White Paper

An Introduction to PCIe® 4.0 performance with KIOXIA CM6 Series SSDs and KumoScale™ Software

1. Executive Summary

1.1 Introduction

This document contains reference material and technical documentation for product testing results using products developed by KIOXIA Corporation (hereinafter referred to as "KIOXIA").

1.2 Summary

NVM Express™ over Fabrics (NVMe-oF™) benchmark tests were conducted with NVM Express™ (NVMe™) SSDs of KIOXIA and KumoScale™ software. This document describes the purpose, configuration, methods, and test findings, to provide a benchmark for disk selection and sizing when deploying KumoScale™ software.

2. Testing Background

2.1 KIOXIA SSDs

KIOXIA SSDs are solid state drives (SSD) capable of using the latest high-speed interface, NVMe™. Table below shows the latest KumoScale™ software compatible KIOXIA SSDs.

Table 2.1: KIOXIA SSDs (as of October 2021)

Enterprise SSD	KIOXIA CM6 Series PCIe® 4.0 NVMe™ SSD
	KIOXIA CM5 Series PCIe® 3.0 NVMe™ SSD
Data Center SSD	KIOXIA CD6 Series PCIe® 4.0 NVMe™ SSD
	KIOXIA CD5 Series PCIe® 3.0 NVMe™ SSD
	KIOXIA XD5 Series PCIe® 3.0 NVMe™ SSD

2.2 NVMe-oF™ Overview

NVMe-oF™ makes NVMe™ SSDs available over an Ethernet or Fibre Channel network. This allows multiple servers to share single or multiple NVMe™ storage systems, providing end-to-end communication via the NVM Express protocol, similar to iSCSI in a SAS-based storage SAN.

It supports multiple transports, including RoCEv2 (RDMA over Converged Ethernet) and NVMe™ over TCP, that can be used depending on user requirements. This document provides data demonstrating that connecting over RoCEv2 enables fast and low latency storage connectivity.

2.3 KumoScale™ Software

KumoScale™ software is a software developed by KIOXIA for NVMe™ based storage management. It provides storage services using NVMe-oF™ technology compatible with RoCEv2 transport and TCP transport, SSD volume pooling, and per-volume performance settings for QoS.

KumoScale™ software runs on storage nodes equipped with NVMe™ SSDs, which serve as NVMe-oF™ targets which are connected and used as block storage by Linux™ initiators. NVMe-oF™ enables performance close to that of DAS (Direct Attached Storage).

Based on lab testing by KIOXIA in June 2020, comparing DAS and NVMe-oF™ configurations demonstrated only a 15 us latency difference with a block size of 4KB.

Volume management capabilities of NVMe-oF™ technology and KumoScale™ software improves overall system utilization by sharing storage, reducing waste, and enhancing flexibility as KumoScale™ software allocates required space from the shared storage pool.

KumoScale™ storage nodes are connected to initiators, such as bare metal, virtual machines, containers, etc. In a Kubernetes™ environment, applications that require persistent storage are connected to KumoScale™ storage nodes via Container Storage Interface (CSI). In a bare metal environment, storage allocation to the nodes can be automated with automation tools, such as Ansible Playbook.

For access management and security of the NVMe-oF™ shared block storage, the storage is visible from all connected compute nodes, but access can be controlled. Additionally, integrated log management with Syslog and telemetry functions for data collection and analysis are provided, enabling users to check the collected information via KumoScale™ software.

NVMe-oF™ deployment with KumoScale™ software, SSD and other components

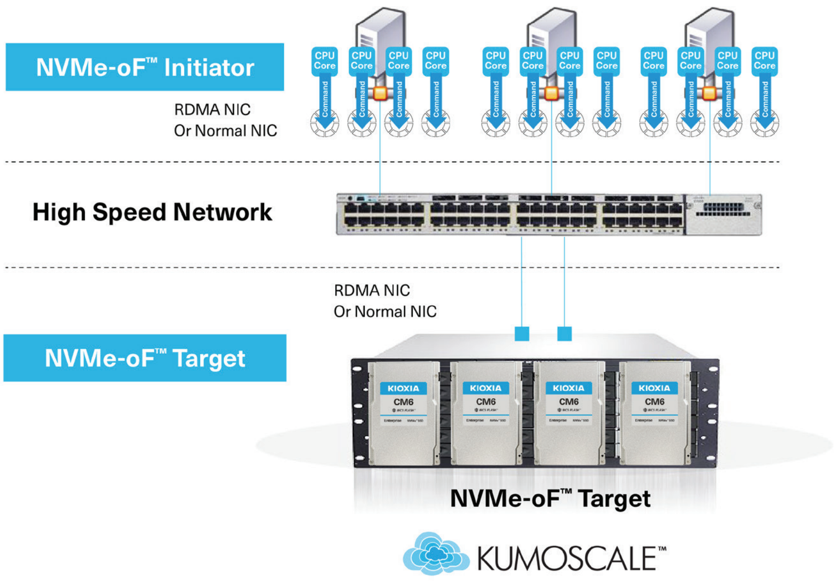

Fig. 2.3: NVMe-oF™ Topology with KumoScale™ Software

Platform requirements for KumoScale™ software.

Table 2.3: Platform requirements

Component	Minimum Configuration
Memory	64GB DDR4
System disk	2 x 128GB SATA DOM
Network interface card	One of the following: MCX516A-CCAT or MCX545A-CCAN ConnectX - 5 EN network interface card, 100 GbE dual/single port QSFP28, PCIe® 3.0 x16 ROHS R6 MCX416A-CCAT ConnectX - 4 EN network interface card, 100 GbE dual/single port QSFP28, PCIe® 3.0 x16 ROHS R6 Solarflare Communications XtremeScale SFC9250 10/25/40/50/100G Ethernet Controller
Power unit	Dual power supply (hot swap)
Management interface	In KumoScale™ software, data ports can be used for management traffic (may be dedicated ports).

Reference: KIOXIA Products / R&D KumoScale Software

3. Test Overview

3.1 Purpose

The purpose of the testing is to provide a guideline for disk selection/sizing in a real-world network storage configuration. Testing compared the performance metrics of local (DAS) KIOXIA disks with the same SSDs in a KumoScale™ disaggregated architecture. It presents disk preference and sizing data under a real-world network storage configuration by quantifying SSD performance with evaluation targets using local NVMe™ SSDs and NVMe-oF™ with KumoScale™ software.

3.2 Overview

In test case 1 (local/DAS), KIOXIA SSDs are mounted as local disks, and 18 patterns of throughput are measured for each SSD, in 4/16/64KiB block sizes, according to the Read/Write ratio.

In test case 2 (NVMe-oF™ configuration), the same KIOXIA SSDs were tested using a networked disaggregated storage configuration with KumoScale™ software via a switch. Bandwidth, IOPS, latency, and CPU utilization was measured for 4/16/64KiB block sizes, according to the Read/Write ratio for each SSD as in test case 1.

3.3 Test Pattern List

The table below shows the measurement patterns.

Table 3.3: Test Pattern

Job No.	SSD Quantity (unit)	Read/Write Ratio (%)	Block size (KiB)
1	1	Read: 100	4
2			16
3			64
4		Read: 75/Write: 25	4
5			16
6			64

Table 3.3: Test Pattern *(Continued)*

Job No.	SSD Quantity (unit)	Read/Write Ratio (%)	Block size (KiB)
7	2	Read: 100	4
8			16
9			64
10		Read: 75/Write: 25	4
11			16
12			64
13	4	Read: 100	4
14			16
15			64
16		Read: 75/Write: 25	4
17			16
18			64

4. Test Case 1

4.1 Test configuration

Configuration for test case 1.

Table 4.1: Test component

Item	Product	Quantity	Remarks
Physical server	Supermicro AS-1114S-WN10RT	1	-
CPU	AMD EPYC™ 7702 64-Core Processor	1	64 core
Memory	-	-	512 GB
Network interface card	Mellanox MT28908 Family [ConnectX-6]	※b	※b Only one port of 100 GbE NIC is used
Local SSD	KIOXIA KCM61VUL3T20	10	32TB (3.2TB*10)
Storage management software	KumoScale™ 3.15	1	Installed, but not used in test case 1.
Measurement tool	fio	-	v3.23

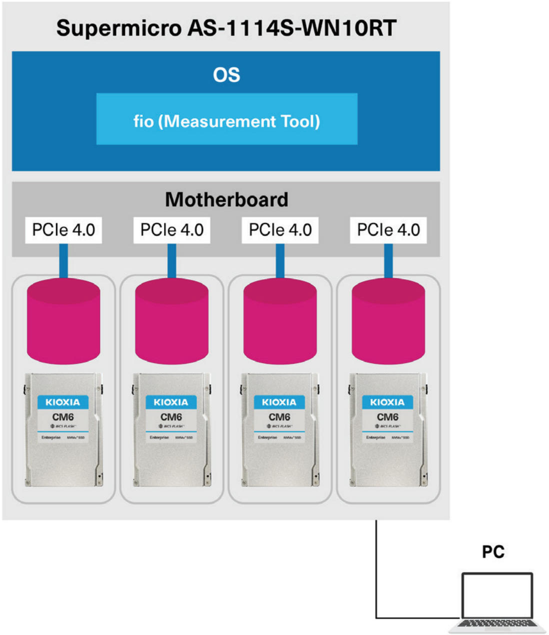

Fig. 4.1: Test case 1 configuration image

4.2. Test Methodology

Creating a volume per SSD.

Run fio on the physical server to measure bandwidth, IOPS, and latency, then compare the results with each pattern. Refer to “3.3 Test Pattern List” for the measurement patterns.

4.3. Result, Comparison

See *“7.1 Test Case 1: Measurement Result Graph”* for the measurement results.

When measured with 4 KiB block size and Read100%, a single SSD yielded approx. 6 GB/s throughput. *(Fig 7.1.1)*

Two SSDs achieved approximately 12 GB/s *(Fig. 7.1.4)*, and four SSDs approximately 24 GB/s *(Fig. 7.1.7)*, confirming that maximum values can be measured for SSD standalone specifications.

Also, when measured with 16KiB and 64KiB block size, maximum specified performance was measured.

