Red Hat Ceph Storage in BBVA

High Performance Workloads

Daniel Parkes
Senior Cloud Consultant, Iberia
06/05/2019
Spain’s second largest bank, **BBVA** have a broad global presence & innovative culture

- **€685 billion in total assets**
- **73 million customers**
- **>30 countries**
- **8,200 branches**
- **31,602 ATMs**
- **131,745 employees**

Data at the end of March 2018. Those countries in which BBVA has no legal entity or the volume of activity is not significant are not included.
BBVA. Why Red Hat Ceph Storage?
Key Storage Decisions
Factors

- Enterprise Class - High Performance
- Openstack Integration
- Multi-Geographic Distribution
- Secure Multitenancy
- Efficiency / Scalability
- Automation
Key Storage Decisions Factors

- Enterprise Class - High Performance
- Openstack Integration
- Multi-Geographic Distribution
- Secure Multitenancy
- Efficiency / Scalability
- Automation

- Cinder, Glance, Swift, Manila, Nova, Keystone
- Single Storage Layer
- Availability Zones - Regions
- Containerized Services
- Security
- Interoperability / API compatibility
Key Storage Decision Factors

- Enterprise Class - High Performance
- Openstack Integration
- Multi-Geographic Distribution
- Secure Multitenancy
- Efficiency / Scalability
- Automation

Bluestore

Next Generation performance flash-native

RBD, RGW, CephFS

RBD-mirroring

Bluestore compression

Erasure Coding

Ansible Driven

Cinder, Glance, Swift, Manila, Nova, Keystone

Single Storage Layer

Availability Zones - Regions

Containerized Services

Security

Interoperability / API compatibility
Partnership
Hardware Architecture
Choosing the right hardware configuration
<table>
<thead>
<tr>
<th><strong>Chassis</strong></th>
<th>5 x Cisco UCS C220-M5SN Rack Server</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU</strong></td>
<td>2 x Intel Xeon Platinum 8180. 28 core @ 2.50 GHz</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>12 x 16GB DIMM Modules(196 GB)</td>
</tr>
<tr>
<td><strong>NIC</strong></td>
<td>2 x Cisco UCS VIC 1387 40GB Dual Port</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>Data: 7x Intel® SSD DC P4500 4.0 TB</td>
</tr>
<tr>
<td></td>
<td>RocksDB/WAL: 1x Intel Optane SSD P4800X 375GB</td>
</tr>
<tr>
<td><strong>Software Configuration</strong></td>
<td>RHEL 7.6, Linux Kernel 3.10, RHCS 3.2(12.2.8)</td>
</tr>
</tbody>
</table>
# Client Hardware Configuration

<table>
<thead>
<tr>
<th><strong>Chassis</strong></th>
<th>7 x Cisco UCS B200 M4 Blade servers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU</strong></td>
<td>2x Intel® Xeon® CPU E5-2640 v4 @ 2.40GHz</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>528 GB</td>
</tr>
<tr>
<td><strong>NIC</strong></td>
<td>Cisco UCS VIC 1387 2 port (20Gb public network)</td>
</tr>
<tr>
<td><strong>Software Configuration</strong></td>
<td>RHOSP 10, RHEL 7.6, Linux Kernel 3.10, Pbench-FIO 3.3</td>
</tr>
</tbody>
</table>
Network Architecture
Software Architecture
RH Ceph Storage Configuration
Red Hat Ceph Storage 3.2 Configuration

- **Software versions:**
  - Red Hat Ceph Storage 3.2 (Luminous 12.2.8)
  - RHEL 7.6
  - Linux Kernel 3.10

- **Ceph-Ansible Containerized deployment**

- **3 Ceph nodes will have collocated MON + MGR + OSD services**

- **Set per OSD container Limit: 7 Vcpus/12Gb Ram**

- **2 OSDs per NVMe drive/ 70 OSDs**

- **WAL and RocksDB configured on Intel Optane P4800X drive**
Fitting Red Hat Ceph Storage in BBVA’s current Red Hat OpenStack
<table>
<thead>
<tr>
<th>Availability Zones</th>
<th>Openstack</th>
<th>IaaS Overlay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cinder Storage AZ1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cinder Storage AZ2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cinder Storage AZ3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#redhat #rhsummit
Availability Zones Openstack IaaS Overlay

- Cinder Storage AZ1
  - Cinder Standard type

- Cinder Storage AZ2
  - Cinder Standard type

- Cinder Storage AZ3
  - Cinder Standard type

- Cinder NFS Backend
- Cinder Standard Type
- Sata Disk Storage
Availability Zones Openstack IaaS Overlay

Cinder Storage AZ1
- Cinder Standard Type
- Sata Disk Storage

Cinder Storage AZ2
- Cinder Standard Type
- Cinder High Performance Type

Cinder Storage AZ3
- Cinder Standard Type
- Cinder High Performance Type

Red Hat Ceph Storage All-Flash Cluster

Cinder AZ1 Pool Crush Rule

Cinder AZ2 Pool Crush Rule

Cinder AZ3 Pool Crush Rule

- Cinder NFS Backend
- Cinder Standard Type
- Sata Disk Storage

- Cinder RBD Backend
- Cinder High Perf Type
- All-Flash Disk Storage
- IOPS/Low Latency
Red Hat Ceph Storage 3.2 Performance
Bluestore on All Flash Clusters Results
Performance Testing Methodology

- 2 OSDs per NVMe drive.

- **Ceph Storage Pool Config.**
  - 2x Replication 4096 PGs. 105x200Gb RBD images. \( \sim 2TB \times 2 = 4 \) TBytes.

