WHITE PAPER

NVMe-oF OVERVIEW

TABLE OF CONTENTS

Contents

INTRODUCTION TO DIRECTFLASH™ FABRIC

Pure Storage® is an innovation and industry leader[1] in all-flash and was one of the first vendors to recognize the value of NVMe™ and NVMe over Fabrics™ (NVMe-oF™). Pure's DirectFlash Software and DirectFlash Modules[2] optimize NVMe technology to manage flash more efficiently when compared to either industry standard NVMe or SAS SSDs. IT organizations now have the ability to take advantage of NVMe in their infrastructure with the introduction of high density NVMe capable servers. Unfortunately wide scale adoption of direct attached NVMe comes at the cost of changing out physical server infrastructure and requires that the application have the ability to use scale out storage infrastructure to provide data protection. This can also lead to data siloing and organizations spending more time and money managing the location of their data.

Instead of building out new physical and application infrastructures, organizations can take advantage of NVMe-oF to leverage an increase in performance and density and realize the benefits of flexible, efficient, and shared network storage. In this white paper, we discuss how using Pure's DirectFlash Fabric with NVMe-oF to connect servers to FlashArray™ allows organizations to drive lower latency performance and higher density storage to servers and applications across a fabric using common storage practices and data center infrastructure. In addition, Pure provides these applications rich, enterprise services normally not available with direct attached storage architectures.

BACKGROUND

Access to persistently stored data at a lower latency is the key to application performance. With the improvement of CPUs, moving stored data into memory across a communications medium has become a focal point for improving the end user experience through better application performance. There continues to be many emerging technologies that provide organizations with a method for accessing stored data faster by placing it closer to the CPU. But these technologies come at a premium in cost and scalability. The industry is always looking for the best balance between performance, scale, cost, and data protection.

The most common technology used to address these requirements has been to use HDD within a local server or in an array located on a common fabric. With the introduction of NAND flash technology many of the applications which used local drives started replacing HDD with SSD flash drives and in some cases PCIe attached flash devices. The introduction of these flash technologies as a persistent storage medium created a revolution. SSD technology simply replaced the spinning platters within a HDD with flash technologies. These drives kept the underlying SAS/ SATA transport technologies and communications mechanisms. While consumers of this technology saw significant

[1] https://www.purestorage.com/resources/type-a/gartner-mq-2018.html
[2] https://www.purestorage.com/products/purity/directflash.html

performance improvements and the ability to integrate with existing SAS and SATA controllers, the drives were not being utilized to their fullest potential.

The inability to achieve maximum performance from flash devices provided an opportunity for a new class of PCIe connected storage devices using NAND technologies. These PCIe connected devices provided better performance but at a higher cost. These devices were also decentralized which created additional overhead in infrastructure management. These initial flash based PCIe devices were also proprietary as no standard communication mechanism yet existed. The proprietary nature contributed to the cost and also added the overhead of managing the device drivers and dealing with the lifecycle of the products and vendor lock-in. Intel began work on a standard in 2007 and released the NVMHCI 1.0 specification in April 2008.[3] In the second half of 2009 more than 90 companies, including Intel, began work on an industry standard for accessing flash. This group was the NVM Express™ Workgroup and Version 1.0 of the NVMe specification was released in March of 2011.[4] In 2012 NVMe devices started becoming publicly available and inbox drivers began appearing in major OS distributions. In the years that followed NVMe devices started appearing in various form factors including a 3.5 drive form factor. In response to the demand, server manufacturers began providing servers with NVMe capable drive slots that were front panel accessible.

One of the primary challenges with NVMe technology using PCIe as a transport mechanism is that the placement was limited in scale and distance. In order to provide more connectivity than a server provided or to add capabilities for a server that did not have PCIe drive slots or limited PCIe slots there needed to be a method for connecting NVMe devices remotely across a fabric. There has been some work to extend the PCIe bus beyond the server. This typically includes adding a PCIe switch to the design. This design introduces a new set of challenges such as redundancy, hot plug of NVMe devices, scalability, and operations overhead.

Given the need to increase scale and distance, the NVMe Express group began work to standardize communications with NVMe devices across fabrics. In 2014 a standard for using Fibre Channel (FC), FC-NVMe[5], was proposed and a group was formed within the INCITS T11 committee. Around the same time the NVMe Express group began work on a transport binding specification for Remote Direct Memory Access (RDMA) over fabrics. Version 1.0 of NVMe over Fabrics was completed in June of 2016 and includes a reference to the work of INCITS as a transport mechanism. The FC-NVMe standard (INCITS 540-2018) was completed in April 2018.

