NVM Express over Fabrics

High Performance SSD Interconnect with NVMe over Chelsio iWARP RDMA

Introduction

NVM Express (NVMe), developed by a consortium of storage and networking companies, is an optimized interface for accessing PCI Express (PCIe) non-volatile memory (NVM) based storage solutions.

With an optimized stack, a streamlined register interface and command set designed for high performance solid state drives (SSD), NVMe is expected to provide significantly improved latency and throughput compared to SATA based solid state drives, with support for security and end-to-end data protection.

NVMe over Fabrics is a specification that extends the benefits of NVMe to large fabrics, beyond the reach and scalability of PCIe. NVMe enables deployments with hundreds or thousands of SSDs using a network interconnect, such as RDMA over Ethernet. Thanks to an optimized protocol stack, an end-to-end NVMe solution is expected to reduce access latency and improve performance, particularly when paired with a low latency, high efficiency transport such as RDMA. This allows applications to achieve fast storage response times, irrespective of whether the SSDs are attached locally or accessed remotely across enterprise or data center networks.

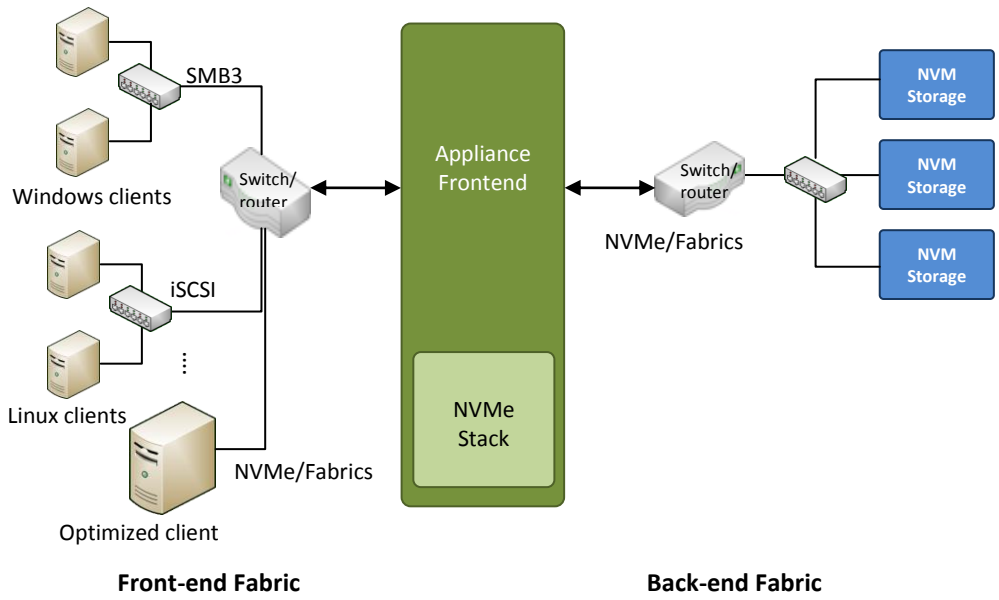

Figure 1 – NVMe over Fabrics using iWARP RDMA can share the network with iSCSI and TCP/IP

The following are key design characteristics of NVMe over Fabrics:

- Transport neutral abstraction, with RDMA as first use case
- Enhanced status reporting and expanded capabilities including live firmware updates
- No translation to or from another protocol such as SCSI eliminates overheads and latency

- Streamlined control and data transfer command set
- Support for efficient parallel I/O in multi-core systems, with up to 64K queues and native support for MSI-X interrupts
- Direct encapsulation for efficient operation over RDMA fabrics

As a result, **NVMe over Fabrics** delivers the following benefits:

- Sustained performance for data-intensive workloads, such as real-time data analytics
- Low latency and high IOPS with reduced CPU and memory utilization
- Scalable storage with direct-attach level performance

In addition, by using iWARP RDMA, NMVe over Fabrics can seamlessly share a large routed network with iSCSI and other traffic as it uses the same reliable and robust TCP/IP foundation.

The Chelsio Terminator 5 ASIC

The Terminator 5 (T5) ASIC from Chelsio Communications, Inc. is a fifth generation, high-performance 2x40Gbps/4x10Gbps server adapter engine with Unified Wire capability, allowing offloaded storage, compute and networking traffic to run simultaneously. T5 provides extensive support for stateless offload operation for both IPv4 and IPv6. It is a fully virtualized NIC engine with separate configuration and traffic management for 128 virtual interfaces, and includes an on-board switch that offloads the hypervisor v-switch.

