architectingIT

NVMe in the Data Centre

A User's Guide to Technology and Solutions

Release 2.0

Chris M Evans

NVMe in the Data Centre Second Edition

Published by Brookend Limited, January 2019.

Document reference number BRKWP0120.

NVME IN THE DATA CENTRE

CONTENTS

Background

Information Technology has an insatiable desire for data, which for today's businesses has become a hugely valuable asset and piece of intellectual property. As the speed of technology increases, getting data into applications for processing has become more critical than ever. Storage has always been a bottleneck compared to processor and memory speeds, so achieving the highest level of performance from storage systems and devices has never been so important.

Unfortunately, we live with a legacy of storage communications that were designed nearly 40 years ago. SCSI[1], the protocol that drives much of our storage today, was developed in the late 1970s. SATA is based on a PC storage standard from the 1980s. In both cases, the performance and efficiency of these protocols are becoming an issue, as new media starts to take over the data centre. Both SCSI and SATA were designed for the age of spinning disks, where latency was measured tens of milliseconds. Today, flash storage operates at around 100µs (microseconds), with new storage-class memory[2] solutions delivering at orders of magnitude faster (around 10µs for 3D XPoint™ or better for MRAM technologies).

Storage Networking

Of course, the performance of individual drives is just one aspect, however we've also been implementing storage networks for at least 20 years. Most recently this has been based on Fibre Channel or Ethernet (iSCSI and FCoE) technologies. Prior to that, mainframe systems implemented ESCON which developed into FICON as part of Fibre Channel. All of these storage networking technologies introduced operational benefits in consolidating storage, improving resiliency, reducing stranded resources and improving performance.

The performance gains came from the introduction of the Integrated Cached Disk Array, a storage appliance combining disk drives and DRAM to accelerate read and write performance - of which EMC Corporation was a pioneer with the Symmetrix platform. In the 25 years since these systems were introduced, performance has increased dramatically, through the use of faster disks and now solid-state media such as NAND flash.

Storage Area Networks (SANs) are now proving to be too slow for some applications. As a result, new architectures are being developed that disaggregate storage and use new networking solutions. We're also seeing the adaption and improvement of existing storage fabrics such as Fibre Channel, addressing some of the legacy issues seen in this technology.

NVM Express

As a replacement for SCSI and SATA, the storage industry has developed a new protocol called Non-Volatile Memory Express, usually shortened to NVMe. This is also sometimes written as NVM Express, especially when referring to the industry body.

NVMe is a replacement for SCSI (and indirectly ATAPI, the protocol behind SATA) both for individual drives and for storage networking. NVMe has been designed for low-latency solid-state media, fixing many of the bottlenecks seen with legacy protocols. Although NVMe won't entirely replace SAS/SATA, it is envisaged that NVMe will gradually become the protocol of choice over time, with legacy protocols retained for the use cases where NVMe is less appropriate.

The Scale of The Problem

So how big a problem are today's legacy protocols? The Architecting IT blog post *"Performance Analysis of SAS/SATA and NVMe SSDs"*[3] provides some additional context. This post examines an academic paper[4] that studies the end-to-end I/O profile for common media, including SATA HDDs and SSDs.

---

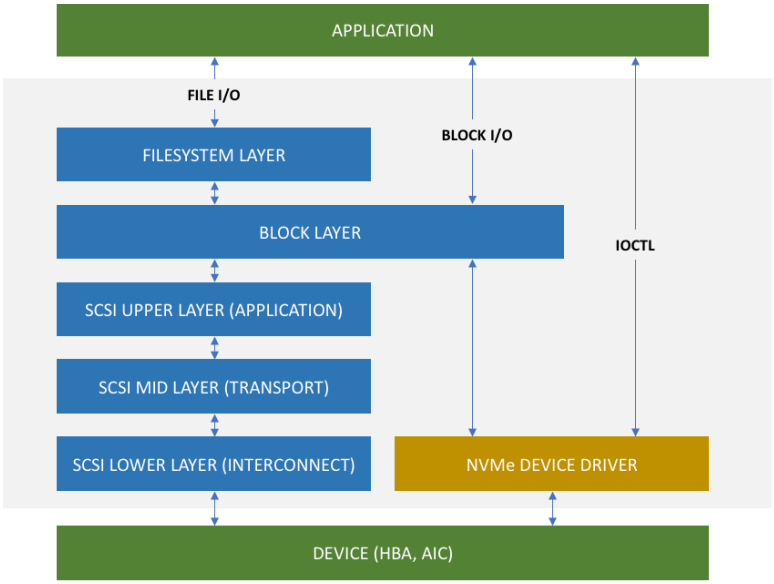
[1] https://storageunpacked.com/2018/11/74-all-about-sas/
[2] https://blog.architecting.it/what-are-scm-and-pm/
[3] https://blog.architecting.it/performance-analysis-sas-sata-nvme/
[4] Performance Analysis of NVMe SSDs and their Implication on Real World Databases

With legacy SATA HDDs, the time spent in software accessing a drive is only a very small proportion of the overall I/O transaction time. Almost all the time seen is in waiting for the mechanical media to respond. Hard drives have to re=position the read/write head on the platter, while waiting for the drive to rotate to the starting position on a track where data will be written or read. Compared to the time taken to access the drive, this mechanical movement represents eons of additional latency.

With SATA SSDs, time in software (processing the I/O request) is a much greater proportion of the I/O transaction time, because NAND flash is orders of magnitude faster than a hard drive. As a result, improvements in media performance have a much more diminished effect on the overall I/O time. Moving to storage class media, for example, would see little benefit in I/O latency and make it hard to justify the increased media costs.

NVMe aims to optimise the I/O software stack, partly through software changes and partly through implementing a more direct connection for storage and the CPU (more on this later).

Figure 1 - Simplified Linux I/O Stack



As the paper referenced in the blog post shows, reducing the I/O response time results in both better application performance and an increase in utilisation for the processor. Improved CPU utilisation can have financial benefits too, as many software applications are licensed by processor cores.

Figure 1 - Simplified Linux I/O Stack shows how NVMe simplifies the I/O path compared to SCSI. Note this image represents NVMe within a server, not the implementation of NVMe over Fabrics.

NVMe Overview

NVM Express is a specification that defines a set of commands to enable communication between the processor in a server and persistent storage media connected to the PCIe bus. As NVMe develops, the specification is being extended to introduce enhanced management features and the ability to support a range of platforms, from mobile to the enterprise data centre.

The development of NVM Express standards is managed by an industry body called NVM Express Inc[5], incorporated in 2014 and run by the industry, currently with over 100-member companies. NVM Express Inc was previously known as the NVM Express Work Group and NVMHCI Work Group before that.