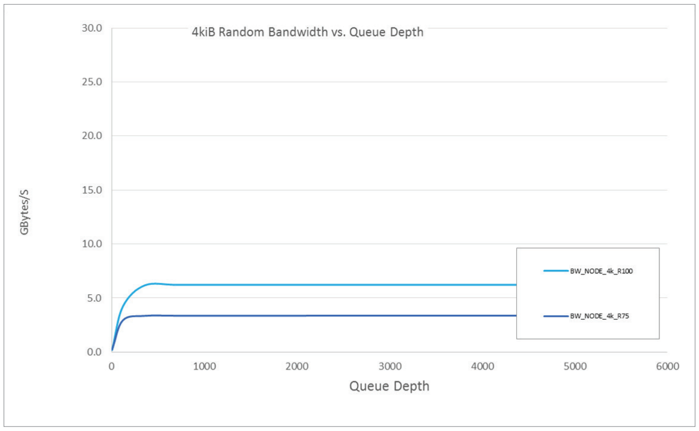

Fig. 7.1.1: Bandwidth (excerpt from p.11)

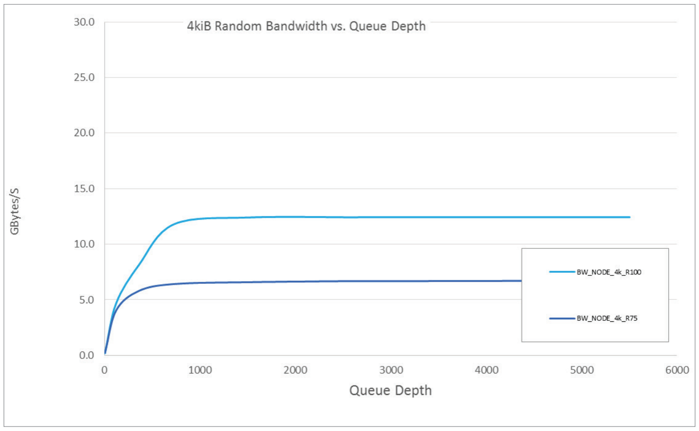

Fig. 7.1.4: Bandwidth (excerpt from p.11)

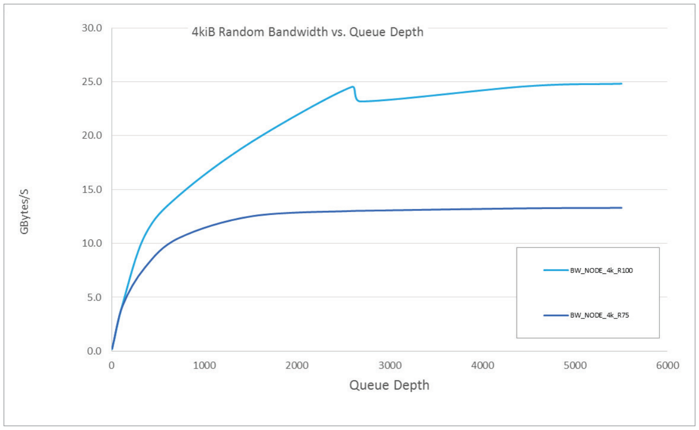

Fig. 7.1.7: Bandwidth (excerpt from p.12)

4.4. Key Takeaways

Local SSDs (DAS) tests demonstrated maximum performance of all of the SSDs, scaling performance linearly with no degradation.
When testing two SSDs, the 100Gbps/1-port network limit was reached.

5. Test Case 2

5.1. Test Configuration

Configuration for test case 2.

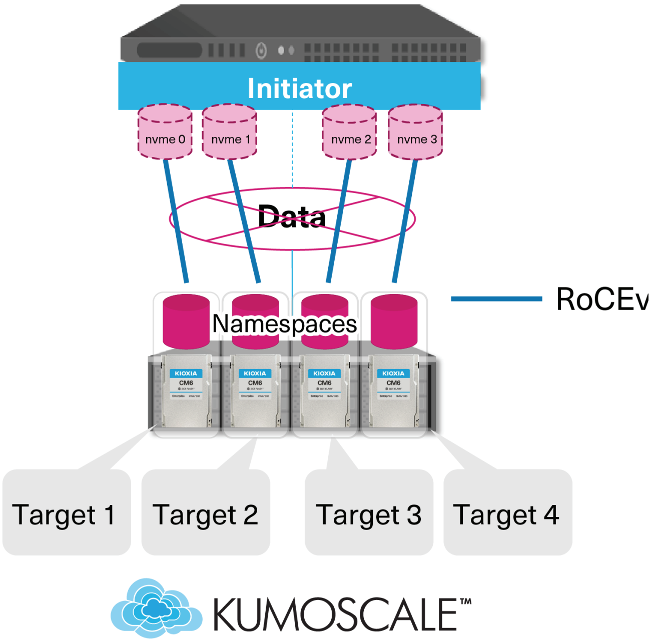

Fig. 5.1.1: Test case 2 configuration

5.1.2. Initiator Configuration (hardware and software)

Initiator configuration.

Table 5.1.2: Initiator

Item	Product	Quantity	Remarks
Physical server	HPE ProLiant DL380 Gen10	1	
CPU	Intel® Xeon® Gold 6148 CPU	2	2 CPU x 20 core x 2 thread
Memory	2.40GHz	-	768 GB
Network interface card	-	※b	※b Only one port of 100 GbENIC is used
OS	Mellanox MT27800 Family [ConnectX™-5] CentOS® 7 ※Kernel: 4.18.0-147.el8.x86_64+nvme host-patched-4 (applied patch by KIOXIA)	-	-
Measurement tool	fio	-	v3.23

5.1.3. Target Configuration (hardware and software)

Target configuration.

Table 5.1.3: Target

Item	Product	Quantity	Remarks
Physical server	Supermicro AS-1114S-WN10RT	1	-
CPU	AMD EPYC™ 7702 64-Core Processor	1	64 core
Memory	-	512 GB	-
Network interface	Mellanox MT28908 Family [ConnectX-6]	※b	※b Only one port of 100 GbE NIC is used
Local SSD	KIOXIA KCM61VUL3T20	10	32TB (3.2TB*10)
Storage management software	KumoScale™ 3.15	-	-v3.15

5.1.4. Network Configuration

Network configuration.

Table 5.1.4: Network

Item	Product	Quantity	Remarks
Switch	Mellanox SN2700	1	Onyx 3.7.1200

5.2. Test Methodology

Perform benchmark tests according to the patterns listed in *"3.3 Test Pattern List"*. Execute fio on the initiator, measure the bandwidth, IOPS, and latency, then compare the results with each pattern.

Additionally, CPU utilization of each initiator and target are measured for each block size when four SSDs are mounted. The following table shows the workloads used in the measurements.

Test case 1 using two SSDs, resulted in reaching the upper limit of the 100Gbps/1 port network. The performance evaluation is limited to one, two, or four SSDs.

Table 5.2. CPU utilization measurement patterns

Job No.	Block size (KiB)	Read/Write Ratio (%)	Workload (Block Size x Queue Depth x # of Jobs)
19	4	Read: 100	4
20			3480
21			6400
22		Read: 75/Write: 25	4
23			3480
24			6400
25	64	Read: 100	4
26			960
27			2208
28		Read: 75/Write: 25	4
29			960
30			2208

5.3. Results, Comparison

See *"7.2 Test Case 2 Measurement Result Graph"* and *"7.3 Test Case 2 CPU Utilization Measurement Graph"* for measurement results.

5.3.1. Transfer Speed

When running a Read 100% job with 16/64KiB block sizes using two SSDs, transfer speed is confirmed to reach out to the maximum of single 100GbE port bandwidth *(Fig. 7.2.5-1)*.

When running a Read100% job with 16/64KiB block sizes using four SSDs, transfer speed is confirmed to reach the maximum of single 100GbE port bandwidth *(Fig. 7.2.8-1)*.

When running a Read 75% /Write 75% job with 16KiB/64KiB block sizes using four SSDs, the total bandwidth for each Read/Write using full-duplex transfer speed can exceed the 100GbE/s network limit *(Fig. 7.2.8-1) (Fig. 7.2.9-1)*.

When measuring Read 100% and 4KiB block size, the bandwidth limit was not reached in all job patterns for SSD one, two, and four units *(Fig.7.2.1-1) (Fig.7.2.4-1) (Fig.7.2.7-1)*.

In addition to the above results, tests were performed using six and eight SSDs. Since the 100GbE/1 port network became a bottleneck with a configuration of two SSDs, very little performance improvement was measured with additional SSDs in this test configuration. Therefore, results for patterns with more than four SSDs, were omitted.

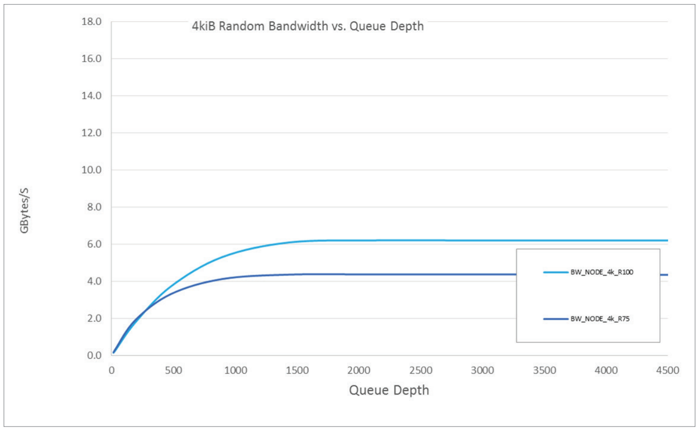

Fig. 7.2.1-1: Bandwidth (excerpt from p.13)

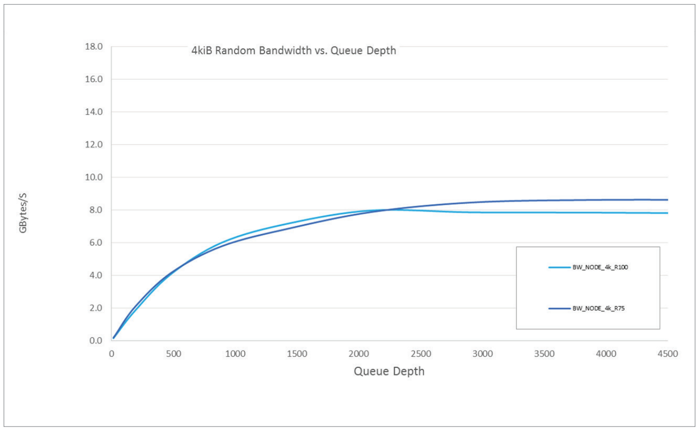

Fig. 7.2.4-1: Bandwidth (excerpt from p.14)

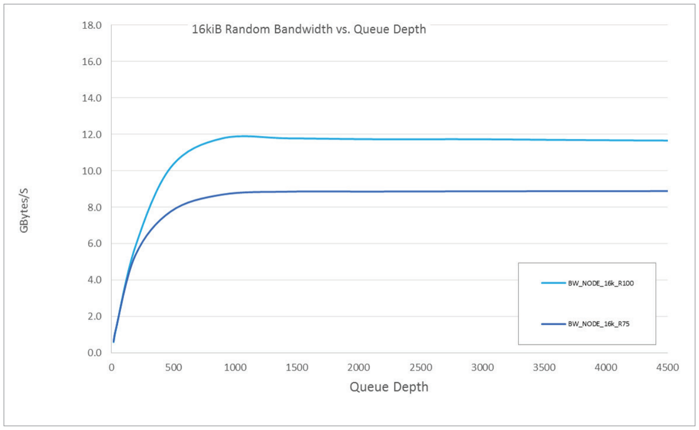

Fig. 7.2.5-1: Bandwidth (excerpt from p.15)

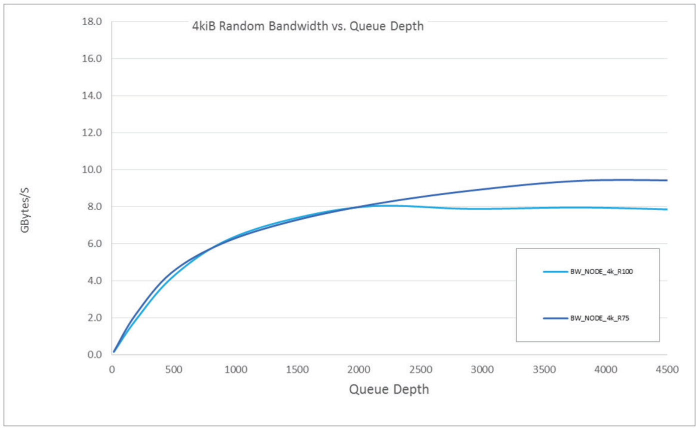

Fig. 7.2.7-1: Bandwidth (excerpt from p.17)

