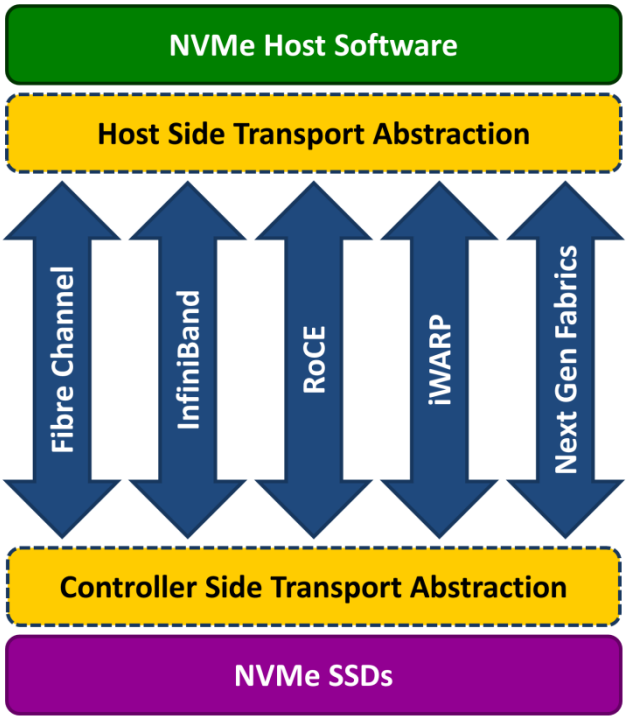
NVM Express® Moves Into The Future

NVM Express® (NVMe™) is a new and innovative method of accessing storage media and has been capturing the imagination of data center professionals worldwide. The momentum behind NVMe has been increasing since it was introduced in 2011. In fact, NVMe technology is expected to improve along two dimensions over the next couple of years: improvements in latency and the scaling up of the number of NVMe devices in large solutions.

NVMe over Fabrics

NVM Express over Fabrics defines a common architecture that supports a range of storage networking fabrics for NVMe block storage protocol over a storage networking fabric. This includes enabling a front-side interface into storage systems, scaling out to large numbers of NVMe devices and extending the distance within a datacenter over which NVMe devices and NVMe subsystems can be accessed.



Work on the NVMe over Fabrics specification began in 2014 with the goal of extending NVMe onto fabrics such as Ethernet, Fibre Channel and InfiniBand®. NVMe over Fabrics is designed to work with any suitable storage fabric technology. This specification was published in June 2016.

Two types of fabric transports for NVMe are currently under development:

- NVMe over Fabrics using RDMA
- NVMe over Fabrics using Fibre Channel (FC-NVMe)

Using RDMA with NVMe over Fabrics includes any of the RDMA technologies, including InfiniBand, RoCE and iWARP. The development of NVMe over Fabrics with RDMA is defined by a technical sub-group of the NVM Express organization.

FC-NVMe, the NVMe over Fabrics initiative relating to Fibre Channel-based transport, is being developed by the INCITS T11 committee, which develops all of the Fibre Channel interface standards. As part of this development work, FC-NVMe is also expected to work with Fibre Channel over Ethernet (FCoE).

The goal of NVMe over Fabrics is to provide distance connectivity to NVMe devices with no more than 10 microseconds (µs) of additional latency over a native NVMe device inside a server. NVMe over Fabrics solutions are expected to begin to become available in 2016.