Providing a method for accessing NVMe over a Fabric using more efficient communication and queueing mechanisms allows consumers of the technology to overcome some of the infrastructure limitations. In particular the NVMe devices no longer need to be installed into individual servers. This eliminates the need to add new server models capable of supporting NVMe drives to the infrastructure. It also eliminates the need for an organization to use a scale out storage infrastructure. Instead of changing the way storage is consumed and maintained, NVMe-oF allows an organization to leverage the benefits of consuming storage from remote devices, like an array, connected to a fabric. These benefits include the advantages of scale, lifecycle management, snapshots, reliability, data mobility, and data protection.

[3] https://nvmexpress.org/wp-content/uploads/NVMHCI-1_0-Gold.pdf
[4] https://nvmexpress.org/resources/specifications/
[5] https://standards.incits.org/apps/group_public/project/details.php?project_id=1705

The added benefit of using NVMe-oF is that the performance rivals that of Direct Attached Storage (DAS) which opens up new workloads and architectures.

Since the formation of the working group there has been an emergence of NVMe-oF products. These products fall into 4 primary groups that may be consumed by end users or manufacturers: JBOF (Just a bunch of flash), NVMe-oF appliances, NVMe-oF arrays, and NVMe-oF storage back end.

The first group and one of the early market products is the JBOF. This product provides a shelf of flash with an NVMe-oF controller that allows users to connect servers to NVMe devices over the fabric. These products are targeted at high performance and scale out applications. The devices typically lack the data reduction efficiency, high availability and data protection offered in enterprise class all-flash arrays and rely on software and application design to handle such deficiencies.

The NVMe-oF appliance is the second group of products. These are devices that connect a shelf of NVMe across a fabric but offer additional data services. These services may include data protection using erasure coding. These devices also typically provide some hardware resiliency to provide high availability. These products also target high performance applications. Some of the challenges with these products are scalability, custom software clients and hardware lifecycle.

The NVMe-oF enabled array is the emerging offering of the NVMe-oF storage products, but ironically it is also the most mature and established of the end products that organizations will consume. An NVMe-oF enabled array provides all of the management, scalability, cost management, and data protection of a traditional iSCSI or FC array with the added benefit of NVMe-oF performance, which is the focus of this paper. Simply put, this solution provides a standard communications transport for initiators to communicate with the storage device. These transports can include any of the NVMe RDMA transports like RDMA over converged Ethernet (NVMe/RoCE™), NVMe over InfiniBand (NVMe/IB™), NVMe over Internet Wide Area RDMA Protocol (NVMe/iWARP™) and can also include NVMe over Fibre Channel (NVMe/FC™), and NVMe over TCP (NVMe/TCP™).

NVMe-oF back end is a product class that is typically used by manufactures. Until recently array manufacturers have used SAS to connect additional shelves to scale up capacity on an array. In an all-flash array SAS becomes the bottleneck as described previously in the introduction. Using NVMe-oF provides the benefit of improving performance by reducing the bottleneck and taking advantage of NVMe standards. Pure Storage was one of the first vendors to create an NVMe-oF backend with the combination of FlashArray//X and DirectFlash Shelf.[6]

[6] https://www.purestorage.com/content/dam/purestorage/pdf/datasheets/ps_ds_flasharray_03.pdf

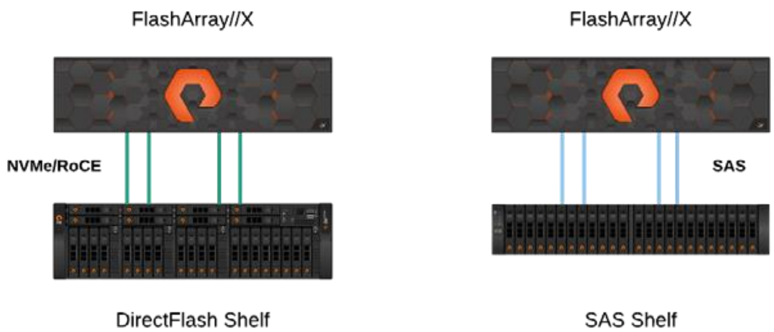

FIGURE 1. FlashArray with DirectFlash and disk shelves

Given the experience and success using NVMe/RoCE with the DirectFlash shelf, Pure has included support for server connectivity to the FlashArray//X using **NVMe/RoCE v2** in Purity version 5.2. While this is the first NVMe-oF transport to be offered on FlashArray//X, it will not be the only transport offered. With the completion of FC-NVMe in April of 2018 and the ratification of NVMe/TCP in November of 2018, there will now be three choices for NVMe over Fabrics: Fibre Channel, RDMA, and TCP. The key differentiator of RDMA technologies is that it can use shared memory to access data. This model closely represents the PCIe transport and provides performance improvements by bypassing several layers of communication within the stack when reading and writing the data.

