White Paper

Supermicro All-Flash NVMe Solution for Ceph Storage Cluster

High Performance Ceph Reference Architecture Based on Ultra SuperServer® and Micron 9300 MAX NVMe SSDs

Table of Contents

2 **Powering Ceph Storage Cluster with Supermicro All-Flash NVMe Storage Solutions**

4 **Supermicro Ceph OSD Ready All-Flash NVMe Reference Architecture**

Planning Consideration

Supermicro NVMe Reference Architecture Design

Configuration Consideration

Ultra All-Flash NVMe OSD Ready System

Micron 9300 MAX NVMe SSDs

Ceph Storage Cluster - Software

7 **Performance Testing**

Establishing Baseline Performance

Testing Methodology

8 **4K Random Workload FIO Test Results**

4K Random Write Workload Analysis

4K Random Read Workload Analysis

4K Random 70Read/30Write Workload Analysis

11 **4M Object Workloads Test Result**

4MB Object Write Workload Analysis

4MB Object Read Workload Analysis

13 **Summary**

Super Micro Computer, Inc.
980 Rock Avenue
San Jose, CA 95131 USA
www.supermicro.com

Powering Ceph Storage Cluster with Supermicro All-Flash NVMe Storage Solutions

NVMe, an interface specification for accessing non-volatile storage media via PCI Express (PCI-E) bus, is able to provide up to 70% lower latency and up to six times the throughput/IOPs when compared with standard SATA drives. Supermicro has a wide range of products supporting NVMe SSDs, from 1U rackmount to 4U SuperStorage servers, to 7U 8-socket mission critical servers and to 8U high-density SuperBlade® server solutions.

In this white paper, we investigate the performance characteristics of a Ceph cluster provisioned on all-flash NVMe based Ceph storage nodes based on configuration and performance analysis done by Micron Technology, Inc. Results published with permission. Results based on testing with Supermicro SuperServer 1029U-TN10RT. Other models and configurations may produce different results.

This type of deployment choice is optimized for I/O intensive applications requiring the highest performance for either block or object storage. For different deployment scenarios and optimization goals, please refer to the table below:

Goals	Recommendation	Optimizations
Cost Effectiveness + High Storage Capacity & Density	Supermicro 45/60/90-bay SuperStorage Systems	High capacity SATA HDDs/SSDs for maximum storage density
Accelerated Capacity & Density	45/60/90 bays SuperStorage Systems with NVMe Journal	By utilizing a few NVMe SSDs as Ceph Journal, the responsiveness of a Ceph cluster can be greatly improved while still having the capacity benefits
High Performance/ IOPS Intensive	All-Flash NVMe Server	Achieving the highest performance with all NVMe SSDs

1U Ultra 10 NVMe

1U Ultra 20 NVMe

2U SuperStorage 48 NVMe

The table below provides some references for where Ceph Block or Object Storage are best suited for different types of workloads.

	Ceph Block Storage	Ceph Object Storage
Workloads	1. Storage for running VM Disk Volumes 2. Deploy Elastic Block Storage with On-Premise Cloud 3. Primary storage for My-SQL & other SQL database apps 4. Storage for Skype, SharePoint and other business collaboration applications 5. Storage for IT management apps 6. Dev/Test Systems	1. Image/Video/Audio Repository Services 2. VM Disk Volume Snapshots 3. Backup & Archive 4. ISO Image Store & Repository Service 5. Deploy Amazon S3 Object Storage like services with On-Premise Cloud 6. Deploy Dropbox like services within the Enterprise
Characteristics	1. Higher I/O 2. Random R/W 3. High Change Content	1. Low Cost, Scale-out Storage 2. Sequential, Larger R/W 3. Lower IOPS 4. Fully API accessible

