White Paper

Disaggregated Storage
Application Device Queues

Scalable, Low-Latency Storage Using NVMe over TCP

Intel and Lightbits Labs helped make NVMe over TCP with Application Device Queues (ADQ) technology an open, industry standard to enable fast storage performance with easy implementation, efficiency, and scalability benefits of disaggregated storage

Executive Summary

Non-Volatile Memory Express over TCP (NVMe/TCP) is an industry storage transport standard developed by the NVM Express consortium that consists of a cross-section of industry players including Intel and Lightbits Labs. NVMe over Fabrics (NVMe-oF) enables disaggregated SSD storage to operate at efficiency levels previously possible only through direct-attached solutions. While prior technologies have made network fabric-based NVMe storage possible, they typically involve a limited ecosystem, specialized hardware, and extra complexity in deployment. NVMe/TCP, combined with the Intel® Ethernet 800 Series Network Adapter with Application Device Queues (ADQ), helps remedy these concerns. ADQ enhances NVMe/TCP by lowering latency while retaining ease of implementation, efficiency, and scalability benefits. When combined with Lightbits Labs' LightOS and Intel® 3D NAND SSDs, this approach provides a comprehensive and convenient NVMe/TCP-based storage solution. Intel Optane persistent memory and Optane SSDs are under evaluation as a way to further extend performance.

ADQ is an open technology designed by Intel that is based on enhancements to the Linux kernel. ADQ provides application traffic optimization to help increase application response time predictability and reduce congestion issues, thereby lowering latency and improving total throughput.

The key benefits of the NVMe/TCP with ADQ solution include:

- Ease of implementation from a Linux kernel-based approach
- Higher IOPS at lower latency than NVMe/TCP by itself
- Optimized storage utilization
- Excellent data reliability
- Easy storage scaling

These benefits can apply across a range of demanding use cases, such as databases, Apache Kafka streams processing, large-scale analytics, and public/private cloud services.

Increases Response Time Predictability[1]
Up To 30%

Reduces Latency[2]
Up To 50%

Improves Throughput[3]
Up To 70%

Table of Contents

Acronyms

ADQ	Application Device Queues
DAS	direct-attached storage
FC-NVMe	NVME over Fibre Channel
NVMe	Non-Volatile Memory Express
NVMe-oF	NVMe over Fabrics
NVMe/TCP	NVMe over TCP
PCIe	PCI Express
RDMA	remote direct memory access
RoCE	RDMA over Converged Ethernet
TCP	Transmission Control Protocol

The Storage Challenge

The development of processors with high core counts and multi-threaded software has enabled significant advances in highly parallelized workload execution requiring storage with high IOPS and low latency. Legacy storage systems have failed to keep pace with today's demanding data center use cases (see Figure 1). Simply adding flash, even NMVe-based flash, to traditional architectures does not alleviate the storage bottlenecks experienced with modern, scale out, and cloud native data center applications.

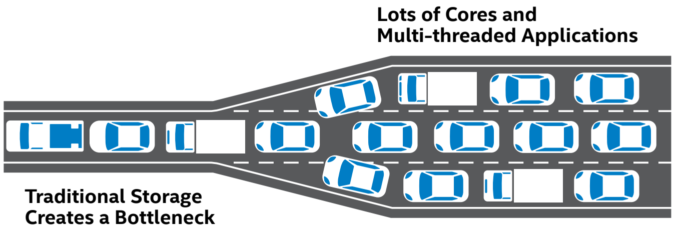


Figure 1. Storage performance has remained challenging as businesses support high-core-count processors and multi-threaded applications.

Thus, developers have moved to flash-based DAS—local NVMe flash that has high bandwidth and low latency. The trade-off is poor utilization, data silos, and increased copy traffic between application servers. Centralized, shared flash is desirable, but it must perform like local flash including NVMe's low-latency characteristics and can't require an alternate fabric; Ethernet offers an attractive basis for a solution.

Keeping up with Modern Data Center Demands

The DAS model (see Figure 2) that deploys NMVe-based drives locally within servers is commonly used today. Because NVMe works over PCIe, DAS is an expedient and high-performance approach to implementing NVMe. However, the DAS model suffers from several drawbacks:

- **Underutilization.** DAS creates islands of storage that are only accessible by the local host. This results in poor overall utilization because not every host uses all, or sometimes any, of its DAS. Some servers may labor at maximum utilization, and others may be grossly underutilized.
- **Lack of scalability.** DAS is difficult to scale cost-effectively. Silos inherently resist scaling, due in part to bandwidth and saturation issues over port, rack-level, and network link resources. Silos of data are difficult to manage because every server and its DAS must be individually managed. This results in data being copied over the network again and again if sharing is needed. DAS scaling limitations illustrate why the market needs infrastructures based on disaggregated models in which compute and storage resources can scale independently. Hyperconverged models fail to address this, because as more storage capacity is added, CPU resources must expand with it. But more compute may not be needed if other CPUs in the cluster aren't being taxed.
- **Inefficiency in Data Protection.** DAS is not inherently redundant. Data protection must be administered on a per-host basis and may consume additional CPU resources. Adding reliability to DAS often involves replicating data across the network. Triple replication is common; it triples the consumption of network bandwidth, required traffic computation, and required storage capacity.

