SNIA®

Storage Networking Industry Association

Optimizing NVMe® over Fabrics (NVMe-oF™)

Comparing Performance of Different Transports with Host Factors using Synthetic & Real-World Workloads

White Paper
April 2021

AUTHORS:
Eden Kim, Calypso Systems, Inc.
Fred Zhang, Intel Corp.

ABSTRACT:
NVMe® over Fabrics (NVMe-oF™) performance is compared using RDMA (iWARP & RoCEv2) and TCP transports across 100Gb Ethernet. Synthetic and real-world workloads are applied across different Fabrics to a storage target using both standard (1500B) and jumbo (9000B) MTU frames. Performance is compared between a six-SSD 3D XPointTM LUN and a 6-SSD 3D NAND LUN. Results show the impact of different workloads (synthetic corner case and real-world workloads), RDMA & TCP transport mechanisms (CPU onload vs CPU offload) and different types of storage LUNs (3D XPoint vs 3D NAND).

SNIA CMSI | COMPUTE, MEMORY, AND STORAGE

SNIA NSF | NETWORKING STORAGE

Table of Contents

Background

White Paper Companion & Update to SNIA/Brighttalk Webcast

This white paper is a companion and update to the SNIA/Brighttalk Webcast "Optimizing NVMe-oF Performance with different Transports: Host Factors" broadcast on September 15, 2020. This webcast was moderated by Tom Friend, principal at Illuminosi, with an introduction by David Woolf, University of New Hampshire. Webcast presenters were Fred Zhang, Intel Corp. and Eden Kim, Calypso Systems, Inc.

In addition to the synthetic Random 4KB & Sequential 128KB Read/Write corner case workloads and the real-world GPS 100% Write Navigation Portal workload presented in the webcast, this white paper is updated with two additional real-world workloads: Retail Web 66% Read Portal and VDI 75% Write Storage Server Cluster.

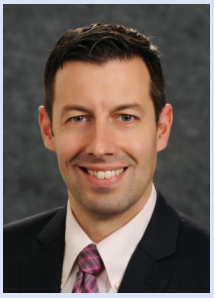
This white paper is a collaboration between the SNIA NSF (Networking Storage Forum), the SNIA SSS TWG (Solid State Storage Technical Working Group) and the SNIA CMSI (Compute, Memory & Storage Initiative).

Click on the following to view the webcast, download the presentation or download the Questions & Answers to the Webcast. Questions concerning this white paper can be sent to Fred.zhang@intel.com or edenkim@calypsotesters.com.

Tom Friend
Illuminosi

Fred Zhang
Intel Corp.

Eden Kim
Calypso Systems, Inc.



David Woolf
Univ. New Hampshire

SNIA Resources

SNIA Resources

The Storage Networking Industry Association (SNIA) is non-profit global organization dedicated to developing standards and education programs to advance storage and information technology. The mission of the Compute, Memory & Storage Initiative (CMSI) is to support the industry drive to combine processing with memory and storage and to create new compute architectures and software to analyze and exploit the explosion of data creation over the next decade. The mission of the Networking Storage Forum (NSF) is to drive the broad adoption and awareness of storage networking solutions.

SNIA, CMSI and the NSF can be found at snia.org, http://www.snia.org/forums/cmsi and at www.snia.org/forums/nsf/technology. Recent white papers can be found at the SNIA Educational Library while podcasts can be heard at snia.org/podcasts. SNIA related videos can also be seen at the SNIAVideo YouTube Channel.

The NVMe website can be found at www.nvmexpress.org, the NVMe Specification can be found at www.nvmexpress.org/developers/nvme-specification/, and the NVMe-oF Specification can be found at https://nvmexpress.org/developers/nvme-of-specification/.

SNIA Technical works including Performance Test Specifications (PTS), can be found at https://www.snia.org/tech_activities/work.

Additional information about SNIA, CMSI or NSF can be found at https://www.snia.org/resources or email can be sent to askcmsi@snia.org.

I. Abstract

NVMe over Fabrics (NVMe-oF) storage is becoming more popular as different Fabrics transports become available. This white paper examines various performance characteristics of different Ethernet NVMe-oF transports: RDMA (iWARP & RoCEv2) and TCP.

In this comparison, we apply both synthetic corner case and real-world workloads. Test IOs are applied across 100GbE NVMe-oF Fabric from the Host Initiator server to the Target Storage server. We test two types of storage that are directly attached to the Target Server – a six-drive 3D XPoint and a six-drive 3D NAND Logical Storage Unit (LUN). Test set-up and conditions are normalized to isolate the impact of Host factors on performance comparisons.

Performance analysis is conducted to compare three key metrics:

- Maximum Transmission Unit (MTU) frame size (standard v. jumbo frames),
- 3D XPoint six-SSD vs 3D NAND six-SSD LUNs
- Synthetic vs real-world workloads across iWARP, RoCEv2 & TCP transports

Synthetic workloads are comprised of random 4K Read/Write (RND 4K R/W) and sequential 128K Read/Write (SEQ 128K R/W) IOs. Real-world workloads are taken from a GPS Nav Portal (100% Write), a VDI Storage Cluster (75% Write) and a Retail Web Portal (65% Read).

Results show substantially similar performance for both standard 1500B (byte) frame and jumbo 9000B frame MTU. The 3D XPoint storage LUN shows substantially higher performance than the 3D NAND storage LUN. RDMA transports (iWARP and RoCEv2), with CPU offload, show significantly higher performance than TCP without CPU offload. RoCEv2 shows slightly lower IO Response Times than iWARP. These tests were all done with the specified hardware and using different network cards could potentially yield different results. Third party testing has shown that using a full NVMe/TCP hardware offload may result in better performance and latency for NVMe/TCP, approximately equivalent to that of RDMA.

iWARP and RoCEv2 show higher performance against 65% Read Retail Web and 100% Write GPS Nav portal workloads while TCP shows higher performance against 75% Write VDI Storage Cluster workloads. This difference may be due to the different IO Stream and block size content of the respective workloads (i.e., different workload IO Stream content will have different performance impact on various transports and/or storage LUNs tested in this study).

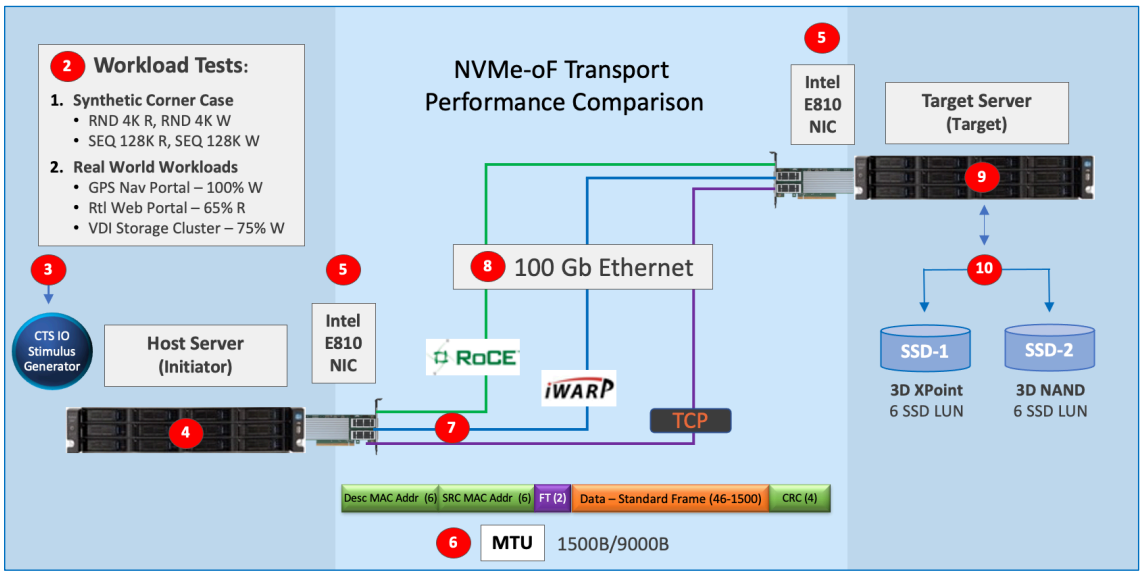


Figure 1 – Set-up & Test Plan

II. Introduction – NVMe over Fabrics (NVMe-oF)

A. NVMe-oF: What is it?

NVM Express (NVMe) is the standard host controller interface for PCIe based Solid State Drives (SSD). The NVMe over Fabrics (NVMe-oF) specification defines a protocol interface and related extensions that enable the NVMe command set to be transmitted over interconnects such as RDMA, Fibre Channel and others. NVMe-oF also extends NVMe deployment from a local to remote host for scale-out NVMe storage.

B. Know your NVMe-oF transports: What's the difference?

There are 3 Ethernet-based transports for NVMe over Fabrics: iWARP RDMA, RoCEv2 RDMA and TCP.

Remote Direct Memory Access (RDMA)

Remote Direct Memory Access (RDMA) is a host-offload, host-bypass technology that enables a low-latency, high-throughput direct memory-to-memory data communication between applications over a network (RFC 5040 A Remote Direct Memory Access Protocol Specification). RDMA usually takes advantage of network hardware offloads and reduces server resources typically dedicated to network functions. There are two main implementations of RDMA on Ethernet for NVMe-oF: iWARP and RoCEv2.

Best Effort vs. Lossless Networks

Best Effort networks are networks that do not guarantee data can be delivered, or delivered in order, or delivered without compromise of integrity. Internet Protocol (IP) network layer is an example of a Best Effort network. IP networks generally rely on an upper-level protocol (e.g., TCP, or Transmission Control Protocol) to provide additional mechanism to achieve a reliable data delivery. Such a mechanism could include, but is not limited to, flow control and congestion management.

Lossless networks, also called "no drop" networks, are so-called because they are designed to be reliable and ensure that no packets will be dropped. Best effort networks, on the other hand, are defined by their inability to guarantee delivery and, as a result, will require re-transmission in the event of packet loss.

Lossless networks can be built on top of Best Effort networks, such as TCP over IP (TCP/IP). UDP over IP (User Datagram Protocol over Internet Protocol or UDP/IP – see below), does not provide flow control and congestion management, nor does it provide *guaranteed* delivery, thus requires additional Ethernet network configuration or mechanisms to avoid dropping packets. These additional configurations include either Priority Flow Control (PFC) at layer 2 (https://www.ieee802.org/1/pages/dcbridges.html) or Differentiated Service Code Point (DSCP) PFC at layer 3 (RFC 2474 Definition of the differentiated services field in the IPv4 and IPv6 headers, RFC 2475 An architecture for differentiated services).

iWARP

iWARP is a computer networking protocol that implements RDMA on top of the pervasive TCP over IP (TCP/IP) protocol. (Note that "iWARP" is not an acronym.) iWARP is layered on Internet Engineering Task Force (IETF) standard congestion-aware protocols such as Transmission Control Protocol (TCP) and Stream Control Transmission Protocol (SCTP). As such, iWARP RDMA runs over standard network and transport layers and works with all Ethernet network infrastructure that supports TCP/IP

Since TCP provides reliable delivery, it can provide a reliable network service to upper level applications on top of an unreliable IP network. iWARP is also known for low-latency hardware offload engines on network adapters but such offload requires iWARP-capable network adapters.

RoCE (RDMA over Converged Ethernet)

Developed in 2009 by the InfiniBand Trade Association (IBTA), RoCEv1 (or RDMA over Converged Ethernet) uses Ethernet data link layers and physical layers to support the InfiniBand (IB) transport and network layer (Annex A16 RoCE, Supplement to InfiniBand Architecture Specification Volume 1 Release 1.2.1). RoCEv2 was further developed to operate on top of UDP/IP (Annex A17 RoCEv2, Supplement to InfiniBand Architecture Specification Volume 1 Release 1.2.1). RoCEv2 provides low latency as well. See Figure 2 below. Today, nearly all implementations of RoCE use RoCEv2 and the term "RoCE" generally means RoCEv2.

The User Datagram Protocol (UDP) is a communication protocol used across the Internet for especially time sensitive transmissions such as video playback or DNS (Domain Name System) lookups. It speeds up communications by not formally establishing a connection before data is transferred. This allows data to be transferred very quickly. "The protocol is transaction oriented, and delivery and duplicate protection are not guaranteed. Applications requiring ordered reliable delivery of streams of data should use the Transmission Control Protocol (TCP)" (RFC768 User Datagram Protocol).

Since UDP does not provide flow control or congestion management and RoCEv2 runs on top of UDP, RoCEv2 typically requires Lossless Ethernet and relies on the use of Priority Flow Control (PFC) or a congestion management solution such as Explicit Congestion Notification (ECN, RFC3168) to minimize packet loss in the event of network congestion. RoCEv2 is ideal for deployment within one data center. RoCEv2 also requires RoCE-capable RDMA network adapters (or *rNICs*) for hardware offload. There are RoCE-capable rNICs that can deliver fast RoCE performance without requiring PFC or ECN, but this capability may be vendor specific and might not operate across RoCE-capable NICs from different vendors.

TCP

TCP, or Transmission Control Protocol, is a widely accepted standard that defines how to establish and maintain network communications when exchanging application data across a network. TCP works in conjunction with Internet Protocol (IP), which determines how to address and route each packet to reach the correct destination.

