Hewlett Packard Enterprise

SC22 Dallas, TX | hpc accelerates.

Update On HPE's Progress With DISTRIBUTED ASYNCHRONOUS OBJECT STORAGE

Lance Evans

Storage Architect
HPC Chief Technology Office
2022-11-14

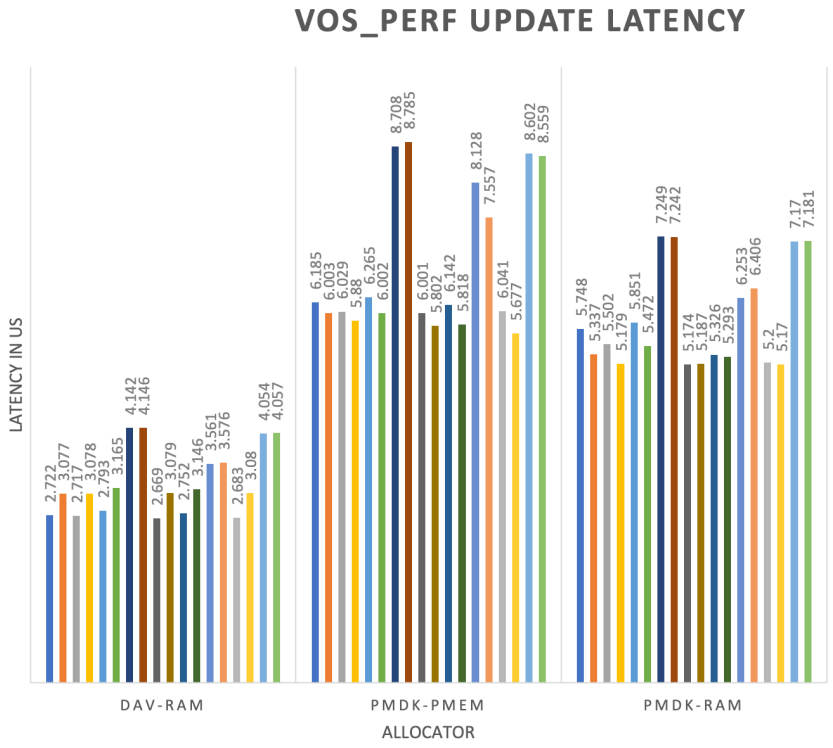
HPE Investment

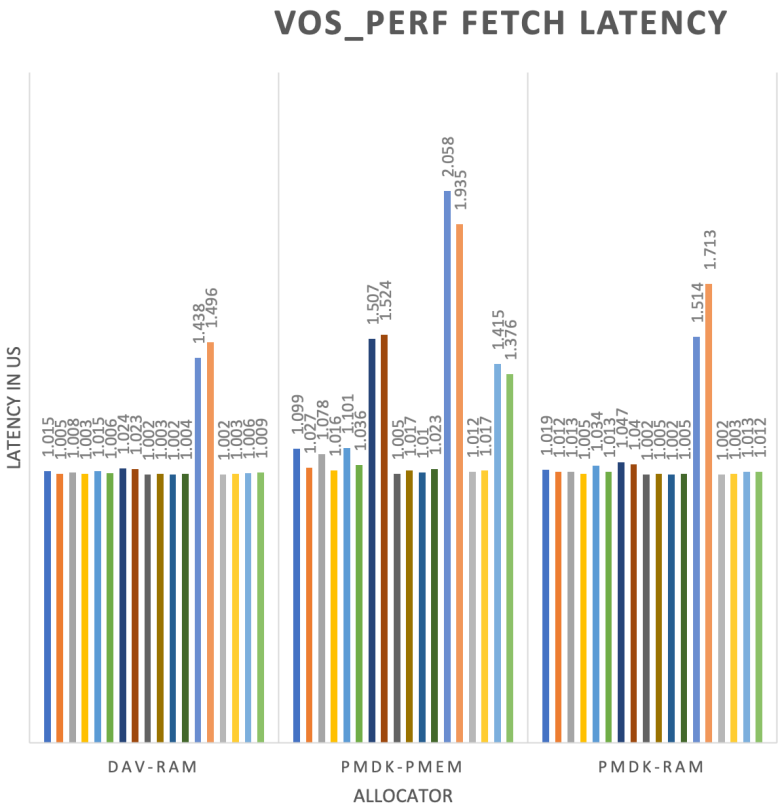
- Chief Technology Office
  - 4 people now allocated to the project
  - Testbed with 20 DAOS server nodes, variety of clients
  - Hosting external and internal POCs
  - Raising DAOS awareness within HPE
- HPC Storage R&D
  - Beginning technology transfer activity
  - Product architecture underway
  - Test engineering engaged
- AI/ML Org
  - Added a DAOS cluster to our GPU testbed
  - Hosting AI/ML POCs

