WHITE PAPER

High-performance All Flash Ceph Cluster on Supermicro X12 Cloud DC platform

Optimize Ceph cluster block storage performance by combining Supermicro® CloudDC servers and Ceph Storage with 3rd Gen Intel® Xeon® Scalable Processors

TABLE OF CONTENTS

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

Supermicro CloudDC with 3rd Gen Intel Xeon Processors

Executive Summary

Enterprise storage infrastructure and related technologies continue to evolve year after year. In particular, as IoT, 5G, AI, and ML technologies are gaining attention, the demand for Software-defined storage (SDS) solutions based on clustered storage servers is also increasing; Ceph is a leading SDS solution that enables high performing workloads to run efficiently. The high throughput and low latency features of modern storage devices such as PCI Gen 4 NVMe, Optane, and a variety of PCIe network cards are essential factors that improve the overall performance of Ceph clusters. Adopting a Ceph cluster that utilizes NVMe Solid State Disks (SSD) maximizes the overall application performance. A team of Supermicro experts designed and assembled a Ceph cluster and then conducted various tests to validate these configurations. The clusters used in the benchmarks are based on the Supermicro CloudDC SYS-620C-TN12R, an all-flash storage server with the 3rd Gen Intel® Xeon® Scalable Processors with all-flash NVMe SSDs.

Introduction to Ceph Community Edition

Whether the need is to provide Ceph Object Storage and/or Ceph Block Device services to Cloud Platforms, deploy a Ceph File System or use Ceph for another purpose, all Ceph Storage Cluster deployments begin with setting up each Ceph Node, your network, and the Ceph Storage Cluster.

Monitors: A Ceph Monitor (ceph-mon) maintains maps of the cluster state, including the monitor map, manager map, the OSD map, the MDS map, and the CRUSH map. These maps are critical cluster states required for Ceph daemons to coordinate with each other. Monitors are also responsible for managing authentication between daemons and clients. At least three monitor nodes are generally needed for redundancy and high availability.

Managers: A Ceph Manager daemon (ceph-mgr) is responsible for keeping track of runtime metrics and the current state of the Ceph cluster, including storage utilization, current performance metrics, and system load. The Ceph Manager daemons also host python-based modules to manage and expose Ceph cluster information, including a web-based Ceph Dashboard and REST API. At least two manager nodes are typically required for high availability.

Ceph OSDs: A Ceph OSD (object storage daemon, ceph-osd) stores data, handles data replication, recovery, rebalancing, and provides some monitoring information to Ceph Monitors and Managers by checking other Ceph OSD Daemons for a heartbeat. At least three Ceph OSD nodes are normally required for redundancy and high availability.

MDSs: A Ceph Metadata Server (MDS, ceph-mds) stores metadata on behalf of the Ceph File System (i.e., Ceph Block Devices and Ceph Object Storage do not use MDS). Ceph Metadata Servers allow POSIX file system users to execute basic commands (like ls, find, etc.) without placing an enormous burden on the Ceph Storage Cluster.

Ceph stores data as objects within logical storage pools. Using the CRUSH algorithm, Ceph calculates which placement group should contain the object and further calculates which Ceph OSD Daemon should store the placement group. The CRUSH algorithm enables the Ceph Storage Cluster to scale, rebalance, and recover dynamically.

Ceph Block Device

A block is a sequence of bytes (often 512). Block-based storage interfaces are a mature and common way to store data on media, including HDDs, SSDs, CDs, floppy disks, and even tape. The ubiquity of block device interfaces is a perfect fit for interacting with mass data storage, including Ceph.

Ceph block devices are thin-provisioned, resizable, and store data striped over multiple OSDs. Ceph block devices leverage RADOS capabilities, including snapshots, replication, and strong consistency. Ceph block storage clients communicate with Ceph clusters through kernel modules or the librbd library.

Ceph's block devices deliver high performance with vast scalability to kernel modules or KVMs such as QEMU and OpenStack that rely on libvirt and QEMU to integrate with Ceph block devices. Specifically, organizations that have chosen a private or hybrid cloud model can provide NVMe-backed Ceph RADOS Block Device (RBD) storage at a price point that is even more favorable than public cloud offerings while retaining essential performance characteristics. You can use the same cluster to simultaneously operate the Ceph RADOS Gateway, the Ceph File System, and Ceph block devices.

Supermicro Benchmarking Setup

Supermicro has run several performance tests with the following setup. Figure 1 shows the Supermicro architecture with three monitor nodes, five Object Storage Daemon (OSD) nodes, and 10 RADOS Block Devices (RBD) load-gen client nodes.

