NEW DATA PATH SOLUTIONS FOR HPC SIMULATION, AI, AND HIGH PERFORMANCE WORKLOADS

Marc Roskow – HPC Storage Product Management
Lance Evans – HPC Chief Technology Office
May 3, 2021
HPE PARALLEL FILE SYSTEM STORAGE: PLACE IN THE ECOSYSTEM

Leverages all surrounding functionality
INTRODUCING HPE PARALLEL FILE SYSTEM STORAGE

Fusion of the leading enterprise parallel file system with the leading industry-standard servers in the HPE factory

- **HPE storage with embedded IBM Spectrum Scale ECE file system running on HPE ProLiant DL325 Gen10 Plus servers** for
  - Clusters of HPE Apollo 2000/HPE ProLiant DL rack servers and HPE Apollo 6500 AI solutions
  - Verticals that often reject Lustre as they need enterprise grade storage features (e.g. Financial Service, Life Sciences, etc.)
  - Home directory storage for large HPC solutions where /scratch directories are on Lustre-based Cray ClusterStor E1000
- **Provides high performance parallel data access for compute nodes concurrently** via
  - Native IBM Spectrum Scale client installed on compute nodes
  - NFS/SMB via Cluster Export Services (CES)
- **High speed connectivity to compute nodes** via
  - InfiniBand HDR100/Ethernet 100 Gb
  - InfiniBand HDR/Ethernet 200 Gb
- **Available in All Flash, All HDD or mixed configurations** based on workload profile
- **Provides broad set of enterprise storage functionality** like
  - Enterprise-grade system availability ("5 Nines") incl. non-disruptive hardware & software upgrades, online expansion/contraction of the file system, etc.
  - Snapshots, compression, data replication, end-to-end data encryption, end-to-end data integrity (from disk to client), audit features for compliance
  - Protocol support beyond POSIX for NFS, SMB, HDF5, Object (S3, SWIFT) and (soon) Nvidia GPUDirect Storage
  - Data life cycle management - policy based data movement and curating and auto-tuning
- **Single price for the HPE storage system** (no file system license per terabyte or per storage drive)
- **Base warranty 3 years** for hardware & 1 year for software – HPE Pointnext Tech Care and HPE Datacenter Care are available
TWO TYPES OF STORAGE SERVERS

Combination of both in the same file system is supported

**HPE ProLiant DL325 Gen10 Plus with 16 x SFF slots**
- # of NVMe SSD per server: 3, 4, 6, 8, 10, 12, 14 or 16
- Capacity points of NVMe SSD in TB: 3.84, 7.68 or 15.36
- (2) InfiniBand HDR/Ethernet 200 Gb 1p adapters

**HPE ProLiant DL325 Gen10 Plus with 8 x LFF slots**
- # of SAS 7.2K RPM HDD per server: 3, 4, 6, or 8
- Capacity points of HDD in TB: 4, 8, 12 or 16
- (1) InfiniBand HDR100/ Ethernet 100 Gb 2p adapter
- Factory installed 1.6 TB NVMe capacity to serve file system metadata and small files from fast NVMe Flash.

**Same configuration rules for both:**
At least 4 and up to 32 storage servers in identical configuration in a RAID cluster.
HPE Parallel File System Storage Installation and Startup Service needs to be ordered with the system.

**11 Flash Storage Servers with**
- 1,037 TB usable capacity
- 260 GB/s Read and 232 GB/s Write bandwidth

<table>
<thead>
<tr>
<th>SKU</th>
<th>Description</th>
<th>Qty</th>
</tr>
</thead>
<tbody>
<tr>
<td>R7R36A</td>
<td>HPE Parallel File System 16Flash BaySvr</td>
<td>11</td>
</tr>
<tr>
<td>R7R40A</td>
<td>HPE PFS 15.36TB NVMe RI SC PM1733 SSD</td>
<td>88</td>
</tr>
<tr>
<td>R7R46A</td>
<td>HPE PFS IB HDR/EN 200Gb 1p QSFP56 Adptr</td>
<td>22</td>
</tr>
</tbody>
</table>

**4 Flash and 10 Storage Servers with**
- 1,008 TB usable capacity
- 101 GB/s Read and 37 GB/s Write bandwidth

<table>
<thead>
<tr>
<th>SKU</th>
<th>Description</th>
<th>Qty</th>
</tr>
</thead>
<tbody>
<tr>
<td>R7R36A</td>
<td>HPE Parallel File System 16Flash BaySvr</td>
<td>4</td>
</tr>
<tr>
<td>R7R38A</td>
<td>HPE PFS 3.84TB NVMe RI SC U.3 PM1733 SSD</td>
<td>32</td>
</tr>
<tr>
<td>R7R46A</td>
<td>HPE PFS IB HDR/EN 200Gb 1p QSFP56 Adptr</td>
<td>8</td>
</tr>
<tr>
<td>R7R35A</td>
<td>HPE Parallel File System 8Disk BaySvr</td>
<td>10</td>
</tr>
<tr>
<td>R7R44A</td>
<td>HPE PFS 16TB SAS 7.2K LFF LP HDD</td>
<td>80</td>
</tr>
<tr>
<td>R7R45A</td>
<td>HPE PFS IB HDR100/EN 100Gb 2p QSFP56 Adptr</td>
<td>10</td>
</tr>
<tr>
<td>R7R37A</td>
<td>HPE PFS 1.6TB Flash Metadata Store</td>
<td>10</td>
</tr>
</tbody>
</table>

