Accelerating the Big Data and cloud storage with Intel® Non-Volatile Memory Technologies

—Intel® Optane™ and Intel® 3D NAND SSDS

Jack Zhang
Storage Solution Architect
Intel Corp

Agenda

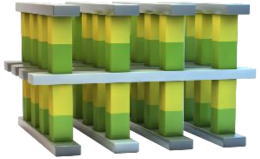
- Intel Non-Volatile Memory Technologies
- Solid Stat Devices (SSD) on Apache Spark
- All Flash Ceph for big data
- Summary

Intel Non-Volatile Memory Technologies

Intel® 3D NAND SSDs and OPTANE SSD Transform Storage

Expand the reach of Intel® SSDs. Deliver disruptive value to the data center.

Optimized STORAGE Solutions



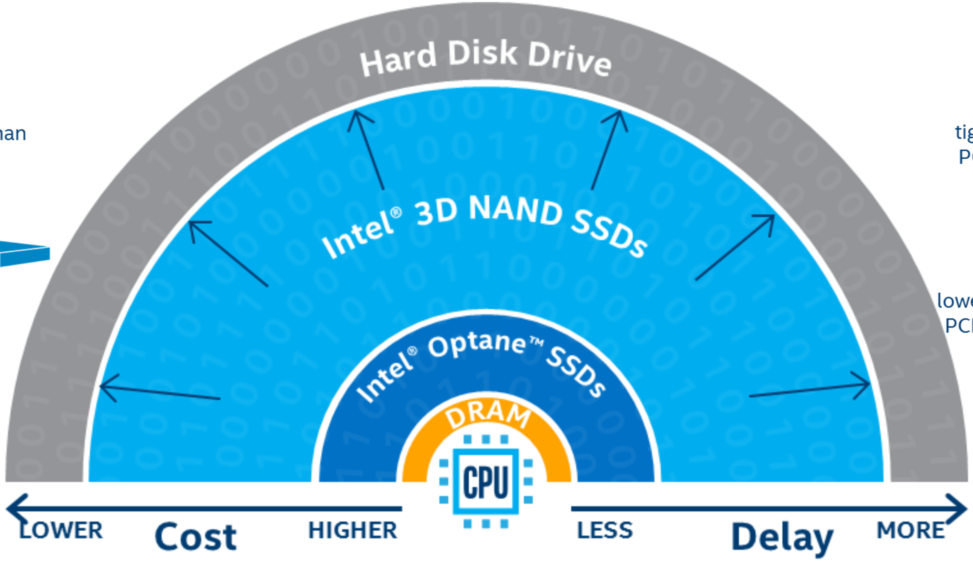
Up to
359x
more IOPS/\$
than 10K HDD[6]

>2X
higher endurance than
2D NAND SSDs[7]

Up to
217x
more IOPS/W
than 10K HDD[6]

More capacity
per rack unit[11]

