Foundation benefits for an Example Science Workflow



The Metadata challenge - Example

- Using Semantic Data Management
  - Hierarchical data locality -> metadata identification.
  - E.g.,

```
# fs:     ../12345/20230509/1200/24/0/T/...
```

  - Using an API to ”translate” locality to unique metadata information.

metadata header: date: 2023-05-09, location: 12345, time: 12.00, step: 24, parameter: T, level: 0
Data payload: • Field data [array of doubles] • GRIB2 • Observations [numerical and non-numerical] • Grid partitions

  - ECMWF have been using this since 1975 and has mor than 400 PiB stored in this format.

Data Mobility “Find” example*

```
# dms_find -global -u torben -n “important_file” -v
fs          path                                        version   tier     state
[lustre]    /fs1:/lustre/tkp/important_file             4         flash    cur
[lustre]    /fs1:/lustre/tkp/arch/important_file        1         hdd      arch
[scale]     /fs4:/root/tkp/imporant_file                3         hdd      bak
[daos]      /ds1:/ds1/team_x/important_file             4         pmem     cur
[dmf7]      /zws:/objectID=112299                       1         hdd      arch
[dmf7]      /tpl1:/index:123456.32                      1         tape     compr
[cloud]     /url:aws.com/tkp.63/arch/important_file     2         cloud    arch
#
```

BUT how do you list or search a Billion files ??

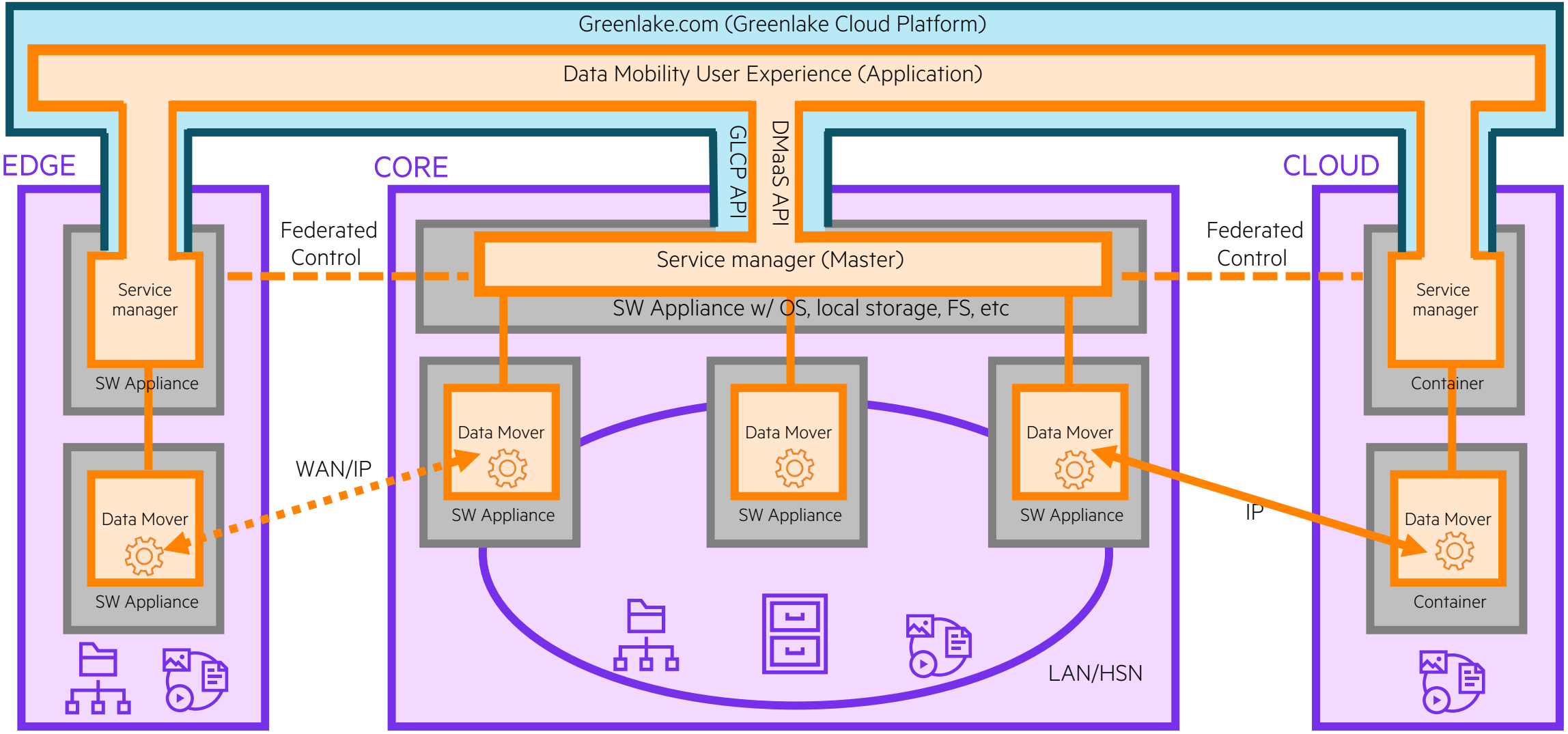
* Artist vew of possible output …

Existing extended metadata solutions

- iRods
  - Open Source !!
  - Parallel file system aware
  - Scalable extensible metadata
  - Mature (been around >15 years)
  - Scalable data movement capabilities
- StarFish
  - HPC and data centric computation focused
  - Parallel, multiprotocol data movement
- MediaFlux
  - Scalable to billions of assets (custom database: XODB), support for 100+ PB…
  - Extensible metadata
  - Multi protocol support
- Nodeum
  - Based in EMEA
  - Some HPC focus (collaborations with Juelich, BCS, CSCS, Fenix RI project etc)

Initial concepts

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 ??

tkp@hpe.com