4U SuperStorage 45-Bay + 6 NVMe

7U 8-Socket 16 U.2 NVMe + AICs

8U 10x 4-Socket SuperBlade®
Up to 14 NVMe per Blade

Supermicro Ceph OSD Ready All-Flash NVMe Reference Architecture

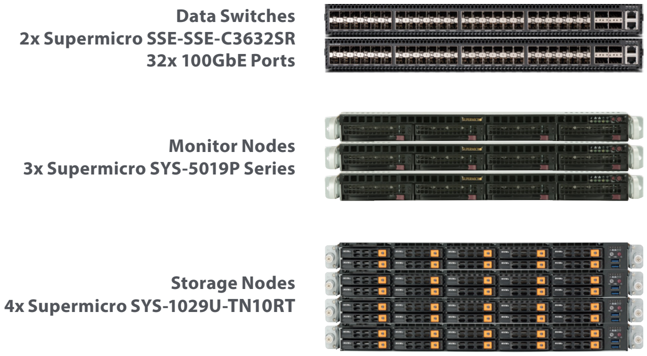
Planning Consideration

- **Number of Ceph Storage Nodes**
  At least three storage nodes must be present in a Ceph cluster to become eligible for Red Hat technical support. Ten storage nodes are the recommended scale for an enterprise Ceph cluster. Four storage nodes represent a valid building block to use for scaling up to larger deployments. This RA uses four storage nodes.

- **Number of Ceph Monitor Nodes**
  At least three monitor nodes should be configured on separate hardware. These nodes do not require high-performance CPUs. They do benefit from having SSDs to store the monitor map data. One monitor node is the minimum, but three or more monitor nodes are typically used in production deployments.

- **Replication Factor**
  NVMe SSDs have high reliability with high MTBR and low bit error rate. 2x replication is recommended in production when deploying OSDs on NVMe versus the 3x replication common with legacy storage.

Supermicro NVMe Reference Architecture Design


Data Switches
2x Supermicro SSE-SSE-C3632SR
32x 100GbE Ports
Monitor Nodes
3x Supermicro SYS-5019P Series
Storage Nodes
4x Supermicro SYS-1029U-TN10RT

Configuration Consideration

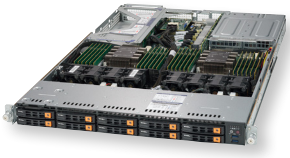
- **CPU Sizing Ceph**
  OSD processes can consume large amounts of CPU while doing small block operations. Consequently, a higher CPU core count generally results in higher performance for I/O-intensive workloads. For throughput-intensive workloads characterized by large sequential I/O, Ceph performance is more likely to be bound by the maximum network bandwidth of the cluster.

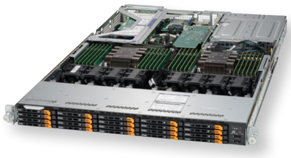
- **Ceph Configuration Tuning**
  Tuning Ceph for NVMe devices can be complex. The ceph.conf settings used in this RA are optimized for small block random performance.

- **Networking**
  A 25 GbE network is required to leverage the maximum block performance benefits of a NVMe-based Ceph cluster. For throughput-intensive workloads, 50GbE to 100GbE is recommended.

Ultra All-Flash NVMe OSD Ready System

Supermicro's Ultra SuperServer® family includes a 10- and 20-drive all-flash NVMe systems in a 1U form-factor, and 24- and 48-drive All-Flash NVMe systems in a 2U form-factor architected to offer unrivaled performance, flexibility, scalability and serviceability and therefore ideally suited to demanding enterprise-sized data processing workloads.


