Dell EMC Ready Solutions for HPC Lustre Storage

Forrest Ling
HPC Enterprise Technologist at Dell EMC
Greater China
2018.10.23
Dell EMC Supports HPC Open Source Software

- Support Open Source Software projects
- Prove Open Source Software in Dell EMC HPC Labs
- Organize HPC Open Source Software Forums
- Provide HPC Platforms running Open Source Software
Dell (China) Awarded “HPC Open Source Software Technical Practice Innovation”

THANKS to:
• CCF TCHPC*
• OpenHPC
• OpenSFS
• OpenPBSpro
• Singularity
• …

CCF TCHPC – China Computer Federation, Technical Committee of HPC
Dell EMC HPC market leadership

10 generations of servers and storage in HPC clusters

1999 2001 2004
1st HPC cluster

2004
Industry’s 1st HPC Ready Solution

2004
Thunderbird

2005
DCS formed

2008
C-series joins PowerEdge

2008
Tungsten

#4

2008
#6

2005

2012

2015
Dell EMC merger — Isilon joins HPC portfolio

2016

2017
Dell EMC AI solutions announced

2017
Zenith System launched at Dell EMC HPC Innovation Lab

2016
Stampede

#7

2015
Dell EMC merger — Isilon joins HPC portfolio

2016

2017
Dell EMC AI solutions announced

2017
Zenith System launched at Dell EMC HPC Innovation Lab

2016
Stampede

#7

2015
Dell EMC merger — Isilon joins HPC portfolio

2016

2017
Dell EMC AI solutions announced

2017
Zenith System launched at Dell EMC HPC Innovation Lab

2016
Stampede

#7
# PowerEdge HPC Server portfolio

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>C6420</td>
<td>Maximizes density, scalability, and energy efficiency per U for high performance hyperscale workloads</td>
<td>Purpose built for HPC</td>
</tr>
<tr>
<td>C6320p: Intel® Xeon Phi™ Processor (KNL)</td>
<td>2 socket, ultra-dense, 4 GPU rack server</td>
<td></td>
</tr>
<tr>
<td>C4140</td>
<td>Ideal combination for dense scale out data center computing and storage in a 1U/2S platform</td>
<td>Infrastructure and I/O</td>
</tr>
<tr>
<td>R640/R440</td>
<td>Ideal for applications requiring best-in-class storage performance, high scalability, and density. Support up to 3 double-wide GPUs.</td>
<td>Large memory</td>
</tr>
<tr>
<td>R740, R740xd</td>
<td>Ideal for mission-critical applications and real-time data and analytics</td>
<td>Modular</td>
</tr>
<tr>
<td>R840, R940, R940xa</td>
<td>R940xa: Support up to 4 double-wide GPUs or up to 8 FPGAs.</td>
<td></td>
</tr>
<tr>
<td>MX7000</td>
<td>Dense compute and optimal memory throughput for demanding HPC workloads</td>
<td></td>
</tr>
</tbody>
</table>

**powerful performance | density | efficiency**

Additional options: DellEMC.com/servers
AMD processor-based PowerEdge HPC Server portfolio

**R6415**
Storage flexibility with up to 10 PCIe NVMe drives. Up to 32 cores in a single socket. 16 DIMMs bringing up to 2TB of total memory capacity. 128 lanes of PCIe 3.0 connectivity.

**R7415**
Up to 24 flash drives, 32 total cores and 16 DDR4 DIMMs.

**R7425**
Up to 30% superior HPC performance. 4x more NVMe for extreme performance storage requirements. Three double wide GPUs, plus up to 8 NVMe drives. Up to 4TB memory capacity optimized.

