White Paper

NVMe-oF™ Network Storage Protocol: NVMe™/TCP vs. RDMA with RoCEV2

September 2023

Table of Contents

Introduction

In the world of data storage, speed and efficiency are crucial. Over the years, storage technologies have evolved significantly to keep up with the growing demand for faster data processing and retrieval. One such revolutionary technology that has transformed the landscape of storage solutions is Non-Volatile Memory Express (NVMe™) attached storage. NVMe has not only solved critical challenges in this field but has also become the go-to standard for modern storage solutions.

History

The history of NVMe dates to the early 2000s when traditional storage technologies, such as hard disk drives (HDD) coupled with serial advanced technology attachment (SATA) were the dominant options for data storage. As the demand for faster storage solutions increased, these technologies started to show limitations in terms of speed, efficiency, and scalability. Solid-state drives (SSD) emerged and were gradually adopted into the enterprise space. They offered considerable performance advantages but relied on legacy architecture. Simply put, serial SATA or the newer parallel serial attached SCSI (SAS), while dependable, were a protocol bottleneck for SSD and ultimately hindered their performance potential. This led to the need for a more advanced storage solution protocol. Niche architectures such as Peripheral Component Interconnect Express (PCIe®) attached SSD were introduced. These devices, while highly performant, used proprietary driver stacks. They became a temporary solution while the industry moved towards ratified standards. This ratification led to the emergence of deployable NVMe devices around 2011.

NVMe is a high-performance interface designed specifically for flash-based storage devices. It was developed by the NVM Express® Working Group, a consortium of leading technology companies, including Intel, Western Digital, and Dell, among others. NVMe uses a streamlined and optimized protocol based around the PCIe interface, which is commonly used for connecting high-speed components in a computer system, such as graphics cards and network cards and critically, the CPU complex.

One of the key challenges that NVMe attached storage has addressed is the issue of speed. NVMe drives offer significantly faster data transfer rates compared to those traditional storage technologies discussed above. The NVMe interface allows for parallel data paths, allowing multiple data requests to be processed simultaneously, resulting in dramatically reduced latency, and improved overall performance. NVMe drives can achieve sequential read and write speeds of over 3,000 MB/s, which are several times faster than SATA or SAS SSDs. This has greatly accelerated data processing and retrieval, making it ideal for applications that require high-speed storage, such as big data analytics, artificial intelligence, and real-time processing.

Another challenge that NVMe has addressed is efficiency. Traditional storage technologies suffer from performance bottlenecks due to limitations in their architecture and protocols. On the other hand, the NVMe protocol takes full advantage of the PCIe interface, enabling efficient data communication between the storage device and the host system. This results in lower CPU utilization, reduced power consumption, and improved overall system efficiency.

NVMe also addresses the challenge of scaling capacity. The exponential growth of data in today's digital age requires storage solutions that can scale seamlessly without compromising performance. NVMe drives come in various form factors, including U.2, M.2, and PCIe cards, which provide flexibility for different use cases. NVMe drives can be easily integrated into existing storage infrastructures, and they can be used in different configurations, such as Just a bunch of Flash (JBOF) storage arrays, to meet the storage requirements of modern data centers and enterprise environments.

The Move to Disaggregate Storage

Storage Area Networks (SAN) aside (which also throttle the performance potential of SSD), storage resources were, and in many instances, still are, tightly coupled to servers. This means that each server had its own dedicated storage. This made it difficult to allocate and manage storage resources efficiently. This approach often led to over-provisioning or underutilization of storage resources, as well as inefficient, therefore a costly use of server resources. Further challenges became apparent such a SKU sprawl with isolated storage pools, often defined by application function.

Disaggregated storage, on the other hand, separates the storage from the compute, allowing storage to be shared across multiple servers and accessed more flexibly. This approach enables the storage to be scaled independently of compute, providing greater flexibility to the enterprise architect in managing the infrastructure.

In addition to providing better resource utilization, disaggregated storage also enables enterprise architects to deploy storage in a more efficient manner, as well as allowing them to use advanced storage technologies, such as flash memory and NVMe, which require high-speed connectivity and low-latency networks.

As disaggregation becomes a key trend in modern infrastructure architecture, choices evolve in relation to the network transport protocol(s) available, while questions must be asked about the most relevant protocol to choose.

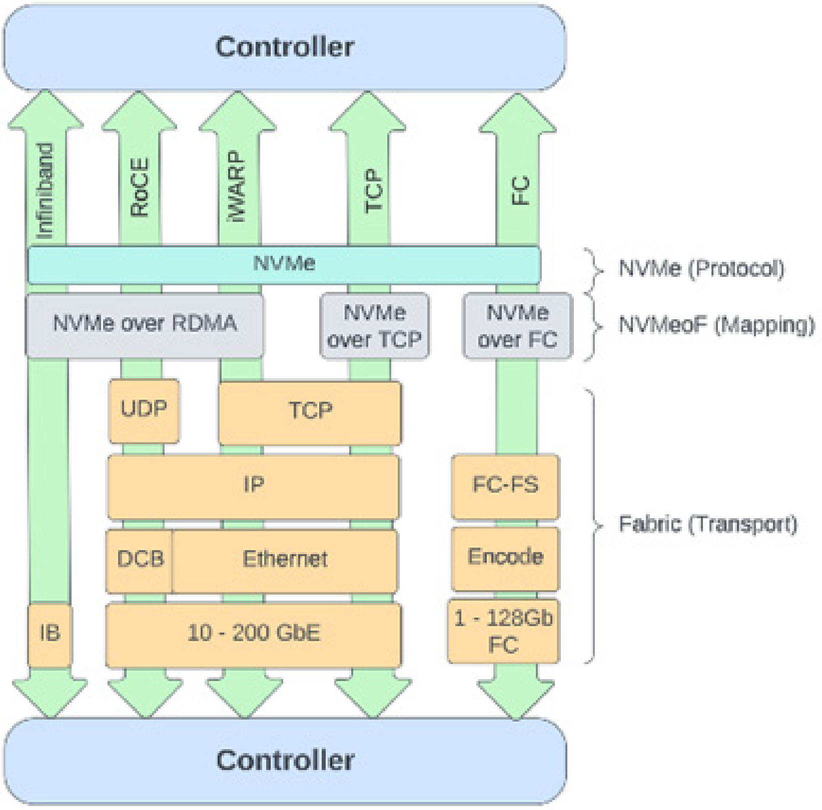
Networked NVMe

NVMe over Fabric (NVMe-oF™) is the storage protocol that expands the NVMe function by enabling host access to remote NVMe storage pools over a network fabric.