Capacity
for Less



Up to
200x
tighter QoS than
PCIe NAND SSD

>3X
higher endurance than
PCIe NAND SSD

Up to
30%
lower power than
PCIe ANND SSD

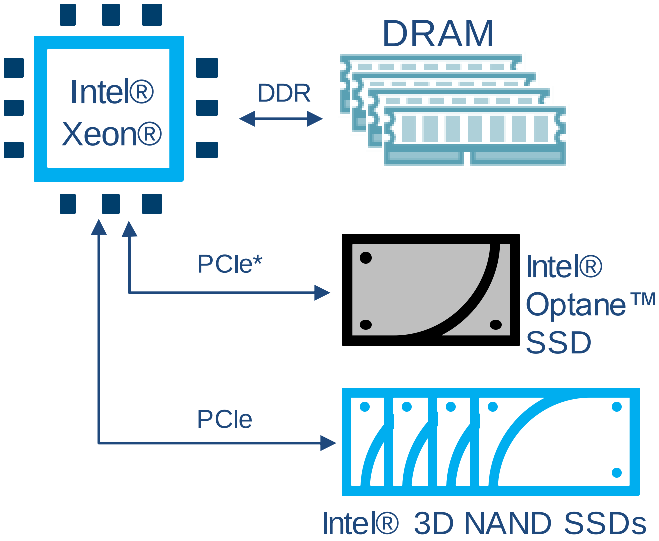
More VMs, Same QoS
per rack

Performance
for Less

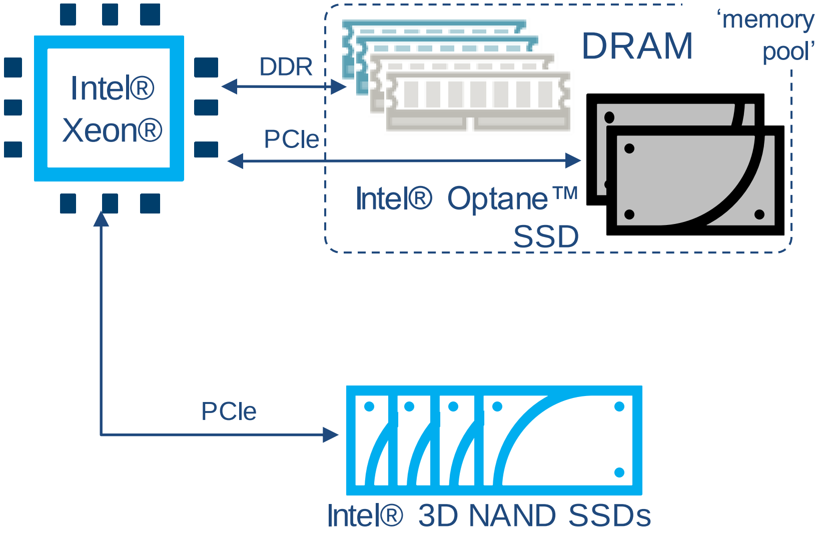
Refer to appendix for footnotes

Intel® Optane™ SSD Use Cases

Fast Storage



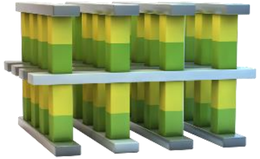
Extend Memory



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

Innovation for Cloud STORAGE : Intel® Optane™ Intel® 3D NAND SSDs

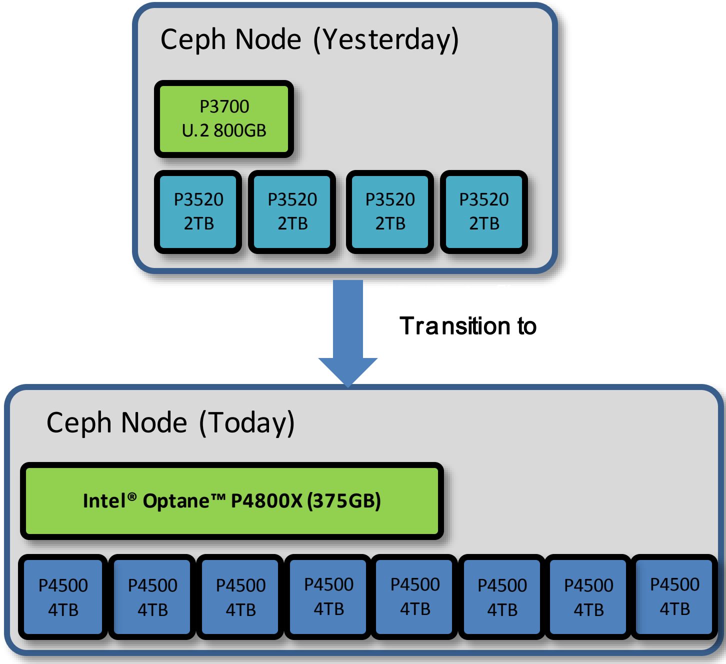
- New Storage Infrastructure: enable high performance and cost effective storage:



+

- **Journal/Log/Cache** **Data**

- Openstack/Ceph:
  - Intel Optane™ as Journal/Metadata/WAL (**Best** write performance, **Lowest** latency and **Best** QoS)
  - Intel 3D NAND TLC SSD as data store (cost effective storage)
  - **Best IOPS/$, IOPS/TB and TB/Rack**



Ceph Node (Yesterday)

P3700 U.2 800GB

P3520 2TB | P3520 2TB | P3520 2TB | P3520 2TB

Transition to

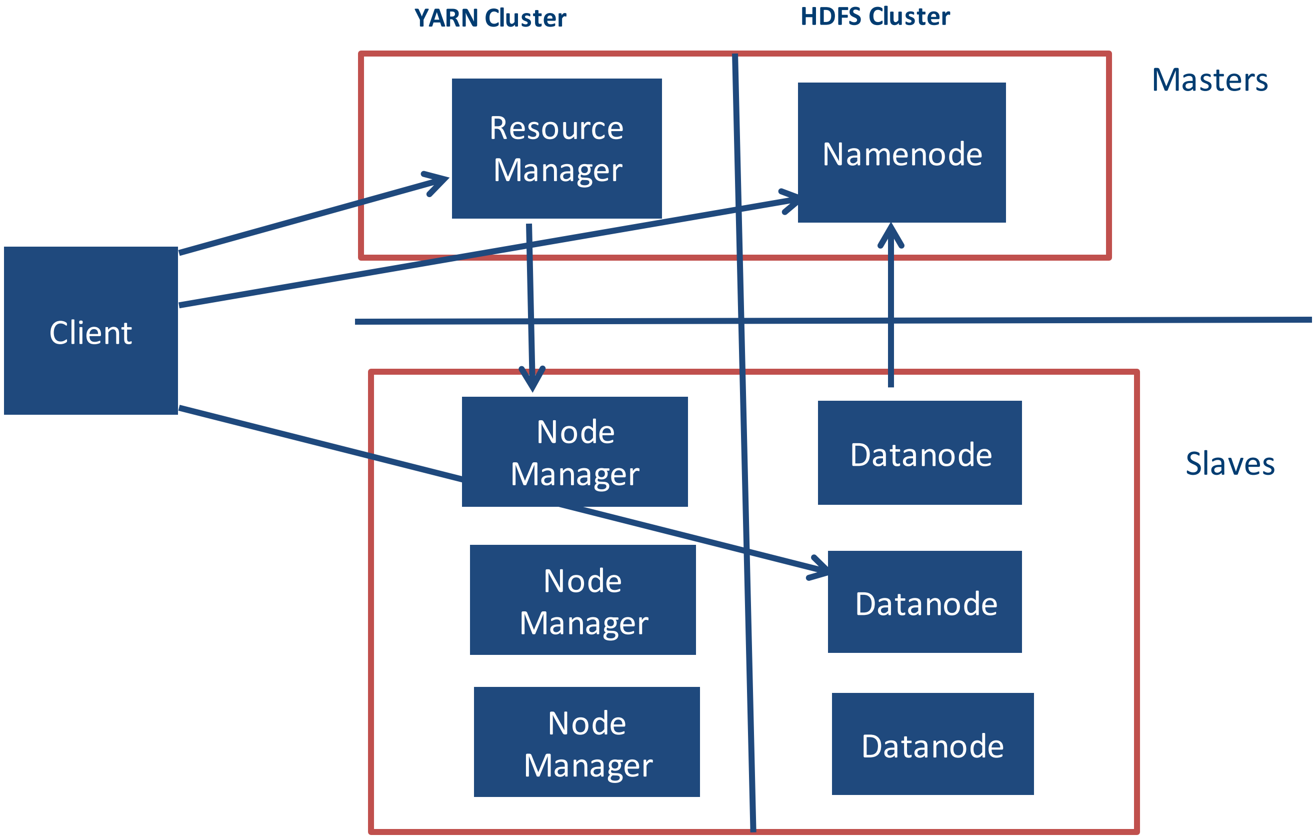
Ceph Node (Today)

Intel® Optane™ P4800X (375GB)