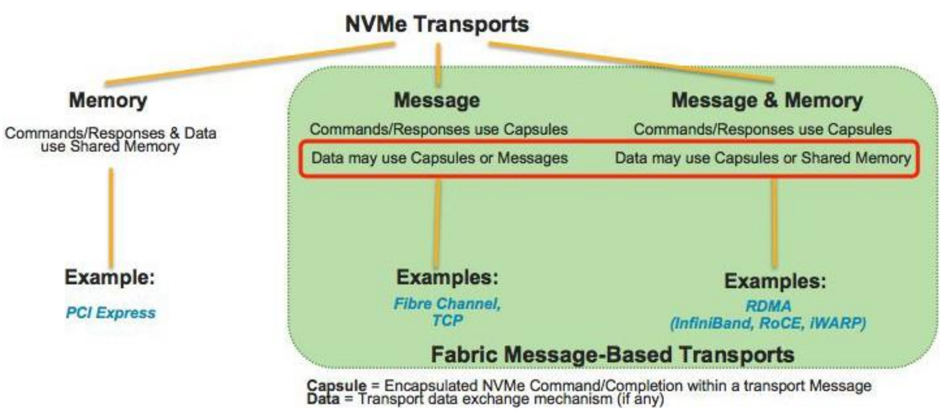

FIGURE 2. NVMe transport options

Choosing the appropriate NVMe Transport will depend on several factors including:

- Application Requirements
- Infrastructure Availability
- Operational Model
- Cost to Implement
- Data Protection
- Business Continuity

The primary goal of this document is to help you understand the use cases for NVMe/RoCE with FlashArray//X. We will identify some of the performance characteristics of NVMe/RoCE with the FlashArray//X and describe some appropriate architectures and their operational and lifecycle benefits. In addition we will review some target application use cases for a NVMe-oF FlashArray using RoCE.

BASELINE PROTOCOL PERFORMANCE

The purpose of the baseline performance test was to create a controlled testing environment where the transport was the only variable. By doing this we were able to clearly identify the performance characteristics of each protocol for a standard set of tests. The baseline testbed setup includes 2 initiators directly connected to a FlashArray//XR90 with a DirectFlash shelf and 46 TB of raw capacity.

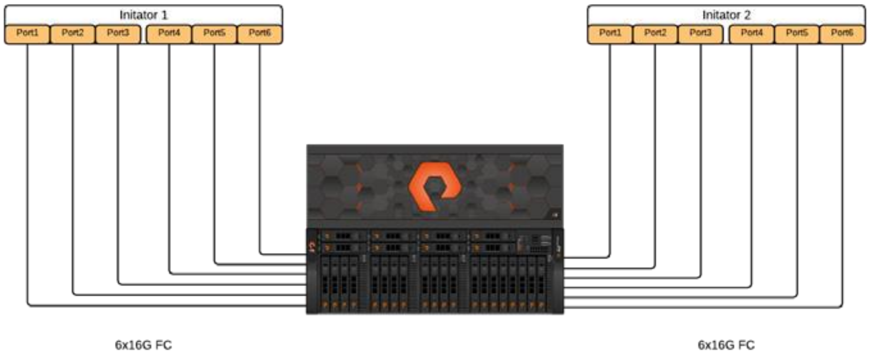

FIGURE 3. Baseline test setup – Fibre Channel

For Fibre Channel, the initiators were directly connected with 6x16G FC connections.

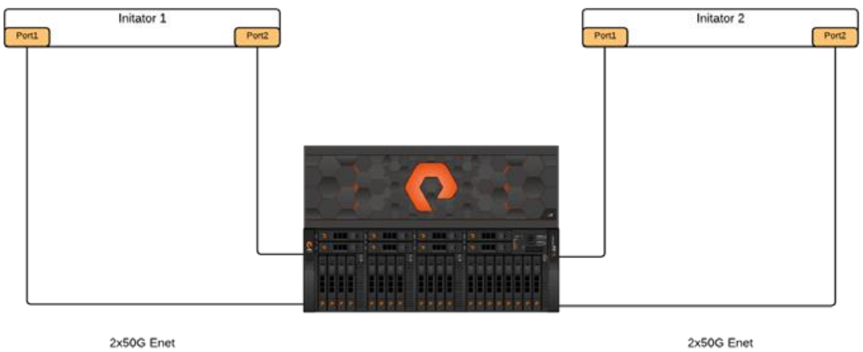

FIGURE 4. Baseline test setup – Ethernet

For NVMe/RoCE and iSCSI, the initiators were directly connected with 2x50G full duplex Ethernet connections.

For the baseline performance metrics between protocols we ran the same job over each protocol. This job consisted of a series of read and write tests at varying block sizes to multiple volumes on the array over a period of several hours. During this time the IOPS from the initiators were incremented until the links were pushed to full saturation. The job was run multiple times on each testbed to validate that the test results were consistent. The average latency and throughput were measured as the baseline indicators. We then combined the data points and averaged the results for read and write operations to establish a baseline metric.

BASELINE METRIC RESULTS

What the baseline metric shows is that NVMe/RoCE is on par with Fibre Channel performance for average latency and IOPS across an averaged set of read and write data points. When looking at NVMe/RoCE vs FC baseline metrics the key takeaway is that, on average, a standard ethernet fabric now has the same latency and I/O properties of a Fibre Channel environment.

