SNIA ESF | ETHERNET STORAGE www.snia-europe.org

How Ethernet RDMA Protocols iWARP and RoCE Support NVMe over Fabrics

By David Fair, SNIA-ESF Chair, Intel, and John Kim, SNIA-ESF Member, Mellanox.

NVM Express (NVMe) is a wholly new storage protocol optimized for Non-Volatile Memory (NVM), including Flash but defined broadly enough to encompass future non-volatile memory technologies. NVMe replaces the venerable SCSI commands optimized over decades for rotating media while delivering substantially improved performance.

NVMe is architected to deliver the high bandwidth and low latency that Flash and future NVM technologies are capable of. Data centers have already begun embracing NVMe in the form of PCIe adapters and NVMe is coming to client laptops over the new M.2 form factor and connector. NVMe is supported in all major operating systems and hypervisors.

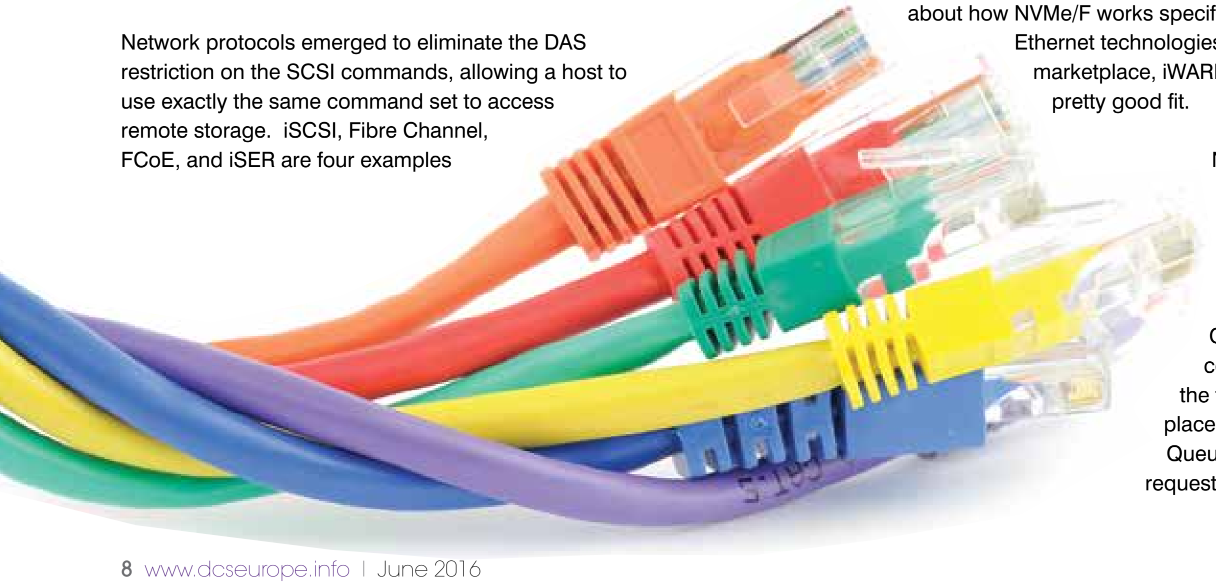
As with the original SCSI command set, NVMe is architected around a Direct Attach Storage (DAS) model, where the storage resides in the host that is accessing it. This provides easy access and fast performance to that host, but is inflexible, making it difficult to reallocate storage, cluster applications, perform failover, or migrate virtual machines from one host to another.

Network protocols emerged to eliminate the DAS restriction on the SCSI commands, allowing a host to use exactly the same command set to access remote storage. iSCSI, Fibre Channel, FCoE, and iSER are four examples of network protocols that enable a host to access remote storage with SCSI commands. The new “NVMe Over Fabrics” (NVMe/F) specification currently under development in the NVM Express Working Group does exactly the same thing for the NVMe command set. NVMe/F enables a host to access remote NVM devices across a network using the NVMe command set.

