IDF15

Ceph: Open Source Storage Software Optimizations on Intel® Architecture for Cloud Workloads

Jian Zhang – Software Engineer, Intel Corporation

DATS005

Make the Future with China!

Agenda

- The Problem
- Ceph Introduction
- Ceph Performance
- Ceph Cache Tiering and Erasure Code
- Intel Product Portfolio for Ceph
- Ceph Best Practices
- Summary

Agenda

- The Problem
- Ceph Introduction
- Ceph Performance
- Ceph Cache Tiering and Erasure Code
- Intel Product Portfolio for Ceph
- Ceph Best Practices
- Summary

The Problem: Data Big Bang

From 2013 to 2020, the digital universe will grow by a factor of 10, from 4.4 ZB to 44 ZB

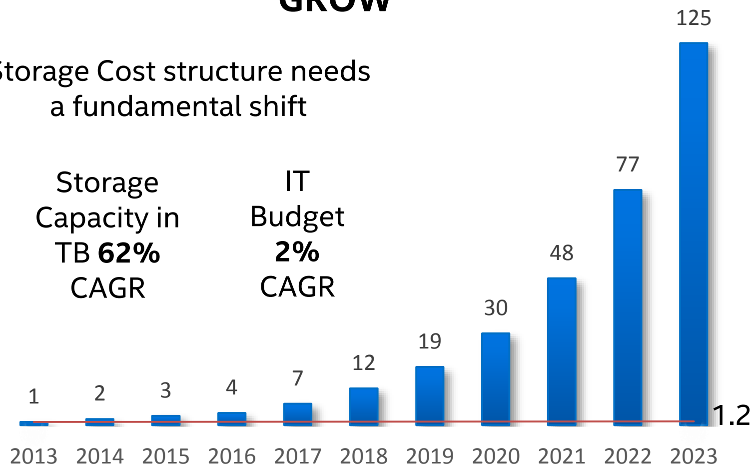
It more than doubles every two years.

COST CHALLENGES COTINUE TO **GROW**

Storage Cost structure needs a fundamental shift

Storage Capacity in TB **62%** CAGR

IT Budget **2%** CAGR



IT PROS WILL SHOULDER A GREATER **STORAGE BURDEN**

230 GB Per IT Pro

1,231 GB Per IT Pro

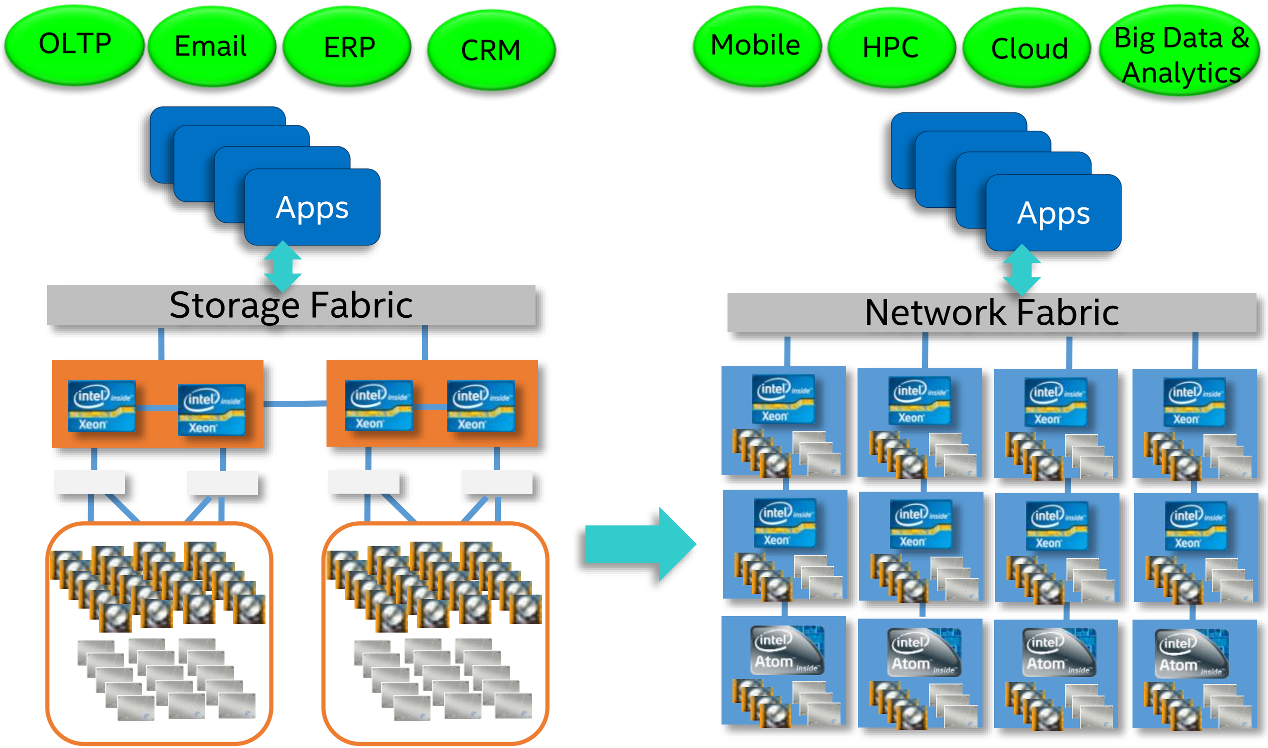
Data needs are growing at a rate unsustainable with today's infrastructure and labor costs

Source: IDC – The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things - April 2014

Diverse Workloads & Cost Drive Need for Distributed Storage

Traditional Workloads

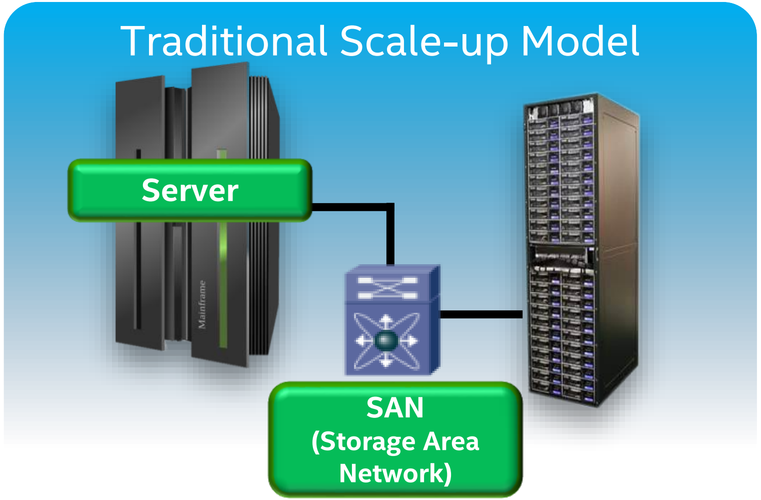
Today's Trends

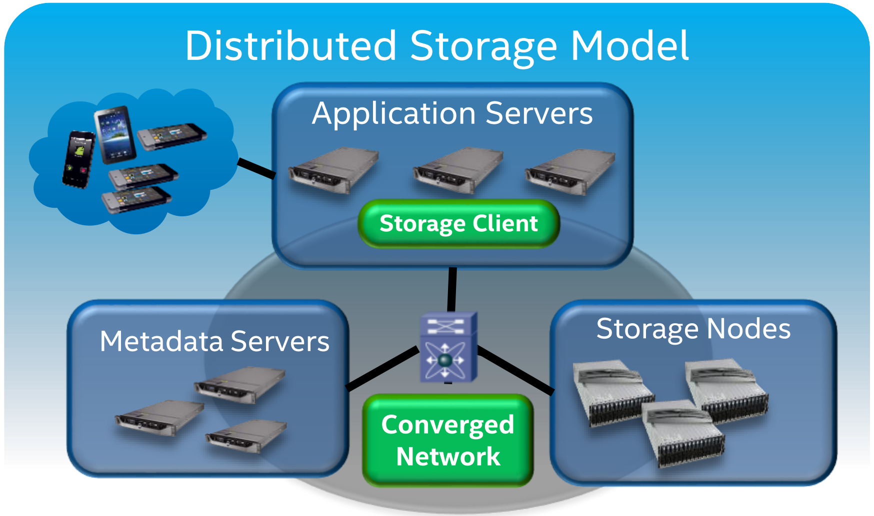


Challenges

- Cost
- Diverse Workloads
- Scale on Demand
- Increasing Complexity
- Management

