Technology Preview:

Intel DAOS

Distributed Asynchronous Object Storage

Dell EMC HPC & AI Innovation Lab - www.hpcatdell.com

Austin, TX

DELL Technologies

What is DAOS?

- **D**istributed **A**synchronous **O**bject **S**torage
- A new, innovative distributed parallel file system based on Intel Optane Persistent Memory (SCM/PMem) and NVMe SSDs
- Coordinates parallel IO across many nodes presented to the user as a single filesystem
- Delivers exceptionally high bandwidth and IOPS on commodity servers
- Can be utilized either as a standalone file system, or as a performance tier integrated with existing storage systems
- Runs in user space via development kits
- Persistent byte-level access to handle more granular transactional IO (4k, random)

Why evaluate it?

- Growing curiosity
- Newer technologies and workflows
- Explore storage system possibilities for exascale systems with newer workloads and technologies

High Value Use Cases

Artificial Intelligence

- AI workloads perform large volumes of reads - data access time becomes critical
- Native AI framework support (Apache Spark) enables AI workloads

High Performance Data Analytics (HPDA)

- HPDA generates large volumes of small random reads/writes
- DAOS provides new rich storage API with native support for unstructured and semi-structured data
- DAOS stores small writes and metadata into byte-granular persistent memory, removing performance bottlenecks & unleashing higher performance