Use Cases

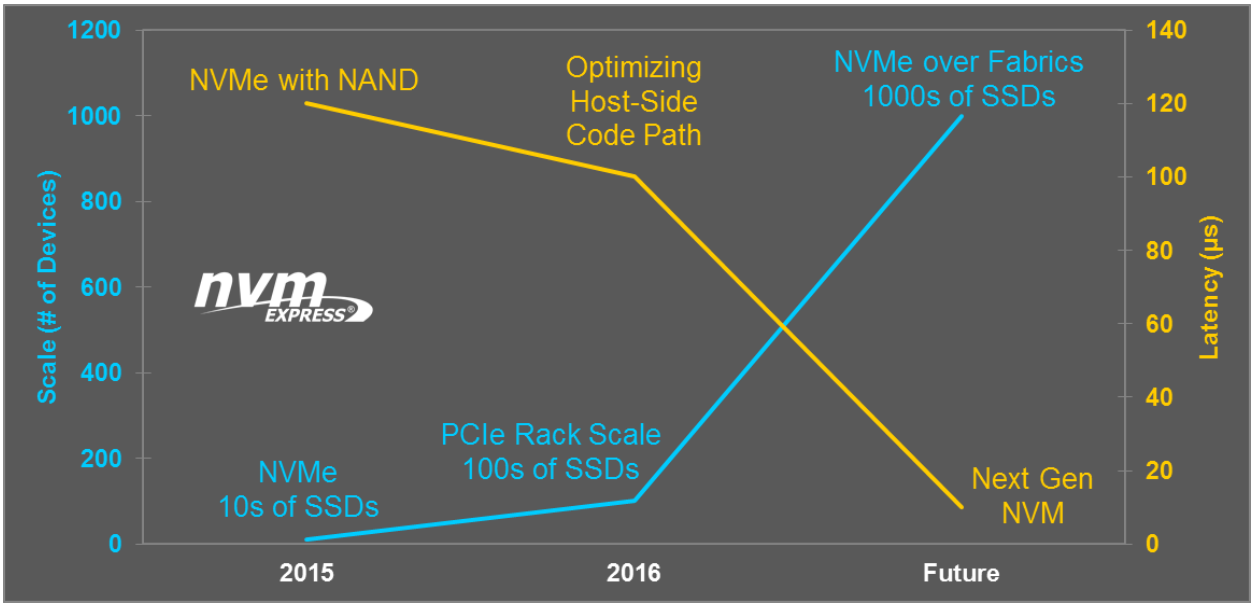
There are several use cases for NVMe over Fabrics. One can easily imagine a storage system comprised of many NVMe devices, using NVMe over Fabrics with either an RDMA or Fibre Channel interface, making a complete end-to-end NVMe storage solution. This system would provide extremely high performance while maintaining the very low latency available via NVMe.

Another implementation would use NVMe over Fabrics to achieve the low latency while connected to a storage subsystem that uses more traditional protocols internally to handle I/O to each of the SSDs in that system. This would gain the benefits of the simplified host software stack and lower latency over the wire, while taking advantage of existing storage subsystem technology.

Peering Over the Horizon into the Future

As we move further into 2016, we can expect additional optimization of the host-side code path for NVMe to become available, further reducing overall latency to approximately 100 µs as seen by the host computer. In addition, the industry will provide solutions that can scale to hundreds of NVMe devices into what might be called “rack scale” shared storage solutions.

Looking ahead into 2017 and beyond, we expect availability of the first of the “post-flash” non-volatile memory solutions that utilize the NVMe protocol running on a new type of storage media, resulting in latencies of approximately 20-25 µs as seen by the host computer. At the same time, the NVMe over Fabrics solutions will begin to provide significant scaling to thousands of NVMe devices in a large shared storage solution providing storage to hundreds or thousands of application processing cores.



NVMe over Fabrics Technical Characteristics

Obviously, transporting NVMe commands across a network requires special considerations over and above those that are determined for local, in-storage memory. For instance, in order to transmit NVMe protocol over a distance, the ideal underlying network or fabric technology will have the following characteristics:

- **Reliable, credit-based flow control and delivery mechanisms.** This type of flow control allows the network or fabric to be self-throttling, providing a reliable connection that can guarantee delivery at the hardware level without the need to drop frames or packets due to congestion. Credit-based flow control is native to Fibre Channel, InfiniBand and PCI Express® transports.
- **An NVMe-optimized client.** The client software should be able to send and receive native NVMe commands directly to and from the fabric without having to use a translation layer such as SCSI.