Distributed Storage





- \+ High Availability (Failover)
- \+ High perf workloads (e.g., database)
- \+ Enterprise Mission Critical Hybrid Cloud
- − Limited Scale
- − Costly (Cap-ex and Op-ex)

- \+ Pay as you Grow, massive on-demand scale
- \+ Cost, Performance optimized
- \+ Open and commercial solutions on x86 servers
- \+ Applicable to cloud workloads
- − Not a good fit for traditional high perf workloads

- **Ceph is the most popular† open source virtual block storage option. Also provides object, *file (experimental)*.**
- **Strong customer interest - several production implementations already.**

† OpenStack User Survey Insights: http://superuser.openstack.org/articles/openstack-user-survey-insights-november-2014

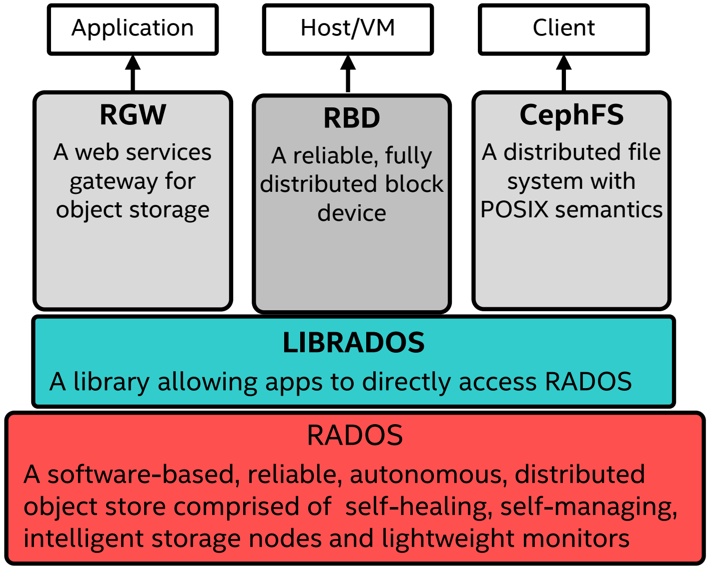
IDF15

Agenda

- The Problem
- Ceph Introduction
- Ceph Performance
- Ceph Cache Tiering and Erasure Code
- Intel Product Portfolio for Ceph
- Ceph Best Practices
- Summary

Ceph Introduction

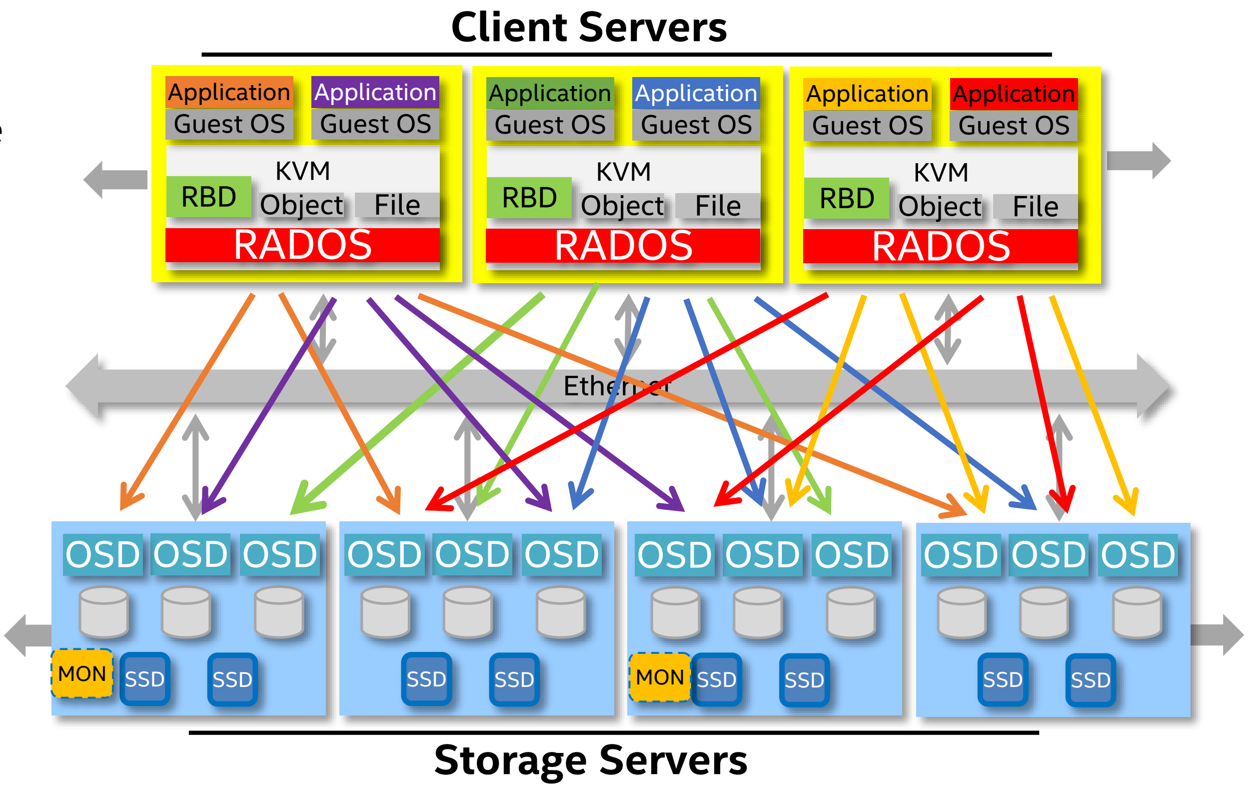
- Ceph is an open-source, massively scalable, software-defined storage system which provides object, block and file system storage in a single platform. It runs on commodity hardware—saving you costs, giving you flexibility—and because it's in the Linux* kernel, it's easy to consume.
- Object Store (RADOSGW)
  - A bucket based REST gateway
  - Compatible with S3 and swift
- File System (CEPH FS)
  - A POSIX-compliant distributed file system
  - Kernel client and FUSE
- Block device service (RBD)
  - OpenStack* native support
  - Kernel client and QEMU/KVM driver