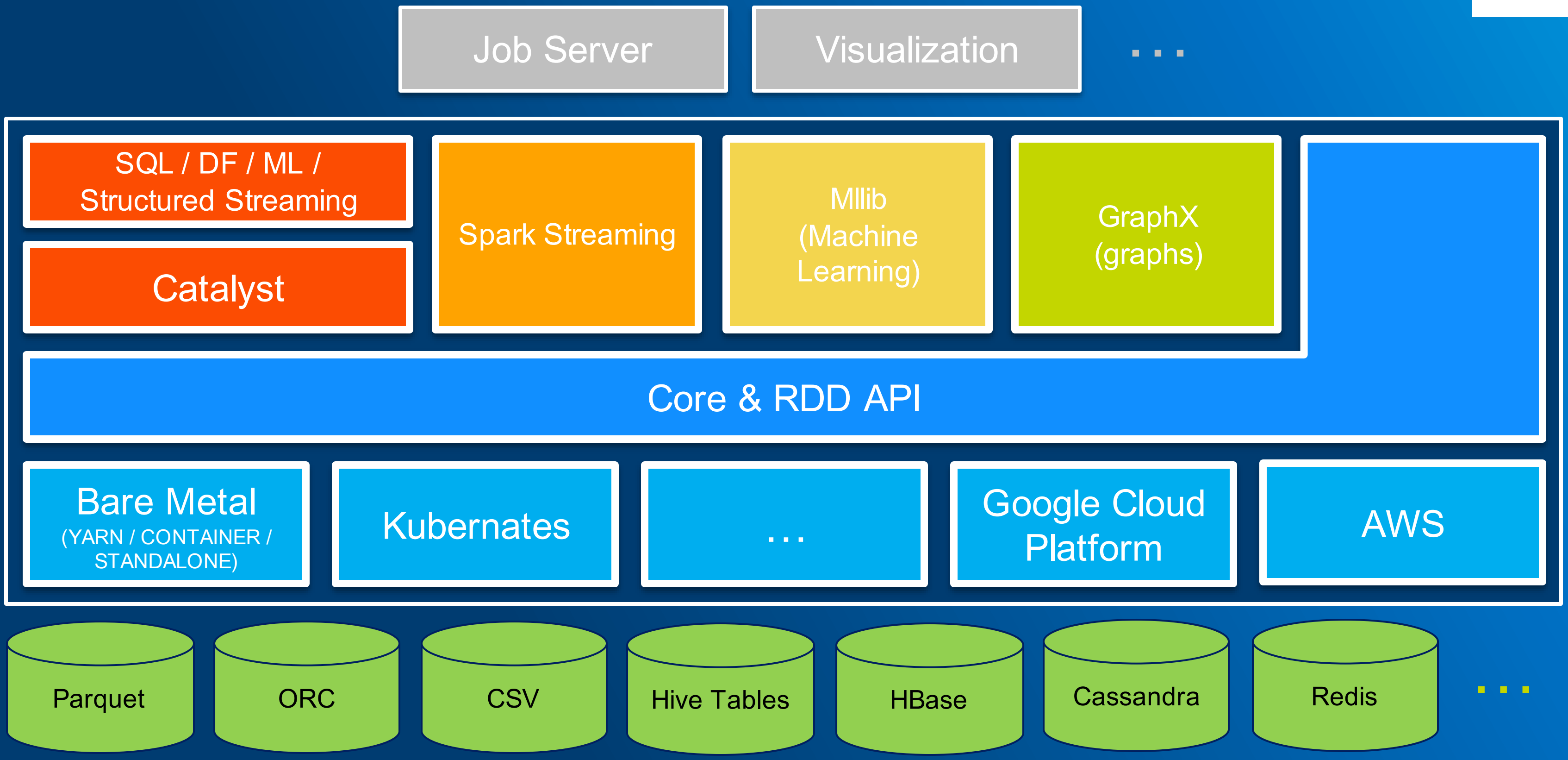
P4500 4TB | P4500 4TB | P4500 4TB | P4500 4TB | P4500 4TB | P4500 4TB | P4500 4TB | P4500 4TB

Solid Stat Devices (SSD) on Apache Spark

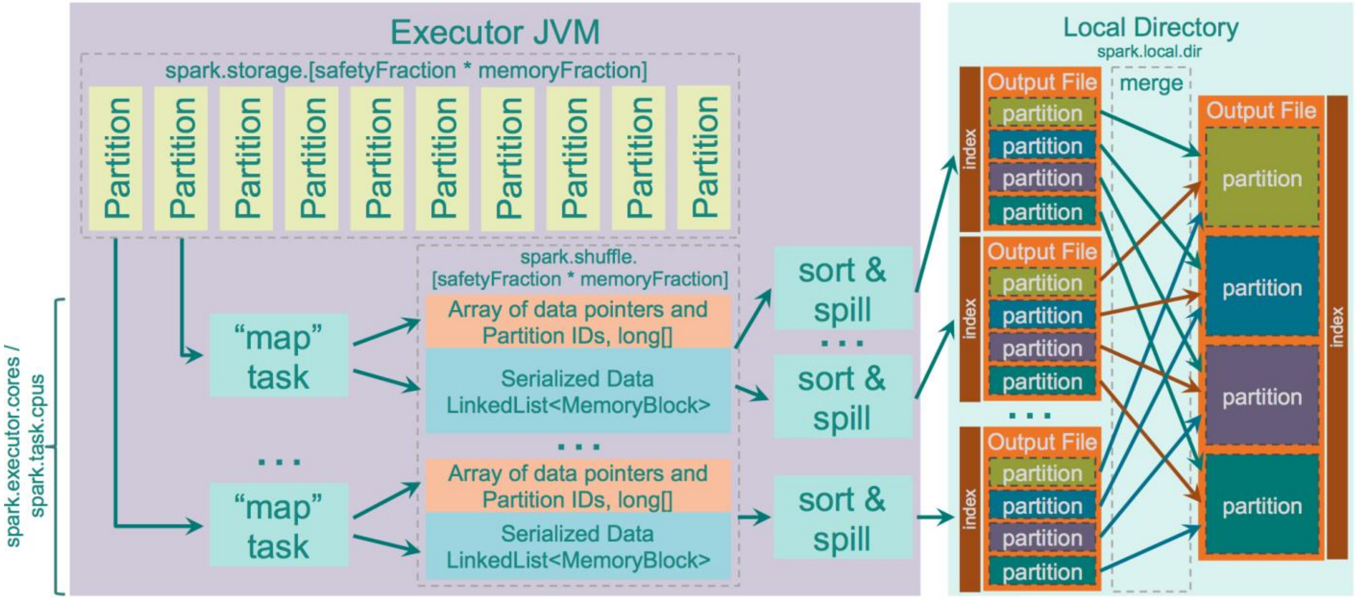
Apache Hadoop Architecture



Apache Spark



Shuffle with SSD in Big Data



For MapReduce or Spark, shuffle process will spill temp files on local disk when memory is not enough to hold all the data. To place the temp data on SSD, it is expected to achieve better performance for MapReduce or Spark workload.