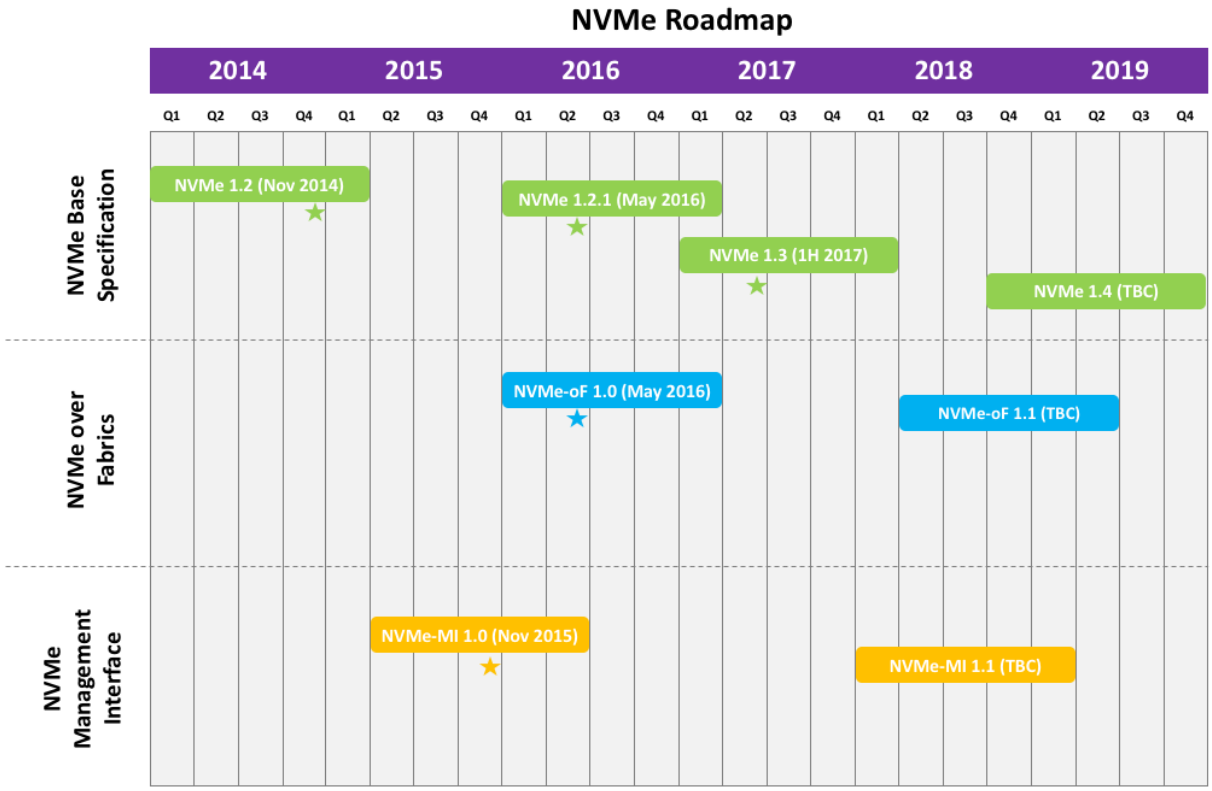


Figure 2 - NVMe Standards Roadmap

The first NVMe base specification 1.0 was released on 1 March 2011 and covered connectivity of PCIe-connected storage devices in a server or PC. Version 1.1 was released on 11 October 2012 and version 1.2 on 3 November 2014. The NVM Express consortium is actively developing 1.3 of the specification, which was completed in 1H2017. The latest version addresses the requirements of NVMe for mobile devices (including low power consumption). Work on release 1.4 is in progress, covering issues such as I/O determinism and multipathing.

The specifications for out-of-band management of NVMe devices is defined under the NVMe-MI (Management Interface) specification. This was first released in November 2015 and continues to be developed to address in-band management and enhanced discovery requirements.

NVMe over Fabrics

As a parallel to existing storage networking technology, work on the NVMe over Fabrics (NVMe-oF™) specification was begun in 2014, with the first release completed in 2016. The specification provides the capability to use NVMe outside of a PCIe bus, using fabric topologies that include Fibre Channel, Ethernet, TCP and InfiniBand. The aim of NVMe-oF is to enable access to both individual NVMe drives and storage systems.

[5] https://nvmexpress.org/

Benefits of NVMe

NVMe provides significant I/O performance and reduced latency compared to legacy protocols like SAS and SATA. The design places storage physically closer to the processor and improves protocol efficiency, reducing the length of the "I/O stack". NVMe devices sit directly on the PCIe bus, which offers higher bandwidth and lower latency than using storage controllers. This is contrast to SAS/SATA drives, that have to be connected to a storage controller, which is, in turn, generally connected via a PCIe interface.

NVMe is optimised to reduce latency in the software I/O path for storage. This includes improving the way in which NVMe devices are managed and by increasing the level of parallelism. As SAS and SATA protocols were designed for slow mechanical media, only a single I/O queue was used to manage read/write requests. NVMe introduces many more queues (65,535) with much greater queue depth (65,535 requests) compared to SAS and SATA that had approximately 253 and 32 requests respectively.

The ability to process I/O requests in parallel has significant benefits with solid-state media. Hard drives were restricted to the movement of the drive head, making it impractical to process multiple I/O queues. NAND flash and SCM devices however, have many back-end channels to write data to persistent media. The controller of an SSD can therefore write and read I/O in parallel, so it makes sense to have the ability to service I/O in the same way at the interface of the device to the host.

While parallelism doesn't necessarily decrease latency (other than when I/Os are queued behind each other) it does allow for much greater throughput, which we see reflected in the performance figures of NVMe devices. Typically, NVMe SSDs offer much higher bandwidth and an increased number of IOPS compared to SATA/SAS devices. As NVMe drives continue to grow in capacity, the ability to write across the device at the back end becomes important in continuing to manage performance. We will discuss more on this later.

Positioning with SAS/SATA

SAS and SATA are older protocols that were developed in the age of spinning media. SAS (Serial Attached SCSI) originates from the SCSI protocol, a parallel hard drive interconnect for servers from the late 1970s. SATA (Serial ATA or Serial Advanced Technology Attachment) derives from PATA (the parallel version of the protocol), which originated in the IBM PC architecture. The origins of SAS/SATA mean the protocols are not well suited for highly parallel access media such as NAND flash and storage-class memory products. However, both protocols still have a place in providing connectivity for hard drive media, sequential access devices and for slower, cheaper solid-state storage devices.

It's envisaged that both SAS and SATA will remain in use for some time, declining in use as NVMe matures and becomes more prevalent on servers. For storage arrays, SAS is currently used as the most popular back-end interconnect interface and protocol. This is likely to be the case for some time, as the interface is mature and well-understood by array manufacturers. Over time, as the cost of transforming to NVMe lowers, it is anticipated that vendors will begin the transition to NVMe. You can find more on this process in the review of individual vendors later on in this document, where we discuss vendors that have already made the transition to back-end NVMe.

NVMe Deep Dive

NVM Express or NVMe is a storage protocol. It enables the communication with storage devices and storage systems across either a PCIe bus or fabric. NVMe can be seen as both a successor and complementing existing SAS and SATA protocols. As already discussed, NVMe was needed to address the shortcomings of traditional protocols where the largest impact on I/O response time was the speed of mechanical media.

Physical Connectivity

NVMe devices sit on the PCIe bus within a server, PC or mobile device. Today, most products on the market support PCIe version 3.0 with four lanes for data transfer (typically written as PCIe 3.0 x4). A lane is made from a couple (pair) of serial point-to-point links. Devices that use multiple lanes will stripe data across the lanes to increase throughput. A PCIe 3.0 x4 device can support up to 3.94GB/s in each direction across the bus.

The first NVMe devices connected directly onto a PCIe motherboard slot of a server using the AIC (Add-in-Card) format. As we will discuss, the range of form-factors for NVMe drives has increased significantly, as the restrictions set by spinning disk media are removed.

Deployment Models

Today in the data centre we can see NVMe being deployed many different scenarios. Most obvious is the use of NVMe devices as DAS (Direct Access Storage) within servers, personal computers and laptops. This has led to a number of new form-factors, which we'll discuss in a moment.

Storage array vendors are also making the move to NVMe. In terms of array design, we see two approaches. The first is to use NVMe internally, connecting persistent media to internal controllers that improve the overall array performance. The second is to use NVMe across the network, reducing the latency of storage networking and again, improving performance.

The term "End-to-End NVMe" has started to emerge as a definition of storage platforms that support front-end NVMe connectivity, with back-end device connectivity. In this architecture, there are expected to be no other storage protocols to get in the way and reduce performance. More background can be found in the Architecting IT blog post *"The Race towards End-to-End NVMe in the Data Centre"*[6].

Form Factors

NVMe has seen the introduction of a range of new form factors, as vendors producing solid-state products are not encumbered by the dimensions of traditional hard drives. Until the adoption of NVMe, SAS and SATA solid-state disks typically mimicked the existing 2.5" and 3.5" drive specification. Initially this made sense, as SSDs could simply be used as a drop-in replacement for HDDs.

Solid-state media is defined by the size of the die (memory chips) used to build the device. Alternative form-factors can offer improvements in density and better heat flow than legacy 2.5" media. As a result, vendors have developed a range of form factors that work in personal computers, laptops, tablets and standard servers.

[6] https://blog.architecting.it/end-to-end-nvme/

AIC

AIC (Add-in-Card) was the original form factor for NVMe devices. In fact, AIC SSDs existed before the NVMe standard was established. Vendors including Fusion-io produced devices that required proprietary host drivers to make the hardware appear as a local drive. AIC devices were the main form factor for NVMe SSDs for many years, offering the largest capacity and highest performance.

Today AIC devices use PCIe 3.0 x4 connectors, as seen in the Intel AIC SSD shown here.