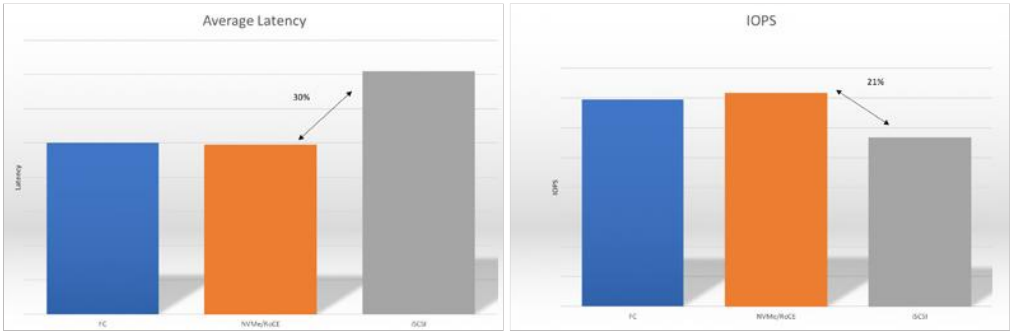

FIGURE 5. Baseline metric results

This data shows that the NVMe/RoCE baseline metric demonstrates a significant improvement over iSCSI: an average of 30% latency improvement for non-saturated traffic and an average of 20% I/O improvement.

By averaging read and writes across the different test the baseline metric value smooths out some of the other interesting values. For example in some instances we saw up to a 50% decrease in latency when comparing NVMe/RoCE to iSCSI and a 20% decrease in latency when comparing NVMe/RoCE to Fibre Channel.

While that data point alone does not represent an endorsement to replace a Fibre Channel fabric with an ethernet fabric, it does show a proof point that ethernet can not only provide on par performance as demonstrated in the baseline metric but can also outperform FC for some workloads. But this data should not be considered as endorsement to move away from FC, it simply validates the expected performance improvements offered by NVMe-oF as opposed to SCSI.

In fact, as mentioned before, the FC-NVMe standard is bringing the NVMe-oF transport to the Fibre Channel fabric (NVMe/FC) and we fully expect that it will provide a performance improvement in that environment as well. The journey to NVMe-oF will vary depending on application and organizational requirements. For organizations that have made the decision to use ethernet and iSCSI for connectivity to block storage devices the data represented in the baseline metrics indicates that they no longer have to compromise performance.

The key takeaway is that NVMe-oF delivers on the promise of increased performance and that NVMe/RoCE provides a significant and measurable performance improvement for typical ethernet based storage connectivity within the data center. For customers who have been considering consolidation and simplification of networks within the data center with a move to ethernet, the NVMe/RoCE baseline metric provides another data point for that consideration. But performance is only a part of the decision process. One of the primary considerations for adopting any NVMe-oF transport would be understanding your particular application use case and architecture.

USE CASES

Like most technologies it is important to identify if and where that technology fits into your organization. As highlighted before there are 3 options for NVMe-oF. NVMe/FC will likely be consumed much like traditional Fibre Channel today. Meaning that customers who have existing FC infrastructure capable of supporting NVMe/FC will simply change the way initiators communicate with NVMe targets by enabling NVMe/FC within their infrastructure. Customers who are leveraging IP over ethernet to communicate with their SCSI targets will have more options. One option will be to wait for the NVMe/TCP to become a mature and mainstream transport and the other option will be to start adopting NVMe/RoCE for applications.

The key decision point between ethernet options will be determined by the application requirements. If customers are looking to provide storage for traditional applications at the lowest cost where performance is not a factor, then iSCSI will continue to be the a popular solution until NVMe/TCP matures as a transport. For organizations who are using or considering NVMe PCIe cards or NVMe drives scaled out across multiple physical servers and for those looking to achieve better latency and throughput with fewer CPU cycles on the initiators serious consideration should be given to an NVMe/RoCE solution.

Many of the next generation cloud-native class of applications lend themselves to a scale out infrastructure. This class of application includes such applications as:

- Hadoop
- MariaDB
- MongoDB
- Cassandra

One of the main challenges organizations have supporting these next generation applications is the physical components required for them to run on. The amount and type of servers required for scale out applications create operational and lifecycle management challenges. In particular there is often inefficient use of the storage due to capacity being locked into compute silos on these devices. This can also be costly because adding more storage to support the application requires adding additional underutilized compute nodes. This all results in more physical devices to manage, more drives to replace, stranded CPU or storage, more maintenance costs, and often disruption during maintenance or upgrades.

