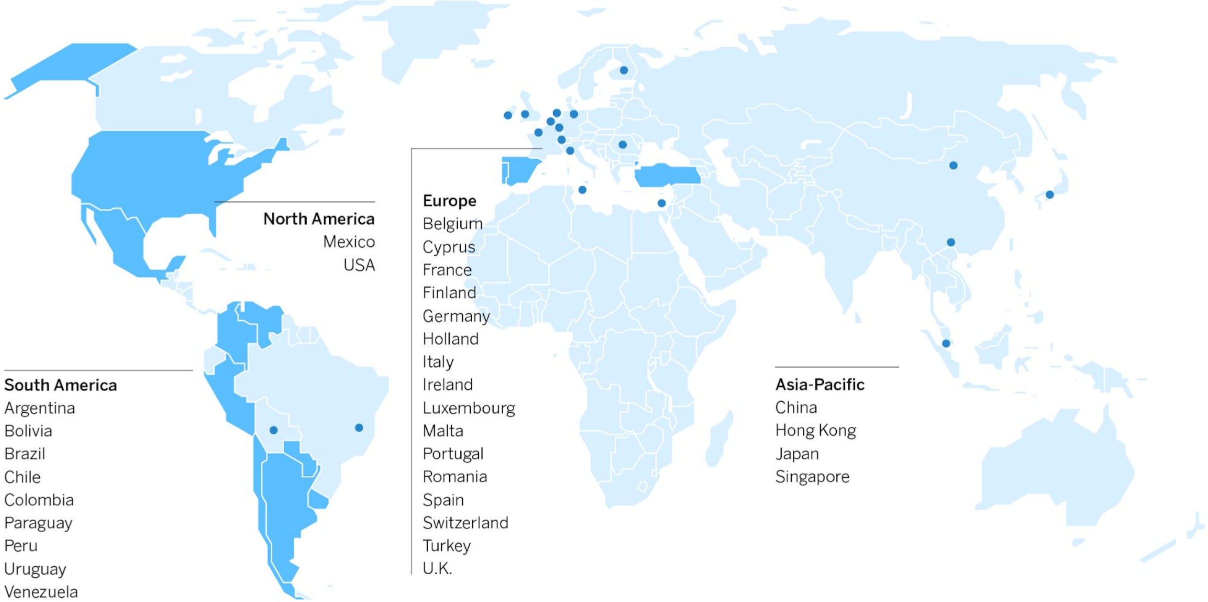
Red Hat Ceph Storage in BBVA

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**



North America
Mexico
USA

Europe
Belgium
Cyprus
France
Finland
Germany
Holland
Italy
Ireland
Luxembourg
Malta
Portugal
Romania
Spain
Switzerland
Turkey
U.K.

South America
Argentina
Bolivia
Brazil
Chile
Colombia
Paraguay
Peru
Uruguay
Venezuela

Asia-Pacific
China
Hong Kong
Japan
Singapore

€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

Six Strategic Priorities

- New standard in customer experience
- Digital sales
- New business models
- Optimize capital allocation
- Unrivaled efficiency
- A first class workforce

BBVA. Why Red Hat Ceph Storage?

BBVA

Key Storage Decisions Factors

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

BBVA

Key Storage Decisions Factors

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

openstack

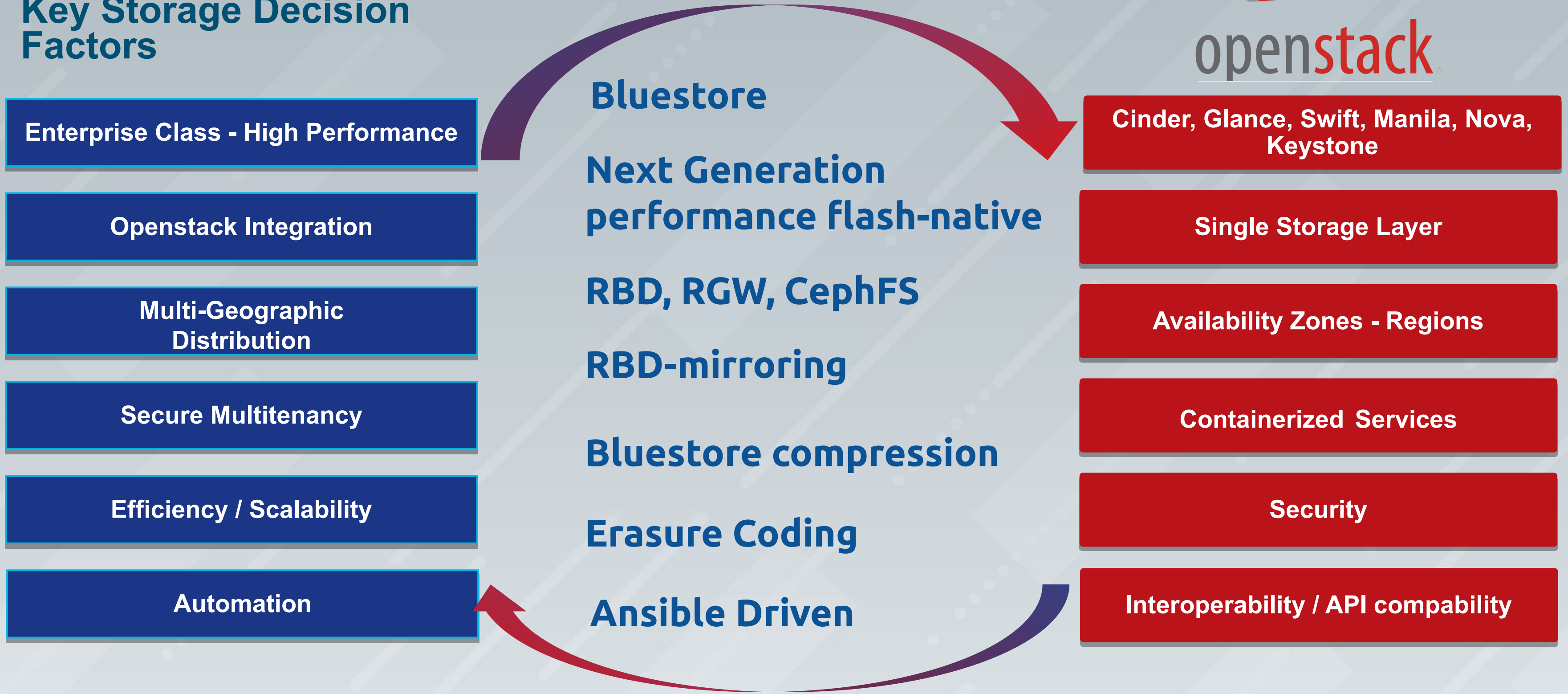
- Cinder, Glance, Swift, Manila, Nova, Keystone
- Single Storage Layer
- Availability Zones - Regions
- Containerized Services
- Security
- Interoperability / API compatibility