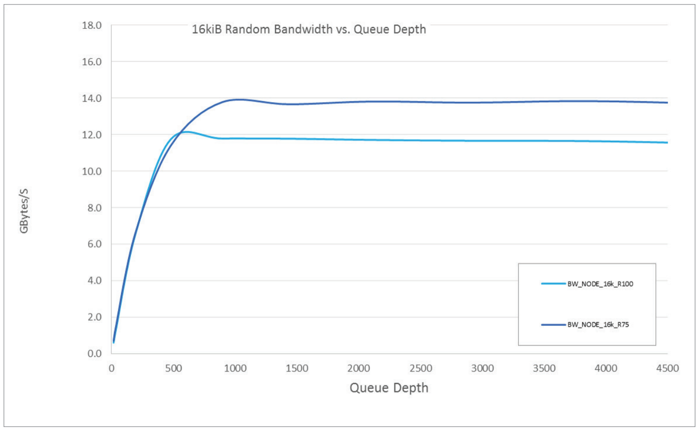

Fig. 7.2.8-1: Bandwidth (excerpt from p.17)

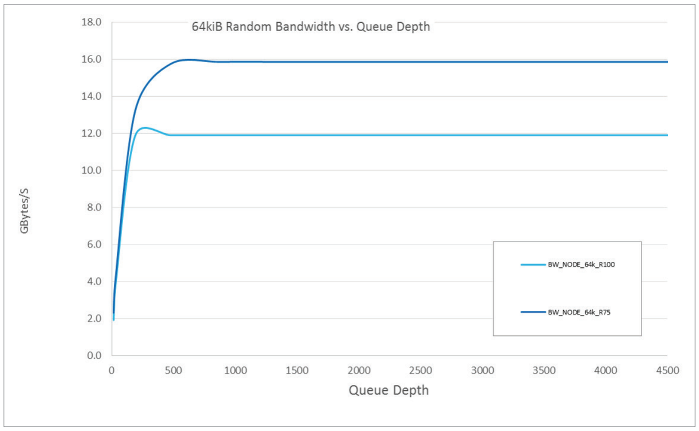

Fig. 7.2.9-1: Bandwidth (excerpt from p.18)

5.3.2. CPU Utilization

CPU utilization increases with higher load for initiator. Testing confirmed that the storage target running KumoScale™ software has CPU utilization margin even with a higher load.

Additionally, testing confirmed that CPU utilization will not reach 100% regardless of block size or Read/Write ratio.

No difference was observed in the Workload 3480/6400 comparison or the Workload 960/2208 comparison *(Fig. 7.3.3-2) (Fig. 7.3.4-2)*.

The initiator showed 100% CPU utilization for the Workload 6400 or 2208, regardless of block size or Read/Write ratio *(Fig. 7.3.3-1)*.

Although measurements are by block size, not much difference was observed comparing Read 100% and Read 75%/Write 25% measurements.

The following graph shows the CPU idle value of the initiator (server); the X-axis shows the time and the Y-axis shows the CPU idle percent, with a 100% idle indicating 0% CPU utilization and a 0% idle indicating 100% CPU utilization.

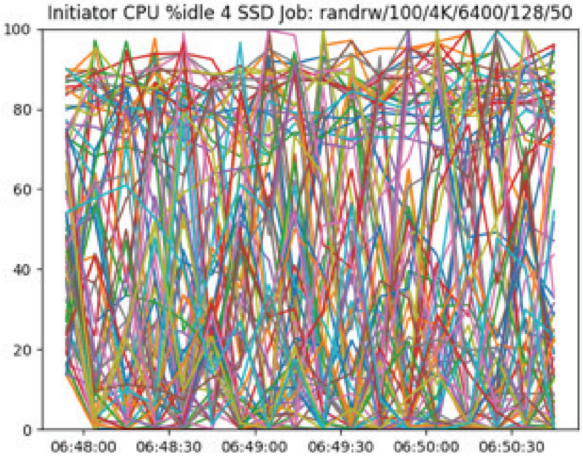

Fig. 7.3.3-1: Initiator (excerpt from p.19)

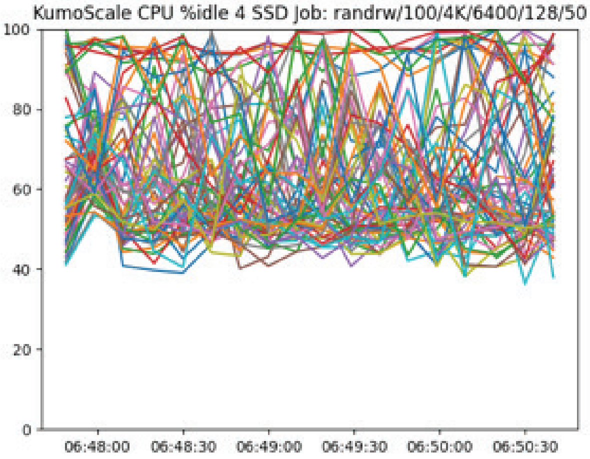

Fig. 7.3.3-2: Target (excerpt from p.20)

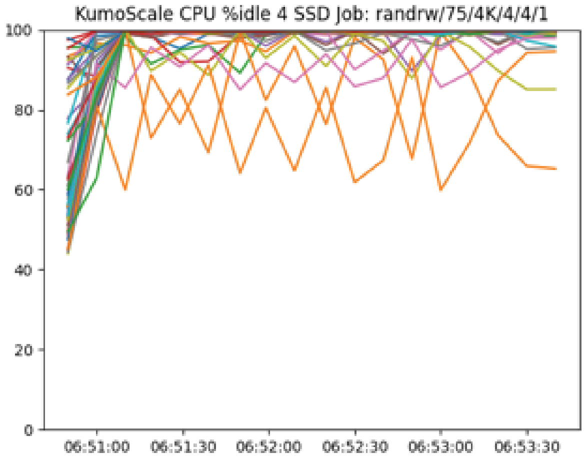

Fig. 7.3.4-2 :Target (excerpt from p.20)

5.4. Key Takeaways

5.4.1. SSD Performance

The networked storage configuration with KumoScale™ software achieves nearly identical performance as the locally connected SSDs, as does the Read/Write performance.

5.4.2. Performance Limiting Factor

The performance limiting factor is not the SSDs, but rather the network. At 100 Gbps, two SSDs can saturate network bandwidth. Adding a 100G RDMA NIC (RNIC) to the evaluation target and making it dual-ported could maximize the performance up to eight SSDs.

5.4.3. CPU (Target Server)

The target can accommodate traffic from additional initiators, since its CPU has enough headroom, and the protocol offloading in the NIC seems to be effective.

5.4.4. CPU (Initiator Server)

The initiator cannot drive enough traffic to saturate the storage target because its CPU does not have enough headroom, and another initiator server is required to make full use of the target server in this configuration.

5.4.5. Remarks

The production environment will use more than a dozen SSDs. The following table shows the network bandwidth required depending on the SSD quantity for a single target server.

Table 5.4.5: SSD quantity and required network bandwidth

SSD Quantity	Network Bandwidth
1	100 Gbps or greater
2	100 Gbps or greater
4	200 Gbps or greater
6	300 Gbps or greater
8	400 Gbps or greater
10	500 Gbps or greater

6. Conclusion

6.1. Conclusion

An NVMe-oF™ disaggregated server and storage configuration utilizing KumoScale™ software and KIOXIA SSDs provide fast and efficient use of shared flash storage. This data is beneficial for disk selection and sizing in network storage to get the most out of KIOXIA SSDs.

7. Appendix

7.1. Test Case 1 Test Results

The following graphs show comparisons between the Read100% and the Read75%/Write25% measurements.