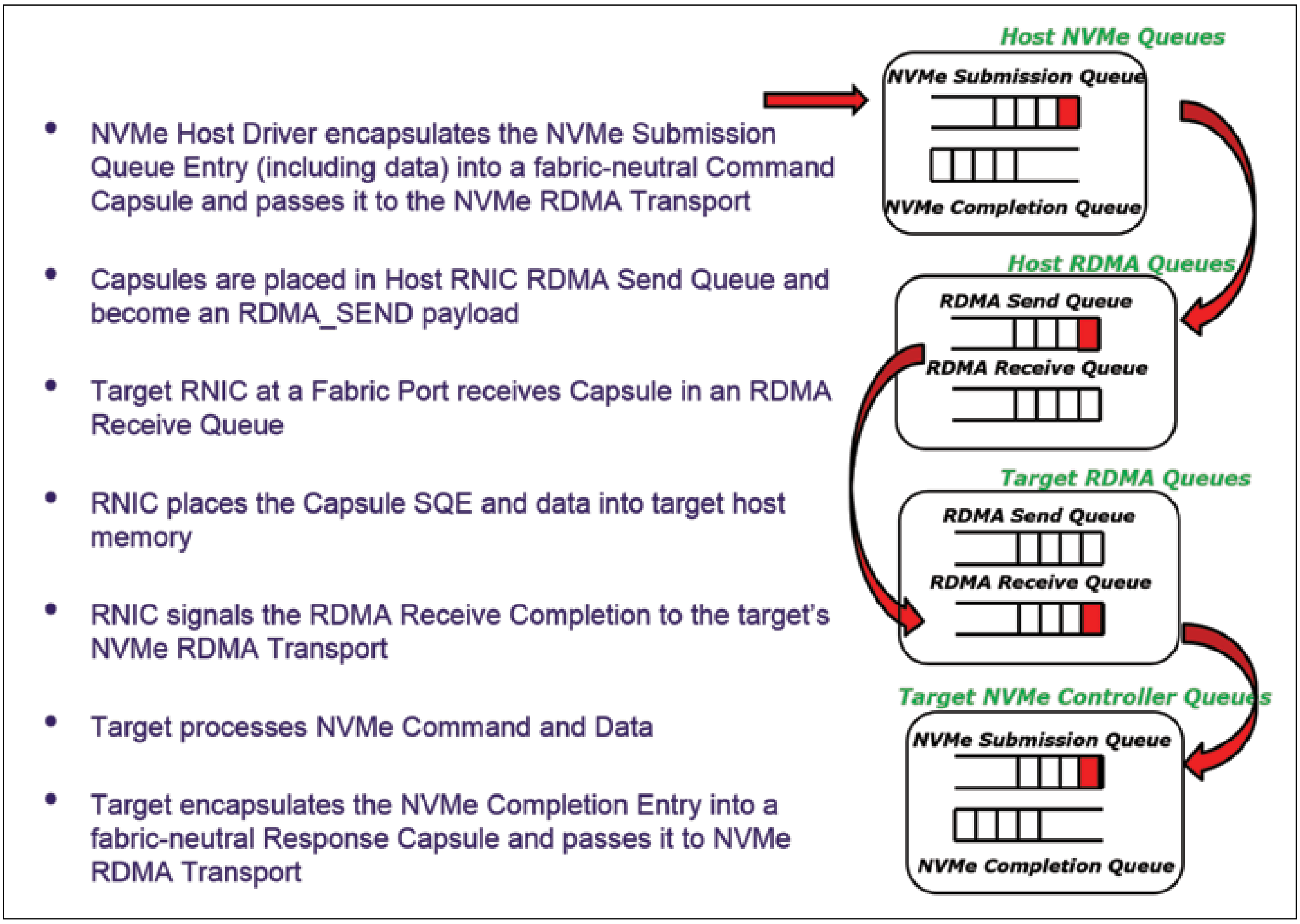
The intent of the NVMe/F specification is to be general enough to work over a range of fabrics. Initial targets are Fibre Channel, InfiniBand, iWARP Ethernet, and RoCE Ethernet with the expectation that this list will grow over time. The specification developers came up with the concept of a generic “Capsule” for commands and data that can work with all four of these transports.

A great introduction to this specification is the SNIA-ESF webcast, **https://www.brighttalk.com/webcast/679/175515**, “Under the Hood with NVMe over Fabrics.”

However, since the specification is intended to be neutral with regard to fabrics, the authors of that webcast said little specific about how NVMe/F works specifically with the two RDMA Ethernet technologies currently available in the marketplace, iWARP and RoCE. Turns out it is a pretty good fit.



NVMe uses a queueing model for reads and writes. To write to NVM, a host writes data (and other things such as control commands) to a Submission Queue in a target-side NVMe controller. Reads occur when the target-side NVMe controller places data in a host’s Completion Queue in response to a read request.



- NVMe Host Driver encapsulates the NVMe Submission Queue Entry (including data) into a fabric-neutral Command Capsule and passes it to the NVMe RDMA Transport
- Capsules are placed in Host RNIC RDMA Send Queue and become an RDMA_SEND payload
- Target RNIC at a Fabric Port receives Capsule in an RDMA Receive Queue
- RNIC places the Capsule SQE and data into target host memory
- RNIC signals the RDMA Receive Completion to the target's NVMe RDMA Transport
- Target processes NVMe Command and Data
- Target encapsulates the NVMe Completion Entry into a fabric-neutral Response Capsule and passes it to NVMe RDMA Transport

At least at a high level, this makes things map pretty nicely into RDMA Ethernet because RDMA also uses a queue model (called "queue pairs") for transmits with Send Queues and receives with Receive Queues.

The illustration from our recent SNIA-ESF webcast shows how these mappings between NVMe Queues and RDMA Queue Pairs work:

In the end, the encapsulation of NVMe commands and data over RDMA Ethernet lets the host "think" it is just using the NVMe command set to talk to local storage when in fact the storage can be remote, accessed over a high-performance RDMA Ethernet network. This is important as a factor to accelerate NVMe/F adoption because software written to take advantage of NVMe in a local device or DAS model will work with NVMe/F without any rewrites.

All that is required from a software point of view are new host NVMe/F drivers and storage target NVMe/F drivers. To help accelerate the market, the NVM Express Working Group is also developing Linux versions of these drivers alongside their specification development. So Linux will be the first system software to support NVMe/F. Presumably, the other operating system and hypervisor software vendors will follow.

To learn more, we encourage you to watch the SNIA-ESF webcast, **http://www.brighttalk.com/webcast/663/185909**, "How Ethernet RDMA Protocols iWARP and RoCE Support NVMe over Fabrics" and check out the follow-up webcast, **http://sniaesfblog.org/?p=507**, Q&A blog where we answer 30 intriguing questions from our live event.

> "To help accelerate the market, the NVM Express Working Group is also developing Linux versions of these drivers alongside their specification development. So Linux will be the first system software to support NVMe/F."