PURESTORAGE® | SPECIAL EDITION

NVMe over RoCE Storage Fabrics for NOOBS

- History of Networked Storage
- Unraveling NVMe over RoCE
- Industry's First Mainstream NVMe over RoCE All-Flash Array
- What's Next

NVMe over RoCE Storage Fabrics

For NOOBs

Table of Contents

New Technology Reference Guides

For NOOBs

What we have witnessed over the years is that early on in the life of new data center IT, it's difficult to find the information needed to quickly assess the situation.

New Technology Reference Guides for Noobs are designed to capture information we think IT pros need to know about an emerging technology.

We crowd-source information in these guides, which means we get a lot of it from people like you. If you think we missed something or simply have something you'd like to contribute, we welcome your input, and if we use it, we'll attribute the input to you.

Frank Berry

Founder & Senior Analyst

IT Brand Pulse

Where IT perceptions are reality

NVMe over RoCE Storage Fabrics

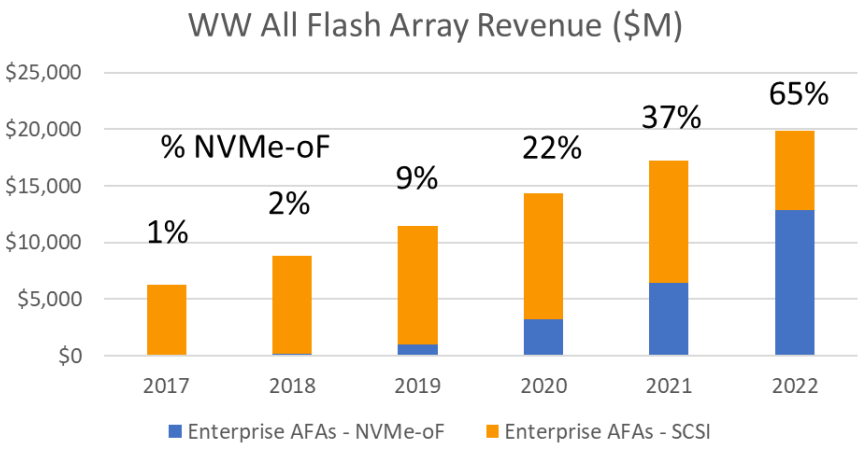
Introduction

While users at nearly 100% of businesses use some type of networked storage, IT Brand Pulse estimates that less than 1% of IT organizations are using NVMe storage fabrics. So, by definition, there are a lot of NVMe over RoCE Noobs out there.

Learning about this technology is urgent because analyst firm G2M Research forecasts that in only three years, two-thirds of all-flash array revenue will be from AFAs for NVMe fabrics.

Read NVMe over RoCE Storage Fabrics for Noobs if you want to bootstrap your knowledge of the market, technology and products in just a few minutes.

Thank you in advance for reading this publication, and for any feedback you send to frank.berry@itbrandpulse.com.



Source: G2M RESEARCH NVMe Ecosystem Snapshot

Chapter 1
History of Networked Storage

2000: Direct-Attached HDDs to Shared SAN/NAS Storage

In the 1990s, client-server computing started displacing monolithic mainframes and minicomputers. To scale storage as data grew, direct attached storage (DAS) was added to servers. IT organizations soon realized that storage and servers were single points of failure, storage was wasted with utilization at an average of only 30%, and network bandwidth was consumed by storage access and backup.

The low utilization of DAS spawned the network attached storage (NAS) architecture, where dedicated file servers were shared on a local area network (LAN). However, network congestion remained a major issue as storage access and backup shared the LAN with application servers.

Then at the turn of the millennium storage area networks (SANs) swept direct-attached storage out of the data center. CFOs loved the cost savings resulting from increased utilization of storage from an average of 30% to over 80%. Data center architects loved how they ran on dedicated storage networks, and how efficiently IT could manage their storage, as resources for hundreds of servers could be provisioned from a single array.

2010: Hyperscalers Change the Applications Landscape

By 2010, huge public cloud service providers like Facebook, AWS, and Google put the "hyper" in scale with distributed data center architectures encompassing a wide range of scale-out applications and infrastructure including databases, servers, networking and storage. At that time the fastest way to access data was to store it on flash memory inside servers. The pendulum swung back towards DAS for cloud-native applications with random IO profiles and varying block sizes such as big-data analytics and machine learning.

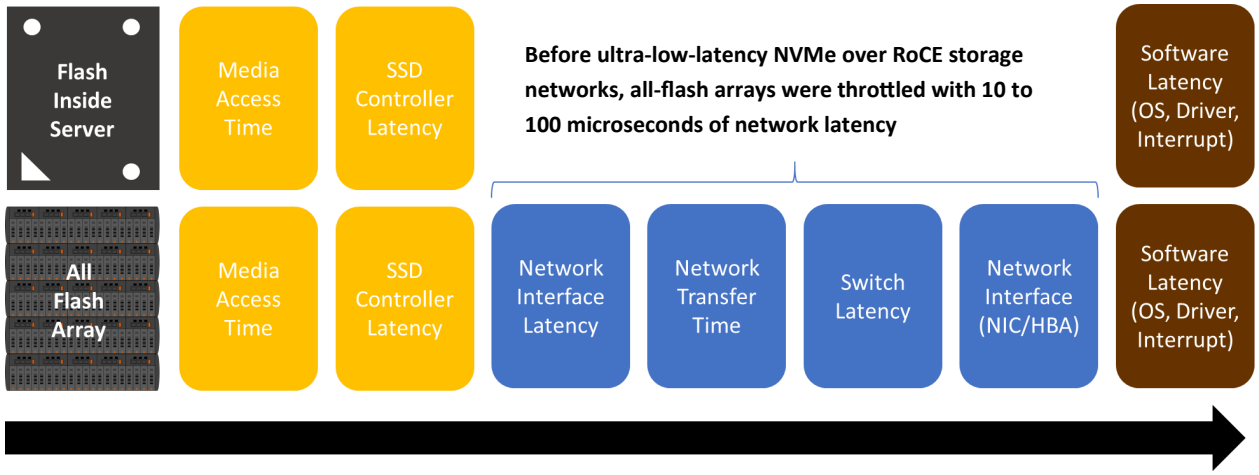
DAS Re-Emerged for Cloud-Native Applications