HPE's DAOS Community Contributions (so far)

Metadata on SSD:	Decouples DAOS from a requirement for persistent memory hardware Contributions to alternative memory allocator and metadata on NVMe
Client-Side Metrics:	Enables application profiling for optimization of IO over DAOS Added DAOS client metric interceptor library exposing counters/stats
Tensorflow IO:	Enables Tensorflow to efficiently interact with DAOS Bug fixes, optimizations, and enabling dynamically loaded DAOS libraries
PyDAOS Tuning:	Exposes DAOS objects as native pythonic data structures Use a single event queue to avoid setup/teardown for every get/put call
Ray Plasma:	Enables distributed in-memory object store to spill data over into DAOS Python plugin for apps using smart_open to access DAOS key-value I/F
YCSB Plugin:	Yahoo Cloud Serving Benchmark enables comparison of various K/V DBs Added support for the native DAOS K/V API to YCSB

Metadata On SSD: DAOS VOS Allocator (DAV) Initial Latency Microbenchmark

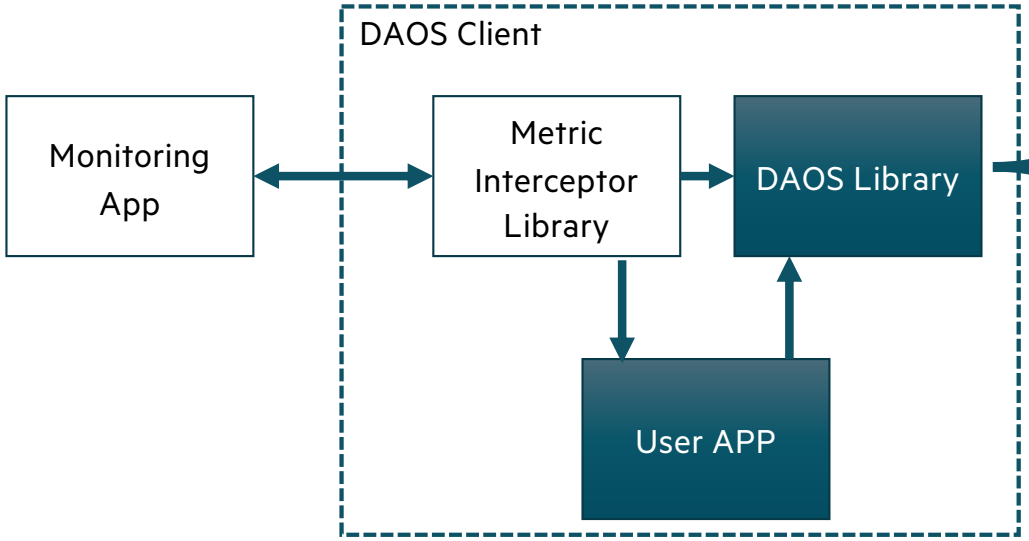




SC22
Dallas, TX | hpc accelerates.