---

**Powerful performance | Density | Efficiency**

Additional options: [DellEMC.com/servers](https://www.dell EMC.com/servers)
# HPC Storage portfolio

## GPFS
Available through DDN and Arcastream partnerships

## Lustre
Building in-house support capability in partnership with Intel; high-end support in partnership with DDN and ClusterStor

## BeeGFS, Ceph and others
PowerVault and PowerEdge platforms can power a variety of storage capabilities

## HPC NFS Storage
High availability storage system with up to 480TB of raw storage capacity
Red Hat HA software stack

## HPC Lustre Storage
Lustre storage starting from 960TB raw storage per object storage server pair and 14.9GB/s of write and 16.8GB/s of read throughput. Scale out performance and capacity with additional OSS pairs

## Isilon
Scale-out NAS storage to store, protect and analyze unstructured data

## Elastic Cloud Storage (ECS)
All the benefits of a public cloud while keeping cost under control

---

### File system

### High performance and flash

### Archival

- **powerful performance | efficiency | scalability**

HPC Storage comparison

- Dell EMC + 3rd-party
- HPC Lustre Storage
- Other SDS
- HPC NFS Storage
- Job Scheduler
- Isilon
- Elastic Cloud Storage (ECS)
- Virtustream

Performance

Capacity

Scratch

Project

Archive
# HPC fabric portfolio

<table>
<thead>
<tr>
<th>Dell EMC Ethernet</th>
<th>Mellanox InfiniBand</th>
<th>Dell EMC H-Series</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Utilized for base management and/or primary interconnect</td>
<td>• High speed, low latency fabric for optimized communication</td>
<td>• Designed specifically for HPC</td>
</tr>
<tr>
<td>• Options range from 1GbE to 100 GbE</td>
<td>• Established HPC fabric</td>
<td>• Higher switch chip port count for added density</td>
</tr>
<tr>
<td>• Low complexity</td>
<td>• Virtual Protocol Interconnect (VPI) offers flexibility for converged fabric</td>
<td>• CPU-fabric integration</td>
</tr>
<tr>
<td>• Smaller lower cost HPC solutions</td>
<td>• Proven high performance fabric</td>
<td>• Maximizes FLOPs/$ keeping costs in balance</td>
</tr>
<tr>
<td>• In-house expertise for Ethernet management</td>
<td>• Connect to existing InfiniBand fabric for HPC storage or compute</td>
<td>• Fewer switch hops in fat-tree</td>
</tr>
</tbody>
</table>

Options included in all Dell EMC Ready Solutions for HPC or custom designed opportunities.
Dell EMC “4 Stages” Methodology Grow HPC in China

Business Life Cycle
- Funding Application
- Commercial Contract
- Production
- Return of Investment

Technology Life Cycle
- Technical Design
- Deployment
- Maintainance
- Innovation

Four - “zation”
- Templatization
- Standardization
- Communityization
- Globalization

Participate & Alliance
- HPC Open Source Software (OpenHPC) Alliance

Participate & Alliance
- Ready Solutions for HPC:
  - Life Sciences;
  - Manufacturing;
  - Lustre
- Services for HPC:
  - ProDeploy
  - ProSupport
  - Partners Support
- Social Media for HPC:
  - WeChat Groups
  - Blogs Web sites
- International exchange:
  - Top500.org
  - SC/ISC
  - Hpcwire.com
  - Insidehpc.com

Make it Real
Simplified design, configuration and ordering

Dell EMC has developed several modular Ready Solutions for HPC

- Dell EMC Ready Solution for HPC Life Sciences
- Dell EMC Ready Solution for HPC Digital Manufacturing
- Dell EMC Ready Solution for HPC Research
- Dell EMC Ready Solution for HPC NFS Storage
- Dell EMC Ready Solution for HPC Lustre® Storage
# Upcoming Lustre solution – PowerVault ME storage

## Server configuration

<table>
<thead>
<tr>
<th>IML, MDS and OSS server model</th>
<th>One DellEMC PowerEdge R640 and Four DellEMC PowerEdge R740.</th>
</tr>
</thead>
</table>
| Processor                     | IML Server: Dual Intel Xeon Gold 5118 @ 2.3GHz  
                                 MDS and OSS Servers: Dual Intel Xeon™ Gold 6136 @ 3.00GHz |
| Memory                        | IML Server: 12 x 8GB 2666MT/s DDR4 RDIMMs.  
                                 MDS and OSS Servers: 24 x 16GiB 2666MT/s DDR4 RDIMMs. |
| InfiniBand HCA (slot 8)        | Mellanox ConnectX-5 EDR PCIe adapter. |
| External storage controller (slot 1,2,4 and 5) | Four Dell 12Gbps SAS HBAs. |