The first NVMe-oF 1.0 specification was released in June 2016 and extended NVMe technology to additional transports beyond PCIe, such as Ethernet, Fibre Channel, and InfiniBand. The NVMe-oF 1.1 specification, released in 2019, added finer grain I/O resource management, end-to-end flow control, support for NVMe/TCP and improved fabric communication.

With several protocol choices available for implementing NVMe-oF, each with its own advantages and use cases, organizations are given the opportunity to explore high performance storage networks with latencies that are close to direct attached solutions.

The simplified NVMe-oF Diagram below helps to visualize the options that are available. Understanding those options, their benefits and drawbacks are the main discussion points for the remainder of this document.



The three main NVMe-oF storage fabrics supported by NVMe are:

1. NVMe over Fabrics using Fibre Channel (FC-NVMe)

FC-NVMe is a protocol that extends NVMe over traditional Fibre Channel networks by encapsulating the NVMe command set in FC Frames. It relies on common FC processes such as zoning and integrates seamlessly into current FC protocol. It provides low-latency, high-bandwidth, and reliable connectivity, making it a consideration for those enterprise storage environments where Fibre Channel is already deployed.

2. NVMe over Fabrics using RDMA (NVMe-oF RDMA)

Remote Direct Memory Access (RDMA) is a technology that allows data to be transferred directly between the memory of two systems without involving the CPU or OS, reducing latency and CPU overhead in data transfer. NVMe-oF RDMA uses RDMA protocols, such as InfiniBand or RDMA over Converged Ethernet (RoCE), to enable high-performance, low-latency access to NVMe storage devices over the network. It is suitable for applications that require ultra-low latency and high throughput, such as high-performance computing (HPC), data analytics, and real-time databases.

The three transport protocols below all use NVMe RDMA:

a. InfiniBand is a well-established high-speed interconnect technology. With native RDMA support, InfiniBand allows for high-speed, low-latency, and efficient communication between servers and remote NVMe storage devices. InfiniBand show wide adoption in dedicated networks such as high-performance computing (HPC) environments while its adoption in the Enterprise space is limited.

b. RDMA over Converged Ethernet (RoCEv2) enables efficient routable RDMA communication over UDP on Ethernet networks. These component technologies allow for low-latency, high-throughput, and low-CPU utilization data transfers, making it ideal for HPC, data center, and storage networking applications. RoCEv2 has superseded RoCEv1 which only supported non routable Layer 2 networking with limited data center bridging feature support. While RoCEv1 is still available, it is essentially obsolete within the data center environment.

c. Internet Wide Area RDMA Protocol (iWARP). Communicating RMDA over TCP/IP, iWARP was offered as highly scalable and functional across standard Ethernet infrastructure without the complexity of Data Center Bridging (DCB) and Priority Flow Control (PFC), making it easy to deploy and maintain. One exception regarding infrastructure is the requirement for iWarp capable NIC's. Early adoption by vendors was limited.

3. NVMe over Fabrics using TCP (NVMe/TCP)

NVMe/TCP uses standard TCP/IP networking protocols to transport NVMe commands and data over Ethernet networks. It leverages existing Ethernet infrastructure and is easy to deploy, making it a cost-effective option for data center environments that already have Ethernet networks in place. NVMe/TCP is typically positioned for less latency-sensitive workloads while being deployable over routable Layer 3 networks.

Which Transport Protocol is the Appropriate Choice?

As is often stated, "it depends." The choice of protocol isn't necessarily a straightforward answer and depends on multiple factors. The specific requirements of the storage environment, including performance, latency, cost, and existing infrastructure interoperability all play their role.

This section looks at those additional factors that may influence a choice decision.

NVMe over Fabrics using Fibre Channel (FC-NVMe)

Despite the maturity and stability of Fibre Channel. Its long-term future needs to be considered. After firm adoption for the last 25 years, it is being challenged by the cloud-based architectures, for example, today's hyperscalers, scale-out cloud-like data centers and hyperconverged infrastructure (HCI) platforms.

It is understandable that end users may want to continue using their FC SAN infrastructures. Significant investment would have been required (that may not be fully returned as it may still have longevity) in the solution that offers tested security, reliability, and predictable lossless performance. That lossless guarantee may be a requirement of the applications using it. Would an experienced storage administrator necessarily want to move away from that segregated environment to an Ethernet based infrastructure that is traditionally the domain of Network administrators where a whole new set of network and security protocols need to be learnt, implemented, and benchmarked?

Regardless, SAN networks still present themselves as a premium cost solution that cannot compete long term with NVMe over TCP/RoCE.

Modern Ethernet hardware that supports TCP and RoCE is broadly available and has demonstrated the performance, reliability, and security that storage networks require. Those solutions offer higher performance, lower latency, significantly reduced costs, greater ease of use, and a more clearly defined roadmap for the future. While SAN fabric bandwidth sits at 64Gb today with more progressive switch vendors stating 128Gb, this lags behind Ethernet fabrics which at the time of writing, has vendor support for 400Gb infrastructure.

While FC-NVMe might present itself as a mid-term solution for FC SAN infrastructures, it is not showing itself as a dominant contender in the adoption of NVMe-oF protocols.