Daos client-side metrics concepts

Potential Collection Approach



- **Counters**
  - **Various RPC calls**
  - **Successes, Failures, In-Flight**
- **Statistics**
  - **Fetches and Updates**
  - **Count, total size, avg size, deviation etc.**
- **Distributions (Histogram)**
  - **Statistics for Several Size Ranges**
  - **Statistics for Protection types**
- Code
  - https://github.com/daos-stack/daos/pull/6497
  - Expected to be released with DAOS 2.4
- Applications
  - Enable Metrics via New Library API calls
  - Allocate, Dump, Reset, Free Counters
  - Daos_test additions to validate

Daos client-side metrics test utility output

```
******************** Dumping Pool RPC Counters ********************
Name                         Inflight         Success         Failure
pool connect                        0              14               5
pool disconnect                     0              14               0
pool attr(get/set)                  0               8               0
pool query                          0              26               0
******************* Dumping Container RPC Counters *******************
Name                         Inflight         Success         Failure
cont create                         0              40               4
cont destroy                        0              39               1
cont open                           0              37               2
cont close                          0              37               0
cont snapshot                       0               5               0
cont snaplist                       0               1               0
cont snapdestroy                    0               5               1
cont attr                           0               8               0
cont acl                            0               3               2
cont prop                           0               4               1
cont query                          0              11               9
cont oidalloc                       0               1               0
cont aggregate                      0               1               0
******************** Dumping Object RPC Counters ********************
Name                         Inflight         Success         Failure
obj update                          0             160               0
obj fetch                           0              84               0
obj enum dkey                       0               3               3
obj enum akey                       0               3               2
obj enum recx                       0              13               6
obj enum obj                        0               0               0
obj punch obj                       0               1               0
obj punch dkeys                     0              29               0
obj punch akeys                     0               7               0
obj query keys                      0               1               0
obj sync                            0               1               0
obj cpd                             0               0               0
*************** Dumping Object IO Stats ***************************
Name                    Count           Sum Size           Sum of Sqrs Size            Min            Max
update                    160           52402115            800682222004243              0       24494592
fetch                      84           24712051            199997103372031              1       12247296
```

Pool-Related RPCs

Container-Related RPCs

Object-Related RPCs

Object IO Statistics

```
*************** Dumping i/o Distribution by Size *****************
Name                         update cnt       fetch cnt
IO_0_1K                              65              34
IO_1K_2K                              1               1
IO_2K_4K                              1               1
IO_4K_8K                             22              34
IO_8K_16K                            41               2
IO_16K_32K                            7               3
IO_32K_64K                            4               1
IO_64K_128K                           1               1
IO_128K_256K                          1               1
IO_256K_512K                         11               1
IO_512K_1M                            1               1
IO_1M_2M                              1               1
IO_2M_4M                              1               1
IO_4M_INF                             3               2
********** Dumping update call Distribution for RP ***************
Name                         update cnt            size
NO_RP                                 5            7050
RP2                                 139        52157970
RP3                                   1           26841
RP4                                   1           21208
RP6                                   1           31566
RP8                                   1           42472
RP12                                  0               0
RP16                                  0               0
RP24                                  0               0
RP32                                  0               0
RP48                                  0               0
RP64                                  0               0
RP128                                 0               0
RPU                                   0               0
********** Dumping update call Distribution for EC ***************
Name                    fstripe/sng cnt            size      pstripe cnt            size
IO_EC2P1                              2           12352                1            8192
IO_EC2P2                              2           16480                1           12288
IO_EC4P1                              2           20544                1            8192
IO_EC4P2                              2           24672                1           12288
IO_EC8P1                              0               0                0               0
IO_EC8P2                              0               0                0               0
IO_EC16P1                             0               0                0               0
IO_EC16P2                             0               0                0               0
IO_ECU                                0               0                0               0
```

