WHAT IS NVME-OF INTEROP TESTING AND WHY IS IT IMPORTANT?

by David Woolf

Nonvolatile memory express (NVMe) has gained incredible adoption as a low-latency interface for connecting host systems with SSDs. NVMe storage attached directly to host systems via peripheral component interconnect express (PCIe) is commonly available in high-end storage servers, laptops, and mobile devices.

A complementary specification, NVMe over fabrics (NVMe-oF) defines the attachment of that same NVM storage media over a fabric with microsecond-level latency. The NVMe-oF specification was designed to be agnostic of the underlying transport, and an additional transport specific binding specification is provided to define exactly how a transport can carry NVMe traffic.

Just as PCIe-connected NVMe SSDs have many use cases and flavors (form factor, number of ports, flash type), NVMe over fabrics enables a variety of implementation types.

Several common datacenter fabric transports have emerged as favorite transports including Fibre Channel (FC), RDMA over Converged Ethernet (RoCE), and most recently TCP. Each of these has a binding specification defined. RDMA and TCP binding specifications are included in the NVMe-oF specification. FC-NVMe, is available from INCITS standards body.

While there are also proprietary solutions for sending NVMe over a fabric, in this article we'll examine those three transports and their characteristics.

UNH-IOL

▲ IOL INTERACT running NVMe Testing

Fibre Channel

Fibre Channel has been a well-known lossless transport designed for storage use cases implemented for decades. As such, it has found a home in critical storage infrastructure. While the end of FC has been predicted for many years, it's clear that as a technology, FC does what it does quite well, as there are many repeat buyers.

Adopting FC-NVMe was a critical step for the FC community, as it enabled an easy migration path from spinning disks connected using Fibre Channel Protocol (FCP), to the low-overhead, low-latency access to NVMe storage via FC-NVMe. In fact, one of the key benefits of FC-NVMe is that it can be run over the same infrastructure as traditional FCP. This means users aren't forced to do a massive rip-and-replace upgrade to their infrastructure to use FC-NVMe. Rather, existing investments in FC infrastructure can be maintained, and storage media can be incrementally upgraded to NVMe as needed.

FC-NVMe allows endpoints to negotiate the number of queues that will be enabled between initiators and targets. This is important, as it takes advantage of one of NVMe's most important characteristics, the ability to support a massive number of I/O queues. Parallel queues enable efficient use of compute and storage resources. For example, with FC-NVMe, different queues can be created for administration commands and data I/O. Further, a core compute node could be configured with an independent queue for each core to access the storage media. This means that threads operating on a given core will not have storage access blocked by storage access from another core.

This isn't a freebie. The SAN engineer needs to understand the workload needs and how to manage those queues, as well as what the FC infrastructure can support. Either way, this parallelism can be leveraged to enable efficiency and performance.

RoCE

A number of NVMe-oF products have been announced using RoCE, and have gained a following with those shops more comfortable with an ethernet-based fabric. Like FC, the queuing methodology for RoCE maps very well to NVMe, enabling minimal protocol translation as data makes it way from the network to the SSD. However, while ethernet is well known, implementation of a lossless fabric using ethernet requires turning on Data Center Bridging (DCB) protocols to handle congestion and flow control. This adds another interoperability vector, and complexity in order to tune the protocols properly to get the most out of the network with a given workload. So, for applications where latency is the primary consideration, RoCE makes sense. Of course, every hop in a network adds latency, therefore some vendors are targeting RoCE as an NVMe-oF solution for single

rack applications, and looking at TCP as a solution for aisle scale or datacenter scale applications.

TCP

The most recent addition, ratified in December 2018, of the NVMe-oF transport family is TCP, or NVMe/TCP. TCP has a few key advantages. First, TCP is a well-known protocol. Network engineers understand its behavior very well, and know how to use it. From that perspective, the 'learning curve' of enabling NVMe/TCP in the datacenter is shorter and smoother.

Next, similar to FC, TCP takes advantage of the queueing characteristics of NVMe. Each TCP connection is mapped to an NVMe queue. Thus, again, admin and I/O commands can be given separate queues so as to not block one another. Multi-threaded applications can leverage multiple CPU cores, with each core having its own queue to access storage. This simplicity in mapping in the transition from the network to within the storage array keeps latency relatively low.

Further, TCP can be run on simple switches without the extended capabilities of RoCE capable switch, which generally speaking, can add cost to a switch, but do offer performance enhancements.

Relative to RoCE, comparisons show that TCP does take a latency hit of several microseconds. However, it's important to remember that not all workloads will be sensitive to that difference. The applications needing the absolute highest performance, where cost is a secondary consideration relative to performance, may likely be better served by RoCE. Again, the IT team needs to understand their workload, and make an informed assessment of their needs.

Conclusions

NVMe-oF is very compelling for many storage-use cases, but the exact flavor of NVMe-oF used will depend on existing infrastructure, and whether the absolute lowest latency is really needed. FC-NVMe allows existing FC users to continue to use FC infrastructure, while getting the benefits of low latency streamlined access to flash. RoCE allows using ethernet networks, which are well known, lower latency than TCP, at a cost and complexity premium relative to TCP. In greenfield applications, TCP is quite attractive, for a lower cost point with a marginal latency sacrifice relative to RoCE. Managing congestion on the TCP network can go a long way to ensuring a high-performing fabric. EE

David Woolf is senior engineer, datacenter technologies at the University of New Hampshire InterOperability Laboratory (UNH-IOL).