	Transactional Database	Virtualized Private Cloud	Analytics/ Data Warehouse	Scale-out Database	Big Data, AI, ML	Test / Development
	ORACLE, Microsoft SQL Server, SAP, Epic, MariaDB	vmware, Microsoft Hyper-V	kafka, HIVE, TERADATA.	cassandra, mongoDB.	Apache Spark, hadoop	DEV OPS
IO Profile	Random	Random & Sequential	Sequential	Random	Random & Sequential	Random
IO Size	Small	Small-Large	Large	Small & Medium	Large	Small-Large
Server Architecture	Scale-up	Scale-up	Scale-up & Scale-out	Scale-out	Scale-out	Scale-out
Storage Architecture	SAN	SAN	SAN & DAS	DAS	DAS	DAS

2010: All-Flash Arrays Arrive

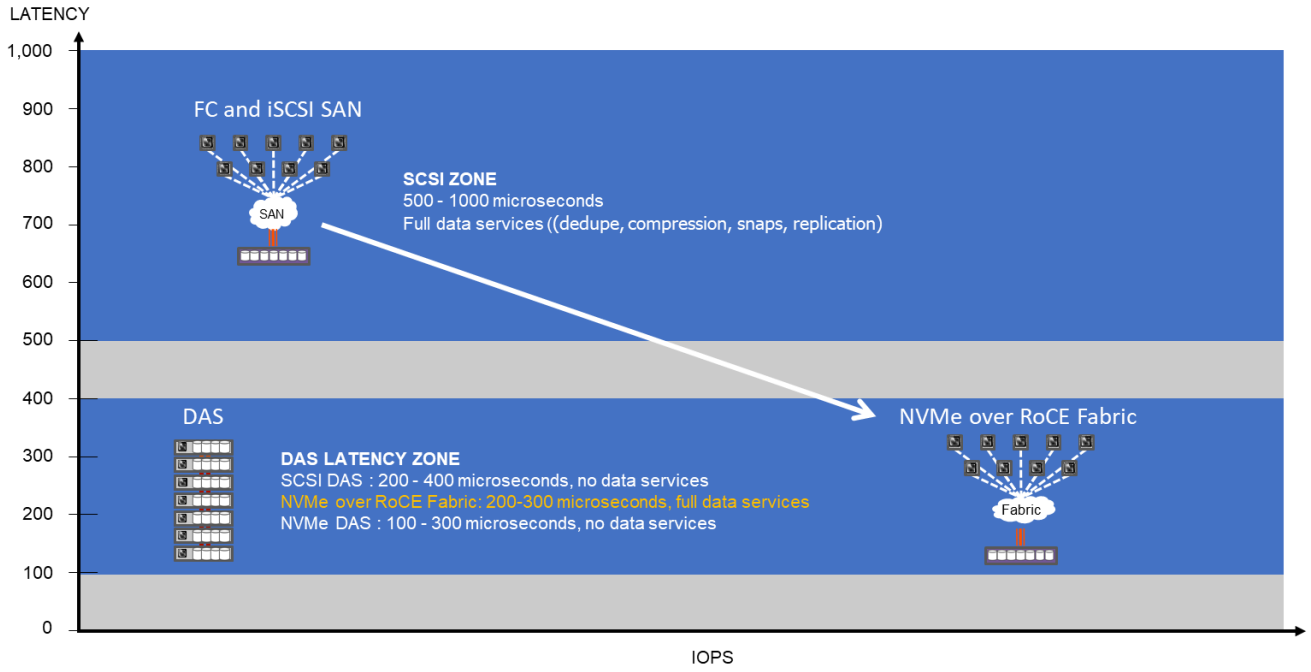
At the same time scale-out data center architectures emerged, all-flash arrays (AFAs) entered the market with SAS and SATA SSDs inside. AFAs were 100 times faster than disk arrays and quickly dominated the "high-performance" array market, relegating disk arrays to "high-capacity" applications. However, for extremely latency-sensitive scale-out apps, flash inside the server performed best. The culprits were common networks used to connect external all-flash arrays. They added somewhere from 10 to 100 microseconds of latency.

SAN vs. DAS Flash Storage Latency



2019: Networked Storage with the Low Latency of DAS

A new chapter in the history of networked storage is starting this year when NVMe over Fabric technology is available from major vendors like Pure Storage®. The pendulum will swing again from DAS back to shared SAN storage as this new class of super-SAN almost completely eliminates the difference in performance between local and networked flash storage.



MacStadium

They Support Software Developers Worldwide.

We provide the enterprise-ready storage platform that enables them to scale without disruption.

MacStadium is the leading Apple Mac hosting provider, supplying dedicated servers and private cloud hosting solutions in over 50 countries. MacStadium uses FlashArray//X to deliver an Enterprise 100% NVMe storage solution for its customers.

"The uniqueness of MacStadium and the performance and resiliency of Pure in a private cloud environment is just a great marriage."

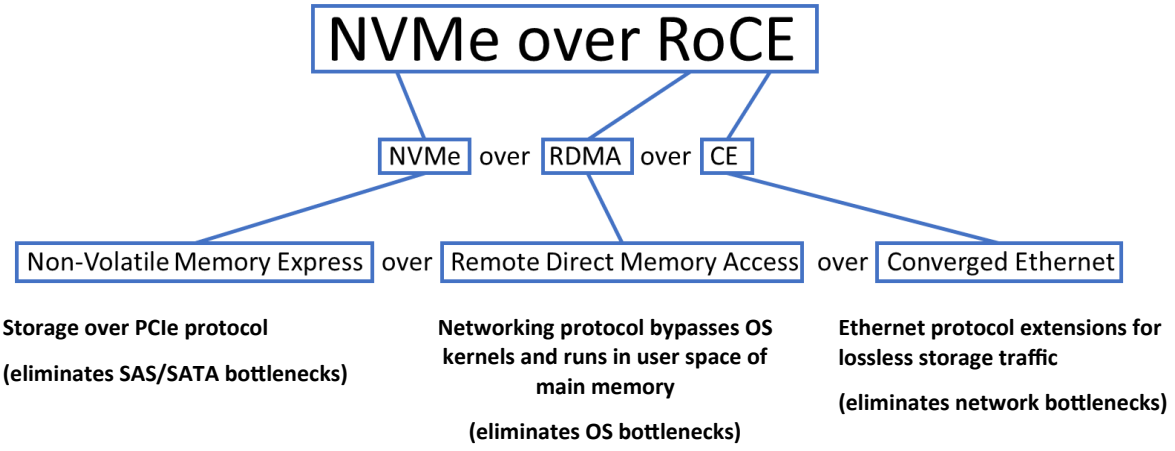
Greg McGraw,
CEO, MacStadium

SEE IT IN ACTION AT
www.purestorage.com/customers/macstadium.html

PURESTORAGE®

Chapter 2
Unraveling "Non-Volatile Memory Express over Remote Direct Memory Access over Converged Ethernet" (NVMe over RoCE)