NVMe over TCP (NVMe/TCP) was added to the NVMe-oF Specification v1.1. NVMe/TCP uses standard TCP as a transport for NVMe-oF, thus it can work with any Ethernet network adapter without additional specific requirements and without having to make network configuration changes or implement special equipment. The TCP transport binding in NVMe-oF defines the methodology used to encapsulate and deliver data between two hosts using normal TCP connections.

NVMe/TCP, however, can have its downsides. For example, the specification can increase system processor loads because TCP—in the absence of an adapter that performs TCP offload--relies on the host CPU and OS to process the protocol stack and thus can require additional host system processing power.

NVMe/TCP can also result in higher latency (or response time) rates because additional copies of data must be maintained in the TCP stack. The extent of this latency depend on how the specification is implemented and the type of workloads being supported, and may be reduced by using a network adapter that supports TCP offload.

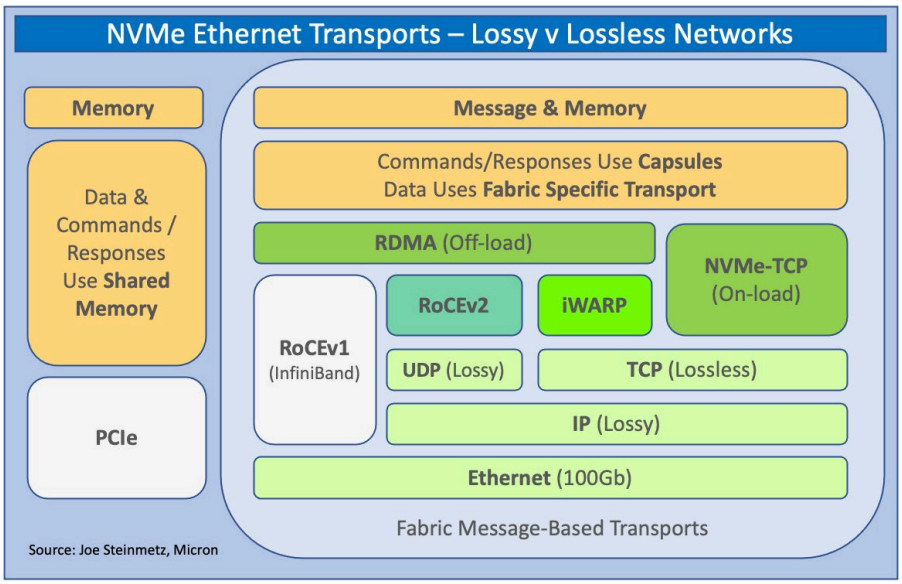


Figure 2 – NVMe Ethernet Transports: UDP v TCP

C. RoCEv2 v iWARP – UDP v. TCP

iWARP and TCP are more tolerant to packet loss than RoCE. iWARP is based on the TCP/IP architecture, which provides flow control and congestion management. Thanks to TCP, in the event of packet loss iWARP supports selective re-transmission and out-of-order packet receive. These technologies can further improve the performance in a Best Effort network.

While the RoCEv2 standard implementation includes a mechanism for recovering from packet loss, it traditionally recommends a "lossless" network because it will experience performance degradation if/when packet loss occurs. To avoid this, RoCE usually uses Layer 2 IEEE Data Center Bridging enhancements (notably Priority Flow Control) and/or Layer 3 ECN to minimize packet loss and ensure in-order-delivery.

D. NVMe-oF: How Mature is It?

NVMe-oF v1.0 specification was released in June 2016 and revised to v1.1 in October 2019 with some refinement and the addition of TCP as a new transport. As of now, there are many Ethernet products on the market supporting NVMe-oF.

There is robust driver support in OS ecosystems. Linux drivers are available for NVMe-oF on both Initiator and Target. VMware has an NVMe-oF initiator. There are also 3rd parties that are providing Microsoft Windows NVMe-oF Initiators.

The University of New Hampshire Inter-Operability Lab (UNH-IOL) also organizes interoperability and conformance tests for various transports among different Ethernet product vendors.

III. Factors Impacting Different Ethernet Transport Performance

A. Scope of Discussion

There are many factors impacting NVMe-oF performance including Host, Switch and Network. This white paper focuses on Host factors and we consider CPU offload vs. onload (software-based) technology, different NVMe drive attributes and their impact on performance, and Maximum Transmission Unit (MTU) frame size (1500B vs. 9000B) in the analysis of RDMA and TCP performance. Accordingly, our testing does not consider Network (e.g., Switch) configurations, settings, topologies, or best practices as test variables.

Host

On the Host server, CPU and memory configuration impact the performance of NVMe-oF, especially NVMe/TCP which relies on the Host OS protocol stack in software-based solutions. NVMe drive attributes also impact the performance of NVMe-oF. Where there are no other performance bottlenecks, NVMe drive performance can still be bottlenecked by IO R/W mix, transfer size and latency attributes present in many workload scenarios.

Switch

Switch settings can impact the overall performance of NVMe-oF. The performance of NVMe-oF can be significantly affected by buffering, oversubscription, the set-up of a dedicated traffic class, as well as congestion control mechanisms for NVMe-oF. This is especially true for NVMe over RoCE, as RoCE usually relies on a lossless network to support high performance. As noted above, this white paper and corresponding test results do not attempt to suggest best practices for various switch conditions or best practices.

Network

Network topologies are other factors to consider. Performance considerations include factors such as: bandwidth over-subscription of the target storage, required fan-in ratios of Initiator and Target, Quality of Service settings, Class of Service configurations, and numerous other conditions. As noted above, this white paper and corresponding test results do not attempt to suggest best practices for various network conditions or best practices.

B. Onload vs Offload

RDMA is a Host bypass and offload technology that results in lower CPU utilization. In NVMe over RDMA, an RDMA engine on an RDMA Network Interface Card (RNIC) bypasses the Operating System (OS) protocol stack and can use direct remote memory-to-memory data access.

Traditional TCP relies on the OS kernel protocol stack. The CPU utilization might not be significant for 1Gb or 10Gb Ethernet but when the network speed moves up to 100Gb, the CPU utilization will go up noticeably. As a result, software-based NVMe/TCP normally consumes more CPU cycles than RDMA for the same workload due to that reliance upon the kernel.

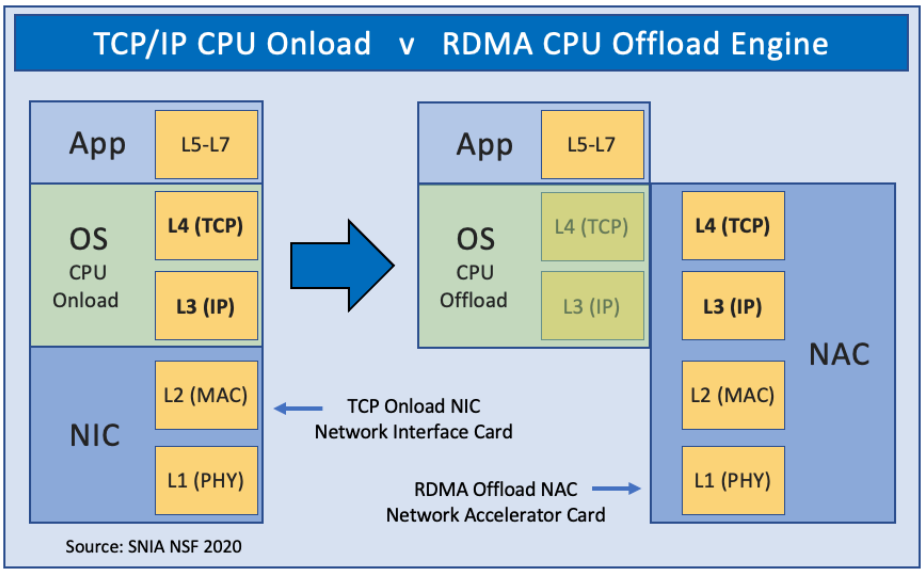


Figure 3 – Onload v Offload Engines

A complete TCP offload engine on network adapter would be able to achieve higher performance with low CPU utilization.

Note: There are also other technologies such as Storage Performance Development Kit (SPDK), (spdk.io), that work in user space and operate in a polling mode with dedicated CPU cores to achieve high throughput and low latency.

C. MTU: 1500B v 9000B

Maximum Transmission Unit (MTU) is the maximum size of the packet, at the Internet Protocol layer, that can be transmitted over a given media without fragmentation. Ethernet frames add an additional 18 byte or 22 byte with IEEE 802.1Q tags. If jumbo frame is supported, the MTU Ethernet frame can be up to 9000 bytes (9KB). Higher MTU size might improve CPU utilization and bandwidth for large IO workloads, but it can also potentially increase latency. The use of Ethernet jumbo frames also requires the jumbo frame setting to be enabled on all servers, switches, and optional routers in the network to ensure proper function.

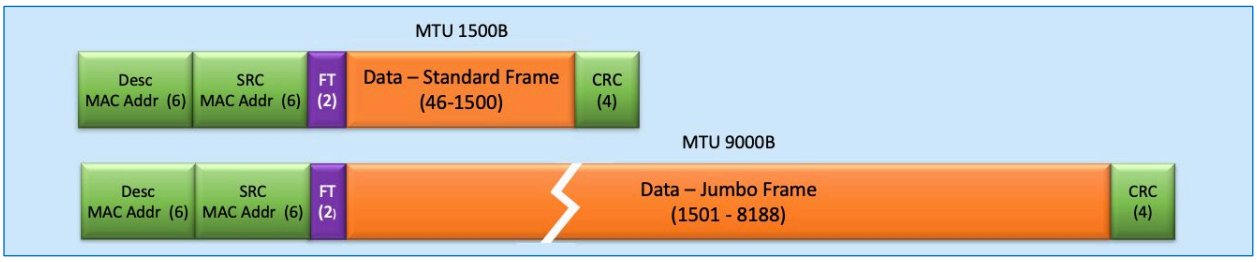


Figure 4 – MTU Frame Size: Standard 1500B v Jumbo 9000B

D. Individual Drive Level Factors

NVMe-oF is very much reliant on underlying NVMe drive performance, especially for Direct Attached Storage (DAS). Other storage systems with storage head nodes on the Target Initiator, such as in Network Attached Storage (NAS) devices add an additional layer of abstraction to the performance discussion. Factors such as Policy settings, Quality-of-Service, and various Erasure Coding and RAID strategies combine with NVMe drive performance and impact the overall performance of NVMe-oF.

Real-world workload IO patterns are also very different. Workloads can be Read intensive, Write intensive or some combination of Reads (R) and Writes (W). IO Streams, which are Random (RND) or Sequential (SEQ) R or W transfers of a specific data size, are also affected by their Demand Intensity (DI), or Outstanding IO (OIO). Here, we use DI, OIO and Queue Depths (QD) interchangeably.

The type of underlying SSD also needs to be considered as NVMe drives perform differently based on these factors. NVMe drives, therefore, need to be selected based on the expected IO Stream content and Demand Intensity to deliver the best performance. Different NVMe drives are designed with different IO characteristics and target performance ranges. Figure 5 below lists the NVMe SSD manufacturer specifications for RND 4K R/W and SEQ 128K R or W performance used in this case study.

Individual drive level characteristics can be masked or modified by each layer of abstraction from storage to Fabrics space and back. For example, using a large cache in front of the SSDs may reduce the observed performance differences between using different types of SSDs. Therefore, SSD-level factors may not have the expected impact on observed Host-level performance. Some examples of SSD-level factors include the following:

Read-Write (RW) Mix. Small amounts of Write IOs may disproportionately impact mixed RW workload performance. Also, drives designed for "Read or Write Intensive" workloads may be based on IO Stream content that is markedly different from the actual application generated workload.

Block Size/Access. Small block RND and large block SEQ IO sizes may have different performance.

IO Streams. Real-world workloads comprised of mixed IO Streams can affect performance differently than synthetic workloads that are comprised of single IO Streams and a fixed Demand Intensity (DI).

Demand Intensity (DI) Saturation. Lower DI can starve IOPS but reduce Response Times (RTs) while higher DI can increase both IOPS & Response Times.

Storage Capacity. Smaller SSD capacity may become saturated, triggering garbage collection and RT spikes.

Bottlenecks in IO Data Path. RTs can be impacted by each component in the IO data path making it difficult to isolate the root cause of RT bottlenecks (see Figure 18 - Response Time Spikes).

Individual Drive Manufacturer Specifications – 3D XPoint v 3D NAND SSD

Figure 5 below lists manufacturer specifications for SSDs used in this study. While 3D XPoint SSDs show symmetric RND 4K and SEQ 128K R/W performance, 3D NAND SSDs show asymmetric RND 4K R/W performance. 3D XPoint SSDs have higher RND W performance and are lower capacity than the 3D NAND SSDs. 3D NAND SSDs have higher SEQ R/W performance and are higher capacity than the 3D XPoint SSDs.

Note that the manufacturer specification optimal Queue Depth (QD) range is lower for 3D XPoint (QD=16) than for 3D NAND (QD=256). This means that while drives can be exposed to any number of QD jobs, the best (optimal) QD and associated performance is listed in the SSD manufacturer specification.