BBVA

ceph

openstack

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

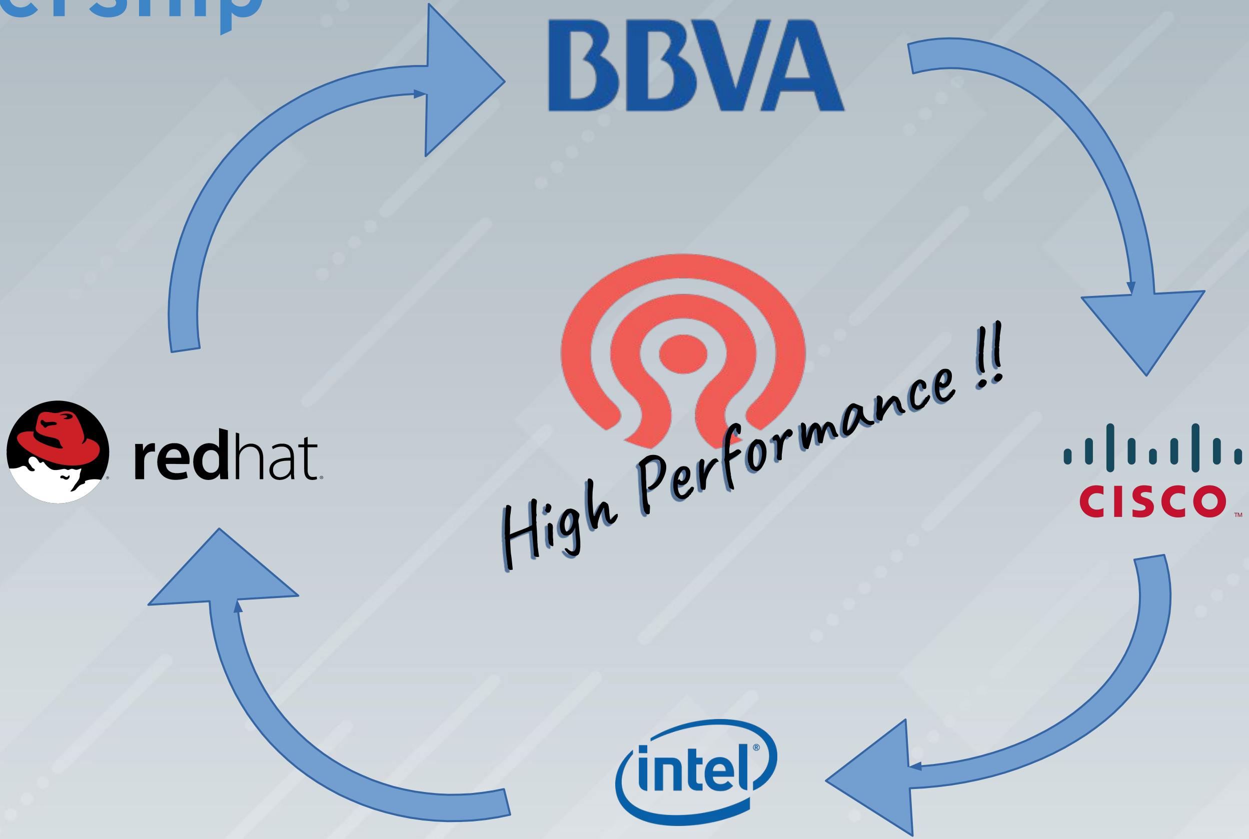
Availability Zones - Regions

Containerized Services

Security

Interoperability / API compability

Partnership



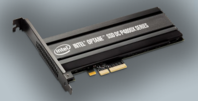
High Performance !!

Hardware Architecture

Choosing the right hardware configuration

Red Hat Ceph Storage Node Configuration

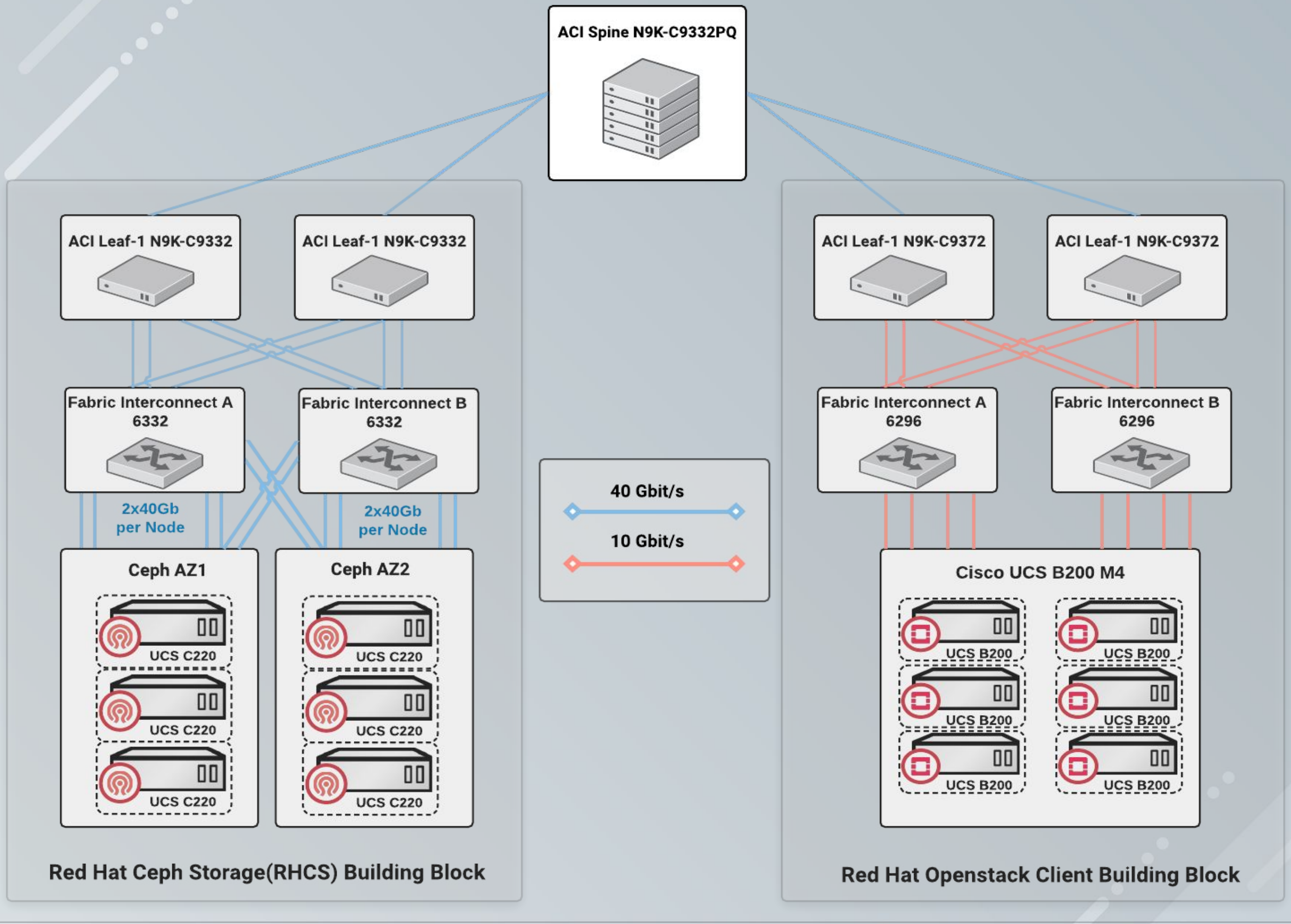
Chassis	5 x Cisco UCS C220-M5SN Rack Server
CPU	2 x Intel Xeon Platinum 8180. 28 core @ 2.50 GHz
Memory	12 x 16GB DIMM Modules(196 GB)
NIC	2 x Cisco UCS VIC 1387 40GB Dual Port
Storage	Data: 7x Intel® SSD DC P4500 4.0 TB RocksDB/WAL: 1x Intel Optane SSD P4800X 375GB
Software Configuration	RHEL 7.6, Linux Kernel 3.10, RHCS 3.2(12.2.8)