**11 HDD Storage Servers with**
- 1,046 TB usable capacity
- 7.1 GB/s Read and 5.4 GB/s Write bandwidth

<table>
<thead>
<tr>
<th>SKU</th>
<th>Description</th>
<th>Qty</th>
</tr>
</thead>
<tbody>
<tr>
<td>R7R35A</td>
<td>HPE Parallel File System 8Disk BaySvr</td>
<td>11</td>
</tr>
<tr>
<td>R7R44A</td>
<td>HPE PFS 16TB SAS 7.2K LFF LP HDD</td>
<td>88</td>
</tr>
<tr>
<td>R7R45A</td>
<td>HPE PFS IB HDR100/EN 100Gb 2p QSFP56 Adptr</td>
<td>11</td>
</tr>
<tr>
<td>R7R37A</td>
<td>HPE PFS 1.6TB Flash Metadata Store</td>
<td>11</td>
</tr>
</tbody>
</table>
KEY DATES FOR NEW HPE PARALLEL FILE SYSTEM STORAGE

- **April 6, 2021**: Announcement/launch
- **May 17, 2021**: GA/Volume shipments
- **August 2, 2021**: Larger drives, Higher HDD density, etc.
HPE DAOS REFERENCE ARCHITECTURE ATTRIBUTES

- Low latency, highly concurrent data path software
  - Intel DAOS 2.0
  - Client libraries for KV, object
  - Direct app integration for HDF5, MPI-IO, Spark, Python,
  - POSIX via IO library, preloaded intercept library, or FUSE

- Hardware
  - Based on Proliant DL-380 or 360 Gen10 Plus
  - SSD plus persistent memory in ratios (100% + 6%)
  - Infiniband HDR, Slingshot 11 interconnect options
  - Possible server-side accelerators for computational storage experiments

- Supporting software
  - HPCM for cluster management
  - Distributed scripting for SW configuration
  - Core cluster and software control

- Installation model
  - Onsite via compliant BOM, cookbook, scripts
  - Factory integration prior to arrival using similar tools

WHAT IT IS

**DAOS Client:** Library With Sharding & Erasure Coding

**DAOS Server:** Horizontally Scalable Storage Engine

App: AI/Analytics/Database/Simulation

POSIX  MPIIO  HDF5  Python  Spark  Etc

Software Latency ~20us

Infiniband, Slingshot (libfabric-based)

MySQL?  Tensorflow?
HPE DAOS REFERENCE ARCHITECTURE BENEFITS

- Differentiated high-performance applications
  - Large shared file/object performance at scale
  - Tiny parallel KV, object, graph projects
  - Random searches across arbitrarily sized datasets
  - Concurrency of multiple users, untuned IO, without interference
  - Develop on laptop, deploy on exascale
  - Highest performance and concurrency at any IO size

- Exposes emerging hardware speeds
  - Achieve full bandwidth of media to client applications
  - Drive media queue depths to operational limits
  - Low latency for reduced time to results in complex query sequences
  - Parallelism of client-side erasure coding

WHAT PROBLEMS IT SOLVES
REFERENCE IMPLEMENTATION, CAPACITY CONFIG

- Unbundled Repeatable Solution Delivery Method
  - Qualified hardware and software BOMs
  - HPCM cluster management software
  - Light installation / configuration scripting packages
  - Reference doc set: for field or factory integration
  - Customer system administration skills required
- Single-Rack Solution Maximums (Capacity Config):
  - 16x DL-380 Gen10 Plus; 128TiB pmem, 2PB raw capacity
  - 4x 200Gb Switches (Mellanox or Slingshot)
  - ~1,400GBps/700GBps raw read/write throughput
  - ~100M/50M read/write OPS
- DAOS Solution Milestones
  - Build it yourself from available HPE SKUs in June
  - Repeatable BOMs, scripts, instructions in October
  - Full Productization In planning phase, target late 2022
- Support
  - Select early POC customers assisted by engineering
  - Official support offering with productization

HOW TO GET IT

2x Aruba 8360 Mgmt Switches max
3x HPE Management Servers max:
  - DL-325 Gen10 single-socket

Max 40RU

4x 200GbE Switches (optional):
  - Mellanox QM8700 (72 uplinks max)
  - HPE Slingshot 1 (80 uplinks max)

16x max HPE DAOS Server Cfg each:
  - DL-380 Gen10 Plus (Ice Lake)
  - 16x Gen4 NVMe SSD 122TB ttl
  - 16x Optane Memory 8TiB ttl (6.7%)
  - 2x 200Gb NIC
REFERENCE IMPLEMENTATION, PERFORMANCE CONFIG

- Unbundled Repeatable Solution Delivery Method
  - Qualified hardware and software BOM
  - HPCM cluster management software
  - Light installation / configuration scripting packages
  - Reference doc set: for field or factory integration
  - Customer system administration skills required

- Single-Rack Solution Maximums (Performance Config):
  - 32x DL-360 Gen10 Plus; 64TiB pmem, 1PB raw capacity
  - 4x 200Gb Switches (Mellanox or Slingshot)
  - ~1,400GBps/700GBps raw read/write throughput
  - ~200M/100M read/write OPS

- DAOS Solution Milestones
  - Build it yourself from available HPE SKUs in June
  - Repeatable BOMs, scripts, instructions in October
  - Full Productization In planning phase, target late 2022

- Support
  - Select early POC customers assisted by engineering
  - Official support offering with productization

HOW TO GET IT

2x Aruba 8360 Mgmt Switches max
3x HPE Management Servers max:
  - DL-325 Gen10 single-socket

4x 200GbE Switches (optional):
  - Mellanox QM8700 (72 uplinks max)
  - HPE Slingshot 1 (80 uplinks max)

32x max HPE DAOS Server Cfg each:
  - DL-360 Gen10 Plus (Ice Lake)
  - 8x Gen4 NVMe SSD 30TB ttl
  - 16x Optane Memory 2TiB ttl (6.7%)
  - 2x 200Gb NIC
THANK YOU

Q&A after the next presentation