Manufacturer Spec	RND 4K R	RND 4K W	SEQ 128K R	SEQ 128K W
SSD-1: 3D XPoint (6) x 375 GB SSD	550,000 IOPS QD 16	550,000 IOPS QD 16	2,500 MB/s QD 16	2,200 MB/s QD 16
SSD-2: 3D NAND (6) x 4.0 TB SSD	636,500 IOPS QD 256	111,500 IOPS QD 256	3,000 MB/s QD 256	2,900 MB/s QD 256

Figure 5 - SSD Characteristics – Mfgr SSD Specifications

IV. Test Comparison: iWARP v RoCEv2 v TCP

A. Test Plan

Objectives

iWARP, RoCEv2 and TCP. The primary test objective in this study is to compare the performance of iWARP, RoCEv2 and TCP transport protocols across 100Gb Ethernet. We assess the impact of RDMA transports (iWARP and RoCEv2) that utilize CPU offload compared to traditional TCP transport without offload (i.e., CPU onload). As noted above, we do not take into consideration additional significant factors for solution performance, such as network topologies, QoS, switch configurations, or other network best practices.

Workloads. The performance of synthetic corner case workloads (RND 4K RW and SEQ 128K RW) is compared to the performance of three real-world workloads (GPS Nav Portal, Retail Web Portal and VDI Storage Cluster). We observe the impact of synthetic workload RW mix for small block (4K) RND and large block (128K) SEQ corner case workloads. We assess the impact of multiple IO Stream real-world workloads of differing RW mixes (100% W v 66% R v 75% W). We also evaluate the impact of differing sequences of IO Stream combinations and Queue Depths (QDs) over the course of the real-world workload IO Capture.

MTU. We compare standard (1500B) and jumbo (9000B) MTU frame size on performance.

Storage LUNs. We measure the difference in performance between 6-drive 3D XPoint storage LUN and 6-drive 3D NAND SSD storage LUN (RAID 0). We consider IO access patterns (IO Stream size, RW mix and RND or SEQ access), the ability of storage design to handle different workload types, and the ability of LUNs to saturate 100Gb Ethernet with IOs.

Host Factors across Initiator & Target Server

We attempt to normalize Host Initiator and Target Storage server configuration and settings to isolate the impact on performance by Host factors across 100Gb Ethernet (see test set-up below). These Host factors include, among other items:

- CPU offload of RDMA transports (iWARP & RoCEv2) versus CPU onload transport (TCP)
- MTU frame size (standard vs jumbo)
- Test workload composition and settings (see Test Workloads & Test Methodology below)
- IO Stream content (IO transfer size, RW mix, RND or SEQ access and QD)
- Outstanding IO (OIO)/Demand Intensity (DI) – measured by Thread Count (TC) x QD
- IO Size (transfer size or block size)
- Performance across a range of Demand Intensity

Test Topology

In our “Back-to-Back Test Topology” we apply test IOs from a Host across the NMVe-oF transport to a Target without use of a Switch (see Figure 6 below). Test Workloads (2) are generated from the Calypso Test Suite (CTS) IO Stimulus generator (3) which is mounted on the Host Initiator server (4). The CTS IO Stimulus generator is a Calypso IO engine that allows direct, remote or Fabrics test of logical storage. The CTS IO Stimulus generator (3) is a Linux based libaio (Asynchronous IO) which utilizes a 48-bit random number generator to load the test threads and queues with non-repeating, random binary data.

Test scripts are generated from a CTS control server database (1,2,3). IO Stream commands, test settings and test steps for synthetic and real-world workloads (2) are compiled from the CTS database (1) and sent to the CTS IO Stimulus generator (3) residing on the Host Initiator server (4).

The CTS IO Stimulus generator (3) then sends test script IO commands to logical storage (10) across the 100Gb Ethernet Fabrics (7,8) via Intel E810-CQDA2 NIC (5) to the target server (9) and the test storage LUNs

(10). Test result data packets are transmitted to, and archived on, the CTS Control Server database (1) for subsequent display, reporting and data analytics.

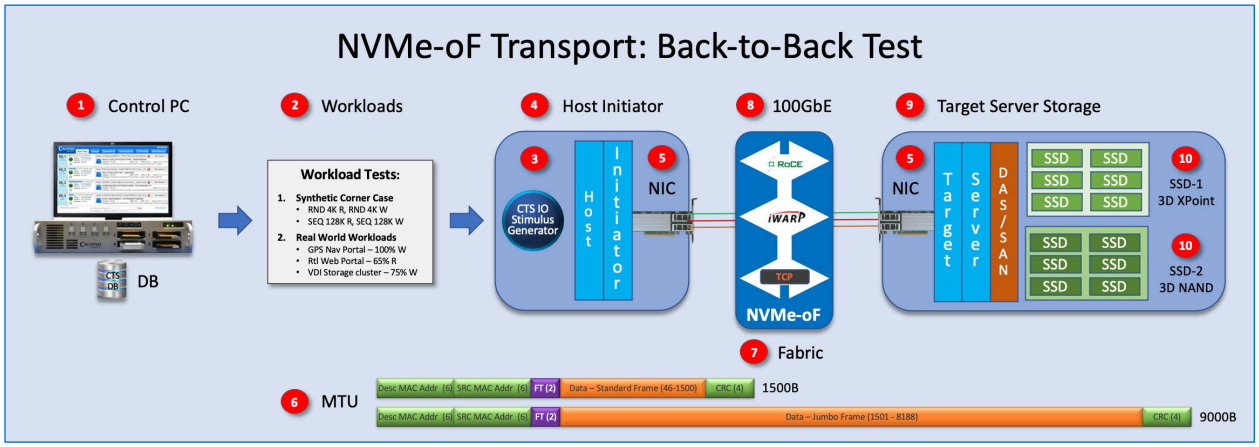


Figure 6 – Back-to-Back Test Topology

B. Test Workloads

NVMe-oF testing applies both synthetic corner case benchmark and real-world application workloads to 3D XPoint SSD and 3D NAND SSD LUNs.

Synthetic Corner Case Workloads

Synthetic corner case benchmark tests apply a fixed and constant workload to the target storage. Corner case tests are designed to saturate target storage to determine performance outside the range of normal operation and to compare IO performance against manufacturer specifications.

The “four corners” benchmark tests are typically small block RND RW (RND 4K) and large block SEQ RW (SEQ 128K). Each IO Stream is applied independently and separately to steady state, usually at a fixed Outstanding IO (OIO) such as OIO = 1 (T1/Q1) or OIO = 128 (T4/Q32).

Real-World Workloads

Real-world application workloads apply various combinations and sequences of IO Streams and Queue Depths (QD) to storage, as observed from real-world workload IO captures. Each real-world workload has a unique composition and sequence of IO Streams and QDs. These IO Streams and QDs are a constantly-changing combination of non-typical block sizes and of varying Outstanding IO (OIO). The intent of real-world application workload testing is to assess the performance of storage in response to an IO workload akin to that observed during application and storage use in real-world deployment. See Figure 7: Real-World Workloads Comparison Table below.

Visualizing Real-World Workloads using IO Stream Maps

IO Stream Maps are used to present the IO Streams, Queue Depths and IO metrics of real-world workloads. IO Stream Maps are derived from IO captures of IO Streams that occur at a given level of the software stack (i.e., file system or block IO) when real-world applications are run. IO Stream statistics are averaged over a given time-step, or time interval of observation. These IO steps are then used to create an IO Stream Map that shows IO Streams, metrics and events (Process IDs) that occurred during the IO capture. Using IO Capture time-step statistics allows viewing of very long duration captures without the associated very large data sets. The real-world application workload testing utilizes the IO Stream statistics derived from the IO Captures.

Real-World Workload Comparison Table

Figure 7 shows the three Real-World Workloads used in this study. While every real-world workload is unique, any workload can be generally characterized by its overall RW mix, the selected or occurring IO Streams, total IOs, total IO Streams that occur during the capture, and by the range of QDs (Minimum, Maximum and Median).

Each workload shows the overall RW mix of the IO Capture, the IO Capture level (file system or block IO), the total IOs observed, the total IO Streams observed, the nine most frequently occurring IO Streams by percentage of IOs, and the minimum, maximum and median QD of the workload.

IO Captures for multiple drives (such as 2-Drive 24-Hour or 6-Drive 12-Hour) means that the IO Streams and metrics for each storage device are consolidated into a single composite IO Stream Map and associated metrics and statistics. Multiple drive composite IO Stream Map consolidation is a feature of the CTS IOProfiler toolset.

Real-World Workload	RW Mix Normalized	IO Capture Level	Total IOs	Total IO Streams	9 Most Frequent IO Streams by % of IOs						Min QD	Max QD	Median QD
Retail Web Portal: 2-Drive, 24-hour	65% R	Block IO	4.5 M	5,086	18.5% 10.0% 4.0% 3.4% 2.7 %	RND RND RND SEQ RND	64K R 8K R 4K W 64K R 8K W	17.0% 8.4% 3.7% 2.9%	SEQ SEQ SEQ RND	0.5K W 8K R 64K W 4K R	5	306	19
GPS Nav Portal: 1-Drive, 24-hour	100% W	Block IO	3.5 M	1,033	21.6% 11.7% 9.6% 3.4% 2.1%	SEQ SEQ RND RND SEQ	4K W 0.5K W 4K W 8K W 1.5K W	12.0% 10.7% 4.9% 2.4%	RND SEQ RND RND	16K W 16K W 8K W 2K W	6	368	8
VDI Storage Cluster: 6-Drive, 12-hour	75% W	Block IO	167 M	1,223	19.3% 9.1% 4.2% 3.3% 2.3%	RND SEQ SEQ SEQ SEQ	4K R 4K R 128K R 4K W 8K R	11.3% 8.2% 3.6% 3.3%	RND SEQ RND RND	4K W 32K R 32K R 8K R	64	1024	128
SNIA CMSI Reference Workloads – Retail Web Portal, GPS Nav Portal and VDI Storage Cluster workloads can be viewed at www.testmyworkload.com													

Figure 7 - Real-World Workloads: Comparison Table

The Retail Web Portal (see Figure 8) can be characterized as 2-drive, 24-hour workload comprised of different retail SQL Server events (such as morning boot storm, opening, daily activities, closing activities and 2 am back-up), a mix of different IO Streams, a range of block sizes from 0.5K to 64K, and a normalized RW mix of 65% R IOs.

The GPS Nav Portal (see Figure 10) can be characterized as a single-drive, 24-hour workload comprised of homogenous IO activity related to GPS Navigation, smaller block size IO Streams, occurrence of periodic SEQ 0.5K IO spikes, and a normalized RW mix of 100% W IOs.

The VDI Storage Cluster (see Figure 12) can be characterized as a six-drive RAID 0 LUN, 12-hour workload comprised of traditional storage block sizes (non-fragmented block sizes), IO Streams with higher QDs (up to 1,024) and a normalized RW mix of 75% W IOs.

Retail Web Portal

Figure 8 below shows an IO Stream Map for a two-drive 24-hour, 66% Read Retail Web Portal which is comprised of different activities and IO Streams over the course of the day. The x-axis indicates *time,* showing both hours and key events over the 24-hour IO Capture. Each data point represents a 5-minute time-step over which the IO statistics are averaged. The y-axis shows *IOs*, *IOPS* and *IO Streams* with their associated IO metrics.

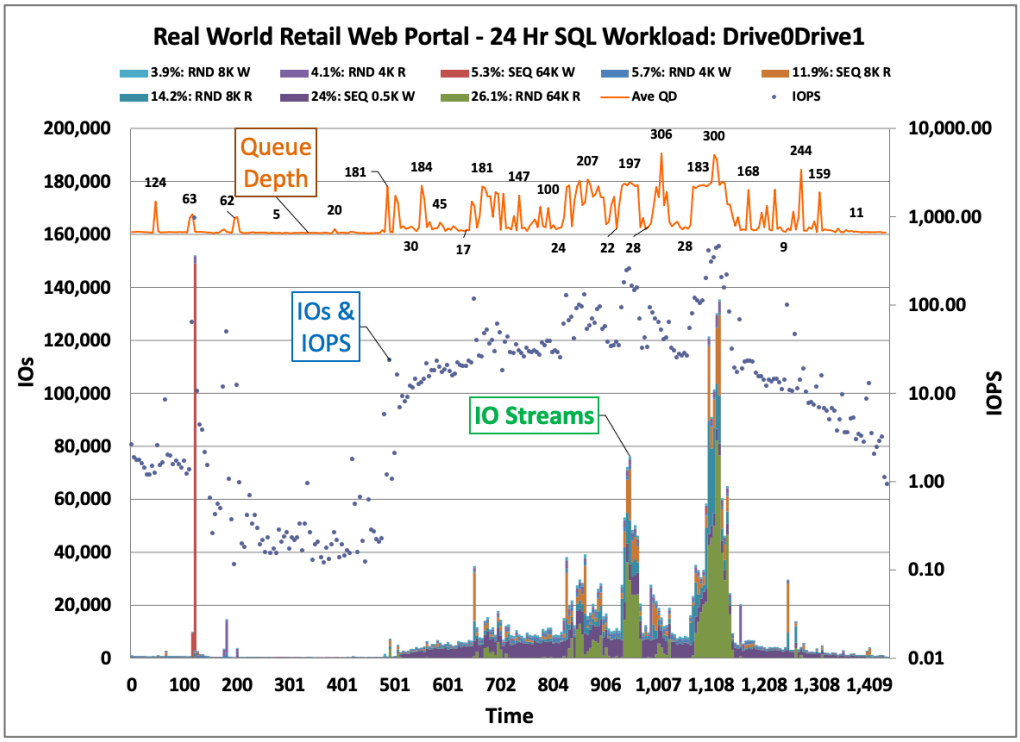


Figure 8 – IO Stream Map: Retail Web Portal

Each color in the stacked columns represents a distinct IO Stream of a RND/SEQ access, block size and RW mix. The orange line shows the average QD for each step of the IO Capture and ranges from QD=7 to QD=306 with a median QD=19. The blue dots are IOs and IOPS that occur over the 24-hour capture.