Client Hardware Configuration

Chassis	7 x Cisco UCS B200 M4 Blade servers
CPU	2x Intel® Xeon® CPU E5-2640 v4 @ 2.40GHz
Memory	528 GB
NIC	Cisco UCS VIC 1387 2 port (20Gb public network)
Software Configuration	RHOSP 10, RHEL 7.6, Linux Kernel 3.10, Pbench-FIO 3.3

Network Architecture


ACI Spine N9K-C9332PQ
ACI Leaf-1 N9K-C9332
ACI Leaf-1 N9K-C9332
Fabric Interconnect A 6332
Fabric Interconnect B 6332
2x40Gb per Node
2x40Gb per Node
Ceph AZ1
UCS C220
UCS C220
UCS C220
Ceph AZ2
UCS C220
UCS C220
UCS C220
Red Hat Ceph Storage(RHCS) Building Block
40 Gbit/s
10 Gbit/s
ACI Leaf-1 N9K-C9372
ACI Leaf-1 N9K-C9372
Fabric Interconnect A 6296
Fabric Interconnect B 6296
Cisco UCS B200 M4
UCS B200
UCS B200
UCS B200
UCS B200
UCS B200
UCS B200
Red Hat Openstack Client Building Block

Software Architecture

RH Ceph Storage Configuration

Red Hat Ceph Storage 3.2 Configuration ceph

❑ **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**

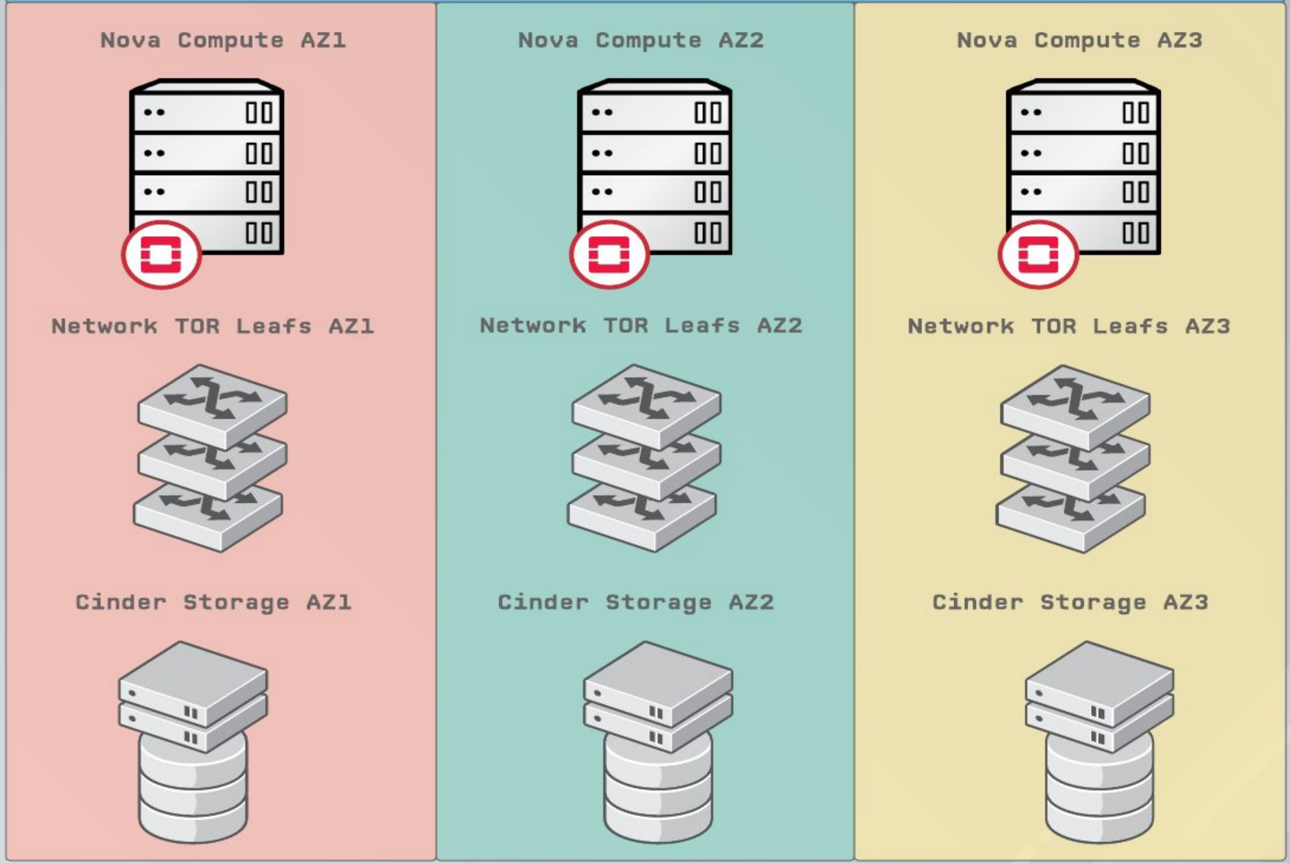
❑ **Set per OSD container Limit: 7 Vcpus/12Gb Ram**

