CW+ Content

NVMe flash storage 101

In this e-guide

In this e-guide:

Flash storage is already much faster than spinning disk, but with the advent of NVMe – a new standard based on PCIe – flash could achieve a potential that has so far eluded it. NVMe is a storage protocol designed for the performance and characteristics of flash and will replace the SAS and SATA protocols that were developed for spinning disk and currently act as a bottleneck to solid state drives. It is, however, early days. So far, NVMe can be deployed without complications in limited scenarios, but the tech industry is fast developing ways of implementing NVMe that will see shared storage arrays with the blistering performance of flash NVMe.

Antony Adshead, storage editor

Storage briefing: NVMe vs SATA and SAS

Antony Adshead, storage editor

The rise of flash storage has been a dominant theme in the datacentre in recent years. And rightly so, with performance in the order of 100 times better than spinning disk media.

But there are always going to be choke points in the input/output (I/O) path from app to media, and a key bottleneck has been the protocols used to address storage on disk.

These have largely settled onto SAS and SATA for spinning disk hard disk drives (HDDs), and these formats have also been adopted as protocols utilised by flash media drives.

But SAS (based on the SCSI command set) and SATA (based on the ATA command set) are historic protocols developed for mechanical media. They do not have the characteristics to take advantage of the benefits of flash media.

So, the industry has come up with a standard, non-volatile memory express (NVMe), built for flash and complete with a number of performance advantages.

NVMe is a standard based on peripheral component interconnect express (PCIe), and is built for physical slot architecture. As they launched PCIe server flash products, suppliers each developed proprietary protocols to manage traffic. NVMe is a largely successful exercise in the replacement of such disparate proprietary protocols with a true standard.

NVMe also allows for the use of 2.5in format solid state drives via the U.2 connector on the card body.

Of course, NVMe-equipped cards can be aggregated into an array-type format in a similar way to EMC's DSSD D5 and startup E8's D24. In fact, NVMe format cards could be added to array controllers now. However, any such array product that wanted to take advantage of NVMe would require controller hardware that did not act as a bottleneck itself.

Key benefits of NVMe over SATA and SAS

In short, NVMe provides much greater bandwidth than SAS and SATA, with vastly improved queuing.

SATA, in its SATA III incarnation, provides 6Gbps and 600Mbps throughput. SAS provides 12Gbps and 8Gbps throughput.

NVMe has data transfer performance characteristics of PCIe Gen 3 and therefore has bandwidth of around 1Gbps per lane – up to 16Gbps in a 16-

lane configuration. PCIe Gen 4 will likely double that and is set for release in 2017.
At the same time, NVMe is built to handle more queues than SAS and SATA at the drive, with 65,000 queues and 65,000 command queue depths possible. This compares to one queue for SAS and SATA, and queue depths of 254 and 32 respectively.

This means NVMe-equipped storage should not experience the performance degradation that SAS and SATA can go through if overloaded with I/O requests.

NVMe key use cases

In summary, NVMe offers blistering performance for flash storage over the existing SAS and SATA drive connection protocols. Its ability to handle a high number of queues and commands makes it eminently suitable for applications that batter storage I/O with large queue depths, such as databases and some web operations.

Currently, however, it is mostly a server flash technology, although its use in arrays and hyper-converged infrastructure and as addressable shared storage (via NVMf – NVMe over fabrics) is coming.

Next article

NVMe over fabrics vs Fibre Channel, Ethernet, Infiniband

Antony Adshead, storage editor

We recently looked at NVMe, a PCIe-based protocol that allows computers to communicate with storage in a way optimised for flash storage, with huge increases in input/output (I/O) and throughput possible compared with spinning disk-era SAS and SATA protocols.

As we saw, currently NVMe is a slot-in for existing PCIe server flash use cases, although storage array suppliers have started to develop arrays that utilise NVMe-connected flash storage, and this is more than likely the future direction of in-array connectivity.

But the I/O chain, obviously, does not end at the storage array. Mostly, datacentre usage sees multiple hosts hooked up to shared storage. So, to preserve the advantages of NVMe between host and array, NVMe over fabrics (NMVf), has been developed.

The key idea behind NVMf is that the advantages of NVMe – high bandwidth and throughput with the ability to handle massive amounts of queues and commands – are preserved end-to-end in the I/O path between server host and storage array, with no translation to protocols such as SCSI in between that would nullify those benefits.