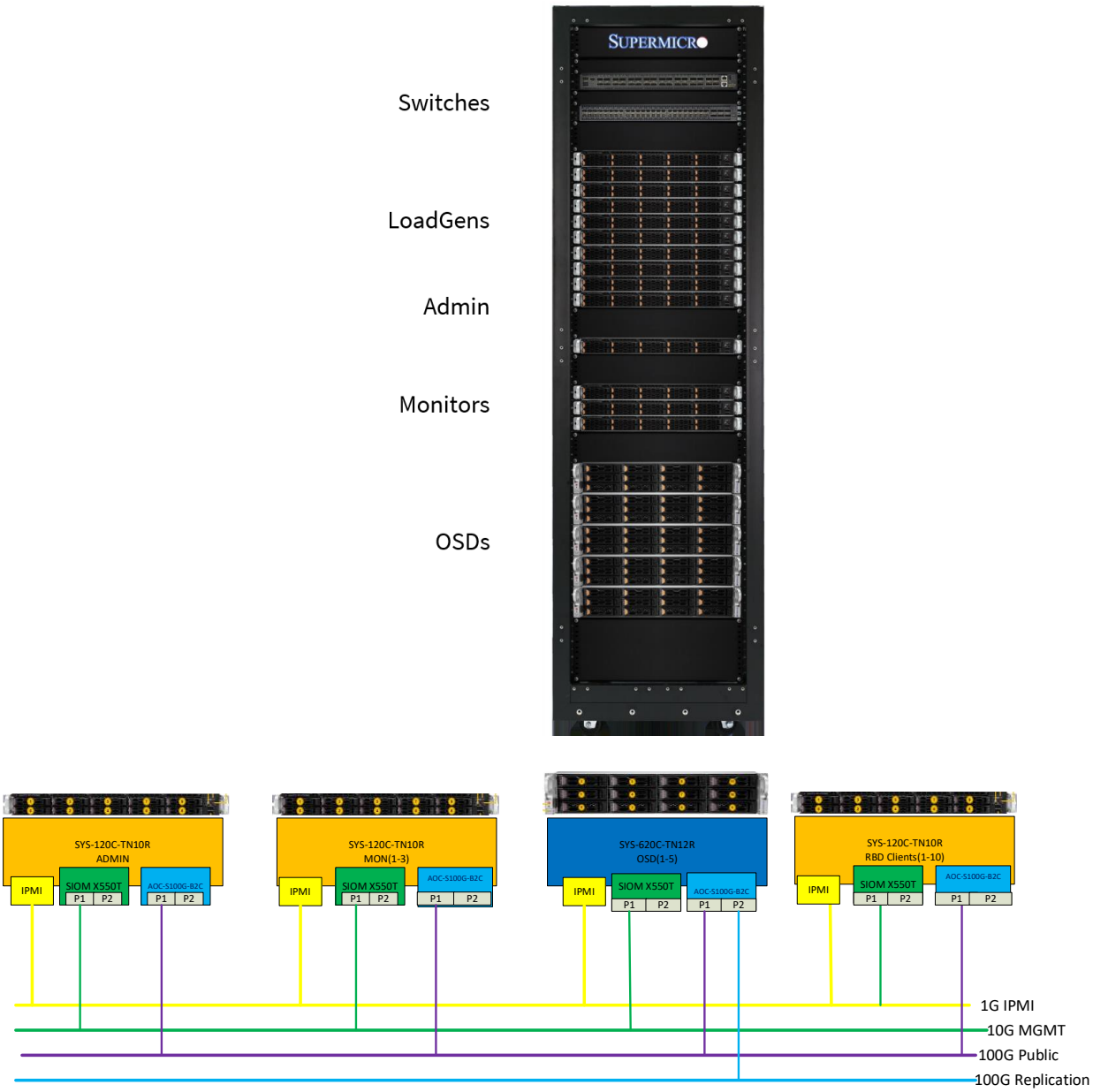

Figure 1– Supermicro Ceph benchmarking setup

Supermicro Hardware BOMS and Software Specifics

Software	Component	Value
Ceph	Version	Ceph Octopus 15.2.8
	Config	/etc/ceph/ceph.conf
	Deployment	cephadm and cephorch
	PG num	4096
	#OfOSD Daemons/NVMe	2
	Pool size	2
FIO	Version	3.25
	IO Engine	RBD
	Ramp time	300 seconds
	Run time	600 seconds
	RDB Size	100 GB
	Block Size	4KB (Random), 128KB (Sequential)
	Read/write mix	70/30 Read/Write
	IO Depth	1-32 (Random) 32 (Sequential)
	Num jobs	1
	Num Of RBD clients	200

Table 1 - Ceph-FIO settings

The Ceph Storage cluster is deployed on the Supermicro CloudDC servers containing 3rd Gen Intel® Xeon® Scalable Processors. The software versions used were Ceph Octopus on SUSE Linux Enterprise Server 15 SP2, and Flexible I/O Tester (fio) 3.25

System / Component	Part Number	Qty / Node	Part Description
Server	**SYS-120C-TN10R**	**1**	**Admin x1 Monitor x3 RBD Clients x10**
CPU	P4X-ICX6330-SRKHM	2	ICX 6330 2P 28C/56T 2.0G 42M 11.2GT 205W 4189 D2
Memory	MEM-DR416L-HL04-ER29	16	16GB DDR4-2933 2Rx8 ECC REG DIMM
NVMe M.2(OS)	HDS-SMP-HFS7T6GETFEID430	2	KXG60ZNV1T02 PCIe® Gen3 x4, NVMe™ 1.3a 1TB
AOC	AOC-S100G-b2C	1	Two QSFP28 100Gbps Ethernet port PCIe 4.0 x 16 host interface,RoHS
AOC	AOC-ATG-i2TM	1	AIOM 2-port 10GBase-T, Intel X550,RoHS
Software License	SFT-DCMS-SINGLE	1	Supermicro System Management Software Suite node license, HF, RoHS/REACH, PBF
Server	**SYS-620C-TN12R**	**1**	**OSD x5**
CPU	P4X-ICX8368-SRKH8	2	ICX 8368 2P 38C/76T 2.4G 57M 11.2GT 270W 4189 D2
Memory	MEM-DR432L-HL03-ER32	16	SK Hynix 32GB DDR4-3200 2Rx8 (16Gb)ECC REG DIMM
NVMe M.2(OS)	HDS-SMP-HFS7T6GETFEID430	2	KXG60ZNV1T02 PCIe® Gen3 x4, NVMe™ 1.3a 1TB
NVMe(OSD Drives)	HDS-SMP-KCM6XRUL3T84	12	Kioxia CM6 3.84TB NVMe PCIe 4x4 2.5" 15mm SIE 1DWP
AOC	AOC-S100G-b2C	1	BCM57508 NetXtreme-E 10Gb/25Gb/40Gb/50Gb/100Gb/200Gb Ethernet
AOC	AOC-ATG-i2TM	1	AIOM 2-port 10GBase-T, Intel X550,RoHS
Software License	SFT-DCMS-SINGLE	1	Supermicro System Management Software Suite node license, HF, RoHS/REACH, PBF