7.1.1. SSD x 1, block size 4 KiB

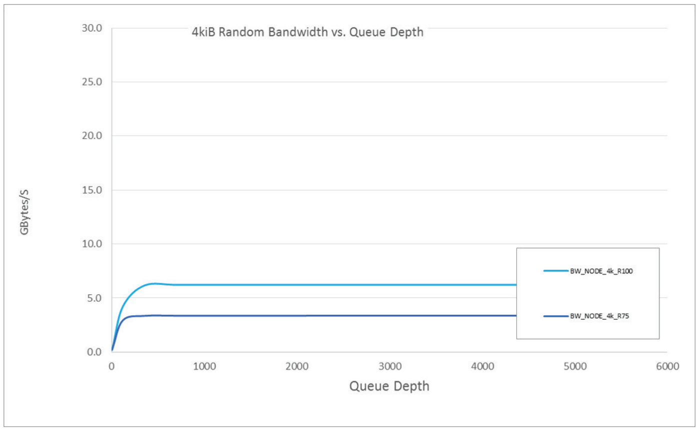

Fig. 7.1.1: Bandwidth

7.1.2. SSD x 1, block size 16 KiB

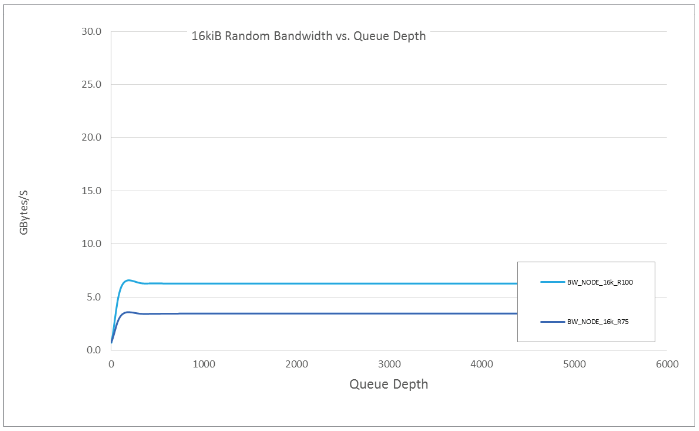

Fig. 7.1.1: Bandwidth

7.1.3. SSD x 1, block size 64 KiB

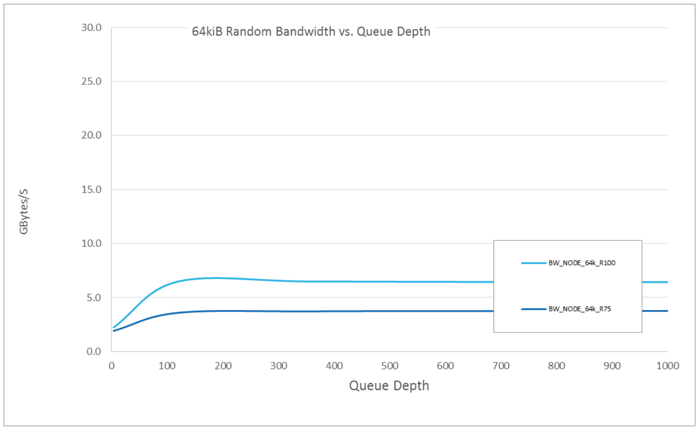

Fig. 7.1.3: Bandwidth

7.1.4. SSD x 2, block size 4 KiB

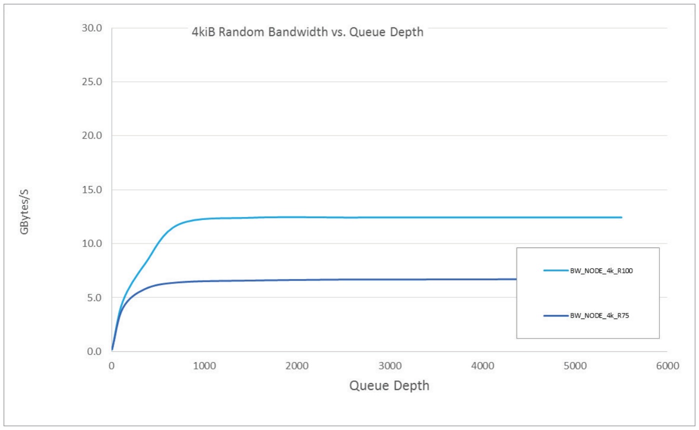

Fig. 7.1.4: Bandwidth

7.1.5. SSD x 2, block size 16 KiB

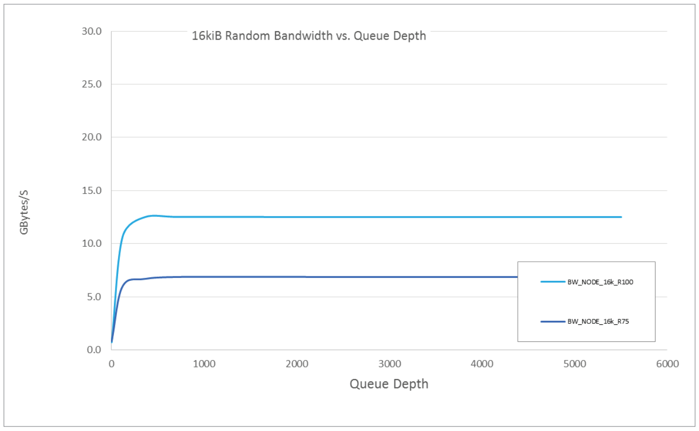

Fig. 7.1.5: Bandwidth

7.1.6. SSD x 2, block size 64 KiB

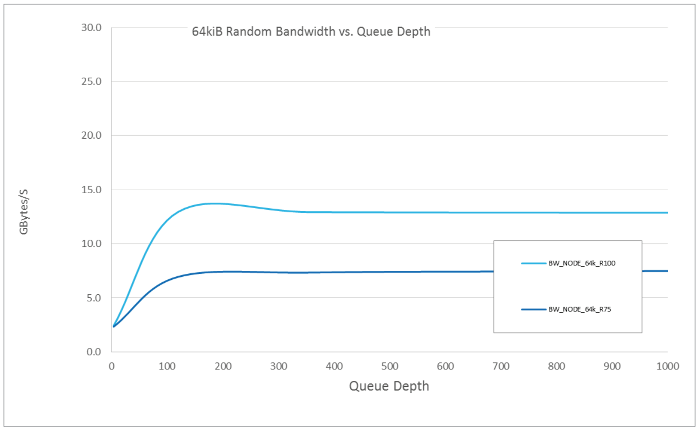

Fig. 7.1.6: Bandwidth

7.1.7. SSD x 4, block size 4 KiB

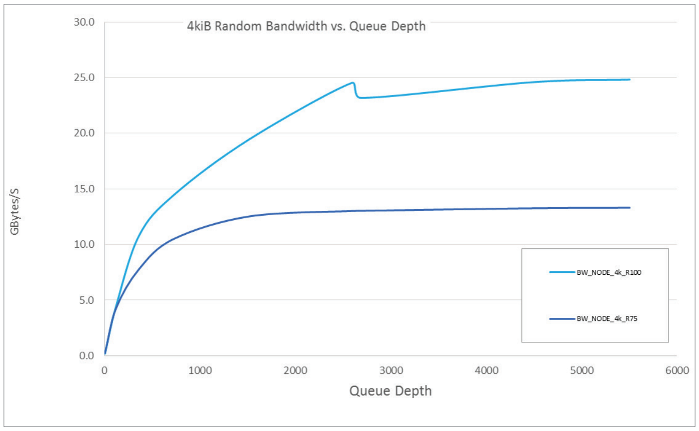

Fig. 7.1.7: Bandwidth

7.1.8. SSD x 4, block size 16 KiB

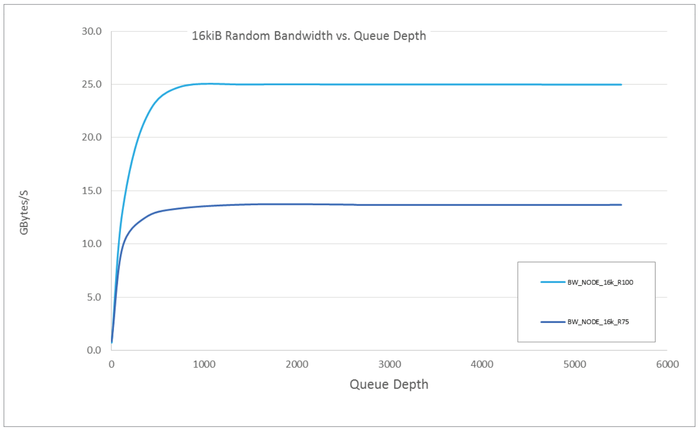

Fig. 7.1.8: Bandwidth

7.1.9. SSD x 4, block size 64 KiB

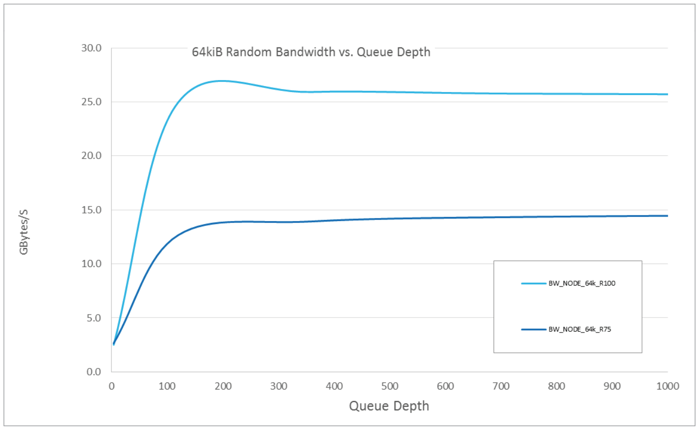

Fig. 7.1.9: Bandwidth

7.2. Test Case 2 Measurements

The following graphs show comparisons between Read100% and the Read75%/Write25% measurements.