The first organizations to understand how to overcome the limitations of physical infrastructure were those running hyperscale data centers. These organizations looked to reinvent physical infrastructure and separate the application from the physical devices. Much of the work they have done has made its way into the Open Compute Project.[7]

The methodology they used has led the way for a new class of architecture known as disaggregated infrastructure. In these infrastructures the goal is to break up the individual components of compute, network, and storage into separate devices which each have an individual life cycle but look and perform for the application as though the resources are part of one physical server.

[7] https://www.opencompute.org/

Because NVMe/RoCE most closely resembles NVMe over PCIe there is an excellent case for using NVMe-oF to disaggregate storage for the infrastructure. Using some of the hyperscale designs as a model it is easy to see how a FlashArray//X could become the disaggregated storage component in a scale out model, while providing enterprise features not typically found in scale-out, direct attached storage architectures.

For example here we show a 42RU rack with 32 compute nodes connected to a FlashArray//X90R2 with a DirectFlash Shelf over a NVMe/RoCE fabric contained within a single rack. Using a pair of IEEE 802.3 standards-based switches in the middle of the rack to serve as dedicated connectivity to the storage within the rack, we are able to provide a storage performance experience that is near the performance of an internal NVMe device (Direct Attached Storage). The top of rack switches provides the IP connectivity for the application.

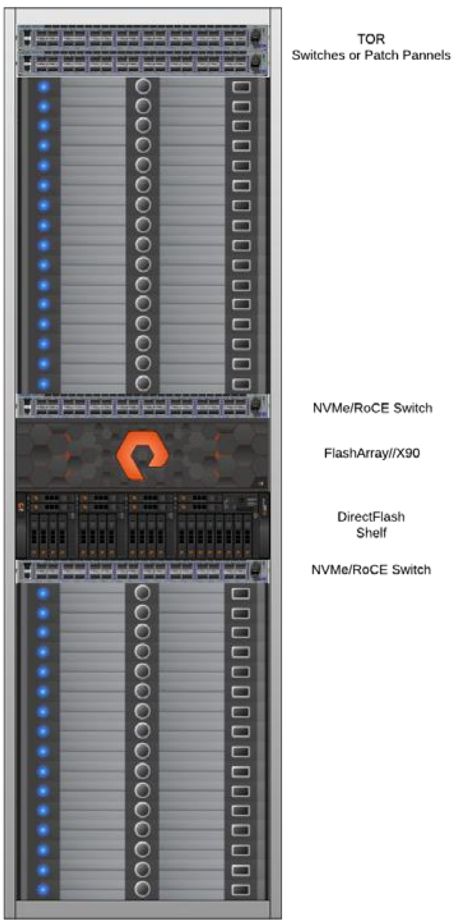

FIGURE 6. Single rack example

Depending on the application data profile this could provide as much as 93TB[8] of usable capacity per compute node. In addition to more usable capacity, the array provides the additional functionalities of snapshots, data protection, non-

[8] Assumes an equal division of total capacity across all compute nodes and a 4.5:1 compression ratio

disruptive upgrades, lifecycle management, and the benefits of Pure's Evergreen Storage™ business model.[9] Racks could also be configured to include more compute nodes using modular server technologies. This type of architecture provides a great deal of flexibility. It is important to note that in addition to physically separating storage and compute that they also become separate fault and maintenance domains where a failure or maintenance of a compute node does not impact storage performance and availability.

Another example of a disaggregated model using NVMe/RoCE and the Pure Storage FlashArray//X would be a pod-based design. As shown in this example we have a pair of 42RU racks each with 40 1RU servers connecting to a middle of pod storage rack. As in the previous example, the switches in the storage rack act as a separate fabric for the NVMe/RoCE network in order to provide a storage performance experience that is near the performance of an internal NVMe device. The top of rack switches for the compute racks act as the IP connectivity for the application.