IDF15

Ceph Cluster Overview

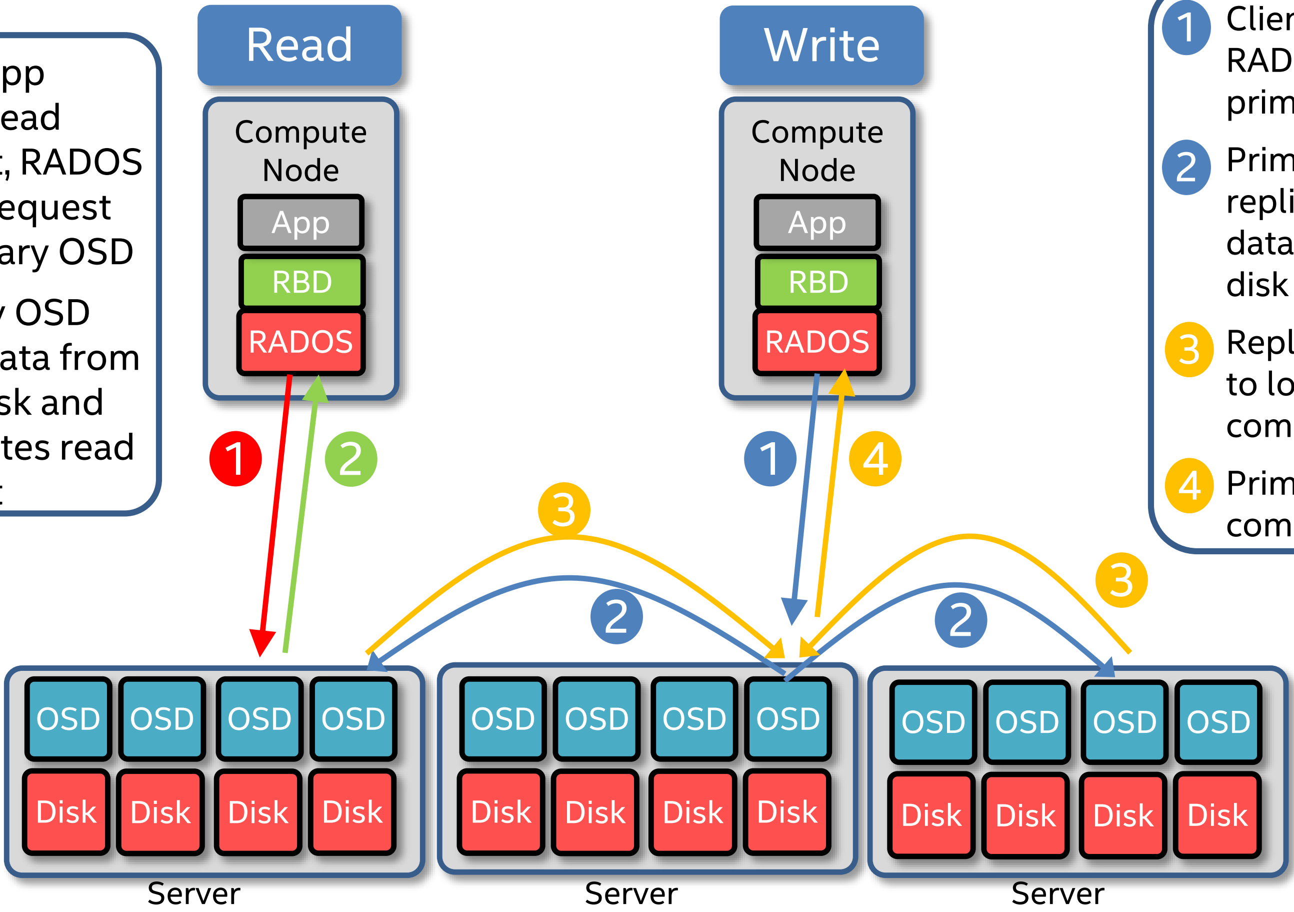
- **Ceph Clients**
  - Block/Object/File system storage
  - User space or kernel driver
- **Peer to Peer via Ethernet**
  - Direct access to storage
  - No centralized metadata = no bottlenecks
- **Ceph Storage Nodes**
  - Data distributed and replicated across nodes
  - No single point of failure
  - Scale capacity and performance with additional nodes



Ceph scales to 1000s of nodes

Object Store Daemon (OSD) Read and Write Flow

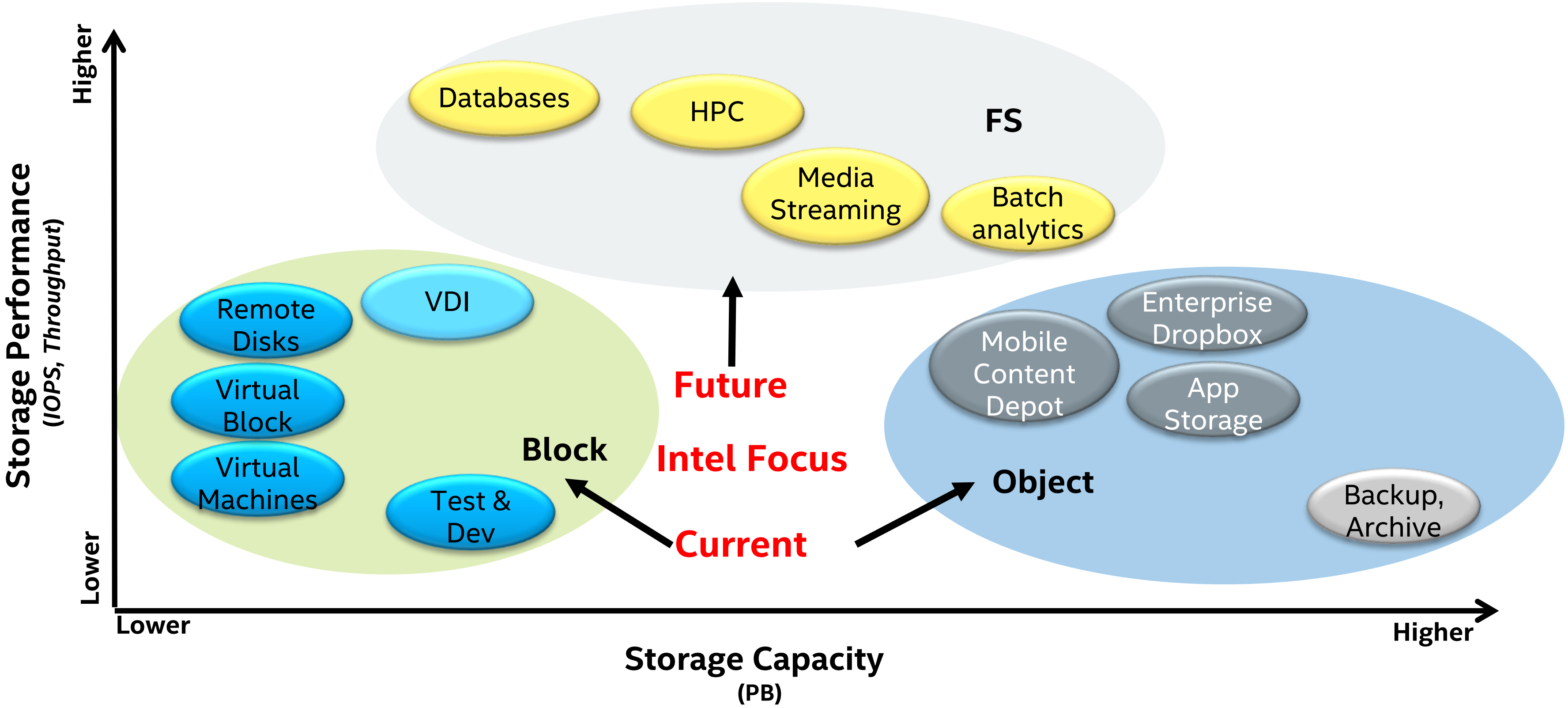
1. Client app issues read request, RADOS sends request to primary OSD
2. Primary OSD reads data from local disk and completes read request



1. Client app writes data, RADOS sends data to primary OSD
2. Primary OSD identifies replica OSDs and sends data, writes data to local disk
3. Replica OSDs write data to local disk, signal completion to primary
4. Primary OSD signals completion to client app