NVMe over RoCE is the name of a storage networking technology and short for "Non-Volatile Memory Express over Remote Direct Access Memory over Converged Ethernet." Wrapped inside those 12 words and 27 syllables are 3 core technologies working together to deliver extraordinary microsecond access time to storage from anywhere in a data center.



Part 1 of 3: Non-Volatile Memory Express (NVMe)

NVMe is a protocol used to access storage on a PCI Express bus. Because the PCIe bus offers several high-speed lanes that can be used in parallel, SSDs with NVMe interfaces support 3x more IOPS than the previous generation of SSDs with Serial-Attached SCSI (SAS) interfaces.

The first NVMe SSDs started shipping inside servers in 2014 because PCIe busses were already there. In 2019, NVMe SSDs will have almost completely displaced SAS SSDs in high-performance enterprise applications.

One command queue.
32 commands per queue .

SAS/SATA interface
SCSI commands

PCIe interface
NVMe Commands

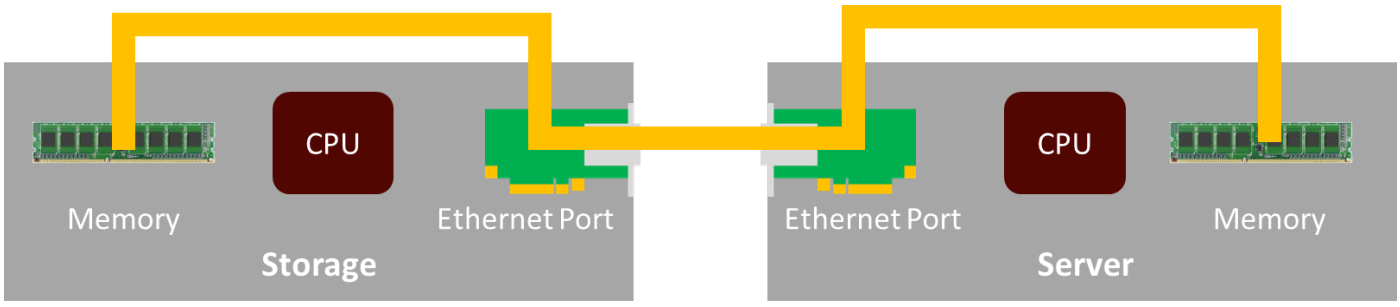
65,535 queues.
65,536 commands per queue.

Part 2 of 3: Remote Direct Memory Access (RDMA)

RDMA is a remote memory-management capability enabling server-to-server and server-to-storage data movement directly between application memory. Bypassing the CPU and operating system results in higher performance, lower latency, and lower CPU utilization.

InfiniBand RDMA is the original technology pervasive in the High Performance Computing industry where it connects the world's biggest and fastest server clusters. Now RDMA is being used in Converged Ethernet (RoCE) networks to eliminate performance-robbing layers of the stack.

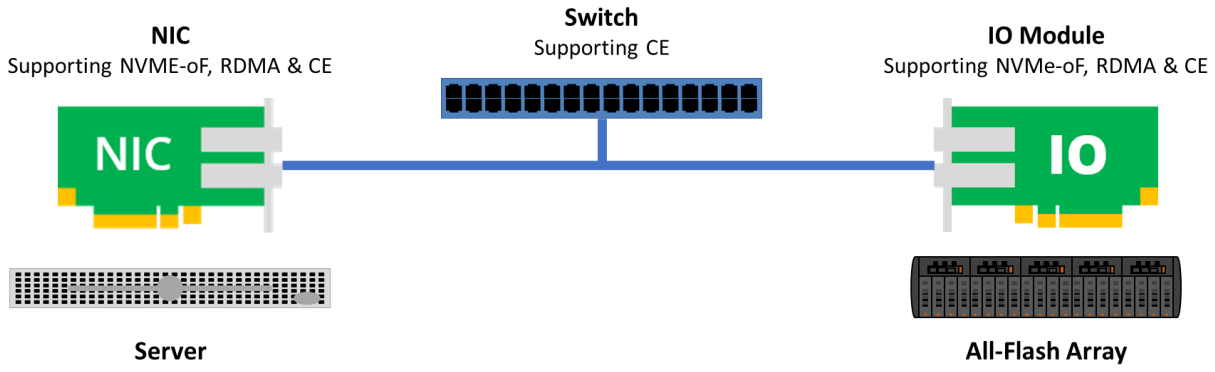
RDMA Accelerates Storage Networking Traffic by Bypassing the CPU and OS



Part 3 of 3: Converged Ethernet (CE)

Once the Ethernet industry started down the path of incorporating RDMA, the requirement for Priority Flow Control emerged to ensure zero loss of packets. The Ethernet industry responded by developing an enhanced version of Ethernet called Converged Ethernet (also known as Data Center Bridging, Data Center Ethernet, and Converged Enhanced Ethernet). Converged Ethernet includes Priority Flow Control which provides a link level flow control mechanism that can be controlled independently for each class priority to ensure zero loss, even when the network is congested.

CE, NVMe, and RDMA Support Needed for a RoCE Storage Fabric



In order to configure an NVMe over RoCE storage fabric, the NIC, switch and all-flash array must support Converged Ethernet. The NIC (sometimes called an R-NIC) and all-flash array must provide support for RoCE.