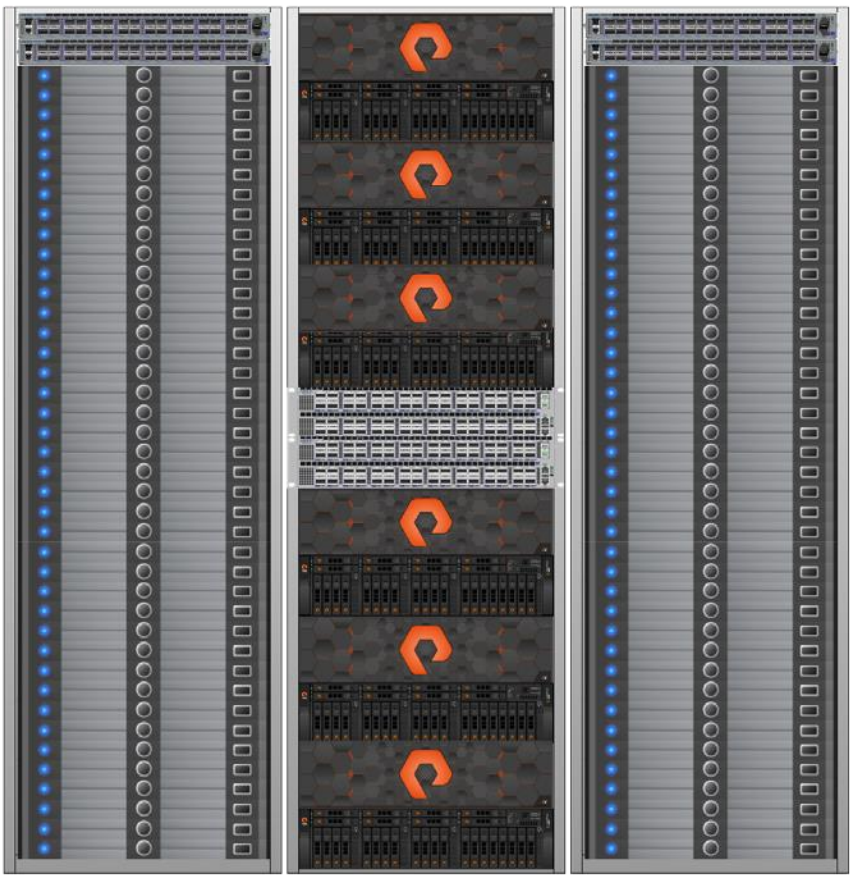

FIGURE 7. Three rack example

In this example there are a total of 6 FlashArray//X90R2 each with a dedicated DirectFlash Shelf. Depending on the application data profile, this can provide as much as 209TB[10] of usable capacity per compute node. Like the previous

[9] https://www.purestorage.com/products/evergreen.html

[10] Assumes an equal division of total capacity across all compute nodes and a 4.5:1 compression ratio

example, the number of server nodes per rack or even the number of racks per pod could vary depending on the application requirements. One of the major benefits of a rack or pod-based architecture is that it provides a repeatable design for scaling out over time. This strategy can help minimize the upfront investment. And like the previous example the advantages of snapshots, data protection, lifecycle management, non-disruptive upgrades, and the Pure Evergreen Storage subscription are available.

These designs are some examples of how NVMe/RoCE can be used the FlashArray//X but should not be considered exclusive designs. We expect that over time, customers will find additional use cases and deployment scenarios.

In addition to building the physical infrastructure, the appropriate application would need to be selected. When selecting the application, a simple baseline performance number would not provide enough data to make an informed decision, instead there would be a need to test specific application performance using NVMe-oF vs. the current fabric or DAS solution.

APPLICATION PERFORMANCE

During the course of testing our solutions team compared the performance of several applications. They tested both enterprise class and next generation applications. For the application performance testing the setup included a fabric to mirror a more production-like environment. Each of the Initiators contained two 2-port FC or ethernet cards and one port from each card was connected to each of the two fabric switches. This provided maximum throughput and redundancy.

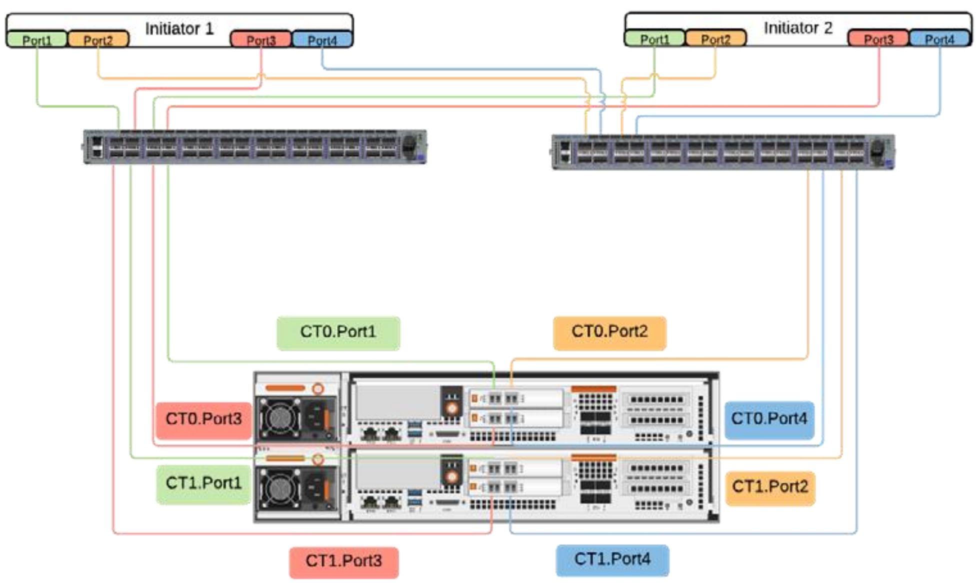

FIGURE 8. Performance testing setup

With Linux being one of the first operating systems to support NVMe-oF, we focused testing on new webscale and cloud-native apps such as MongoDB, Cassandra, and MariaDB. In every case, applications on NVMe/RoCE consistently outperformed iSCSI. In the case of MongoDB there was 25% faster application access over iSCSI. Cassandra showed 24% lower latency than iSCSI and MariaDB tests showed 30% faster application transaction times.