IDF15

Ceph Workloads

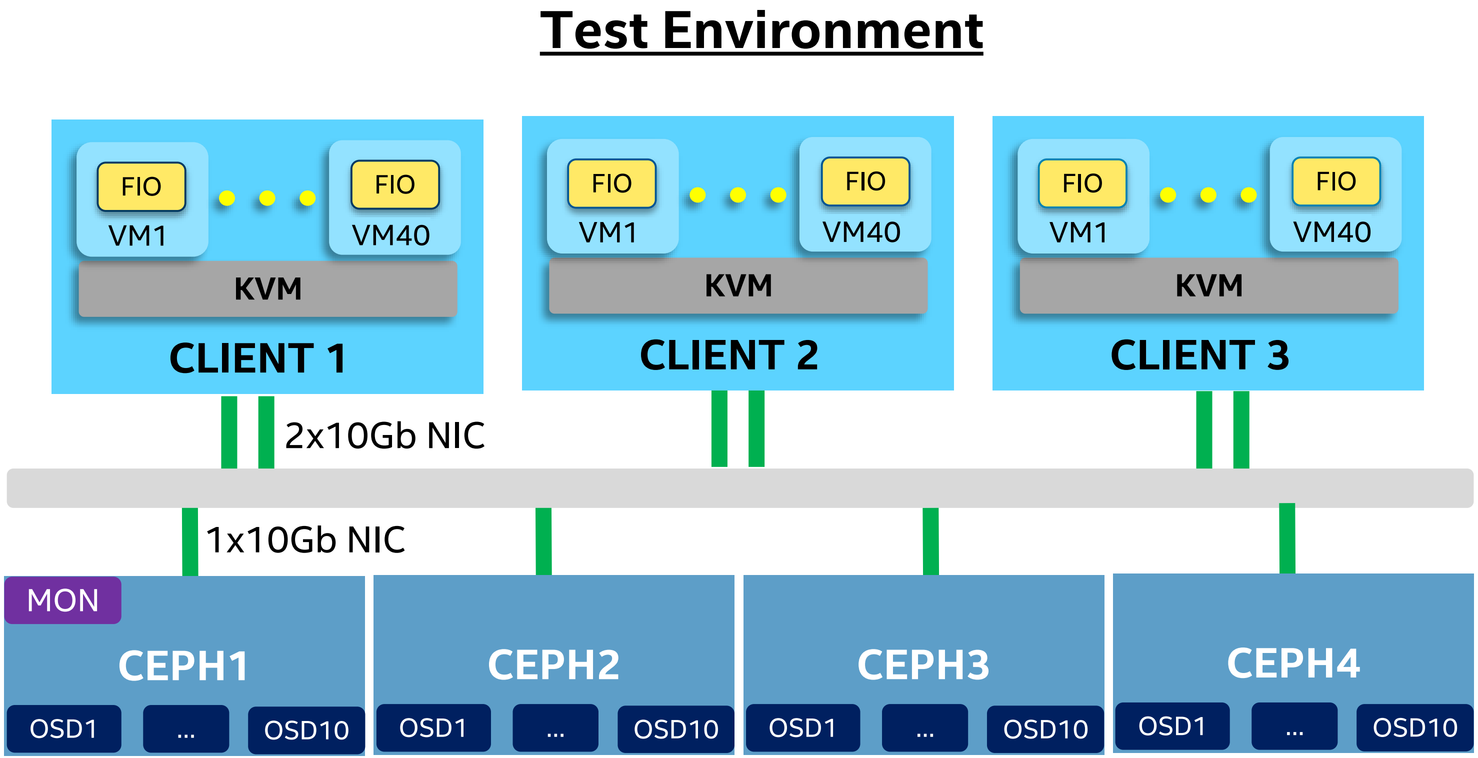


IDF15

Agenda

- The Problem
- Ceph Introduction
- Ceph Performance
- Ceph Cache Tiering and Erasure Code
- Intel Product Portfolio for Ceph
- Ceph Best Practices
- Summary

Ceph Block Performance – Configuration



Note: See page #37, #38, #39 for system configuration and benchmark data

Compute Node

- 2 nodes with Intel® Xeon™ processor x5570 @ 2.93GHz, 128GB mem
- 1 node with Intel Xeon processor E5 2680 @2.8GHz, 56GB mem

Storage Node

- Intel Xeon processor E3-1275 v2 @ 3.5 GHz
- 32GB Memory
- 1xSSD for OS
- 10x 3 TB 7200rpm
- 2x 400GB Intel® SSD DC S3700

Ceph Block Performance – Measure Raw Performance

1 Run FIO on one HDD, collect disk IO performance

3 TB
7200 rpm

Random 512K Write	Random 512K Read	Random 4K Write	Random 4K Read
70 MB/s	**80 MB/s**	**270 IOPS**	**160 IOPS**

Note: Sequential 64K (Client) = Random 512K (Ceph OSD)

2 Estimate cluster performance (include replication overhead for writes – 2x in this test)

40 HDDs (Reads)
20 HDDs (Writes)

Random 512K Write	Random 512K Read	Random 4K Write	Random 4K Read
1400 MB/s	**3200 MB/s**	**5400 IOPS**	**6400 IOPS**

 Note: See page #37, #38, #39 for system configuration and benchmark data

Ceph Block Performance – Test Results

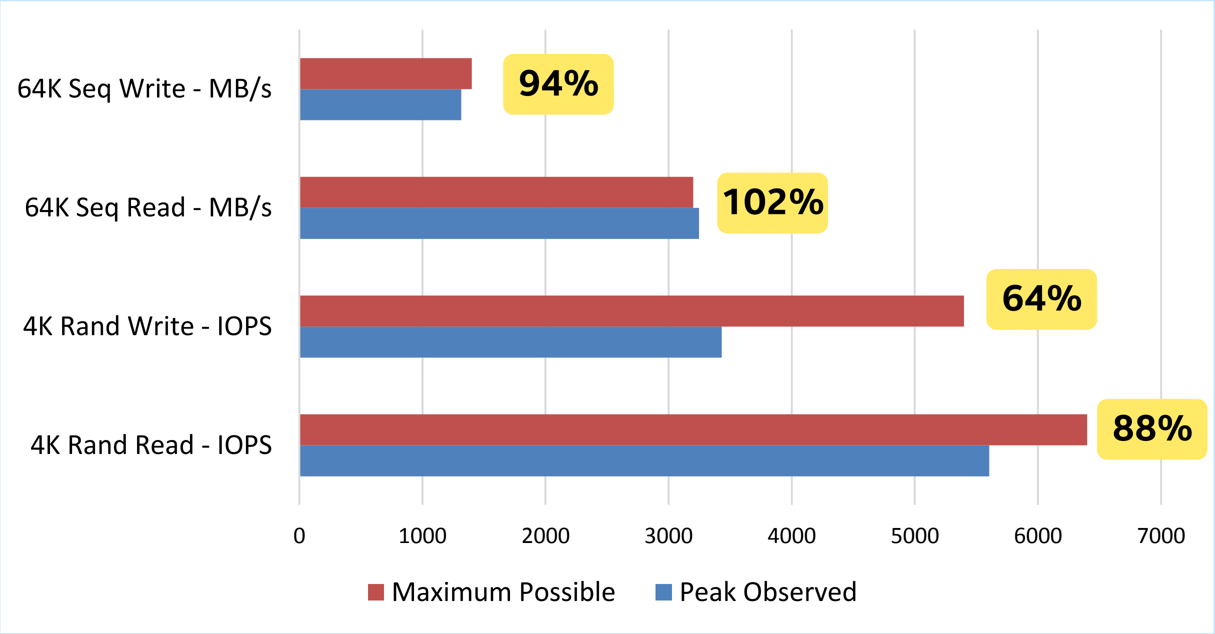
Drop OSD Cache

Prepare Data (dd)

Run FIO

1. 60GB Span
2. 4 IOs: Sequential (W,R), Random (W, R)
3. 100s warm-up, 600s test
4. RBD images – 1 to 120

CEPH Cluster Performance



Note: Random tests use **Queue Depth=8**, Sequential tests use **Queue Depth**=64
See page #39, #40, #41 for system configuration and benchmark data