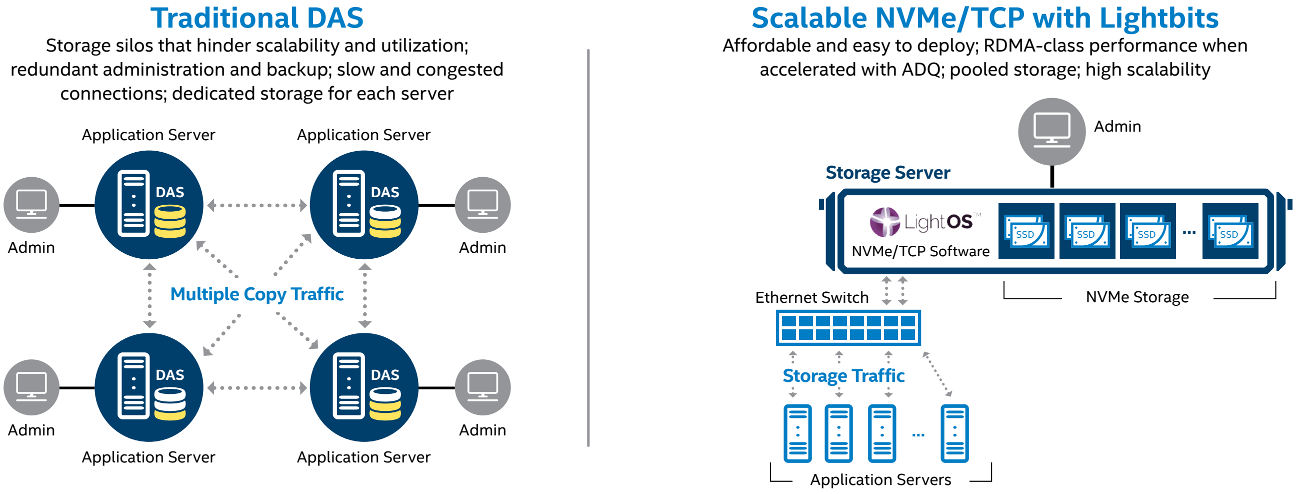

Traditional DAS
Storage silos that hinder scalability and utilization;
redundant administration and backup; slow and congested
connections; dedicated storage for each server
Application Server
Application Server
DAS
DAS
Admin
Admin
Multiple Copy Traffic
DAS
DAS
Admin
Admin
Application Server
Application Server
Scalable NVMe/TCP with Lightbits
Affordable and easy to deploy; RDMA-class performance when
accelerated with ADQ; pooled storage; high scalability
Admin
Storage Server
LightOS
NVMe/TCP Software
SSD
SSD
SSD
...
SSD
NVMe Storage
Ethernet Switch
Storage Traffic
...
Application Servers

Figure 2. The DAS model (left) illustrates the drawbacks of conventional data center storage. In contrast, NVMe/TCP with ADQ technology and Lightbits LightOS allow for easy to deploy, modular, scalable, centrally managed storage over Ethernet connections.

NVMe-oF Solves Some Issues but Raises Others

To address some of the limitations of the DAS model, the industry needed a way to deliver disaggregated storage using NVMe across networks efficiently. In 2014, the NVM Express standards body started work on the NVMe over Fabrics (NVMe-oF) specification, which was published in 2016 and utilized a variety of industry-standard storage transport protocols (see Figure 3). The goal of NVMe-oF was to provide remote connectivity to NVMe devices while incurring no more than an additional 10 microseconds (µs) of latency compared to the same NVMe devices accessed in a local server.

The first two defined standards to achieve this goal were:

1. NVMe-oF using remote direct memory access (RDMA) technologies, including InfiniBand, Intel® Omni-Path Architecture (Intel® OPA), iWARP, and RDMA over Converged Ethernet (RoCE)
2. NVMe over Fibre Channel (FC-NVMe)

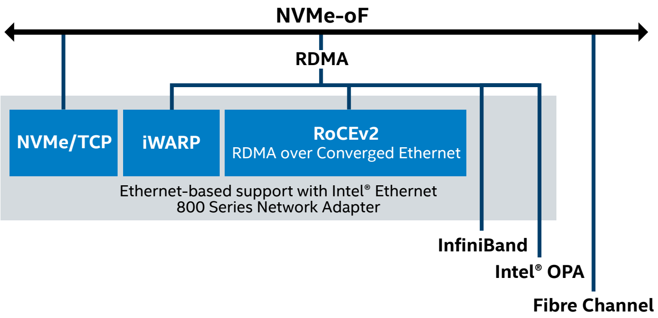

NVMe-oF
RDMA
NVMe/TCP
iWARP
RoCEv2
RDMA over Converged Ethernet
Ethernet-based support with Intel® Ethernet 800 Series Network Adapter
InfiniBand
Intel® OPA
Fibre Channel

Figure 3. The NVME-oF framework has gradually added transport protocols and makes provisions for the next-generation protocols.

These approaches are worthy advances; however, they have some disadvantages. RoCE and iWARP are open, published standards for RDMA over Ethernet, with high-performance capability. However, both technologies require dedicated expertise to implement and may require specialized hardware. RoCE requires the use of a lossless Ethernet network that requires special Ethernet switch settings. As a result, RoCE deployments are difficult to scale and often confined to a single rack. FC-NVMe offers greater scaling ability but can increase infrastructure cost.

NVMe/TCP Is Easy to Implement and Scale, but Increases Latency

Meanwhile, development began on NVMe/TCP, which would also operate over Ethernet. The specification was released in November 2018. NVMe/TCP uses the standard TCP stack running on the Linux kernel within the host CPU, is easy to deploy, and is supported by a wide ecosystem. As a result, it's an NVMe-oF solution that is easily implemented and scalable. The technology is compatible with most Ethernet adapters and requires no special switch settings. Because it runs on the Linux stack and does not bypass the host CPU, as RDMA protocols do, latency remains a drawback along with increased host CPU utilization. Hence, NVMe/TCP exhibits higher latency relative to RDMA options such as RoCE and iWARP for network storage technology.

Enter ADQ Technology

While working on NVMe/TCP, Intel was also developing Application Device Queues (ADQ) technology. It became clear that, while ADQ can be applied across a range of use cases, NVMe/TCP would make an excellent initial application of ADQ.

ADQ is an open technology analogous to express lanes on a freeway. Surface streets are fraught with unpredictable delays like streetlights and construction. Freeways could support bumper-to-bumper traffic at high speeds if all traffic movement was absolutely consistent and predictable, which, of course, rarely happens, especially during rush hour. But dedicated express lanes on the freeway that allow specific types of traffic to travel from point A to point B can provide a fast, predictable commute.

Similarly, without dedicated queues, networking throughput and latency are unpredictable under heavy traffic conditions and always changing. Some operations will be slower than others. Because the slowest operations (known as tail latency) determine overall application response time, too many slow operations scaled across many systems can choke application performance. On a single system, tail latency is typically not a problem. However, across a broadly distributed platform, as with a cloud service, excessive tail latency can make it more difficult to fulfill customer service-level agreements (SLAs).