7.2.1. SSD x 1, block size 4 KiB

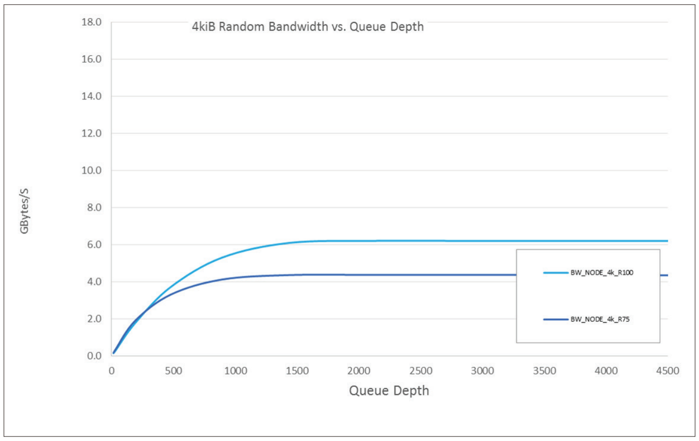

Fig. 7.2.1-1: Bandwidth

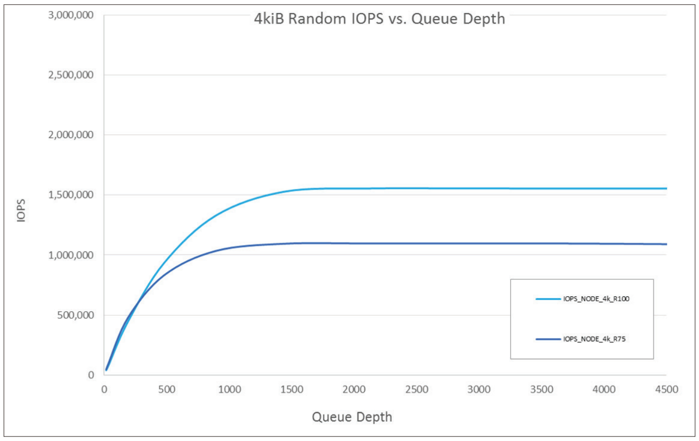

Fig. 7.2.1-2: IOPS

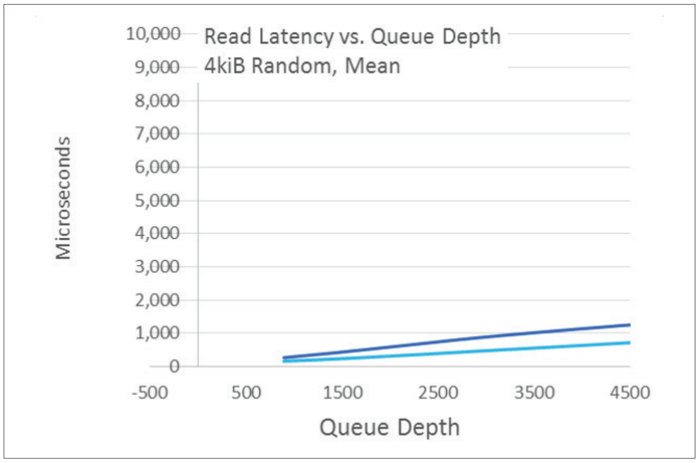

Fig. 7.2.1-3: Latency

7.2.3. SSD x 1, block size 64 KiB

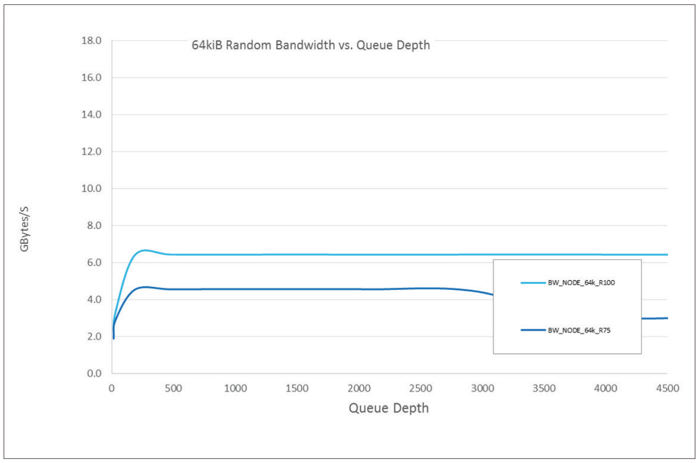

Fig. 7.2.3-1: Bandwidth

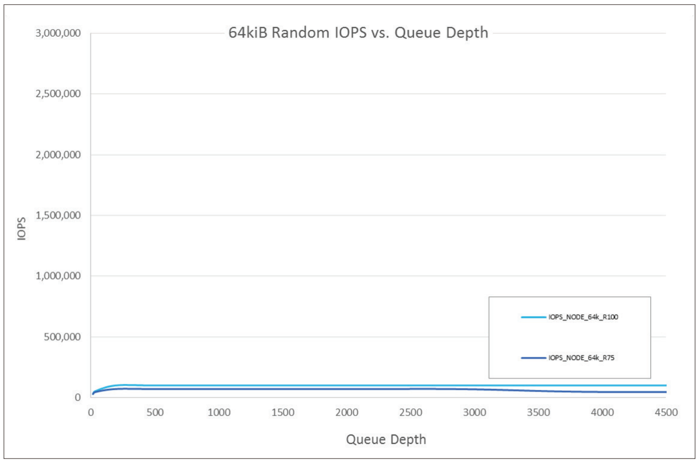

Fig. 7.2.3-2: IOPS

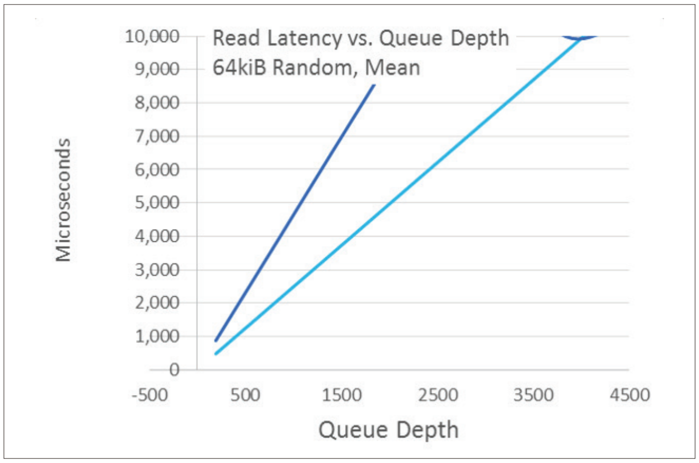

Fig. 7.2.3-3: Latency

7.2.4. SSD x 2, block size 4 KiB

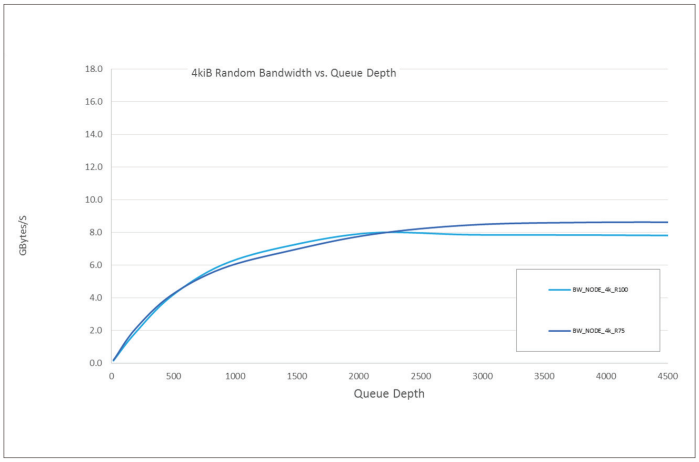

Fig. 7.2.4-1: Bandwidth

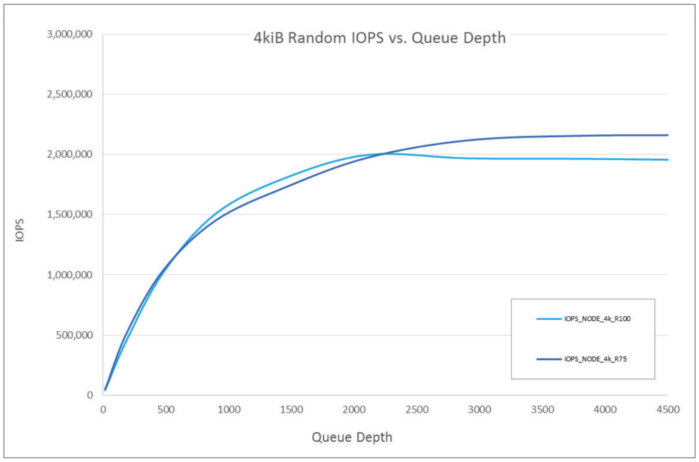

Fig. 7.2.4-2: IOPS

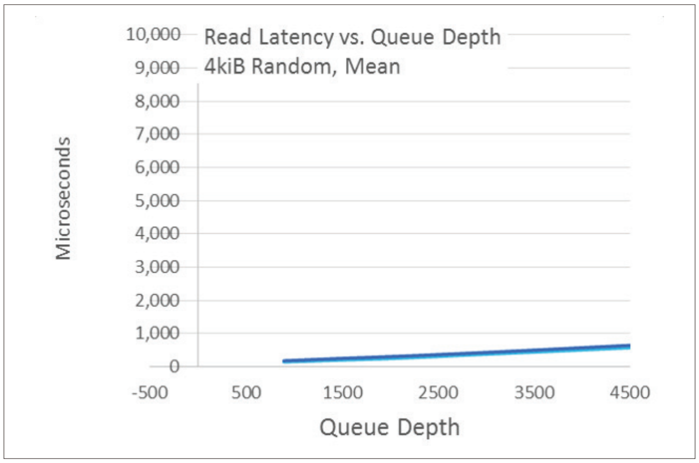

Fig. 7.2.4-3: Latency

7.2.5. SSD x 2, block size 16 KiB

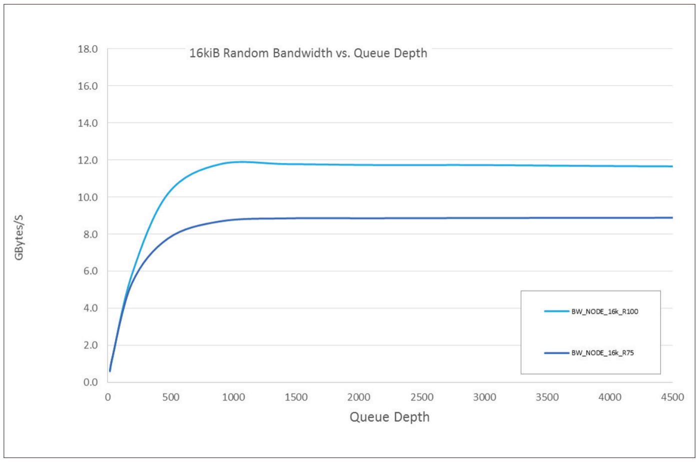

Fig. 7.2.5-1: Bandwidth

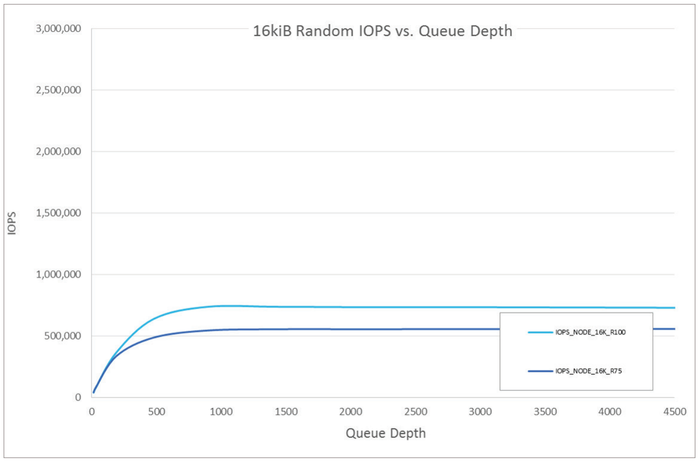

Fig. 7.2.5-2: IOPS

7.2.5. SSD x 2, block size 16 KiB *(Continued)*

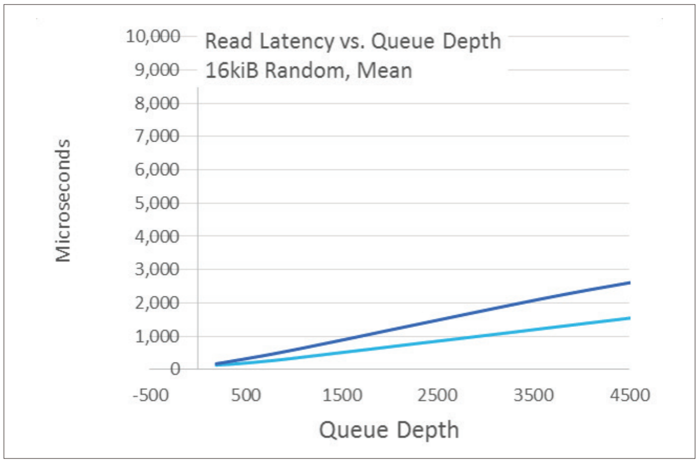

Fig. 7.2.5-3: Latency

7.2.6. SSD x 2, block size 64 KiB