M.2

M.2, originally known as NGFF (Next Generation Form Factor) is an NVMe device that looks similar in size to a memory DIMM and is sometimes referred to as a "gumstick".

The difference compared to a DIMM is in the placement of the PCIe connector at the end of the device. M.2 comes in a range of formats, the most popular of which is 2280 (22mm wide by 80 mm long). However, width dimensions range from 12-30mm and length from 16 to 110mm, depending on the type of device.

M.2 devices have a number of "key" positions, which denotes the place in the physical connector where a gap (or notch) is placed to ensure only the correct devices can be plugged into a motherboard socket. NVMe SSDs use the *M* key format (also known as socket 3), as shown in the Samsung EVO drives to the right.

As M.2 devices aren't hot-swappable, they are not used in scale-out enterprise platforms, but instead find uses as boot drives or secondary drives in servers and desktops. The M.2 format also supports SATA 3.0 and USB 3.0 protocols.

U.2

The U.2 format, previously known as SFF-8639, looks like a traditional 2.5" hard drive, albeit with an NVMe interface. The aim of developing to existing 2.5" and 3.5" standards was to provide backward compatibility with existing mechanical drive enclosures.

U.2 drives have the benefit of being hot-swappable and have the ability to be connected via cabling or a backplane rather than physically plugged into a slot on a motherboard. This makes them more suitable for scalable enterprise storage solutions where the drives can be placed in a more practical arrangement. This is the type of layout seen with NetApp AFF A800 and IBM FlashSystem 9100 appliances.

NF1

Previously known as NGSFF (Next Generation Small Form Factor) the NF1 format has been introduced by Samsung. NF1 is slightly larger than M.2 at 30.5mm wide, allowing two NAND die to be placed side-by-side on the device, as can be seen in the drive on the right.

This small change in dimensions should provide better device capacity density than M.2. Samsung has already released an 8TB NF1 drive that uses 16 dies mounted on both sides of the PCB. This is equivalent to the number of dies expected to be used within a standard 2.5" form-factor drive, making NF1 capacities comparable with U.2 and potentially larger than M.2.

More information can be found in the blog post *"Samsung Introduces 8TB NF1 NVMe SSDs"*[7]

Ruler

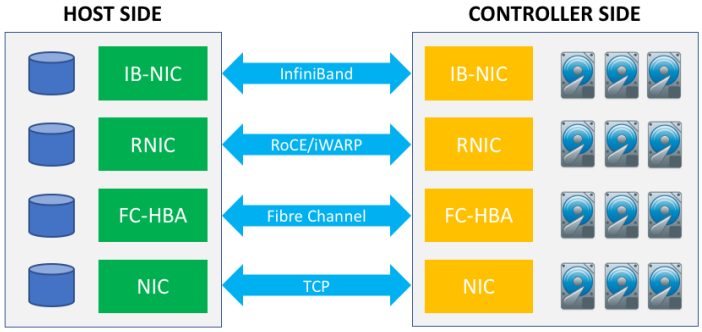
The Ruler[8] format (also known as EDSFF, Enterprise & Datacenter SSD Form Factor) has been introduced by Intel, Lenovo and others as part of the EDSFF Working Group.[9] The aim is to address the density and airflow issues of mounting SSDs in enterprise servers. Ruler offers better front-to-back airflow at high density within a server chassis and can more easily be accessed for maintenance, as devices are hot-pluggable, typically from the front.

There are currently form factors defined for both 1U and 2U servers, with orientations either horizontal (lying flat) or vertical (upright). Specifications define short medium and long versions of the drive. Intel has released a long form factor drive (E1.L) which today offers 8TB of capacity and can be packed in 32-wide in a 1U server.

Drive capacities are expected to increase quickly, with 32TB SSDs already being discussed by the industry. This would represent 1PB of capacity in just 1U of rack space.

NVMe over Fabrics

Where NVMe devices are connected to the PCIe bus within a server, NVMe over Fabrics (NVMe-oF) extends the protocol over a fabric or network. This can be thought of as being analogous to Fibre Channel physical networks that transport both the Fibre Channel Protocol (FCP), which is essentially SCSI, and FICON for mainframe.



[7] https://blog.architecting.it/samsung-8tb-nf1-nvme-ssd/
[8] https://blog.architecting.it/flash-capacities-failure-domains/
[9] https://edsffspec.org/

NVMe-oF is supported across a number of fabric technologies. Today this includes Fibre Channel, InfiniBand, RoCE (RDMA over Converged Ethernet), iWARP (Internet Wide-Area RDMA Protocol) and TCP. The latter uses standard Ethernet NICs as the physical transport layer.

FC-NVMe

FC-NVMe is the name given to NVMe implemented on Fibre Channel networks. As physical Fibre Channel is a transport medium that simply encapsulates higher-level storage protocols, NVMe can easily be supported on the latest Fibre Channel infrastructure. The FC-NVMe standard is owned and developed by INCITS (InterNational Committee for Information Technology Standards) that also manages the SCSI standard through the T10 sub-committee. The T11 sub-committee owns and develops both the Fibre Channel and FC-NVMe standards.

Figure 3 shows how Fibre Channel is implemented in a layered model similar to OSI. Data transmitted over a fibre channel network is broken down into frames.

FC-2 defines the framing protocol, with each frame (see Figure 4) broken down into a header and payload of up to 2112 bytes. The header determines the protocol "type", while the payload encapsulates both the NVMe commands and data and is defined in the FC-4 layer.

Separation of the frame transmission from the storage protocol allows multiple storage protocols to be carried across the same physical network at the same time.

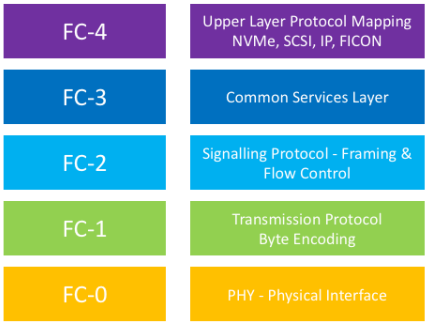


Figure 3 - Fibre Channel Protocol Layers

Figure 4 - Fibre Channel Frame

All Gen6 (32Gb) and most Gen5 (16Gb) Fibre Channel technology should support FC-NVMe, making it relatively simple for existing enterprises to make use of the new protocol. However, some components such as name services and multi-pathing are not fully implemented and only partially supported at the operating system level.

Why use Fibre Channel for NVMe? Many enterprises have heavily invested in Fibre Channel from the perspective of physical cabling and infrastructure, switching, HBAs, monitoring equipment, design and skills. As a result, this investment will have built up knowledge and understanding of how Fibre Channel works and more critically, how to fix things when they go wrong. FC-NVMe provides the ability to leverage that investment, without much effort in migration and with reduced risk. Enterprises can therefore gain immediate performance improvements and plan for Ethernet-based NVMe solutions if appropriate.

For the leading storage vendors, the greatest traction in moving to NVMe will be seen in FC-NVMe enabling existing products because this allows their install base to easily adopt NVMe. Therefore, we're likely to see FC-NVMe as the initial transition for many enterprises while Ethernet implementations are used more tactically.

NVMe over Ethernet & InfiniBand

Both Ethernet and InfiniBand provide support for NVMe, using protocol-specific host adapter cards. Vendor solutions include standard and bespoke RDMA NICs and InfiniBand adaptors, typically running at 25Gb, 40Gb and 100Gb speeds.

Ethernet solutions use RoCE, or RDMA over Converged Ethernet, which has two versions. RoCE v1 is a link layer protocol and so not routable. This means storage traffic would be limited to devices in the same Ethernet

broadcast domain. RoCE v2 is implemented on top of UDP and so can be routed. To date, vendors are favouring RoCE v2 as the choice of protocol for NVMe-oF implementations.

NVMe over TCP

The ability to use NVMe over standard TCP (NVMe/TCP) has been in development for some time and was ratified in November 2018. This additional transport binding will allow traditional Ethernet hardware to process NVMe requests and is analogous to the use of iSCSI over traditional Ethernet networks.

NVMe/TCP is important because it allows existing data centres with high-speed networking to implement NVMe-oF without additional hardware. All that's required is a software NVMe host driver. This could mean separating out management functionality to an IP network (which wouldn't have to be the fastest available) too.