In short, the huge parallels that NVMe offers are retained across the storage network or fabric.

NVMe is based on PCIe communications protocols, and currently has the performance characteristics of PCIe Gen 3. But traffic can't travel natively between remote hosts and NVMe storage in an array. There has to be a messaging layer between them.

That messaging layer is essentially what NVMf comprises. So far, the NVM Express group has devised fabric transports to allow remote direct memory access (RDMA) and Fibre Channel-based traffic with the aim of not increasing latency by more than 10 microseconds, compared with an NVMe device in a PCIe slot. RDMA allows a direct connection from one device to the memory of another, without involving the operating system.

RDMA-based protocols include RoCE or RDMA over Converged Ethernet; iWARP or internet wide area RDMA protocol, which is roughly RDMA over TCP/IP; and Infiniband.
NVMf performance should be governed by that of the networking protocol used, so bandwidth with Ethernet will be in the hundreds of gigabits per second, Infiniband at tens of gigabits per second per-lane and Fibre Channel hitting 128Gbps with its Gen 6.

NVMf products

It's early days, so there isn't much in the way of products at the time of writing.

NVMf will be supported by default at the hardware level in NVMe cards and drives, as well as NVMe-enabled arrays. But, so far, the products necessary to convert existing networks and fabrics, such as NICs and HBAs, for hosts as well as switches, are thin on the ground.

In February 2016, Broadcom released samples of Gen 6 Fibre Channel adapters to manufacturers, but there's no sign of generally available products yet.

On the RDMA NIC front, there seems to be little available yet, although Mellanox (RoCE) and Chelsio (iWARP) have product pages on their websites. Meanwhile, Mangstor makes arrays with NVMe storage that can be connected to hosts via Mellanox NVMe-capable switches.

Next article

NVMe: PCIe card vs U.2 and M.2

Antony Adshead, storage editor

Non-volatile memory express (NVMe) is a peripheral component interconnect express (PCIe)-based standard protocol that can allow solid state storage to work to its full potential by hugely increasing drive connectivity performance.

NVMe can be deployed via PCIe form factors that include add-in cards (AiC), M.2 and U.2.

Methods to allow NVMe's fast access performance are being developed for use across storage networks and fabrics, but for the time being NVMe is mostly a replacement for serial advanced technology attachment (SATA) and serial-attached SCSI (SAS) drives in servers or possibly storage arrays.

Here, Antony Adshead, storage editor at Computer Weekly, talks to Greg Schulz, founder and senior consulting analyst of independent IT advisory consultancy firm Server StorageIO about NVMe deployment options.

Antony Adshead: Why use a U.2 slot when you could use the PCIe AiC?

Greg Schulz: Simple. Your server or storage system may be PCIe slot constrained yet have more available U.2 slots. There are U.2 drives from various suppliers including Intel and Micro, as well as servers from Dell, Intel and Lenovo among many others.

Adshead: Why and when would you use an NVMe M.2 device?

Schulz: As a local read/write cache, or perhaps a boot and system device on servers or appliances that have M.2 slots. Many servers and smaller workstations including Intel NUC support M.2. Likewise, there are M.2 devices from many different suppliers including Micron and Samsung, among others.

Adshead: Where and why would you use NVMe PCIe AiC?

Schulz: Whenever you can and if you have enough PCIe slots of the proper form factor, including the number of lanes – for example x1, x4, x8, x16 – to support a particular card.

Adshead: Can you mix and match different types of NVMe devices in the same server or appliance?

Schulz: As long as the physical server and its software (BIOS/UEFI, operating system, hypervisors, drivers) support it, then yes. Most server and appliance suppliers support PCIe NVMe AiCs, but you need to pay attention to precise form factor. You should also verify operating system and hypervisor device driver support. PCIe NVMe AiCs are available from Dell, Intel, Micron and many other suppliers.

Adshead: Does having M.2. mean you have NVMe?

Schulz: That depends. Some systems implement M.2 with SATA, while others support NVMe. Read the fine print or ask for clarification.

Adshead: Who should we keep an eye on in the NVMe ecosystem?

