Enterprise Ceph: Everyway, your way!

Amit Bhutani @ Dell
Kyle Bader @ Red Hat
Red Hat Summit
June 2016
Agenda

• Overview of Ceph Components and Architecture

• Evolution of Ceph in Dell-Red Hat Joint OpenStack Solution

• Red Hat Ceph Storage – General Hardware Recommendations

• Performance and sizing Dell DSS 7000² for Capacity Optimized Object Storage

• Performance and sizing Dell PE R730xd for Cost/Capacity and Throughput optimized Block/Object Storage
Ceph – Architecture

APP

RGW
A web services gateway for object storage, compatible with S3 and Swift

HOST/VM

RBD
A reliable, fully-distributed block device with cloud platform integration

CLIENT

CEPHFS
A distributed file system with POSIX semantics and scale-out metadata management

LIBRADOS
A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby, PHP)

RADOS
A software-based, reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors
Ceph – Components

OSDs:
- 10s to 10000s in a cluster
- One per disk (or one per SSD, RAID group…)
- Serve stored objects to clients
- Intelligently peer for replication & recovery

Monitors:
- Maintain cluster membership and state
- Provide consensus for distributed decision-making
- Small, odd number
- These do not serve stored objects to clients
Ceph – Enlightened Application

RADOS CLUSTER
Ceph – Native Protocol

LIBRADOS:
- Direct access to RADOS for applications
- C, C++, Python, PHP, Java, Erlang
- Direct access to storage nodes
- No HTTP overhead
- Object Classes
Ceph – RESTful Object Storage

RADOS CLUSTER

APPLICATION

RADOSGW
LIBRADOS

RADOSGW
LIBRADOS

APPLICATION
Ceph – Make It Webby

RADOS Gateway:
- REST-based object storage proxy
- Uses RADOS to store objects
- API supports buckets, accounts
- Usage accounting for billing
- Supports S3 and Swift protocols
- Consistent within zone
- Multi-master zone based async replication
Ceph – Block Storage

RADOS CLUSTER
Ceph – Block Storage

RADOS BLOCK DEVICE:

- Storage of disk images in RADOS
- Decouples VMs from host
- Images are striped across the cluster (pool)
- Snapshots (full / differential)
- Copy-on-write clones
- Async mirroring
Evolution of Ceph in Dell-Red Hat Joint OpenStack Solution

- Ceph was part of the first generation of Dell-Red Hat joint Cloud Solution release
  - Based on OSP 5 ("Icehouse") and Inktank Enterprise Ceph 1.2 ("Firefly")
  - Ceph integrated with OpenStack Cinder, Glance, Ephemeral storage and Foreman

- Currently in the fifth generation of the Dell-Red Hat joint Cloud Solution
  - Based on OSP 8 ("Liberty") and RHCS 1.3.2 ("Hammer")
  - Single Ceph cluster (3+ nodes) with pools for Cinder, Glance, Ephemeral storage and Object storage

- Dell Enhancements
  - Ceph integrated into OSP8 Director
  - Placement Groups customizable on a per pool basis
  - Calamari Deployment
  - Load balanced and HA RADOS Gateway
    - Swift protocol endpoint automatically configured
    - Configured for Keystone authentication

Dell-Redhat OpenStack Reference Architecture URL: [http://www.dell.com/openstack](http://www.dell.com/openstack)
# Red Hat Ceph Storage – General Hardware Recommendations

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Properties</th>
<th>Example Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IOPS Optimized</strong></td>
<td>• Lowest cost per IOPS</td>
<td>• Typically block storage&lt;br&gt;• 3x replication (HDD) or 2x replication (SSD)&lt;br&gt;• MySQL on OpenStack clouds</td>
</tr>
<tr>
<td></td>
<td>• Highest IOPS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Meets minimum fault domain recommendation (single server is ≤ 10% of the cluster)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• MySQL on OpenStack clouds&lt;br&gt;• Active performance storage for video, audio, and images&lt;br&gt;• Streaming media</td>
</tr>
<tr>
<td><strong>Throughput Optimized</strong></td>
<td>• Lowest cost per given unit of throughput&lt;br&gt;• Highest throughput&lt;br&gt;• Highest throughput per BTU/watt&lt;br&gt;• Meets minimum fault domain recommendation (single server is ≤ 10% of the cluster)</td>
<td>• Block or object storage&lt;br&gt;• 3x replication&lt;br&gt;• Active performance storage for video, audio, and images&lt;br&gt;• Streaming media</td>
</tr>
<tr>
<td><strong>Capacity Optimized</strong></td>
<td>• Lowest cost per TB&lt;br&gt;• Lowest BTU per TB&lt;br&gt;• Lowest watt per TB&lt;br&gt;• Meets minimum fault domain recommendation (single server is ≤ 15% of the cluster)</td>
<td>• Typically object storage&lt;br&gt;• Typically erasure coded&lt;br&gt;• Object archive&lt;br&gt;• Video, audio, and image object archive repositories</td>
</tr>
</tbody>
</table>
Benchmark Tooling