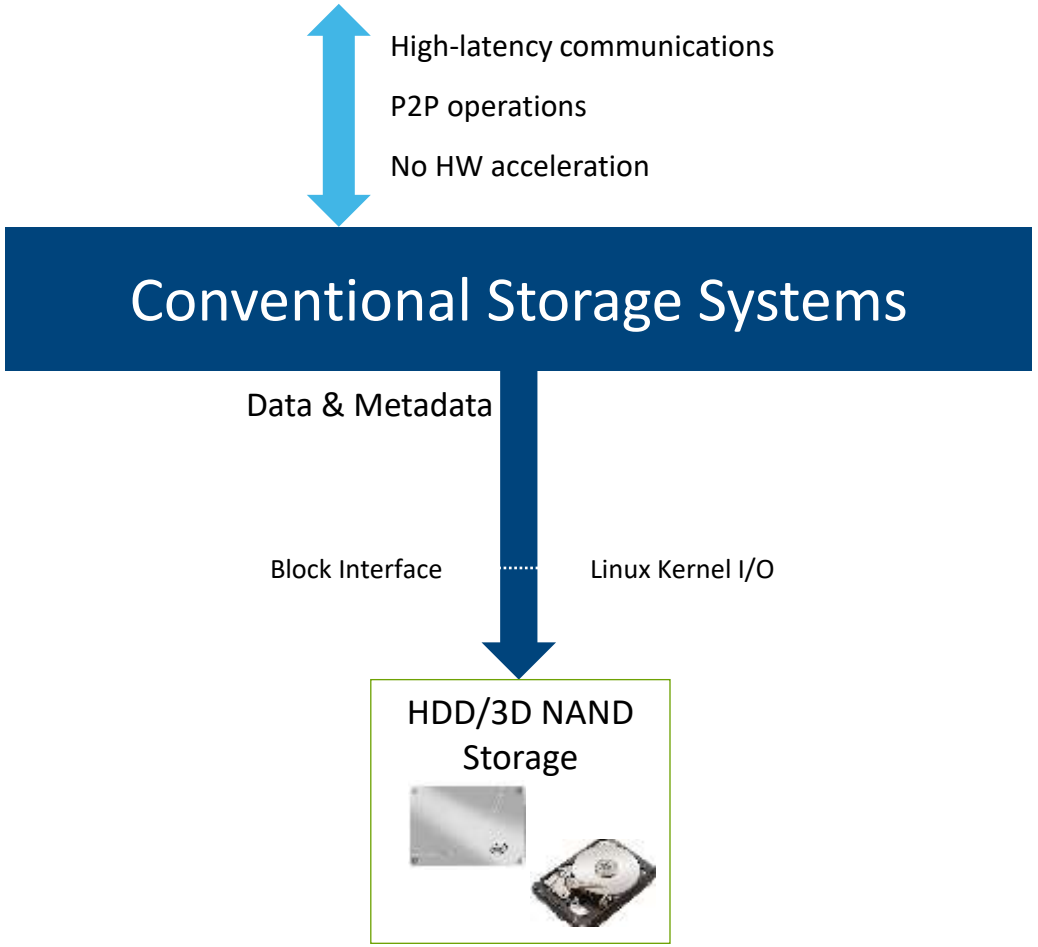
Traditional Modelling & Simulation

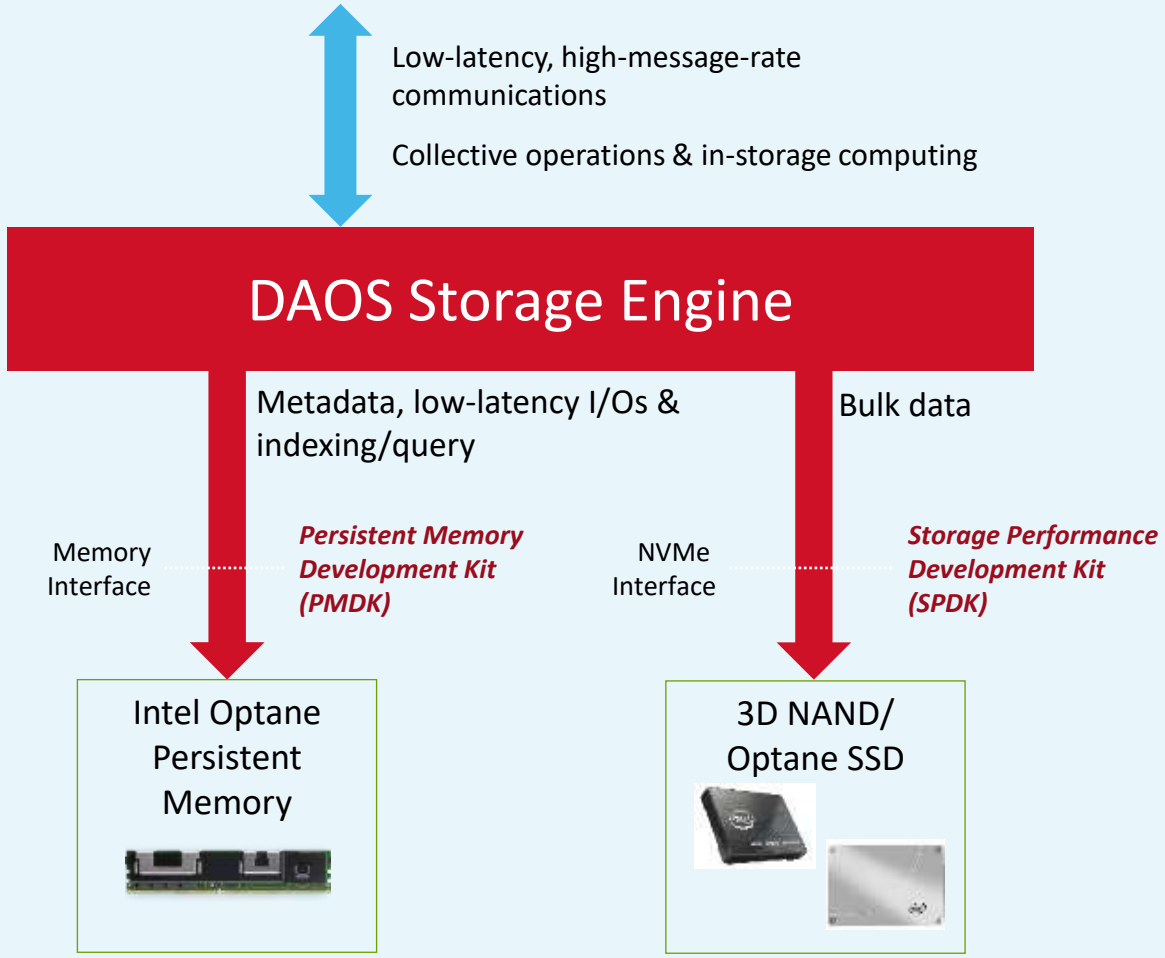
- Workloads with small file I/O, large IOPs, or low-latency requirements struggle under existing PFS that are optimized for large streaming reads and writes
- Traditional HPC centers aligning on IO-500 benchmarks are placing more and more requirements in RFPs on unaligned I/Os that can't be addressed by Spectrum Scale or Lustre
- Direct integration of the domain-specific data models (e.g. oil & gas, meteorology, animation…) over the DAOS API to accelerate applications

Convergence of AI, HPDA, and traditional HPC

- Need a storage system that can simultaneously support next gen workflows, where the different workflows can exchange data and communicate at high levels of performance