## Storage configuration

<table>
<thead>
<tr>
<th>Object Storage Enclosure</th>
<th>Four DellEMC PowerVault ME4084 enclosures fully populated for 336 drives. 2.69PB raw capacity solution if equipped with 8TB SAS drives.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata Storage Enclosure</td>
<td>DellEMC PowerVault ME4024 enclosure fully populated with 24 x 960GB SAS SSD</td>
</tr>
<tr>
<td>RAID controllers</td>
<td>Duplex RAID controllers in the DellEMC ME4084 enclosure. Duplex RAID controllers in the DellEMC ME4024 enclosure.</td>
</tr>
</tbody>
</table>
| Hard Disk Drives               | 84 - 8TB 7200 rpm NL SAS drives per ME4084 enclosure.( ME4 supports up to 12TB HDDs )  
                                 24 – 960GB SAS SSDs per ME4024 enclosure. |
Initial Performance Data

Sequential IOzone (2TB aggregate data size)

- Throughput in GB/s
- #Threads

Throughput measurements for Sequential IOzone tests with different numbers of threads. The y-axis represents throughput in GB/s, and the x-axis shows the number of threads.

Mdtest File Metadata: 2M Files, 2x MDTs (DNE)

- File create 2MDTs
- File Stat 2MDTs
- File Remove 2MDTs

Graph showing Mdtest File Metadata with 2M Files and 2x MDTs in DNE. The x-axis represents the number of threads, and the y-axis shows operations per second (Ops/sec).

File MDTest 2M files, 2 MDTs in DNE

Graph illustrating File MDTest with 2M files and 2 MDTs in DNE. The x-axis displays the number of threads, and the y-axis shows the number of operations per second (Ops/sec).
## Upcoming Lustre solution – PowerVault ME storage

<table>
<thead>
<tr>
<th>Category</th>
<th>Improvement over previous generation</th>
<th>Lustre with ME4</th>
<th>Lustre with ME3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hard drives in base config</td>
<td>40%</td>
<td>336 drives</td>
<td>240 drives</td>
</tr>
<tr>
<td>Capacity</td>
<td>43%</td>
<td>Up to 4PB</td>
<td>Up to 2.8 PB</td>
</tr>
<tr>
<td>Peak Sequential Writes</td>
<td>42%</td>
<td>21.3 GB/s</td>
<td>15 GB/s</td>
</tr>
<tr>
<td>Peak Sequential Reads</td>
<td>25%</td>
<td>22.5 GB/s</td>
<td>18 GB/s</td>
</tr>
<tr>
<td>Peak File create metadata</td>
<td>107%</td>
<td>≈ 60K op/s</td>
<td>≈ 29K op/s</td>
</tr>
<tr>
<td>Peak File remove metadata</td>
<td>50%</td>
<td>≈ 240K op/s</td>
<td>≈ 160K op/s</td>
</tr>
<tr>
<td>Peak File stat metadata</td>
<td>23%</td>
<td>≈ 669K op/s</td>
<td>≈ 544K op/s</td>
</tr>
</tbody>
</table>

ME4012: 12-drive RBOD (DPE)  
ME4024: 24-drive RBOD (DPE)  
ME4084: 84-drive RBOD (DPE)  

ME412: 12-drive Expansion (DAE)  
ME424: 24-drive Expansion (DAE)  
ME484: 84-drive Expansion (DAE)
HPC Solution Support and Deployment Services

Deployment

ProDeploy for HPC

1. HPC Add-on: Individual nodes
2. HPC Add-on: M1000e
3. HPC Add-on: Storage

Support

Asset-level support

ProSupport

or

ProSupport Plus

Solution support

ProSupport Add-on for HPC

Supported Hardware and Software Technology and Local Services Partners

Cluster Management

Operating System

Networking

Storage

Server

Local Partners
Thanks!