1U 10 NVMe


1U 20 NVMe

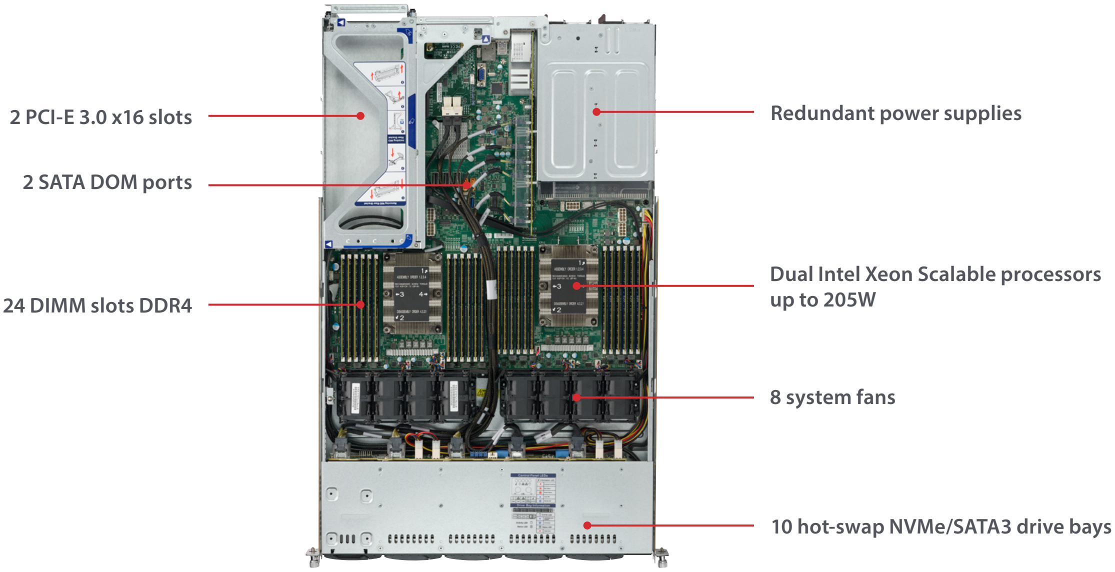
2U 24 NVMe

2U 48 NVMe

For this reference architecture, we utilized four highly flexible 10-drive systems as the Ceph Storage Nodes with the hardware configurations shown in the table below.

Table 1. *Hardware Configurations*

Component	Recommendation
Server Model	Supermicro® Ultra SuperServer® 1029U-TN10RT
Processor	2x Intel® Xeon® Platinum 8168; 24 cores, 48 threads, up to 3.7GHz
Memory	12x Micron 32 GB DDR4-2666 DIMMs, 384 GB total per node
NVMe Storage	10x Micron 9300 MAX NVMe 12.8 TB SSDs
SATA Storage	Micron 5100 SATA SSD
Networking	2x Mellanox ConnectX-5 100 GbE dual-port
Power Supplies	Redundant 1000W Titanium Level digital power supplies



Micron® 9300 MAX NVMe SSD

Micron 9300 MAX NVMe SSDs

The Micron® 9300 series of NVMe SSDs is Micron's flagship performance family with the third generation NVMe SSD controller. The 9300 family has the right capacity for demanding workloads, with capacities from 3.2TB to 15.36TB in mixed-use and read-intensive designs.

Table 2. *Micron 9300 MAX NVMe Specifications*

Model	Micron® 9300 MAX	**Interface**	PCI-E 3.0 x4
Form-Factor	U.2 SFF	**Capacity**	12.8 TB
Sequential Read	3.5 GB/s	**MTTF**	2M device hours
Sequential Write	3.5 GB/s	**Random Read**	850,000 IOPS
Endurance	144.8 PB	**Random Write**	310,000 IOPS

Note: GB/s measured using 128K transfers, IOPS measured using 4K transfers. All data is steady state. Complete MTTF details are available in Micron's product datasheet.

Ceph Storage Cluster - Software

- **Red Hat Ceph Storage 3.2**
  Red Hat collaborates with the global open source Ceph community to develop new Ceph features, then packages changes into predictable, stable, enterprise-quality releases. Red Hat Ceph Storage 3.2 is based on the Ceph community 'Luminous' version 12.2.1, to which Red Hat was a leading code contributor.