In a software-defined data centre world, where cloud providers are looking to standardise on hardware, this can both reduce costs and provide the ability to logically map NVMe devices to virtual instances, rather than having connect physical NVMe drives as happens today.

Reworking Figure 1, we can see in Figure 5 - NVMe-oF Integration how NVMe, whether connected over a network or not and irrespective of the protocol, can be mapped to host block or file I/O without the application needing to be aware of the specifics of the implementation. This mirrors the way in which Fibre Channel and iSCSI LUNs are made visible to host in existing storage networking solutions.

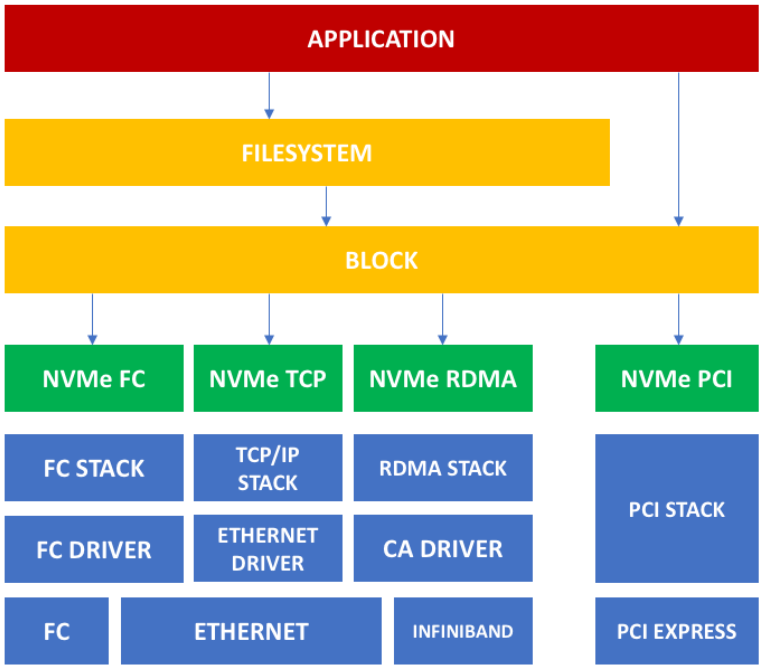


Figure 5 - NVMe-oF Integration

Vendor Solutions

Storage device and array vendors have been using NVMe in the data centre for some time. At the device level, there are products from all the major manufacturers, plus a host of smaller suppliers. At the systems level, vendors are starting to introduce NVMe as a front-end protocol and within systems to connect "back-end" storage media. The logical conclusion to this transition is systems that are fully "end-to-end NVMe"[10], with no legacy protocols being used.

NVMe is also making the move into HCI (Hyper-converged Infrastructure), as vendors look to gain the performance benefits of low internal device latency combined with high-speed networking. In HCI designs, storage is much closer to the processor compared to accessing across a network, reducing latency and improving workload performance.

Custom or Commodity?

The idea of using custom-built platforms or whether to use off-the-shelf commodity components has been a subject of much debate in the storage industry in recent years. Reliable componentry allowed vendors to deliver solutions that were purely software-based – commonly called software defined storage (SDS). This approach has significant cost advantages for the customer, as it separates the pricing of hardware and software components. Customers can source their own hardware, which for many reasons can be both more practical and cost effective. Software is simply added on as an additional cost.

As the drive for lower latency and high performance continues, there's a question to be asked as to whether commodity is still the best approach. Many vendors are bringing products to market that deconstruct traditional SSDs, instead choosing to build custom hardware centred around FPGAs and custom chip designs.

For these vendors, the benefit is quantified by the ability to achieve ever lower latency numbers. However, there may be another goal at play. We already know that public cloud hyperscalers don't buy traditional storage. Therefore, making products that the hyperscalers would buy can provide access to a huge market and potential acquisition.

Architectural Choices

NVMe introduces significant performance benefits over traditional drives. As a result, a single NVMe drive can easily soak up the bandwidth available with modern Xeon processors. The impact of this is seen in the design of storage systems that traditionally direct all I/O through one or more shared controllers[11].

Traditional storage architectures use one or more shared controllers because this provides a single point of control and management of metadata. The controller stores state information on how physical media is used and how logical devices are presented to connected hosts. With metadata in a single place, functionality like data protection (snapshots/clones) and data reduction (thin provisioning, de-duplication) are easier to implement.

In order to gain performance, much or all of the metadata is retained in DRAM[12]. Metadata needs to be kept in DRAM because performance of flash is so fast that any staging of metadata into DRAM to read or update it could add to the latency of read and write I/O. Unfortunately, this creates scalability limits on architectures based on dual controller configurations where DRAM is limited by processor count in the server.

Vendors have started to develop new architectures that remove some of the legacy architectural bottlenecks. We now see disaggregated solutions[13] that make drives more directly accessible to the host and push responsibility of metadata management and data services to those hosts.

Other solutions have returned to focus on hardware, implementing FPGAs and MIPS-based processors in their designs. Using hardware components in this way can provide a lower latency than implementing functionality in

[10] https://blog.architecting.it/end-to-end-nvme/
[11] https://blog.architecting.it/avoiding-all-flash-missing-iops/
[12] https://blog.architecting.it/flash-storage-dram-curve/
[13] https://storageunpacked.com/2017/09/garbage-collection-005-disaggregated-storage/

software, so reducing the impact on performance of the I/O stack. This hardware focus also applies to the use of fast networking and RNICs rather than traditional network cards. As a counter to the hardware approach, some vendors are staying purely software-focused, only selling hardware solutions as part of hardware vendor partnerships.

Benchmarks & Hero Numbers

NVMe has allowed system vendors to compete in delivering the fastest storage products available. Industry benchmarks are typically based on throughput at the lowest latency possible. This is the perfect challenge for NVMe. We're seeing records being broken on an almost weekly basis, as vendors tune their products to get the best scores possible.

How should we evaluate these so-called "hero" numbers? The first aspect to recognise is that performance typically follows a bathtub curve[14]. At either end of the scale, systems are great at all-read or all-write I/O. Some vendors use 100% read cache performance as a measure of the capability of their systems, which is clearly a false premise because performance derives from a workload mix of read and write I/O. With mixed workloads, there tends to be a drop in overall system performance, creating a typical 'bathtub" curve.

So, when looking at benchmarks, keep in mind that vendors are positioning their solutions in the best light possible. This includes building and tuning systems to get the best results possible. This isn't cheating but providing an indication of what could be achieved. More real-world data is achieved by looking at how platforms and systems cope with application workloads, such as databases or analytics. This is where we see the real-world capabilities of storage systems and can provide a better indication of how one solution can be better than another.

Of course, benchmarks and hero numbers are only one aspect to making a purchasing decision. It's important to have a good set of requirements before evaluating vendor products and to rate those requirements in order of importance during the evaluation process.

Enterprise Platform Vendor Analysis

This section covers vendors that have introduced NVMe into their products, either partly or fully as a replacement for existing protocols. The list of vendors included is based on those that, to date, have provided briefings on their products to a level of detail that provides sufficient insight into how the products are differentiated in the market. The list of vendors under analysis is continually reviewed and upgraded regularly.

Apeiron Data (ADS1000)

Apeiron Data Systems has developed a NVMe-oF storage solution using a proprietary networking protocol known as NVMe over Ethernet (NoE). The ADS1000 is a 2U appliance, housing up to 32 2.5" NVMe SSDs with 32 40GbE ports. The difference between ADS1000 and traditional shared storage is that the architecture doesn't use any shared controllers but instead connects drives directly with the host as the NVMe drives and controllers sit on the same PCIe midplane,

The vastly simplified architecture of ADS1000 means the performance overhead of the solution is as low as 2-3µs, allowing latency as low as 12µs for Optane media and 100µs for MLC NAND flash. Of course, the trade-off is in dispensing with data services, so features like data protection need to be added in at the host layer. Host servers must also use Apeiron proprietary HBAs that are configured for NoE.

Analysis. Apeiron has essentially developed a storage consolidation solution, taking drives out of host servers and providing the management benefits of a shared chassis. There are lots of scenarios where this will prove beneficial, but potential customers expecting a traditional storage array might be disappointed. NVMe over Ethernet is a proprietary protocol and we've seen that before with Coraid and ATA over Ethernet. This technology failed to catch on, possibly because of the lack of data services customers had come to expect from SAN storage.