But in most cases these applications use Direct Attached Storage (DAS) on the servers as opposed to iSCSI. In an effort to understand how NVMe/RoCE would compare the team tested the performance of a multi node Cassandra cluster using SAS connected SSD within each node to that of the same nodes connecting to NVMe devices on a FlashArray//X with the testbed as described above.

The results for this test were compelling. The latency decreased by more than 28% while operations per second increased by more than 33%. While this was using SAS based SSDs in the server instead of NVMe it should be pointed out that unless an organization has made the choice to purchase servers that support high numbers of NVMe drives, the more common deployments would be SAS-DAS.

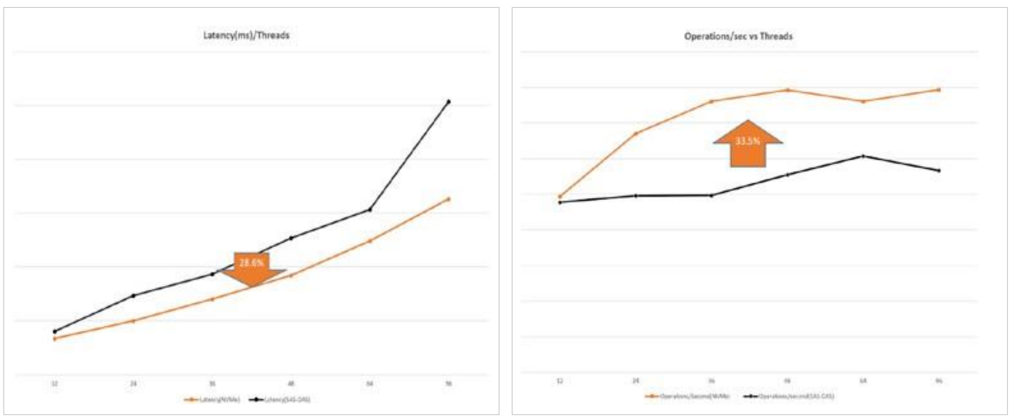

FIGURE 9. Performance testing results

For MongoDB the team also tested NVMe/RoCE performance vs DAS. As the thread count increased beyond 32, we saw an increase in operations per second. At 96 threads the performance using FlashArray//X and NVMe/RoCE was 54% higher.

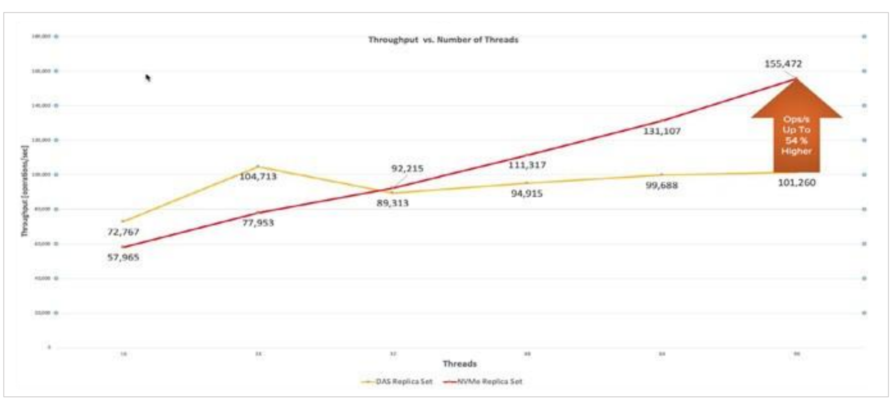

FIGURE 10. Performance testing results vs DAS

Overall, the application performance testing lines exceeded the expected baseline metrics and line up with some of the higher performance data points that were seen across the multiple test points. What all of the tests show is that NVMe-oF provides benefits for application performance for both applications that may currently be using iSCSI or even DAS. The key to understanding how this may benefit your organization is by testing for your application and architecture needs.

CUSTOMER FEEDBACK

During the development of the NVMe/RoCE feature we were fortunate enough to have some of our valued and trusted customers test the solution with their applications. These customers were looking to build a type of disaggregated infrastructure where the application could take advantage of the NVMe-oF performance improvements for an ethernet fabric. For these customers the applications are internally developed and scaling the storage to meet the application demand can be difficult and time consuming. In both cases they were looking to NVMe-oF solutions to help them meet the needs of their organization.

One of these customers was also exploring JBOF and appliance NVMe/RoCE solutions that do not provide rich data services. The main metric they were interested in for this application was write latency. The feedback was that they consistently saw on the order of 200us-250us write latency. This was significantly lower than the iSCSI solution that they currently use in their environment. When comparing to the appliance and JBOF solutions using NVMe/RoCE their results ranged from 50us to 200us, but none of those solutions provided the efficiency or high availability offered by the FlashArray//X. For their specific application they were willing to sacrifice some latency to gain the data services, redundancy, support, and non-disruptive upgrades provided by Pure Storage.