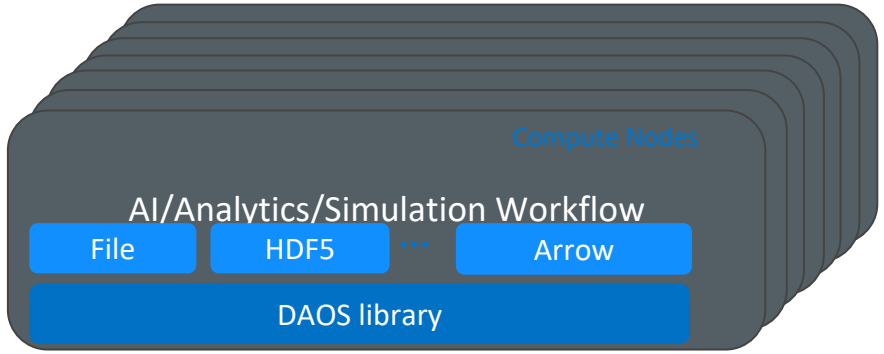
DAOS Architecture

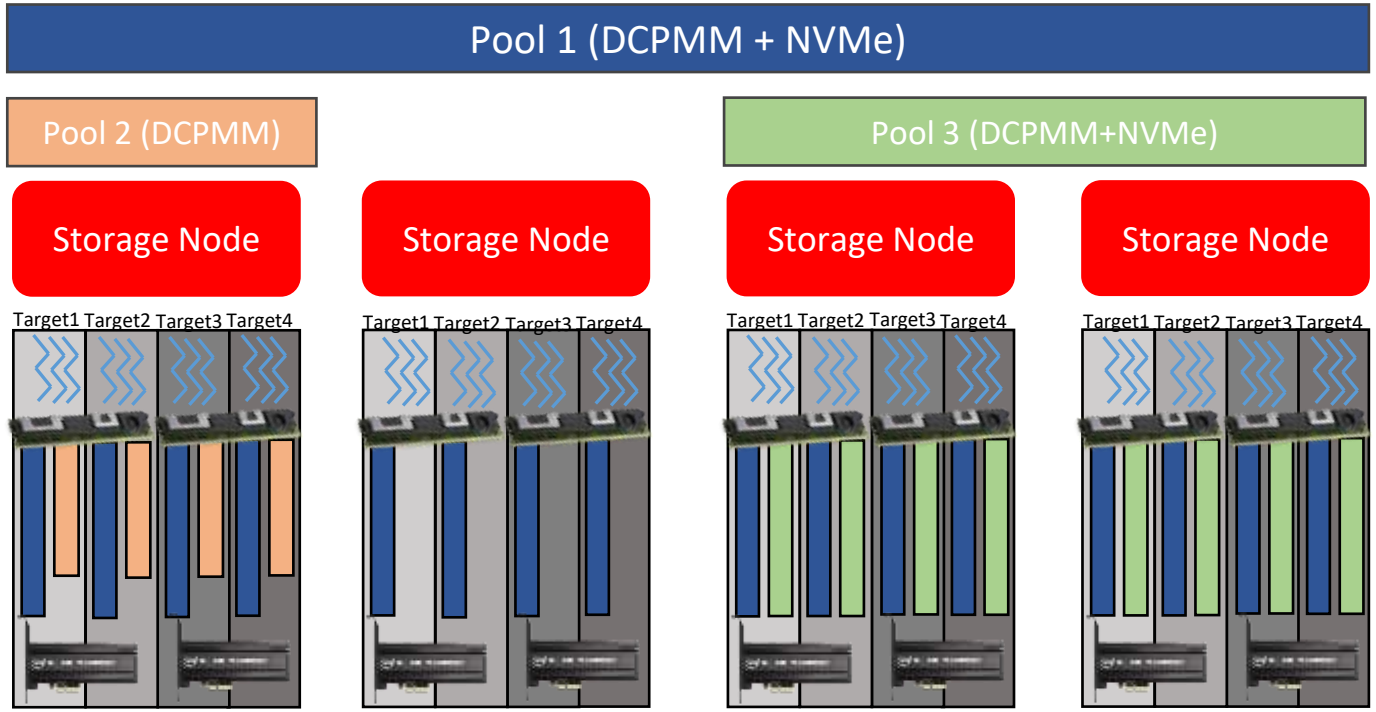




DAOS Pool

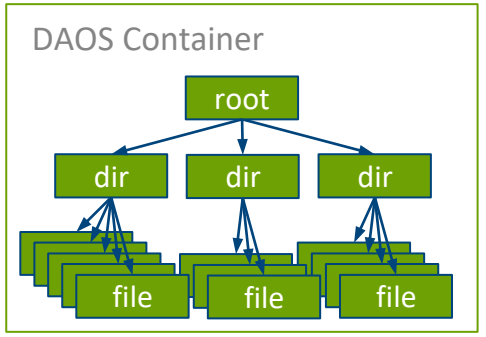
- Provides storage virtualization
- Distributed storage reservation
  - Persistent memory / DCPMM
  - NVMe SSD
- Predicatable capacity
  - Can be resized
  - Can be extended (i.e. span more servers)
- Multi-tenancy
  - NFSv4-type ACLs
- Typically 1 pool = 1 project
  - Can have a single pool or 100’s



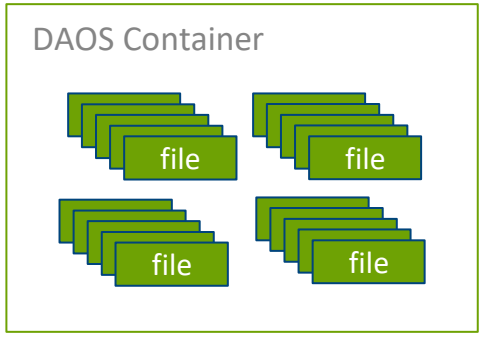


Datasets: DAOS container

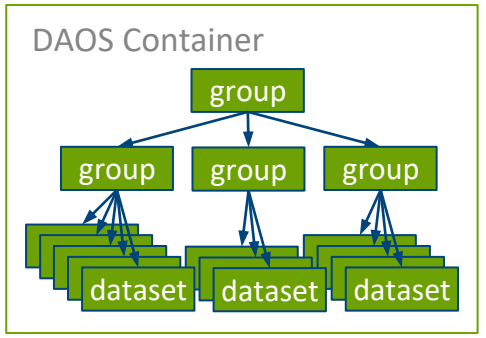
- Manageable and coherent entities
  - Stored in a pool
  - Simplified data management
    - Cross-tier migration
    - Query capability to identify recently accessed containers
    - Container indexing
  - Snapshot and rollback support
  - Built-in producer/consumer workflow pipeline support
  - NFSv4-type ACLs