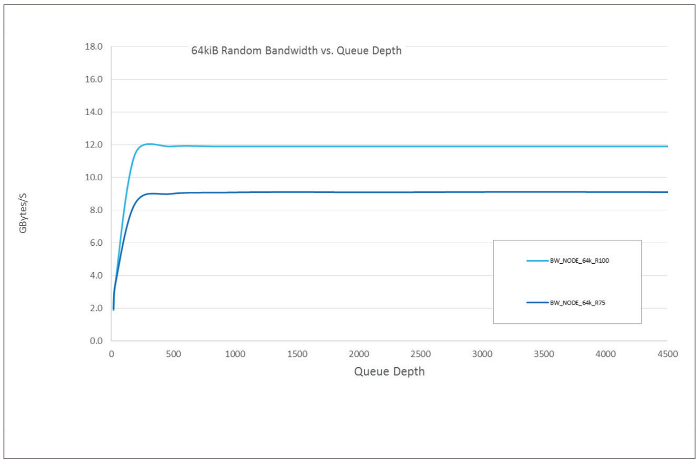

Fig. 7.2.6-1: Bandwidth

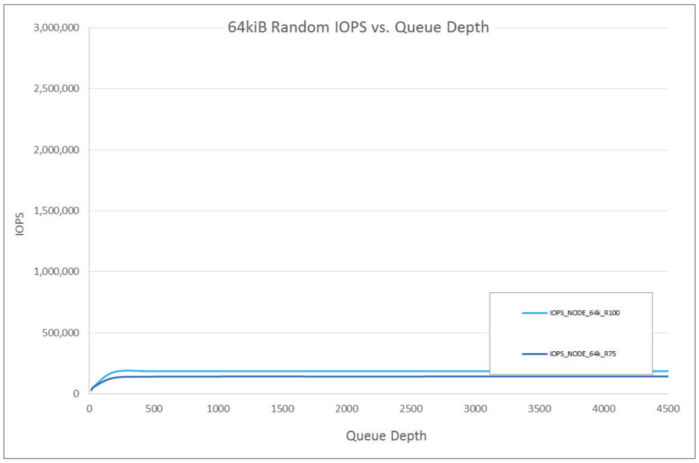

Fig. 7.2.6-2: IOPS

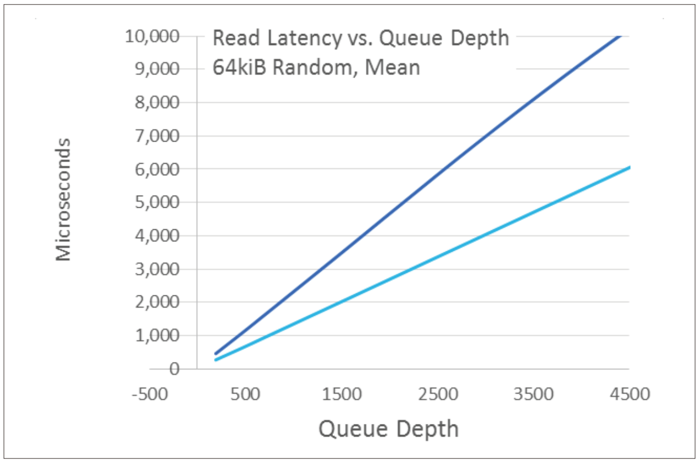

Fig. 7.2.6-3: Latency

7.2.7. SSD x 4, block size 4 KiB

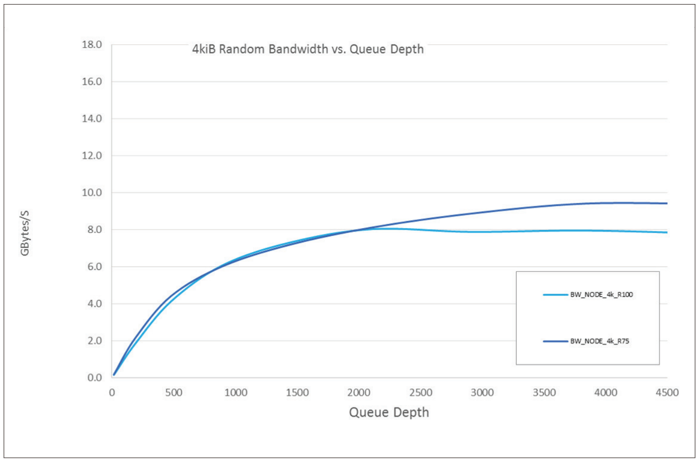

Fig. 7.2.7-1: Bandwidth

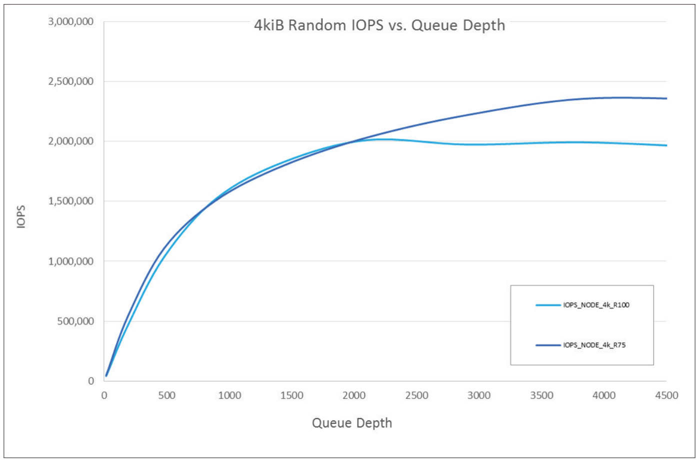

Fig. 7.2.7-2: IOPS

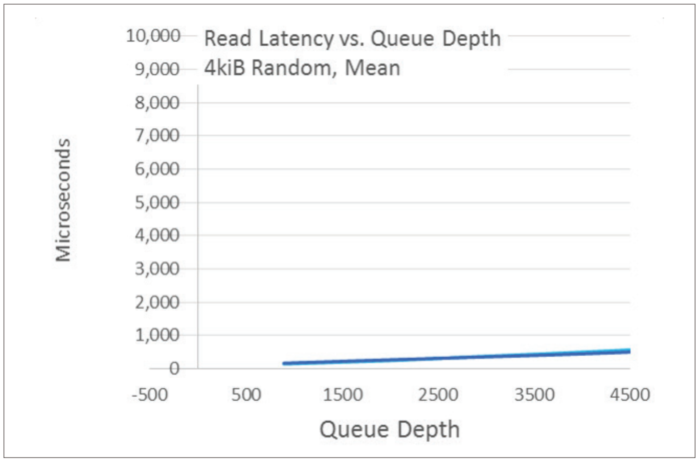

Fig. 7.2.7-3: Latency

7.2.8. SSD x 4, block size 16 KiB

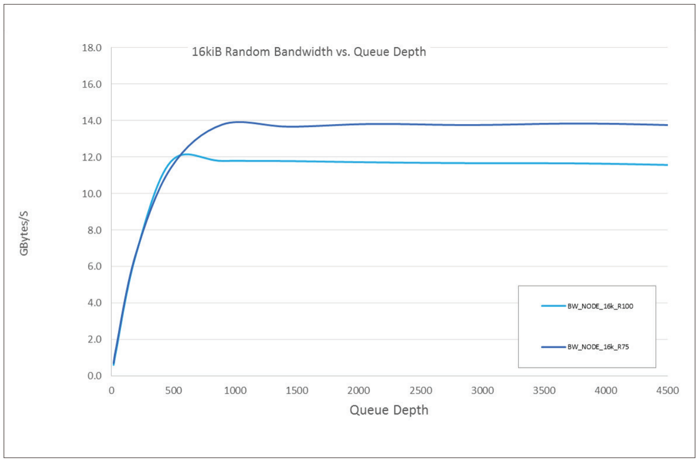

Fig. 7.2.8-1: Bandwidth

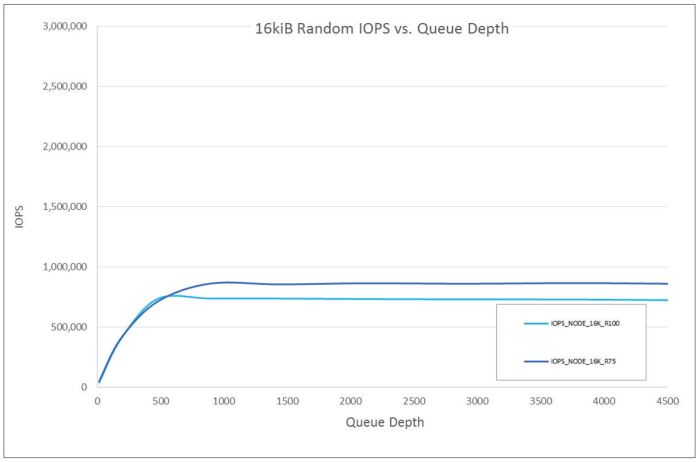

Fig. 7.2.8-2: IOPS

7.2.8. SSD x 4, block size 16 KiB *(Continued)*

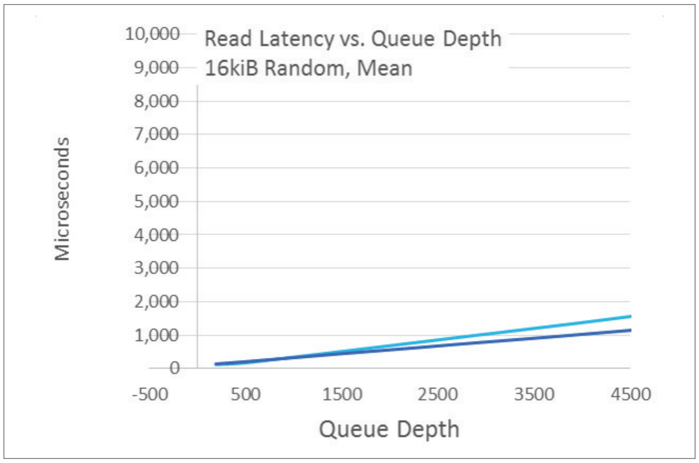

Fig. 7.2.8-3: Latency

7.2.9. SSD x 4, block size 64 KiB

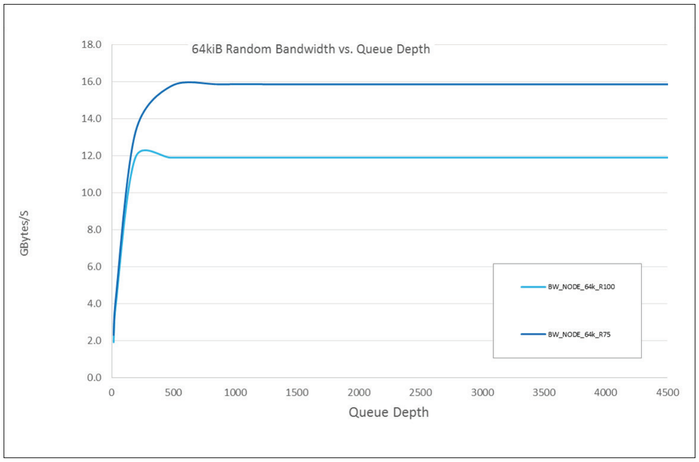

Fig. 7.2.9-1: Bandwidth

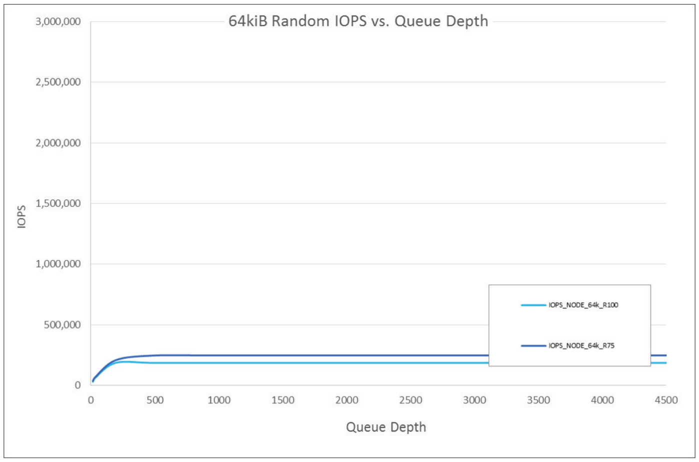

Fig. 7.2.9-2: IOPS

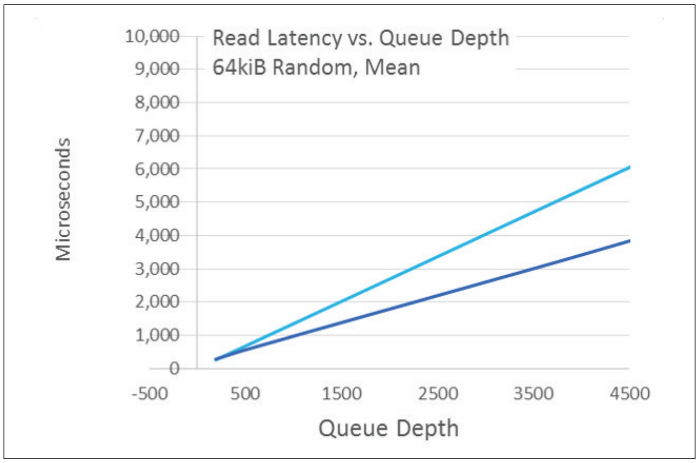

Fig. 7.2.9-3: Latency

7.3. Test Case 2 CPU Utilization Measurements

Measurements of CPU utilization with four SSDs.

