White Paper

Planning for the Transition to Production-Ready NVMe over Fabrics Deployments in the Enterprise

Sponsored by: Broadcom

Eric Burgener
April 2020

IDC OPINION

Businesses worldwide are undergoing digital transformation (DX). DX is the move to much more data-centric business models, and it makes data a strategic asset that enterprises can mine to help inform better business decisions. As information technology (IT) organizations make this transition, they are deploying new next-generation applications (NGAs) that leverage mobile computing, social media, big data and analytics, and cloud technologies, and these workloads are in turn demanding more performance, availability, agility, and efficiency from IT infrastructure.

Nonvolatile memory express (NVMe)-based storage is a new high-performance storage technology that is increasingly being deployed for not only NGAs but also other performance-sensitive workloads in the enterprise. NVMe offers much lower latency, higher bandwidth and throughput, and significantly higher parallelism than the older SCSI-based storage. Since enterprise-class storage systems started to ship with persistent NVMe storage in 2017, revenue has grown at a rapid rate. By the end of 2019, all established storage vendors (Dell EMC, Hitachi Vantara, HPE, Huawei, IBM, NetApp, Pure Storage, etc.) offered NVMe storage in their flagship arrays, and the market for NVMe-based all-flash arrays (NAFAs) crested $2 billion in revenue.

Strong demand for NAFAs is also driving the deployment of a new higher-performance storage networking technology called NVMe over Fabrics (NVMe-oF). This protocol enables internal storage latencies (sub-100 microseconds) for enterprise-class shared storage, providing access to much higher capacities, much better capacity utilization, and a range of enterprise-class data services not available with internal NVMe storage. IDC expects a transition from SCSI to NVMe technology for enterprise storage over the next several years, and NVMe-oF will be a part of that. IT organizations with performance- and availability-sensitive workloads should have a strategy in place for how they will be leveraging NVMe-oF over time to ensure they stay competitive as their industry digitally transforms.

A key decision in enterprise NVMe strategy will be what type of network transport to use to network NAFAs, and enterprises should carefully consider the capabilities of Fibre Channel (FC) as they make this decision. FC networking is already in place for many performance- and availability-sensitive applications, and the ability of the protocol to deliver "performance at scale" for these types of workloads has been proven across two decades of production use. Given FC's performance, availability, scalability, simple manageability, and efficiency, it is an excellent fit as the transport protocol on which to base NVMe-oF networks, and for those customers already running FC, the move to the higher-connection host protocol can be as simple as a software upgrade.

IN THIS WHITE PAPER

Enterprises undergoing DX are encountering many new performance requirements as they add NGAs, and many of them are implementing NVMe-based storage technologies. As enterprises transition to NVMe, NVMe-oF will have to be part of that strategy at some point. The performance of NVMe-based technologies significantly heightens the demands on storage networking technology. This white paper reviews the considerations IT managers should take into account before moving to NVMe-oF, making the case for why FC is the best transport protocol option for the types of high-performance, high-availability workloads that will be run on NVMe-oF storage networks.

SITUATION OVERVIEW

For enterprises undergoing DX — the transition to much more data-centric business models — storage infrastructure performance requirements have been on the rise. While IT managers have to maintain many legacy applications, NGAs built around mobile computing, social media, big data and analytics, and cloud technologies are demanding new levels of not only performance but also agility. The business value of many of the newer, more real-time analytics applications significantly decreases with the time it takes to produce results, so both speed and agility are essential. To meet these requirements, the use of persistent solid state storage has become mainstream, and as IT managers evolve their storage infrastructure, they are also looking to leverage new technologies like NVMe to improve infrastructure density and efficiency.

Starting in the late 1980s, enterprises began to use the SCSI protocol to attach storage devices. Over time, drive manufacturers evolved their hard disk drive (HDD) products to work more efficiently with this protocol. The first solid state disks (SSDs) used in enterprises were also attached with the SCSI protocol, but SCSI did not leverage the solid state media very efficiently. A new protocol standard that had been developed specifically for persistent solid state storage devices was first introduced in 2011, and by 2019, enterprise-class external storage systems using that standard (called NVMe) generated $2 billion in revenue. Compared with SCSI, NVMe delivers roughly an order of magnitude better performance and many orders of magnitude higher parallelism (an important feature given the broad use of multicore CPUs today), as well as enabling higher-density storage devices and much better utilization of solid state storage resources. For performance-sensitive workloads, IDC expects enterprises to transition from SCSI to NVMe, and IDC predicts that by 2021, NAFAs will be driving over 50% of all primary external storage revenue.