Table 2- Supermicro CloudDC Ceph Block Storage BOM

X12 CloudDC: All-in-one Rackmount Platform for Cloud Data Centers

Fig 2- CloudDC SKUs

Supermicro's X12 CloudDC platform offers tool-less mechanical design for rapid cloud deployment and easy maintenance, featuring:

- Ultimate flexibility on I/O and storage with 2 or 4 PCI-E 4.0 x16 slots and dual AIOM slots (OCP 3.0 compliant) slots for maximum data throughput, X12 CloudDC is designed to have excellent serviceability with tool-less brackets, hot-swap drive trays, and redundant power supplies that ensure rapid deployment and more efficient maintenance in data centers.
- Redundant high-efficiency Platinum/Titanium Level power supplies for resiliency and lower carbon footprint.
- Rich Security Features: TPM 1.2/2.0, signed firmware, Silicon Root of Trust, Secure Boot, System Erase, Runtime FW protection, FIPS Compliance, Trusted Execution Environment.

Baseline Test Results

The purpose of this first test is to measure the pure I/O performance of the storage at each node where the Ceph package is not installed. Each node has a 12 KIOXIA CM6 3.84TB NVMe SSD, and the performance was measured using the Fio (Flexible I/O tester) benchmark tool with the libaio IO engine. IOPS performance was evaluated for random IO workloads of a small IO size (4 KB). Sequential performance was also assessed for sequential IO workloads of a large IO size (128 KB). The test was performed three times, and the results were averaged. The table below shows the baseline test results.

	4K Random Write			
	Avg. Throughput (KIOPS)	Avg. 99.99%th Latency (ms)	Avg. Latency (ms)	Avg. Throughput (GB/s)
OSD1	7948.33	8.12	0.39	30.30
OSD2	8523.00	1.96	0.36	32.53
OSD3	8656.33	1.92	0.35	33.00
OSD4	8664.33	1.88	0.35	33.07
OSD5	8660.67	1.87	0.35	33.03

	4K Random Read			
	Avg. Throughput (M IOPS)	Avg. 99.99%th Latency (ms)	Avg. Latency (ms)	Avg. Throughput (GB/s)
OSD1	13.00	6.59	0.24	49.53
OSD2	13.50	1.24	0.23	51.57
OSD3	13.33	0.90	0.23	50.97
OSD4	13.13	0.87	0.23	50.00
OSD5	13.40	0.89	0.23	51.10
	128K Seq Write			
	Avg. Throughput (GB/s)	Avg. 99.99%th Latency (ms)	Avg. Latency (ms)	Avg. Throughput (K IOPS)
OSD1	43.77	21.98	1.07	358.00
OSD2	47.37	4.34	0.99	388.00
OSD3	47.50	4.29	0.99	389.00
OSD4	47.50	4.29	0.99	389.00
OSD5	47.50	4.29	0.99	389.00
	128K Seq Read			
	Avg. Throughput (GB/s)	Avg. 99.99%th Latency (ms)	Avg. Latency (ms)	Avg. Throughput (K IOPS)
OSD1	75.97	5.32	0.62	622.33
OSD2	69.60	0.67	2.77	570.33
OSD3	77.90	0.60	1.01	638.00
OSD4	77.90	0.60	1.04	638.00
OSD5	77.90	0.60	1.04	638.00

Table 3 -Baseline Results

Benchmark Configurations and Results

The following sections provide the results of synthetic benchmark performance for all-flash based Ceph clusters using the KIOXIA HDS-SMP-KCM6XRUL3T84 NVMe SSD. The test was conducted in the RBD-based storage pool, which is the block storage component for Ceph. Workloads were generated using the Fio benchmark with ten client servers.

Before starting the test, Supermicro engineers created 200 RBD images that generated a total of 20 TB of data. A 2x replication was applied, resulting in the total size of the data stored in the cluster being 40 TB.

• 10 Clients x 20 RBD images per client x 100 GB RBD image size = 20 TB (2x Replication: 20 TB x 2 = 40 TB)

A random test was created and run with a 4 KB small IO workload and with the number of jobs equal to 1 and IODepth 1 to 32 per Fio instance. A sequential test was also created and run with a 128 KB large IO workload with the number of jobs equal to 1 and IODepth 32 per Fio instance. We also measured latency variation across each test.

4 KB Random Read Workload

We measured the performance and latency of 4 KB random reads with increasing IODepth (1to32) on 200 clients. At an IODepth of 32, 4KB random read performance was measured at an average of 3.31 Million IOPS, with an average latency of 1.93ms and a tail latency (99.00th percentile latency) of 3.92ms. As the IODepth increased, IO performance and latency tended to increase. Maximum 3.32 Million IOPS was observed at an IODepth of 16 with an average latency of 0.96ms.