- **RBD block tests where run using FIO RBD IOengine.**
  - Pbench-fio version 3.3 used to generate the load.
  - 1 RBD image per client, 105 clients spread among the 7 hypervisors available.
  - 3 workloads used: Random Read, Random Write and Mixed 70% Read/30% Write

- **RBD block test duration and execution.**
  - The RBD images were pre-conditioned writing the full size of each volume.
  - Each tests was run 4 times during 10 minutes.
  - The results presented are the average of these 4 runs.
Peak Performance With Small Block Workloads (4kb)

**RHCS 3.2 BlueStore on All-Flash: IOPS vs. Average Latency vs. IO-Depth**

5 x Ceph Nodes | 4KB Block Size | 105 x RBD Volumes

- **IO-Depth 64**
  - RR. 2.2 Million IOPS@3ms average latency
  - RW. 463K IOPS@11ms average latency
  - MIXED. 691K IOPS@5.8ms average latency

- **IO-Depth 32**
  - RR. 2 Million IOPS@1.8ms average latency
  - RW. 415K IOPS@6.2ms average latency
  - MIXED. 611K IOPS@3.2ms average latency

#redhat #rhsummit
Small Block (4k) Workload CPU Utilization

RHCS 3.2 BlueStore on All-Flash: IOPS vs. CPU Utilisation vs. IO-Depth

5 x Ceph Nodes | 4KB Block Size | 105 x RBD Volumes

<table>
<thead>
<tr>
<th>IO-Depth 64</th>
<th>RR. 92% CPU usage</th>
<th>RW. 83% CPU usage</th>
<th>MIXED. 83% CPU usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>IO-Depth 32</td>
<td>RR. 89% CPU usage</td>
<td>RW. 76% CPU usage</td>
<td>MIXED. 75% CPU usage</td>
</tr>
</tbody>
</table>

#redhat #rhsummit
Peak Throughput With Big Block Workloads (1MB/4MB)

**Big Block 4MB throughput**
- **RR.** Limited to 10Gbytes/s by network
- **RW.** Limited to 10Gbytes/s by network
- **MIXED.** Limited to 10Gbytes/s by network

**Big Block 1MB throughput**
- **RR.** Limited to 10Gbytes/s by network
- **RW.** 6.2 Gigabytes/s
- **MIXED.** 8.5 Gigabytes/s

RHCS 3.2 BlueStore on All-Flash: Throughput vs. Media Utilisation vs. CPU Utilisation

5 x Ceph Nodes | 1MB / 4MB | IODepth 32 | 105 x RBD Volumes

- Throughput GB/s:
  - Random Write: 6.2 GB/s
  - Random ReadWrite (70/30): 8.5 GB/s
  - Random Read: 6.2 GB/s

- CPU & Media Utilisation:
  - Random Write: 13%
  - Random ReadWrite (70/30): 13%
  - Random Read: 13%
RHCS 3.2 Bluestore performance Scalability

Scale Out. IOPS results
RR. 55% Increase with 5 nodes
RW. 90% Increase with 5 nodes
MIXED. 77% Increase with 5 nodes

Scale Out. Latency results
RR. 29% Lower latency
RW. 40% Lower latency
MIXED. 46% Lower latency
How Many OSDs Per Drive 2 or 4?

4 OSD higher CPU Percentage Utilization

Very similar IOPS and Latency results.

#redhat #rhsummit
Performance Boost Using Intel Optane P4800X for the RocksDB/WAL device

RocksDB/WAL with Optane

IOPS, 8KB Block Size
RR. 29% Increase in IOPS
RW. 37% Increase in IOPS
MIXED. 51% Increase in IOPS

Avg Latency, 8KB Block Size
RR. 17% Lower latency
RW. 27% Lower latency
MIXED. 43% Lower latency

Tail Latency, 8KB Block Size
RR. 75% Lower latency
RW. 71% Lower latency
MIXED. 51% Lower latency
SQL workload: All-Flash vs Spinning Drives

MySQL on Ceph RBD: Write Queries Per Seconds (QPS) & 99th Percentile Tail Latency (ms)

60 x HDD Ceph Cluster, 42 x NVMe Ceph Cluster, 10 x MySQL Instances, 1 x 100G Cinder RBD per MySQL Instance

- Ceph Bluestore on HDD Media
- Ceph Bluestore on Flash Media
- Ceph Bluestore on HDD P99 Tail Lat.
- Ceph Bluestore on Flash Media P99 Tail Lat.

Queries Per Second (QPS)
Tail Latency (ms)
BlueStore (Default vs. Tuned) Performance Comparison
red.ht/RHCS-Bluestore-Performance-Blog1
Configuration Files Used During Benchmarking

Red Hat Ceph Storage ceph.conf file used during tests
red.ht/ceph-conf

Red Hat Ceph Storage ceph-ansible group_vars/all.yml
red.ht/ceph-ansible-conf

Red Hat Enterprise Linux custom Tuned profile
red.ht/rhel-tuned-conf

Flexible I/O tester configuration template file
red.ht/fio-template-conf
FIND US AT RED HAT SUMMIT

- At the Storage lockers
- At the Red Hat booth
- At one of Storage dedicated sessions (red.ht/storageatsummit)
- At the Community Happy Hour (Tues 6:30, Harpoon Brewery)
- At the Hybrid Cloud Party (Wed, 7:30, “Committee” restaurant)

redhat.com/storage
@redhatstorage
redhatstorage.redhat.com

Red Hat OpenShift Container Storage
red.ht/videos-RHOCS

Red Hat data analytics infrastructure solution
red.ht/videos-RHDAIS

Red Hat Hyperconverged Infrastructure
red.ht/videos-RHHI
THANK YOU

linkedin.com/company/Red-Hat

facebook.com/RedHatInc

youtube.com/user/RedHatVideos
twitter.com/RedHat