• Network testing – full graph iperf
• CBT – Ceph Benchmark Tool
  • RADOS Bench
    • Various block sizes
    • 3x and Erasure Coded Pools
• COSBench
  • Swift benchmarks
Performance and sizing Dell DSS7000

Storage server built to scale

DSS 7000 / 7500
Ultra-dense Storage Server

- Ideal for high-capacity object storage workloads including archival
- 4U chassis with 720 TB of storage using 90x 3.5” drives (8TB per drive)
- Single or dual server solution in a single chassis
- Ceph 2.0 with decreased deployment time, improved security/authentication, Enhanced compat with S3 and Swift
# Dell DSS 7000² Detailed Specifications

<table>
<thead>
<tr>
<th><strong>Form Factor</strong></th>
<th>4U rack mount (EIA standard 19” width)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Processors</strong></td>
<td>2x Intel® Xeon® E5-2600 v4 per node</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>12 x DIMMs per node: 16GB/32GB DDR4 RDIMM</td>
</tr>
</tbody>
</table>
| **Storage Controllers** | Avago MegaRAID 9361-8i 12Gb/s SAS and SATA  
PMC Sierra PMC8805 12Gb/s SAS and SATA |
| **Drive Bays**  | 90 x 3.5” hot swappable drive options  
2 x 2.5” hot swappable boot drives per node |
| **Drive Type**  | Enterprise SATA HDD, SATA SSD           |
| **PCIe Slots**  | 3 x8; 1 x16 slots (low profile and full height) |
| **Embedded Networking** | 4x 1GbE LOM per node                  |
| **Systems Management** | IPMI 2.0, BMC with vKVM and 1 x 1GbE management port |
| **Power Supply** | Two Platinum efficiency 1100W or 1600W AC hot-plug redundant power supplies per node |
| **Availability** | Select countries                       |
Dell DSS 7000² Conclusions

1. 90 MB/s per OSD on replicated reads
2. 25 MB/s per OSD on replicated writes
3. 16250 MB/s per server on replicated reads
4. 4325 MB/s per server on replicated writes
5. Highest density per rack unit
6. Excellent performance with Red Hat Ceph Storage 2.0
Incredible range of configurability ideal for variety of workloads

- 6 distinct storage configurations
- Support four ultra fast, ultra-low latency Express Flash NVMe PCIe SSDs
- Supports up to 6 PCIe 3.0 expansion slots
# Dell R730xd Detailed Specifications

<table>
<thead>
<tr>
<th><strong>Form Factor</strong></th>
<th>2U rack mount</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Processors</strong></td>
<td>2x Intel® Xeon® E5-2600 v4 product family (up to 22 cores)</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>Up to 1.5 TB (24 DIMM slots): 4 GB/8 GB/16 GB/32 GB/64 GB DDR4 up to 2400MT/s</td>
</tr>
</tbody>
</table>
| **Storage Controllers** | Internal controllers: PERC H330, PERC H730, PERC H730P  
                          External HBAs (RAID): PERC H830  
                          External HBAs (non-RAID): 12Gbps SAS HBA |
| **Drive Bays**       | Internal hard drive bay and hot-plug backplane:  
                          Up to 16 x 3.5” SAS, SATA, nearline SAS, SSD + 2 x 2.5” drives  
                          Up to 18 x 1.8” SAS, SATA, nearline SAS, SSD drives + 8 x 3.5” SAS, SATA, nearline SAS, SSD drives, + 2 x 2.5” HDD  
                          Up to 26 x 2.5” SAS, SATA, nearline SAS, SSD, PCIe SSD drives |
| **PCIe Slots**       | Up to 6 x PCIe 3.0 slots plus dedicated RAID card slot |
| **Embedded Networking** | 4 x 1GbE, 2x10+2GbE, 4 x 10GbE NDC |
Dell R730xd Configurations

- Measured the throughput optimization and price/performance in the 3x replication and erasure-coded nodes:

- Some of the configurations tested and their legends are:
  1. RAD_12+3_raid_0_10gbe_3xRep
  2. RAD_12+3_raid_0_10gbe_EC
  3. RAD_16+1_raid_0_10gbe_3xRep
  4. RAD_16+1_raid_0_10gbe_EC
  5. RAD_16+1_jbod_10gbe_3xRep
  6. RAD_16+1_jbod_0_10gbe_EC
  7. RAD_16+0_raid_0_10gbe_3xRep
  8. RAD_16+0_raid_0_10gbe_EC