Putting it all Together into NVMe over RoCE Storage Fabrics

The three technologies discussed on the previous pages—NVMe, RDMA and CE—come together to form a new class of storage network which the industry is calling an NVMe over RoCE storage fabric. In many respects, all-flash arrays for NVMe-oF fabrics are an evolution of existing "SANs." At the core of new AFAs for NVMe fabrics are the same hardware and data services used with previous generations of Fibre Channel, Ethernet and InfiniBand arrays.

NVMe over RoCE being another class of SAN is worth mentioning because one of the primary benefits of NVMe fabrics is they allow IT organizations to match the performance of flash storage inside servers while providing the same enterprise data services, efficiency, and flexibility of SAN hardware and software.

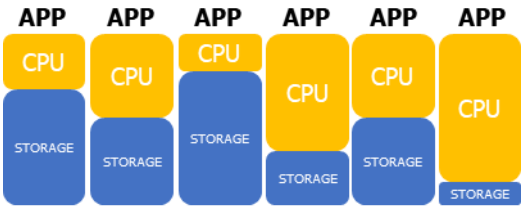
Consolidating DAS with NVMe over RoCE Fabrics will Improve Storage Utilization

The Problem with DAS

Unreliable

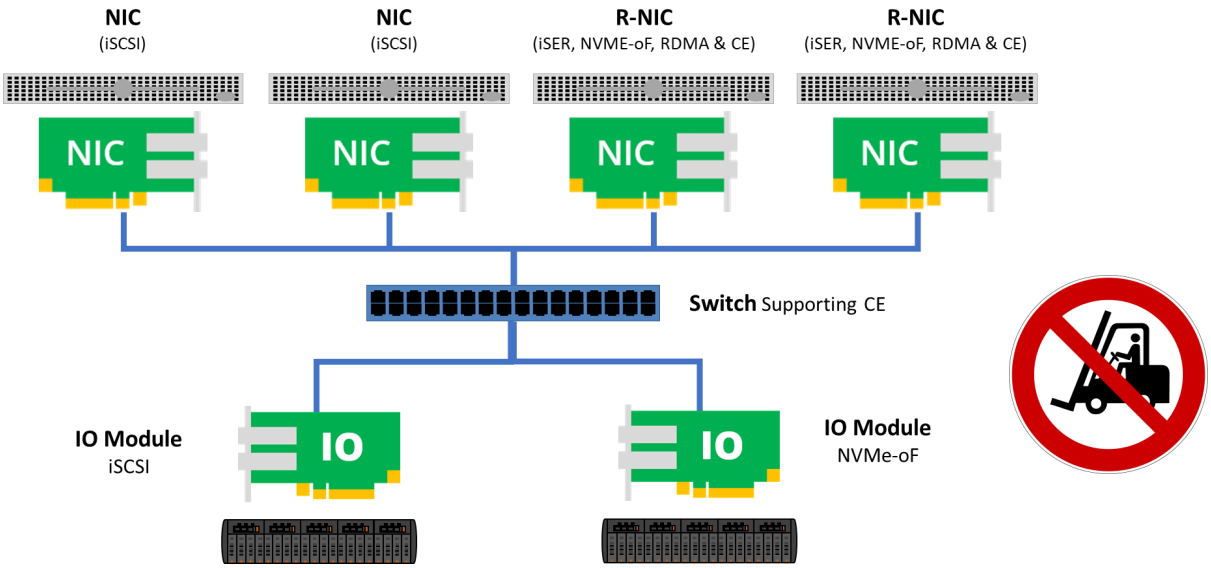
No Data Services

Inefficient Scaling



Another benefit of NVMe-oF is that unlike implementing a DAS architecture, there is no need to rip-out your SANs to accommodate scale-out apps. NVMe fabric segments can be added to your existing SAN, share the same data services, while simultaneously delivering the performance of local flash storage.

No Forklift Upgrades: SANs and NVMe over RoCE Fabrics Co-Exist in Super-SANs



Servers with R-NICs and all-flash arrays with NVMe over RoCE interfaces will plug-and-play with installed CE switches.

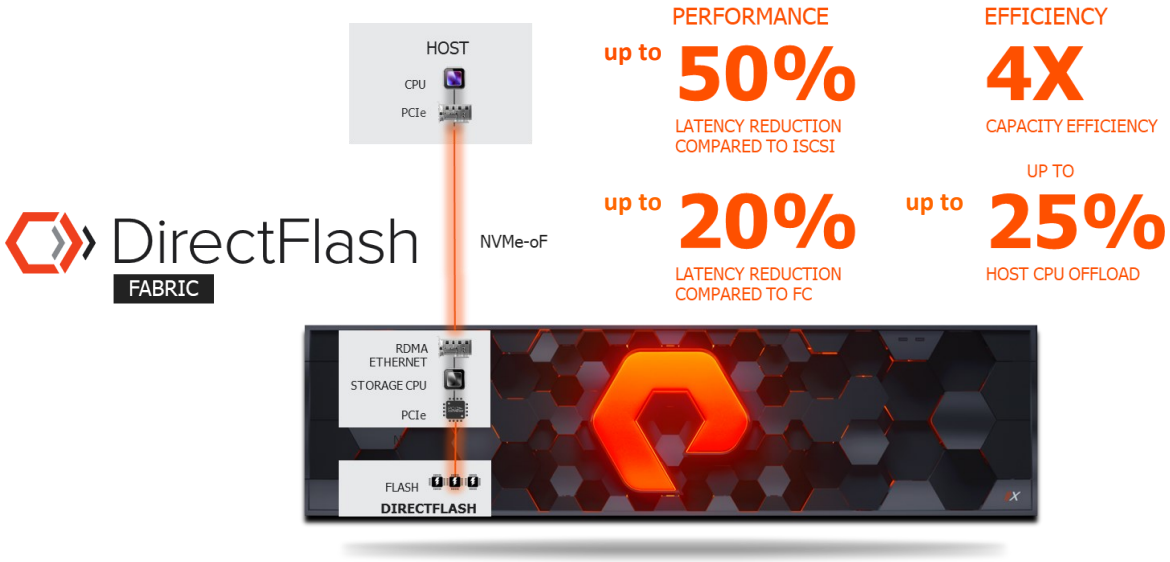
Chapter 3
The Industry's 1st Mainstream NVMe over RoCE All-Flash Array

Pure Storage DirectFlash Fabric

The Pure Storage FlashArray™ family delivers software-defined, all-flash, 100% NVMe power and reliability from the entry-level FlashArray//X10 to the FlashArray//X90, all with end-to-end NVMe and NVMe-oF performance and enterprise class data services.