[14] https://blog.architecting.it/avoiding-the-storage-performance-bathtub-curve/

Dell EMC (PowerMax)

Dell EMC recently upgraded their high-end enterprise products with the introduction of the NVMe-based PowerMax. As a successor to VMAX, PowerMax systems are all-flash and based around NVMe technology, although the platform retains a lot of VMAX heritage. PowerMax is also the first of the Dell EMC storage products to implement NVMe for persistent media.[15]

The PowerMax architecture uses the concept of a brick as the building block of a system. Each brick consists of dual directors, two DAEs (Disk Array Enclosure) and battery backup. Currently two models are offered – the PowerMax 2000, with one or two bricks and the PowerMax 8000 that scales up to eight bricks.

In making the transition to NVMe, Dell EMC has centralised the architecture around PCI Express. Each director within a brick is connected by PCIe to front-end I/O modules (supporting Fibre Channel, 10Gb Ethernet for iSCSI and FICON), back-end I/O modules, vault persistent memory (NVMe), new data reduction hardware and of course the other director in the brick. Bricks are cross-connected using InfiniBand (Virtual Matrix).

At the time of writing, PowerMax supports 1.92TB, 3.84TB and 7.68TB drive. This equates to a maximum raw capacity of 737TB for PowerMax 2000 and 2211TB for PowerMax 8000. Obviously actual usable and effective capacities will depend on the RAID implementation, spare drives and de-duplication ratios achieved. PowerMax systems are expected to ship in 2019 with storage class memory (Intel Optane) as a tier of storage.

In terms of performance, Dell EMC claims Power Max 2000 can achieve 1.7 million IOPS. PowerMax 8000 systems can scale to 10 million IOPS, with 150GB/s at 300µs latency.

Analysis. Dell EMC continues to evolve the original Symmetrix line, which now shows almost nothing of the original architecture. However, as a platform, PowerMax retains the legacy of reliability and enterprise-class that has been seen in Symmetrix, DMX and VMAX. Dell EMC has effectively replaced the existing SAS back-end with NVMe, while introducing software enhancements to exploit the increased bandwidth and parallelisation. Customers comfortable with VMAX will see PowerMax as the expected evolution of the platform range. The next question to ask is where Dell EMC will go with XtremIO (which seems to compete heavily with PowerMax), with VMAX hybrid and mid-range products. It's possible to envisage a stripped down PowerMax eventually replacing XtremIO, with the mid-range platforms fitting the hybrid requirements for customers.

E8 Storage Inc. (E8 Appliances and Software)

E8 Storage Inc has built a disaggregated architecture that takes the traditional functions of a storage array and divides them between a shared storage appliance and the host server. This design, which separates the control path from the I/O data path, removes the bottlenecks caused by storage systems that channel all data through one or more linked controllers.

Hosts and appliances are connected to each other using a high-speed network (typically InfiniBand or Ethernet) that delivers extremely low latency at high throughput. The distributed architecture uses up to a single CPU core on each host, which creates a distributed controller across the connected hosts.

Scalability in this architecture is linear, as each host added to the configuration brings resources that scale the "virtual" controller. The E8 storage appliances act as a PCIe to Ethernet bridge, connecting SSDs in the appliance to each host using NVMe over Fabrics. Rather than use dedicated hardware, storage appliances are built around a standard x86 server. This means customers can deploy the E8 software on their own hardware solution and in fact E8 Storage does have a software-only offering. The x86 server within the appliance also hosts the metadata server, providing a single point of reference for configuration information across the distributed architecture.

There are currently two hardware models available. The E8-D24 appliance has dual motherboards for redundancy, connecting through a passive mid-plane to up to 24 NVMe SSDs in a 2U chassis. Each of the "canisters" provides connectivity from either 100GbE or 100Gb InfiniBand to the dual-ported NVMe drives. Within each canister, system DRAM is used to process write I/O, therefore battery backup is provided to protect the data in the event of system power loss. This system is capable of delivering up to 10 million IOPS and 40GB/s throughput. Data is protected within the appliance typically using a RAID-6 scheme.

[15] https://blog.architecting.it/powermax-vmax-xtremio/

The smaller E8-S10 has a single canister, up to 10 single-port NVMe drives and can deliver 4 million IOPS and 16GB/s read throughput in 1U. Within the appliance, data is protected typically using a RAID-5 scheme, with appliance redundancy implemented at the host layer through replicating to multiple appliances using techniques such as LVM mirroring.

A single E8 storage appliance can support up to 126 host servers with latencies close to the native NVMe drives themselves. The architecture is not restricted to a single appliance and can be scaled with multiple appliances connecting to a single host.

Clearly, having a distributed architecture means consuming resources on the host, however, this is seen as a small price to pay for the ability to gain horizontal scalability. The architecture also provides the ability to completely minimise the impact of typical storage management functions on the I/O path. E8 Storage appliances simply route data from host to drive, with no translation layer in the appliance itself.

One solution to the consumption of resources on the host has been to offload some functionality to the network interface card in the host. At Flash Memory Summit in 2018, E8 Storage demonstrated their solution in conjunction with the Broadcom Stingray SmartNIC.

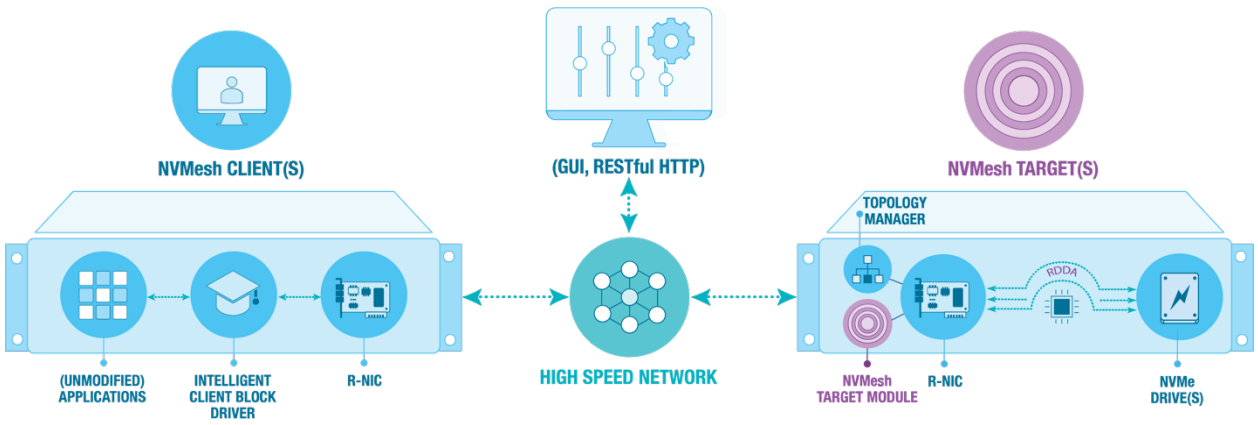
E8 Storage has been strong in demonstrating the power of their architecture, producing multiple benchmark test results, including the integration of file-based solutions such as IBM Spectrum Scale[16].

Analysis. The E8 architecture removes the scalability issues of traditional storage, significantly reducing the I/O path from host to media. However, the trade-off is in pushing some features and functionality to the host, with a resultant consumption of resources. This could be mitigated by using a SmartNIC to offload this work but does result in a more bespoke solution. E8 has seen particularly good results as a block storage basis for scale-out file systems, which offers good applicability to analytics and other ML/AI applications.

Excelero Inc (NVMesh)

Excelero has developed a solution that uses a cluster or mesh of servers acting as both consumers and providers of storage. NVMesh allows NVMe storage resources across many servers to be accessed across the network using NVMe over Fabrics and a proprietary protocol called RDDA, a form of remote direct memory access (RDMA). Each consumer and provider of storage uses a dedicated RNIC (RDMA NIC) that provides direct access to drives without needing to access the target server's CPU.

Removing the bottleneck of accessing the processor in the server where NVMe storage is attached means scalability in individual servers and across the mesh is not a problem. Architectures that require host processing also require system memory and that can place limits on scalability.[17]



[16] https://www.spec.org/sfs2014/results/res2018q3/sfs2014-20180706-00041.html
[17] https://blog.architecting.it/flash-storage-dram-curve/

NVMesh presents storage to consuming hosts as block devices. This allows other software (for example distributed file systems) to be layered over the generic block storage capacity or to a hyper-converged architecture.