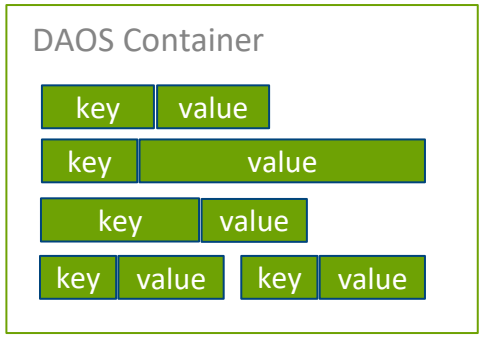
Encapsulated POSIX Namespace



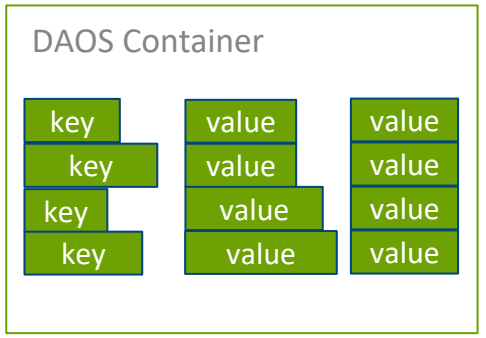
File-per-Process



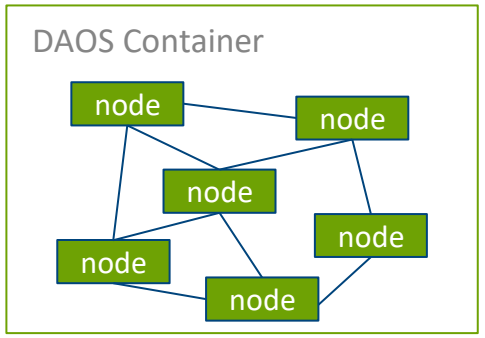
HDF5 « File »