As implemented in the Intel® Ethernet 800 Series, ADQ establishes up to 2,048 lanes, or queues, for network traffic. To improve high-priority application performance, multiple queue pairs (transmit and receive) can be dedicated to certain types of high-priority application traffic. For example, one hundred twenty-eight queue pairs (1 pair per processor core) might be reserved only for a specific database application's traffic. Unlike general-purpose lanes, these reserved lanes feature one type of packet zipping along a stream to a single destination. No large trucks cutting in, no unexpected lane changes. Admins assign as many ADQ lanes to an application as are needed for an application's bandwidth needs while reserving a handful that may be needed for general-purpose traffic.

The net result is a dramatic reduction of the application's response time latency, higher overall storage data throughput, and most importantly, increased application response time predictability through a reduction in the tail latency, enabling greater consistency in meeting customer SLAs.

ADQ helps enable NVMe/TCP to achieve distributed storage performance results in the same range as RDMA-based protocols. Working together, Intel and Lightbits Labs deliver NVMe/TCP with ADQ for distributed storage by combining Intel Ethernet 800 Series Network Adapters with Lightbits Labs LightOS.

By applying ADQ to the standard, open-source NVMe/TCP block driver, any application using an NVMe/TCP block device, such as those noted in Figure 4, can obtain ADQ's storage benefits. Some applications may require modification to work directly with ADQ. Others, such as open-source Redis, may not require any changes. For more information on which applications have been enabled to support ADQ, visit the ADQ Resource Center.

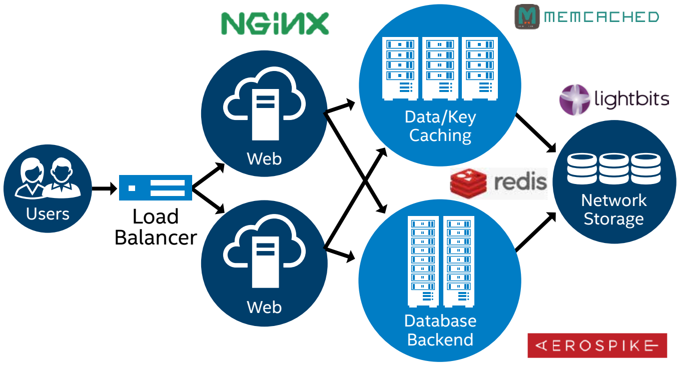


Figure 4. NVMe/TCP with ADQ technology can help accelerate storage performance across a range of initial data center applications, with more to come.

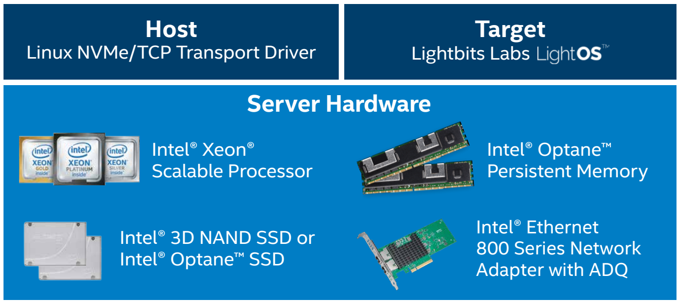

Host
Linux NVMe/TCP Transport Driver
Target
Lightbits Labs LightOS™
Server Hardware
Intel® Xeon® Scalable Processor
Intel® Optane™ Persistent Memory
Intel® 3D NAND SSD or Intel® Optane™ SSD
Intel® Ethernet 800 Series Network Adapter with ADQ

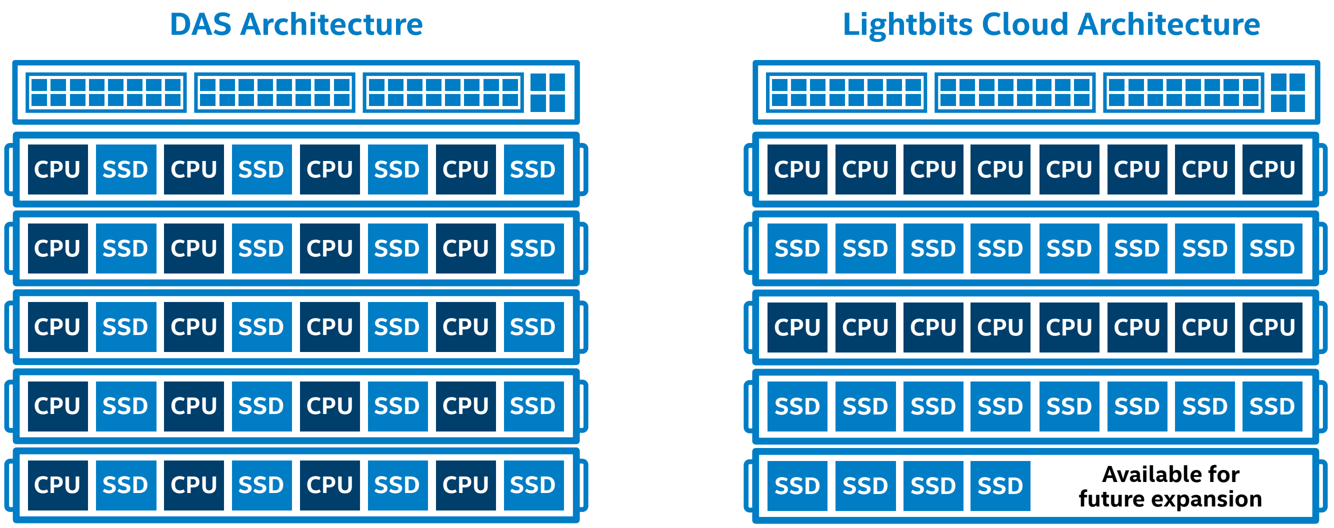
Figure 5. Technologies from Lightbits Labs and Intel can create a performant and highly scalable NVMe/TCP platform.