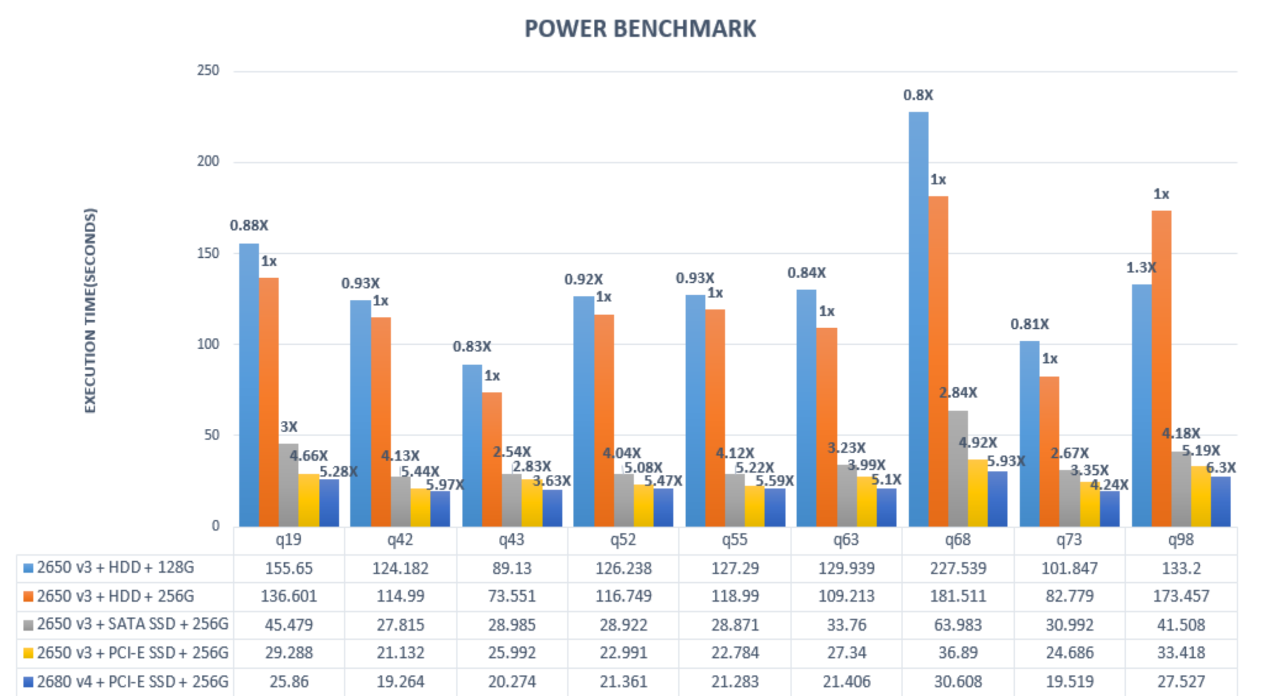
Benchmark Configurations & Workloads

Nodes	Master	Slave
Roles	Hadoop Name Node, Spark Master	Hadoop Data Node, Spark Slaves
Services	Name Node, Resource Manager	Data Node, Node Manager
Numbers	1	7
Processer	Intel Xeon E5-2650 v3 (HSW) / Intel Xeon E5-2680 v4 (BDW) (Dual Socket / node)	
Memory	256GB	256GB
Storage	OS Disk: 480GB SSD	OS Disk: 480GB SSD Data Disk: 1TB SATA HDD x 8 / Data Disk: Intel S3520 SSD x 8 / Data Disk: Intel P3600 SSD x 3
Network	10Gb	10Gb

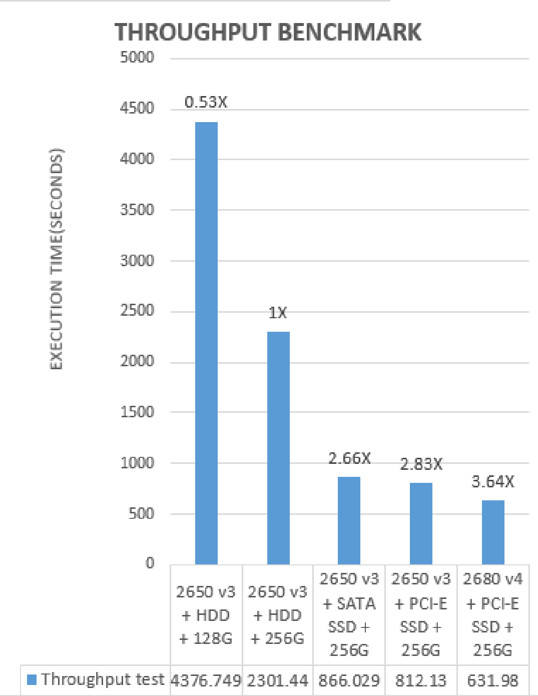
Hadoop/Spark Configuration	
Hadoop version	2.7.3
Spark version	2.1.0
Executor memory	25~40 GB
Executor Cores	8 – 10 / executors
Executor Number	5 / nodes
Spark Mode	yarn-client
JDK Version	1.8.0_112
memory.Overhead	10% Executor Memory
Shuffle Partition #	200
Broadcast threshold	30MB
broadcastTimeout	3600 sec
GC	Parallel GC

Workload (TPC-DS)	
Queries	19,42,43,52,55,63,68,72,98
Data Scale (Raw Data)	10 TB
Data Format	Parquet
Compression Codec	Snappy
Data Size	~3TB

Latest Intel's Platform powers Apache Spark (SQL)



	q19	q42	q43	q52	q55	q63	q68	q73	q98
■ 2650 v3 + HDD + 128G	155.65	124.182	89.13	126.238	127.29	129.939	227.539	101.847	133.2
■ 2650 v3 + HDD + 256G	136.601	114.99	73.551	116.749	118.99	109.213	181.511	82.779	173.457
■ 2650 v3 + SATA SSD + 256G	45.479	27.815	28.985	28.922	28.871	33.76	63.983	30.992	41.508
■ 2650 v3 + PCI-E SSD + 256G	29.288	21.132	25.992	22.991	22.784	27.34	36.89	24.686	33.418
■ 2680 v4 + PCI-E SSD + 256G	25.86	19.264	20.274	21.361	21.283	21.406	30.608	19.519	27.527



	2650 v3 + HDD + 128G	2650 v3 + HDD + 256G	2650 v3 + SATA SSD + 256G	2650 v3 + PCI-E SSD + 256G	2680 v4 + PCI-E SSD + 256G
■ Throughput test	4376.749	2301.44	866.029	812.13	631.98

DISK is the bottleneck other than CPU for SQL queries; we can observe 2.8x performance gain when upgrade the 8 * HDD -> 8 * S3520 SSD, and another ~35% performance while upgrade the E5 2650 v3 + 8*S3520 SSD -> E5 2680v4 + 3 P3600 SSD. We believe the more PCI-E SSD may give even better acceleration.