The Purity Operating Environment software within every FlashArray model enables organizations to provide the highest levels of business continuity with ActiveCluster™ along with proven 99.9999% availability, completely non-disruptive operations, and AI-driven, white-glove support.

For customers who need the full benefit of NVMe, Pure offers DirectFlash Fabric. This feature provides support of NVMe-oF RoCE with full data services with FlashArray//X, Purity Operating Environment 5.2 and RDMA-enabled NICs.



Future Proof, Ready When You Are

Pure Storage has always created products that capitalize on advances in storage and flash media. When FlashArray//M was launched in 2015, Pure engineered a product with future technologies into the product - including readiness for NVMe.

It's Evergreen technology and business model meant customers could benefit first from NVMe within NVRAM and then from 100% NVMe in DirectFlash Modules, with optimized DirectFlash Software to globally manage the flash for optimum performance and efficiency. With every innovation, customers could upgrade from prior generations of product, completely non-disruptively, protecting their investments and avoiding forklift upgrades and data migrations.

PURESTORAGE®

IDC TECHNOLOGY SPOTLIGHT

WHAT'S NEXT IN ENTERPRISE STORAGE?

NVMe AND NVMe-oF FOR ALL

The primary Flash Market is Evolving to Next Generation Architectures.

WHAT DOES THE FUTURE HOLD?

As the IT industry enters the cloud era, next-gen flash-driven storage architectures will be needed to enable unprecedented levels of performance and density. Download this IDC report to understand how the market is evolving to 100% NVMe-oF and software driven flash and how Pure Storage plays in this market.

DOWNLOAD THE IDC REPORT AT

www.purestorage.com/content/dam/purestorage/pdf/AnalystReport/IDC/IDC_for_X.pdf

IDC

IDC TECHNOLOGY SPOTLIGHT

Primary Flash Market Evolving to Next-Generation Architectures

April 2017

Adapted from *Worldwide All-Flash Array Market Shares, 3Q16: Top Five Market Share Holders Consolidate Gains*, by Eric Burgener, Liz Conner, and Jingwen Li, IDC #US42374817

Sponsored by Pure Storage

As the information technology (IT) industry enters the cloud era, where hybrid IT is the dominant deployment model in organizations of all sizes, the capabilities of primary all-flash arrays (AFAs) will need to evolve to handle cloud scale and agility. Datacenters of all types and sizes will be using AFAs as general-purpose storage platforms more and more and will be increasing the workload densities that these systems must support. A key challenge is that current-generation AFAs have fundamental bottlenecks that both limit their consolidation and cloud-scale capabilities and create architectural risk as future memory technologies emerge. A next generation of AFAs — what could be called cloud-era flash arrays — will be best positioned to meet these requirements cost-effectively, will employ flash-driven rather than flash-optimized architectures, and will have NVMe rather than SCSI technology at the heart of their designs. This Technology Spotlight examines the evolving primary flash array market with a particular emphasis on what next-generation flash-driven enterprise storage architectures will look like. It also looks at the role Pure Storage, with its FlashArray//X, plays in this strategically important market.

Introduction

As flash storage has permeated mainstream computing, enterprises are coming to better understand not only its performance benefits but also the secondary economic benefits of flash deployment at scale. This combination of benefits — lower latencies, higher throughput and bandwidth, higher storage densities, much lower energy and floor space consumption, higher CPU utilization, the need for fewer servers and their associated lower software licensing costs, lower administration costs, and higher device-level reliability — has made the use of AFAs an economically compelling choice relative to legacy storage architectures initially developed for use with hard disk drives (HDDs). As growth rates for hybrid flash arrays (HFAs) and HDD-only arrays fall off precipitously, AFAs are experiencing one of the highest growth rates in external storage today — a compound annual growth rate (CAGR) of 26.2% through 2020.

Already, AFAs drive over 70% of primary storage spend, and over the next year, 76% of those organizations that do not already use an AFA in production have plans to evaluate and/or deploy one. AFAs are widely used for the mixed consolidation of primary workloads, with 47% of deployed AFAs hosting from 5 to 9 workloads and 36% hosting 10 or more. This number is expected to increase precipitously as organizations continue to retire aging storage equipment and move those workloads to AFAs over the next 12 months to gain the benefits of flash deployment at scale.

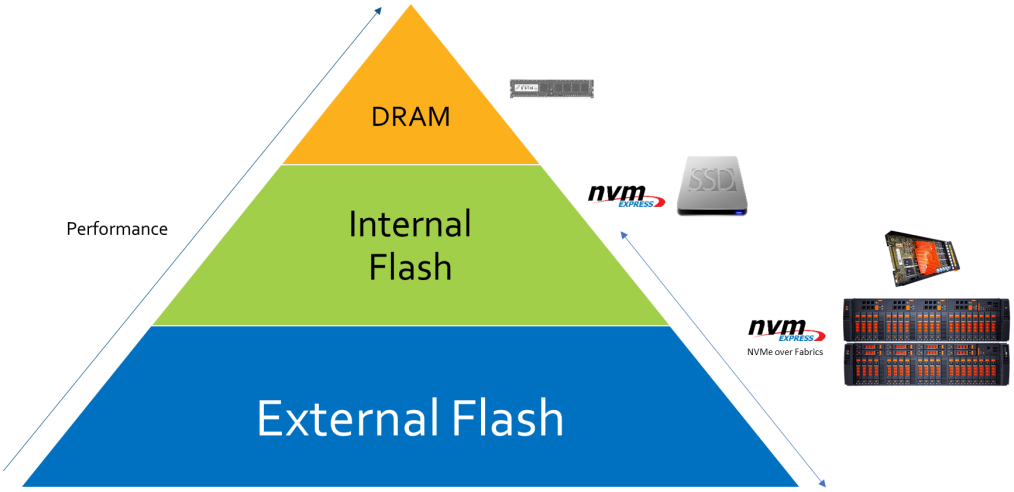
US42424617

Chapter 4
What NVMe over RoCE Means to You

A New Data Center Storage Hierarchy

NVMe over RoCE storage fabrics promise to disrupt the data center storage hierarchy by pushing into a space previously reserved for flash storage inside servers. Consolidating DAS into NVMe over RoCE fabrics will become a best-practice because they will meet comparable performance service levels at a much lower cost of ownership.