- **A low-latency fabric.** The fabric itself should be optimized for low latency. The fabric should impose no more than 10 µs of latency end-to-end, including the switches.
- **Reduced latency and CPU utilization adapters or interface cards.** The adapter should be able to register direct memory regions for the applications to use so that the data to be transmitted can be passed directly to the hardware fabric adapter.
- **Fabric scaling.** The fabric should be able to scale out to tens of thousands of devices or more.
- **Multi-Host support.** The fabric should be able to support multiple hosts actively sending and receiving commands at the same time. This also applies to multiple storage subsystems.
- **Multi-port support.** The host servers and the storage systems should be able to support multiple ports simultaneously.
- **Multi-path support.** The fabric should be able to support multiple paths simultaneously between any NVMe host initiator and any NVMe storage target.

The large number (64K) of separate I/O queues and inherent parallelism of these NVMe I/O queues can work well with the type of fabrics described above. Each of the 64K I/O queues can support 64K commands simultaneously, making it capable of implementation in very large fabrics. Furthermore, the small number of commands in the NVMe command set makes it relatively straightforward to implement in a variety of fabric environments.

Differences between local NVMe and NVMe over Fabrics

Approximately 90% of the NVMe over Fabrics protocol is the same as the local NVMe protocol. This includes the NVMe namespaces, I/O and administrative commands, registers and properties, power states, asynchronous events, reservations and others. The key differences are in four areas, listed in the table below.

Differences	PCI Express (PCIe)	NVMe over Fabrics
Identifier	Bus/Device/Function	NVMe Qualified Name (NQN)
Discovery	Bus Enumeration	Discovery and Connect commands
Queueing	Memory-based	Message-based
Data Transfers	PRPs or SGLs	SGLs only, added Key

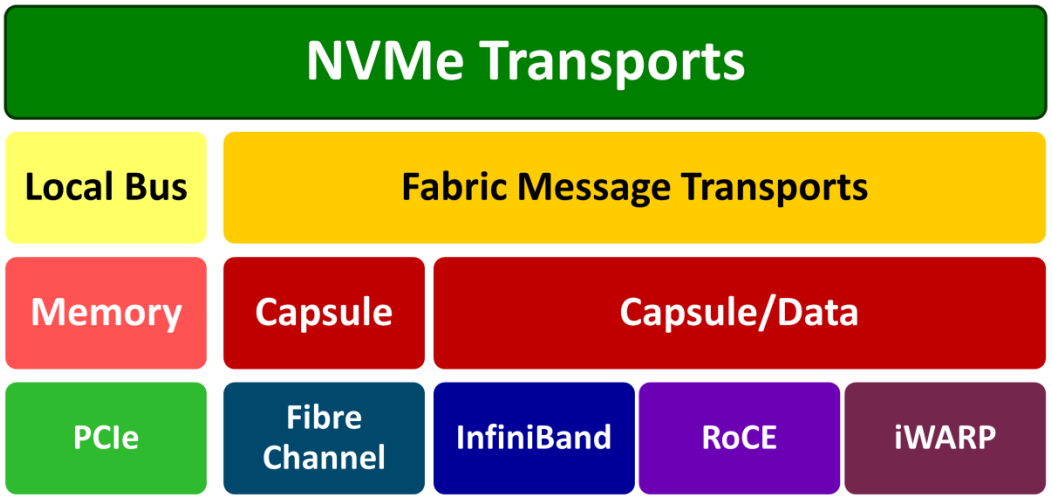
PRP: Physical Region Page (physical memory page address, PCIe transport only)
SGL: Scatter-Gather List (list of locations and lengths for read or write requests)

These differences are primarily interesting to developers of NVMe products, as their device drivers need to properly handle both local NVMe devices and remote NVMe devices. Some of these items, such as the Identifier, may be exposed to end-users to help identify specific NVMe devices for specific applications. The discovery mechanism is designed to work with multiple types of transports.