As a software solution, understanding the performance capabilities of NVMesh compared to other vendors isn't a simple task. Excelero has partnered with Boston Ltd in the UK to create a set of pre-configured NVMe appliances that use Micron 9200 NVMe devices in either 1U or 2U servers. The "Talyn" platform demonstrates sample configurations for Utility, Capacity and Performance requirements, claiming up to 24 million IOPS with 96GB/s throughput at less than 100µs.

Analysis. Excelero has chosen to be software-only which has good and bad repercussions. On the good side, any servers with appropriate NIC hardware can be used to build out NVMesh. On the negative side, customers can't see any reference architectures with hero numbers. Excelero perhaps needs more partnerships with solution providers to show the validity and capability of the software.

IBM (FlashSystem 9100)

Until recently, IBM claimed that back-end NVMe storage wasn't fast enough to deliver the performance required of modern all-flash systems. Using technology from the Texas Memory Systems acquisition in 2012, IBM took the RamSan platform and developed a range of products based on a proprietary flash component known as a MicroLatency module.

The IBM FlashCore architecture as seen in platforms like the FlashSystem 900 use end-to-end hardware with custom ASICs and FPGAs to write directly to NAND. As a result, IBM deemed the use of NVMe would slow down this architecture. [18]

Figure 6 - IBM FlashCore Module

With the release of the FlashSystem 9100 platform in July 2018, IBM has decided that NVMe drives do have some merit. The 9100 system uses either commodity NVMe SSDs or IBM's new FlashCore NVMe module that uses the U.2 form factor. The new FlashCore drives are available in capacities of 4.8TB, 9.6TB and 19.2TB, which is slightly larger than existing commodity SSDs.

IBM claims the use of Everspin MRAM for power loss protection in replacement of super-capacitors provides the ability to add more NAND chips into the standard U.2 form-factor design.

Both the FlashSystem 9110 and 9150 models support up to twenty-four 2.5" NVMe drives in a 2U chassis. Expansion shelves continue to use SAS drives. Platforms currently support iSCSI (up to 25GbE) and Fibre Channel (16Gb/s). IBM has issued a Statement of Direction that states FC-NVMe will be supported in the future on FC 16Gb/s adaptors and NVMe-oF will be supported on 25Gb Ethernet, although at the time of writing these remained future enhancements.[19]

IBM headline numbers claim single node performance of 2.5 million IOPS, although this is based on 4K read cache hits. More conservative figures of 1.1 million IOPS are quoted with 4K read cache misses. Throughput is 34GB/s per node (256K blocks). IBM also claims latency "as low as" 100µs, although this figure is not in the online datasheet.[20]

At the end of 2017, IBM demonstrated NVMe-oF using Power9 servers, InfiniBand and IBM FlashSystem. QDR InfiniBand (40Gb/s) is now supported on the FlashSystem 900 platform. IBM claims read/write latencies of 95/155µs respectively with 1.1 million IOPS (100% random read) and 600,000 IOPS (random write). Throughput is 10GBs (100% sequential read) and 4.5GB/s (100% sequential write).

Analysis. IBM's move to commodity drives seems a little at odds with the way the company was originally developing all-flash products. FlashCore modules now only have a sight capacity gain over commodity SSD, although IBM does gain some advantages in offloading functionality and having the ability control the FTL within

[18] https://www.youtube.com/watch?v=qqvmy7Bte7g
[19] https://www-01.ibm.com/common/ssi/rep_ca/6/877/ENUSZG18-0076/ENUSZG18-0076.PDF
[20] https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=10017110USEN&

the devices. Moving to NVMe perhaps gains broader market recognition but doesn't reflect any improvement in product performance compared to previous FlashSystem products. It remains to be seen where IBM is headed.

Lightbits Labs (NVMe/TCP Solution)

Lightbits Labs is a start-up currently still in stealth that is developing a shared storage solution based on NVMe/TCP (NVMe over TCP). The aim is to bring to market a solution that doesn't depend on more expensive host-based hardware such as InfiniBand or RoCE adaptors and so can be deployed across generic infrastructure in the data centre.

Lightbits is also integrating hardware acceleration at the back-end of their product solution that implements a global flash translation layer for data services.

Analysis. It's early days for the Lightbits technology, however start-ups are starting to see NVMe/TCP as a valid architecture that can perhaps straddle the current Fibre Channel and iSCSI solutions and the high-end RoCE and InfiniBand offerings. Building in a proprietary FTL resembles more what's been seen by Pure Storage and Violin Systems. Exactly what benefit this offers remains to be seen.

NetApp (AFF A800 and EF570)

NetApp has had a long history of using solid-state media as an acceleration solution for the company's storage platforms. PAM cards (Performance Acceleration Module) based on PCIe flash storage were introduced as early as 2008 to provide read caching. The technology was subsequently renamed Flash Cache.

In terms of putting all-flash into existing platforms, NetApp was slower to react than their competitors. Dave Hitz, NetApp co-founder has openly admitted that the strategy was initially focused on caching rather than using solid-state media as a tier.[21] This is reflected in the use of Flash Cache (acceleration within the array/appliance) and Flash Accel, which offered host-based acceleration for hypervisors through host-based storage and an ESXi plugin (VIB).

In 2015, NetApp introduced the "All-Flash FAS" or AFF product line. AFF systems are purely based on flash storage and use a set of features and software enhancements and improvements called ONTAP FlashEssentials to optimise the I/O path for solid-state media. The AFF line has continued to grow. In May 2018, NetApp introduced the AFF A800, an all-flash platform that uses NVMe solid-state disks.

AFF A800 supports up to 48 NVMe SSDs within a 4U chassis that contains one controller pair. A single A800 cluster can be expanded to 12 HA pairs (NAS protocols) or 6 HA pairs (SAN protocols), which means up to 576 NVMe drives (SAN) or 1,152 NVMe drives (NAS) in one system. NetApp claims to support 15.36TB NVMe drives, which would provide around 740TB of raw capacity in one controller pair. As no vendor is currently shipping 15.36TB NVMe drives, it is more likely that systems will ship with at best, 7.68TB drives, with around 370TB of raw capacity per node. Internally, each SSD is connected to each controller through two PCIe Gen 3 lanes, with four switches per controller. Expansion shelves are still SAS connected.

In terms of performance, NetApp claim a single node-pair can achieve up to 1.1 million IOPS and 25GB/s throughput at 200µs latency. These numbers scale up linearly with additional cluster nodes, assuming that workloads are able to be distributed across the available node pairs. NetApp has submitted the AFF A800 to Storage Performance Council testing, achieving a 2.4 million SPC-1 IOPS score.[22]

At the front end of the AFF A800, NetApp supports FC-NVMe using 32Gb (Gen 6) Fibre Channel. This allows the company to claim that AFF A800 is the first "end-to-end" NVMe enabled storage array in the market. The platform also supports 100Gb Ethernet.

NetApp has a second NVMe-enabled product line. The EF-series products have two models that were introduced in September 2017 with NVMe-oF support. The EF570 (all-flash) and EF5700 (hybrid) systems both support NVMe-oF using InfiniBand EDR (100Gb/s). NetApp claims performance figures of one million IOPS with 21GB/s bandwidth at sub-100µs latency. This is effectively not much more than the latency of the underlying

[21] https://www.youtube.com/watch?v=Ybg4bSyJ8H8
[22] http://spcresults.org/sites/default/files/files/executive_summary/A32007_ES.pdf

storage media. EF-series products are much more basic in terms of data services, so low latencies are likely to be achievable with this platform.

NetApp is working on host-side connectivity using NVMe-oF, with technology from the acquisition of Plexistor in May 2017. This is now being marketed as Memory Accelerated Data or MAX Data for short. MAX Data uses host-based storage-class memory (SCM) and the Plexistor software to create an extremely low-latency local file system, capable of delivering single digit microsecond latencies. The local file system acts as a write-back cache, offloading data via snapshots to an AFF appliance over 100Gb/s RDMA. MAX Data is still in preview, so product details are subject to change before final release.