IODepth	Average Latency(ms)	IOPS(millions)	P99.00 Latency (ms)	P99.90 Latency (ms)
1	0.56	0.36 m	1.75	4.25
2	0.58	0.69 m	1.82	4.6
4	0.49	1.65 m	1.8	5.05
8	0.49	3.24 m	2.34	7.02
16	0.96	3.32 m	3.73	186.31
32	1.93	3.31 m	3.92	590.29

Table 4- 4K Random Reads

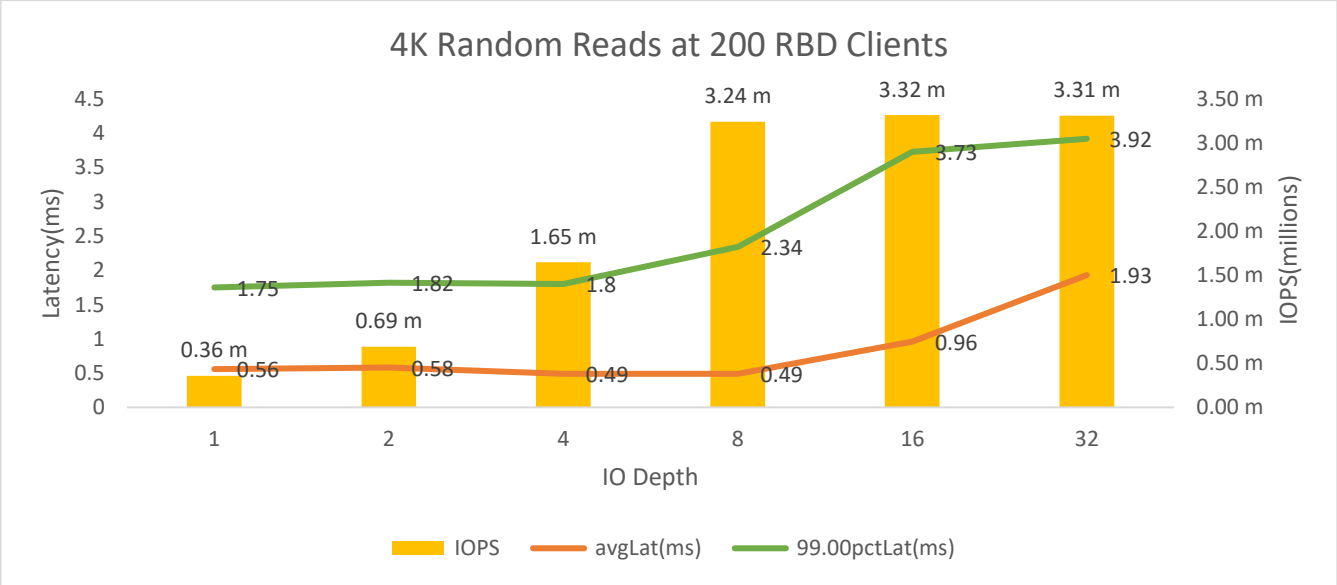

Fig 3- 4K Random Reads at 200 RBD Clients

4 KB Random Write Workload

The benchmarks measured the performance and latency of 4 KB random writes with increasing IODepth (1to32) on 200 clients. At an IODepth of 32, 4 KB random write performance was measured at an average of 641K IOPS, with an average

latency of 9.75 ms and an average tail latency (99.90th percentile latency) of 937.73ms. As IODepth increased, IO performance and latency tended to increase. Tail latency (99.90th percentile latency) increased significantly at IODepth of 32.

IODepth	Bandwidth(GB/s)	K IOPS	Average Latency (ms)	P99.00 Latency (ms)	P99.90 Latency (ms)
1	1.18	301	0.65	1.53	3.69
2	1.29	329	1.19	6.73	13.43
4	1.78	455	1.72	8.88	16.04
8	2.26	577	2.71	13.34	22.08
16	2.45	627	4.99	23.61	39.09
32	2.50	641	9.75	39.55	937.73

Table 5 – 4K Random Writes

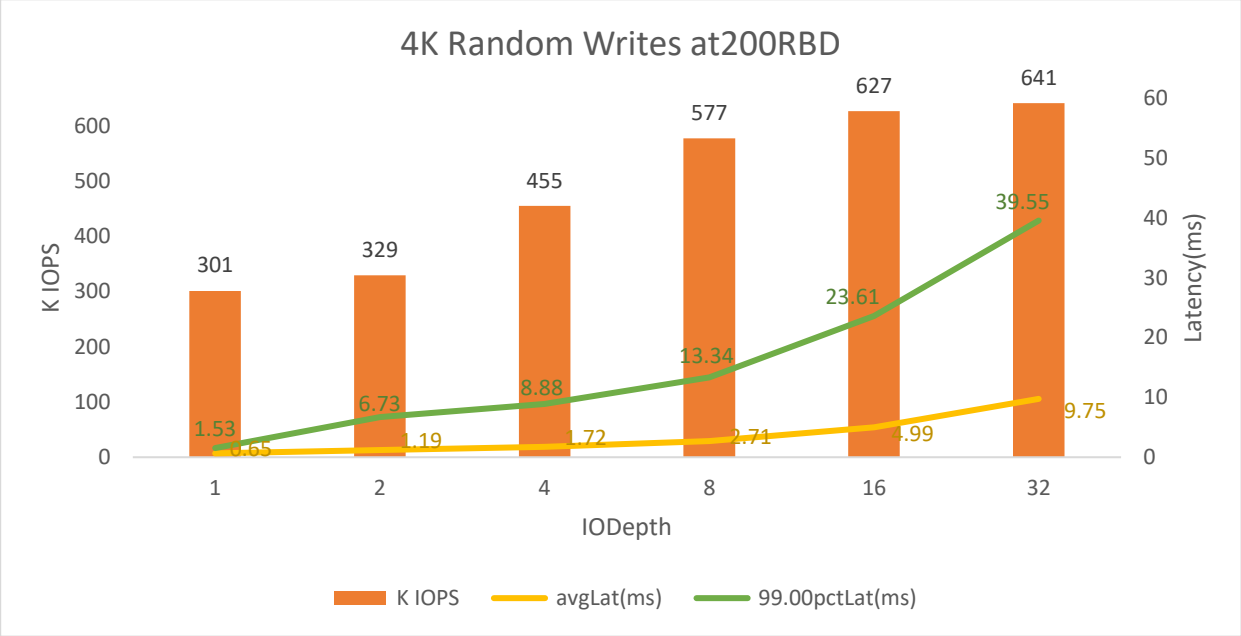

Fig 4 - 4K Random Writes at 200RBD Clients

4 KB Random ReadWrite 70/30

We measured the performance and latency of 4 KB random readwrites 70/30 with increasing IODepth (1to32) at 200 clients. At an IODepth of 32, 4 KB random readwrites 70/30 performance was measured at an average of 1.315million IOPS, with an average latency of 4.76ms and an average tail latency (99.90th percentile latency) of 1340.71ms. As IODepth increased, IO performance and latency tended to increase. Tail latency (99.90th percentile latency) increased significantly at IODepth of 16 and higher.