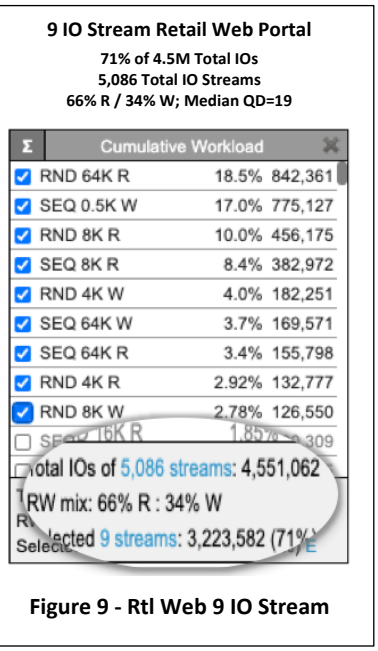


Figure 9 - Rtl Web 9 IO Stream

Note: The red SEQ 64K IO Stream spike at 2 am for Back-up, the low level of blue IOPs and IOs during early morning hours of limited use, the purple SEQ 0.5K W dominant IO Streams during morning boot storm and the mixed IO Streams and peak IOs and IOPS that occur over the course of daily transactions and activities.

Figure 9 shows the Cumulative Retail Web Portal 9 IO Stream Workload selected for display in the IO Stream Map. There were 5,086 total IO Streams observed over the 24-hour IO Capture with an overall 66:34 RW mix. The 9 most frequently occurring IO Streams represent 71% of the total 4.5 million IOs that occur.

The normalized RW mix for the 9 IO Stream workload is 65:35 RW, i.e., while the overall workload has a 66:34 RW mix for all 5,086 IO Streams, there are 65% Reads based on only the 9 selected IO Streams.

The 4 dominant IO Streams by % of IO occurrence are RND 64K R (18.5%), SEQ 0.5K W (17%), RND 8K R (10%) and SEQ 8K R (8.4%). The key characteristics of each real-world workload are summarized in Figure 7: Real-World Workloads Comparison Table.

GPS Navigation Portal

Figures 10 and 11 below show the single-drive IO Stream map and 9 IO Stream Workload for a 24-hour GPS Navigation Portal. The IO Stream map shows a more homogenous composition and occurrence of IO Streams as compared to the multiple event-based Retail Web Portal in Figure 8 above.

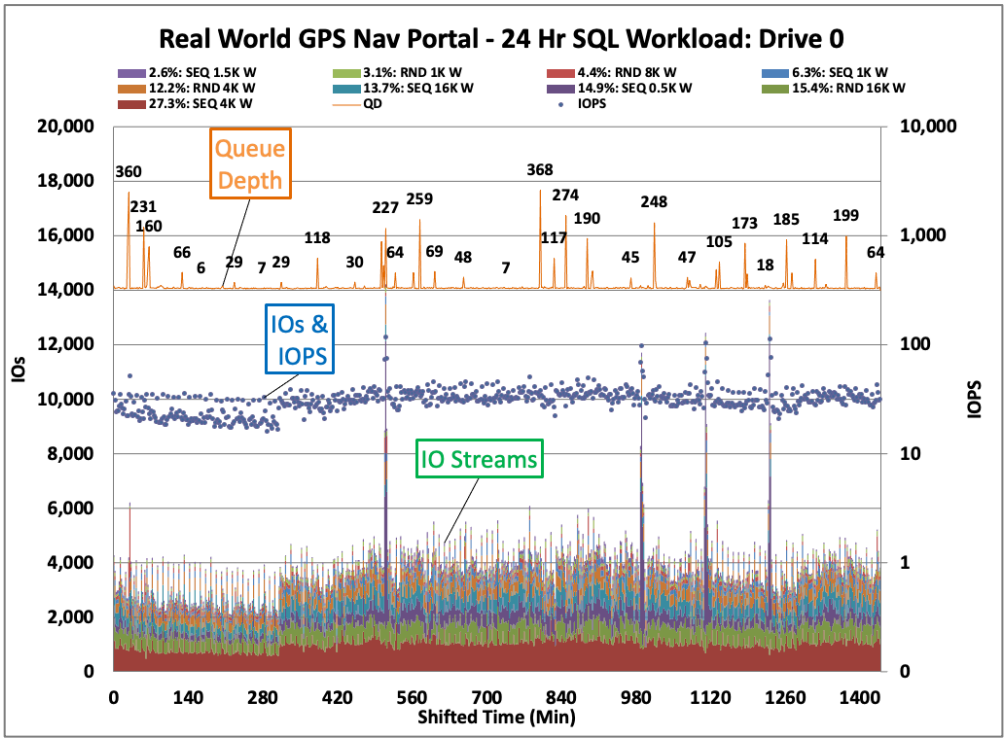


Figure 10 - IO Stream Map: 24-hour GPS Nav Portal

Note: IO Stream maps can display one or more drive IO captures. The CTS IO Stream Map feature can combine the IO Streams and statistics from multiple concurrent drive IO captures into a single composite IO Stream map.

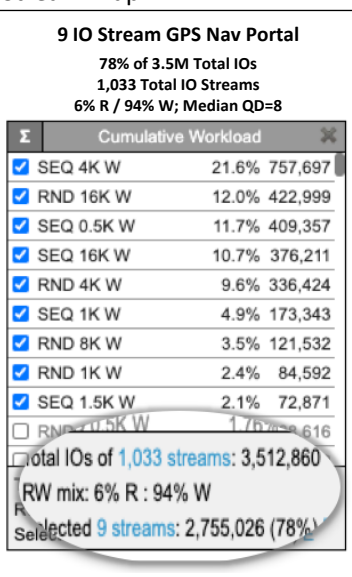


Figure 11 - GPS Nav 9 IO Stream

Figure 10 shows four SEQ 0.5K W spikes of 12,000 IOs and a QD range of QD=6 to QD=368 with a median QD=8. Note that the IOs are more tightly clustered in a band around 10,000 IOs as compared to the Retail Web Portal IO range between 40,000 and 160,000 IOs.

The 9 most frequently occurring IO Streams represent 78% of the total 3.5 million IOs that occur. The normalized RW mix for the 9 IO Stream workload is 100% W, i.e., a RW mix of 100% W is observed when based only on the 9 selected IO Streams as opposed to 94% W for the overall 1,033 IO Streams.

Block sizes for the 9 IO Streams are smaller than the Retail Web Portal IO Streams and range up to 16K (compared to 64 K in the Retail Web Portal workload).

The 4 dominant IO Streams by % of IO occurrence are SEQ 4K W (21.6%), RND 16K W (12%), SEQ 0.5K W (11.7%) and SEQ 16K W (10.7%). See Figure 7: Real-World Workloads Comparison Table.

VDI Storage Server Cluster

Figures 12 and 13 below show the six-drive IO Stream Map and 9 IO Stream Workload for a 12-hour, 75% Write VDI 6-Drive Storage Cluster. The IO Stream map shows a homogenous composition of IO Streams and QDs with varying IO peaks.

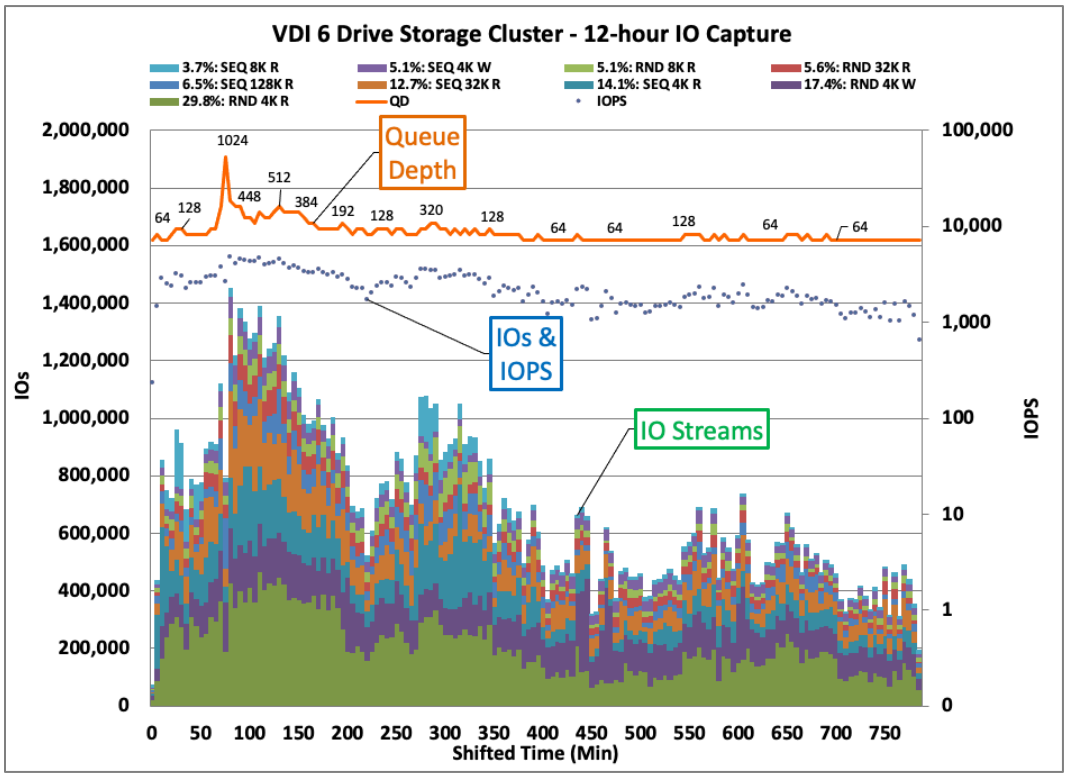


Figure 12 - IO Stream Map: 12-hour VDI Storage Cluster

In Figure 12 above, IOPS vary as QD changes such that there is a peak of 1.4M IOs when QD peaks at 1,024. The QD range is QD=7 to QD=1,024 with a median QD=128. Note that IOs are tightly clustered in a band around 1.4 M IOs as compared to the Retail Web Portal IO range between 40,000 and 160,000 IOs and GPS Nav Portal band of 10,000 IOs.

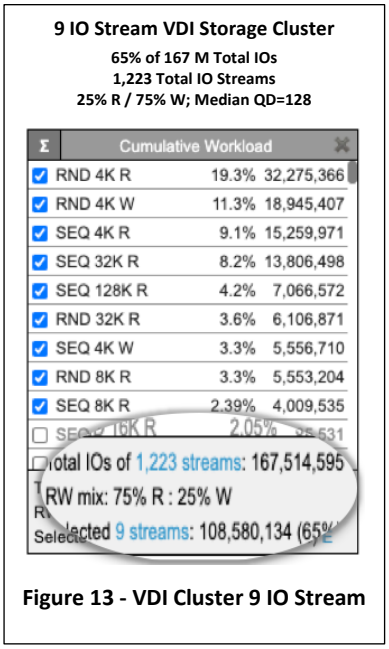


Figure 13 - VDI Cluster 9 IO Stream

In Figure 13 VDI Storage Cluster Cumulative Workload, the block sizes for the 9 IO Streams are predominantly RND/SEQ 4K RW, RND/SEQ 8K RW with some SEQ 32K RW and SEQ 128K R. This reflects the IO block sizes more typically associated with block IO storage.

The 9 most frequently occurring IO Streams represent 65% of the total IOs that occur. Here, there are 167 M IOs compared to 4.5 M and 3.5 M IOs in the Retail Web and GPS Nav Portal workloads. The normalized RW mix for the 9 IO Stream workload is 75% W, i.e., the 75% W RW mix based only on the 9 selected IO Streams, not on all 1,233 IO Streams.

The 4 dominant IO Streams by % of IO occurrence are RND 4K R (19.3%), RND 4K W (11.3%), SEQ 4K R (9.1%) and SEQ 32K R (8.2%). See Figure 7: Real-World Workloads Comparison Table.

C. Test Set-Up

Normalization for NVMe-oF Host Factors. Because IO Stream workloads and composition are affected by each layer of software and abstraction, every effort is made to normalize the impact of the hardware/software stack on performance measurements. This allows us to assess the impact of NVMe-oF Fabrics transport *Host Factors*. See Figure 6: Back-to-Back Test Topology.