Because NAFAs are so much more powerful than SCSI-based AFAs (SAFAs), the SCSI-to-NVMe transition is going to place much higher demands on storage networking infrastructure. A new host connection protocol called NVMe-oF was developed to address this issue, and early versions of this protocol started to ship in late 2017. Some of today's most performance-sensitive workloads in financial services, retail, healthcare, and social media are already using NVMe-oF to enable enterprise-class shared NVMe storage to consistently deliver sub-100 microsecond and lower latencies to attached servers. While customers can achieve these kinds of latencies by directly attaching PCIe SSDs as internal server storage, enterprise-class shared storage supports much larger capacities, makes much more efficient use of storage capacity, and enables the use of array-based data services (RAID, thin provisioning, compression, deduplication, snapshots, encryption, replication, etc.) that make storage easier to manage.

The efficiencies of end-to-end NVMe between servers and shared storage deliver not only much higher performance than SCSI can, but NVMe enables a streamlining of IT infrastructure as well by improving server CPU utilization, increasing storage densities, and lowering datacenter floor space consumption. Those deploying NVMe for performance-sensitive workloads also note that a given level of performance can be met with far fewer devices (relative to both HDDs and SCSI-attached SSDs), resulting in more efficient IT infrastructure. NVMe drives higher-performance densities in storage platforms while evolving accelerated compute technologies can handle higher bandwidth and more data on the server side. Together, these enable real-time and big data and analytics workloads that were not possible in the past and will demand significantly higher storage network throughput.

While the transition from SCSI to NVMe in arrays will happen much faster than the transition from SCSI to NVMe-oF for host connections, the industry is still clearly moving in the direction of NVMe-oF. There are enterprises that demand end-to-end NVMe performance today based on workload requirements, but other enterprises want to ensure that they purchase platforms on storage technology refresh that are "NVMe ready." The NVMe readiness enables a smooth transition to NVMe-oF technology for both devices and storage networks as it becomes necessary to meet evolving service-level agreements (SLAs). Enterprises will need to understand how they will be leveraging NVMe-oF technology for storage networking over the next three to five years, particularly as they are adding performance- and availability-sensitive NGAs as part of their DX. It will be critical to ensure that they have the right storage networking infrastructure in place to meet performance, availability, and scalability requirements as successful businesses evolve over time.

Planning the Transition to NVMe-oF

Many established enterprise storage providers (Dell EMC, IBM, NetApp, Pure Storage, etc.) are already shipping NAFAs as well as NVMe-oF. As customers deploy NAFAs from many of these established vendors to replace legacy SCSI-based arrays, they can choose to keep existing SCSI host connections or implement NVMe-oF. When near-term workload performance requirements are not hindered by SCSI-based storage networking, customers can make the end-to-end NVMe transition in two phases, providing extra scheduling flexibility. Customers do, however, want to know just exactly what is required to do the NVMe-oF upgrade and how application availability may be impacted during that process. How easy the transition to NVMe-oF will be with different types of systems and infrastructure in place should be a consideration as IT managers plan this evolution. The three network transport options for NVMe-oF are Ethernet, InfiniBand, and FC, and enterprises should plan carefully to ensure that they deploy the higher-performance host connection protocol with the transport option that best supports their performance, availability, and scalability requirements over time.

NAFAs and NVMe-oF will clearly be deployed first for an IT organization's most performance-sensitive workloads. For these types of workloads, most of these organizations are already using FC. FC was originally designed for demanding workloads that require low latency, deterministic delivery, and high availability at scale. Successful businesses will clearly grow, and FC was built to deliver these capabilities not only in single switch pods (i.e., a top-of-rack switch that just needs a high-performance network to attach servers to storage within a single rack) but also as environments grow to more complex switch fabrics that support hundreds to thousands of servers across many racks. While some early NGAs were deployed in single switch pods, most of these workloads will outgrow those environments and will need the "performance at scale" capabilities for which FC was originally designed.