- Eg. #1 - PowerEdge R730XD with 12 hard disk drives (HDDs) and 3 solid state drives (SSDs) in 3X Data Replication and single-drive RAID0 mode. Similarly, ‘jbod’ indicates PERC passthrough mode.
Dell R730xd – Benchmark Environment
Configuration Drive / Server Throughput
Dell R730xd Conclusions

1. 16+1 wins for replicated writes, 12+3 wins for replicated reads
2. 13+3 replicated reads cost/performance is lowest
3. 16+1 wins for replicated writes in cost/performance
4. 16+0 wins for cost/capacity with erasure coding
5. 12+3 replicated wins for per drive and per server reads, 16+1 for writes with erasure coding
6. For throughput, JBOD and RAID0 performance is similar
7. 16+1 replicated does well for per drive and per server reads, also writes with erasure
8. 16+0 wins for reads, 16+1 for writes
# Red Hat Ceph Storage – Dell Hardware Recommendations

<table>
<thead>
<tr>
<th></th>
<th>Extra Small</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100TB+</td>
<td>500TB+</td>
<td>1PB+</td>
<td>2PB+</td>
</tr>
<tr>
<td><strong>IOPS</strong></td>
<td>Future Direction</td>
<td>Future Direction</td>
<td>Future Direction</td>
<td>Future Direction</td>
</tr>
<tr>
<td><strong>Throughput</strong></td>
<td>• 7x R730xd (14U)</td>
<td>• 21x R730xd (42U)</td>
<td>• 6x DSS7000² (24U)</td>
<td>• 12x DSS7000²</td>
</tr>
<tr>
<td></td>
<td>• 1x server/chassis</td>
<td>• 1x server/chassis</td>
<td>• 2x server/chassis</td>
<td>• 2x server/chassis</td>
</tr>
<tr>
<td></td>
<td>• 12x 4TB HDD</td>
<td>• 12x 6TB HDD</td>
<td>• 45x 6TB HDD</td>
<td>• 45x 6TB HDD</td>
</tr>
<tr>
<td></td>
<td>• 3x 200GB SSD</td>
<td>• 3x 200GB SSD</td>
<td>• 2x 800GB NVMe SSD</td>
<td>• 2x 800GB NVMe SSD</td>
</tr>
<tr>
<td></td>
<td>• 2x 10GbE</td>
<td>• 2x 10GbE</td>
<td>• 2x 40GbE</td>
<td>• 2x 40GbE</td>
</tr>
<tr>
<td></td>
<td>• 3x Replication</td>
<td>• 3x Replication</td>
<td>• 3x Replication</td>
<td>• 3x Replication</td>
</tr>
<tr>
<td><strong>Capacity</strong></td>
<td>N/A</td>
<td>N/A</td>
<td>• 15x R730xd (30U)</td>
<td>• 6x DSS7000²</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• 1x server/chassis</td>
<td>• 2x server/chassis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• 16x 6TB HDD</td>
<td>• 45x 6TB HDD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• 0x SSD</td>
<td>• 0x SSD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• 2x 10GbE</td>
<td>• 2x 40GbE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• 8:3 Erasure Coding</td>
<td>• 8:3 Erasure Coding</td>
</tr>
</tbody>
</table>
The power to do more
Backup
Replicated vs Erasure Coded Pools
HBA JBOD vs RAID0

Drive Throughput (MBps avg) - HBA JBOD vs Single-Disk RAID0

Server Throughput (MBps avg) - HBA JBOD vs Single-Disk RAID0

Drive Throughput (MBps avg) - HBA JBOD vs Single-Disk RAID0

Server Throughput (MBps avg) - HBA JBOD vs Single-Disk RAID0
Block Sizes

Which IO Patterns Benefit from >10GbE on R730XD OSD - Server Throughput (MBps avg)

How Does Workload Block Size Affect Ceph Performance - Drive Throughput (MBps avg)

Which IO Patterns Benefit from >10GbE on R730XD OSD - Server Throughput (MBps avg)

How Does Workload Block Size Affect Ceph Performance - Drive Throughput (MBps avg)
More Block Sizes

**Which IO Patterns Benefit from >10GbE on R730XD OSD - Server Throughput (MBps avg)**

- Line graphs showing the performance differences between 12+3_rep and 12+3_EC for both read and write operations.

**How Does Workload Block Size Affect Ceph Performance - Server Throughput (MBps avg)**

- Line graphs comparing the performance of different block sizes for both read and write operations.
Capacity / Throughput Analysis

Which R730XD and Ceph Pool Configurations are Optimized for Capacity Archives - Server Throughput (MBps avg)

Which R730XD and Ceph Pool Configurations are Throughput Optimized - Server Throughput (MBps avg)