We have seen that Real-world workloads are comprised of a constantly changing combination of IO Streams and QDs. The capture and curation of Real-world workloads thus requires faithful and accurate compilation, reproduction and curation of IO Capture steps and IO metrics into test workloads and scripts.

Control PC, Database & CTS Scripting. The Calypso Control Server supports testing of Real-world workloads with the Calypso Test Software (CTS) Database 4.0.1, CTS 6.5 programmatic test scripting and IOProfiler Real-world application workload IO Capture Module 6.5. The CTS Control Server prepares and compiles IO Captures into Real-world workload test scripts.

The CTS Control Server is a SuperMicro X11SSH-F-O LGA 1151 Micro ATX Server, Intel Xeon Quad-Core E3-1225 v5 3.3GHz CPU, 80W 8MB L3, 32 GB 2133 Mhz DDR4 ECC RAM, 64-bit Windows 7 Pro OS, Calypso CTS 6.5, CTS DB 4.0.1 and 10/100Mb/s Ethernet which connects remotely over TCP to the Host Initiator server.

CTS IO Stimulus Generator. The CTS IO Stimulus generator 2.0.2 is mounted on the Host Initiator server. Compiled test scripts are sent from the Control PC to the CTS IO Stimulus generator which then applies test IOs to Target logical storage. Test measurement data packets are transmitted to, and archived on, the Control Server database for subsequent replay, reporting and data analytics.

Host Initiator Intel Server. The Host Initiator is an Intel Server Board S2600WF, Intel Xeon Platinum 8280 2.7 GHz 28 core CPU, 198 GB 2166 Mhz DDR4 ECC RAM, RHEL OS 8.1 kernel 5.7.8 and Intel Ethernet Network Adapter E810-CQDA2.

Intel Ethernet Network Adapter E810-CQDA2. Test IOs are applied to logical storage via Host server NIC card across 100Gb Ethernet to the Target server NIC. The Intel Ethernet Network Adapter E810-CQDA2 Network Interface Card (NIC) supports all Ethernet based transports for NVMe over Fabrics. This allows use of a single NIC in the test set-up and eliminates the need to change NICs when there are changes in Ethernet transports.

A single adapter can handle all Ethernet based traffic including RDMA iWARP, RDMA RoCEv2 and standard TCP. The Intel E810-CQDA2 also features 8 SerDes and MAC which can support multiple port configurations including: 100Gb/50Gb/25Gb/10Gb. The Intel E810 also supports up to 4x 25Gb ports or 8x 10 Gb ports per server.

100Gb Ethernet Cable. QSFP28 Direct Attach 100Gb Ethernet Cable is a high density, low power, passive, direct attach 100Gb Ethernet cable designed for short distance direct interconnect. Here, a 1-meter cable connects the Intel E810 NICs for the Host and Target server in a “back-to-back” (without a switch) configuration.

Target Server. The Target server is an Intel Server Board S2600WF, Intel Xeon Platinum 8280 2.7 GHz 28 core CPU, 198 GB 2166 Mhz DDR4 ECC RAM, RHEL OS 8.1 kernel 5.7.8 and Intel Ethernet Network Adapter E810-CQDA2.

Target Storage LUNs. Target storage consists of two separate six-drive RAID 0 LUNs. SSD-1 is comprised of six-375 GB 3D XPoint SSDs with LUN capacity 2.25 TB. SSD-2 is comprised of six-4TB 3D NAND SSDs with LUN capacity 24 TB. SSD-1 is a lower capacity (2.25 TB) LUN based on 3D XPoint SSDs while SSD-2 is a higher capacity (24 TB) LUN based on 3D NAND SSDs.

Note: Manufacturer specified QD refers to the optimal QD setting to obtain the listed performance values and does not indicate a minimum or maximum QD required for normal drive operation.

Test Set-Up: NVMe-oF Transport Test		
Item	Description	Note
Control Server	Calypso CTS Control Server; CTS test Software 6.5; CTS Database 4.0.1; IOProfiler IPF 6.5	CTS Software, Database & Test Scripting IO Capture; Curation & Creation of Real-world workload scripts; Archival, Analytics & Reporting of Test Results
Real-World Workload IO Capture	Calypso IOProfiler (IPF) Real-World Application Workload IO Capture	Time-step IO Capture of Real-World Application Workloads: Block IO level IO Captures
Test Workloads	Synthetic Corner Case Real-World Application	RND 4K RW; SEQ 128K RW – Single Stream T4Q32 Rtl Web; GPS Nav; VDI Storage Cluster – 9 IO Stream
IO Stimulus Workload Generator	CTS IO Stimulus Generator 2.0.2; Libaio 48-bit RND number generator	Host based AIO Stimulus Generator; Application of Test IOs across NVMe-oF Fabrics to Target Storage; Transmits test results data to CTS Control Server
Host Initiator Server	Intel Server Board S2600WF; Intel Xeon Platinum 8280, single 28 core CPU, 198 GB DDR4 ECC RAM	Intel Xeon 8280 2.7Ghz 28 core CPU, 198GB 2166 Mhz DDR4 ECC RAM, RHEL 8.1 kernel 5.7.8
Network Interface Card (NIC)	Intel Ethernet Network Adapter E810-CQDA2Multi-transport NIC	Intel Ethernet Network Adaptor E810-CQDA2Host NIC & Target NIC; ice-1.1.3, rdma-1.1.21, NVM 2.1 0x8000433E; iWARP, RoCEv2, TCP; Link Flow Control On
100Gb Ethernet Cable	QSFP28 Direct Attach 100Gb Ethernet cable	High density, low power, passive, short distance (1.0m) direct attach 100Gb cable
Maximum Transmission Unit (MTU)	1500B 9000B	Standard frame Jumbo frame
Ethernet Transport	RDMA (iWARP, RoCEv2) TCP	RDMA offset (iWARP, RoCEv2) No RDMA non stateful offload (TCP)
Target Storage Server	Intel Server S2600WF; single 28 core CPU, 198 GB DDR4 ECC RAM	XEON 8280 2.7Ghz 28 core CPU, 198 2166 Mhz DDR4 ECC RAM, RHEL 8.1 kernel 5.7.8
Target Storage LUN – SSD-1	3D XPoint – LUN capacity 2.25TB RAID 0 SSD LUN - 6 x 375 GB SSD	Mfgr Spec: RND 4K IOPS: 550K IOPS R; 550K IOPS W – QD16 SEQ 128K MB/s: 2500 MB/s R; 2200 MB/s W – QD16
Target Storage LUN - SSD-2	3D NAND - LUN capacity 24.0TB RAID 0 SSD LUN - 6 x 4TB SSD	Mfgr Spec: RND 4K IOPS: 636K IOPS R; 111K IOPS W – QD256 SEQ 128K MB/s: 3000 MB/s R; 2900 MB/s W – QD256

Figure 14 - Test Set-up: NVMe-oF Transport Test

D. Test Methodology

Our Test Set-up attempts to normalize the hardware/software environment to isolate the impact of Host Factors on NVMe-oF performance. We apply different Synthetic and Real-World Application Workloads across two RDMA (iWARP and RoCEv2) transports and across TCP. We evaluate performance differences using two types of storage LUNs (SSD-1 3D XPoint v SSD-2 3D NAND) and MTU frame size (standard 1500B v jumbo 9000B).

Single IO Stream synthetic corner case tests are generated by the test software. Multiple IO Stream Real-world application workloads are based on IO Streams observed by IO Step Capture tools, in this case at the block IO level. IO Captures are archived in the Control Server database and used to compile three 9 IO Stream test workloads and to create real-world workload test scripts.

Test scripts are transmitted from the Control Server database to the IO Stimulus Generator on the Host Initiator. The IO Stimulus Generator applies test IOs across the Host Initiator NIC - 100Gb Ethernet cable – NIC to the Target Storage LUNs (SSD-1 or SSD-2). Test measurement data packets are transmitted back across the NVMe-oF to the Control Server database for display, data analytics, post processing and reporting.

Metrics

IOPS, MB/s & Response Times (RT) are referred to and used per SNIA PTS definitions. Higher IOPS & MB/s and lower RTs indicate better performance. Average RT (ART) averages all RTs while Maximum RT (MRT) is the single highest RT observed. "Five Nines" (5 9s) RT Quality-of-Service (QoS) evaluates 99,999 (5 9s) of each 100,000 IO RTs (or drops 1 out of every 100,000 IO RTs.) 5 9s QoS is often referred to as the "long tail" RTs and more accurately reflects RT distributions and hence is known as the IO "Quality-of-Service".

Outstanding IO (OIO), or Demand Intensity (DI), is the total TC x QD of the test IOs being applied to storage at a given time. Corner case test Max OIO=128, real-world workload Replay test Max OIO ranges from 306 to 1,024 while the real-world workload TC/QD Sweep test has a Max OIO=576.

Real-World Workload IO Capture Methodology

The Real-world application workloads used herein are derived from real-world workload IO captures taken at the block IO level using IO Capture tools. Real-world workloads are designed to assess how storage and applications perform to the IO Streams and QDs observed during real-world application and storage use.

Real-world workload source captures are captures of IO Stream activity whose statistics are averaged over a series of pre-defined steps, or time-steps. No personal or actual data is recorded. IO Stream metrics are averaged for each time-step and used to characterize IO Stream activity over time periods (seconds, hours, days or weeks) and to re-create IO Stream Maps from the database. SNIA CMSI reference source workloads and IO Capture tools can be viewed and downloaded for free at www.TestMyWorkload.com.

Pre-Conditioning & Steady State

The Real-world workload Thread Count/Queue Depth Sweep (TC/QD Sweep) test is used as a pre-conditioning and steady state determinant for all tests. This test is first run to steady state after which all other tests are immediately run without a device Purge or pre-conditioning. The application of subsequent test workloads is based on the steady state previously achieved in the initial TC/QD Sweep test.

The TC/QD Sweep test steady state is determined by applying a SEQ 128K W pre-conditioning for twice the LUN User capacity followed by running the SNIA Real World Storage Workload (RWSW) Performance Test Specification (PTS) steady state methodology until steady state is reached. See SNIA RWSW PTS v1.0.7 here.

After SEQ 128K W pre-conditioning, the Applied Test Workload (or the IO Streams selected as the test workload – see RWSW PTS Definitions 2.1.2) is run until five consecutive round measurements of the TC/QD tracking variable meets a maximum 20% data excursion/10% slope steady state window from which all data measurements are reported. In this case, the highest OIO (or TC/QD) combination of the Applied Test Workload is used as the steady state tracking variable.

Synthetic Corner Case Benchmark Tests

Synthetic corner case tests are comprised of four, single IO Stream access pattern tests: RND 4K R, RND 4K W, SEQ 128 K R and SEQ 128K W. In each case, the storage is pre-conditioned to steady state by first applying the Real-world workload TC/QD Sweep test (see above).

After TC/QD Sweep pre-conditioning/steady state is achieved, each corner case workload is run for 10 minutes at an Outstanding IO (OIO) of 128 by setting Thread Count to 4 and Queue Depth to 32 (T4Q32).

The synthetic corner case benchmark tests are used to compare IOPS and RT QoS performance for MTU frame size, storage LUN and NVMe-oF transport. The test results can also be used to compare NVMe-oF LUN performance to the manufacturer performance specifications for the underlying SSDs (Figure 5 above).

Real-World Workload Replay Test

We run two types of Real-world workload tests in this study: Replay test and the TC/QD Sweep test. In the Replay test, we re-create and run the sequence and combination of IO Streams and QDs observed in the IO Capture. In the TC/QD Sweep test, we apply a fixed composite of IO Streams for each step of the TC/QD Sweep test while varying the range of OIO (Outstanding IO or Demand Intensity) to assess IO and response time saturation of the application and storage.

The Replay test is designed to measure how storage performs to an actual workload observed during real-world use. Once a real-world workload is captured as an IO Capture (see IO Capture Methodology), the time-step statistics are used to compile a Replay test script. The test script time-step duration can be set by the user to allow for a very long real-world capture (min, hours, days) to be replayed over a shorter test duration or for a shorter fine grain resolution capture (uSec, Sec) to be replayed over a longer test duration.

The IO operations generated and performed by the Replay test script reflect the various time-step statistics. The Replay test script does not replicate the exact order and timing sequence of all the IO operations as observed within the IO Capture.

Replay test results can be viewed as a "single number" (or average value across time) for easy comparison, i.e., the comparison of IOPS, MB/s and response times averaged over an entire Replay test. However, more importantly, the Replay test allows the test operator to observe measurements for various subsets of time, i.e., for specific time periods, events or Process IDs (PIDs) that occur during the IO Capture. This allows for the analysis of discrete events of interest such as back-up, boot storm, daily activities and more.

Real-World Workload TC/QD Sweep Test

The TC/QD Sweep test is designed to evaluate real-world workload performance saturation as Demand Intensity (or OIO) is increased. In this test, some number of IO Streams of interest, usually the dominant IO Streams by percentage of IO occurrence over the course of the IO Capture, are used to construct a fixed combination of IO Streams for each step of the TC/QD Sweep test. Here, we have selected the 9 most frequently occurring IO Streams for each real-world workload as our composite Applied Test Workload. See Figures 7, 9, 11 - 9 IO Stream Workloads and Figure 13 Real-World Workload Comparison Table.