For most customers that already have FC installed, moving to NVMe-oF will be very easy. For those organizations with a NAFA already networked using 32 Gb/s Gen 6 (which began shipping in 2016), moving to NVMe-oF can be as simple as just a software upgrade. Because of the simplicity of this migration, most enterprises already using FC will continue with it as they transition to end-to-end NVMe. Because FC was originally developed for the kind of performance-sensitive workloads that are most likely to benefit from end-to-end NVMe, it is a very good fit as a transport protocol for NVMe-oF. IDC has seen Ethernet used as a transport for NVMe-oF for greenfield deployments in single switch pod, single-workload environments, but before deploying NVMe-oF in this manner, enterprises should closely consider the issues discussed in the following bulleted list. InfiniBand is not expected to be used except in niche workloads in the enterprise, although it may achieve broader acceptance in technical computing environments as a transport for NVMe-oF. In detail:

- **Performance:** FC is built upon a foundation designed from day one to provide deterministic delivery. Ethernet does not provide deterministic delivery. In FC, traffic is not sent unless it can be immediately received (i.e., end-to-end buffer credits are available), and packets always traverse a known optimal and deterministic path to the receiver. Zero copy is an inherent part of the FC wire protocol, a feature that contributes to the low-latency packet delivery supported by FC networks. A separate remote direct memory access (RDMA) feature must be added on top of Ethernet to get it to support zero copy, and this feature is much less mature than the zero-copy capability that was built into the FC wire protocol from the beginning.
- **Scalability:** By default, FC cascades traffic across available links, thus protecting against congestion on any single link becoming a bottleneck that impacts the ability of FC to consistently provide deterministic, low-latency packet delivery even as network traffic scales. With FC, recovery does not require broadcasts and reconvergence, which with Ethernet can result in deadlock situations and get much worse as Ethernet networks grow. From congestion to flow control to link aggregation and recovery, FC is built to deliver "performance at scale," while Ethernet networks have to be specifically designed and configured for it using functionality that sits on top of the Ethernet wire protocol.
- **Management:** It's much easier to manage FC networks, particularly as they scale in size, because of the design of fabric services such as discovery, link aggregation, congestion and flow control, routing, and recovery, which were built in as part of the original FC wire protocol. The need to address the problems these features resolve in storage networks is why new higher-level protocols such as RDMA over Converged Ethernet (RoCE) and iWARP (both of which first started to ship in 2017) were introduced to run over the Ethernet transport. They must be separately configured during an Ethernet storage network deployment and require more manual tuning from experienced administrators to support low-latency storage packet delivery. Once environments grow beyond a single switch pod, it can become very difficult to meet performance and availability SLAs using NVMe-oF implementations such as RoCE and iWARP. IT managers should also note that the FC ecosystem is more standardized across vendors than Ethernet when it comes to supporting NVMe-oF, and administrators will have more to learn about vendor-specific NVMe-oF implementations when using Ethernet as the transport protocol.
- **Security:** FC networks are based on a "point-to-point design" where each attached server can see all storage shared to it but not any of the other attached servers. This makes the FC environment much more secure in the event a server in the network gets compromised. Ethernet is built around an "any-to-any design" where every attached server sees not only the shared storage but every other attached server as well. This is not the only determinant of security in fabrics, and both Ethernet and FC include other software-driven capabilities (virtual

networks, etc.) that help enforce security, but the "point-to-point nature" of FC does make it inherently more secure.

- **Efficiency:** The efficiency with which FC handles performance- and availability-sensitive workloads drives key advantages for both ease of management and cost relative to Ethernet. Because FC's fabric services are embedded as part of the FC wire protocol, FC's default settings drive high performance without requiring administrators to work with separately configured add-on facilities. Ethernet administrators will need to work with a variety of add-on services such as RoCE to tune Ethernet networks for better performance with storage workloads, an exercise that is increasingly difficult to perform as networks scale because Ethernet lacks deterministic delivery as a foundation. Because of this deterministic delivery in FC, flow control and recovery mechanisms require less overhead and scale more easily as networks are expanded. Packets rarely have to be resent to ensure reliable delivery. Congestion is identified and resolved at a network (rather than at a session) level so it is dissipated more rapidly. Ethernet storage networks, on the other hand, are often heavily overprovisioned (in terms of bandwidth) to meet SLAs and manage traffic at a session level — both factors that increase the costs of these networks.