IODepth	Bandwidth(GB/s)	IOPS(millions)	Average Latency (ms)	P99.00 Latency (ms)	P99.90 Latency (ms)
1	1.90	0.50 m	0.4	0.86	2.23
2	3.21	0.84 m	0.47	1.27	3.08
4	4.42	1.16 m	0.69	2.66	5.56
8	5.05	1.32 m	1.21	7.53	22.32
16	5.14	1.35 m	2.37	12.15	375.84
32	5.14	1.35 m	4.76	12.45	1340.71

Table 6 – 4K Random ReadWrites 70/30

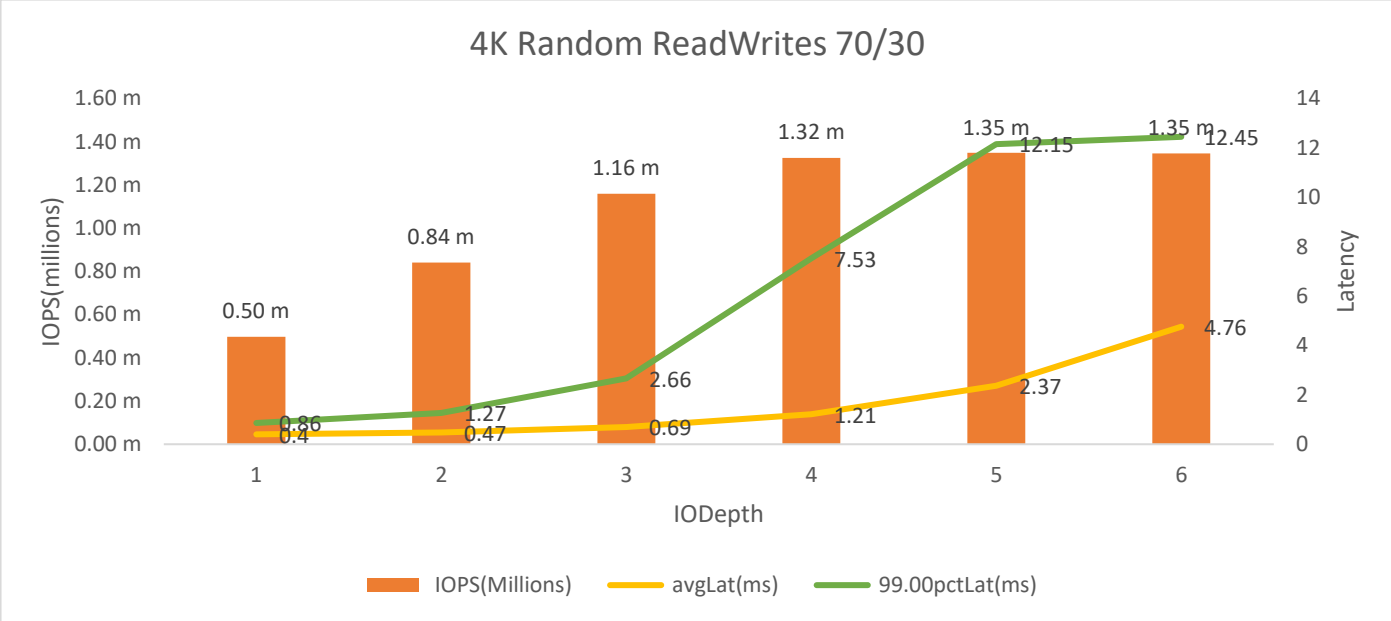

Fig 5 - 4K Random ReadWrites 70/30

Random Workloads CPU utilization

The CPU utilization has increased gradually as IO Depth increased for all three random workloads, as shown in Figures 6, 7, and 8. There's still headroom for more RBD clients, and that even though IOPS started to taper at 32 IO depth, the total aggregate throughput has room to scale with more clients.