This 9 IO Stream composite is applied to test storage for each step of test script. After pre-conditioning the storage by applying SEQ 128K W for twice the LUN capacity, the 9 IO Stream workload is run while changing the Demand Intensity for each one-minute period from a low OIO to a high OIO (e.g., OIO=1 to OIO=576). A steady state tracking variable, in this case OIO=576, is observed until performance stays within a steady state window defined as no more than a 20% data excursion and 10% slope for the best linear fit curve for five consecutive OIO rounds.

TC/QD Sweep test results are presented, and can be easily compared, as Demand Intensity (DI) Outside (OS) Curves. The DI OS Curve presents IOPS as a function of increasing OIO. IOPS are on the x-axis while average response times (ARTs) are on the y-axis. IOPS and ARTs are plotted in a best linear fit curve from minimum to maximum OIO. The resultant DI OS Curve shows an increase in IOPS and ARTs as OIO increases.

The figure of merit in a DI OS Curve is the optimal OIO point just before the "knee" of the DI OS Curve where IOPS are high and ARTs have not yet rapidly increased. While the knee of the DI Curve is algorithmically determined (60% slope increase formula), DI OS Curves typically show a sharp increase in ARTs with a commensurate levelling or decrease (or "foldback") in IOPS. See Figure 15 – Demand Intensity Curve and Figure 16 – Demand Intensity Outside Curve below.

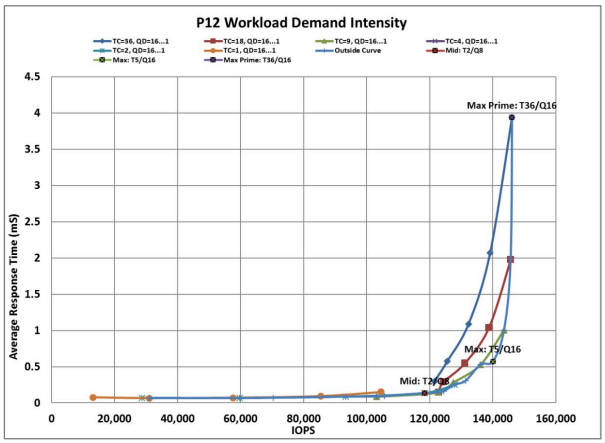


Figure 15 - Demand Intensity Curve

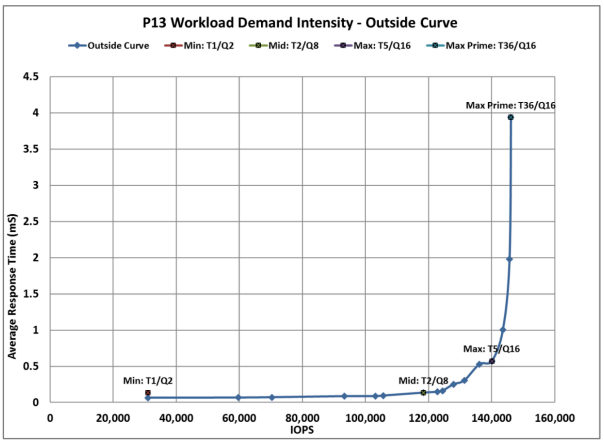


Figure 16 - Demand Intensity Outside Curve

Figure 15 DI Curve above shows each Thread Count (TC) as a data series where the QD increases from QD=1 to QD=16. The yellow line is TC=1 while the blue line is TC=36. In this DI Curve, Min IOPS OIO=T1Q1, Mid IOPS OIO=T2Q8 with Max IOPS OIO=T5Q16. Max Prime IOPS OIO (T36Q16) is defined as the maximum IOPS without regard to ART, i.e., IOPS at the max TC/QD. The figure of merit is the Max IOPS OIO point (T5Q16), or the OIO point where IOPS are high and ART has not yet significantly risen.

Figure 16 above shows the DI Outside Curve which algorithmically determines the Min, Mid, Max and Max Prime IOPS OIO and ARTs and creates the best linear fit outside curve. Here, Min IOPS OIO=T1Q2, Mid IOPS OIO=T2Q8, Max IOPS OIO=T5Q16 and Max Prime IOPS OIO=T36Q16. The simplified DI OS Curve allows for easy comparison of multiple DI OS Curves, tests or storage units (see Test Results – TC/QD Sweep below).

The interpretation of the DI OS Curve tells us that with this specific 9 IO Stream workload (Retail Web Portal), IOPS will continue to rise with increasing OIO until we reach a saturation point of OIO=80 (T5Q16) with 140,000 IOPS and 0.5 mS ART. After this, IOPS level off to 145,000 but ARTs continue to increase with increasing OIO until OIO=576 (T36Q16) where ART reaches 4 mS. To state it another way, after reaching optimal OIO=80 (with 140,000 IOPS and 0.5 mS ART), increasing OIO to OIO=576 nominally increases IOPS but at a cost of increasing ART by 800% (0.5 mS to 4.0 mS), i.e., the cost of an additional 5,000 IOPS is 3.5 mS ART.

Test Flow

This study applies (7) synthetic and real-world workload tests while changing (2) MTU frame size, (3) NVMe-oF transports and (2) storage LUNs. This results in a 7 x 2 x 3 x 2 test matrix of 84 tests.

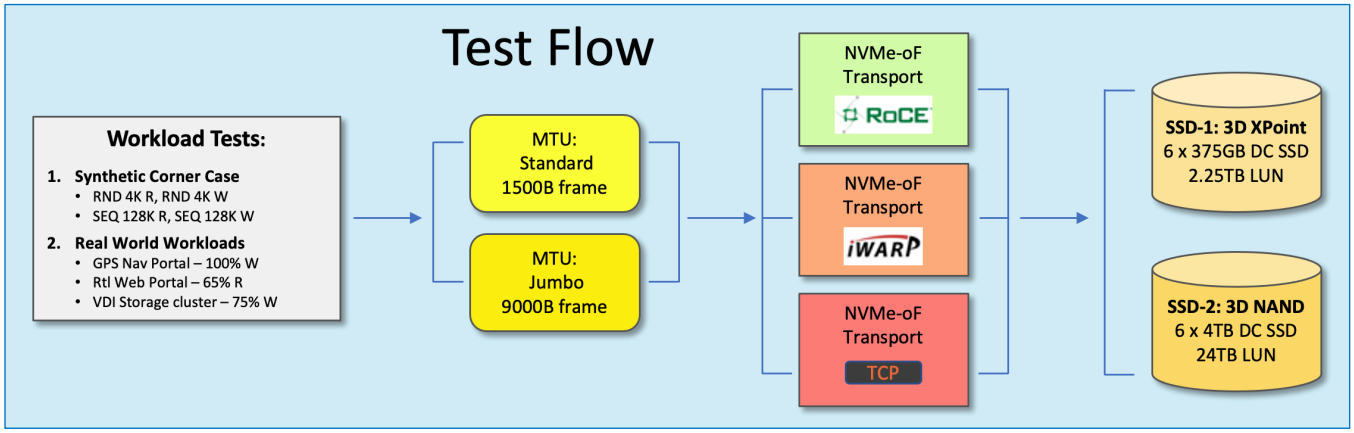


Figure 17 - Test Flow: 84 Test - Test Matrix

E. Test Results

Synthetic Corner Case: RND 4K & SEQ 128K RW

We compare 1500B standard (blue) vs 9000B jumbo (red) frame size across iWARP, RoCEv2 & TCP 100Gb Ethernet transports. We apply single IO Stream synthetic RND 4K RW & SEQ 128K RW workloads to 3D XPoint and 3D NAND storage LUNs. OIO is set to OIO=128 (T4/Q32) as we measure IOPS and 5 9s Response Time Quality-of-Service (RT QoS). Note - SEQ 128K RW IOPS can be converted to MB/s by dividing by 8.

Summary. Standard and jumbo frames show substantially equivalent IOPS except where noted below. iWARP read workloads show very large RT QoS spikes (147mS to 429mS). RDMA (iWARP, RoCEv2) performance with CPU offload is substantially equivalent. RDMA is significantly higher performance than software-based (no-offload) TCP. 3D XPoint performance is higher than 3D NAND except for RND 4K & SEQ 128K R RTs over iWARP and RND 4K R IOPS over TCP.

3D XPoint LUN

- **iWARP** - RND 4K R jumbo IOPS are higher than standard frame. Read workloads show very high RT QoS spikes, perhaps due to Host factors above the SSD storage level (see Figure 18).
- **RoCEv2** - RND 4K RW & SEQ 128 K RW show substantially similar IOPS & RT QoS for standard and jumbo frame size. RoCEv2 does not show RT QoS spikes for Read workloads (see Figure 19).
- **TCP** - RND 4K W IOPS has higher jumbo frame IOPS but substantially similar IOPS for RND 4K R and SEQ 128K RW IOPS. TCP SEQ 128K RW show high RT QoS (see Figure 20).

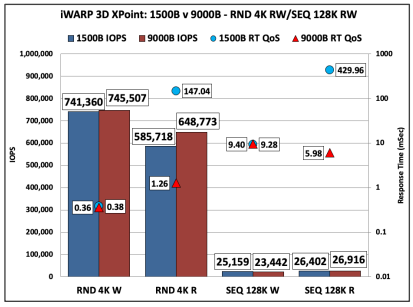


Figure 18 - Synthetic Corner Case
iWARP 3D XPoint: IOPS & QoS – 1500B v 9000B

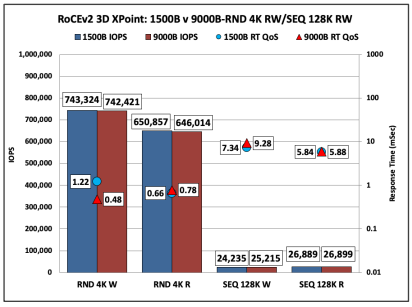


Figure 19 - Synthetic Corner Case
RoCEv2 3D XPoint: IOPS & QoS – 1500B v 9000B

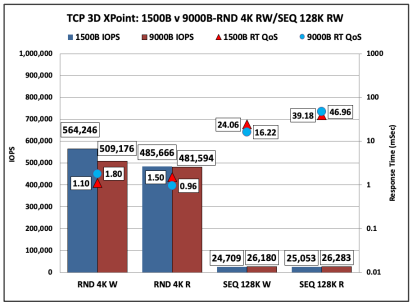


Figure 20 - Synthetic Corner Case
TCP 3D XPoint: IOPS & QoS – 1500B v 9000B

3D NAND LUN

- **iWARP** - RND 4K R & SEQ 128K W standard frame IOPS are higher than jumbo frame. RND 4K W & SEQ 128K R IOPS are similar. SEQ 128K W has high RT QoS (see Figure 21).
- **RoCEv2** - RND 4K RW & SEQ 128 K RW show substantially similar IOPS & RT QoS for standard and jumbo frame size. RoCEv2 SEQ 128K W show high RT QoS similar to iWARP (see Figure 22).
- **TCP** - RND 4K R & SEQ 128K W jumbo frame IOPS are higher than standard frame. SEQ 128K RW show high RT QoS (see Figure 23).

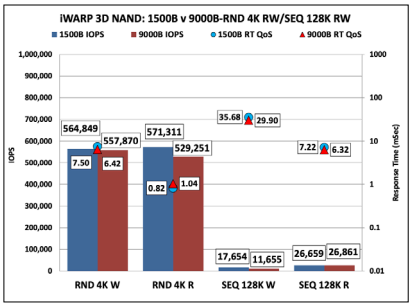


Figure 21 - Synthetic Corner Case
iWARP 3D NAND: IOPS & QoS – 1500B v 9000B

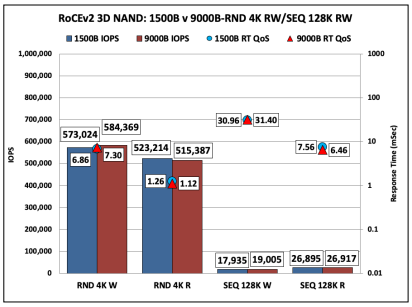


Figure 22 - Synthetic Corner Case
RoCEv2 3D NAND: IOPS & QoS – 1500B v 9000B

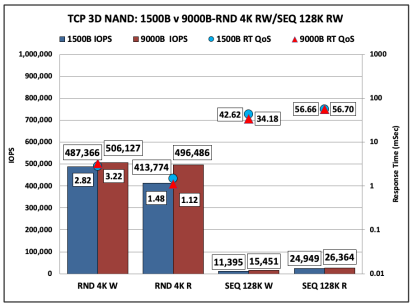


Figure 23 - Synthetic Corner Case
TCP 3D NAND: IOPS & QoS – 1500B v 9000B

Real-World Workloads: Replay Test

We compare 3 Real-world workloads: Retail Web (65% R), GPS Nav (100% W) and VDI Storage Cluster (75% W). We run the 9 IO Stream workloads observed during the IO Capture. Replay tests are run against 3D XPoint and 3D NAND storage using standard (blue) and jumbo (red) MTU frame size. Results show IOPS and RT QoS averaged over the entire Replay test. Max OIO are 306 (Rtl), 368 (GPS) and 1,024 (VDI).