  As a self-healing, self-managing, unified storage platform with no single point of failure, Red Hat Ceph Storage decouples software from hardware to support block, object, and file storage on standard servers, HDDs and SSDs, significantly lowering the cost of storing enterprise data. Red Hat Ceph Storage is tightly integrated with OpenStack services, including Nova, Cinder, Manila, Glance, Keystone, and Swift, and it offers user-driven storage lifecycle management.

Table 3. *Ceph Storage and Monitor Nodes: Software*

Operating System	Red Hat Enterprise Linux 7.6
Storage	Red Hat Ceph Storage 3.2: Luminous 12.2.1
NIC Driver	Mellanox OFED Driver 4.5-1.0.0

Table 4. *Ceph Load Generation Nodes: Software*

Operating System	Red Hat Enterprise Linux 7.6
Storage	Red Hat Ceph Storage 3.2: Luminous 12.2.1
NIC Driver	FIO 3.1 w/ librbd enabled Mellanox OFED Driver 4.5-1.0.0

Performance Testing

Establishing Baseline Performance

Local NVMe performance shall be measured for both 4KB blocks and 4MB objects for reference.

4KB Block

Each storage node was tested using FIO across all 10 9300 MAX 12.8TB NVMe SSDs. 4KB random writes were measured with 8 jobs at a queue depth of 10. 4KB random reads were measured with 50 jobs at a queue depth of 10.

Table 5. *4KB Random Workloads: FIO on 10x 9300 MAX NVMe SSDs*

Storage Node	Write IOPS	Write Avg. Latency (ms)	Read IOPS	Read Avg. Latency (ms)
Node 1	6.47M	0.12	5.95M	0.13
Node 2	6.47M	0.11	5.93M	0.13
Node 2	6.47M	0.12	5.95M	0.13
Node 4	6.47M	0.12	5.93M	0.13

4MB Object

4MB writes were measured with 8 jobs at a queue depth of 8. 4MB reads were measured with 8 jobs at a queue depth of 8.

Table 6. *4MB Workloads: FIO on 10x 9300 MAX NVMe SSDs*

Storage Node	Write Throughput	Write Avg. Latency (ms)	Read Throughput	Read Avg. Latency (ms)
Node 1	32.7 GB/s	0.31	32.95 GB/s	0.31
Node 2	32.66 GB/s	0.31	32.95 GB/s	0.31
Node 2	32.69 GB/s	0.31	32.95 GB/s	0.31
Node 4	32.67 GB/s	0.31	32.95 GB/s	0.31

Testing Methodology

- **4KB Random Workloads: FIO + RBD**
  4KB random workloads were tested using the FIO synthetic IO generation tool and the Ceph RADOS Block Device (RBD) driver. 100 RBD images were created at the start of testing. When testing on a 2x replicated pool, the RBD images were 75GB each (7.5TB of data); on a 2x replicated pool, that equals 15TB of total data stored. When testing on a 3x replicated pool, the RBD images were 50GB each (5TB

of data); on a 3x pool, that also equals 15TB of total data stored. The four storage nodes have a combined total of 1.5TB of DRAM, which is 10% of the dataset size.

- **4MB Object Workloads: RADOS Bench**
  RADOS Bench is a built-in tool for measuring object performance. It represents the best-case object performance scenario of data coming directly to Ceph from a RADOS Gateway node. Object workload tests were run for 10 minutes, three times each. Linux file system caches were cleared between each test. The results reported are the averages across all test runs.