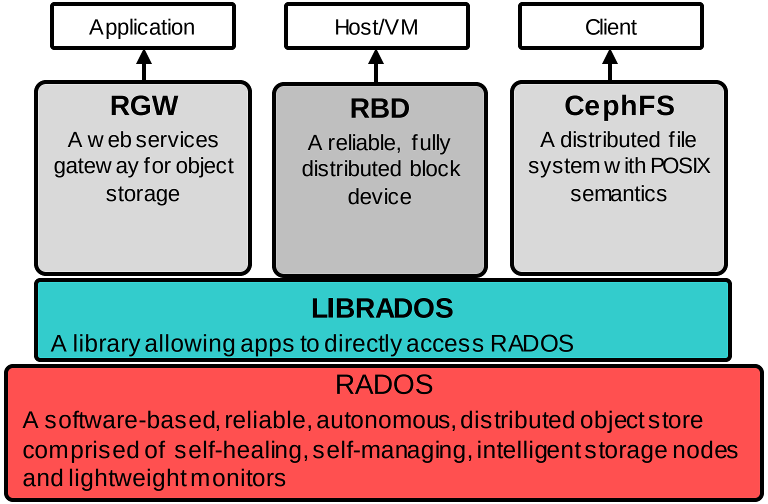
TCO Model (Sequential)

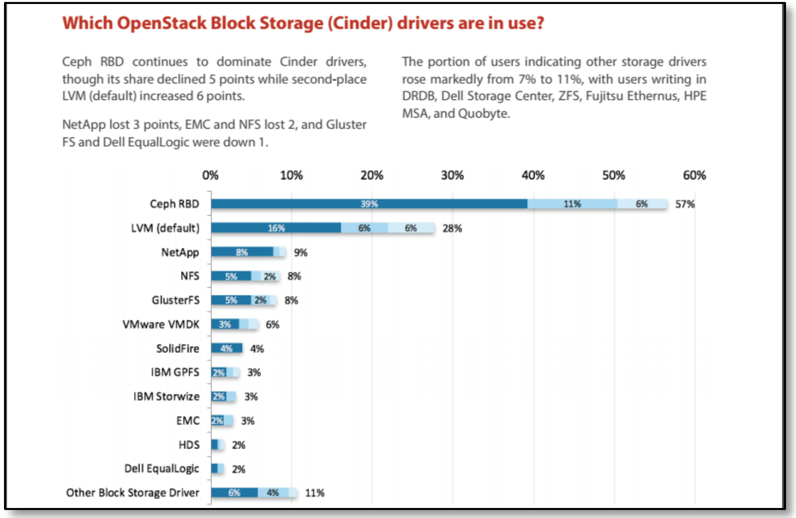
	Intel HSW(2650v3)			Intel BDW(2680v4)
Disk Types	HDD(1TB SATA)	SATA SSD(s3520)	PCIe SSD(p3600)	PCIe SSD(p3600)
Numbers of Drives	8	8	3	3
Total Capacity	1TB x 8	1.2TB x 8	1.6TB x 3	1.6TB x 3
Performance Gain	1x	~2.65x	~2.83x	~3.64x
Drive Cost	$800	$4224	$4782	$4782
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Cooling Cost	$505	$82	$42	$42
Enclosure Cost	$3943	$3943	$0	$0
Reliability	$1008	$339	$287	$287
Total Cost	$6677	$8656	$5146	$5146
Cost (per GB)	1x	1.08x	-	-
Perf (per Dollar)	1.0x	~2.4x	1.0x	~1.28x

Latest Intel’s Platform powers Apache Spark (SQL)

All Flash Ceph for big data

About Ceph


Application
Host/VM
Client
RGW
A web services gateway for object storage
RBD
A reliable, fully distributed block device
CephFS
A distributed file system with POSIX semantics
LIBRADOS
A library allowing apps to directly access RADOS
RADOS
A software-based, reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors



Which OpenStack Block Storage (Cinder) drivers are in use?

Ceph RBD continues to dominate Cinder drivers, though its share declined 5 points while second-place LVM (default) increased 6 points.

NetApp lost 3 points, EMC and NFS lost 2, and Gluster FS and Dell EqualLogic were down 1.

The portion of users indicating other storage drivers rose markedly from 7% to 11%, with users writing in DRDB, Dell Storage Center, ZFS, Fujitsu Ethernus, HPE MSA, and Quobyte.

- Open-source, object-based scale-out storage
- Object, Block and File in single unified storage cluster
- Highly durable, available – replication, erasure coding
- Runs on economical commodity hardware
- 10 years of hardening, vibrant community

- Scalability – CRUSH data placement, no single POF
- Replicates and re-balances dynamically
- Enterprise features – snapshots, cloning, mirroring
- Most popular block storage for Openstack use cases
- Commercial support from Red Hat

References: http://ceph.com/ceph-storage, http://thenewstack.io/software-defined-storage-ceph-way,

Who is using Ceph?