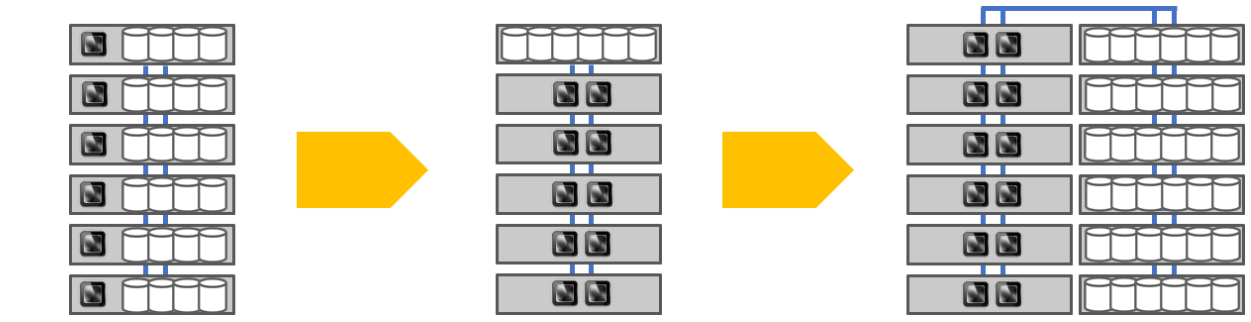
All Flash Arrays Become a Favorable Alternative to Flash Inside Servers



Dis-Aggregation of Servers and Storage

For IT organizations that have implemented a rack-pod architecture, all-flash arrays with NVMe over RoCE interfaces can be deployed to support a group of servers in a rack. For even greater efficiency, a pool of all-flash arrays can be deployed and provisioned to support a large server farm. In either case, dis-aggregated storage accessed over a over a high-speed NVMe over Fabrics network will provide comparable levels of performance as local flash storage.

Dis-Aggregation: Storage No Longer Needs to be Next to Servers



Storage & servers aggregated

Storage & servers dis-aggregated at rack level

Storage & servers dis-aggregated at data center level

Chapter 5
Killer Apps for NVMe Storage Fabrics

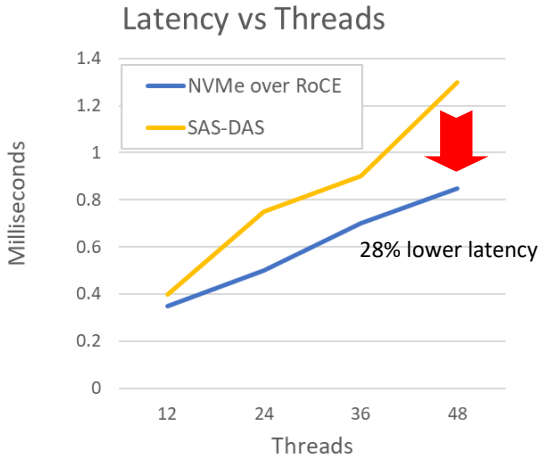
Apache Cassandra is a free and open-source, distributed, wide column store, NoSQL database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure.

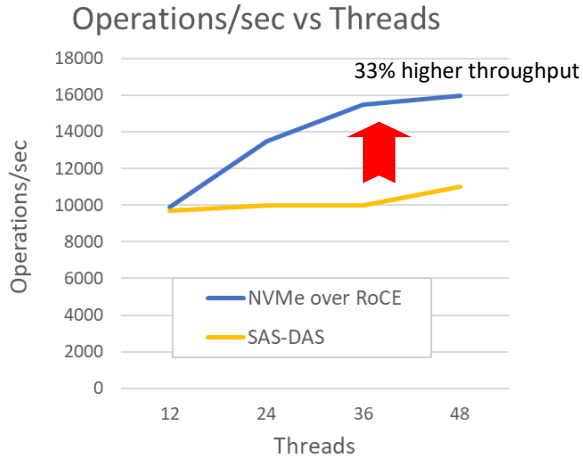
One major benefit of Pure Storage FlashArray//X for Cassandra is data reduction. For database workloads, the typical data reduction achieved via always-on deduplication and compression is 3:1. This results in significant storage footprint savings, and with non-disruptive capacity upgrades, there is no need to pre-purchase loads of underutilized capacity, as you would have to do with DAS.

Cassandra also offers native snapshots. Most customers cannot take more than 3 native Cassandra snapshots before they run out of storage space. With Pure Storage FlashArray//X , snapshots are instantaneous, do not consume much space, and you can take thousands of snapshots with FlashArray without any issues.

Flexible Pure Storage solutions are available with NVMe-oF RoCE that provides performance equivalent to DAS, but with full enterprise data services. Below are test results of a 3 node Cassandra cluster, with replication factor set to 3 for keyspaces. The results show Pure FlashArray//X with DirectFlash achieves up to 30% lower latency and just under 30% greater operations/sec vs SAS SSD - DAS.