4K Random Workload FIO Test Results

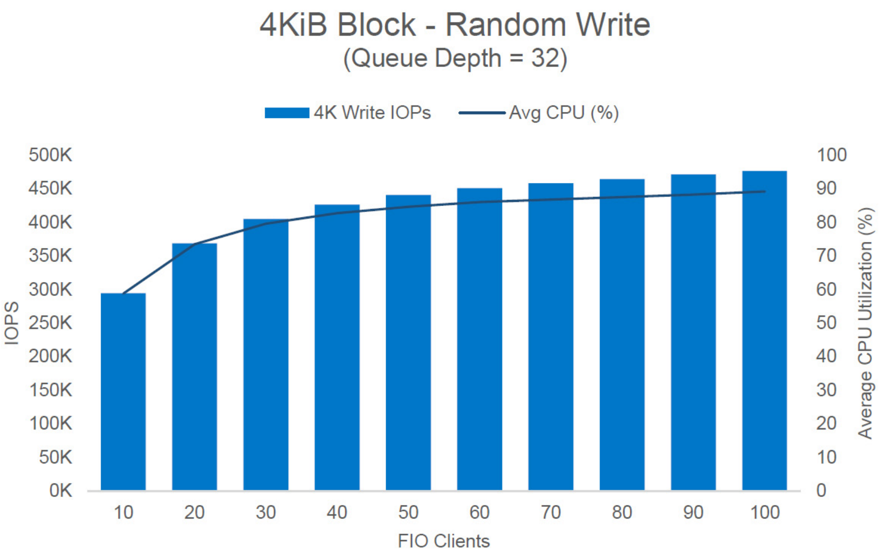
4KB random workloads were tested using the FIO synthetic IO generation tool and the Ceph RADOS Block Device (RBD) driver

4K Random Write Workload Analysis

4KB random writes reached a maximum of 477K IOPS at 100 clients. Average latency ramped up linearly with the number of clients, reaching a maximum of 6.71ms at 100 clients. Figures show that IOPS increased rapidly, then flattened out at 50 – 60 clients. At this point, Ceph is CPU-limited. 99.99% tail latency increased linearly up to 70 clients, then spiked upward, going from 59ms at 70 clients to 98.45ms at 80 clients.

Table 7. *4KB Random Write Results*

FIO Clients	4KB Random Write IOPS	Average Latency	95% Latency	99.99% Latency	Average CPU Util.
10 Clients	294,714	1.08 ms	1.48 ms	22.97 ms	58.8%
20 Clients	369,092	1.73 ms	2.60 ms	34.75 ms	73.6%
30 Clients	405,353	2.36 ms	4.09 ms	40.03 ms	79.6%
40 Clients	426,876	3.00 ms	6.15 ms	44.84 ms	82.8%
50 Clients	441,391	3.62 ms	8.40 ms	50.31 ms	84.7%
60 Clients	451,308	4.25 ms	10.61 ms	55.00 ms	86.1%
70 Clients	458,809	4.88 ms	12.63 ms	59.38 ms	86.8%
80 Clients	464,905	5.51 ms	14.46 ms	98.45 ms	87.6%
90 Clients	471,696	6.11 ms	16.21 ms	93.93 ms	88.3 %
100 Clients	477,029	6.71 ms	17.98 ms	70.40 ms	89.3 %