7.3.1. Block size 4 kB, Read 100%, workload 4

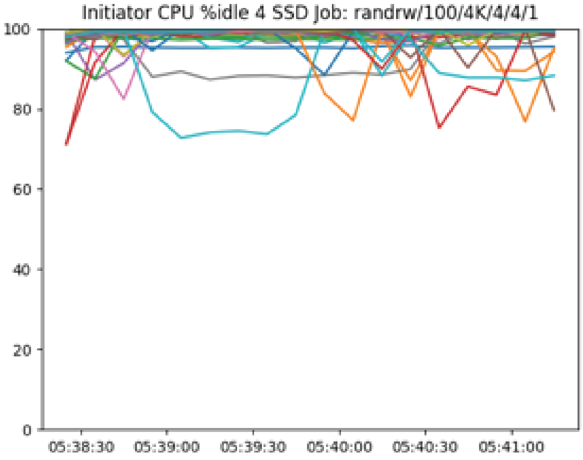

Fig. 7.3.1-1: Initiator

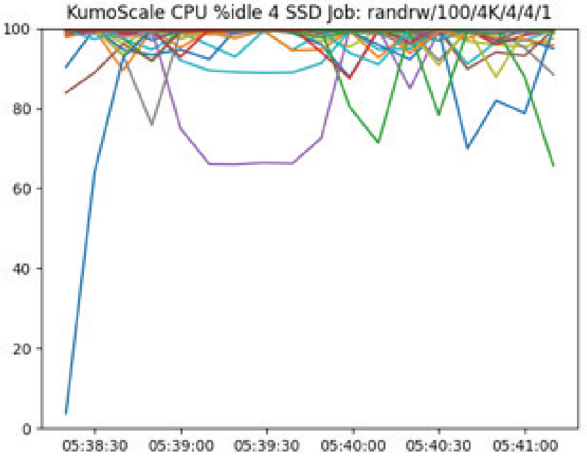

Fig. 7.3.1-2: Target

7.3.2. Block size 4 kB, Read 100%, workload 3480

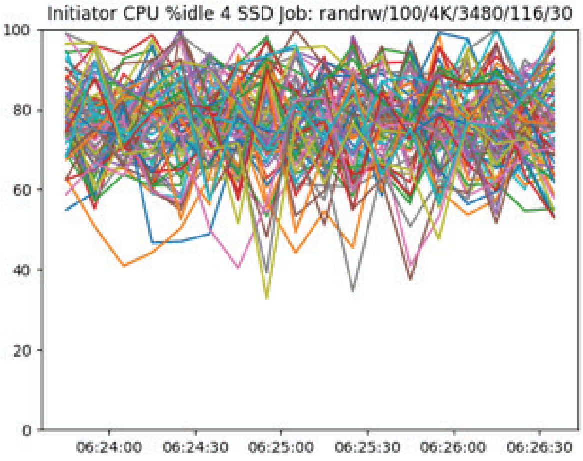

Fig. 7.3.2-1: Initiator

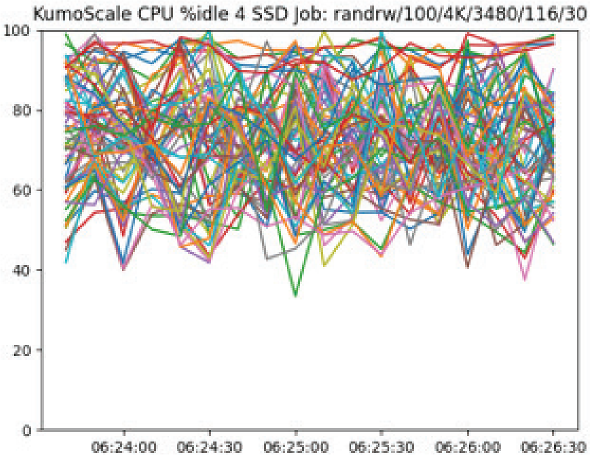

Fig. 7.3.2-2: Target

7.3.3. Block size 4 kB, Read 100%, workload 6400

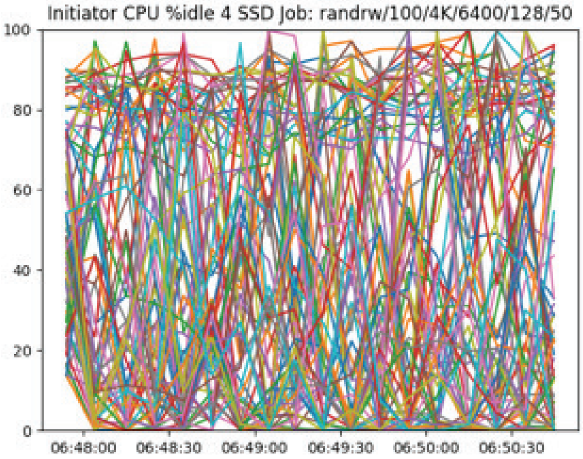

Fig. 7.3.3-1: Initiator

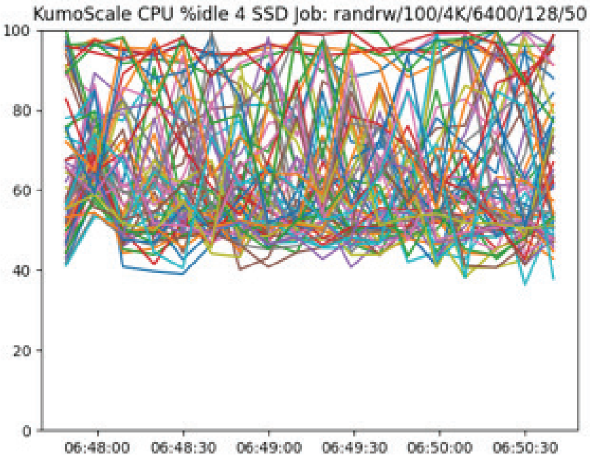

Fig. 7.3.3-2: Target

7.3.4. Block size 4 kB, Read 75%, workload 4

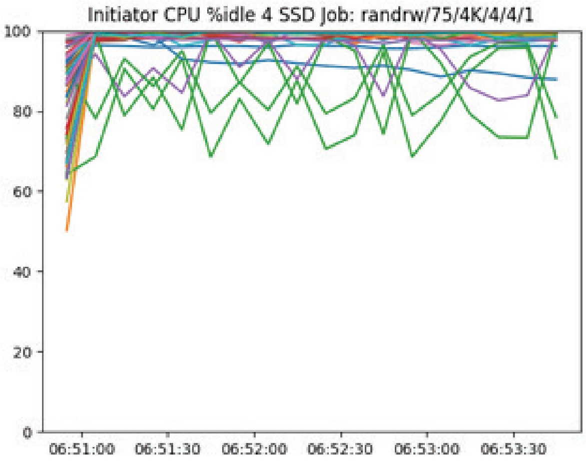

Fig. 7.3.4-1: Initiator

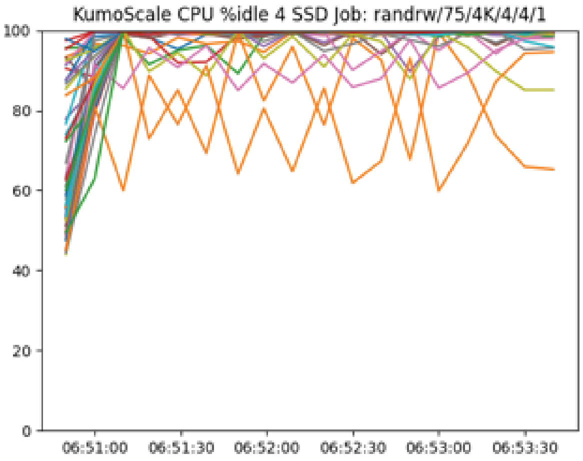

Fig. 7.3.4-2: Target

7.3.5. Block size 4 kB, Read 75%, workload 3480

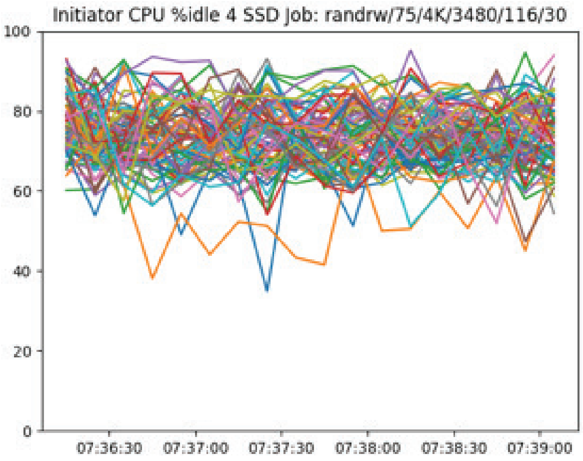

Fig. 7.3.5-1: Initiator

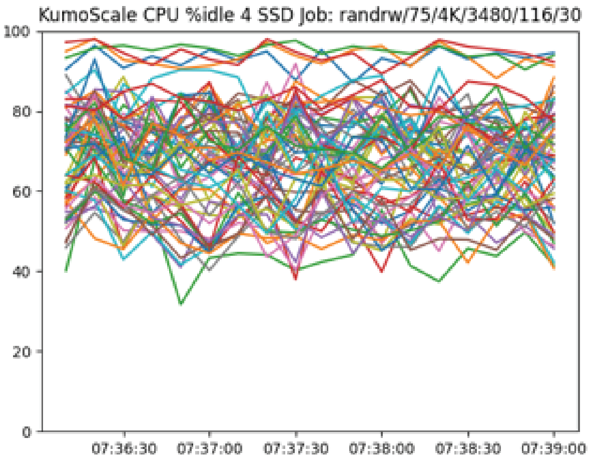

Fig. 7.3.5-2: Target

7.3.6. Block size 4 kB, Read 75%, workload 6400

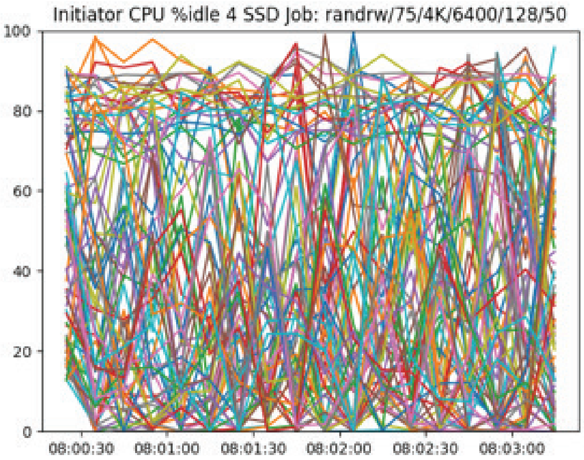

Fig. 7.3.6-1: Initiator

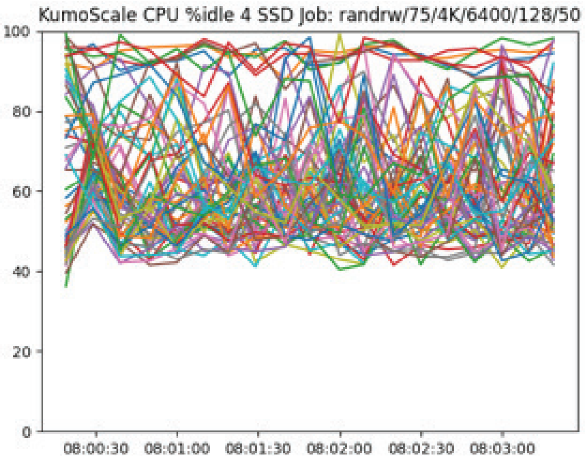

Fig. 7.3.6-2: Target

7.3.7. Block size 64 kB, Read 100%, workload 4

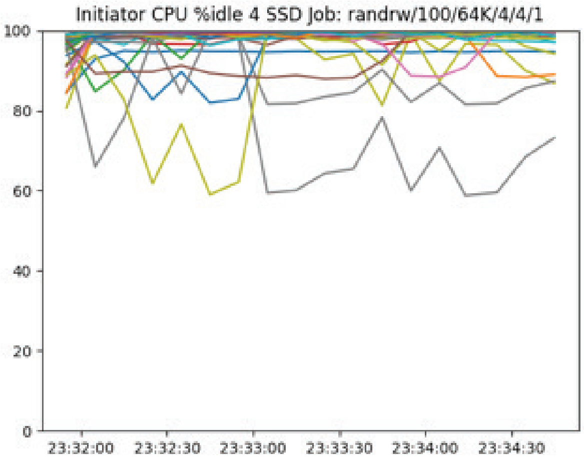

Fig. 7.3.7-1: Initiator

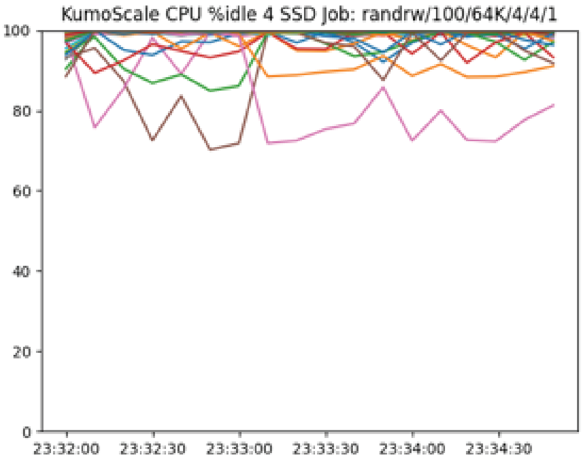

Fig. 7.3.7-2: Target

7.3.8. Block size 64 kB, Read 100%, workload 960

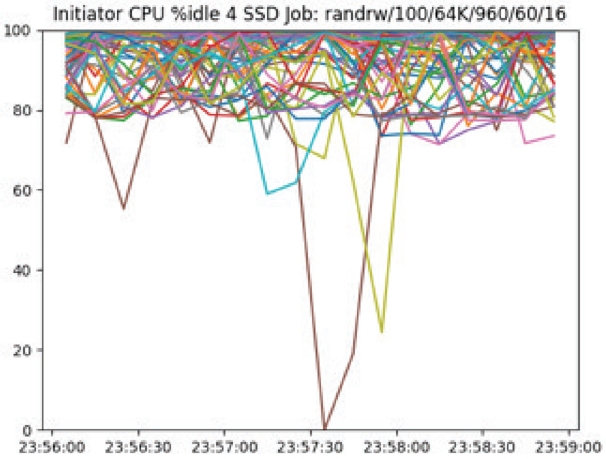

Fig. 7.3.8-1: Initiator

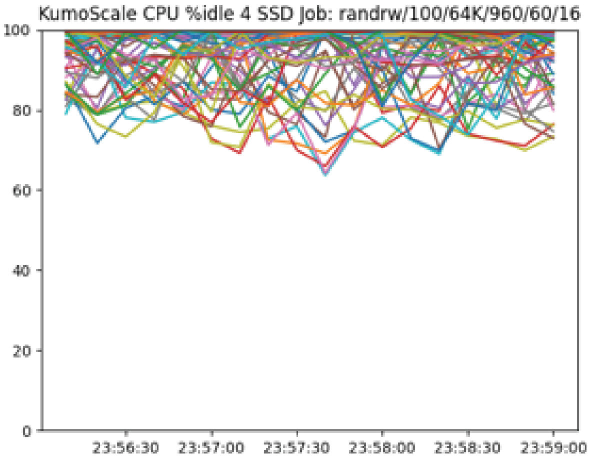

Fig. 7.3.8-2: Target

7.3.9. Block size 64 kB, Read 100%, workload 2208

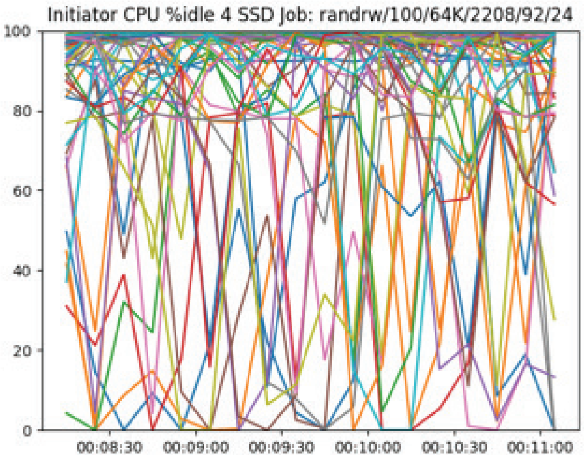

Fig. 7.3.9-1: Initiator

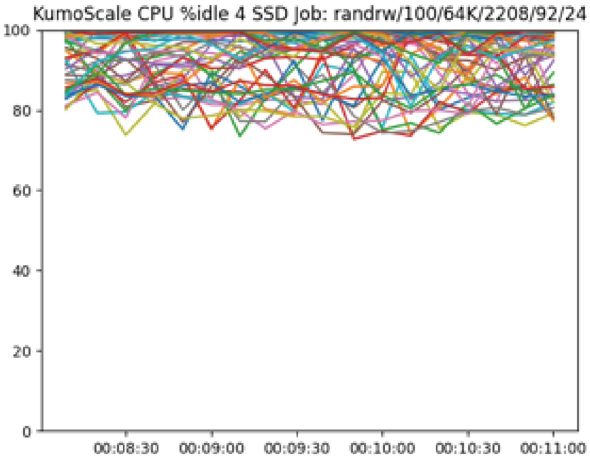

Fig. 7.3.9-2: Target

7.3.10. Block size 64 kB, Read 75%, workload 4

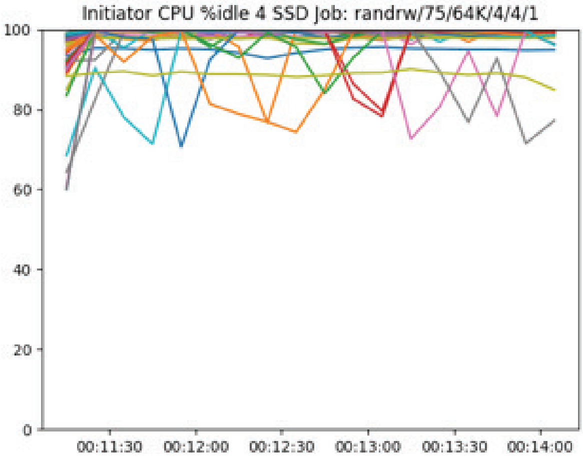

Fig. 7.3.10-1: Initiator

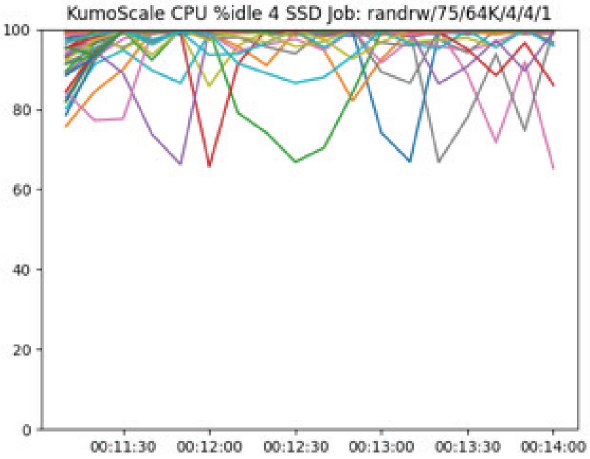

Fig. 7.3.10-2: Target

7.3.11. Block size 64 kB, Read 75%, workload 960

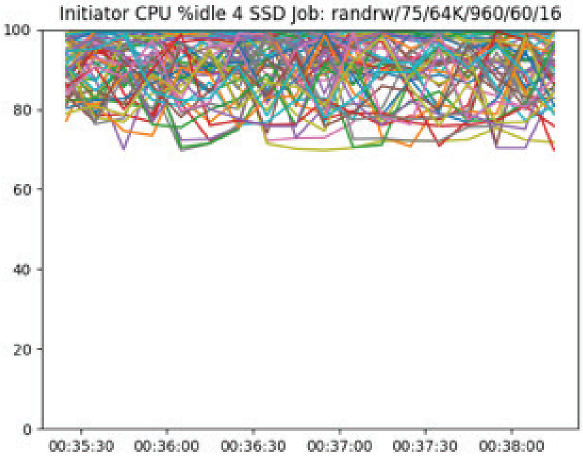