❑ **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

Availability Zones Openstack IaaS Overlay



Nova Compute AZ1

Network TOR Leafs AZ1

Cinder Storage AZ1

Nova Compute AZ2

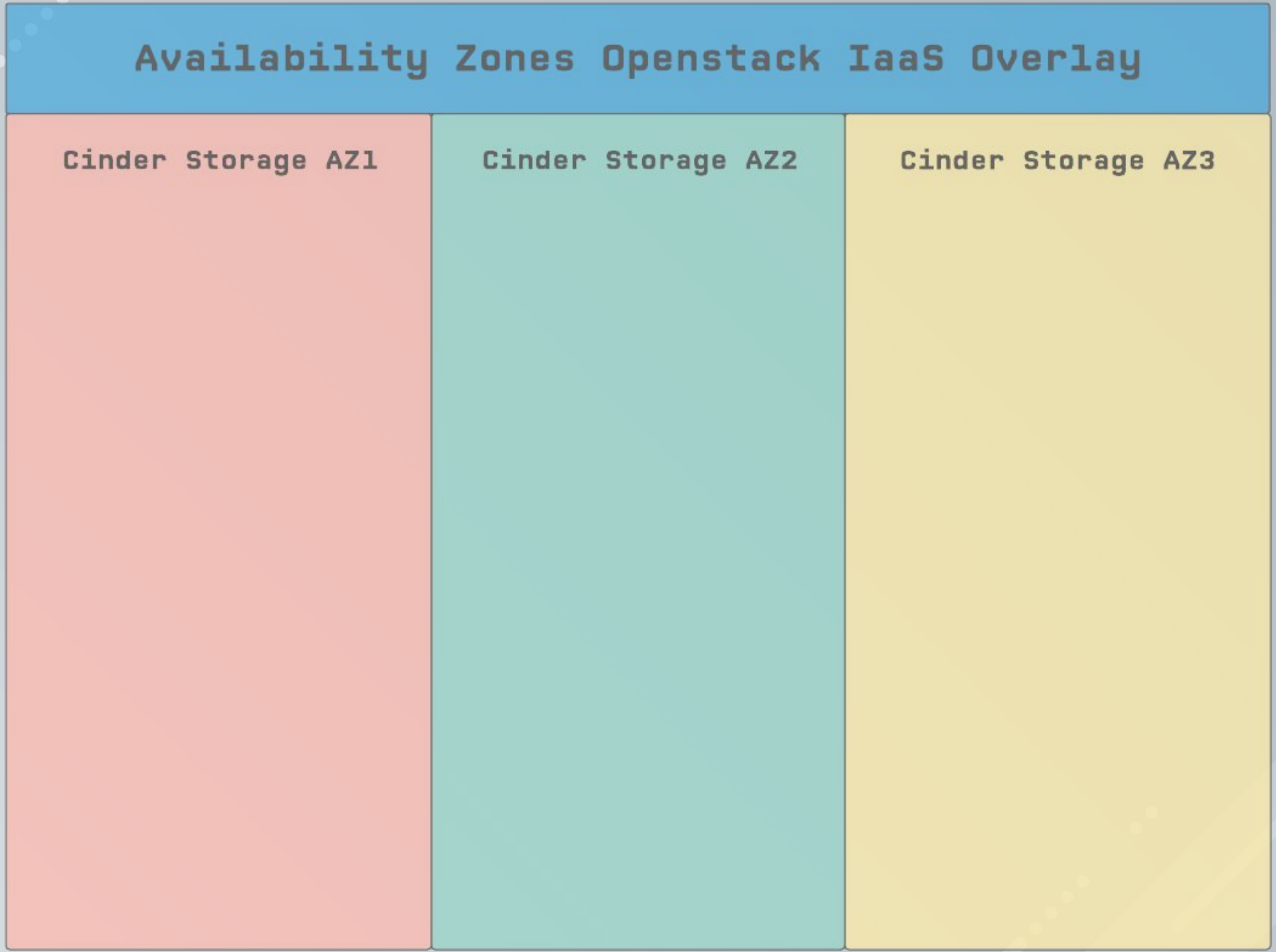
Network TOR Leafs AZ2

Cinder Storage AZ2

Nova Compute AZ3

Network TOR Leafs AZ3

Cinder Storage AZ3



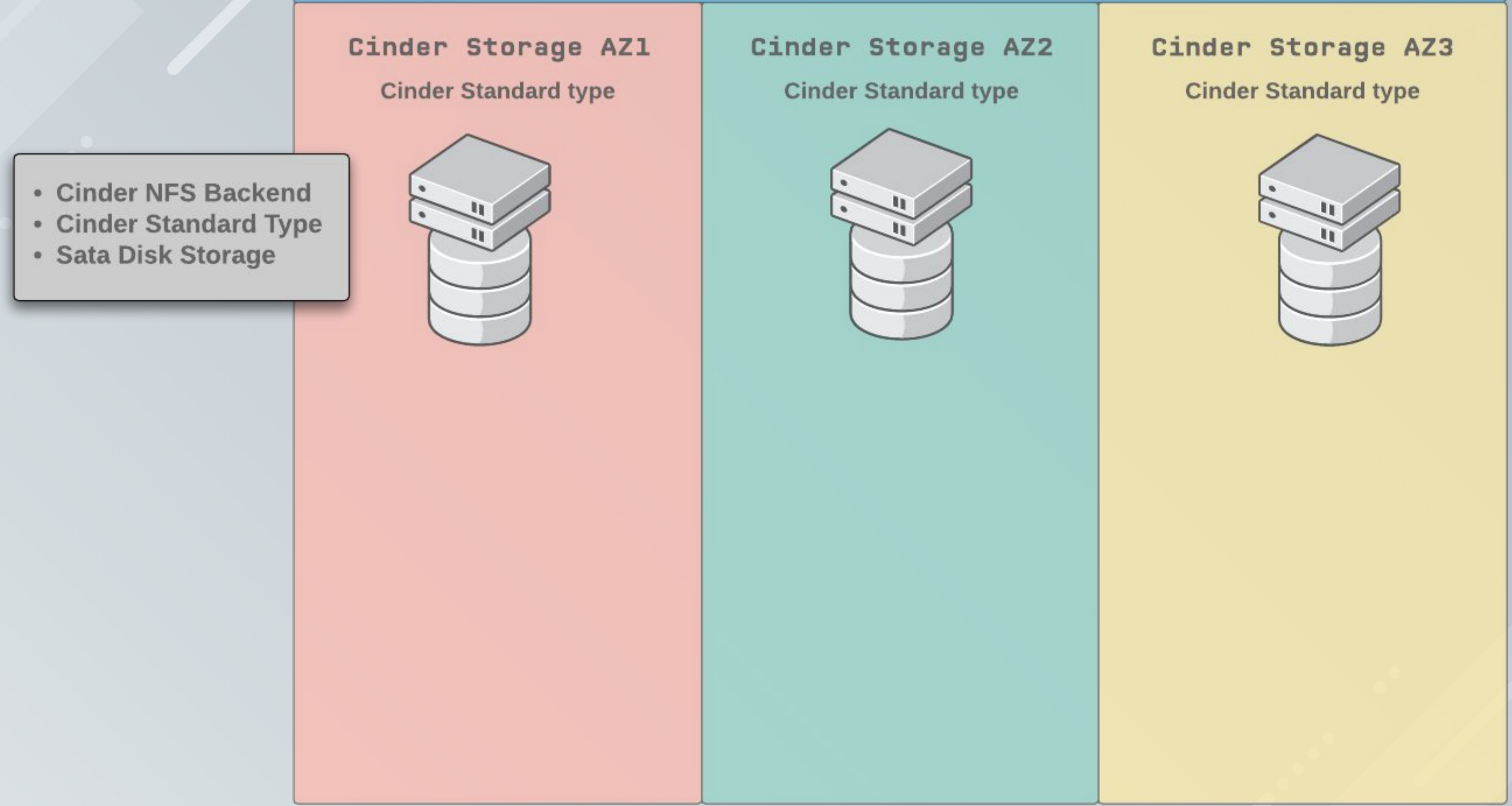
Availability Zones Openstack IaaS Overlay