Key-value Store

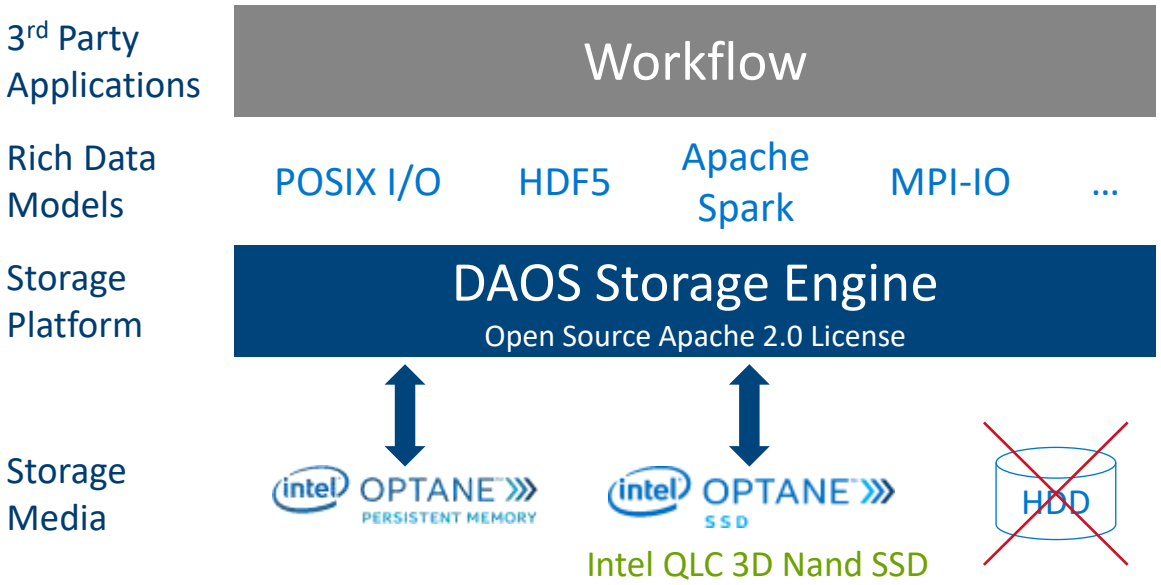


Columnar Database

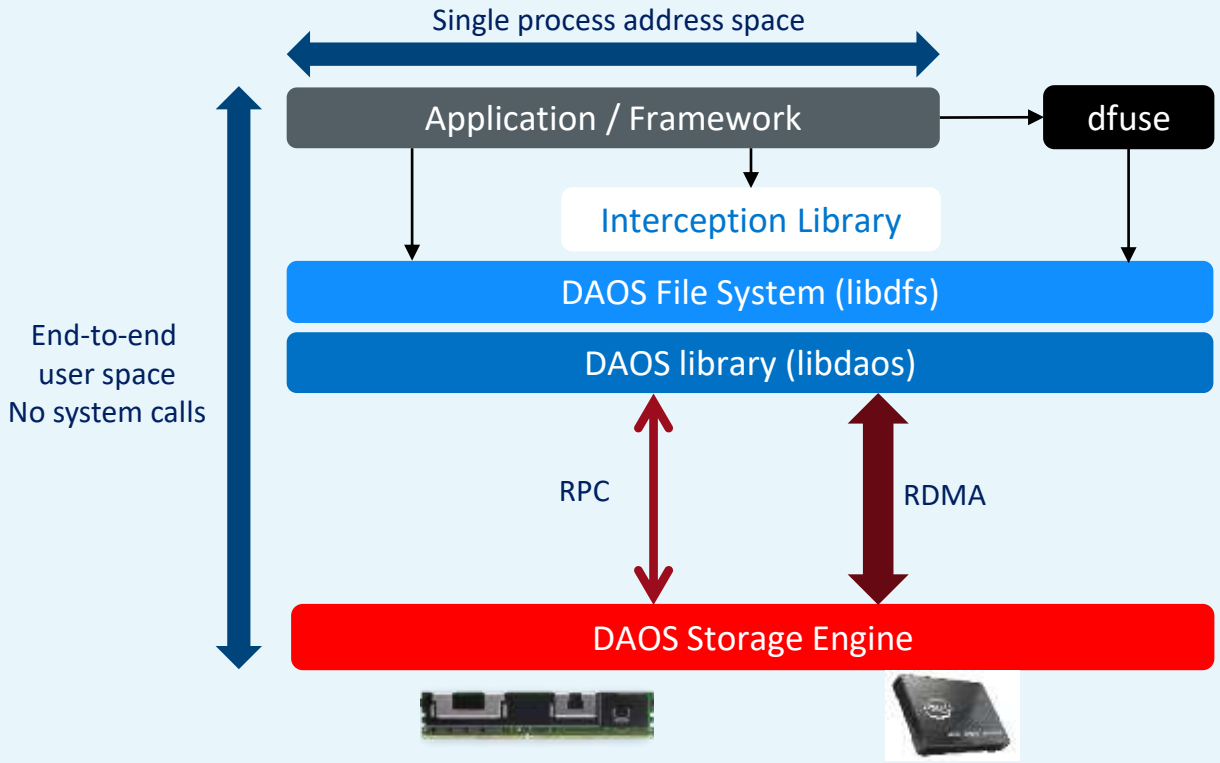


Graph

DAOS Stack/Middleware



POSIX I/O Support



Middleware Code

- **DAOS Containers**

```
$ daos container create --path=/mnt/project1/userA/NS1 --pool=uuid --type=POSIX/HDF5/etc
$ dfuse --pool pool_uuid --container cont_uuid -m /tmp/daos
```

- **DFS (DAOS File System) API Library**

- I/O Interception Library (no code changes) (2x Boost)

```
$ export D_LOG_MASK=ERR
$ export D_LOG_FILE=/home/daosadmin/ERR-run-ior.log
$ export LD_PRELOAD=/path/to/daos/install/lib/libioil.so
```

```
Original:
fd = open(file_name, O_CREAT|O_RDWR, 0600);
/** set up iov */
pwritev(fd, iov, 1, offset);
preadv(fd, iov, 1, offset);
close(fd);

DFS Change:
dfs_open(dfs, NULL, file_name, 0600, O_CREAT|O_RDWR, 0, 0, NULL, &file);
/** setup sgl (DAOS iov) */
dfs_write(dfs, file, &sgl, offset, NULL);
dfs_read(dfs, file, &sgl, offset, &bytes_read, NULL);
dfs_release(file);
```

- **HDF5 Vol**

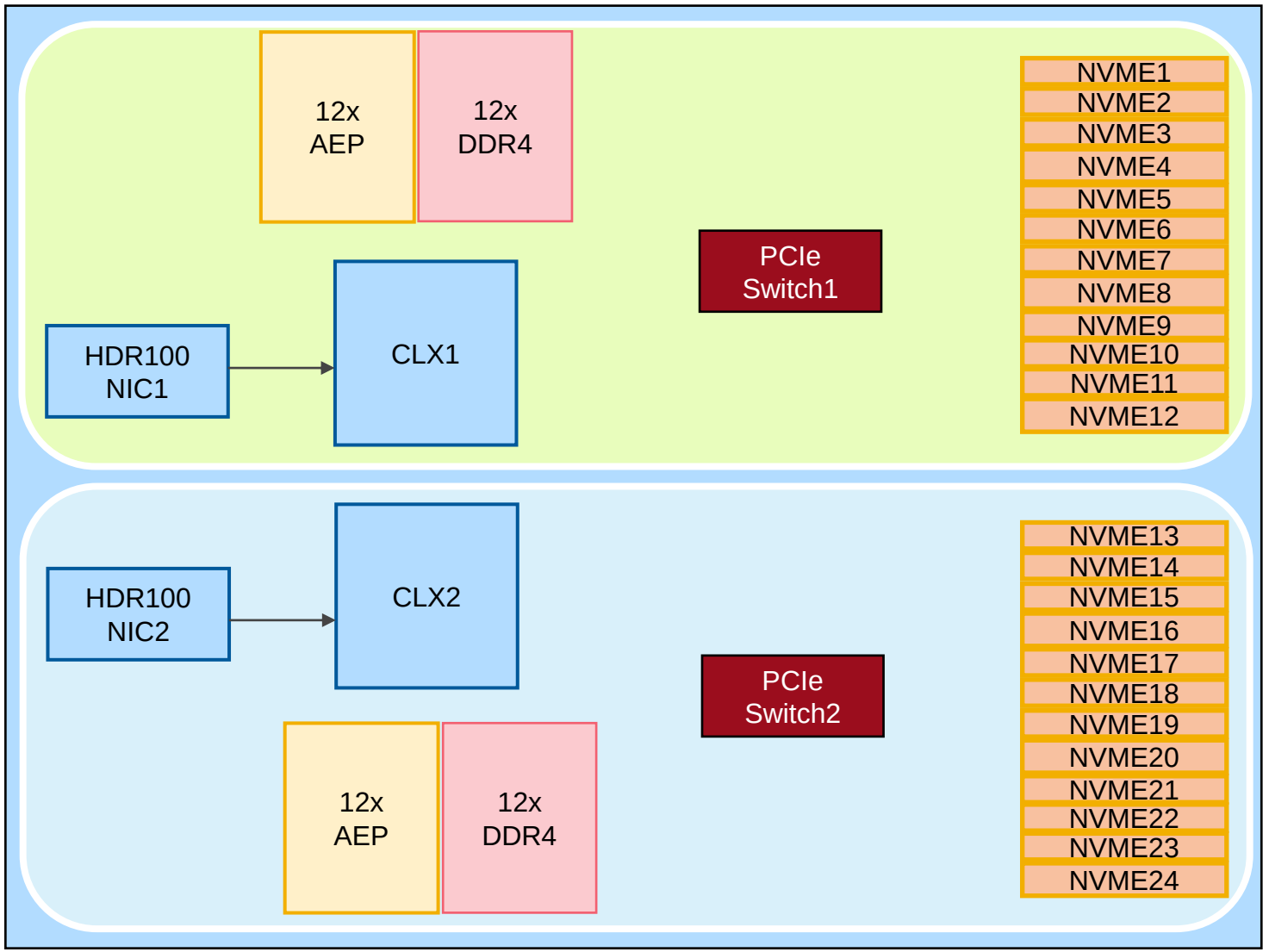
Dynamically loaded plugin (No change to user application):
How to set/get pool & container uuid then?
1. Use env variables (User passes those).
2. Unified Namespace with special file storing pool/container as extended attributes

```
export HDF5_PLUGIN_PATH=/path/daos-vol/lib/
export HDF5_VOL_CONNECTOR=daos
```