  With FC, a high-performance NAFA requires fewer ports to meet a defined performance target than Ethernet (whether it is iSCSI or an NVMe-oF enabler such as RoCE or iWARP), resulting in configurations that much more efficiently utilize resources and require fewer links (and possibly switches, depending on scale). Given that an IT organization already has FC expertise, it is easier to configure and manage FC networks to meet SLAs for performance- and availability-sensitive workloads and easier to maintain that performance as a network scales than with Ethernet. And because of FC's deterministic delivery foundation, latency is less impacted by load, recoveries are faster, and overall network availability is higher than with Ethernet networks for these workloads.

- **Maturity:** The statement that Ethernet has been around longer, costs less on a component basis, and is more scalable than FC is generally true, although it is not necessarily true when it comes to block-based storage workloads. FC was designed for performance- and availability-sensitive block-based storage workloads and has been running them for two decades; Ethernet has been adapted for it with add-on facilities over time that are less mature than the Ethernet standard itself. When it comes to NVMe-oF, the Ethernet enhancements needed to support it (e.g., RoCE, iWARP) only started to be deployed in 2017 and some (TCP) only became available in 2019. Not only is this use case much newer on Ethernet, but there is more vendor-specific variability in how it is implemented than with FC. And in spite of implementations such as RoCE or iWARP, Ethernet still lacks the deterministic delivery foundation of FC, which means that even when those enhanced protocols are used, storage performance and availability on Ethernet networks are still much more subject to the size of the network and the amount of traffic and congestion identification and resolution and recovery operate less efficiently.

Additional NVMe-oF Considerations

There has been much discussion in the industry around whether Ethernet can be used as a transport for storage networks. Protocols such as iSCSI or FC over Ethernet (FCoE) have made this possible in the past, but it has been generally known that when storage workloads require high performance at scale or high availability or must accommodate unpredictable growth over time, FC provides a solution that better accommodates these types of demanding storage workloads. This is not to say that Ethernet shouldn't be used as a transport for storage networks but that enterprises should be cautious when deploying Ethernet to ensure that it will adequately meet the performance at scale, availability, and ease of manageability requirements over time. Because of the type of traffic for which NVMe-oF

will be deployed, all concerns about Ethernet performance, scalability, manageability, security, and efficiency for storage networking are only exaggerated. Maturity is also a concern because NVMe-oF implementations using Ethernet as a transport are quite new (with the latest NVMe over TCP being the youngest of all at this point). It is true that NVMe-oF has been deployed in production using RoCE, but these implementations have tended to be dedicated workloads in single switch pod environments over which very little other network traffic is flowing.

One might ask whether support for RDMA over Ethernet, while still young, offers the promise of resolving the "performance at scale" and availability concerns in the near term. RoCE and iWARP both support RDMA, while NVMe over TCP does not. It is true that Ethernet storage networks that use the RoCE or iWARP specifications can actually deliver latencies lower than FC in single switch pods, but in storage fabrics that require multiple hops, those latencies will still vary significantly based on network traffic (e.g., congestion) and what else is going on in the network (e.g., recovery), and they do still lack deterministic delivery. As network traffic increases (for a given configured bandwidth) or networks are expanded, latency variability will become worse in Ethernet networks even with RDMA.

IT managers should also note that implementing RDMA over Ethernet requires new smart NICs, which will put those servers using them on a different maintenance path than servers using "industry-standard Ethernet." It will also require that all network switches support the Data Center Bridging Exchange Protocol (DCB), an issue which may require updating network infrastructure. NVMe over TCP addresses this last concern by using the standard converged Ethernet adapters and drivers that are becoming available now, but without the RDMA support, the TCP implementation retains many of the traditional overprovisioning, performance, availability, and efficiency issues that were evident with iSCSI for more demanding storage workloads. FC, on the other hand, does not require new host bus adapters.