Summary. Standard and jumbo frames show equivalent IOPS and RT QoS. RDMA (iWARP, RoCEv2) IOPS & RT QoS performance is significantly higher than software-based TCP. 3D XPoint storage performance is substantially higher than 3D NAND storage especially for RDMA Retail Web and GPS Nav Portal workloads. However, 3D XPoint and 3D NAND performance is are similar for VDI Storage Cluster workload.

Note that Replay test IOPS here are substantially lower than synthetic corner case test IOPS due, in part, to the different OIO and different type and sequence of IO Streams present in real-world Replay workloads.

3D XPoint LUN

- iWARP standard and jumbo frames show substantially similar performance (see Figure 24).
- RoCEv2 standard frames show higher IOPS for GPS Nav and VDI Cluster Storage (see Figure 25).
- TCP jumbo frames have higher IOPS for Retail Web and VDI Cluster Storage IOPS (see Figure 26).
- RT QoS are similar for standard & jumbo frames. RT QoS is lower for iWARP & RoCEv2 vs TCP.
- 75% W VDI workloads show higher IOPS than 100% W GPS Nav & 65% R Retail Web workloads.

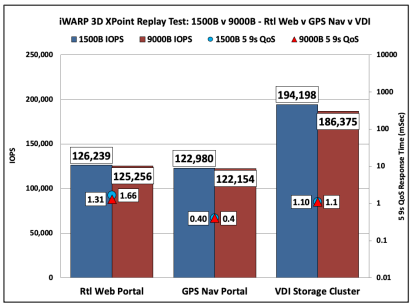


Figure 24 - Replay Test: Rtl v GPS v VDI
iWARP 3D XPoint: IOPS & QoS - 1500B v 9000B

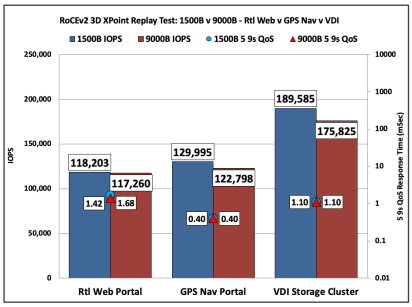


Figure 25 - Replay Test: Rtl v GPS v VDI
RoCEv2 3D XPoint: IOPS & QoS - 1500B v 9000B

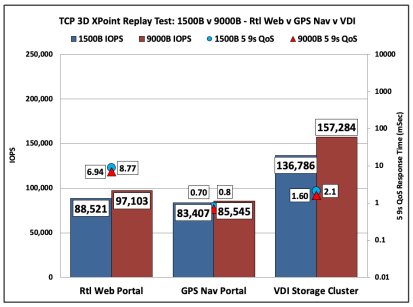


Figure 26 - Replay Test: Rtl v GPS v VDI
TCP 3D XPoint: IOPS & QoS - 1500B v 9000B

3D NAND LUN

- iWARP standard and jumbo frames show substantially similar performance (see Figure 27).
- RoCEv2 standard and jumbo frames show substantially similar performance (see Figure 28).
- TCP jumbo frames have higher IOPS for VDI Cluster Storage IOPS (see Figure 29).
- RT QoS are similar for standard & jumbo frames for iWARP, RoCEv2 & TCP.
- IOPS for Rtl Web & GPS Nav 3D NAND RDMA are lower than Rtl Web & GPS Nav 3D XPoint RDMA
- 75% W VDI workloads show higher IOPS than 100% W GPS Nav & 65% R Retail Web workloads.

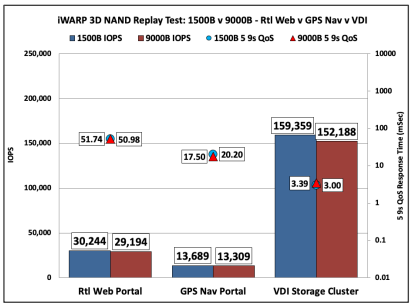


Figure 27 - Replay Test: Rtl v GPS v VDI
iWARP 3D NAND: IOPS & QoS - 1500B v 9000B

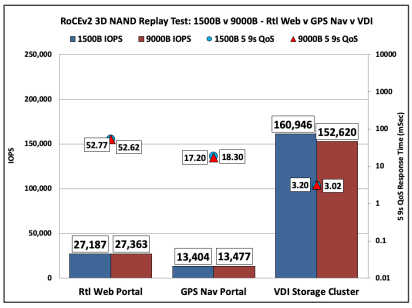


Figure 28 - Replay Test: Rtl v GPS v VDI
RoCEv2 3D NAND: IOPS & QoS - 1500B v 9000B

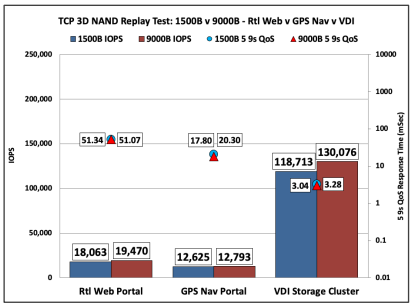


Figure 29 - Replay Test: Rtl v GPS v VDI
TCP 3D NAND: IOPS & QoS - 1500B v 9000B

Real World Workloads: TC/QD Depth Sweep Test

Thread Count/Queue Depth (TC/QD) Sweep test compares Real-world workloads for Retail Web Portal (65% R), GPS Nav Portal (100% W) and VDI Storage Cluster (75% W) as fixed IO Stream workloads. The 9 IO Stream Applied Test Workload is applied as a fixed composite for each step of the test while OIO is changed across a range from OIO=1 (T1Q1) to OIO=576 (T36Q16). We present standard 1500B frame size results only.

The objective of the TC/QD Sweep test is to use the Demand Intensity Outside Curve (DI OS) to observe the IOPS & ART saturation as OIO increases. The key figure of merit, indicated by the enlarged data point, is the Max IOPS OIO point where IOPS are highest while ART has not yet dramatically increased.

Summary. RDMA shows a smoother DI OS Curve with lower max ART and higher IOPS. However, TCP shows higher IOPS at the Max IOPS OIO point across all 3 workloads. RDMA show higher IOPS than TCP as OIO increases beyond the Max IOPS OIO point, i.e., RDMA IOPS increase with OIO but, in this case, at an unacceptable increase in ART. All transports are similar but show differences depending on the workload.

3D XPoint LUN

- **Retail Web 65% R** - TCP shows superior Max IOPS OIO point ART but higher max ART at T36Q16. RDMA shows higher IOPS than TCP beyond the knee of the curve (see Figure 30).
- **GPS Nav 100% W** - TCP shows superior Max IOPS OIO point IOPS & ART (see Figure 31).
- **VDI Cluster 75% W** - TCP shows superior Max IOPS OIO point IOPS ART but higher max ART QoS at T36Q16. RDMA shows higher IOPS than TCP beyond the knee of the curve (see Figure 32).

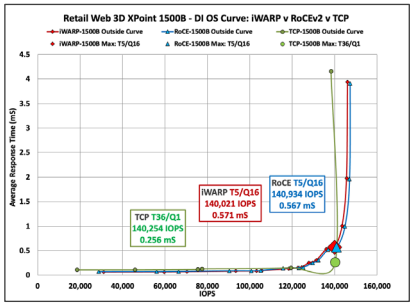


Figure 30 - DI OS Curve: Rtl Web 3D XPoint: iWARP v RoCEv2 v TCP

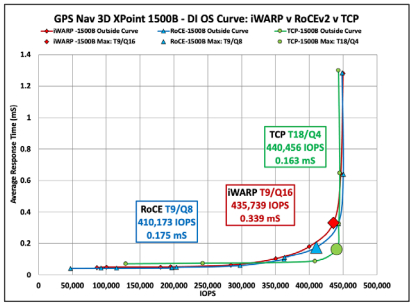


Figure 31 - DI OS Curve: GPS Nav 3D XPoint: iWARP v RoCEv2 v TCP

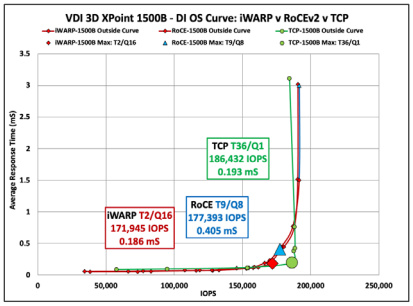


Figure 32 - DI OS Curve: VDI Cluster 3D XPoint: iWARP v RoCEv2 v TCP

3D NAND LUN

- **Retail Web 65% R** – RDMA & TCP show substantially similar DI OS Curves (see Figure 33).
- **GPS Nav 100% W** – RoCEv2 has the lowest RT QoS but lower IOPS, iWARP has higher IOPS & ART while TCP has the highest IOPS and RT at the knee of the curve (see Figure 34).
- **VDI Cluster 75% W** - TCP has the lowest RT and highest IOPS while iWARP & RoCEv2 show lower IOPS and similar RT (see Figure 35).

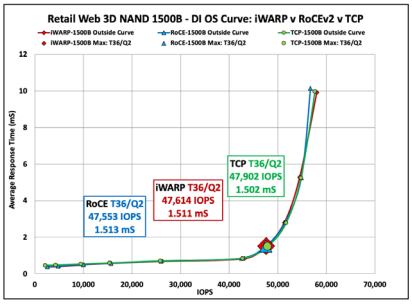


Figure 33 - DI OS Curve: Rtl Web 3D NAND: iWARP v RoCEv2 v TCP

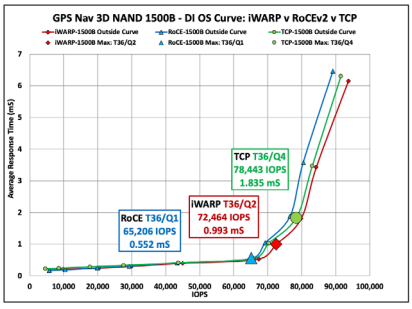


Figure 34 - DI OS Curve: GPS Nav 3D NAND: iWARP v RoCEv2 v TCP

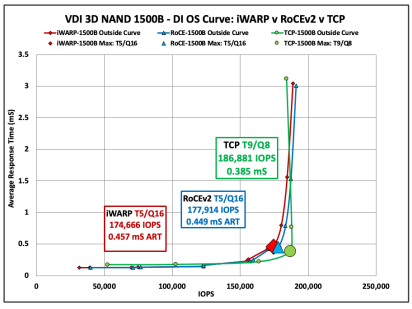


Figure 35 - DI OS Curve: VDI Cluster 3D NAND: iWARP v RoCEv2 v TCP

3D XPoint Storage LUN v 3D NAND Storage LUN

We compare 3D XPoint (blue) vs 3D NAND (red) storage LUNs for synthetic corner case RND 4K & SEQ 128K RW and TC/QD Sweep real-world Retail Web Portal (65% R), GPS Nav Portal (100% W) and VDI Storage Cluster (75% W) workloads. We present MTU standard 1500B frame size results only.

Max OIO (or Demand Intensity) for synthetic corner case workloads is OIO=128 (T4Q32). Max OIO for real-world workloads is OIO=576 (T36Q16). Corner case workloads show IOPS v RT QoS while real-world workload Demand Intensity Outside Curves (DI OS Curves) show IOPS v Average Response Times (ART).

Summary – Synthetic Corner Cases. IOPS for 3D XPoint are higher than 3D NAND especially for Write workloads (RND 4K/SEQ 128K). iWARP 3D XPoint has high RT QoS spikes for Reads (RND 4K/SEQ 128K).

- **iWARP** – 3D XPoint Write workloads have significantly higher IOPS. IOPS for Read workloads are substantially equivalent (see Figure 36). 3D XPoint Read workloads have very high RT QoS, due to Host level, Switch or Network topology factors (see Section III.D. Individual Drive Level Factors).
- **RoCEv2** – 3D XPoint has substantially higher IOPS except that SEQ 128K R IOPS are similar. 3D NAND has significantly higher RT QoS for RND 4K W & SEQ 128K W (see Figure 37).
- **TCP** - 3D XPoint has significantly higher IOPS for Write workloads (see Figure 38).

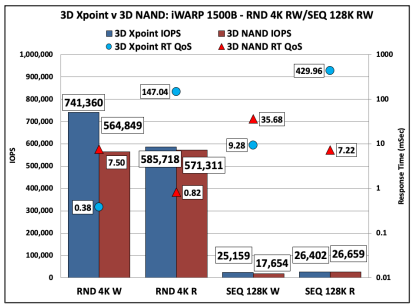


Figure 36 - Synthetic Corner Case
iWARP: 3D XPoint v 3D NAND - IOPS & QoS

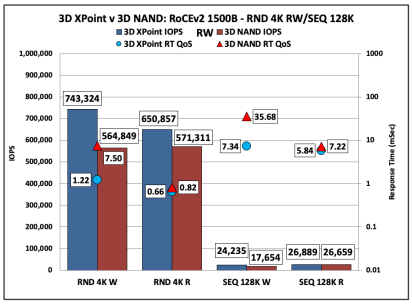


Figure 37 - Synthetic Corner Case
RoCEv2: 3D XPoint v 3D NAND - IOPS & QoS

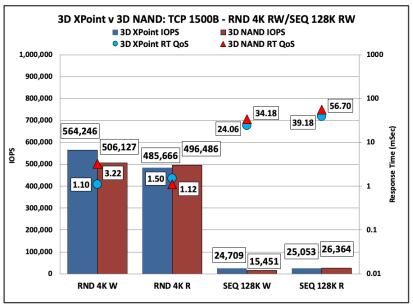


Figure 38 - Synthetic Corner Case
TCP: 3D XPoint v 3D NAND - IOPS & QoS

Summary – Real-World Workloads. DI OS Curves clearly present differences in 3D XPoint and 3D NAND performance and the impact of OIO saturation for different RW-IO Stream-high DI workloads. 3D XPoint RDMA (iWARP & RoCEv2) has substantially higher IOPS and substantially lower ART. For VDI 75% W workload, 3D XPoint and 3D NAND have substantially equivalent IOPS but ART are higher for 3D NAND. In all cases, maximum Response Times at fully saturated OIO (T36Q16) are much higher for 3D NAND than for 3D XPoint.