NVMe over Fabrics using RDMA (NVMe-oF RDMA)

In the enterprise sector, Ethernet is by far the most popular transport technology which in turn relegates InfiniBand to a niche offering.

a. InfiniBand: Traditionally the protocol of choice in HPC environments, InfiniBand as an option, would require a forklift upgrade to an existing Ethernet infrastructure – thus making it way too costly for most enterprise data centers. Given its low level of adoption outside of HPC, InfiniBand will not be discussed further in this document.

b. RoCE V2 is easily the most popular method for deploying RDMA in Ethernet based data centers. RoCE provides high performance and low latency at the adapter (in the 1-5us range) but requires a lossless Ethernet network to achieve low latency operation. This means that the Ethernet switches integrated into the network must support data center bridging and priority flow control mechanisms so that lossless traffic is maintained. It is therefore likely that existing infrastructure will have to be upgraded / reconfigured to use RoCE. The preference for custom RDMA capable NIC's (RNIC) and drivers are an additional cost consideration. The availability of Soft RoCE does allow RDMA functionality on systems that do not have dedicated hardware support for RDMA. Soft RoCE emulates the RoCE protocol in software, which can be useful when dedicated RDMA hardware is impractical or cost prohibitive. It's important to note that Soft RoCE may introduce additional latency compared to hardware based RoCE implementations since it relies on the host system's CPU for processing RDMA operations.

The challenge with the lossless or converged Ethernet environment is those configurations are a complex process. Without the use of Layer 3 congestion control mechanism like ECN, scalability can be very limited in a modern enterprise context. RoCE V2 is by no means a plug and play solution that would necessarily fit into an existing data center infrastructure without significant cost and skill set investment.

c. iWARP was originally positioned as an alternative to RoCE. iWARP has seen a lack of industry adoption and limited vendor support. Its stated advantages are its ability to support RDMA while integrate into existing TCP network infrastructure, with reasonable latencies of 10-15us, along with the ability to scale more easily than RoCE. That integration, however, involves additional complex transition layers in the transport section of the network stack. This complexity introduced inefficiencies in the stack and so compromised the fundamental targets of high throughput, low latency, and low CPU utilization. Ultimately iWARP demonstrated that it was more complex and expensive to implement while also requiring some dedicated hardware (NIC). All while being less performant than RoCE. Given its low level of adoption, iWARP will not be discussed further in this document.3. NVMe over Fabrics using TCP (NVMe/TCP).

NVMe/TCP

Emerging more recently as an available alternative to RoCEv2, NVMe over TCP claims high performance with lower deployment costs and reduced complexity in deployment. A key differentiator here is that NVMe over TCP extends NVMe across the entire data center using existing infrastructure and standard drivers – an area where RoCE v2 struggles by default.

TCP is very common, well understood, and highly scalable. There is a huge ecosystem of vendors in the TCP world, making major investments in improving its performance capabilities. While RoCEV2 currently offers superior performance and lower latency, that gap may close significantly as NVMe over TCP efficiencies are explored and improved by vendors.

It should be noted, that overall, the NVMe/TCP protocol, like RoCEv2, is still early in its adoption. While vendors support does appear strong, implementation challenges have been seen in relation to vendor interoperability and network performance. While NVMe/TCP can be implemented in an existing network, that network design and utilization can have a huge impact on NVMe/TCP performance. The NVMe/TCP protocol itself has potential networking challenges that may contribute to latency issues. Examples such as head-of-line blocking and incast scenarios need to be appreciated and practical workarounds explored.

Head-of-line blocking can occur when multiple packets are sent over the network, and a packet in the middle of a sequence is delayed or lost. This delay or loss of a single packet can cause all subsequent packets to be held up or blocked at the receiving end until the delayed or lost packet is received or retransmitted. This is a consequence of the in-order delivery guarantee provided by TCP.

Incast is a networking challenge that can occur in data center environments when multiple servers simultaneously request data from a single target, resulting in a congestion collapse as the incoming traffic exceeds the network's capacity to handle it. Both scenarios are undesirable and while there are approaches available to mitigate the impact such as traffic engineering and flow control, it adds administrative burden and the need for proactive network monitoring and traffic analysis.

Tool such as Data Center TCP (DCTCP) aims to address the challenges of congestion control in data center networks, where traditional TCP congestion control algorithms may not be suitable due to their conservative behavior. DCTCP aims to minimize packet loss and improve network utilization by detecting and reacting to congestion proactively. DCTCP operates by monitoring the congestion level in the network and adjusting the TCP sender's transmission rate accordingly. It achieves this by calculating the congestion level based on the difference between the average queue length and a congestion threshold. By utilizing Explicit Congestion Notification (ECN), DCTCP allows network switches to notify the sender about congestion before it leads to packet loss. This enables DCTCP to respond quickly and reduce its transmission rate, preventing further congestion and maintaining a high throughput.

While NVMe over TCP can be considered somewhat plug and play, it is evident that true performance and latencies are very much dependent on the implementation and utilization of the existing infrastructure and so expectations may need to be tempered accordingly.

The next section highlights the current storage platform offering from Western Digital; the OpenFlex™ Data24 3200 Series NVMe-oF Storage Platform.

OpenFlex Data24 3200 Series NVMe-oF Storage Platform

Western Digital's OpenFlex Data24 3200 series NVMe-oF storage platform extends the high performance of NVMe flash to shared storage. Similar to the original OpenFlex Data24, it provides low latency sharing of NVMe SSDs over a high-performance Ethernet fabric to deliver similar performance to locally attached NVMe SSDs. Unsurpassed connectivity in its class using Western Digital RapidFlex™ NVMe-oFcontrollers, allows up to six hosts to be attached without a switch like a traditional JBOF. The OpenFlex Data24 3200 series uses Western Digital's RapidFlex C2000 Fabric Bridge Adapters to provide 2, 4, or 6-ports of 100GbE which can now connect to RDMA (RoCEv2) and/or TCP configured host initiators, offering flexibility of connectivity to either RoCE or TCP host ports for optimum usage.

By enabling applications to share a common pool of storage capacity, data can be easily shared between applications or needed capacity can be allocated to an application to respond to application needs.