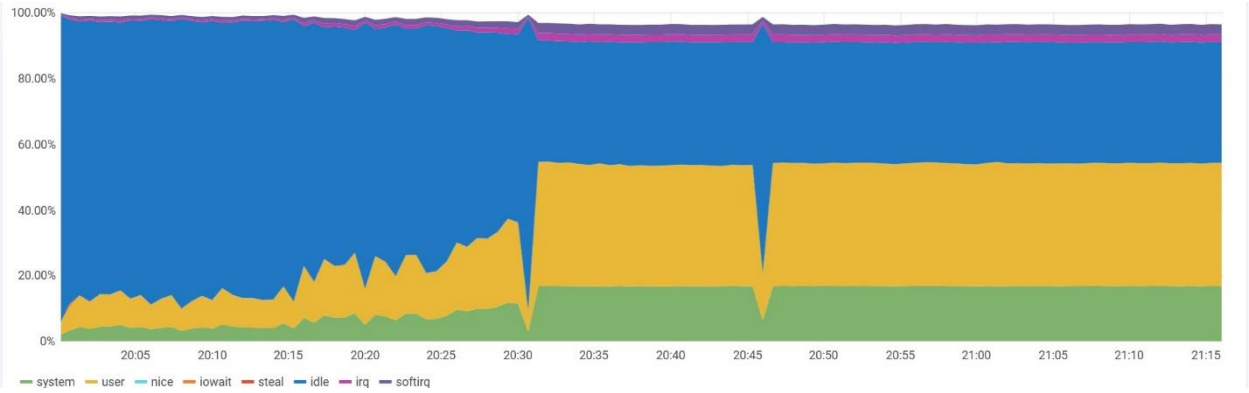

Fig 6 - OSD Nodes CPU Utilization-4K Random Reads IODepth 1-32 Jobs

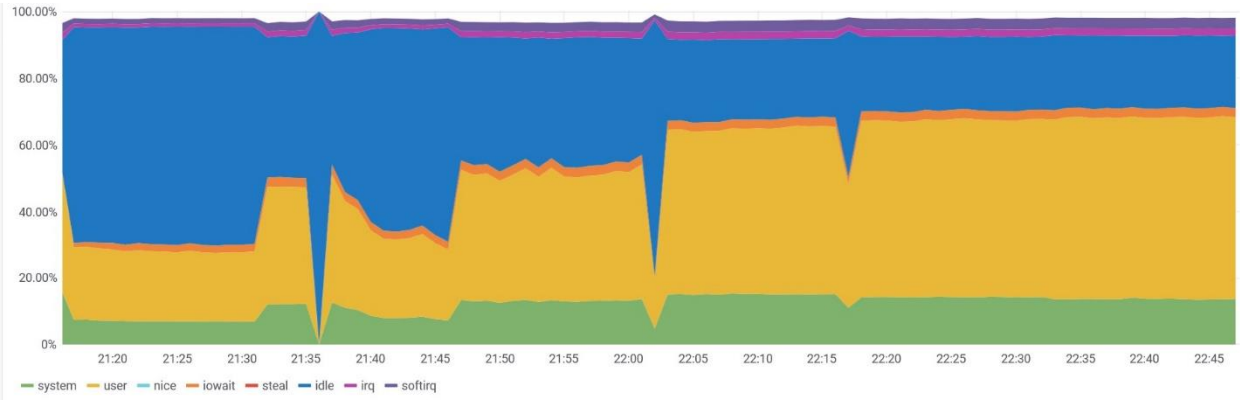

Fig 7 - 4K Random Writes OSD CPU Utilization IODepth 1-32

Fig 8 - 4K Random ReadWrites 70/30 OSD CPU Utilization IODepth 1-32

128 KB Sequential Read Workload

The maximum throughput of 54.82 GB/s was reached with180 RBD clients for the 128 KB sequential reads. Latency increased steadily as the number of clients increased, while throughput remained relatively constant once the number of clients reached 180. The sequential reads workload caused the network bandwidth to approach its limits.

Procs	Bandwidth(GB/s)	K IOPS	Average Latency (ms)	P99.00 Latency (ms)
20	44.37	363	1.76	4.49
40	51.97	426	3.01	17.31
60	52.95	434	4.43	31.44
80	54.45	446	5.74	60.04
100	53.49	438	7.3	59.58
120	53.47	438	8.79	91.07
140	52.85	433	10.35	116.04
160	54.69	448	11.43	115.64
180	54.82	449	12.83	168.53
200	52.29	428	14.94	219.73