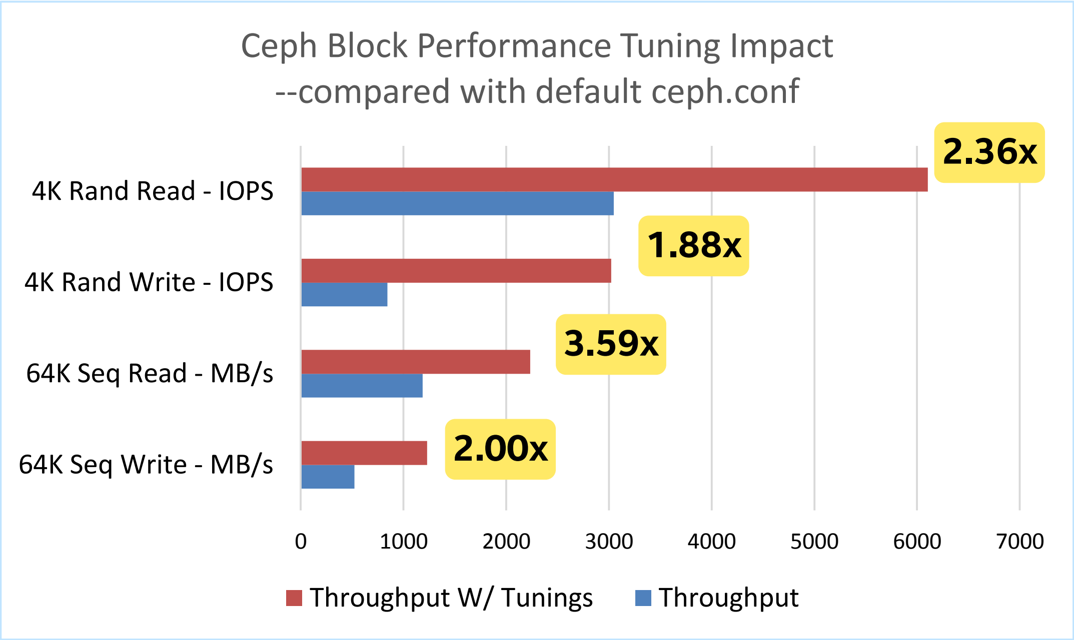
Ceph performance is close to max cluster IO limit for all but random writes – room for further optimizations

IDF15

Ceph Block Performance – Tuning effects

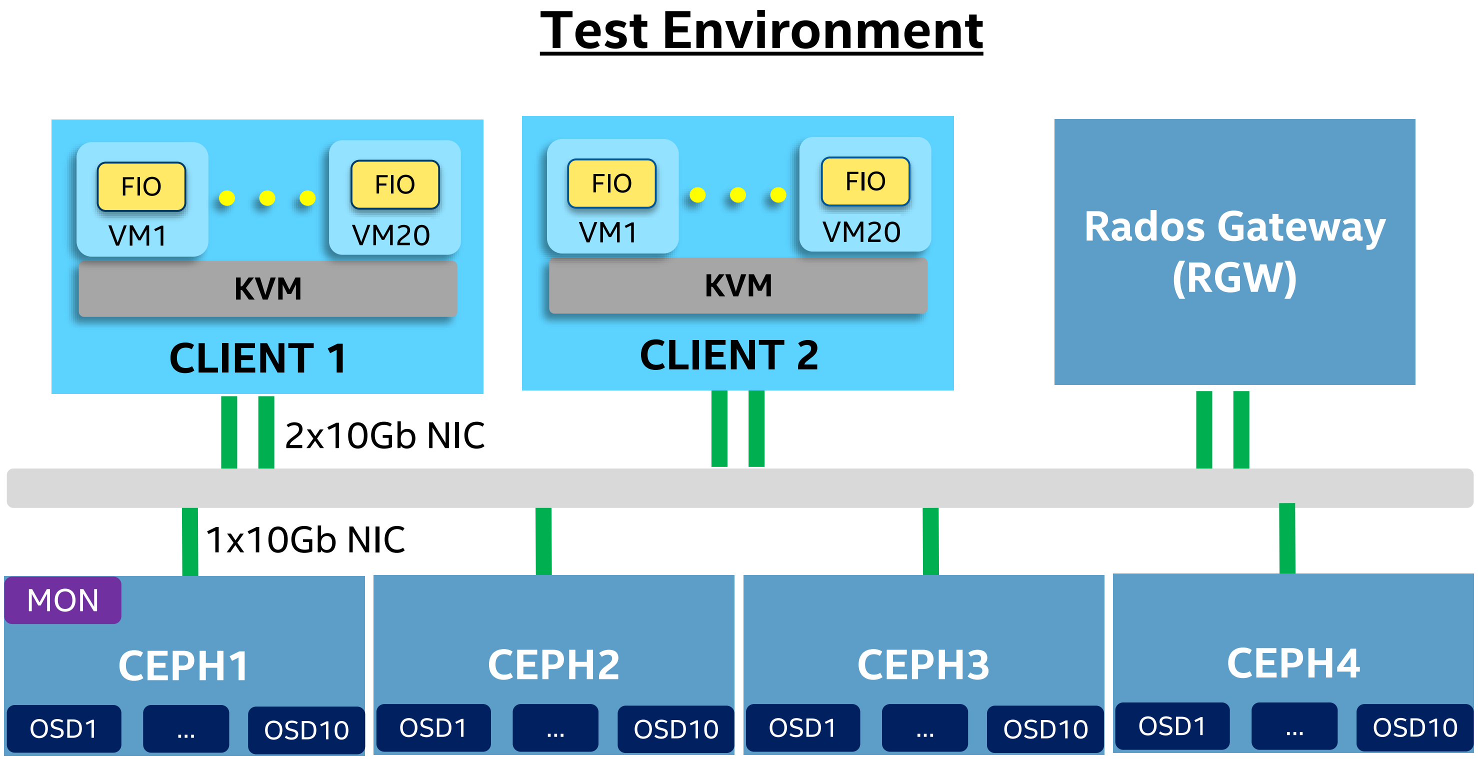
Best Tuning Knobs

- Large pg number: 81920
- Omap data on a separate Disk
- Read ahead = 2048
- I/O merge & write cache



Note: See page #37, #38, #39 for system configuration and benchmark data

Ceph Object Performance – Configuration



Note: See page #40, #41 for system configuration and benchmark data

Client Node

- 2 nodes with Intel® Xeon™ processor x5570 @ 2.93GHz, 24GB mem

Rados Gateway

- 1 nodes with Intel Xeon processor E5-2670 @ 2.6GHz, 64GB mem

Storage Node

- Intel Xeon processor E3-1280 v2 @ 3.6 GHz
- 16 GB Memory
- 1xSSD for OS
- 10x1 TB 7200rpm
- 3x 480GB Intel® SSD S530

Ceph Object Performance – Test Results

CEPH Cluster Performance

Prepare the Data

Run COSBench

1. 100 containers x 100 objects each
2. 4 IOs: 128K Read/Write; 10M Read/write
3. 100s warm-up, 600s test
4. COSBench workers – 1 to 2048

#con x # obj	Object-Size	RW-Mode	Worker-Count	Avg-ResTime	95%-ResTime	Throughput	Bandwidth	Bottleneck
	--	--	--	ms	ms	op/s	MB/s	--
100x100	128KB	Read	80	10	20	7,951	971	RGW CPU
		Write	320	143	340	2,243	274	OSD CPU
	10MB	Read	160	1,365	3,870	117	1,118	RGW NIC
		Write	160	3,819	6,530	42	397	OSD NIC

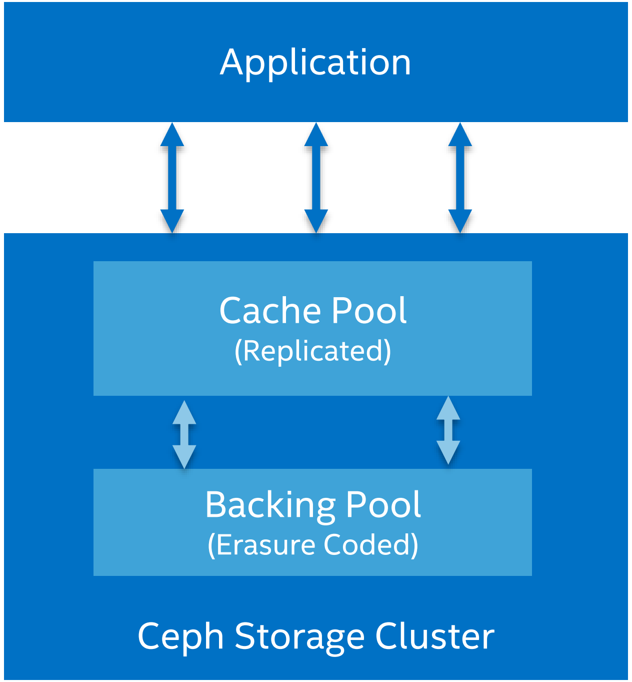
Note: COSBench is an Intel development open source cloud object storage benchmark
https://github.com/intel-cloud/cosbench
Note: See page #42, #43 for system configuration and benchmark data