Telcom

CISCO · TIVIT synapsis · China unicom中国联通 · 中国电信 CHINA TELECOM · américa móvil

CSP/IPDC

YAHOO! · ebay · DigitalOcean · Ctrip · 360 · Letv · SHANDA 盛大集团

OEM/ODM

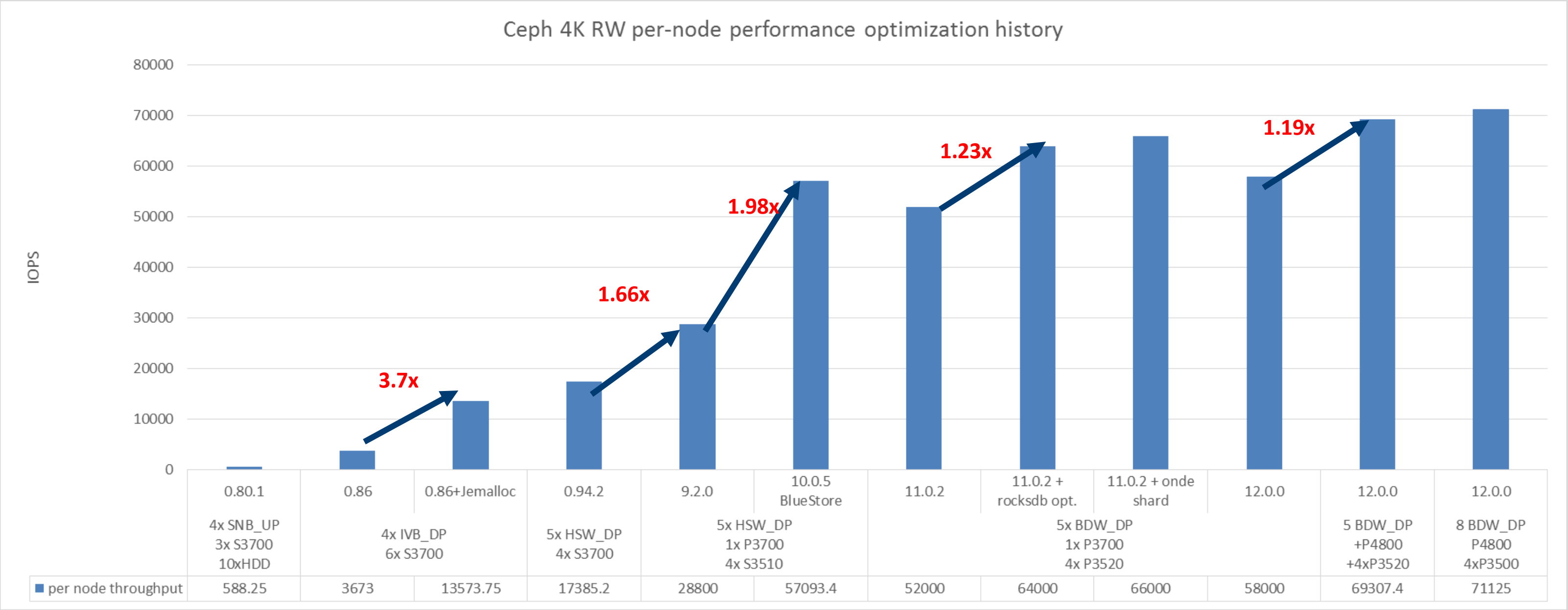
H3C · QCT Quanta CLOUD TECHNOLOGY · Hewlett Packard Enterprise · DELL

Enterprise, FSI, Healthcare, Retailers

AMX · Walmart · GE imagination at work · Adobe · TARGET · Bloomberg

Searchable Examples: http://www.slideshare.net/inktank_ceph

Ceph* performance trend with SSD – 4K Random Write



	0.80.1	0.86	0.86+Jemalloc	0.94.2	9.2.0	10.0.5 BlueStore	11.0.2	11.0.2 + rocksdb opt.	11.0.2 + onde shard	12.0.0	12.0.0	12.0.0
	4x SNB_UP 3x S3700 10xHDD	4x IVB_DP 6x S3700		5x HSW_DP 4x S3700	5x HSW_DP 1x P3700 4x S3510		5x BDW_DP 1x P3700 4x P3520				5 BDW_DP +P4800 +4xP3520	8 BDW_DP P4800 4xP3500
■ per node throughput	588.25	3673	13573.75	17385.2	28800	57093.4	52000	64000	66000	58000	69307.4	71125

38x performance improvement in Ceph all-flash array!

Suggested Configurations for Ceph* Storage Node

Standard/good (baseline):

Use cases/Applications: that need high capacity storage with high throughput performance

- NVMe*/PCIe* SSD for Journal + Caching, HDDs as OSD data drive

Better IOPS

Use cases/Applications: that need higher performance especially for throughput, IOPS and SLAs with medium storage capacity requirements

- NVMe/PCIe SSD as Journal, High capacity SATA SSD for data drive

Best Performance

Use cases/Applications: that need highest performance (throughput and IOPS) and low latency/QoS (Quality of Service).

- All NVMe/PCIe SSDs

More information at Ceph.com (new RAs update soon!)
http://tracker.ceph.com/projects/ceph/wiki/Tuning_for_All_Flash_Deployments

Ceph* storage node --Good	
CPU	Intel(R) Xeon(R) CPU E5-2650v3
Memory	64 GB
NIC	10GbE
Disks	1x 1.6TB P3700 + 12 x 4TB HDDs (1:12 ratio) P3700 as Journal and caching
Caching software	Intel(R) CAS 3.0, option: Intel(R) RSTe/MD4.3