NVMe over RoCE up to 30% Lower Latency and Higher Throughput than SAS - DAS





Killer Apps for NVMe over RoCE Storage Fabrics

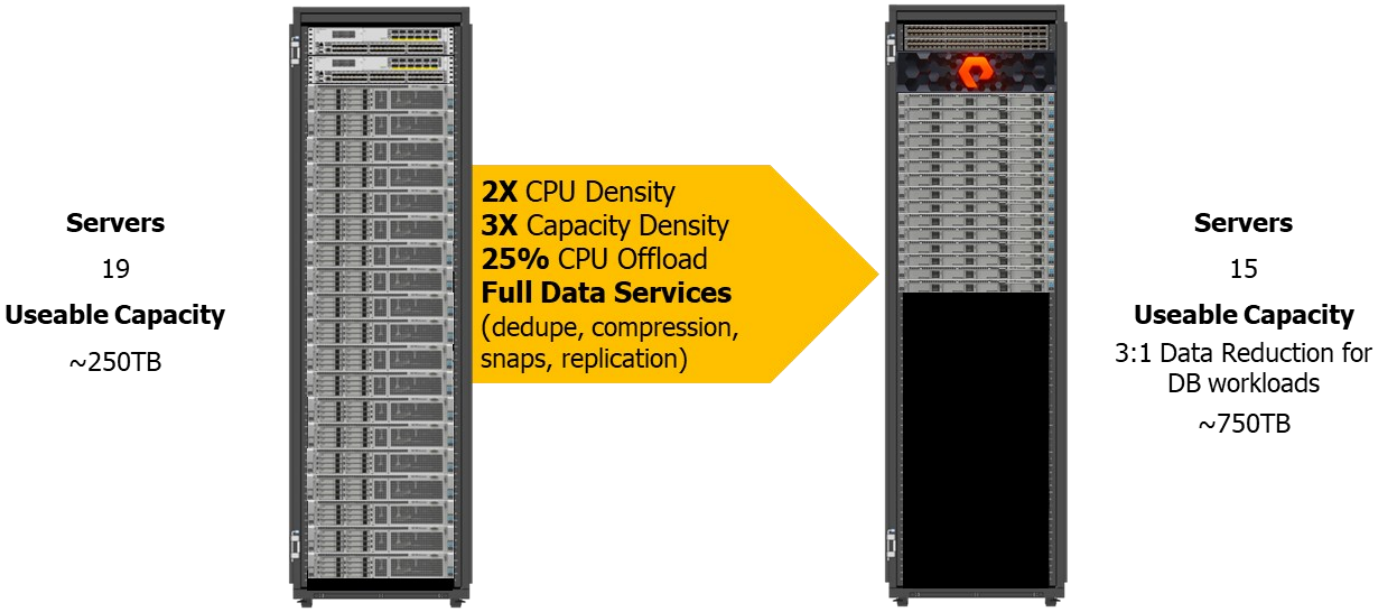
MariaDB is a community-developed, commercially supported fork of the MySQL relational database management system, intended to remain free and open-source software under the GNU GPL. Development is led by some of the original developers of MySQL.

Several Linux and BSD distributions now include MariaDB, while some default to MariaDB, such as Arch Linux, Manjaro, Debian, Fedora, Red Hat Enterprise Linux, CentOS, Mageia, openSUSE , SUSE Linux Enterprise Server, OpenBSD, and FreeBSD.

FlashArray//X with DirectFlash Fabric—tested with MariaDB and HammerDB—resulted in 30% more transactions per second when compared to SAS SSD - DAS.

DirectFlash Fabric delivers dramatically higher CPU density, useable capacity, and throughput as threads scale, as well as enterprise data services.

Higher CPU Density, Higher Capacity, Full Data Services



Killer Apps for NVMe over RoCE Storage Fabrics

MongoDB is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemata. MongoDB is developed by MongoDB Inc. and licensed under the Server Side Public License (SSPL).

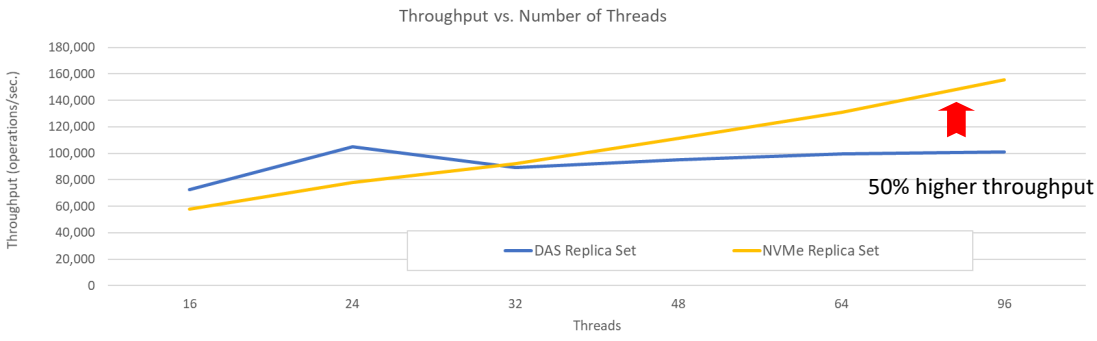
Pure Storage FlashArrays and DirectFlash Fabrics are perfect companions to support MongoDB when the working set no longer fits in the memory and spills over to the storage sub-system which still requires low latency and consistently high performant response times.

Testing by Pure Storage with MongoDB in a multi-node replica set revealed the highest throughput was accomplished with higher number of nodes which is a validation for the scalability aspect that is very critical for NoSQL databases like MongoDB. The tests also validated the Pure Storage architecture enables sustained low latency, very close to DAS, but with full data services.

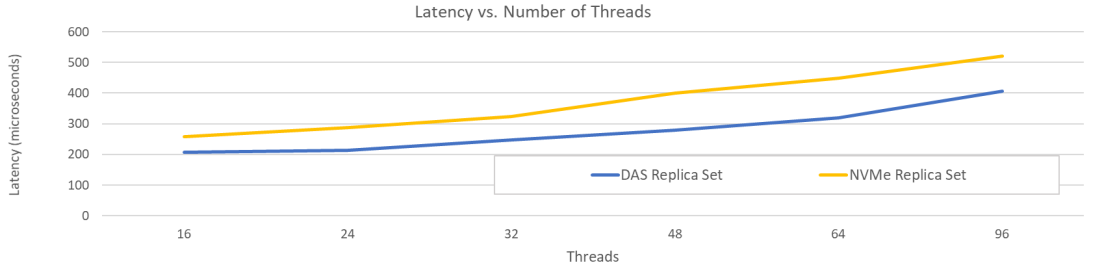
MariaDB is being used in increasingly larger clusters with their Galera functionality. As these clusters grow in size a key benefit Pure Storage is able to provide is universal storage access for all hosts, minimizing per host management.