- **MPI-IO**

Application works seamlessly by just specifying the use of the driver by appending ***daos:*** to the path.

DAOS v1.1.1 Preview - PowerEdge 740xd



Components:

- 2x 6248R (CLX)
  - 24+ Core/Lower Freq
- 12x 16GB RDIMM
  - 2900 MT/s
- 12x Intel Optane100 (AEP)
  - 128 GB Pmem/SCM
- 24x NVMe
  - Intel P4610 SSDs (Gen3)
- 2x HDR100 (200Gb/s)

Evaluation Phase DAOS v1.1.1

Software Installation & DAOS Setup @ www.daos.io

- The DAOS code is hosted on GitHub and can be compiled from Source or downloaded as prebuilt RPMs.
  - **Software Install**: https://daos-stack.github.io/admin/installation/
  - **System Deployment**: https://daos-stack.github.io/admin/deployment/

- Don't have PMem/SCM or SSDs?
  - Try ***tmpfs*** on available ram

```
$ free -g
$ mount -t tmpfs -o size=64G tmpfs /mnt/daos/tmpfs
$ df -h /mnt/daos/tmpfs/
Filesystem      Size  Used Avail Use% Mounted on
tmpfs            64G     0   64G   0% /mnt/daos/tmpfs
```

```
daos_server.yml
 scm_mount: /mnt/daos/tmpfs
 scm_class: ram
 scm_size: 64
```

- Device NUMA Bindings

```
$ grep -E '.*' /sys/class/net/ib*/device/numa_node
/sys/class/net/ib0/device/numa_node:0
/sys/class/net/ib1/device/numa_node:3

$ for dev in $(lspci -nn | grep -i nvme | awk '{print $1}'); do echo -n " $dev on numa "; cat
/sys/bus/pci/devices/0000\:${dev}/numa_node; done
 68:00.0 on numa 2
 69:00.0 on numa 2
 b7:00.0 on numa 1
 b8:00.0 on numa 1
```

- **Tips:**
  - Check logs files and debug levels
  - Use log level DEBUG to check for problems
  - Try insecure mode at first (avoid debugging certs)
  - Confirm device-numa bindings for IB & NVMe
  - Test basic commands as root
    - dmg query
    - dmg pool get-acl
    - daos help *command*
  - ***Online documentation is updated frequently***