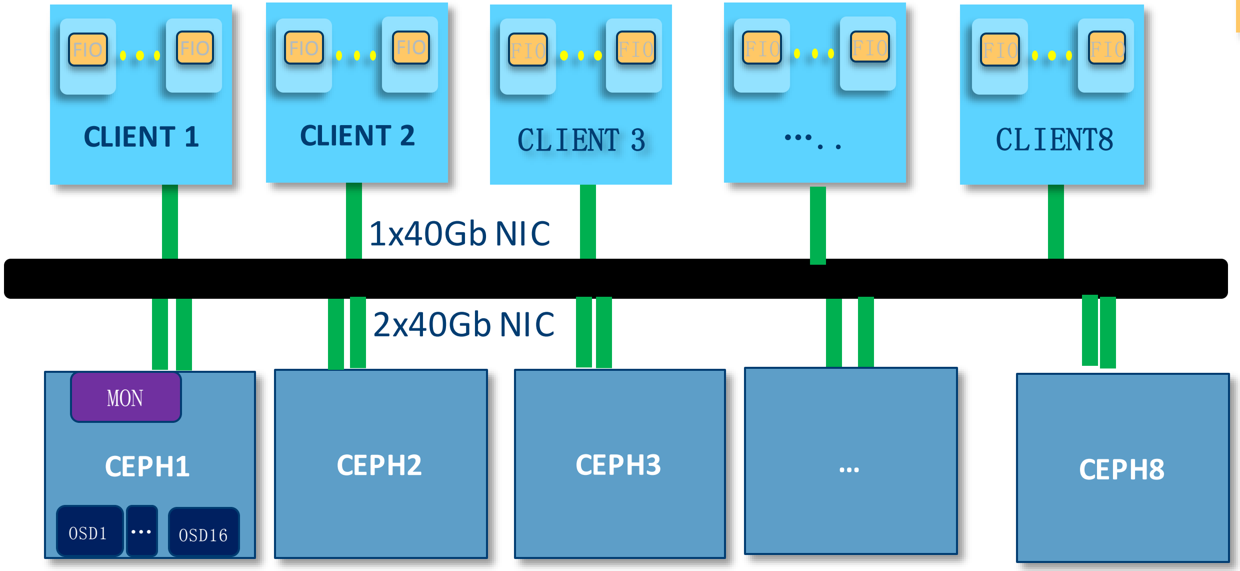
Ceph* Storage node --Better	
CPU	Intel(R) Xeon(R) CPU E5-2690
Memory	128 GB
NIC	Duel 10GbE
Disks	1x Intel(R) DC P3700(800G) + 4x Intel(R) DC S3510 1.6TB Or 1xIntel P4800X (375GB) + 8x Intel® DC S3520 1.6TB

Ceph* Storage node --Best	
CPU	Intel(R) Xeon(R) CPU E5-2699v4
Memory	>= 128 GB
NIC	2x 40GbE, 4x dual 10GbE
Disks	1xIntel P4800X (375GB) + 8x Intel® DC P4500 4TB

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

Intel® SSD DC P4800X + Intel® SSD DC P4500

Test Environment



Workloads

- Fio with librbd
- 20x 30 GB volumes each client
- 4 test cases: 4K/8K/16K random read & write; 64K Sequential read & write

8x Client Node

- Intel® Xeon™ processor E5-2699 v3 @ 2.3GHz, 64GB mem
- 1x X710 40Gb NIC

8x Storage Node

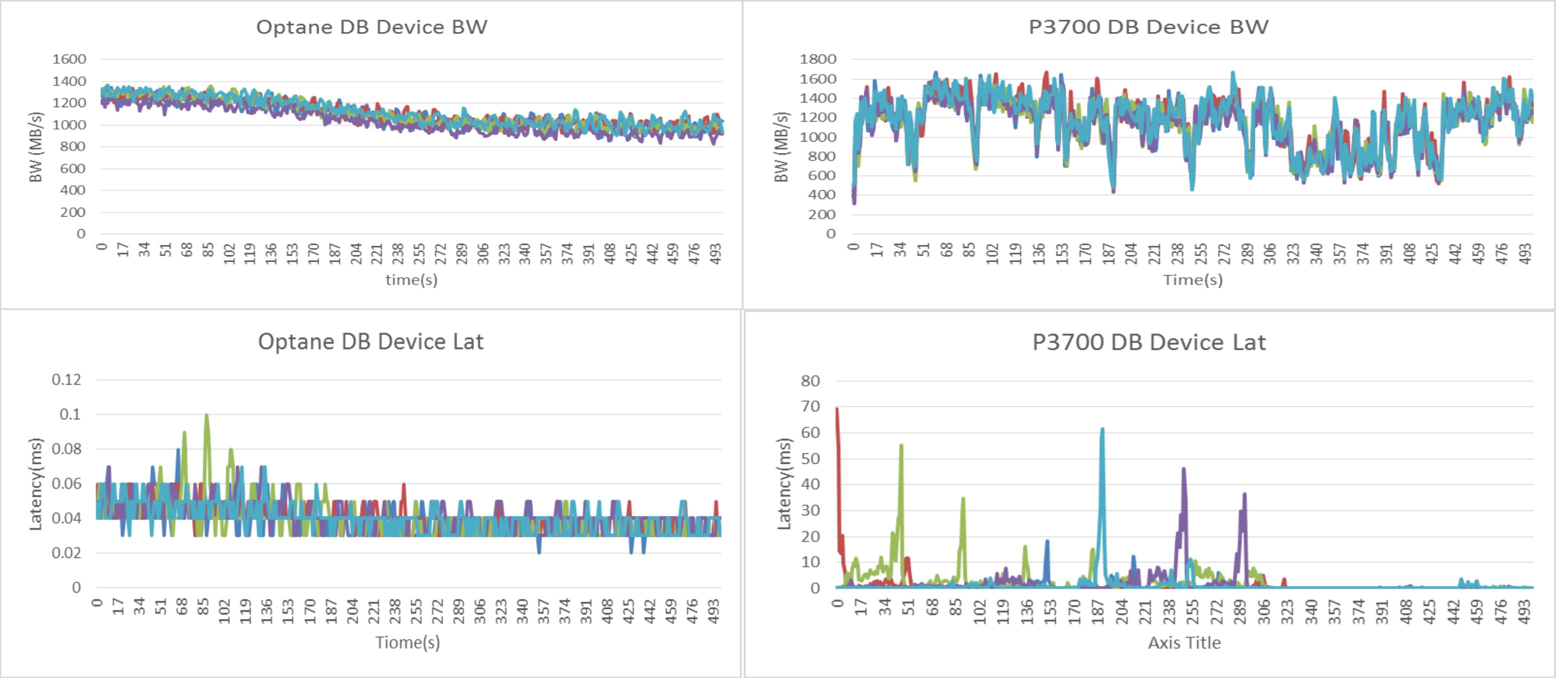
- Intel Xeon processor E5-2699 v4 @ 2.3 GHz
- 256GB Memory
- 1x 400G SSD for OS
- 1x Intel® DC P4800 375G SSD as WAL and rocksdb
- 8x 4.0TB Intel® SSD DC P4500 as data drive
- 2 OSD instances one each P4500 SSD
- Ceph 12.0.0 with Ubuntu 16.04LTS

Performance Results:

	Throughput	Latency (avg.)	99.99% latency
4K Random Read	2876K IOPS	0.9 ms	2.25
4K Random Write	610K IOPS	4.0 ms	25.435
64K Sequential Read	27.5 GB/s	7.6 ms	13.744
64K Sequential Write	13.2 GB/s	11.9 ms	215

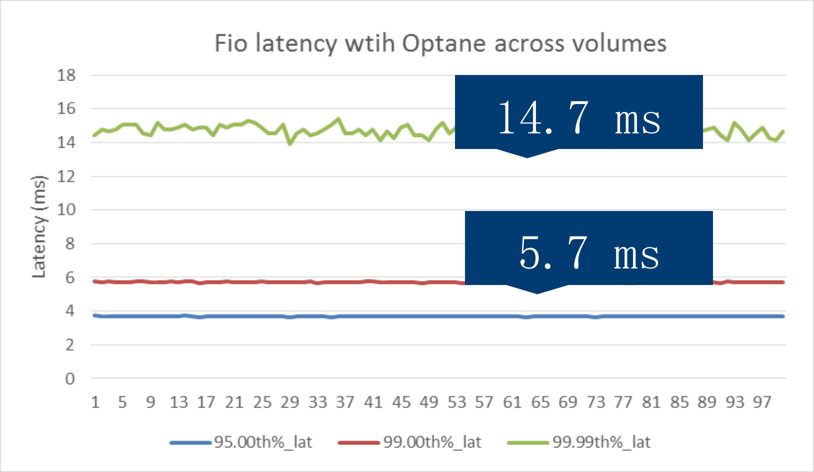
- Excellent performance on Optane cluster, performance was throttled by HW bottlenecks

Ceph* Performance – Performance improvement



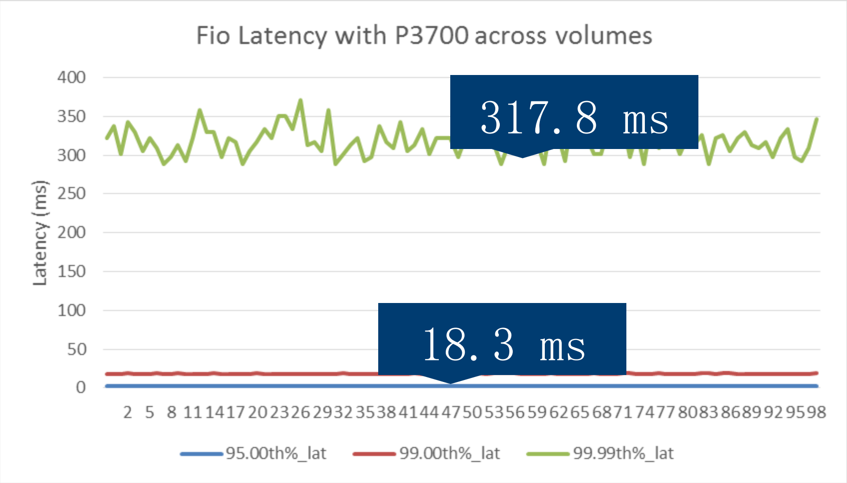
- The breakthrough high performance of Optane eliminated the WAL & rocksdb bottleneck
  - 1 P4800X or P3700 covers up to 8x P4500 data drivers as both WAL and rocksdb

Ceph* Performance – latency improvement


Fio latency wtih Optane across volumes

14.7 ms

5.7 ms

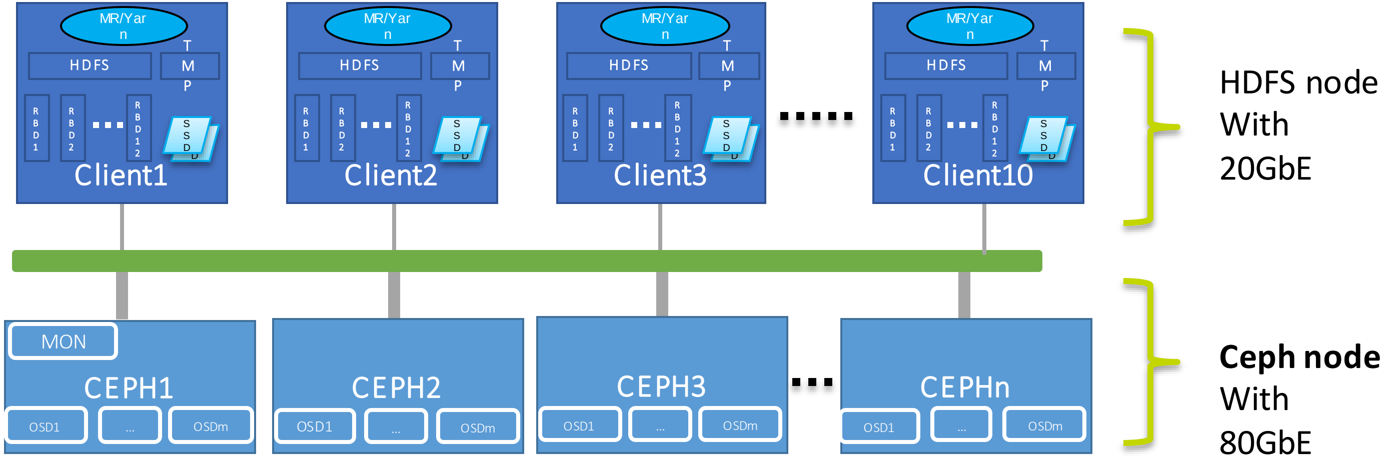

Fio Latency with P3700 across volumes

317.8 ms

18.3 ms

- Significant tail latency improvement with Optane
  - 20x latency reduction for 99.99% latency

Big Data On Ceph



Note:
Ceph node: 2x2683v4, 256GB
80+x SATA HDD, NVMe
HDFS node: 2x2683v3, 256GB,
4x SATA SSD

- Separate Compute and Storage for stability
- HDFS backend with Ceph

Summary

- Storage Innovations: Optane + 3D TLC SSDs = high performance + cost effective storage
- Better performance on Spark SQL with SSDs
- All flash Ceph is being used as backend storage for high IOPS/SLA sensitive workloads such as OLTP, SQL DB etc
- Big data over Ceph for scalability and performance

THANK YOU!

飞天·智能

APSARA INTELLIGENCE