Schulz: Suppliers to look out for include E8, Enmotus (micro-tiering software), Excelero, Magnotics, Mellanox, Microsemi, Microsoft (Windows Server 2016 S2D, ReFS), NVM Express trade groups (for example, nvmexpress.org), Seagate, VMware (there is a native NVMe driver available as part of vSphere ESXi) and WD/Sandisk.

Next article

NVMe gives "shared DAS" as an answer for analytics; but raises questions too

Antony Adshead, storage editor

Go back 10 or 20 years and direct-attached disk was the norm. IE, just disk in a server.

It all became a bit unfashionable as the virtualisation revolution hit datacentres. Having siloed disk in servers was inherently inefficient and server virtualisation demanded shared storage to lessen the I/O blender effect.

So, shared storage became the norm for primary and secondary storage for many workloads.

But in recent years, we saw the rise of so-called hyperscale computing. Led by the web giants this saw self-contained nodes of compute and storage aggregated in grid-like fashion.

Unlike enterprise storage arrays these are constructed from commodity components and an entire server/storage node swapped out if faulty, with replication etc handled by the app.

The hyperscale model is aimed at web use cases and in particular the analytics - Hadoop etc - that go with it.

Hyperscale, in turn, could be seen as the inspiration for the wave of hyper-converged combined server and storage products that has risen so quickly in the market of late.

Elsewhere, however, the need for very high performance storage has spawned the apparently somewhat paradoxical direct-attached storage array.

Key to this has been the ascendance of NVMe, the PCIe-based card interconnect that massively boosts I/O performance over the spinning disk-era SAS and SATA to something like matching the potential of flash.

From this vendors have developed NVMe over fabric/network methods that allow flash plus NVMe connectivity over rack-scale distances.

Key vendors here are EMC with its DSSD D5, E8 with its D24, Apeiron, Mangstor, plus Excelio and Pavilion Data Systems.

What these vendors offer is very high performance storage that acts as if it is direct-attached in terms of its low latency and ability to provide large numbers of IOPS.

In terms of headline figures - supply your own pinches of salt - they all claim IOPS in the up to 10 million range and latency of <100Î¼s.

That's made possible by taking the storage fabric/network out of the I/O path and profiting from the benefits of NVMe.

In some cases vendors are taking the controller out of the data path too to boost performance.

That's certainly the case with Apeiron - which does put some processing in HBAs in attached servers but leaves a lot to app functionality - and seems to be so with Mangstor.

EMC's DSSD has dual "control modules" that handle RAID (proprietary "Cubic RAID") and presumably DSSD's internal object-based file layout. E8 appears to run some sort of controller for LUN and thin provisioning.

EMC and Mangstor run on proprietary drives while E8 and Apeiron use commodity cards.

A question that occurs to me about this new wave of "shared DAS" is: Does it matter whether the controller is taken out of the equation?

I tend to think that as long as the product can deliver raw IOPS in great numbers then possibly not.

But, we'd have to ask how the storage controller's functions are being handled. There may be implications.

A storage controller has to handle - at a minimum - protocol handling and I/O. On top of that are LUN provisioning, RAID, thin provisioning, possibly replication, snapshots, data deduplication and compression.

All these vendors have dispensed with the last two of these, and mangstor and Apeiron have ditched most of the rest, Apeiron, for example, offloading much to server HBAs and the app's own functionality.

So, a key question for potential customers should be over how the system handles controller-type functionality. The more processing that is done over and above the fundamentals has to be done somewhere and potentially hits performance, so is there over-provision of flash capacity to keep performance up while the controller saps it?

Another question is, despite the blistering performance possible with these shared NVMe-based DAS systems, will it be right for leading/bleeding edge analytics environments?

The workloads aimed at - such as Hadoop but also Splunk and Spark - are intensely memory hungry and want their working dataset all in one place. If you're still having to hit storage - even the fastest "shared" storage around -

will it make the grade for these use cases or should you be spending money on more memory (or memory supplement) in the server?

Next article

Getting more CW+ exclusive content

As a CW+ member, you have access to TechTarget's entire portfolio of 120+ websites. CW+ access directs you to previously unavailable "platinum members-only resources" that are guaranteed to save you the time and effort of having to track such premium content down on your own, ultimately helping you to solve your toughest IT challenges more effectively – and faster – than ever before.

Take full advantage of your membership by visiting www.computerweekly.com/eproducts

Images; Fotalia