Fig. 7.3.11-1: Initiator

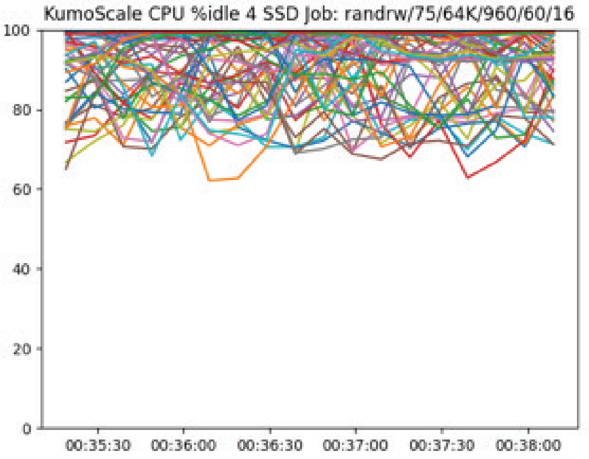

Fig. 7.3.11-2: Target

7.3.12. Block size 64 kB, Read 75%, workload 2208

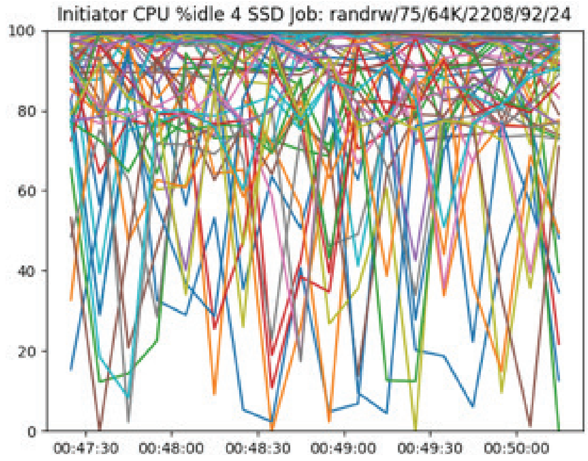

Fig. 7.3.12-1: Initiator

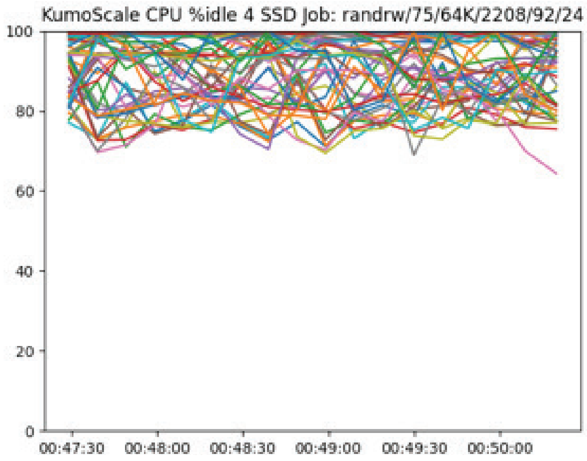

Fig. 7.3.12-2: Target

Notes:

Definition of capacity - KIOXIA Corporation defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes and a terabyte (TB) as 1,000,000,000,000 bytes. A computer operating system, however,reports storage capacity using powers of 2 for the definition of 1Gbit = 2^{30} bits = 1,073,741,824 bits, 1GB = 2^{30} bytes = 1,073,741,824 bytes and 1TB = 2^{40} bytes = 1,099,511,627,776 bytes and therefore shows less storage capacity. Available storage capacity (including examples of various media files) will vary based on file size, formatting, settings, software and operating system, and/or pre-installed software applications, or media content. Actual formatted capacity may vary.

Intel is a trademark of Intel Corporation or its subsidiaries. Linux is a trademark of Linus Torvalds in the U.S. and other countries. NVMe and NVMe-oF are trademarks of NVM Express, Inc. in the United States and other countries. PCIe is a registered trademark of PCI-SIG. Kubernetes is a trademark of The Linux Foundation in the United States and other countries. Ansible is a trademark of Red Hat, Inc. in the United States and other countries. All other trademarks or registered trademarks are the property of their respective owners.

© 2022 KIOXIA Corporation. All rights reserved. Information in this performance brief, including product specifications, tested content, and assessments are current and believed to be accurate as of the date that the document was published, but is subject to change without prior notice. Technical and application information contained here is subject to the most recent applicable KIOXIA product specifications.