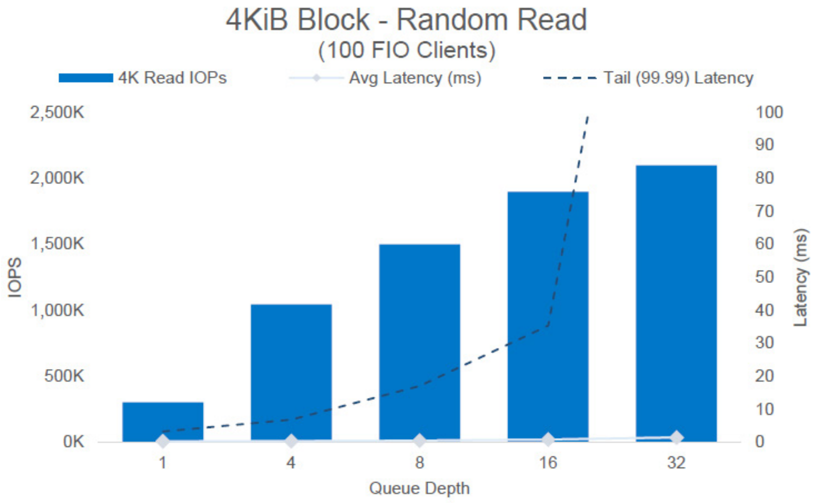


4K Random Read Workload Analysis

4KB random reads scaled from 302K IOPS up to 2 Million IOPS. Ceph reached maximum CPU utilization at a queue depth of 16; Average latency showed an increase as the queue depth increased, reaching a maximum average latency of only 0.33ms at queue depth 32. Tail latency spiked from 35.4ms at queue depth 16 to 240ms at queue depth 32, a result of CPU saturation above queue depth 16.

Table 8. *4KB Random Read Results*

Queue Depth	4KB Random Read IOPS	Average Latency	95% Latency	99.99% Latency	Average CPU Util.
QD 1	302,598	0.33 ms	2.00 ms	3.30 ms	14.1%
QD 2	1,044,447	0.38 ms	2.52 ms	6.92 ms	49.9%
QD 8	1,499,703	0.53 ms	3.96 ms	17.10 ms	75.3%
QD 16	1,898,738	0.84 ms	8.87 ms	35.41 ms	89.3%
QD 32	2,099,444	1.52 ms	20.80 ms	240.86 ms	93.2%

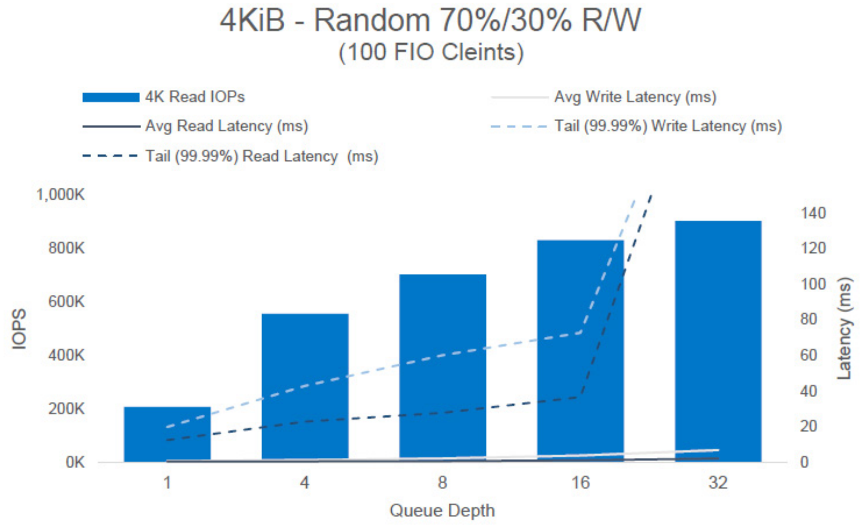


4K Random 70Read/30Write Workload Analysis

70% read/30% write testing scaled from 207K IOPS at a queue depth of 1 to 904K IOPS at queue depth 32. Read and write average latencies are graphed separately, with maximum average read latency at 2.11ms and max average write latency at 6.85ms. Tail latency spiked dramatically above queue depth 16, as the 70/30 R/W test hit CPU saturation. Above queue depth 16, there was a small increase in IOPS and a large increase in latency.

Table 9. *4KB 70/30 Random Read/Write Results*

Queue Depth	R/W IOPS	Avg. Read Latency	Avg. Write Latency	99.99% Read Latency	99.99% Write Latency	Avg. CPU Util
QD 1	207,703	0.72 ms	0.37 ms	19.83 ms	12.45 ms	19.38%
QD 2	556,369	1.23 ms	0.49 ms	42.86 ms	22.91 ms	61.00%
QD 8	702,907	2.12 ms	0.71 ms	60.13 ms	27.77 ms	77.92%
QD 16	831,611	3.77 ms	1.13 ms	72.67 ms	36.56 ms	88.86%
QD 32	903,866	6.85 ms	2.11 ms	261.93 ms	257.25 ms	92.42%