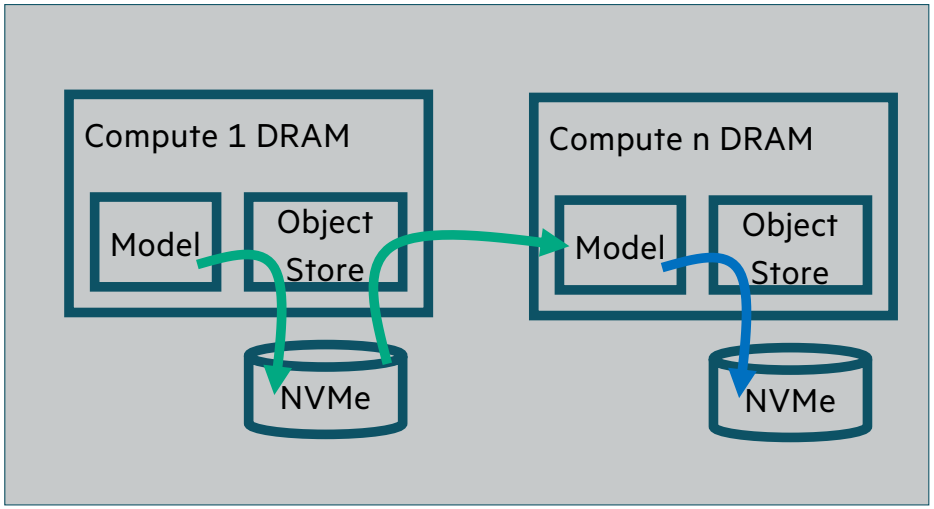
Updates and Fetches For Varying Size Ranges

Updates Using Varied Replication Factors

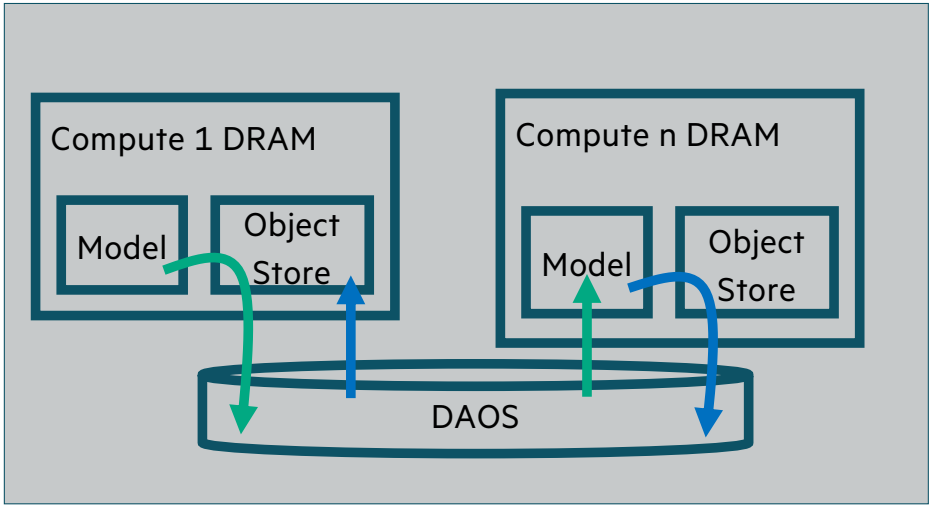
Updates Using Varied Erasure Coding (With Partial vs Full Stripe)