The OpenFlex Data24 3200 series NVMe-oF storage platform can also be used as a disaggregated storage resource in an open composable infrastructure environment using the Open Composable API. The platform can also be specified with just two RapidFlex adapters for simpler environments and as a direct replacement for SAS external storage.

The OpenFlex Data24 3200 Series design exposes the full performance of the dual port NVMe SSDs to the network. With 24 Western Digital Ultrastar DC SN840 3.2 TB devices, the enclosure can achieve up to 71.4 GB/s of 128K bandwidth and over 16.7 MIOPS at 4K block size.

OpenFlex Data24 3200 Series NVME-oF Storage Platform

With TCP or RoCEv2, the OpenFlex Data24 3200 is effectively protocol agnostic. This is important from a customer standpoint, as stands to offer the best fit for their application criteria.

OpenFlex™ Data24 3200 Series NVMe-oF Storage Platform–RoCE vs. TCP Benchmarks

Quality of Service Benchmarks

Quality of Service (QoS) analysis serves to measure the consistency of performance in a test environment. That QoS data is represented graphically via Exceedance Charts. An Exceedance Chart is a logarithmic scaling chart where (in this instance) we look at "Nines" on the vertical (y) axis and latency (response time) on the horizontal (x) axis.

The "Nines" correspond to the percentage of completed IO's for a given response time. Fundamentally we are observing how the test infrastructure performs under the highest levels of load.

Some points for reference:

- Generally, 4 nines (0.9999) is the accepted reference point for observing characteristics.
- Referencing beyond the 6 nines (0.999999 - yellow section of the chart) becomes less reliable for statistical reference due to the sheer volume of IO's that would need to be captured.
- The more vertical the curve, the better as you are increasing the Nines without an increase in latency.
- The more horizontal the curve, the worse as you are increasing latency without improving the Nines.
- In general, an Exceedance Curve that is more jagged or angular suggests that an insufficient number of data points have been generated.
- While beyond the scope of this document, in general (with sufficient IOs), every change in the slope represents:
  - A change in technology (e.g., moving between different caches).
  - A characteristic of the architecture (e.g., different bandwidths for different buses).
  - Different algorithms (e.g., a change in bias or behavior between workloads – in block sizes, arrival rates, etc.).
  - Firmware efficiency and firmware errors.
  - NAND management techniques, etc.

Benchmark Results: RoCE vs. TCP QOS Exceedance Chart

Key Observations

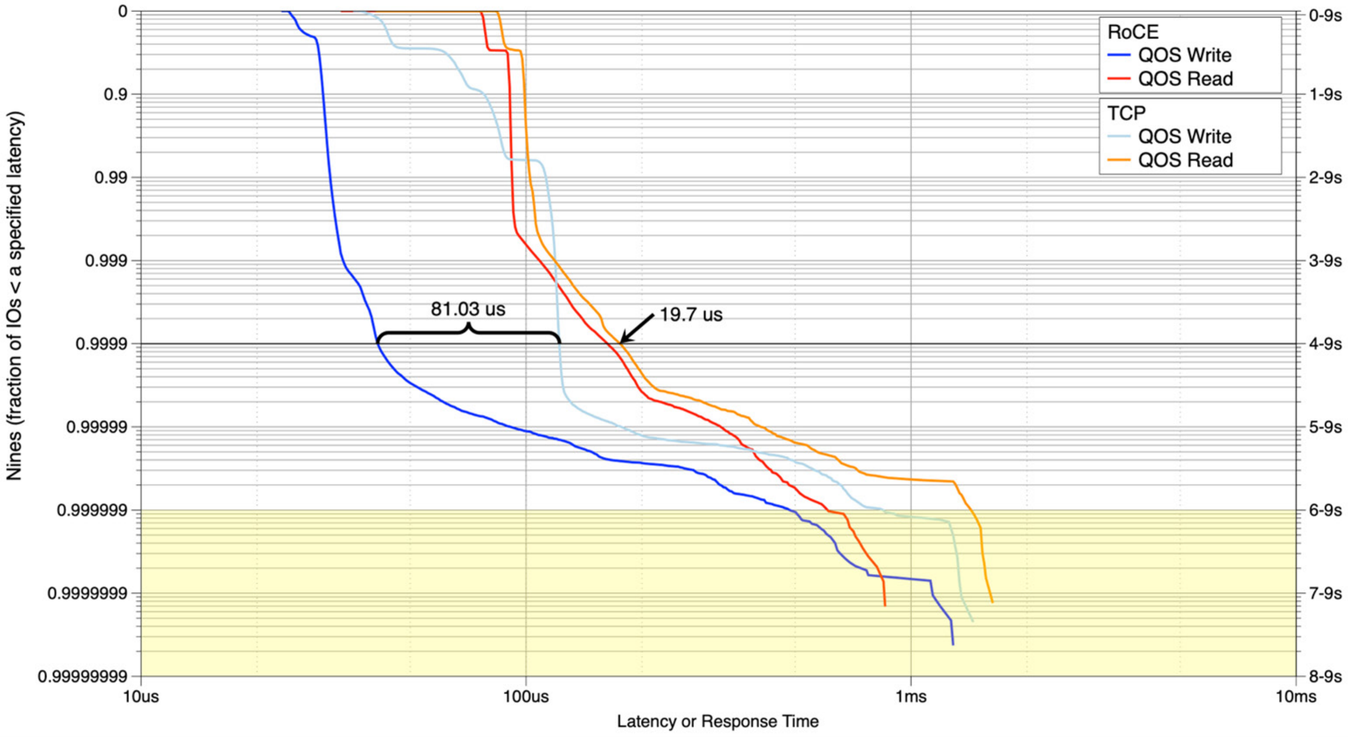
Test Parameters:

- Workload Profile: Random
- Block Size: 4k
- Number of Jobs: 1
- Queue Depth: 1
- Run Time: 20 minutes

All results were measured at the 4 nines (0.9999).

- RoCE QOS Writes: 41.22 us
- TCP QOS Writes: 122.37 us
- RoCE QOS Reads: 162.82 us
- TCP QOS Reads: 177.15 us

In both test scenarios, we see that RoCE outperforms TCP with 8% lower Read latency and 66% lower Write latency.