Cinder Storage AZ1

Cinder Storage AZ2

Cinder Storage AZ3

Availability Zones Openstack IaaS Overlay



Cinder Storage AZ1

Cinder Standard type

Cinder Storage AZ2

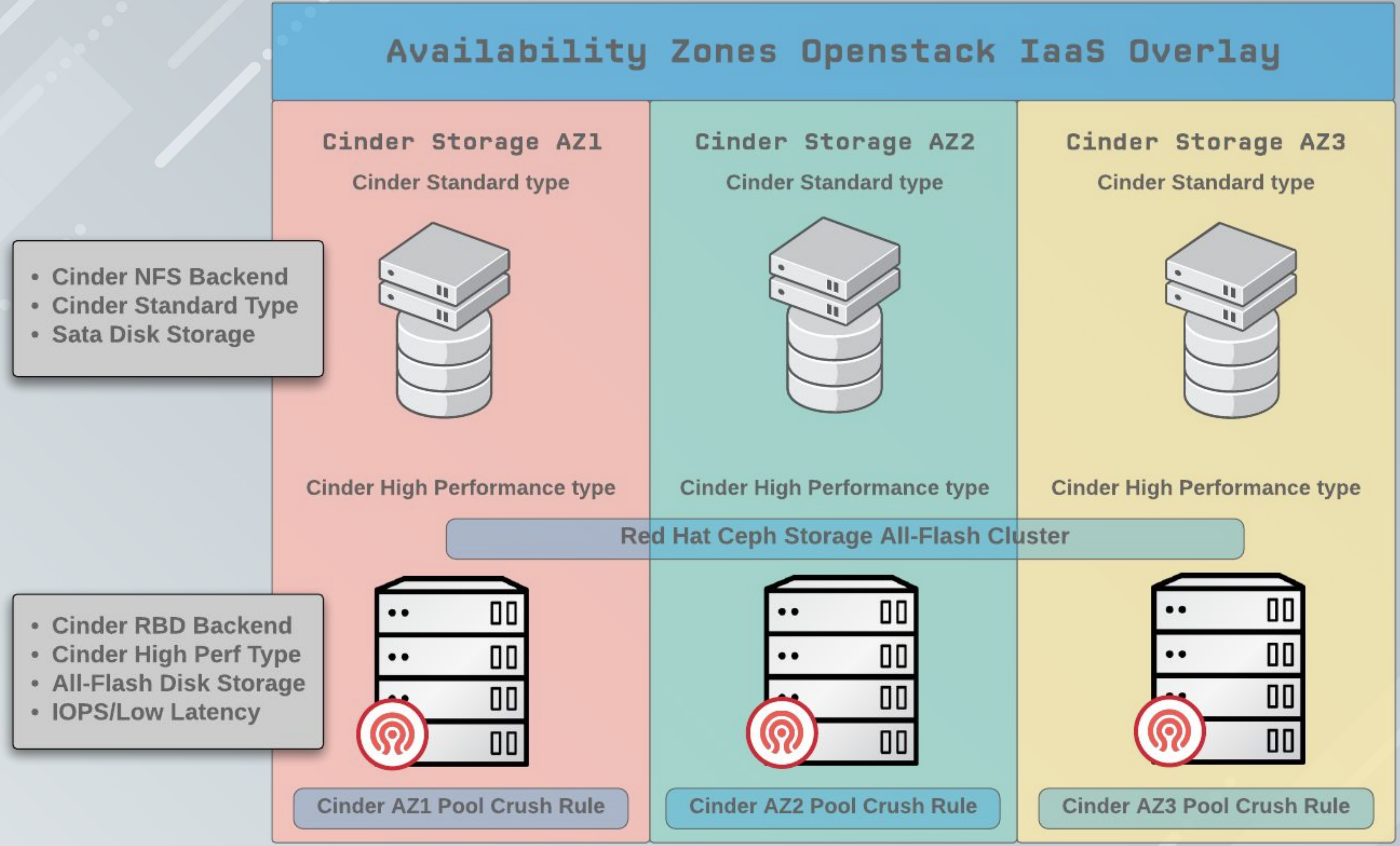
Cinder Standard type

Cinder Storage AZ3

Cinder Standard type

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

Availability Zones Openstack IaaS Overlay



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

- Cinder RBD Backend
- Cinder High Perf Type
- All-Flash Disk Storage
- IOPS/Low Latency

Cinder Storage AZ1	Cinder Storage AZ2	Cinder Storage AZ3
Cinder Standard type	Cinder Standard type	Cinder Standard type
Cinder High Performance type	Cinder High Performance type	Cinder High Performance type
Cinder AZ1 Pool Crush Rule	Cinder AZ2 Pool Crush Rule	Cinder AZ3 Pool Crush Rule

Red Hat Ceph Storage All-Flash Cluster

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. ~2TB x 2 = 4 TBytes.

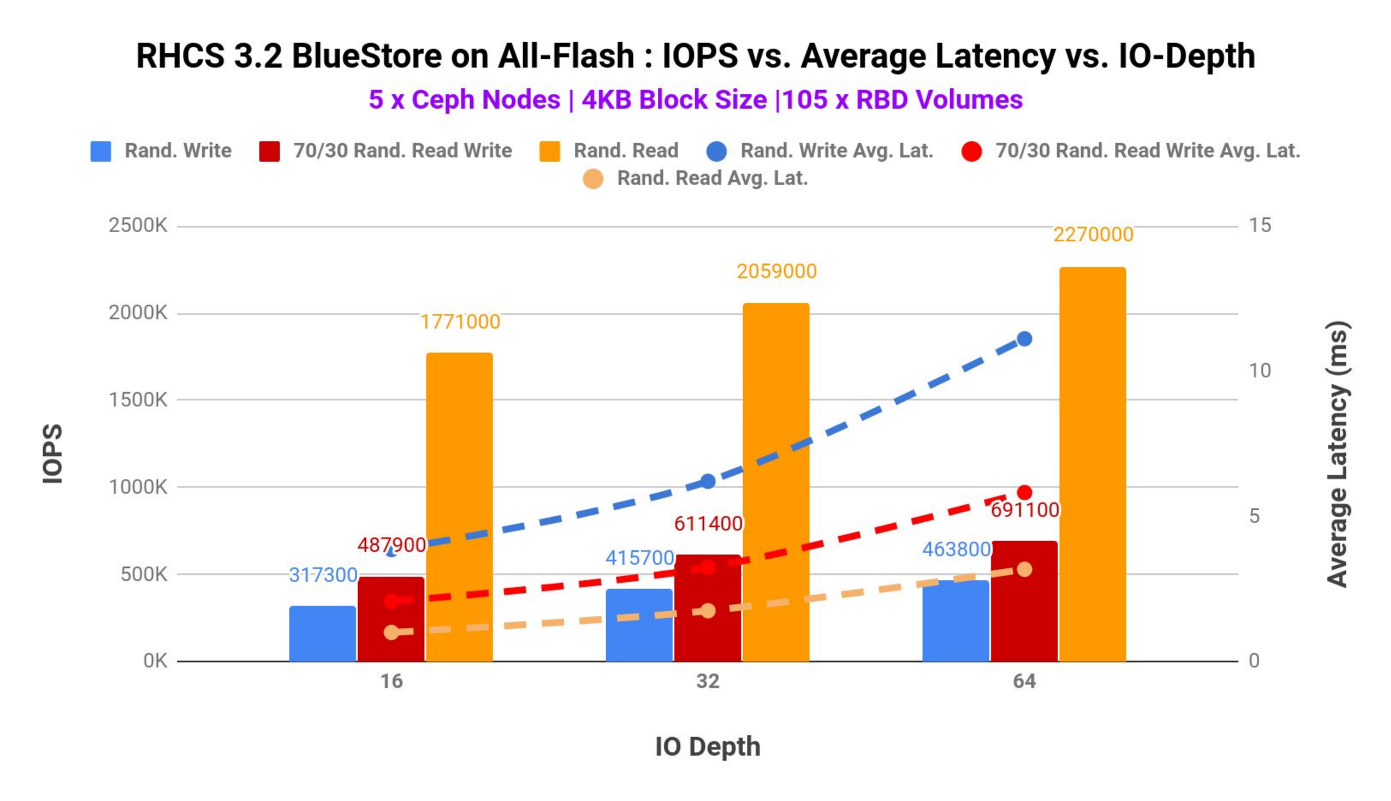
❑ **RBD block tests where run using FIO RBD IOengine.**

- 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)



