SC23
DENVER   NOV 12–17

HPE DAOS Update

Lance Evans

Storage Architect
HPC Chief Technology Office
2023-11-13

HPE's DAOS Involvement

Metadata on SSD: Decouples DAOS from a requirement for persistent memory hardware
Memory allocator contributions

PyTorch: Investigation to optimize DAOS IO for PyTorch
Evaluation is based on fsspec over DFS (& adding dlio_benchmark)

Solution HW: Hardware refresh available to current Proliant servers
Roughly doubles single-rack performance

DAOS Foundation: Founding member of the Foundation
Widens the community of DAOS developers and adopters

POC Systems: DAOS testbed availability for application integration and tuning
Assists customers, partners, ISVs in integrating with DAOS

SC23
Denver, CO | i am hpc.

DAOS POC System Update

- A Single-Rack Solution with Maximums:
  - 32 DL-360 Gen11; 1TB DDR5, 5PB flash
  - Four 200Gb (Slingshot) or 400Gb (Nvidia) Switches
  - ~1.4TBps/1.4TBps raw read/write throughput
  - *~64M/32M peak read/write operations per second*
  - Double these maximums available with chilled doors
- 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 Gen11 single-socket

Up to 4 200 or 400GbE Switches:
- HPE Slingshot 200
- Mellanox QM8700

Up to 32 HPE DAOS Server Cfg:
- DL-360 Gen11
- 1-Socket Sapphire Rapids Cfg *
- Max 10x Gen5 NVMe SSD 3.2TB+ *
- 16x DDR5 *
- 200Gb or 400Gb NIC *

* Double density possible with chilled doors

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

SC23
Denver, CO | i am hpc.

Thank you