Ceph Object performance is close to max cluster IO limit

Agenda

- The Problem
- Ceph Introduction
- Ceph Performance
- Ceph Cache Tiering and Erasure Code
- Intel Product Portfolio for Ceph
- Ceph Best Practices
- Summary

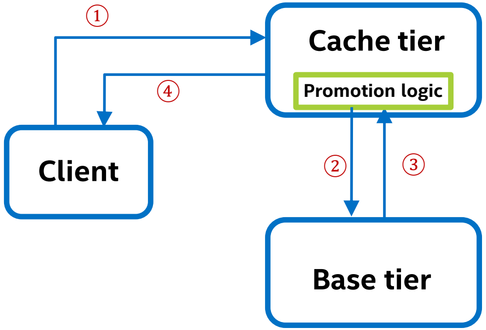
Ceph Cache Tiering



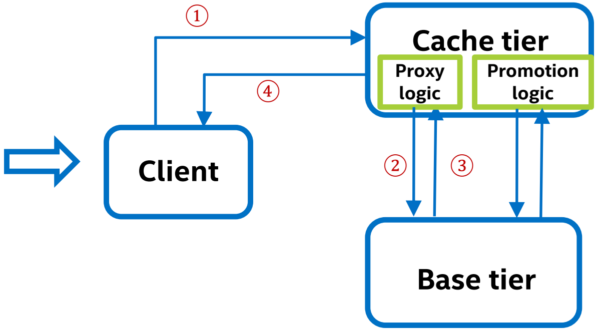
- An important feature towards Ceph enterprise readiness
  - Cost-effective
  - High performance tier W/ more SSDs
  - Use a pool of fast storage devices (Typically SSDs) and use it as a cache for an existing larger pool
    - E.g., Reads would first check the cache pool for a copy of the object, and then fall through to the existing pool if there is a miss
- Cache tiering mode
  - Read only
  - Write back

Ceph Cache Tiering Optimization – Proxy Read/write

Current design

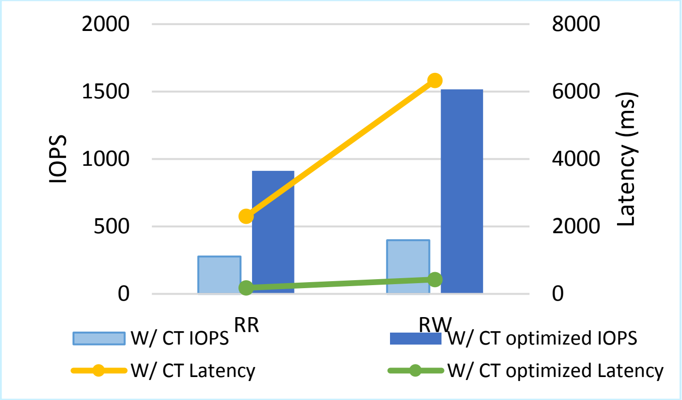


Proxy design



Cache Tiering optimization

4K random write



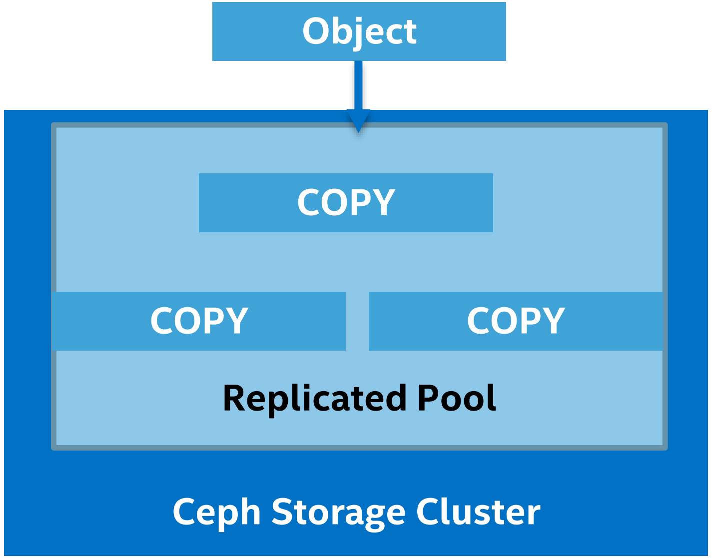
Proxy the read/write operations while the object is missing in cache tier, promotion in background

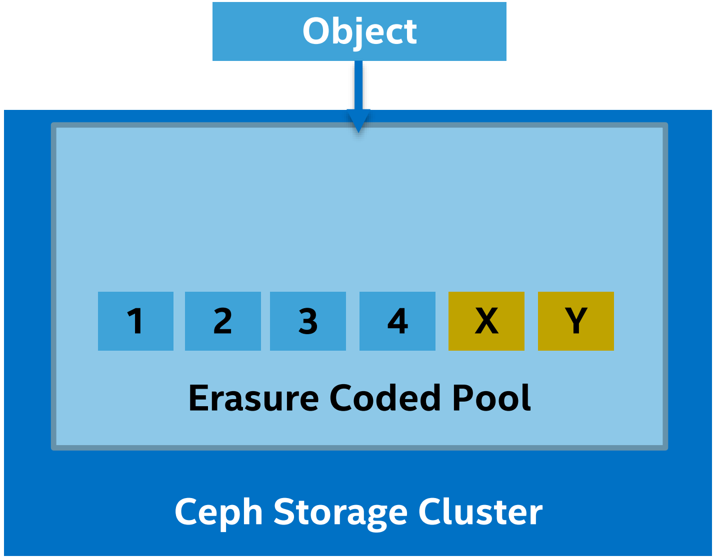
- **3.8x** and **3.6x** performance improvement respectively with proxy-write and proxy-read optimization

Proxy-read and write significantly improved cache tiering performance

 Note: See page #37, #38, #39 for system configuration and benchmark data

Ceph Erasure Coding





Full Copies of stored objects

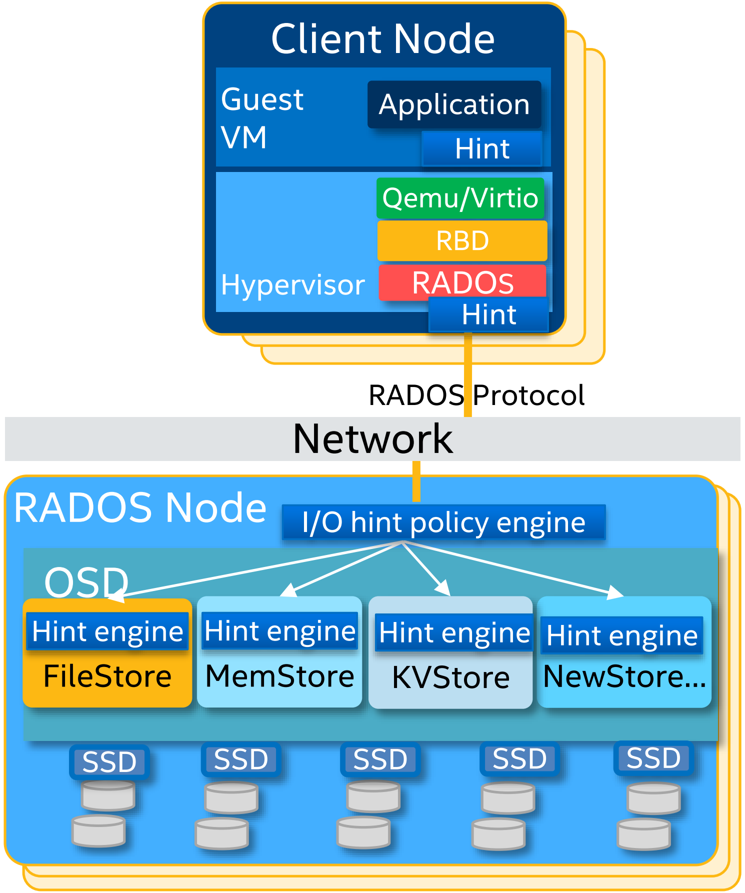
- Very high durability
- 3x (200% overhead)
- Quicker recovery

One Copy plus parity

- Cost-effective durability
- 1.5X (50% store overhead)
- Expensive recovery

Ceph EC optimization – I/O hint

- ISA-L EC library merged to Firefly
- EC performance
  - Acceptable performance impact
    - <10% degradation for 10M objects large scale read/write tests
  - But: Compared with 3x replica, we now tolerate **40%** object loss with **1.6x** space
- Rados I/O hint
  - Provide a hint to the storage system to classify the operate type brings differentiated storage services
  - Balance throughput and cost, and boost Storage performance
  - With rados I/O hint optimization, we can get even higher throughput compared W/O EC!