IO-Depth 64

RR. **2.2 Million IOPS@3ms** average latency

RW. **463K IOPS@11ms** average latency

MIXED. **691K IOPS@5.8ms** average latency

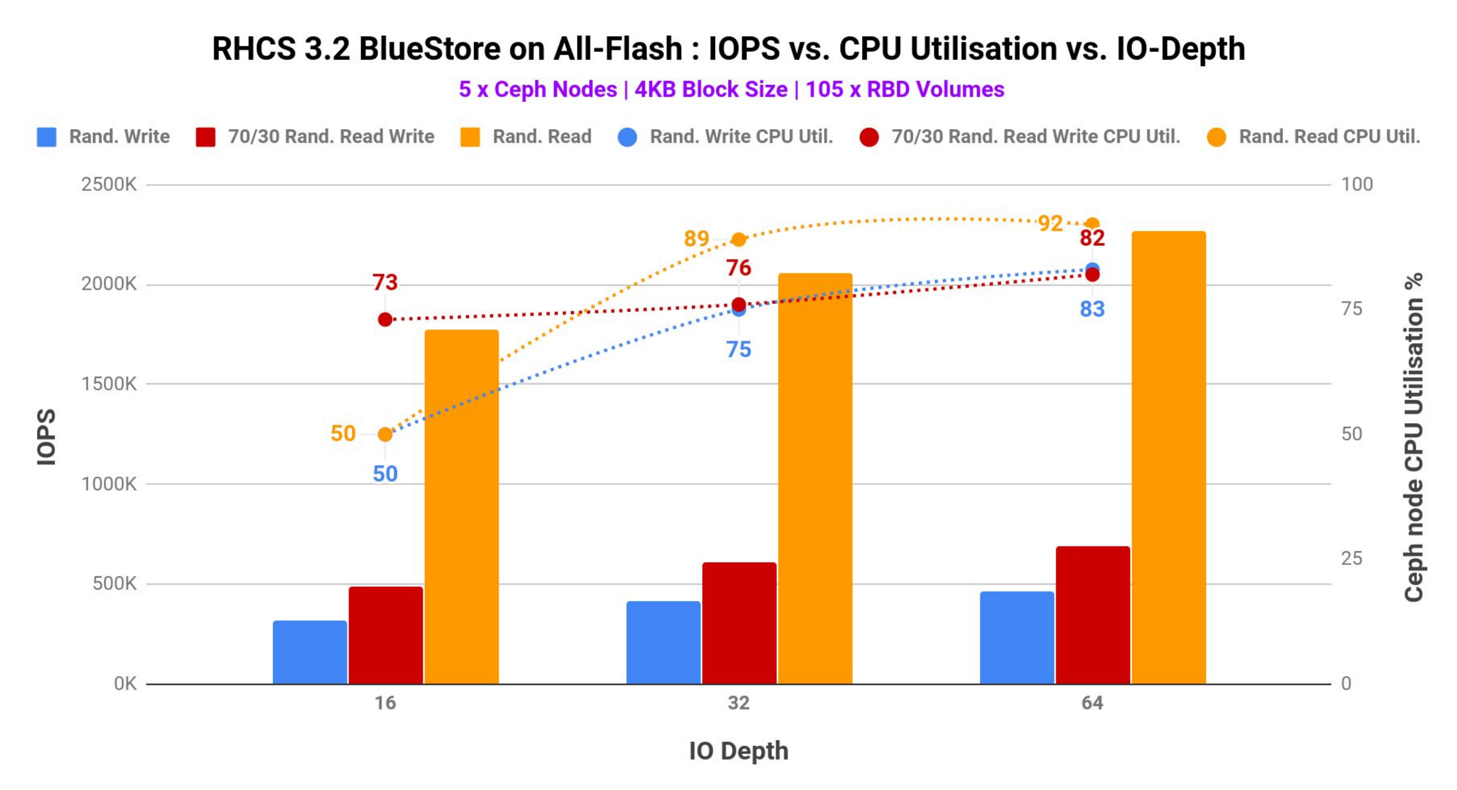
IO-Depth 32

RR. **2 Million IOPS@1.8ms** average latency

RW. **415K IOPS@6.2ms** average latency

MIXED. **611K IOPS@3.2ms** average latency

Small Block(4k) Workload CPU Utilization



IO-Depth 64

RR. **92% CPU usage**

RW. **83% CPU usage**

MIXED. **83% CPU usage**

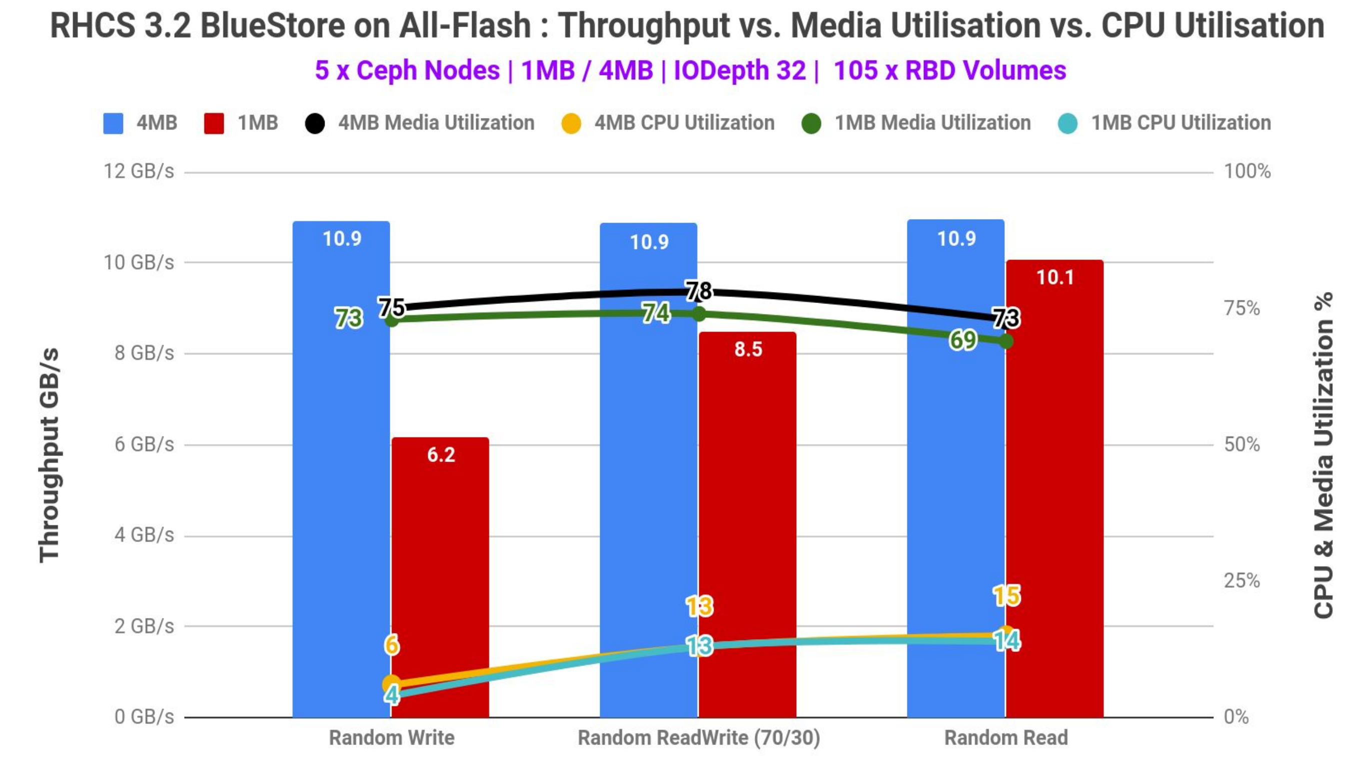
IO-Depth 32

RR. **89% CPU usage**

RW. **76% CPU usage**

MIXED. **75% CPU usage**

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