Analysis. NetApp initially took a different path with flash and NVMe, preferring to add persistent media as an acceleration tool, rather than as a tier of storage. This approach left NetApp behind the curve, however, the company has rapidly recovered, introducing All-Flash FAS and now NVMe-based systems. It's possible that ONTAP wasn't initially capable of using flash as a tier of storage, which is why the original path was chosen, giving the software teams time to work out how to introduce flash as a persistent tier. Putting that aside, customers have gained the ability to migrate to ever faster and lower-latency hardware, without changing from a software platform (ONTAP) that was well known and understood. There's a lot to be said for keeping APIs, CLIs and operational processes consistent as the hardware improves.

Pavilion Data (NVMe-oF Storage Platform)

Pavilion Data Systems has developed a rack-scale NVMe-oF storage solution based on what resembles more of a network switch architecture than storage. The NVMe-oF Storage Platform (which doesn't appear to have any other specific name) is based around a modular design using storage controllers and drives that are deployed into a 4U chassis.

Each controller provides four 100Gb Ethernet connections with capacity for up to 72 NVMe SSDs. Customers can start with two and expand up to 20 controllers in a single 4U chassis.

Pavilion claims performance characteristics of up to 120GB/s bandwidth at 100µs latency and 20 million 4KB read IOPS. Host connectivity uses standard commodity 40GbE or 100GbE RoCE network adaptors and NVMe-oF host drivers.

Analysis. Pavilion has taken a different route in their choice of licensing model. Customers can bring their own media and are charged a licence per media slot, rather than on capacity. This enables the customer to upgrade capacity over time, without paying additional charges. In lots of ways this aligns to the networking model of charging per physical port and will no doubt be attractive to some rack-scale customers.

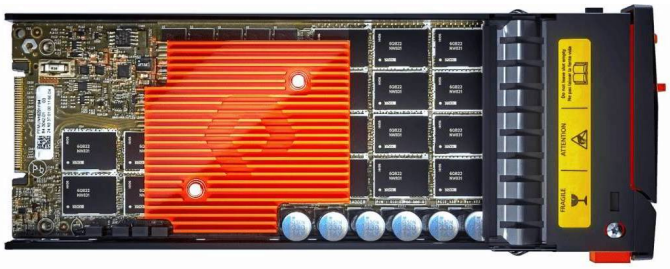
Pure Storage Inc. (FlashArray and FlashBlade)

Pure Storage first introduced NVMe into the FlashArray product line in 2015 with the release of the FlashArray//M series. FlashArray//M changed the architecture of the product line to use dual 2.5" SSD caddies that were connected through either NVMe or SAS. The controllers, SSDs and NVRAM components were all connected through a PCIe midplane, enabling the controllers to remain stateless while moving the drive persistence data from SLC drives on each shelf to the NVRAM in the controller chassis.

With the release of FlashBlade in 2016, Pure removed the dependency on SSDs, building out custom compute modules with storage on a daughterboard on each NAS controller. This talked directly to the NAND, using custom FPGA and ARM cores. Daughterboard capacities were initially 8TB or 52TB and used a proprietary protocol over PCIe to talk to each controller.

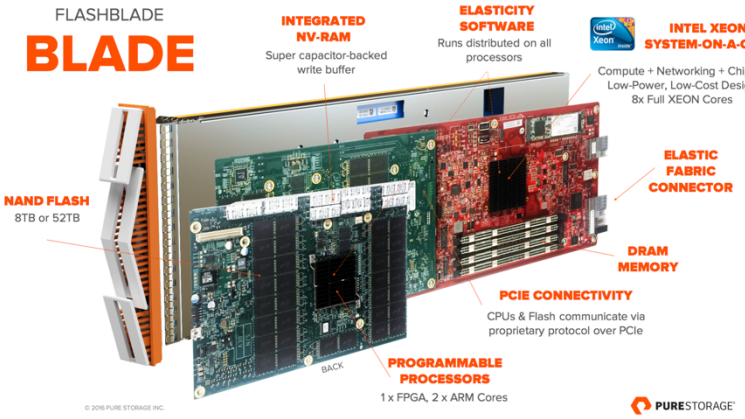
In April 2017, Pure Storage announced FlashArray//X. This uses the same chassis as FlashArray//M but offers the capability to use new DirectFlash modules connected over NVMe in place of the dual-SSD SAS-connected caddies. DirectFlash moves the Flash Translation Layer of software within traditional SSDs into the Purity Operating Environment.

Figure 7 - DirectFlash Module



Now, instead of drives making independent decisions on performance impacting tasks such as wear levelling and garbage collection, the process is managed from within Purity. This enables FlashArray systems to deliver a more consistent (deterministic) performance experience, reduce the amount of over-provisioning used and manage endurance more efficiently. The ability to do "global flash management" is divided into three features, called *Adaptive I/O Control*, *Smart Endurance* and *Predictive Resiliency*.

Figure 8 - FlashBlade Components



At Pure Accelerate 2018, Pure Storage announced the future availability of system expansion based on NVMe over Fabrics. A controller chassis will be capable of connecting to dedicated expansion shelves with up to 28 DirectFlash modules using NVMe-oF based on RoCE v2 over 25Gb Ethernet. This feature is expected to be available in 4Q2018. At the time of writing, NVMe-oF for host connectivity was a future option with no specific timescales on release.

Pure Storage clearly took an early decision to use NVMe as the core of their FlashArray and FlashBlade platforms. The result is that customers who deployed FlashArray//M can upgrade to the latest FlashArray//X systems with new controllers and NVMe drives without disruption or downtime. This is because the same chassis is used for both models. Pure offers a supported upgrade path to replace both the controllers and drives over time. During Pure's 2Q2019 earnings call on 21 August 2018, the company was able to announce that more than 50% of shipments during the quarter were all-NVMe systems, with the majority of systems expected to be shipped all-NVMe by the end of the year. [23]

Pure Storage no longer produces performance figures for FlashArray products, dropping throughput numbers with the introduction of the //X platform. In 2017, the performance of generation one FlashArray//X systems was quoted as "hundreds of microseconds". With generation two, this has been updated to 250µs. The previous fastest system was the FlashArray//M70, delivering 370,000 32K IOPS, 11.5GB/s bandwidth at sub-millisecond latency.

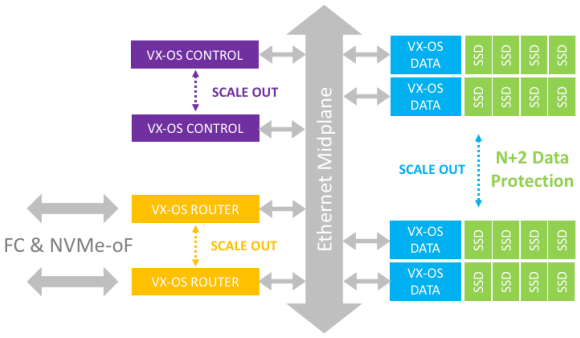
More details on the FlashArray architecture and implementation of NVMe can be found in the Architecting IT blog post ***"Pure Storage – Seeding the NVMe Market"***.

Analysis. Pure Storage saw the move to NVMe early on in their product development. As a result, customers are now reaping the rewards of an architecture that allows every component to be replaced with minimal disruption. It's important not to force operational change by introducing new solutions that require new knowledge, training and coding work. Pure Storage took that a stage further and has enabled customers to do upgrades without even needing rack swing space. It can't be understated how significant reducing operational risk actually is. The question for Pure, though, is where the technology goes next. FlashBlade is scale-out, whereas FlashArray is still effectively scale-up. Will Pure release a completely new architecture to scale-out for future storage requirements? We can only wait and see.

[23] Pure Storage 2Q 2019 Earnings Call Transcript

Vexata Inc. (VX-100M and VX-100F)

Vexata has developed a storage platform that implements either NAND flash or storage class memory as the persistent storage component. The Vexata architecture is a hardware-focused design (although with commodity components) that uses Enterprise Storage Modules at the back-end and controllers (IOCs) at the front-end. Both are centred around a high-speed Ethernet fabric midplane and can be scaled independently.