With integrated, standards based FCoE, iSCSI and RDMA offload, T5-based adapters are high performance drop in replacements for FibreChannel storage adapters and InfiniBand RDMA adapters. Thanks to its TCP/IP based foundation, T5's iWARP RDMA support is a robust, reliable, routable plug-and-play RDMA solution that works over any Ethernet network, from LANs to long distance WAN links, and is a native to private and public clouds.

NVMe over Fabrics Prototype

A prototype of NVMe over Fabrics using Chelsio's T580 2x40Gbps iWARP RDMA NIC was demonstrated by Intel at the Intel Developers Forum, September 9-11 2014 in San Francisco. The prototype delivered identical performance (450K IOPs) for both the local and remote NVMe devices (up to 7x the performance of a SATA SSD), showing no negative impact in accessing remote storage.

Additionally, accessing the remote NVMe device only added 8µsec round trip time latency compared to local NVMe access, handily meeting the goal of 10µsec. Further optimization is expected to reduce this latency further.

The following figure shows the demo setup, with a client and target system connected back-to-back. The demo used Intel P3700 NVMe SSD drives, and compared the direct attach latency and IOPS to those achieved by going over the RDMA transport, through the RNICs on both ends, to the host memory where it was handed down to the remote SSD. The responses traversed the same path in the reverse direction.

Latency through the whole path tallied up to less than 8µsec, including two traversals of the RNICs in each direction, four traversals of the PCI bus for the RDMA operations, in addition to the wire latency.

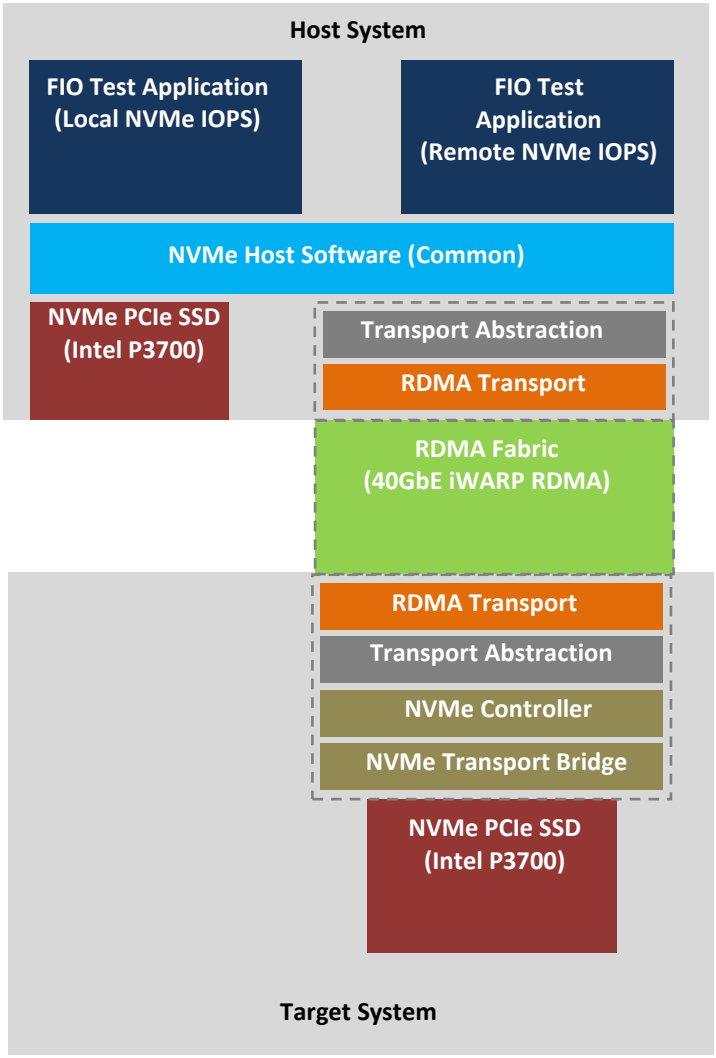

Figure 2 – NVMe over Fabrics Prototype

Summary

This paper provided a brief introduction to **NVMe**, an optimized standard interface for high performance SSD storage, and **NVME over Fabrics**, an NVMe extension for scalable high performance storage. Today, Chelsio is at the leading edge of the NVMe over Fabrics wave, with the first prototype demonstration of the technology shown by Intel at the 2014 IDF using Chelsio's high performance iWARP RDMA over 40Gbps Ethernet solution. This solution has thus been shown to deliver the promise of scalable high performance storage over simple, standard Ethernet gear. Chelsio's iWARP RDMA provides a plug-and-play solution for connecting high performance SSDs over a scalable, congestion controlled and traffic managed fabric, with no special configuration needed.

Related Links

[iWARP: From Clusters to Cloud RDMA]
[SMBDirect 40 GbE iWARP vs 56G InfiniBand]
[Windows Server 2012 R2 SMB Performance]
[The Chelsio Terminator 5 ASIC]