Big Block 4MB troughput

RR. **Limited to 10Gbytes/s** by network

RW. **Limited to 10Gbytes/s** by network

MIXED. **Limited to 10Gbytes/s** by network

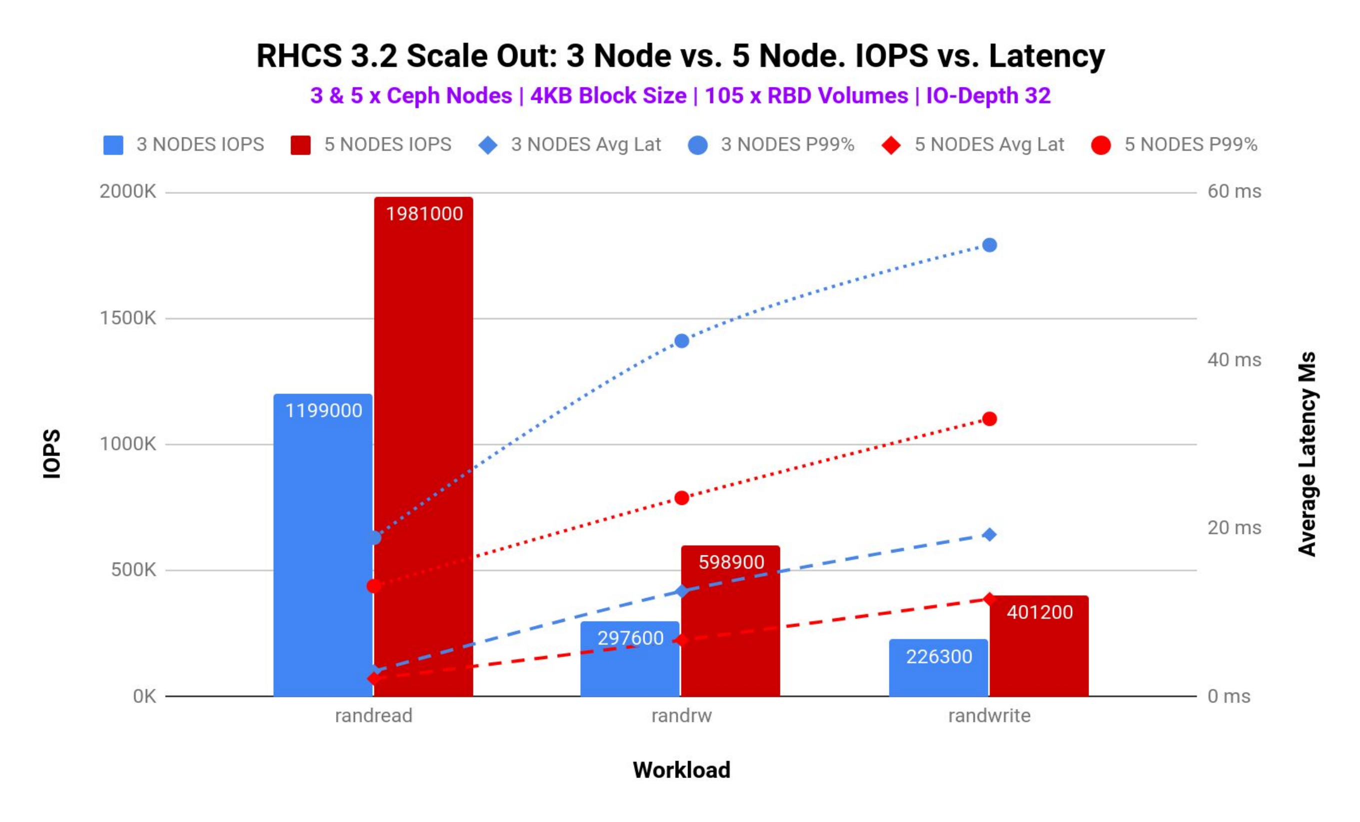
Big Block 1MB troughput

RR. **Limited to 10Gbytes/s by network**

RW. **6.2 Gigabytes/s**

MIXED. **8.5 Gigabytes/s**

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

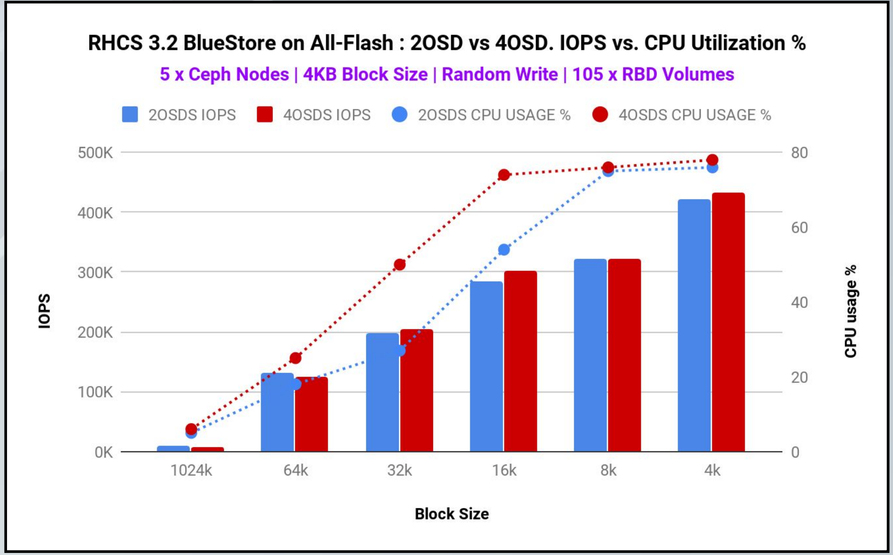
Scale Out. Latency results

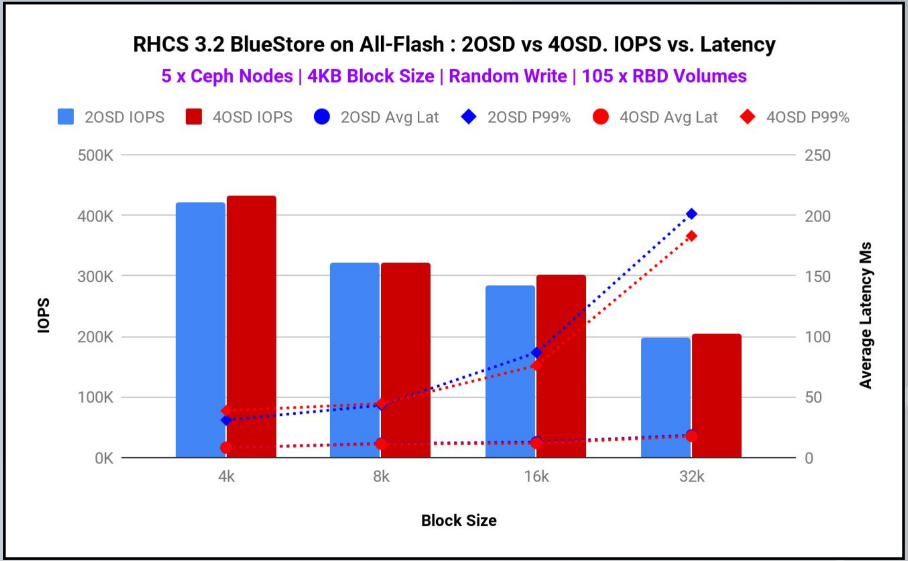
RR. **29% Lower** latency

RW. **40% Lower** latency

MIXED. **46% Lower** latency

How Many OSDs Per Drive 2 or 4 ?