Solution Components and Architecture

In 2020, Intel and Lightbits Labs teamed up to deliver an ADQ-accelerated NVMe/TCP solution to the community. The primary components are the Lightbits Labs NVMe/TCP target platform, the Intel Ethernet 800 Series Network Adapters with ADQ, 3D NAND SSDs, and a compatible Linux kernel. Additionally, Intel Optane persistent memory and Optane SSDs are under evaluation as a way to further extend performance (Figure 5). The following sections detail the various solution components.

Lightbits Labs LightOS

Open-source software exists for implementing NVMe/TCP, but providers such as Lightbits offer refined and differentiated applications for creating and managing scalable NVMe/TCP-based disaggregated storage solutions (Figure 6). LightOS is an NVMe/TCP target solution that manages and virtualizes pools of NVMe drives. Management tasks include providing logical volumes, administering different QoS levels across different SSD pools, and distributing I/O loads intelligently across those pools for maximum efficiency. LightOS can avoid sending writes to drives engaged in garbage


DAS Architecture

Lightbits Cloud Architecture
CPU SSD CPU SSD CPU SSD CPU SSD
CPU SSD CPU SSD CPU SSD CPU SSD
CPU SSD CPU SSD CPU SSD CPU SSD
CPU SSD CPU SSD CPU SSD CPU SSD
CPU SSD CPU SSD CPU SSD CPU SSD
CPU CPU CPU CPU CPU CPU CPU CPU
SSD SSD SSD SSD SSD SSD SSD SSD
CPU CPU CPU CPU CPU CPU CPU CPU
SSD SSD SSD SSD SSD SSD SSD SSD
SSD SSD SSD SSD
Available for future expansion

Figure 6. The "disaggregation" of storage through an NVMe/TCP and LightOS solution leads to a more efficient, scalable architecture in which compute and storage resources serve as flexible, modular blocks rather than inherently bound-together resources.

collection to maintain performance consistency, reduce tail latency, and extend the lifecycle of storage assets.[4] LightOS also addresses target server failover, data reduction, thin provisioning, erasure coding protection, and other features while maintaining 100Gb/s performance levels.

Linux NVMe/TCP Transport Driver

Intel and Lightbits Labs collaborated on this solution's TCP block driver and Linux kernel patches. In November 2018, NVMe ratified the NVMe/TCP transport standard.[5] However, adoption of the standard leaped forward a few months later with the integration of NVMe/TCP transport drivers in Linux kernel 5.0.[6] (For those using distributions that have yet to adopt the new kernel, Lightbits provides downloadable back-ported side drivers on its website.[7]) NVMe/TCP is now accepted as an industry standard and is freely available to everyone.

Intel® Ethernet 800 Series Network Adapter with ADQ[8]

Intel is first to market with support for the full range of Ethernet-based NVMe-oF protocols in a single network adapter: iWARP RDMA, RoCEv2 RDMA, and NVMe/TCP with ADQ acceleration. The Intel Ethernet 800 Series Network Adapter supports port speeds from 1 to 100GbE. This high performance and versatility pairs well with the high bandwidth rates NVMe can achieve. In fact, multiple NVMe devices can run simultaneously through a single 25GbE connection, and NVMe/TCP with ADQ allows them all to maintain low latencies. Additionally, the Intel Ethernet 800 Series includes Intel's first network adapters that support ADQ. Intel has open-sourced ADQ by upstreaming ADQ-related enhancements (known within the community as "patches") to the Linux kernel to encourage broad adoption throughout the industry.

Lightbits performance testing using ADQ revealed an average mean latency across six queue depth measurements of 215 µs without ADQ compared to 146 µs with ADQ, for an average difference of 69 µs (see the Performance Testing section).

Intel® Optane™ Technology

Intel Optane persistent memory and SSDs can provide a significant performance boost in certain software-defined, disaggregated storage solutions using NVMe/TCP. This is especially true for low-latency workloads when NVMe/TCP is bolstered with ADQ. Transactional workloads that demand low latency inherently tie well into the strengths of Intel Optane persistent memory and SSDs.

Intel Optane persistent memory (PMem) is an innovative technology that delivers a unique combination of large-capacity memory with data persistence. Intel Optane PMem significantly reduces the latency to data access because it does not require the same file system, device driver, and bus protocols needed to access disk-based storage.

To illustrate, the average read latency of a NAND SSD is 80 µs,[9] which reduces to 10 µs with Intel Optane SSD, and drops to between 100 and 340 nanoseconds with Intel Optane persistent memory.

Intel Optane technology can provide extra large persistent data structures closer to the processor, minimizing the wait time for data and speeding application execution. Applications that entail low-latency handling of high-capacity workloads, such as database transaction logs or file system metadata operations may benefit from Intel Optane persistent memory or SSDs in NVMe/TCP deployments.

Potential Use Cases

Until recently, most NVMe/TCP work has focused on development, integration with the surrounding storage ecosystem, and laying the groundwork for widespread adoption. Early performance results remain encouraging, and even now we see how NVMe/TCP with ADQ acceleration will serve particularly well in certain markets and use cases. Here are a few examples.

Databases

Mapping the performance demands of different databases on a spectrum is common. Many traditional databases do not have the ultra-low-latency needs that demand in-memory architectures, but others do. However, even in-memory databases will likely run into misses when the application needs to seek data outside of memory and in persistent storage; low-latency storage will remain a priority and is the reason why using NVMe-based storage is increasingly considered a database best practice.

Similarly, start-up and shut-down operations require databases to read or write large amounts of data from memory to storage. High storage bandwidth is needed for such processes. Fortunately, NVMe/TCP does well with both low latency and high bandwidth, and many databases benefit from both qualities, including Aerospike, Cassandra, CouchDB, MongoDB, MySQL, Redis, and PostgreSQL.

Lightbits has shown that Cassandra can achieve even better performance with NVMe/TCP than from a local NVMe drive.[10] In part, this benefit stems from the distribution of storage loads across multiple drives. Workload distribution performs much like RAID striping, which creates less impact than placing the entire workload on a single DAS drive. Workload distribution avoids overhead operations such as garbage collection.

NVMe/TCP with ADQ acceleration may be a particularly good fit for cloud-native databases, most of which are highly scalable and perform their own data protection but can still benefit from low-latency, centralized storage.