DAOS Engine Layout

daos_server.yml

```
access_points: ['10.140.0.7']
port: 10001
control_log_mask: ERROR # DEBUG|ERROR|INFO
transport_config:
 allow_insecure: true

servers:
-
 targets: 6
 pinned_numa_node: 0
 fabric_iface: ib0
 fabric_iface_port: 31316
 scm_mount: /mnt/daos/pmem0
 scm_class: dcpm
 scm_list: [/dev/pmem0]
 bdev_class: nvme
 bdev_list: ["0000:62:00.0","0000:63:00.0","0000:64:00.0","0000:65:00.0","0000:66:00.0","0000:67:00.0"]
-
 targets: 6
 pinned_numa_node: 3
 fabric_iface: ib1
 fabric_iface_port: 31417
 scm_mount: /mnt/daos/pmem1
 scm_class: dcpm
 scm_list: [/dev/pmem1]
 bdev_class: nvme
 bdev_list: ["0000:b3:00.0","0000:b4:00.0","0000:b5:00.0","0000:b6:00.0","0000:b7:00.0","0000:b8:00.0"]
```

```
$ sysctl -w net.ipv4.conf.{ib0,ib1,eth0,all}.accept_local=1
$ sysctl -w net.ipv4.conf.{ib0,ib1}.arp_ignore=2
$ sysctl -w net.ipv4.conf.{ib0,ib1}.rp_filter=2
$ sysctl -w net.ipv4.conf.{ib0,ib1,all}.arp_announce=2
```

```
$ dmg storage scan -verbose
----------
10.140.0.7
----------
SCM Namespace Socket ID Capacity
------------- --------- --------
pmem0         0         799 GB
pmem1         1         799 GB

NVMe PCI     Model                FW Revision Socket ID Capacity
--------     -----                ----------- --------- --------
0000:62:00.0 Dell Express Flash N VDV1DP21    2         1.6 TB
0000:63:00.0 Dell Express Flash N VDV1DP21    2         1.6 TB
0000:64:00.0 Dell Express Flash N VDV1DP21    2         1.6 TB
0000:65:00.0 Dell Express Flash N VDV1DP21    2         1.6 TB
0000:66:00.0 Dell Express Flash N VDV1DP21    2         1.6 TB
0000:67:00.0 Dell Express Flash N VDV1DP21    2         1.6 TB
0000:68:00.0 Dell Express Flash N VDV1DP21    2         1.6 TB
0000:69:00.0 Dell Express Flash N VDV1DP21    2         1.6 TB
0000:6a:00.0 Dell Express Flash N VDV1DP21    2         1.6 TB
0000:6b:00.0 Dell Express Flash N VDV1DP21    2         1.6 TB
0000:6c:00.0 Dell Express Flash N VDV1DP21    2         1.6 TB
0000:6d:00.0 Dell Express Flash N VDV1DP21    2         1.6 TB
0000:b3:00.0 Dell Express Flash N VDV1DP21    1         1.6 TB
0000:b4:00.0 Dell Express Flash N VDV1DP21    1         1.6 TB
0000:b5:00.0 Dell Express Flash N VDV1DP21    1         1.6 TB
0000:b6:00.0 Dell Express Flash N VDV1DP21    1         1.6 TB
0000:b7:00.0 Dell Express Flash N VDV1DP21    1         1.6 TB
0000:b8:00.0 Dell Express Flash N VDV1DP21    1         1.6 TB
0000:b9:00.0 Dell Express Flash N VDV1DP21    1         1.6 TB
0000:ba:00.0 Dell Express Flash N VDV1DP21    1         1.6 TB
0000:bb:00.0 Dell Express Flash N VDV1DP21    1         1.6 TB
0000:bc:00.0 Dell Express Flash N VDV1DP21    1         1.6 TB
0000:bd:00.0 Dell Express Flash N VDV1DP21    1         1.6 TB
0000:be:00.0 Dell Express Flash N VDV1DP21    1         1.6 TB
```

DAOS 101

Dual I/O Engines:

```
[daosadmin@node015 ~]$ dmg system query --verbose
Rank UUID                                 Control Address  State  Reason
---- ----                                 ---------------  -----  ------
0    9e7ac125-326d-41b8-92df-67ea3fca63fe 10.140.0.7:10001 Joined
1    88683034-3a4d-4340-a10f-f64fb415ed0a 10.140.0.7:10001 Joined
```

```
[daosadmin@node015 ~]$ dmg pool query --pool
Pool space info:
- Target(VOS) count:12
- SCM:
  Total size: 400 GB
  Free: 400 GB, min:33 GB, max:33 GB, mean:33 GB
- NVMe:
  Total size: 16 TB
  Free: 16 TB, min:1.3 TB, max:1.3 TB, mean:1.3 TB
```