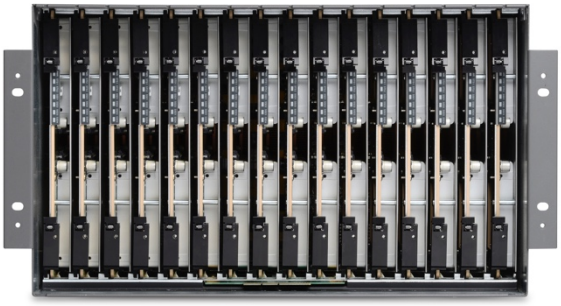
VX-OS ARCHITECTURE

Enterprise Storage Modules (ESMs) are built from commodity SSDs (either NAND flash or Optane) and a dedicated MIPS processor that provides the translation from Ethernet to NVMe. The MIPS processor handles metadata and other offloaded functions such as data protection. The current VX-100 products support from 3 to 16 hot-pluggable ESMs in a 6U chassis.

At the front-end of the system, I/O controller modules use a feature called VX-Router, based on FPGAs. The IOCs provide connectivity to external hosts using Fibre Channel. In the future the platform will support NVMe over Fabrics using Fibre Channel (FC-NVMe) and Ethernet.

The Vexata VX-OS architecture provides for scaling of front and back-end components around the Ethernet midplane. Although today the VX-100 products support dual IOCs and up to 16 ESMs, the architecture will scale to multiple front and back-end components.

Figure 9 - VX-100 ESM View



VX-OS uses many techniques to remove the bottlenecks seen with traditional storage. Control and data paths are separate across the platform by splitting the functionality of the IO Controllers into two components. VX-OS Router handles host I/O, while VX-OS Control looks after management functionality, including the GUI, CLI and RESTful API interfaces.

The distributed architecture of VX-OS stores metadata across ESMs, removing the scalability issues seen with traditional architectures that store metadata in DRAM on the front-end controller. The distributed nature of the design means a VX-100 system can perform much more I/O in parallel, fully exploiting the benefits of NVMe storage. As a result of using a hardware-based design, Vexata claims that end-to-end, VX-OS adds only 5-10µs of additional latency to that of the persistent storage media.

In terms of specifications, at the time of writing the VX-100F, based on NAND flash offers up to 435TB of capacity and 7 million IOPS at just 200µs of latency. The Optane platform (VX-100M) delivers even better performance, with 8 million IOPS at 40µs of latency. Capacity is lower for VX-100M as Optane drives currently offer much lower capacity than NAND flash.

*A more detailed deep dive into the Vexata can be found in an independently authored Architecting IT white paper **"Vexata VX-OS Architecture"**, available for free download online.*

Analysis. The Vexata platform shows what can be achieved by using commodity hardware components and rethinking the design of how front and back-end components communicate. A "matrix-style" architecture is not new, but some of the components in VX-OS are definitely reworked to remove the issues of legacy designs. Specifically, the distribution of metadata to back-end ESMs. VX-OS seems to sit on the cusp of becoming a fully-meshed architecture, which could position the platform well for large-scale enterprise deployments, especially those with low-latency requirements (think ML/AI).

WekaIO (Matrix)

WekaIO is a start-up vendor that has developed a software-based scale-out file system solution. Matrix is deployed onto multiple servers or nodes as a cluster, delivering a single global namespace. The technology uses NVMe drives in order to deliver the most performance and lowest latency compared to traditional solutions.

Matrix is deployed on commodity hardware running the Linux operating system. The architecture implements a virtual file system called MatrixFS that runs in the Linux kernel. MatrixFS is exposed to applications through what looks like a local file system supporting standard ext4 or xfs formats. However, the file system is distributed across and available to all nodes in the cluster.

Where direct access isn't possible, Matrix offers connectivity through a variety of standard storage networking protocols, including SMB, NFS, S3 (object) and HFDS (Hadoop). This is delivered through front-end components running in user space as containers. Both the front-end and back-end software components (as shown in Figure 10) run in user space under LXE. This allows WekaIO to more accurately control the performance of the platform. Back-end components talk to NVMe media directly over PCIe and to network interface cards (NICs) and adaptors using SR-IOV. This means there is no additional copying to/from Linux I/O and networking stack buffers, improving performance. Matrix supports Ethernet (10GbE and upwards) as well as InfiniBand.

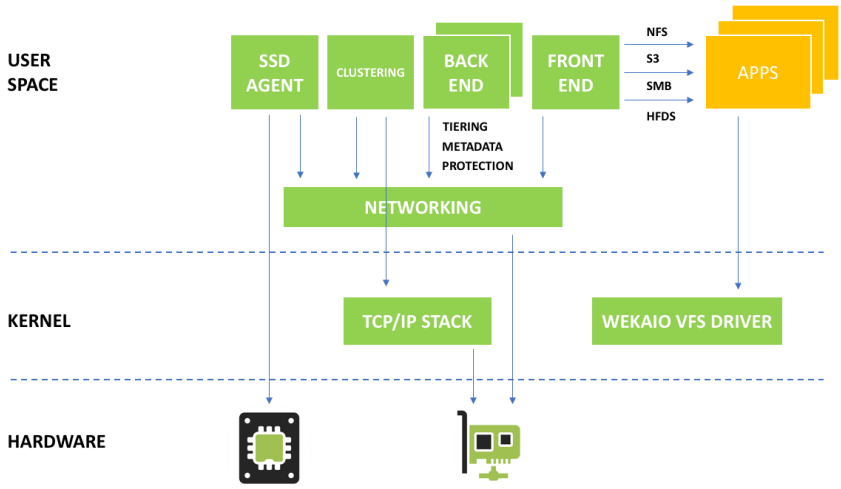


Figure 10 - WekaIO Matrix Architecture

Matrix offers protection from media, server, rack or even metro data centre failure through a proprietary erasure coding system called DDP or Distributed Data Protection. The coding mechanism provides for schemes that use 4D+2P (Data and Parity/Protection) up to 16D+4P, or essentially 4-way protection.

The use of erasure coding is much more efficient than simple mirroring or using replicas but obviously needs to be implemented in a way that doesn't impact performance. To date, WekaIO has not released any details about the proprietary parts of their solution as patents are still pending.

Matrix offers the capability to take data snapshots of a file system. Snapshots are space efficient and have no performance overhead as all write I/O is implemented by a redirect-on-write scheme. WekaIO claims that this scheme does not have the same garbage collection issues as (for example) NetApp WAFL, however, the specific details of this difference have not been made public.

Matrix offers the capability to offload snapshots to an S3-compatible object store. This includes on-premises solutions; however, the feature is really aimed at being able to push data into the public cloud. Matrix snapshots are self-describing and can be re-constituted either as a read-only image to another cluster or as a file system clone.

WekaIO has an existing partnership to sell Matrix through HPE, running the software on HPE Apollo servers. Matrix is also available on the AWS Marketplace and can be deployed on either r3 or i3 instances with local SSD or NVMe drives. [24]

Analysis. Matrix is designed to be highly scalable, but more important, implements distributed metadata. This means a Matrix cluster is well-suited to HPC and ML/AI workloads that are more demanding on small-file

[24] https://www.weka.io/press-releases/wekaio-achieves-amazon-web-services-storage-competency-status-primary-storage/

access. Nodes are connected through a proprietary version of NVMe over Fabrics, which reduces latency low enough that performance can be better than running local NVMe SSDs. WekaIO has proven out the capabilities of Matrix through the results of SPECsfs testing[25], although we need to remember that this benchmark doesn't include a pricing component and therefore can make vendor comparisons difficult. There's no doubt that the platform is capable of delivering on the claims made by WekaIO. The challenge for the company will be to produce reference architectures and solution architectures that go head to head with vendors like NetApp and Pure Storage, both of which have ML/AI reference architectures already in the market.

[25] https://www.spec.org/sfs2014/results/res2019q1/sfs2014-20181218-00055.html

More Information

For additional technical background or other advice on replication technologies, contact enquiries@brookend.com for more information. Architecting IT is a brand name of Brookend Ltd and independent consultancy, working for the business value to the end customer.

Email: architectingit@brookend.com
Twitter: @architectingit

The Author

Chris M Evans has worked in the technology industry since 1987, starting as a systems programmer on the IBM mainframe platform. After working abroad, he co-founded an Internet-based music distribution company during the .com era, returning to consultancy in the new millennium. Chris writes a popular blog at http://blog.architecting.it, attends many conferences and invitation-only events and can be found providing regular industry contributions through Twitter (@chrismevans) and other social media outlets.