Apache Kafka

Kafka is an open, low-latency, high-throughput bus messaging system that can assist with processing real-time data feeds for rapid decision making. For example, if trending news creates a sudden spike of interest in a certain type of product, an infrastructure around Kafka can factor these queries into customer product searches and influence the priority of search returns. Kafka also excels in helping detect anomalous behavior, which can be applied in fraud-detection solutions. Previously, such analysis was inherently rear-looking, relying on the compilation of multiple databases and searching for historical pattern anomalies. Kafka, combined with analytic tools and low-latency storage, makes real-time analysis possible.

Lightbits Labs tested Kafka in an environment that included NVMe/TCP and LightOS (see Figure 7). Test results showed that I/O can achieve the same performance level as local NVMe SSDs.[11] Kafka storage achieves high utilization, improved service levels and security, and fast rebuild times with high resiliency. Notably, implementing the solution required no changes to the network infrastructure or application servers.

ADQ has the potential to benefit Kafka implementations. Where the Kafka brokers are shown, there are also Kafka streamers, consumers, and producers. Kafka brokers receive read/write requests from the network and then read/write those to back-end NVMe/TCP. Having a separate set of ADQs and traffic controls for network traffic should benefit performance in a typical multi-application setup.

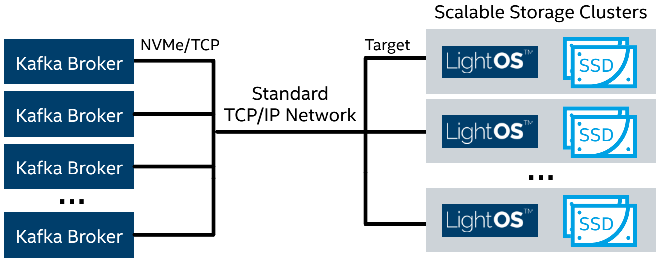


Figure 7. LightOS allows the local NVMe SSDs that would typically reside on Kafka brokers to be replaced by logical volumes over NVMe/TCP. These logical volumes can increase flexibility and avoid many typical DAS drawbacks.

Analytics

Kafka integrates well into analytics engines like Apache Spark, which can be applied to fraud detection, mass-scale patient record analysis, genetic sequencing data, and many more analytics use cases. Apache Spark is an in-memory database, but genomics data can be massive. The accuracy of genomics analytics improves with larger datasets, so even in-memory databases will achieve more accurate results from spilling data into storage. For example, consider searching for a link between disease susceptibility and a certain genetic mutation. The patient dataset spans 200 genetic sequences, but only 20 sequences can fit in one server's memory. One approach would be for a data scientist to manually divide the dataset into 20-patient blocks. Unfortunately, this approach could decrease analysis accuracy and increase project time and complexity. A better approach is to use NVMe/TCP with ADQ against a large NVMe pool to allow the entire 200-patient dataset to be run at local performance levels with no change to the analytics infrastructure. Intel Optane SSDs can help in this situation, as their low latency can affordably make the media a natural NVMe extension of system memory.

Private Cloud Services

Enterprises across all industries are increasingly deploying internal cloud services. Some large firms might build their own solutions from scratch, but most will reach for off-the-shelf cloud solutions and then tweak them to their needs. In effect, even small enterprises become cloud service providers. In such environments, NVMe/TCP with ADQ provides more flexibility. For instance, some companies opt for bare-metal implementations to avoid the complexity and overhead of virtualization. These companies can avoid deploying large amounts of SSD storage across every system. Also, as noted earlier, NVMe/TCP helps alleviate any pressure to overprovision, because adding more high-performance storage becomes a modular, drop-in issue. NVMe/TCP with ADQ helps eliminate the need to predict how much storage might be required, while also assuring that scaling won't entail sacrificing DAS-class performance.

Performance Testing

With guidance from Intel, Lightbits Labs researched how much benefit could be derived from the NVMe/TCP with ADQ solution. In particular, Lightbits examined IOPS performance and latency (both mean and tail) in single-connection and multi-connection contexts across a range of queue depths.

Test Setup and Results

Lightbits conducted testing to ascertain the potential benefit of using NVMe/TCP with ADQ with a Lightbits storage target. They tested NVMe/TCP using the Intel Ethernet 800 Series Network Adapter with and without ADQ on the host for a Lightbits cluster target as represented in Figure 8. Full configuration information is available in the Appendix at the end of this paper.

The test setup contains three switches to represent a three-hop leaf/spine/leaf network configuration. ADQ is enabled and disabled only on the host; ADQ is not implemented on the target servers. The LightOS cluster contains up to three nodes. Measurements were taken for IOPS and latency for a single connection and multiple (three) connections. Workloads were generated with the Flexible IO (FIO) open-source synthetic benchmark tool.

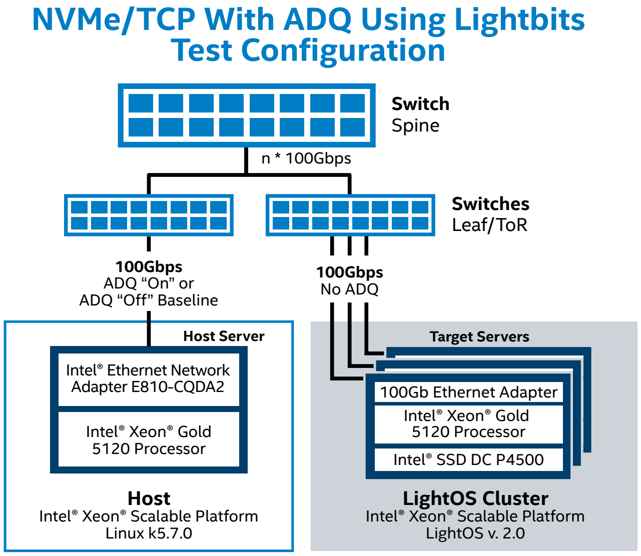

NVMe/TCP With ADQ Using Lightbits Test Configuration
Switch
Spine
n * 100Gbps
Switches
Leaf/ToR
100Gbps
ADQ "On" or
ADQ "Off" Baseline
100Gbps
No ADQ
Host Server
Intel® Ethernet Network Adapter E810-CQDA2
Intel® Xeon® Gold 5120 Processor
Host
Intel® Xeon® Scalable Platform
Linux k5.7.0
Target Servers
100Gb Ethernet Adapter
Intel® Xeon® Gold 5120 Processor
Intel® SSD DC P4500
LightOS Cluster
Intel® Xeon® Scalable Platform
LightOS v. 2.0

Figure 8. This diagram illustrates the topology Lightbits used in performing its NVMe/TCP with ADQ cluster testing. Note the use of the Intel® Ethernet Network Adapter E810 with and without ADQ enabled on the host server.