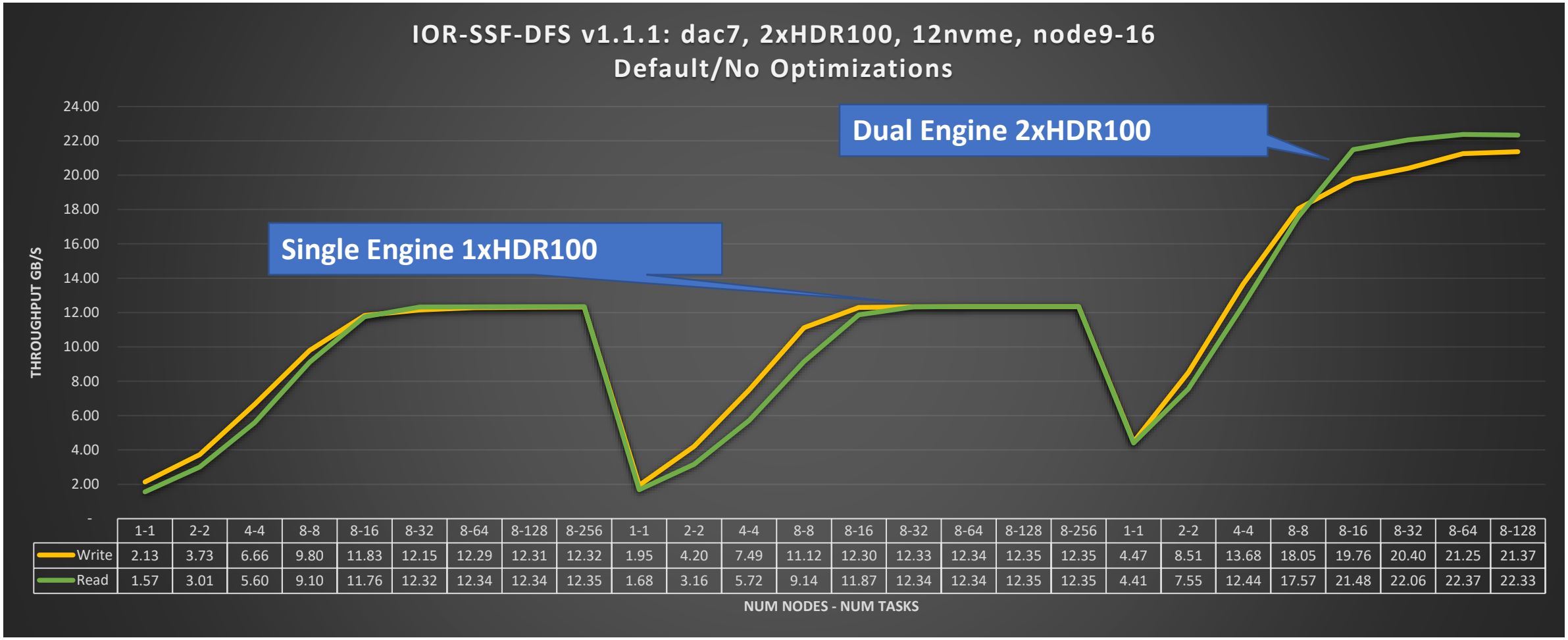
```
[daosadmin@node015 ~]$ dmg pool create --scm-size=200G --nvme-size=8000G
Creating DAOS pool with 200 GB SCM and 8.0 TB NVMe storage (2.50 % ratio)
Pool-create command SUCCEEDED: UUID: 8818aac1-fd48-431b-ade1-063cec014b91, Service replicas: 0

[daosadmin@node015 ~]$ dmg pool get-acl --pool=8818aac1-fd48-431b-ade1-063cec014b91
# Owner: daosadmin@
# Owner Group: daosadmin@
# Entries:
A::OWNER@:rw
A:G:GROUP@:rw

[daosadmin@node015 ~]$  daos container create --svc=0 --type=POSIX --chunk_size=4K --pool=8818aac1-fd48-431b-ade1-063cec014b91
Successfully created container b9735b76-90a7-43a3-892d-54403628a7f7

[daosadmin@node015 ~]$ dfuse -s 0 --pool=8818aac1-fd48-431b-ade1-063cec014b91 --cont=b9735b76-90a7-43a3-892d-54403628a7f7 --
mountpoint=/home/daosadmin/dfuse

[daosadmin@node015 ~]$ df -h /home/daosadmin/dfuse/
Filesystem      Size  Used Avail Use% Mounted on
dfuse            15T  616K   15T   1% /home/daosadmin/dfuse

[daosadmin@node015 ~]$ cd /home/daosadmin/dfuse/
[daosadmin@node015 dfuse]$ echo "Hello DAOS" > /home/daosadmin/dfuse/testfile.txt
[daosadmin@node015 dfuse]$ cat /home/daosadmin/dfuse/testfile.txt
Hello DAOS
[daosadmin@node015 dfuse]$ ls -la /home/daosadmin/dfuse/testfile.txt
-rw-rw-r-- 1 daosadmin daosadmin 11 May 14 07:18 /home/daosadmin/dfuse/testfile.txt

[daosadmin@node015 ~]$ fusermount3 -u /home/daosadmin/dfuse/
```

Considerations:

- Size is per engine
  - 2x in this case
- Let user create the containers
- Multi-Rail on same subnet

Confirm DAOS Engine Config with IOR

ior **-a DFS -o /iortest1** --dfs.svcl 0 --dfs.pool ${pool} --dfs.cont ${cont} --w -r -b 1g -t 1m
ior **-a POSIX -o /home/daosadmin/dfuse/ior-dfuse** --dfs.svcl 0 --dfs.pool ${pool} --dfs.cont ${cont} -w -r -b 4g -t 1m (only provided as example)



	1-1	2-2	4-4	8-8	8-16	8-32	8-64	8-128	8-256	1-1	2-2	4-4	8-8	8-16	8-32	8-64	8-128	8-256	1-1	2-2	4-4	8-8	8-16	8-32	8-64	8-128
Write	2.13	3.73	6.66	9.80	11.83	12.15	12.29	12.31	12.32	1.95	4.20	7.49	11.12	12.30	12.33	12.34	12.35	12.35	4.47	8.51	13.68	18.05	19.76	20.40	21.25	21.37
Read	1.57	3.01	5.60	9.10	11.76	12.32	12.34	12.34	12.35	1.68	3.16	5.72	9.14	11.87	12.34	12.34	12.35	12.35	4.41	7.55	12.44	17.57	21.48	22.06	22.37	22.33

- Instructions to build IOR with DFS libraries:
  - https://github.com/hpc/ior/blob/main/README_DAOS

What's Next?

- Continue research phase with 15th generation hardware
  - PowerEdge R750 with next gen Xeon
  - Optane 200 Series
  - P5500 NVMe
  - Future release of DAOS v2.0
    - Erasure Coding
- Evaluate and optimize performance
- Expand access to internal application engineers for Data Analytics, ML, AI
- Discover and explore any Pros or Cons

Source code on GitHub

- https://github.com/daos-stack/daos

Admin Guide

- https://daos-stack.github.io/

Community mailing list on Groups.io

- daos@daos.groups.io

Slack

https://daos-stack.slack.com/

Support

- https://jira.hpdd.intel.com

How to Install

- https://daos-stack.github.io/admin/deployment/

Questions?

www.hpcatdell.com

Find us by searching: **“Dell HPC and AI Innovation Lab”**