Ray Plasma

Without DAOS Integration



With DAOS Integration

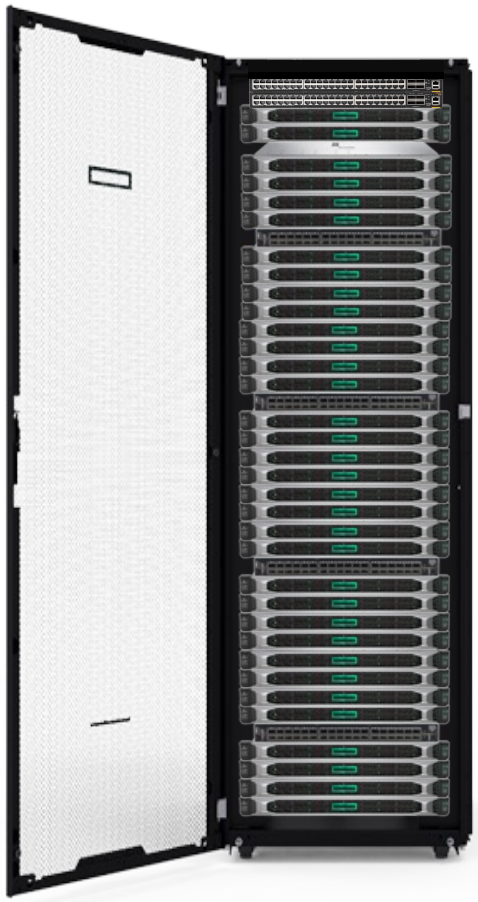


DAOS Administration enablement

- HPE Performance Cluster Manager (HPCM)
  - Server cluster management & monitoring via top-of-rack admin node
  - Can optionally manage compute nodes attached to DAOS as well
- DCM command set augments HPCM
  - Supports multiple logical DAOS systems / clusters within one physical cluster of HPE Proliant nodes
  - Programmatically sets up and tears down mini-clusters on subgroups of nodes
  - Operates / administers DAOS on each of the configured mini-clusters
  - Familiar to HPCM administrators using similar commands
- Cluster Setup Process
  - Compliant HW is pre-assembled onsite or in HPE Manufacturing with firmware / BIOS leveled / configured
  - Admin node's OS/HPCM is installed & added to customer admin network
  - OS distro to be deployed to DAOS servers is added to admin node's HPCM repository
  - BMC & server OS access MACs, and BMC login info are added
- Cluster deployment Process
  - Optionally configure a firewall/gateway from our private admin network thru the admin server to the customer network
  - Discover target nodes found in the config, and install a distro OS, verify the HW
  - Install the DCM package on the admin server
  - Create a DAOS repo on the admin server (may be from web or local DAOS repo mirror/copy)
  - Create and deploy DAOS server images
    - Clone the distro OS on the admin server
    - Install network drivers and DAOS RPMs into the DAOS server image
    - Deploy the image to all the running nodes
  - Use DCM commands to configure DAOS nodes for use
  - Later, DAOS upgrades can be deployed directly to running nodes without re-imaging

DAOS POC System

- A Single-Rack Solution with Maximums:
  - 32 DL-360 Gen10 Plus; 32TB SCM, 512TB flash
  - Four 200Gb Switches (Mellanox or Slingshot)
  - ~700GBps/350GBps raw read/write throughput
  - *~64M/32M peak read/write operations per second*
- Unbundled Repeatable Solution Delivery Method
  - Qualified hardware and software BOM
  - HPCM cluster management software
  - Light installation / configuration scripting
  - Reference doc set: for field or factory integration
  - Customer system administration skills required
- Individual elements sold/supported separately



Up to 2 HPE Management Servers:

- DL-325 Gen10 single-socket

Up to 4 200GbE Switches:

- HPE Slingshot 1
- Mellanox QM8700

Up to 32 HPE DAOS Server Cfg:

- DL-360 Gen10 Plus
- 1-Socket Ice Lake Cfg
- 4x Gen4 NVMe SSD 16+TB
- 8x Optane Memory 1TiB
- 200Gb NIC

SC22 Dallas, TX | hpc accelerates.

Hewlett Packard Enterprise

SC22 Dallas, TX | hpc accelerates.

IO500 BOF:

Tuesday Nov 15

5:15-6:15p

Location: D174

Thank you