Throughput Improves up to 70 Percent at Higher Queue Depths[12]

Figure 9 shows IOPS against queue depth (QD), with single-connection results in the left chart and multi-connection results on the right, as well as with ADQ enabled (dark blue) and disabled (orange) on the host in both cases. We can make several observations here.

At very low queue depths, ADQ offers little advantage. By QD4, though, ADQ's assistance becomes more obvious, and this advantage increases as queue depth increases.

Throughput scales as queue depth increases. However, the scaling is not linear. This is to be expected. A doubling of queue depth does not yield a doubling of IOPS, and going from one thread to 32 does not yield a 32x improvement. The question is how much improvement sustains as resources increase.

In single-thread applications, ADQ delivers the highest proportional advantage at high queue depths. For example, at QD4, the data shows ADQ and non-ADQ results of 31,000 and 21,000 IOPS, respectively, for an improvement of 48 percent. At QD32, though, the numbers scale to 141,000 and 81,300 IOPS, an increase of 73 percent when ADQ is enabled.

Now, examine the same comparison with 32 threads. The results show 1,111,000 IOPS at QD4 with ADQ and 840,000 IOPS without ADQ—a 32 percent improvement. At QD32, the results were 2,884,000 (with ADQ) and 2,176,000 (without), a 32 percent improvement. The NVMe/TCP with ADQ solution can be expected to deliver the highest throughput efficiency at lower core counts with higher queue depths. However, modern workloads require significant throughput, and even the 141,000 IOPS realized by one connection at QD32 can hardly compare to the nearly 3 million IOPS realized at QD32 in the multi-thread configuration.

The key take-away is that ADQ continues to scale throughput with both connection and queue depths.

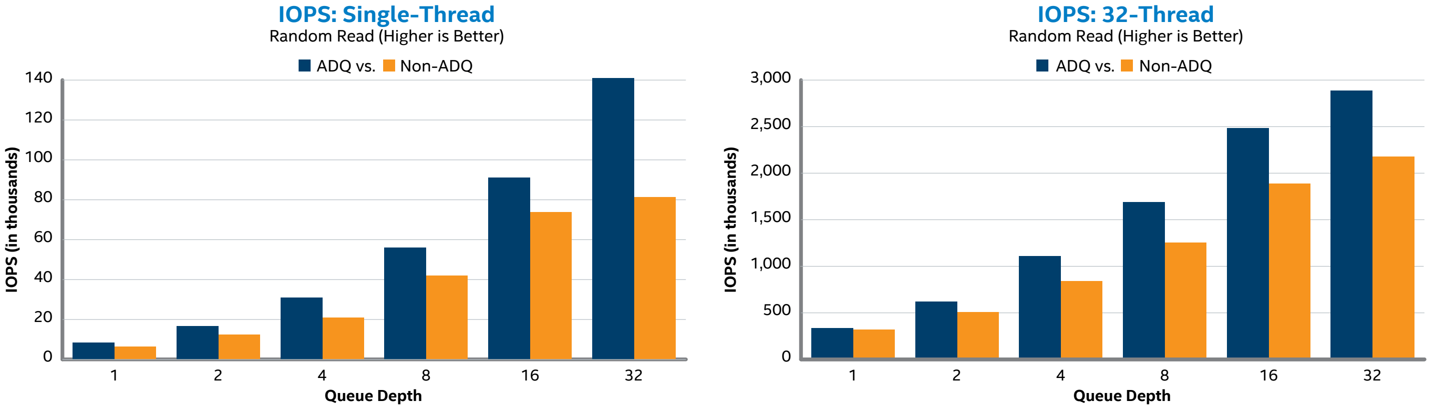


Figure 9. Especially at high queue depths, ADQ provides a significant improvement in throughput.

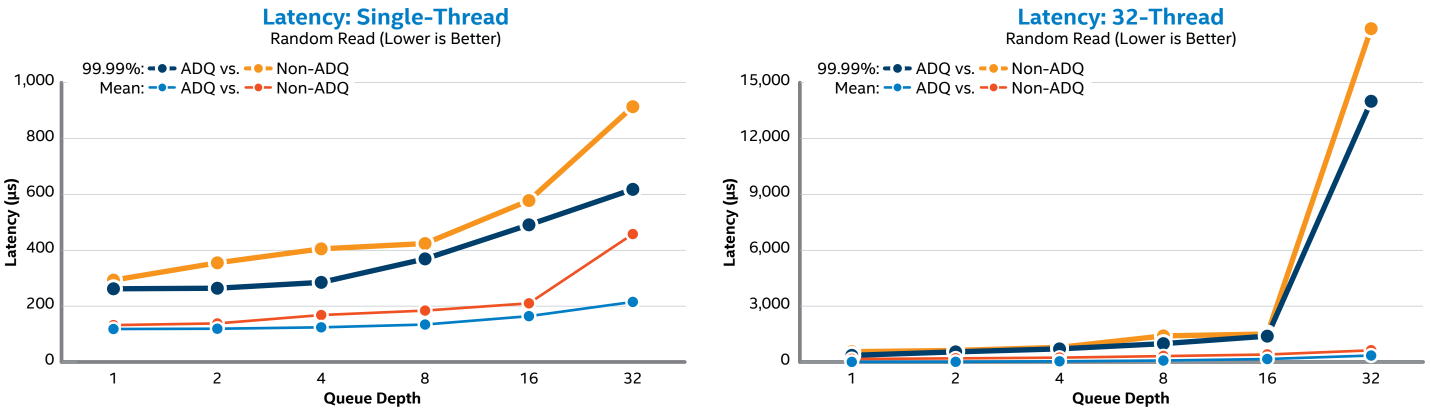


Figure 10. ADQ technology improves latency across all queue depths for both single- and multi-thread workloads.