While performance and availability may be more pressing concerns with the types of workloads that will be run over NVMe-oF networks, cost is also a consideration. Ethernet components are definitely cheaper than FC ones on a component basis. Storage networks running over Ethernet, however, require more configured bandwidth (i.e., overprovisioning), more storage ports, and more sophisticated and manual management to achieve performance and availability objectives. They are likely to require dedicated networks to meet storage traffic needs and vendor-specific training for RDMA-based implementations. Any total cost of ownership comparison of the two approaches for performance- and availability-sensitive storage workloads will need to take into account not just a component cost comparison but also all the costs of configuring and managing a storage network to meet workload SLAs. And they should also take scalability considerations into account if the storage network is likely to grow.

CHALLENGES/OPPORTUNITIES

Although this white paper has not delved into a highly detailed level in exploring the "FC versus Ethernet" decision, it should be clear that there are many technical layers to be considered and that each of these factors becomes even more important when moving to a higher-performance host connection protocol such as NVMe-oF. With increases in Ethernet bandwidth and improvements in the ease of configuring Ethernet networks for performance- and availability-sensitive storage workloads, the transport will likely evolve to handle more demanding storage requirements over time. The key question for enterprises transitioning to NVMe-oF is: What is the best underlying transport for those workloads critical to the business today and necessary for DX in the future? The technical nature of any "FC versus Ethernet" discussion for storage networking complicates the decision, and the

challenge for customers is to cut through the technical details and make the best business decision given their objectives. The challenge for vendors such as Broadcom will be to clearly and concisely contrast the two approaches to help customers make the best decision for their specific environments.

For a vendor such as Broadcom, the breadth and maturity of its FC networking portfolio provide a good opportunity to meet performance- and availability-sensitive workload requirements as the industry transitions from SCSI to NVMe (and NVMe-oF). Because of the ease of migration, existing FC customers are likely to continue with FC as they make their initial transition to NVMe-oF. Customers already committed to Ethernet storage networking may need to consider adding FC networking as they deploy more demanding and real-time workloads during DX. It will be crucial for vendors, however, to be able to highlight the critical business benefits of FC as a transport for NVMe-oF if they are going to appeal to new customers.

CONCLUSION

NVMe-oF will be a part of the storage networking strategy for any enterprise deploying NVMe technology in shared storage platforms. IT managers have three network transport options to consider when deploying NVMe-oF: Ethernet, InfiniBand, and FC. For storage networking environments that will grow beyond a single switch pod environment, Ethernet and InfiniBand present real challenges to providing "performance at scale" for performance- and availability-sensitive block-based storage workloads, even when coupled with newer RDMA functionality. FC, on the other hand, was specifically designed for these types of workloads, and key features needed to meet stringent SLAs such as discovery, link aggregation, congestion and flow control, routing, and recovery were built in as part of the original FC wire protocol, and FC has a proven ability to meet demanding business requirements over two decades of use in enterprises worldwide.

As IT organizations move to NVMe-oF, they will be transitioning their more performance- and availability-sensitive workloads first. It is highly likely these workloads already reside on FC networks, and for these customers, migrating to NVMe-oF can literally be just as simple as a software upgrade. The NVMe protocol was specifically developed for solid state media, and NVMe drives lower latency, higher throughput, increased storage densities, lower floor space consumption, and higher media endurance and reliability much more efficiently than SCSI. Shared storage systems using end-to-end NVMe (i.e., NVMe devices in the array connected to hosts over NVMe-oF networks) can deliver an order of magnitude lower latency and many orders of magnitude higher parallelism than SCSI-based arrays. Any challenges an enterprise experienced in working with Ethernet-based storage networking in the past will only be magnified on NVMe-oF networks because the protocol enables so much more performance than SCSI. For NVMe-oF networks that are expected to grow over time and be highly utilized, enterprises should think very carefully before deploying Ethernet as the transport protocol, although there will clearly be workloads whose specific characteristics (performance, availability, scalability, etc.) are a good fit for it. For workloads that are performance- and availability sensitive and will be experiencing growth over time, the "performance at scale" that FC offers is a safer bet.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com