Note: See page #40, #41 for system configuration and benchmark data

35% performance improvement for EC write with Rados I/O Hint

Agenda

- The Problem
- Ceph Introduction
- Ceph Performance
- Ceph Cache Tiering and Erasure Code
- Intel Product Portfolio for Ceph
- Ceph Best Practices
- Summary

How is Intel helping?

- **Deliver Storage workload optimized products and technologies**
  - Optimize for Silicon, Flash* and Networking technologies

- **Working with the community to optimize Ceph on IA**
  - Make Ceph run best on Intel® Architecture (IA) based platforms (performance, reliability and management)

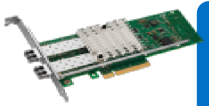
- **Publish IA Reference Architectures**
  - Share best practices with the ecosystem and community

Intel's Product Portfolio for Ceph

Server Boards & Systems

Software Libraries

Intel® Storage Acceleration Library (Intel® ISA-L)



Networking

Software Products

Intel® Cache Acceleration Software (Intel® CAS)

NVM

Opensource

- Virtual Storage Manager (VSM)
- Ceph contributions: No.2
- Cephperf†
- Reference Architectures

ceph

CPU Platforms

Ecosystem

- OEMs / ODMs
- ISVs

Solution focus with Intel® platform and software ingredients. Deep collaboration with Red Hat* and Inktank* (by Red Hat) to deliver enterprise ready Ceph solutions.

†Cephperf to be open sourced in Q2'15

VSM– Ceph Simplified

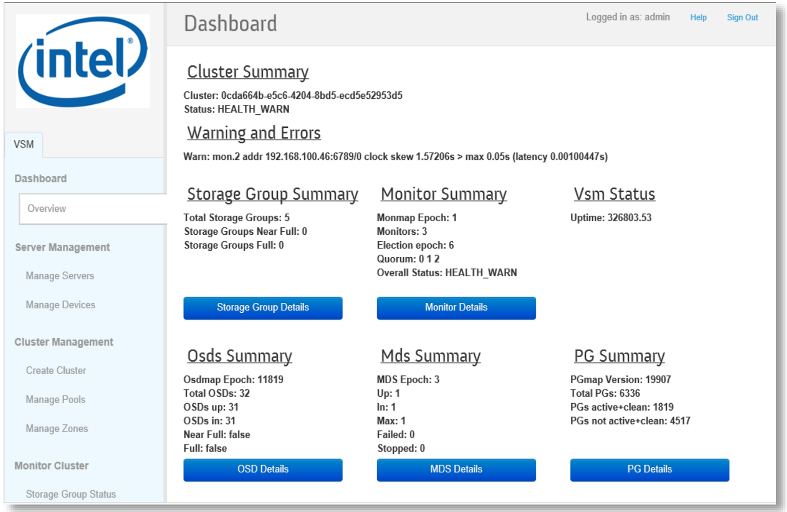
VSM (**V**irtual **S**torage **M**anager) - An open source Ceph management tool developed by Intel, and announced on 2014 Nov's OpenStack* Paris summit, designed to help make day to day management of Ceph easier for storage administrators.

Home page:
https://01.org/virtual-storage-manager

Code Repository:
https://github.com/01org/virtual-storage-manager

Issue Tracking:
https://01.org/jira/browse/VSM

Mailing list:
http://vsm-discuss.33411.n7.nabble.com/



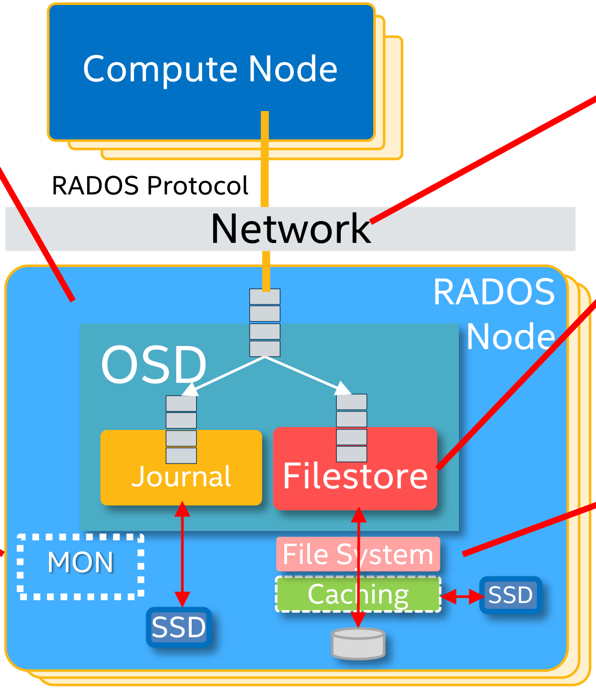
IDF15

Agenda

- The Problem
- Ceph Introduction
- Ceph Performance
- Ceph Cache Tiering and Erasure Code
- Intel Product Portfolio for Ceph
- Ceph Best Practices
- Summary

Ceph – Best Deployment Practices

- One OSD process per disk
- Approximately 1GHz Intel® Xeon™-class CPU per OSD
- 1GB memory per OSD



Jumbo Frames for 10Gbps

Use 10x of default queue params

- In small cluster, can co-locate with OSDs with 4GB-6GB
- Beyond 100 OSDs (100 HDDs or 10 nodes), deploy monitors on separate nodes
- 3 monitors for < 200 nodes

- XFS, deadline scheduler
- Tune *read_ahead_kb* for sequential I/O read
- *Queue Depth* – 64 for sequential, 8 for random

Note: Refer to backup for detailed tuning recommendations.

Agenda

- The Problem
- Ceph Introduction
- Ceph Performance
- Ceph Cache Tiering and Erasure Code
- Intel Product Portfolio for Ceph
- Ceph Best Practices
- Summary

Summary

- Cloud workloads and cost is driving the need for distributed storage solutions

- Strong customer interest and lots of production implementations in Ceph

- Intel is optimizing CEPH for Intel® Architecture

Next Steps

- Take advantage of Intel software optimizations and reference architectures for your production deployments
  - Pilot "cephperf" in Q2'15 and give us feedback http://01.org/cephperf
- Engage with open source communities for delivering enterprise features
- Innovate storage offerings using value added features with Ceph

Additional Sources of Information

- A PDF of this presentation is available from our Technical Session Catalog: www.intel.com/idfsessionsSZ. This URL is also printed on the top of Session Agenda Pages in the Pocket Guide.

- More web based info: http://ceph.com

- Intel® Solutions Reference Architectures www.intel.com/storage

- Intel® Storage Acceleration Library (Open Source Version) - https://01.org/intel%C2%AE-storage-acceleration-library-open-source-version

Legal Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, Xeon and the Intel logo are trademarks of Intel Corporation in the United States and other countries.

*Other names and brands may be claimed as the property of others.