Latency and Predictability

The performance results for the P50 mean latency (red and light blue lines) and P99.99 tail latency (orange and dark blue lines) are shown in Figure 10. A 99.99 percent latency refers to those results at the end of the latency measurements—the slowest 0.01 percent, or "tail latency." P99.99 can be used as a proxy for the predictability of application response times, as this marks the lowest/slowest boundary for sustained application performance. Some applications will want to focus on the P50 mean latency while others, like those bound to maintaining minimum SLAs, will spotlight the P99.99 tail latency.

In every case, ADQ improves both the average and the tail latency. This applies across both single- and multi-thread tests, but the difference between with and without ADQ is more pronounced in single-thread testing.

As expected, latency increases with queue depth. Without ADQ, single-thread latency takes a heavy toll in moving from QD16 to QD32. This was not observed in multi-thread results, which show a much narrower gap between ADQ and non-ADQ results across all queue depths.

Multi-thread mean latency scales in a nearly linear fashion while multi-threaded tail latency takes a sharp upward turn in going from QD16 to QD32. This upward to turn is due to somewhere between QD16 and doubling to the queue depth to QD32 the throughput for ADQ hits wire rate of ~2.8M IOPS (see figure 9, 32-thread). Increasing queue depth after achieving wire rate adds latency.

QD reflects the average number of in-flight I/O requests between an initiator and a target. A higher QD means that the host can optimize the order of requests sent to a storage system, with those requests often being sent in parallel. However, this optimization process takes time, which contributes to latency.

As long as latency levels don't impede an application's SLA, a certain amount of latency can be tolerated, but solution engineers and administrators need to monitor this trade-off.

QD can also serve to throttle traffic so targets are not overwhelmed, which would cause queueing and degrade performance. This, combined with the latency issue just described, is why we don't max out queue depths from the outset. It's a workload-specific balancing act. Queue depth throttling can be especially beneficial when working with multiple initiators, as in a storage cluster.

Noted in Dell's Oracle Database Best Practices paper,[13] a default value of QD32 is often sufficient for most Oracle applications, although "there are specific use cases where changing the queue depth may improve performance." An example is when "a storage array is connected to a few Linux servers with large-block, sequential-read application workloads."

Similarly, when optimizing solutions such as ESXi/ESX hosts, VMware's knowledgebase states that increasing QD can help with large-scale workloads with intensive I/O patterns.[14]

The *predictability* of data processing can also be important to data centers. If given conditions call for workloads to run on certain systems at a certain QD, then administrators want to be able to predict how those workloads will behave. In addition to the predictability of the application's response time, as measured by the P99.99 latency, nearly linear scaling is another type of predictability. As such, we see that the straighter plots of I/O latency across QD of both threading types indicate a higher level of QD scaling predictability when ADQ is used.

Key Take-Aways

Predictability Increased up to 30 Percent;[15] Average Latency Reduced up to 50 Percent[16]

The broadest tail latency split seen in this NVMe/TCP with ADQ examination was with the single-thread configuration at QD4, when a non-ADQ P99.99 latency of 405 µs and an ADQ P99.99 latency of 285 µs was measured, for a 30 percent acceleration. The broadest mean latency difference observed was with the single-thread configuration at QD32, when a non-ADQ mean latency of 458 µs and an ADQ mean latency of 215 µs was measured, for a 53 percent reduction. Naturally, other configurations with other workloads will realize different results, some of which may favor ADQ even more.

These results tend to show that throughput scaled higher with increasing QD, and also that ADQ helped minimize latency penalties as that throughput increased. Ultimately, a significant discovery is that ADQ improves tail latency-based predictability at higher queue depths, which can help accelerate a wide range of multi-threaded data center applications and provide much higher performance reliability at scale.

Also note that, taking all tested queue depths into account, results showed an average 30 percent IOPS performance gain from using ADQ in a multi-threaded environment. Again, the implications for improving data center application performance are compelling.

Note that increased storage traffic carries a processing cost, because CPU utilization increases as traffic scales and also because TCP (unlike RDMA) does not offload the transport protocol in hardware.[17]

Conclusion

Today, data centers need to contain the costs of storage in the face of exploding datasets and demands for real-time responsiveness. NVMe-oF RDMA technologies have helped advance the industry for highly demanding cases. While RDMA-level performance can be excellent, NVMe/TCP offers the potential of broad adoption because of its ease-of-implementation and scalability. When paired with ADQ, NVMe/TCP can provide low-latency performance similar to RoCE and iWARP.

Now with extensive, standardized Linux support, NVMe/TCP can be further utilized with refined solutions, such as Lightbits Labs' LightOS. And it can be pushed to even higher performance levels with Intel Optane technology. Distributed storage can now realize comparable performance levels to DAS, but with significantly improved efficiency and scalability.

Learn More

If you liked this paper, you may also be interested in these related items:

- Intel Optane Technology
- NVM Express Group's NVMe over Fabrics paper
- NVMe/TCP Specification Announcement, November 2018
- Faster, More Predictable Ethernet with the Intel Ethernet 800 Series with ADQ Technology Brief
- ADQ Resource Center
- Lightbits Labs

Find the solution that is right for your organization. Visit **intel.com/ADQ** or contact your Intel representative.

Appendix: Test Configuration

Table A1. NVMe with ADQ Using LightOS: Details of the Test Equipment and Configuration

Note: Further performance improvement may be possible by adding to or replacing NAND storage with Intel® Optane™ SSDs.