Table 7 – 128K Seq Reads at IODepth 32

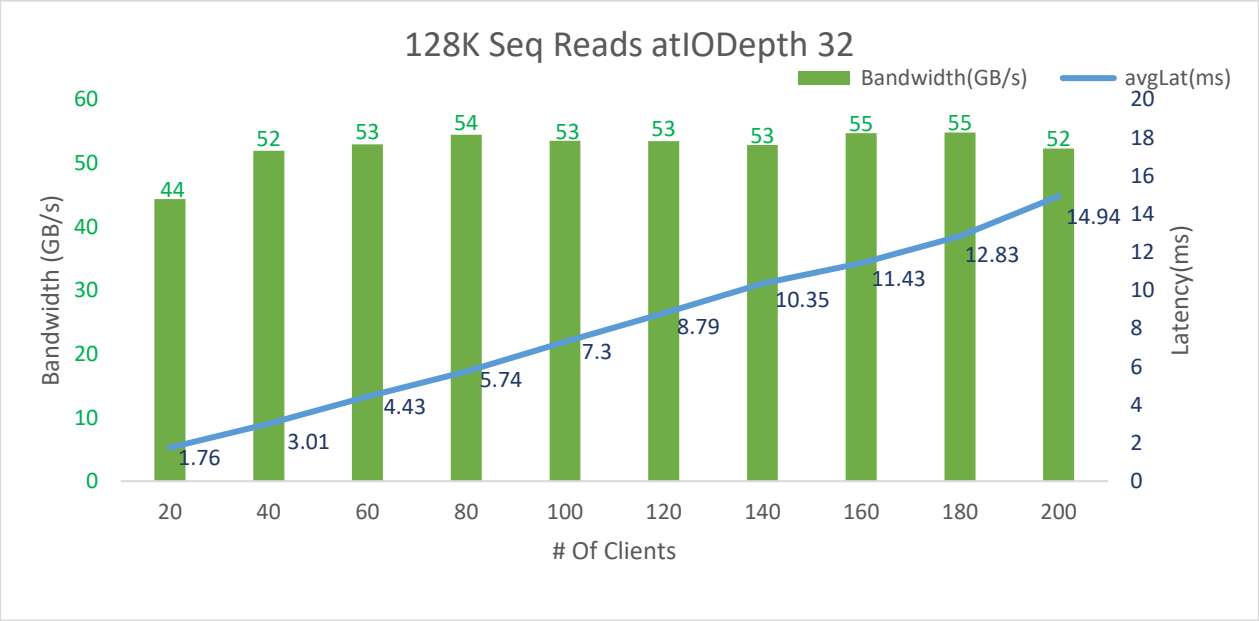

Fig 9 - 128K Seq Reads at IODepth 32

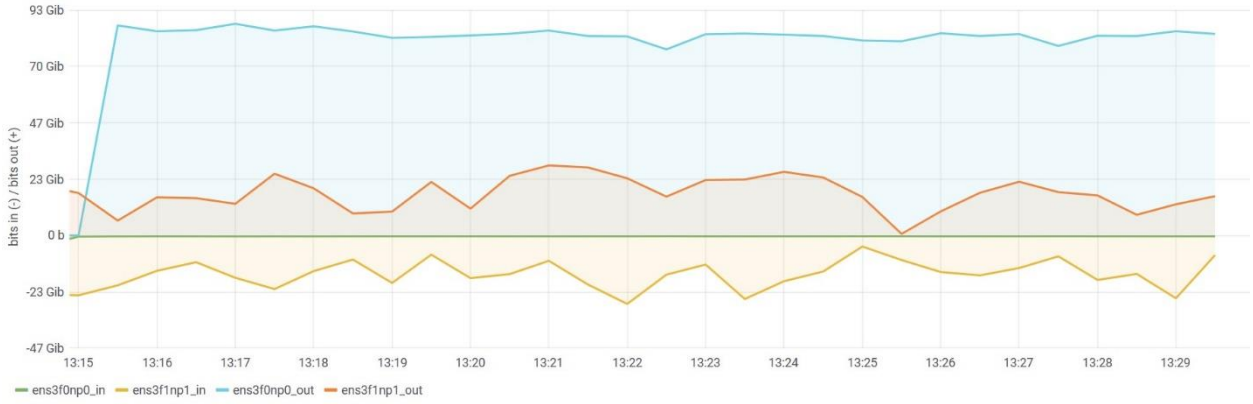

Fig 10 - 128K Seq Reads at 32 IODepth with 200RBD clients- OSD Node 100G Public/Replication Network Utilization

128 KB Sequential Write Workload

The maximum sequential write throughput of 28.36 GB/s reached at 180 RBD clients for the 128 KB write workload. Latency increased steadily as the number of clients increased, while throughput remained relatively constant once the number of clients reached 180.

Procs	Bandwidth(GB/s)	K IOPS	Average Latency (ms)	P99.00 Latency (ms)
20	12.40	102	6.3	17.63
40	18.01	147	8.68	27.27
60	21.13	173	11.09	39.05
80	23.46	192	13.32	53.99
100	24.73	203	15.8	63.66
120	25.35	208	18.49	81.07
140	25.94	212	21.08	119.16
160	26.83	220	23.29	177.78
180	28.36	232	24.8	117.73
200	27.63	226	28.28	195.01