4M Object Workloads Test Result

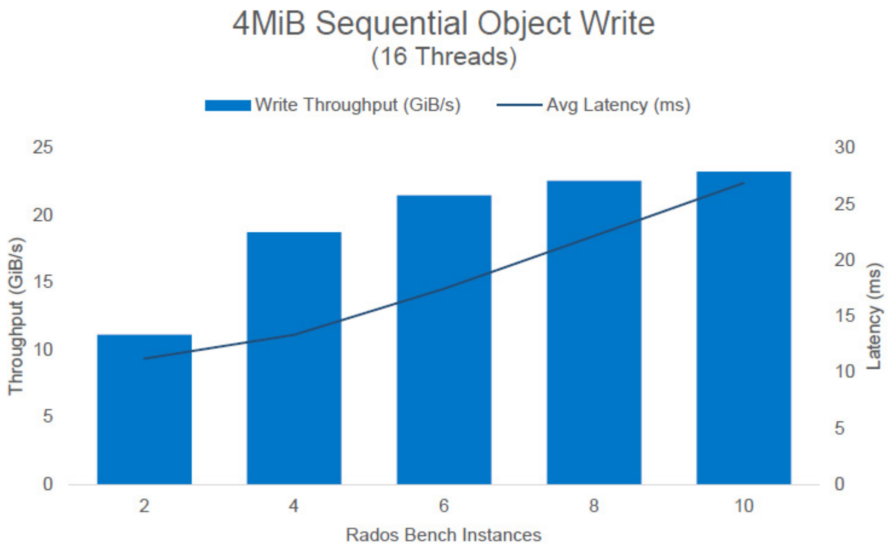
Object workloads were tested using RADOS Bench, a built-in Ceph benchmarking tool. These results represent the best-case scenario of Ceph accepting data from RADOS Gateway nodes. The configuration of RADOS Gateway is out of scope for this RA.

4MB Object Write Workload Analysis

Writes were measured by using a constant 16 threads in RADOS Bench and scaling up the number of clients writing to Ceph concurrently. Reads were measured by first writing out 7.5TB of object data with 10 clients, then reading back that data on all 10 clients while scaling up the number of threads used.

Table 10. *4MB Object Write Results*

Clients @ 16 Threads	Write Throughput	Average Latency (ms)
2 Clients	12.71 GB/s	11.22
4 Clients	20.12 GB/s	13.34
6 Clients	23.06 GB/s	17.46
8 Clients	24.23 GB/s	22.15
10 Clients	24.95 GB/s	26.90