The Cold Hard Numbers

DirectFlash Fabric and NVMe over RoCE delivers more MongoDB operations per second versus DAS as you scale



NVMe over RoCE with full data services near DAS read latency with no data services



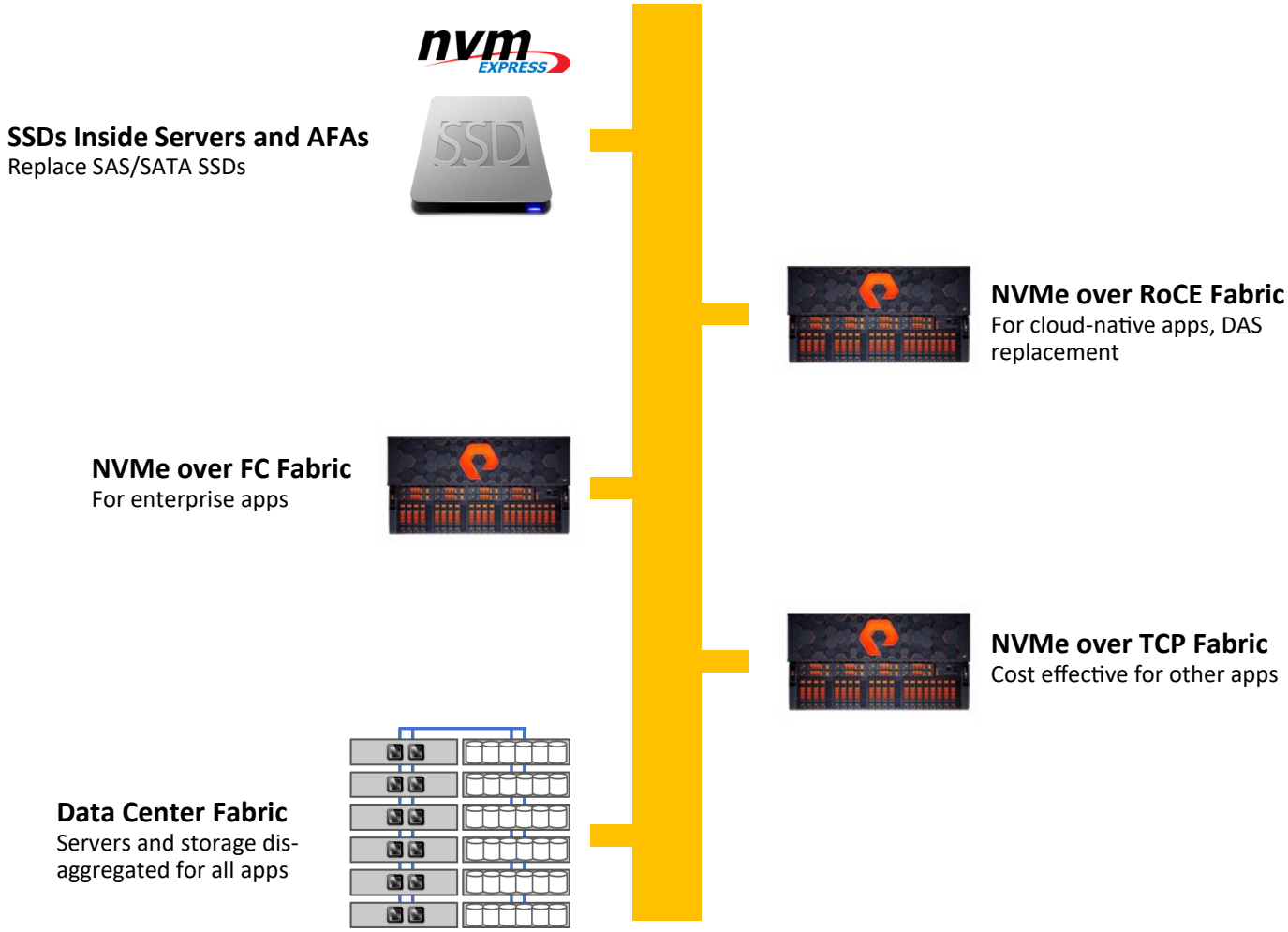
Chapter 6
What's Next

Your applications can benefit from NVMe technology in multiple ways. The following is a suggested road map for implementation that ends with multiple NVMe solutions delivering data to every server in your data center.

First, if your application infrastructure needs modernization and/or applications are not delivering the results your business demands, investigate NVMe all-flash arrays with NVMe capability. While some vendors charge more for these features, Pure includes 100% NVMe with no additional premium compared to SAS SSD options, and now offers NVMe over Fabrics for end-to-end NVMe benefit.

Second, the decision on what type of NVMe over Fabrics protocol to deploy, depends on the applications in your data center and the readiness of those applications for NVMe over RoCE, NVMe over FC, or NVMe over TCP. Pure Storage will support all of these protocols, but is starting with RoCE as they believe the ecosystem support for RoCE is ready now for Linux and applications such as MongoDB, MariaDB, and Cassandra.

NVMe Storage Fabric Implementation Road Map



Chapter 7
Summary

NVMe over RoCE is a Matter of "When" Not "If"

If flash storage is not an important part of your data center infrastructure, the availability of NVMe over RoCE is a non-event. But if you're in the camp where flash storage plays an important role today, and sees the percentage of flash storage in the data center growing, then NVME and NVMe over RoCE is a matter of when, not if.

Sooner is better than later if application performance is a competitive weapon for your organization. Two-thirds of all-flash arrays shipped in three years will include one or more NVMe-oF interfaces. That means your competitors are already at some phase of investigating, evaluating or deploying the technology.

Resources

Analyst report: 2018 Gartner Magic Quadrant for Solid State Arrays

Video: FlashArray//X Shared Accelerated Storage

Video FlashArray//X How It Works

Datasheet: FlashArray//X

Product Information: FlashArray//X 100% NVMe Block Storage

Blog: Pure Brings Hyperscale Architecture to the Enterprise

Blog: Pure Delivers DirectFlash Fabric: NVMe-oF for FlashArray

Blog: Analyzing the Possibilities of MariaDB and DirectFlash Fabric

PURESTORAGE®

Gartner

Three major industry reports.

One thing in common.

See where Gartner and our customers said we stand in 2018.

SEE THE REPORTS AT

www.purestorage.com/resources/type-a/gvartner-ssa-recognition/thank-you.html