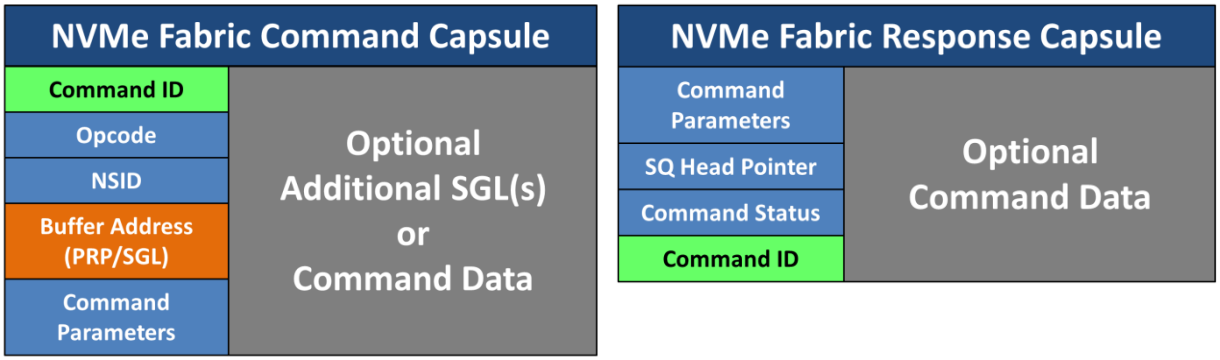
NVMe Transport Mapping

In a local NVMe implementation, NVMe commands and responses are mapped to shared memory in a host over the PCIe interface. However, fabrics are built on the concept of sending and receiving messages without shared memory between the end points. The NVMe fabric message transports are designed to encapsulate NVMe commands and responses into a message-based system by using "capsules" that include one or more NVMe commands or responses. The capsules or combination of capsules and data are independent of the specific fabric technology and are sent and received over the desired fabric technology.

For NVMe over Fabrics, the entire NVMe multi-queue model is maintained, using normal NVMe submission queues and completion queues, but encapsulated over a message-based transport. The NVMe I/O queue pair (submission and completion) is designed for multi-core CPUs, and this low-latency efficient design is maintained in NVMe over Fabrics.



When sending complex messages to an NVMe over Fabrics device, the capsules allow multiple small messages to be sent as one message, which improves the efficiency of the transmission and reduces latency. The capsule is either a submission queue entry or a completion queue entry combined with some amount of data, metadata or Scatter-Gather Lists (SGLs). The content of the elements is the same as local NVMe protocol, but the capsule is a way to package them together for improved efficiency.



NVMe Qualified Name (NQN)

One of the key benefits of a storage fabric is the inherent intelligence used to maintain consistency across all devices. In this case, NVMe over Fabrics uses a familiar qualified naming addressing convention.

The NVMe Qualified Name (NQN) is used to identify the remote NVMe storage target. It is similar to an iSCSI Qualified Name (IQN). Additional details on NVMe qualified names are described in section 7.9 of the NVMe Base specification, available at http://www.nvmexpress.org/specifications/.

Protocol	Type	Example
NVMe	NQN	nqn.2014-08.com.vendor:nvme:nvm-subsystem-sn-d78432
iSCSI	IQN	iqn.1991-05.com.microsoft:dmrtk-srvr-m

Conclusion

NVMe over Fabrics is poised to extend the low-latency efficient NVMe block storage protocol over fabrics to provide large-scale sharing of storage over distance. NVMe over Fabrics maintains the architecture and software consistency of the NVMe protocol across different fabric types, providing the benefits of NVMe regardless of the fabric type or the type of non-volatile memory used in the storage target. The next few years will be very exciting for the industry!