- **Retail Web 65% R** – 3D XPoint IOPS are substantially higher and ART are lower (see Figure 39).
- **GPS Nav 100% W** – 3D XPoint IOPS are substantially higher and ART are lower (see Figure 40).
- **VDI 75% W** – IOPS are substantially equivalent but 3D NAND has higher ART (see Figure 41).

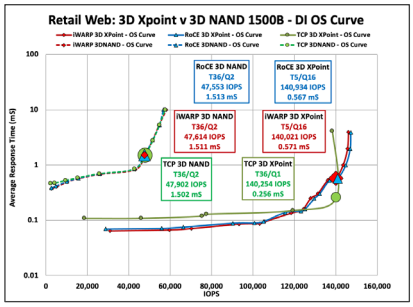


Figure 39 - DI OS Curve: Rtl Web
Retail Web: 3D XPoint v 3D NAND – IOPS & OIO

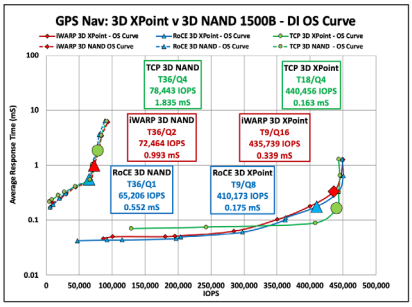


Figure 40 - DI OS Curve: GPS Nav
GPS Nav: 3D XPoint v 3D NAND – IOPS & OIO

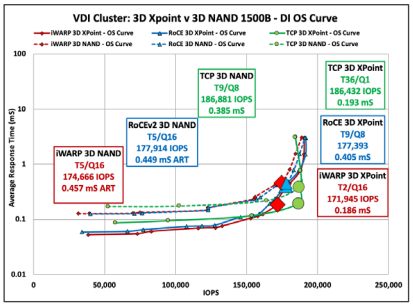


Figure 41 - DI OS Curve: VDI Cluster
VDI Cluster: 3D XPoint v 3D NAND – IOPS & OIO

V. Conclusions

Our key test objective is to evaluate performance differences between RDMA (iWARP & RoCEv2) and TCP 100Gb Ethernet transport while varying workloads, MTU and target storage. We attempt to normalize the HW/SW environment to evaluate the impact of Host Factors (such as MTU frame size, workload, DI, IO Stream content and RW mix). We do not evaluate the impact of Switch or Network topology in this study.

Because we apply test IOs across NVMe Fabrics to logical storage, we can observe the test IO traffic at the Host Initiator level but not directly across, or below, the 100Gb Ethernet cable. Accordingly, test performance results are limited to logical storage at the Host level without discerning the impact or contributions to performance by Switch, Network topology, Storage LUNs or individual SSD Storage.

Underlying SSD performance differences may be obscured by intervening layers of abstraction. For example, in NVMe-oF, RAM cache resides on the target storage server where storage IOs are pooled. This can affect performance when Write IOs are stored in faster RAM cache while certain Read IOs may require accessing the underlying storage directly and thus show slower relative response times. Additionally, while 3D NAND SSDs may have higher Read IOPS (635K v 550K) and lower Write IOPS (111.5K v 550K) specs than 3D XPoint SSDs, actual performance to varying RW mix and IO workloads may not reflect this expected difference in performance.

Notwithstanding the above, we generally observe that RDMA transports share superior performance compared to non-offloaded TCP except for certain cases where corner case Read IOs (iWARP Read workloads) display large RT Spikes. For synthetic corner case workloads, single IO stream performance is superior for 3D XPoint storage over 3D NAND storage. However, there is generally not much difference in performance due to MTU frame-size as both 1500B standard and 9000B jumbo frame sizes show substantially equivalent performance (in IOPS, ART and 5 9s QoS) for both synthetic and real-world workloads.

Test of three-9s IO Stream real-world workloads allows us to evaluate performance differences of multiple IO Streams, varying Demand Intensity, differing RW mix and the impact of constantly changing combinations of IO Streams and Queue Depths on storage and NVMe-oF transport performance.

RDMA performance is superior to non-offloaded TCP for the 100% Write GPS Nav workload and the 65% Read Retail Web workloads. Non-offloaded TCP performance is equal to, if not better than, RDMA for the 75% Write VDI Storage Cluster workload. This difference is likely due to the IO Stream content of the various real-world workloads.

Evaluation of real-world workload Demand Intensity Outside Curves (DI OS Curves) shows the impact of increasing DI on performance. As we saturate storage with increasing Outstanding IOs we see IOPS level off, or even decrease, and RTs begin to dramatically rise. Because real-world workloads have multiple IO Stream content (and changing combinations of IO Streams and QDs), we evaluate the change in Average Response Times (ART) when interpreting DI OS Curves. This allows us to understand the overall impact of the real-world workloads on storage and Fabrics performance.

RDMA DI OS Curves seem to display more deterministic behavior than non-offloaded TCP as IOPS and ART appear to respond more directly and consistently to changing OIO (notwithstanding previously mentioned underlying SSD performance factors). Additionally, RDMA shows lower ART across the DI OS Curve and lower maximum response times at the maximum OIO saturation point as compared to non-offloaded TCP.

Finally, when comparing RDMA transports, we see that iWARP shows nominally higher IOPS in corner case and real-world workloads while RoCEv2 shows lower and more consistent RTs. However, it should be noted that the performance observed between iWARP and RoCEv2 in this study is substantially equivalent.

Other factors that may favor implementation of one or another NVMe-oF transport are not included as test variables in this study. The impact of specific Fabrics configurations, overall NVMe-oF hardware, server and CPU Node configuration, underlying storage devices, CPU resource allocation and network architectures are beyond the scope of this study. Switch and Network topology are, however, planned to be run as test variables in follow-on studies. Questions on this white paper may be directed to the authors at the email addresses first listed above.

About the Authors

Fred Zhang, Intel Corp.

Fred Zhang is a Product Marketing Manager at Intel Corporation where he manages Intel Ethernet Controller products. Fred has over 20 years of experience managing products including silicon products and software products.

Fred is a member of SNIA Network Storage Forum (NSF) and is actively promoting Ethernet based storage solutions to the industry.

Eden Kim, CEO Calypso Systems, Inc.

Eden Kim, CEO of Calypso Systems, is the Chair of the SNIA SSS Technical Working Group, primary author of the SNIA PTS specifications on SSD and Datacenter storage https://www.snia.org/tech_activities/work and author of many SNIA White Papers https://www.snia.org/forums/cmsi/knowledge/whitepapers

Mr. Kim also presents and publishes at world-wide industry trade association events in the US and China and speaks regularly at the SNIA PM Summit, SNIA SDC, Santa Clara Flash Memory Summit (FMS) and the China trade show series sponsored by DOIT in Beijing, Shanghai, Shenzhen, Wuhan and other locations.

Calypso Systems, Inc. is a supplier and manufacturer of advanced workload analysis, test and measurement software, hardware and test services and hosts the www.testmyworkload.com site. Calypso is also the supplier of the SNIA Solid State Storage Performance Test Specification (PTS) Reference Test Platforms (RTP) for SSD and Datacenter testing. Calypso SSD RTPs are standard tools at SSD ODM & OEMs and Calypso IPF servers are fully functioned for RWWs for Cloud, Datacenter and Enterprise customers.

Appendix A: Transport Comparison - Synthetic Workloads

Appendix A: NVMe-oF Transport Comparison [1] – Synthetic Workloads [2,3,4]												
	IO Rate IOPS			**Bandwidth** MB/sec			**Average Response Time** (ART) mSec			**5 9s Quality of Service** (QoS) mSec		
Synthetic Workloads [5,6]	iWARP	RoCE	TCP	iWARP	RoCE	TCP	iWARP	RoCE	TCP	iWARP	RoCE	TCP
RND 4K W – QD128	564,849	573,024	487,366	2,206	2,238	1,904	0.227	0.224	0.263	7.560	6.860	2.820
RND 4K R – QD128	571,311	523,214	413,774	2,232	2,044	1,616	0.224	0.245	0.309	0.820	1.260	1.480
SEQ 128K W – QD128	17,654	17,935	11,395	2,207	2,242	1,424	7.364	7.196	11.256	35.680	30.969	46.620
SEQ 128K R – QD128	26,659	26,895	24,949	3,332	3,362	3,119	4.801	4.759	5.130	7.220	7.560	56.660

Notes

1 RDMA iWARP & RDMA RoCEv2 NAC with Offload; TCP NIC - no Offload
2 Back-to-Back NVMe-oF Transport Topology – 100GbE, Intel E810-CQDA2, No Network Switch; MTU Frame Size 1500B
3 Intel Server S2600WF; XEON 8280 2.7 Ghz 28 core single CPU, 198 GB 2166 Mhz DDR 4 ECC RAM, RHEL 8.1, kernel 5.7.8
4 SSD-2 Storage LUN – 3D NAND NVMe SSD x 6; CTS IO Stimulus Generator, CTS Test Software
5 Synthetic Workloads Pre-conditioned to Steady State per SNIA Solid State Storage Performance Test Specification (PTS) v2.0.2
6 Calypso Test Software (CTS), CTS IO Synthetic Workload Generator and IOProfiler Real World Workload IO Capture toolset

Appendix B: Transport Comparison - Real World Workloads

Appendix B: NVMe-oF Transport Comparison [1] - Real World Workloads [2,3,4]												
	IO Rate IOPS			**Bandwidth** MB/sec			**Average Response Time** (ART) mSec			**5 9s Quality of Service** (QoS) mSec		
Thread Count/Queue Depth Sweep Test [5,6]	iWARP	RoCE	TCP	iWARP	RoCE	TCP	iWARP	RoCE	TCP	iWARP	RoCE	TCP
GPS Nav Portal [7] QD=72/144	72,465	72,465	72,465	477	423	485	0.990	0.550	1.840	13.500	7.400	21.900
Rtl Web Portal [8] QD=72/144	47,614	47,614	47,614	1,205	1,204	1,191	1.510	1.510	1.500	14.050	13.200	14.150
Storage Cluster [9] QD=96/144	174,666	174,666	174,666	2,962	3,031	3,200	0.460	0.450	0.390	2.600	2.200	2.6540
Replay Test [5,6]	iWARP	RoCE	TCP	iWARP	RoCE	TCP	iWARP	RoCE	TCP	iWARP	RoCE	TCP
GPS Nav Portal [7] QD=72/144	13,689	13,404	12,625	96	94	94	0.991	1.015	1.036	20.200	18.300	20.300
Rtl Web Portal [8] QD=72/144	30,234	27,187	18,063	320	298	298	9.778	9.970	9.841	50.980	52.615	51.073
Storage Cluster [9] QD=96/144	159,359	160,946	118,713	2,623	2,648	2,648	0.401	0.398	0.553	3.000	3.020	3.279

Notes

1 RDMA iWARP & RDMA RoCEv2 NAC with Offload; TCP NIC - no Offload
2 Back-to-Back NVMe-oF Transport Topology – 100GbE, Intel E810-CQDA2, No Network Switch; MTU Frame Size 1500B
3 Intel Server S2600WF; XEON 8280 2.7 Ghz 28 core single CPU, 198 GB 2166 Mhz DDR 4 ECC RAM, RHEL 8.1, kernel 5.7.8
4 SSD-2 Storage LUN – 3D NAND NVMe SSD x 6; CTS IO Stimulus Generator, CTS Test Software
5 TC/QD Sweep & Replay Tests per SNIA Real World Storage Workload Performance Test Specification (RWSW PTS) v1.0.7
6 Calypso Test Software (CTS), CTS IO Synthetic Workload Generator and IOProfiler Real World Workload IO Capture toolset
7 GPS Navigation Portal – 24-hour IO Capture; 9 IO Stream; 2 min resolution; 720 steps; Drive0; Block IO level; QD Range 6-368
8 Retail Web Portal - 24-hour IO Capture; 9 IO Stream; 5 min resolution; 290 steps; Drive0 & Drive1; Block IO level; QD Range 5-306
9 Storage Cluster - 13-hour IO Capture; 9 IO Stream; 5 min resolution; 158 steps; Drive0,1,2,3,4,5; Block IO level; QD Range 64-1,024