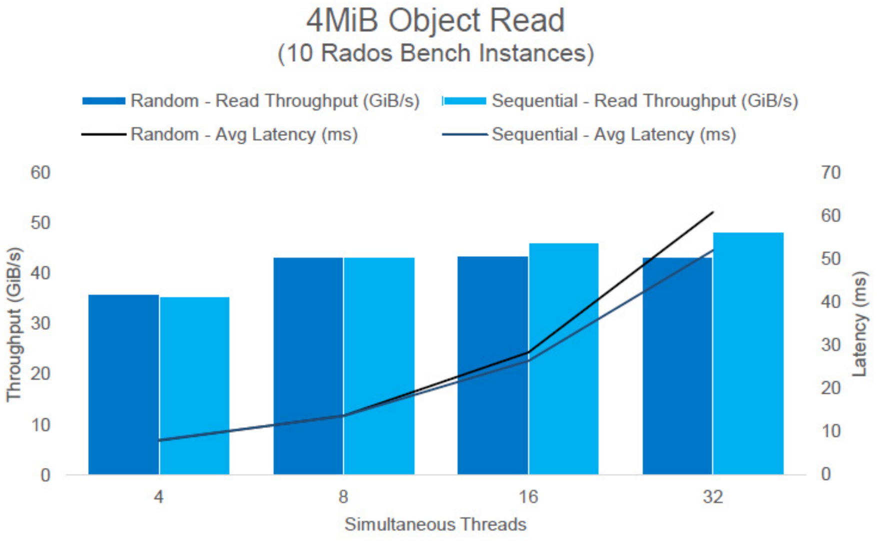


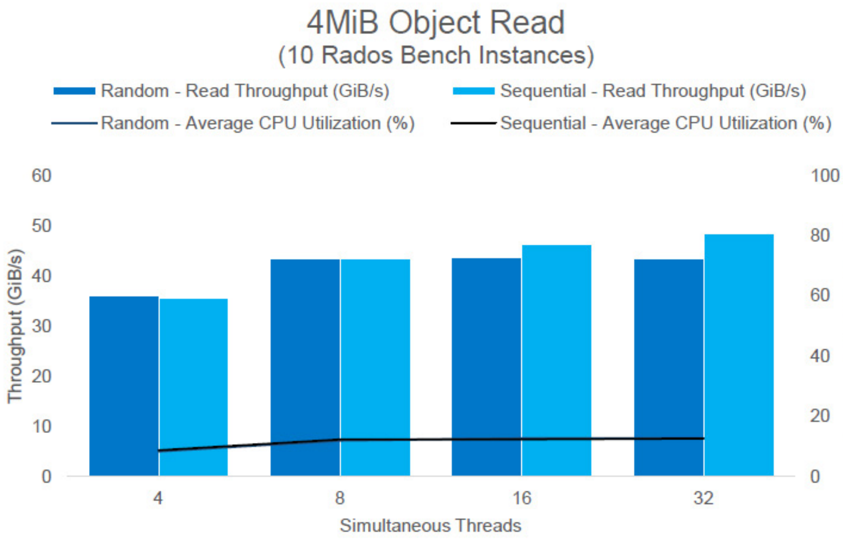
4MB Object Read Workload Analysis

4MB Object reads reached their maximum of 48.4 GB/s and 26.5ms average latency at 16 clients. CPU utilization was low for this test and never reached above 50% average CPU%.

Table 11. *4MB Object Read Results*

10 Clients @ Varied Threads	Read Throughput	Average Latency (ms)
4 Threads	38.74 GB/s	8.38%
8 Threads	46.51 GB/s	11.99%
16 Threads	46.86 GB/s	12.75%
32 Threads	46.65 GB/s	12.77%





Summary

NVMe SSD technologies can significantly improve Ceph performance and responsiveness, especially where Ceph Block Storage is preferred where massive amount of small block random access are found.

Supermicro has the most optimized NVMe enabled server and storage systems in the industry across different form-factors and price points. Please visit www.supermicro.com/nvme to learn more.

About Super Micro Computer, Inc.

Supermicro® (NASDAQ: SMCI), the leading innovator in high-performance, high-efficiency server technology is a premier provider of advanced server Building Block Solutions® for Data Center, Cloud Computing, Enterprise IT, Hadoop/Big Data, HPC and Embedded Systems worldwide. Supermicro is committed to protecting the environment through its "We Keep IT Green®" initiative and provides customers with the most energy-efficient, environmentally-friendly solutions available on the market.

Learn more on www.supermicro.com

About Micron Technology, Inc.

We are an industry leader in innovative memory and storage solutions. Through our global brands — Micron®, Crucial® and Ballistix® — our broad portfolio of high-performance memory and storage technologies, including DRAM, NAND, NOR Flash and 3D XPoint™ memory, is transforming how the world uses information to enrich life. Backed by 40 years of technology leadership, our memory and storage solutions enable disruptive trends, including artificial intelligence, machine learning, and autonomous vehicles, in key market segments like cloud, data center, networking, mobile and automotive. Our common stock is traded on the NASDAQ under the MU symbol. To learn more about Micron Technology, Inc., visit www.micron.com.

Learn more at www.micron.com

Printed in USA ♻ Please Recycle 14_Ceph-Ultra_190912_Rev2