Risk Factors

The above statements and any others in this document that refer to plans and expectations for the first quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Words such as "anticipates," "expects," "intends," "plans," "believes," "seeks," "estimates," "may," "will," "should" and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel's actual results, and variances from Intel's current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be important factors that could cause actual results to differ materially from the company's expectations. Demand for Intel's products is highly variable and could differ from expectations due to factors including changes in the business and economic conditions; consumer confidence or income levels; customer acceptance of Intel's and competitors' products; competitive and pricing pressures, including actions taken by competitors; supply constraints and other disruptions affecting customers; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Intel's gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; segment product mix; the timing and execution of the manufacturing ramp and associated costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; and product manufacturing quality/yields. Variations in gross margin may also be caused by the timing of Intel product introductions and related expenses, including marketing expenses, and Intel's ability to respond quickly to technological developments and to introduce new features into existing products, which may result in restructuring and asset impairment charges. Intel's results could be affected by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Results may also be affected by the formal or informal imposition by countries of new or revised export and/or import and doing-business regulations, which could be changed without prior notice. Intel operates in highly competitive industries and its operations have high costs that are either fixed or difficult to reduce in the short term. The amount, timing and execution of Intel's stock repurchase program and dividend program could be affected by changes in Intel's priorities for the use of cash, such as operational spending, capital spending, acquisitions, and as a result of changes to Intel's cash flows and changes in tax laws. Product defects or errata (deviations from published specifications) may adversely impact our expenses, revenues and reputation. Intel's results could be affected by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues. An unfavorable ruling could include monetary damages or an injunction prohibiting Intel from manufacturing or selling one or more products, precluding particular business practices, impacting Intel's ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. Intel's results may be affected by the timing of closing of acquisitions, divestitures and other significant transactions. A detailed discussion of these and other factors that could affect Intel's results is included in Intel's SEC filings, including the company's most recent reports on Form 10-Q, Form 10-K and earnings release.

Backup

Block Test Environment – Configuration Details

Client Nodes	
CPU	2 x Intel ®Xeon™ x5570 @ 2.93GHz (4-core, 8 threads) (Qty: 2) 2x Intel Xeon E5 2680 @2.8GHz (16-core, 32 threads) (Qty: 1)
Memory	128 GB (8GB * 16 DDR3 1333 MHZ) or 56GB (8GB * 7) for E5 server
NIC	10Gb 82599EB

Client VM	
CPU	1 X VCPU VCPUPIN
Memory	512 MB

Ceph Nodes	
CPU	1 x Intel Xeon E3-1275 V2 @ 3.5 GHz (4-core, 8 threads)
Memory	32 GB (4 x 8GB DDR3 @ 1600 MHz)
NIC	2 X 82599 10GbE
HBA/C204	{SAS2008 PCI Express* Fusion-MPT SAS-2} / {6 Series/C200 Series Chipset Family SATA AHCI Controller}
Disks	1 x SSDSA2SH064G1GC 2.5'' 64GB for OS 2 x Intel SSDSC2BA40 400 GB SSD (Journal) 10 x Seagate* ST3000NM0033-9ZM 3.5'' 3TB 7200rpm SATA HDD (Data)

Ceph cluster	
OS	CentOS 6.5
Kernel	2.6.32-431
Ceph	**0.61.2 built from source**

Client host	
OS	Ubuntu* 12.10
Kernel	3.6.3

Client VM	
OS	Ubuntu 12.10
Kernel	3.5.0-17

- **XFS** as file system for Data Disk
- Each Data Disk (SATA HDD) was parted into 1 partition for OSD daemon
- Default replication setting (2 replicas), **7872** pgs.
- Tunings
  - **Set read_ahead_kb=2048**
  - **MTU= 8000**
- **Change i/o scheduler to [deadline]:**

  # Echo deadline >/sys/block/[dev]/queue/scheduler

Ceph RBD Tuned Configuration

```
[global]
   debug default = 0
   log file = /var/log/ceph/$name.log
   max open files = 131072
   auth cluster required = none
   auth service required = none
   auth client required = none
[osd]
   osd mkfs type = xfs
   osd mount options xfs =
rw,noatime,inode64,logbsize=256k,delaylog
   osd mkfs options xfs = -f -i size=2048
   filestore max inline xattr size = 254
   filestore max inline xattrs = 6
   osd_op_threads=20
   filestore_queue_max_ops=500
   filestore_queue_committing_max_ops=5000
   journal_max_write_entries=1000
   journal_queue_max_ops=3000
   objecter_inflight_ops=10240
   filestore_queue_max_bytes=1048576000
   filestore_queue_committing_max_bytes
=1048576000
   journal_max_write_bytes=1048576000
   journal_queue_max_bytes=1048576000
   ms_dispatch_throttle_bytes=1048576000
   objecter_infilght_op_bytes=1048576000
   filestore_max_sync_interval=10
   filestore_flusher=false
   filestore_flush_min=0
   filestore_sync_flush=true
```

Testing Methodology

Storage interface

Use QemuRBD as storage interface

Tool

- Use “dd” to prepare data for R/W tests
- Use fio (ioengine=libaio, direct=1) to generate 4 IO patterns: sequential write/read, random write/read
- Access Span: 60GB
- For capping tests, Seq Read/Write (60MB/s), and Rand Read/Write (100 ops/s)
- QoS Compliance:
  - For random 4k read/write cases: latency <= 20ms
  - For sequential 64K read/write cases: BW >= 54 MB/s

Run rules

- Drop osds page caches (“1” > /proc/sys/vm/drop_caches)
- 100 secs for warm up, 600 secs for data collection
- Run 4KB/64KB tests under different # of rbds (1 to 120)

Space allocation (per node)

- Data Drive:
  - Sits on 10x 3TB HDD drives
  - So 4800GB/40 * 2 = 240GB data space will be used on each Data disk at 80 VMs.
- Journal:
  - Sits on 2x 400GB SSD drives
  - One journal partition per data drive, 10GB

Object Test Environment - Configuration Details

Client & RGW	
CPU	Client: 2 x Intel ®Xeon™ x5570 @ 2.93GHz (4-core, 8 threads) (Qty: 2) GRW: 2x Intel Xeon E5 2670@2.6GHz (16-core, 32 threads) (Qty: 1)
Memory	128 GB (8GB * 16 DDR3 1333 MHZ) or 56GB (8GB * 7) for E5 server
NIC	10Gb 82599EB

Ceph OSD Nodes	
CPU	1 x Intel Xeon E3-1280 V2 @ 3.6 GHz (4-core, 8 threads)
Memory	32 GB (4 x 8GB DDR3 @ 1600 MHz)
NIC	2 X 82599 10GbE
HBA/C204	{SAS2308 PCI Express* Fusion-MPT SAS-2} / {6 Series/C200 Series Chipset Family SATA AHCI Controller}
Disks	1 x SSDSA2SH064G1GC 2.5" 64GB for OS 3 x Intel SSDSC2CW480A3 480 GB SSD (Journal) 10 x Seagate* ST1000NM0011 3.5" 1TB 7200rpm SATA HDD (Data)

Ceph cluster	
OS	Ubuntu 14.04
Kernel	3.13.0
Ceph	**0.61.8 built from source**

Client host	
OS	Ubuntu* 12.10
Kernel	3.6.3

- **XFS** as file system for Data Disk
- Each Data Disk (SATA HDD) was parted into 1 partition for OSD daemon
- Default replication setting (3 replicas), **12416** pgs.
- Tunings
  - **Set read_ahead_kb=2048**
  - **MTU= 8000**

IDF15

Ceph Rados Gateway (RGW) Tuning

ceph-gateway

- –rgw enable ops log = false
- –rgw enable usage log = false
- –rgw thread pool size = 256
- –Log disabled

apache2-http-server

- –<IfModule mpm_worker_module>
- – ServerLimit 50
- – StartServers 5
- – MinSpareThreads 25
- – MaxSpareThreads 75
- – ThreadLimit 100
- – ThreadsPerChild 100
- – MaxClients 5000
- – MaxRequestsPerChild 0
- –</IfModule>

GW instances

- –5 instances in one GW server
- –Access log turned off