Another customer running a custom application is currently using a disaggregated rack design with multiple compute nodes connecting to a FlashArray via iSCSI. They have been testing a NVMe/RoCE deployment and have been very pleased with the initial results. In fact, they reported double the performance, in some instances, over what they have seen with their similar iSCSI deployment.

As other customers begin to explore NVMe-oF from Pure Storage we are seeing some commonalities. Most of these early adopters are looking to solve a specific problem like:

- Disaggregating compute and storage
- Simplifying data center operations
- Reducing overall cost
- Improving performance
- Reducing downtime
- Efficient use of resources

CONSIDERATIONS

Making the decision to implement an NVMe-oF solution requires planning and analysis. A key thing to remember is that NVMe-oF is a new technology. Outside of the hyperscale environments new technologies take time to incubate and this presents obstacles for most organizations. One of the first barriers to entry is the NVMe-oF ecosystem. The first consideration is the operating system. Currently NVMe-oF is supported in many Linux distributions. Specifically, distros that are using the nvme-core from Kernel 4.19 or later have the necessary components to run NVMe-oF (RoCE or FC). Users are always welcome to utilize an upstream kernel or even compile their own, but that is not a standard practice in most IT shops. At the time of writing this white paper, many of the major Linux distributions have the appropriate support including RHEL/CentOS 7.6 and SLES 15. VMware has demonstrated NVMe-oF capabilities but has not yet publicly announced when that support will be generally available.

In addition to operating system support, there is the need for application vendor support. It will take user pressure and adoption for some vendors to seriously consider certifying their applications on NVMe-oF. As an example, iSCSI support is still lagging for many applications, but some of that is due to performance. NVMe-oF does have the performance advantages, but adoption is just beginning. This is one of the main reasons why some of the early interest in NVMe/RoCE comes from custom application use cases.

Many of these applications benefit from a disaggregated model and users are looking to simplify the design when it comes to storage connectivity. In the rack or pod design, which follows the hyperscale model, the goal is to have minimal latency between the storage and compute nodes. For a NVMe/RoCE design that means creating a lossless fabric for connectivity of compute and storage. The ethernet market has seen a dramatic increase in speeds and switches that support 100G/50G/25G are economical and widely available. In addition, these switches are primarily data center class products that support capabilities like Priority-based Flow Control (PFC) and Explicit Congestion Notification (ECN) which provide lossless fabric capabilities. These well-established features work in conjunction with the IP stack and adapters on the servers and array to provide the lossless capabilities. The configuration practices for a RoCE environment is very consistent and easy to implement. Many RoCE best practices have been established and are well tested in very large scale deployments. Pure Storage works with our networking partners to provide NVMe/RoCE best practices and configuration examples across multiple vendor platforms.

SUMMARY

NVMe over Fabric is a significant step in the evolution of storage and overall data center modernization. Even though there has been much talk about the topic since the ratification of the standard in 2016, there has been little real traction of the technology until now. It appears that with more products, an increase in the size of the ecosystem, and more interest that adoption will only go up. With the 2018 ratification of NVMe/TCP and the FC-NVMe standard which enables NVMe/FC, there are more transport options beyond NVMe/RDMA (which includes NVMe/RoCE). The increase in transport options will create new use cases and ultimately drive further adoption of the technology.

The key thing to consider in these early days is whether the solution meets the requirements. Customers need to evaluate how NVMe/FC, NVME/RoCE, or NVMe/TCP will solve their current business challenges, reduce the complexity of their IT investments and deliver innovation in their organizations. For Pure Storage adopting NVMe/RoCE was the logical step for the DirectFlash Shelf because it allowed us to scale our all-NVMe array, while maintaining the NVMe transport advantages outside of a single chassis. When we examined what we were seeing in the industry concerning scale out storage and disaggregation, NVMe/RoCE made sense as our entry point into NVMe-oF. Our guidance for customers is to evaluate both the benefits and the challenges against their long-term strategy. The evidence is clear that there are benefits to NVMe-oF but choosing when and what transport to adopt remains unique to each customer's current and future business and application requirements.

At Pure Storage, our commitment is to deliver provide simple and efficient storage solutions that our customers love. We have led the flash revolution right from the beginning, and with FlashArray//X and the adoption of advanced flash memory technologies- like NVMe in our DirectFlash Software and DirectFlash Modules, and NVMe/RoCE in the DirectFlash Shelf - we have lead the adoption of NVMe-oF for production workloads. We are now extending DirectFlash into the fabric and out to the server with the introduction of NVMe/RoCE. We remain committed to providing our customers improved performance and rich enterprise data services by leveraging emerging technologies in a world class enterprise array.

ps_wp16p_nvme-of-overview_ltr_01