	System Under Test	LightOS Cluster
Test By	Lightbits Labs	Lightbits Labs
Test Date	July 15, 2020	July 15, 2020
Platform	Supermicro SYS-2029U-TN24R4T	Supermicro SYS-2028U-TN24R4T+
# Nodes	1	1 to 3
# Sockets	2	1
CPU	Intel® Xeon® Gold 5120 Processor @ 2.2 GHz	Intel® Xeon® Processor E5-2648L v4 @ 1.8 GHz
Cores/Socket, Threads/Socket	14 cores/socket, 28 threads/socket	14 cores/socket, 28 threads/socket
Microcode	0x2000065	0xb000036
Hyper-Threading	On	On
Turbo	On	On
BIOS Version	3.2	American Megatrends Inc. (3.1c)
System DDR Mem Config: slots/cap/run-speed	16 slots/16 GB/2133 MT/s DDR4	16 slots/16 GB/2133 MT/s DDR4
Total Memory/Node (DDR+DCPMM)	256 GB	256 GB
Storage - Boot	128 GB SATADOM-SL 3ME3	128 GB SATADOM-SL 3ME3
Storage - Application Drives	N/A	8x Intel® SSD DC P4500
Network Adapter	1x Intel® Ethernet Network Adapter E810-CQDA2 @ 100Gbps	**Single-Node:** 1x Intel® Ethernet Network Adapter E810-CQDA2 @ 100Gbps **Multi-Node:** Add 2 x Mellanox ConnectX-4 EN Ethernet Adapter @ 100Gbps
PCH	N/A	N/A
Other Hardware (Accelerator)	N/A	N/A
OS	CentOS 7.7	LightOS version 2.0 (CentOS 7.7)
Kernel	5.7.0+.x86_64	4.14.189_00172587149ee079f0f16_rel_lb-7.x86_64
Workload and version	FIO 3.20	N/A
NVME/TCP with ADQ Patch	Pull request until put into the main branch: All upstream	N/A
Compiler	N/A	N/A
Network Adapter Driver	1.0.4-1.x86_64, firmware version: 1.40 0x80003ab8 1.2735.0, iproute-4.11.0-25.el7_7.2.x86_64	1.0.4-1.x86_64, firmware version: 1.40 0x80003ab8 1.2735.0, iproute-4.11.0-25.el7_7.2.x86_64
NVMe/TCP Settings	MTU set to 1500 Connected to targets with 32 polling queues	MTU set to 1500
LightOS Settings	N/A	Default
Network Switches	**Host Leaf:** Accton 7712-32X/AOS **Spine:** Mellanox MSN2700-CS2F	**Cluster Leaf:** Accton 7712-32X/AOS **Spine:** Mellanox MSN2700-CS2F
SSD Pool	N/A	8x Intel® SSD DC P4500 1 TB (2.5" U.2)

Table A2. System Under Test Network Adapter Settings

	ADQ "Off" Baseline	ADQ "On"
System Settings		
Interrupt Moderation	adaptive-rx	rx_usecs=0 tx_usecs=50
IRA Balance	Off	Off
Interrupt Affinitization	Linear	Linear
ADQ Settings		
Epoll Busy Poll	N/A	N/A
Socket Option for NAPI ID	N/A	N/A
TC-Mqprio Hardware Offload and Shaper	None	On
TC- Cloud Filter Enabling with TC-flower	None	On
Symmetric Queueing	Off	On

Endnotes

[1] Up to 30% predictability increase as measured by P99.99% improvement for ADQ "On" vs. ADQ "Off" Baseline. Source: Lightbits Labs testing conducted July, 2020. See Appendix for test configuration details. Calculation for single-thread at QD4: (ADQ "On" - ADQ "Off" Baseline)/(ADQ "Off" Baseline) = (285 us - 405 us)/405 us *100% = -30% reduction in P99.99 latency or + 30% increase in predictability.

[2] Up to 50% reduction in latency as measured by mean latency reduction for ADQ "On" vs. ADQ "Off" Baseline. Source: Lightbits Labs testing conducted July, 2020. See Appendix for test configuration details. Calculation for single-thread at QD32: (ADQ "On" - ADQ "Off" Baseline)/(ADQ "Off" Baseline) = (215 us - 458 us)/458 us *100% = -53% reduction in mean latency.

[3] Up to 70% improvement in throughput as measured by IOPS for ADQ "On" vs. ADQ "Off" Baseline. Source: Lightbits Labs testing conducted July, 2020. See Appendix for test configuration details. Calculation for single-thread at QD32: (ADQ "On" - ADQ "Off" Baseline)/(ADQ "Off" Baseline) = (141,000 IOPS - 81,300 IOPS)/81,300 IOPS *100% = 73% improvement in throughput.

[4] Source: Lightbits Labs: lightbitslabs.com/news/lightbits-adds-nvme-tcp-clustered-storage-solution-to-lightos

[5] Lightbits, "Lightbits Labs Celebrates NVM Express Ratification of NVMe/TCP Transport Standard," globenewswire.com/news-release/2018/12/03/1660878/0/en/Lightbits-Labs-Celebrates-NVM-Express-Ratification-of-NVMe-TCP-Transport-Standard.html

[6] Source: KernelNewbies: kernelnewbies.org/Linux_5.0#Storage

[7] Source: Lightbits downloads: lightbitslabs.com/nvme-tcp-drivers

[8] Intel Ethernet Technology page: intel.com/content/www/us/en/architecture-and-technology/ethernet.html

[9] Intel, "Restoring the Balance Between Bandwidth and Latency," intel.com/content/www/us/en/architecture-and-technology/optane-technology/balancing-bandwidth-and-latency-article-brief.html

[10] Source: Lightbits Labs paper, "Disaggregation of Cassandra nodes No Drama Lightbits LightOS SDS," lightbitslabs.com/wp-content/uploads/2020/04/SB_Cassandra-1.pdf

[11] Lightbits, "Apache Kafka and LightOS," lightbitslabs.com/ty-solutions-brief-kafka

[12] See endnote 3.

[13] Source: Dell EMC: dellemc.com/en-us/collaterals/unauth/white-papers/products/storage/h18200-dell-emc-powerstore-oracle-best-practices.pdf

[14] Source: VMware: kb.vmware.com/s/article/2053145

[15] See endnote 1.

[16] See endnote 2.

[17] Source: NVM Express: nvmexpress.org/answering-your-questions-nvme-tcp-what-you-need-to-know-about-the-specification-webcast-qa

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit intel.com/performance.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.