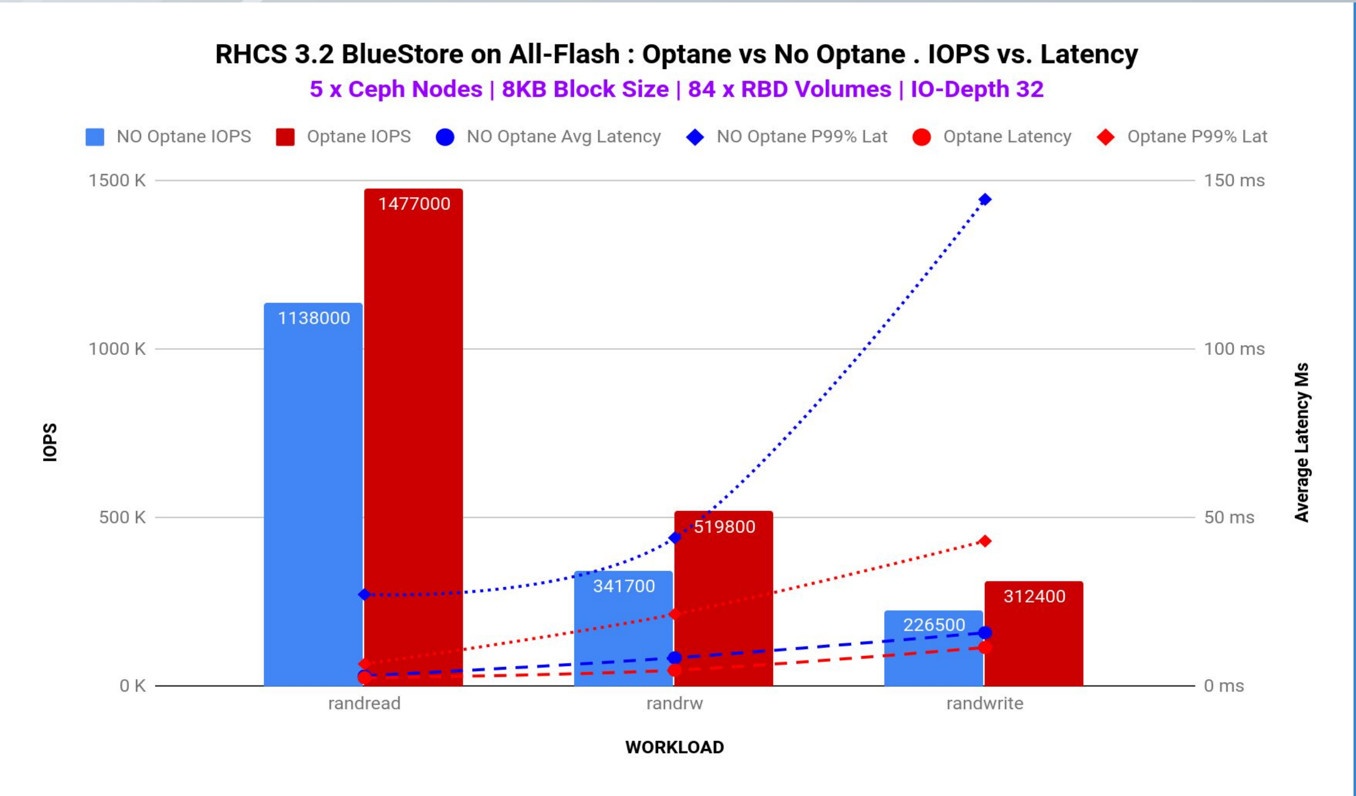




4 OSD higher CPU Percentage Utilization

Very similar IOPS and Latency results.

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

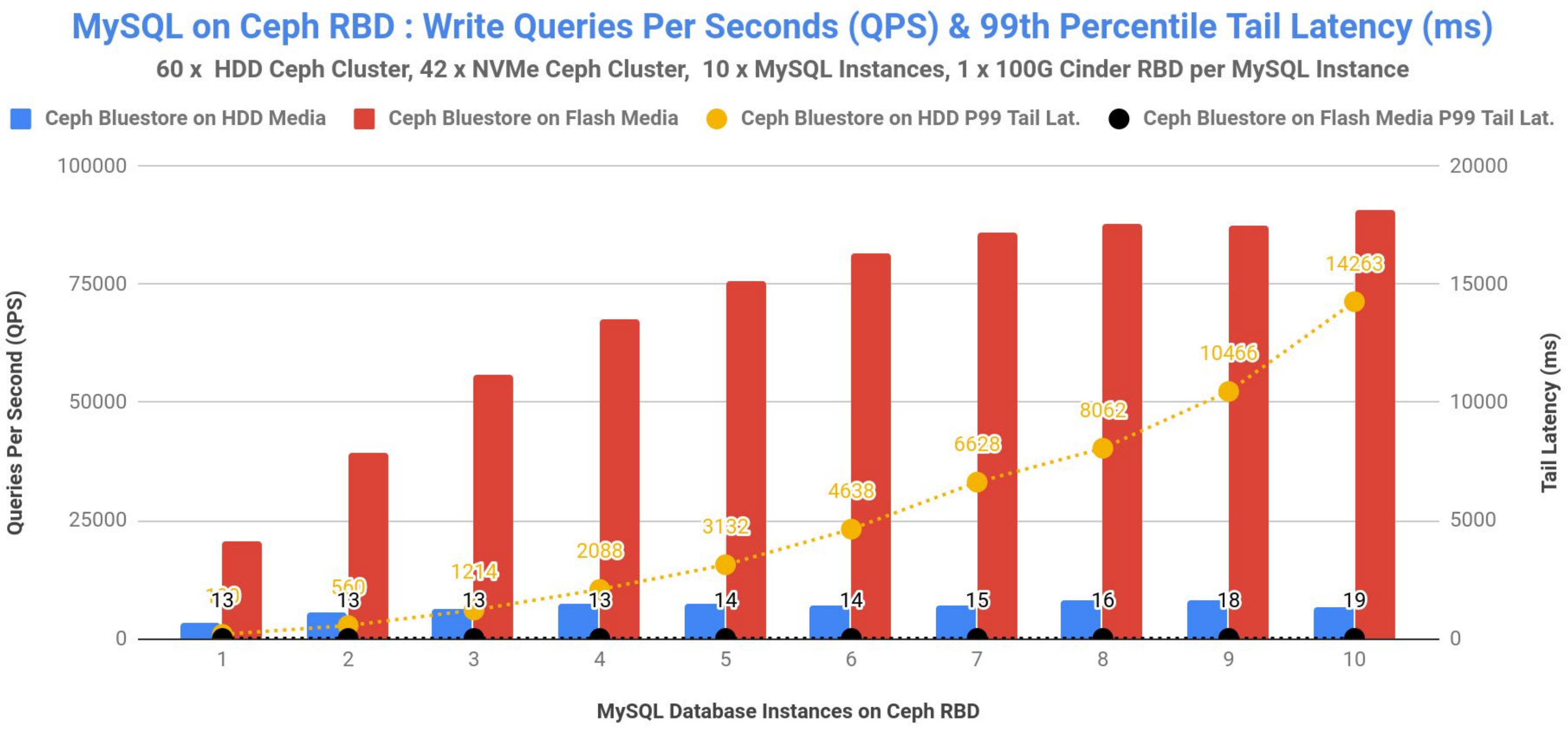
Avg Latency. 8KB Block Size

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

RHCS on All-Flash Cluster: Performance Blog Series

BlueStore (Default vs. Tuned) Performance Comparison
red.ht/RHCS-Bluestore-Perfomance-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

redhat.com/storage

@redhatstorage

redhatstorage.redhat.com

- 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)

▶ **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

RED HAT SUMMIT

THANK YOU

in linkedin.com/company/Red-Hat

f facebook.com/RedHatinc

YouTube youtube.com/user/RedHatVideos

twitter.com/RedHat