Benchmark Results: RoCE vs. TCP Sequential IO Exceedance Chart

Key Observations

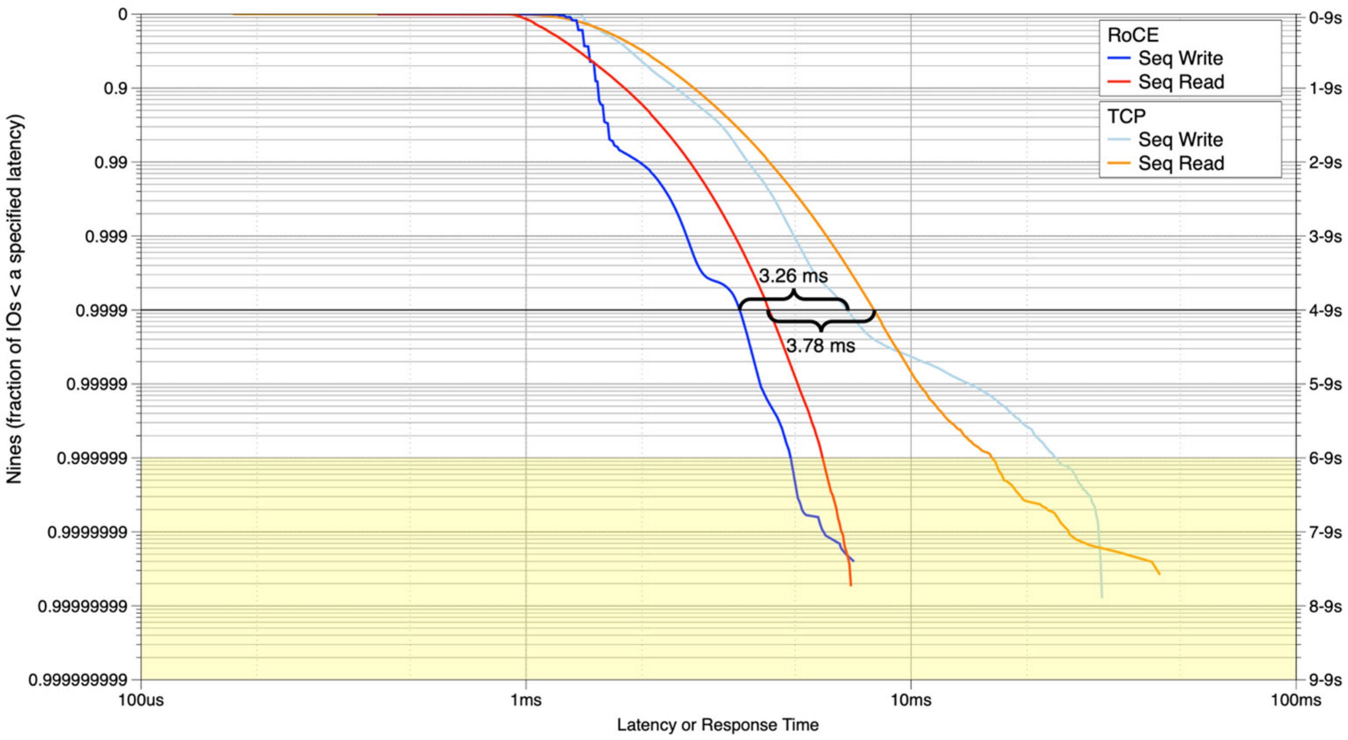
Test Parameters:

- Workload Profile: Sequential
- Block Size: 128k
- Number of Jobs: 1
- Queue Depth: 16
- Run Time: 20 minutes

All results were measured at the 4 nines (0.9999).

- RoCE Sequential Writes: 3.61 ms
- TCP Sequential Writes: 6.39 ms
- RoCE Sequential Reads: 4.29 ms
- TCP Sequential Reads: 8.35 ms

In both test scenarios, we see that RoCE outperforms TCP with 49% lower Read latency and 43% lower Write latency.



Benchmark Results: RoCE vs. TCP Random IO Exceedance Chart

Key Observations

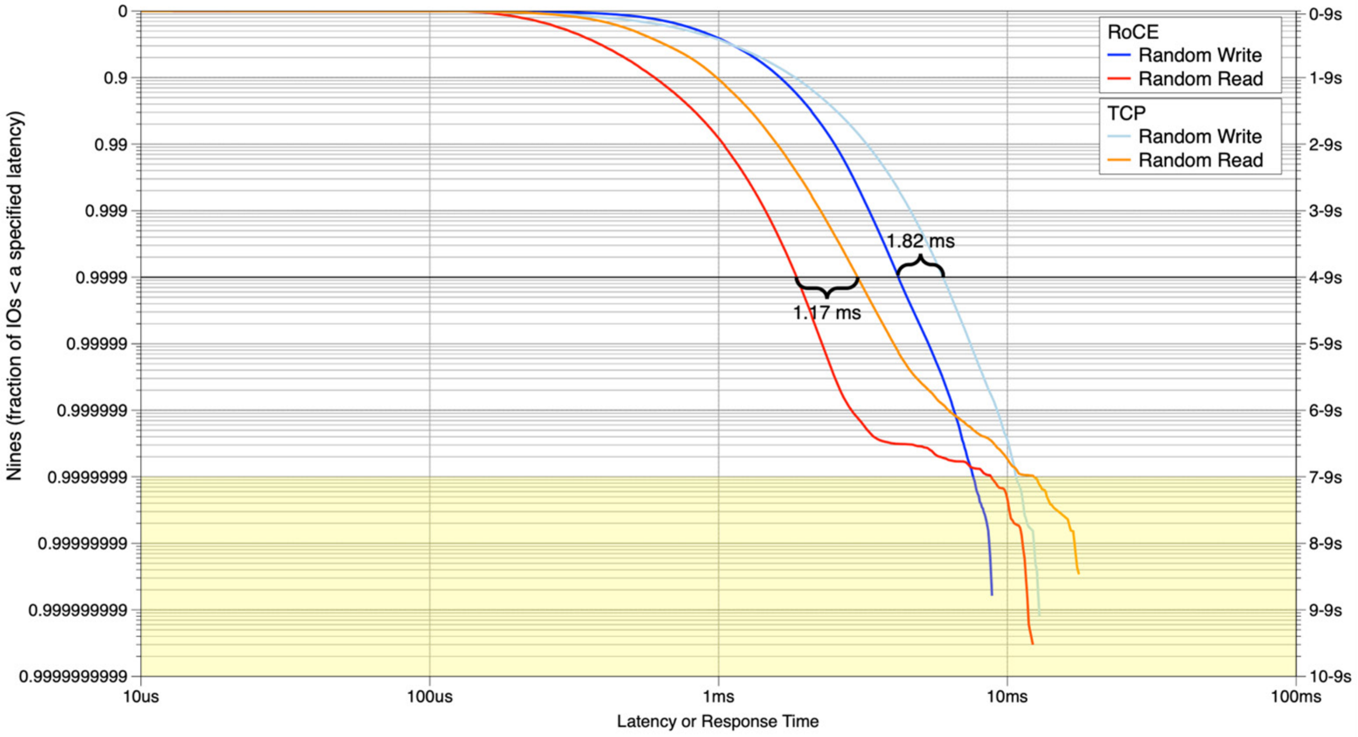
Test Parameters:

- Workload Profile: Random
- Block Size: 4K
- Number of Jobs: 32
- Queue Depth: 32
- Run Time: 20 minutes

All results were measured at the 4 nines (0.9999).

- RoCE Sequential Writes: 4.22 ms
- TCP Sequential Writes: 5.99 ms
- RoCE Sequential Reads: 1.87 ms
- TCP Sequential Reads: 3.03 ms

In both test scenarios, we can see that RoCE outperforms TCP with 38% lower Read latency and 30% lower Write latency.



Benchmark Results: RoCE vs. TCP Random IO Exceedance Chart

Key Observations

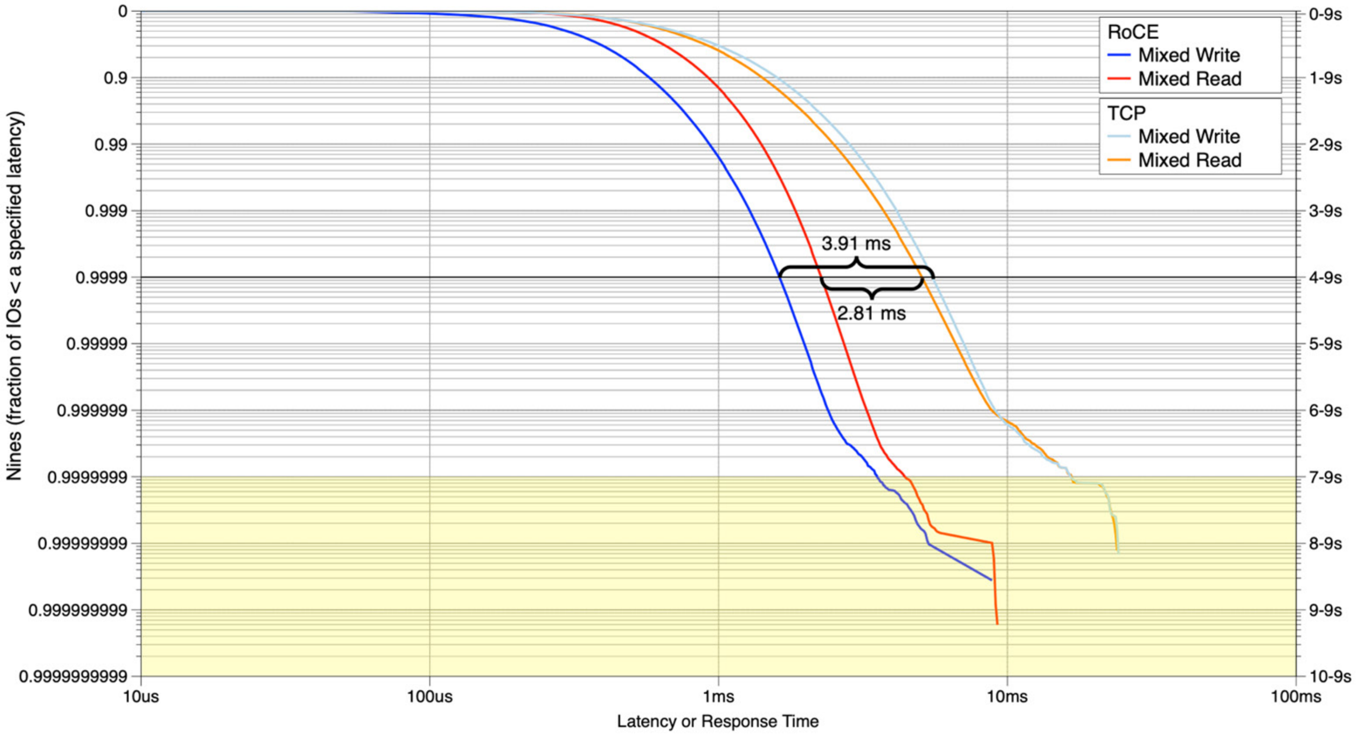
Test Parameters:

- Workload Profile: Random
- Block Size: 4K
- Number of Jobs: 32
- Queue Depth: 32
- Read / Write Mix: 70 / 30
- Run Time: 20 minutes

All results were measured at the 4 nines (0.9999).