Table 8 -128K Seq Writes at IODepth 32

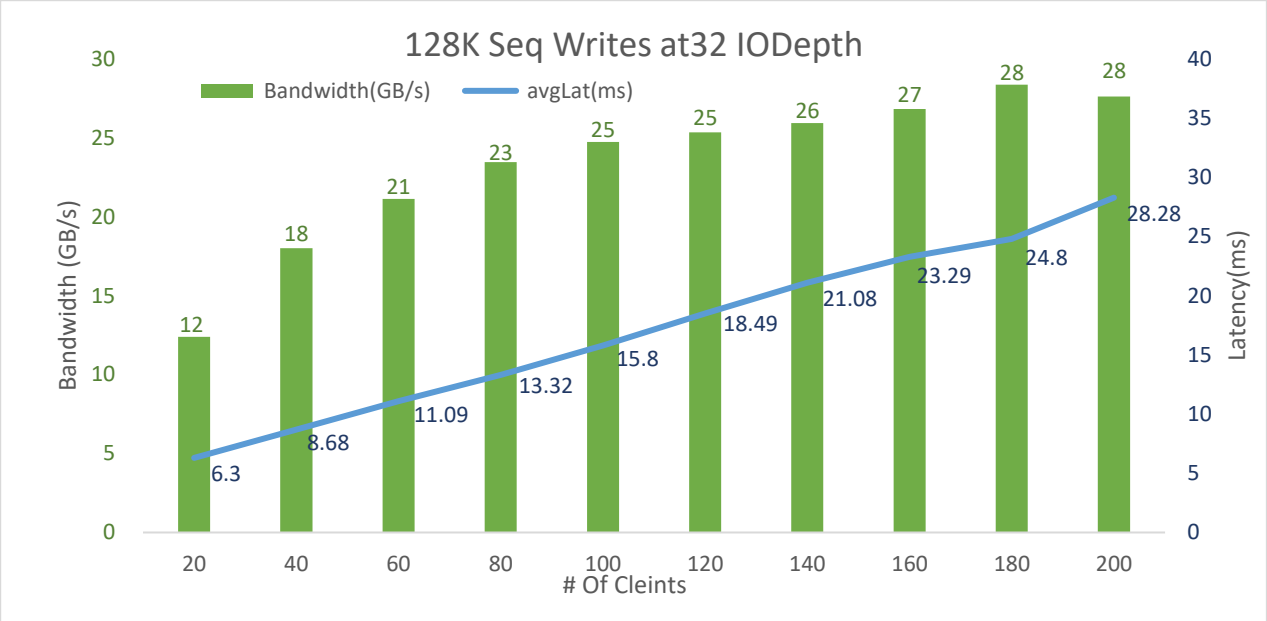

Fig 11 - 128K Seq Writes at 32 IODepth

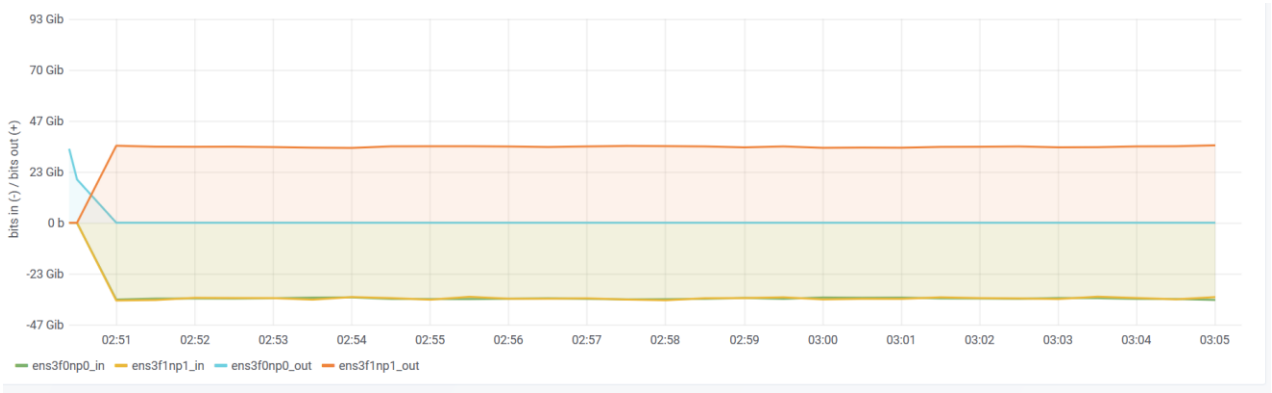

Fig 12 - 128K Seq Writes at 32 IODepth with 200RBD clients- OSD Nodes 100G Public/Replication Network Utilization

Ceph Configuration Files

ceph-conf.ini	**ceph_spec.yml**
[global]	service_type: osd
mon_max_pg_per_osd=1000	service_id: nvme_group
#mon_allow_pool_delete=true	placement:
public network 10.5.5.0/24	host_pattern: 'osd[1-5]'

cluster network 192.168.5.0/24 osd_max_pg_log_entries=10 osd_min_pg_log_entries=10 osd_pg_log_dups_tracked=10 osd_pg_log_trim_min=10 osd_enable_op_tracker=False max_open_files=500000 bluefs_buffered_io=false rgw_list_buckets_max_chunk=999999 rgw_override_bucket_index_max_shards=199 rgw_dynamic_resharding=0 rgw_bucket_index_max_aio=4096 rgw_put_obj_max_window_size=134217728 objecter_inflight_op_bytes=1048576000 objecter_inflight_ops=102400 ms_dispatch_throttle_bytes=1048576000 rgw_obj_stripe_size=262144 rgw_max_chunk_size=262144 rbd_readahead_disable_after_bytes=0 rbd_readahead_max_bytes=4194304 bluestore_default_buffered_read=false mon_allow_pool_delete=true	data_devices: model: 'KCM6XRUL3T84' osds_per_device: 2 #db_devices:

mutex_perf_counter=false throttler_perf_counter=false mutex_perf_counter=false throttler_perf_counter=false bluestore_cache_autotune=0 bluestore_rocksdb_options="compression=kNoCompression,max_write_buffer_number=32,min_write_buffer_number_to_merge=2,recycle_log_file_num=32,compaction_style=kCompactionStyleLevel,write_buffer_size=4MB,target_file_size_base=4MB,max_background_compactions=64,level0_file_num_compaction_trigger=16,level0_slowdown_writes_trigger=128,level0_stop_writes_trigger=256,max_bytes_for_level_base=512MB,compaction_threads=32,flusher_threads=8,compaction_readahead_size=2MB" bluestore_cache_meta_ratio=0.8 bluestore_cache_kv_ratio=0.2	

Summary of Ceph cluster RBD Storage Performance

Random Workload Pattern at 32 IODepth	**IOPS**	**Latency**
4K 100% Random Reads	3.31mil	1.93 ms
4K 100% Random Writes	641K	9.75 ms
4K 70%/30% Read/Write Mix	1.35 mil	4.76 ms
Seq Workload Pattern at 180 RBD Clients	**Bandwidth**	**Latency**
128K Seq Reads	54.82 GB/S	12.83 ms
128K Seq Writes	28.36 GB/S	24.8 ms

Table 9 - Summary

Conclusion

Supermicro CloudDC servers are optimized for large cloud data center deployments and enterprise environments and deliver consistently high performance, making them an ideal solution for software-defined storage such as Ceph Storage. Supermicro has designed a performance-optimizing, all-flash-based Ceph cluster using the CloudDC servers SYS-120C-TN10R & SYS-620C-TN12R. Both use the 3rd Gen Intel Xeon Scalable Processors 8638 CPUs, PCI-E 4.0 NVMe SSDs, and Ceph Storage-Octopus. This solution achieves over 3.24 million IOPS for the 4 KB random read workloads and excellent sequential read throughput.

With a balanced architecture between CPUs and optimized for scalable compute, database, GPU, tiered storage, and I/O intensive applications, cost, and performance can be optimized down to component level. The Supermicro optimized the CPU-to-drive ratios to unlock the maximum balanced bandwidth on the latest U.2 and E1.S NVMe drives with Supermicro Ultra and BigTwin™ systems. All-flash NVMe-based configurations deliver extremely high-performance storage with the highest IOPS per system and per Gigabyte to enable a rich set of data services across the IT infrastructure.

For more information, please visit www.supermicro.com