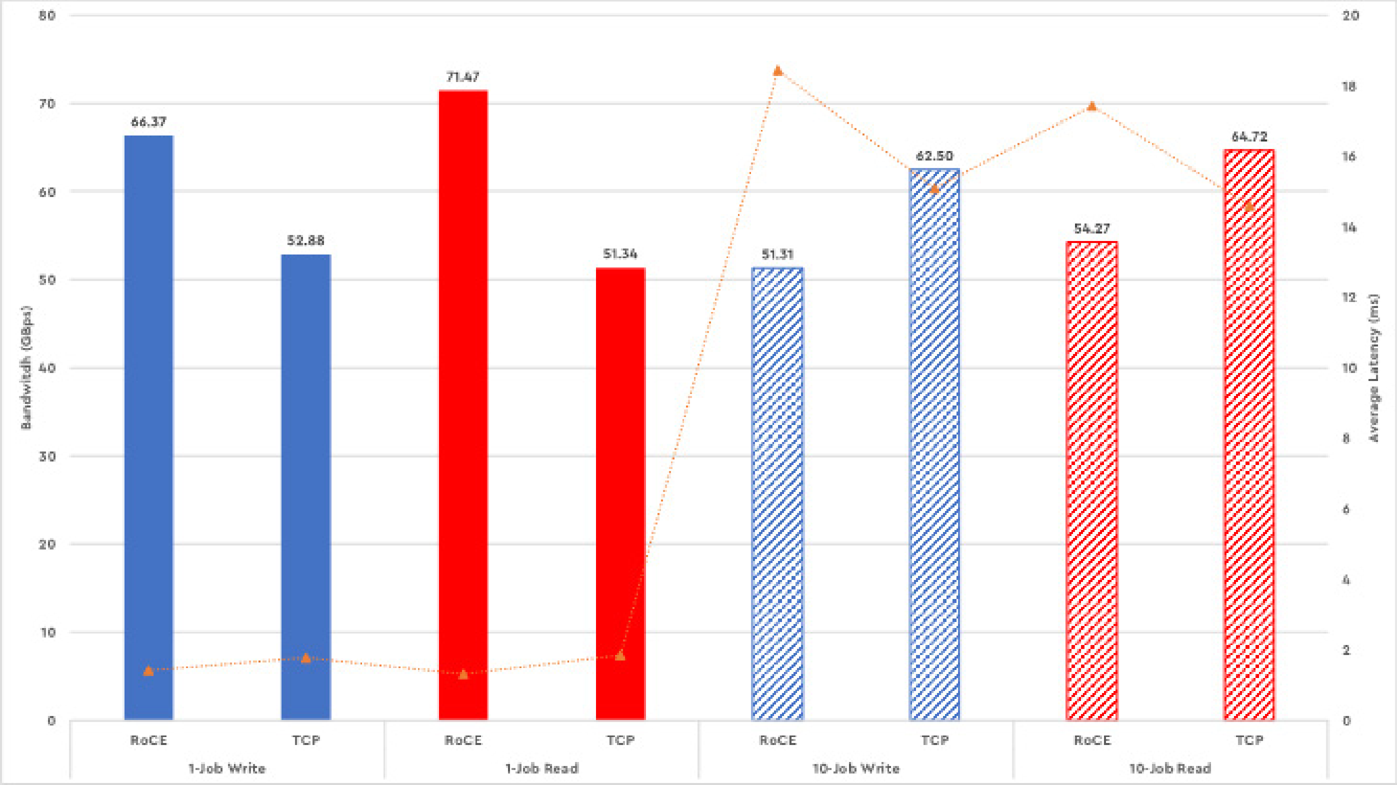
- RoCE Sequential Writes: 1.63 ms
- TCP Sequential Writes: 5.54 ms
- RoCE Sequential Reads: 2.28 ms
- TCP Sequential Reads: 5.08 ms

In both test scenarios, we can see that RoCE outperforms TCP with 55% lower Read latency and 71% lower Write latency.



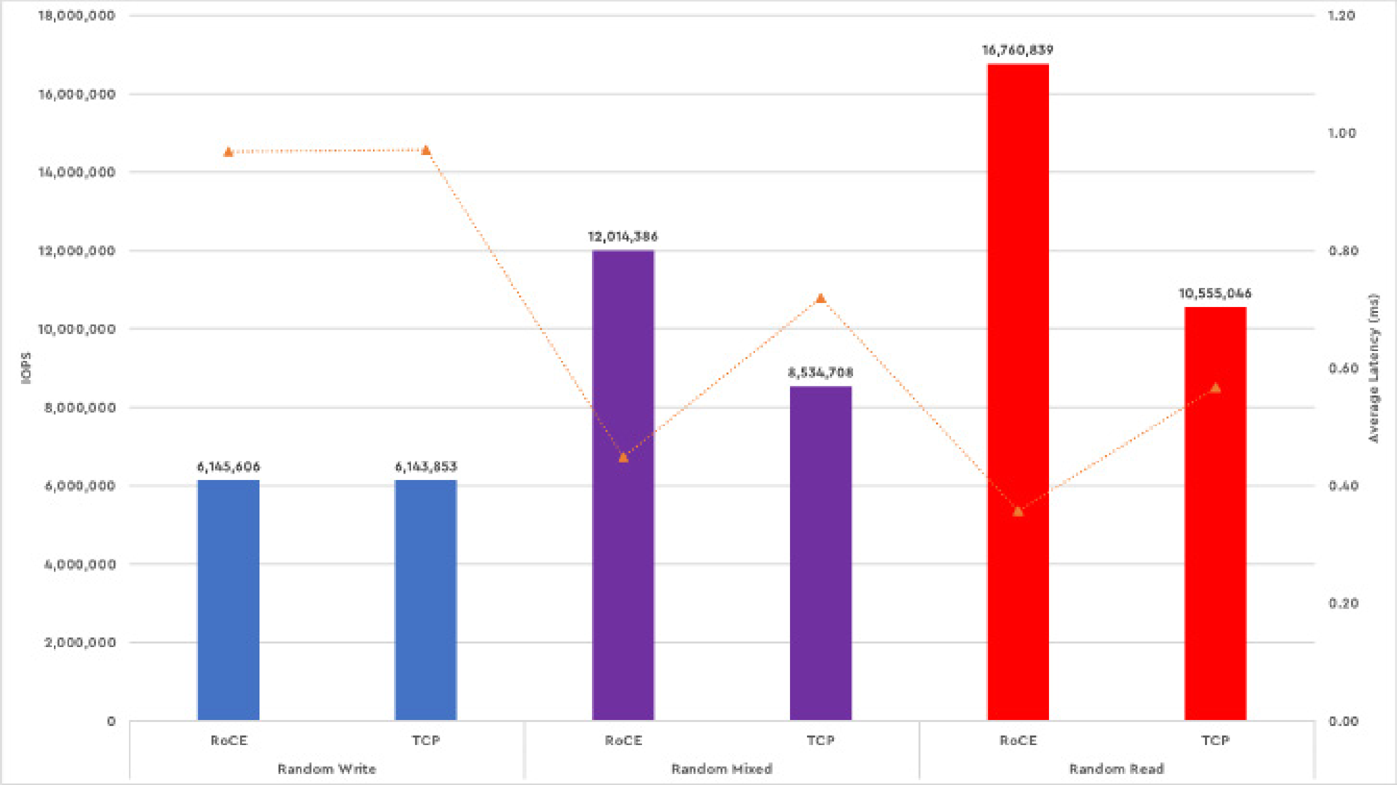
Benchmark Results: RoCE vs. TCP Sequential Bandwidth vs. Average Latency

The low job count and queue depth significantly favor RoCE and shows better latency. Nominal TCP performance is achieved by increasing the workload from one job to ten. This also impacts latency for both protocols. One test that TCP leads in both performance and latency is in ten job reads. Generally, RoCE has better bandwidth at a lower latency.



Benchmark Results: RoCE vs. TCP Random IOPS vs. Average Latency

Read IOPS performance for TCP is about two-thirds that of RoCE. Mixed IOPS sees a similar trend, mainly from the 70% read portion of the workload. Write IOPS are within 0.1% of each other, with RoCE and TCP performing similarly. Generally, RoCE has better IOPS at a lower latency.



The Contenders: NVMe/TCP vs. RDMA with ROCEv2 – Conclusion

There is not necessarily a singular criterion that defines one option over the other. Understanding acceptable latency is certainly key. If the solution is unable to perform to requirements, then it's the wrong solution.

The performance requirements are ultimately dictated by the use case, application. The lowest latency applications are those use cases where even the slightest delay can have significant consequences. Some examples would be:

1. Finance and High-Frequency Trading: In the world of finance, microseconds can make a massive difference. High-frequency trading relies on extremely low-latency applications to execute trades at the optimal moment.

2. Gaming: Online gaming, especially in competitive multiplayer settings, requires low latency to ensure that players' actions are reflected accurately and immediately in the game world.

3. Telecommunications: Services like voice and video calls, as well as applications involving real-time communication, demand low latency to maintain smooth interactions and prevent delays in conversations.

4. Virtual Reality (VR) and Augmented Reality (AR): VR and AR applications require low latency to provide users with a seamless and immersive experience where the virtual world responds to their actions instantly.

5. Content Delivery Networks (CDN): CDNs strive for low latency to deliver online content (such as videos and websites) quickly to users all around the world.

6. Internet of Things (IoT): IoT devices that need to interact with each other in real time, like smart grids or connected vehicles, rely on low latency to ensure smooth communication.

At this time, it would be fair to consider RDMA with RoCEv2 as the NVMe-oF performance gold standard. This superior performance can routinely be demonstrated, particularly when measuring latency. This in turn, lends RDMA with RoCEv2 to the above use cases.

Of course, no performance is free. RoCEv2 requires the deployment of a lossless Ethernet network. This might be moderate to difficult to achieve. There are the infrastructure upgrade costs to consider, interoperability challenges and potential skill set shortfalls to manage. Generally, the usage of NVMe/RoCE has been limited to smaller, high end deployments or large, sophisticated cloud service providers for which having the lowest latency is important.

NVMe/TCP, on the other hand, can be used at every scale, even outside the data center (such as for edge deployments) with few or no changes to the network configuration. The flexibility of TCP makes NVMe/TCP a compelling transport option. Users may see additional microseconds of latency but then, will that be relevant to those where slight delays in data processing or communication do not have critical or immediate consequences? Here are some examples:

1. E-commerce: While e-commerce platforms benefit from efficient data processing, the tolerance for latency is generally higher compared to industries like finance or gaming. Small delays in loading product pages, for instance, might not have a significant impact on user experience.
2. Social Media: While social media platforms aim for responsive user interfaces, minor delays in displaying posts, comments, or notifications are generally more acceptable compared to industries requiring real-time interactions.
3. Retail: Traditional brick-and-mortar retail businesses do not rely heavily on low-latency applications since customers can physically interact with products and staff.
4. Energy and Utilities: Many aspects of energy generation and distribution, like monitoring power plants or analyzing consumption trends, do not require instantaneous responses.
5. Supply Chain Management: While efficiency is important, supply chain processes often involve longer timeframes where minor communication delays are tolerable.
6. Data Analysis and Research: While timely analysis can be important, not all data analysis tasks require instant results. Researchers can often work with slight delays in processing and reporting.
7. Education: While remote learning platforms benefit from responsive interfaces, not all educational activities require real-time interactions.
8. Entertainment Streaming: Streaming platforms (music, video, etc.) can tolerate some buffering time as long as the overall experience remains smooth.
9. Non-Critical Communication Systems: Certain communication systems, like internal corporate communication or non-urgent email, can tolerate latency without significant issues.
10. Non-Critical IoT Applications: Some IoT applications, such as smart home devices for convenience (e.g., smart thermostats), can function well with moderate latency.

While it is still early days for NVMe/TCP adoption, also with its share of implementation and management challenges, its lower cost and relative ease of deployment will likely see it becoming the transport option of choice for broad adoption and deployment of NVMe-oF. In short, at this time, it's not the most performant, but it's likely to be good enough – particularly when compared to legacy options such as iSCSI.

Western Digital.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
www.westerndigital.com

September 2023 / Revision 01