Consideration for adopting NVMeF for Enterprise Storage

Sanjeev Kumar
Software Product Engineering, HiTech, Tata Consultancy Services

26 May 2017
Agenda

1. Current Storage Architecture & Network limitation
2. NVMe Over Fabric Solution
3. Comparison for different NVMe fabric
4. Requirements for running End to End NVMeF solutions
Recap of Communication Protocols

- SAS and SATA performance increased over time but protocols have not changed much.
- NVMe was created to allow direct access to the logical block based architecture of SSDs and to do highly-parallel IO enabling SSD to execute more IO threads than any other protocol before.
Traditional SAS/SATA Storage Architecture

Storage Controller

HBA

PCIe

FC

NIC HBA

CPU

SAS/SATA Controller

SAS

SAS Expender

Disk Array

SATA Disk

SATA Disk

SSD

SAS Disk

SAS Expender

SSD
Next Gen NVMe Storage Architecture
Data Flow with Enterprise Storage Over Network: Limitations

Protocol conversion bridge is required to access the data over network which increases the NVMe latency.
Why NVMe Over Fabric Solution?

- Defines a common architecture that supports a range of storage networking fabrics for NVMe block storage protocol over a storage networking fabric.
- No translation to or from another protocol like SCSI.
- Inherent parallelism of NVMe multiple I/O Queues is exposed to the host.
- NVMe commands and structures are transferred end-to-end.
- Maintains the NVMe architecture across a range of fabric types.
- Maintains architecture and software consistency between fabric types by standardizing a common abstraction and encapsulation definition.

Design goal of NVMe over Fabrics:
Provide distance connectivity to NVMe devices with no more than 10 microseconds (μs) of additional latency.
- Standardization began in 2009

- Standardizes the interface between CPU and PCIe attached SSD

- Version 1.0 was released in 2011. Version 1.3 launched on 1st May 2017. Complete versions are publically available at http://www.nvmexpress.org/specifications/

- Tuned for performance and strong support for many CPU architecture and OS.
Working Principle of NVMe Protocol

1. Host writes command to submission queue
2. Host writes updated submission queue tail pointer to doorbell
3. Controller fetches command
4. Controller processes command
5. Controller writes completion to completion queue
6. Controller generates MSI-X interrupt
7. Host processes completion
8. Host writes updated completion queue head pointer to doorbell

NVMe commands is just 64 bytes of data in your memory while response comes in 16 byte

Source: nvmexpress.org
Two types of fabric transports for NVMe are currently under development:

- NVMe over Fabrics using RDMA (Infiniband, iWARP, RoCE)
- NVMe over Fabrics using Fibre Channel (FC-NVMe)

Source: nvmexpress.org
What is RDMA (Remote Direct Memory Access)?

- RDMA is a combination of hardware, OS code and user space software that enables efficient, high-performance access to remote memory.

- RDMA presents a standard programming interface to user and then uses hardware to perform memory access at distance and scale.

- RDMA runs over a variety of different physical layers and can sit on top of TCP/IP (iWARP), Ethernet (RoCE) and Infiniband.

- Initially developed for HPC, it is now becoming popular in data center environment.

- Applications communicate with the RDMA NIC using dedicated Queue Pairs (QPs) and Completion Queues (CQs):
  - Each application can have many QPs and CQs.
  - Each QP has a Send Queue (SQ) and Receive Queue (RQ).
How does RDMA works?

1. Client establish connection to server using rdma command
2. Both client and server register memory region on their own DRAM
3. Client and server exchange permission and security information on those memory regions
4. Using client and server then ping-pong incrementing data back and forth between two memory regions
NVMe Over fabric Protocol – An Overview

Source: nvmexpress.org
NVMe Over fabric commands flow

- Host fetch the 64 Bytes of commands from Driver
- Host creates the capsules that contain the commands and optional data
- Capsules are transported over relevant fabric (FC or RDMA)
- Capsules support admin commands, IO commands (NVMe commands) and fabric specific commands
- When data is not contained in the capsules, the command contains information like key/offset to locate that data on host

* 90% of NVMe Over Fabric commands are same as NVMe

Source: nvmexpress.org
How Data Flows from Host to Target in NVMeF?

Host
- Linux File System
- Linux Block IO
- NVMe Device Driver
- NVMe Fabrics
- RDMA Verbs
- RDMA Driver
- RDMA Offload
- TCP Offload

Target
- NVMe Fabrics
- RDMA Verbs
- RDMA Driver
- NVMe Device Driver
- NVMe Drive
- NVMe Drive

Kernel
- NIC

Copyright © 2017 Tata Consultancy Services Limited
End to End NVMe Over Fabric Solution for Enterprise Storage

Flash Storage

NVMe Over Fabric

NVMe Subsystem

NVMe Bridge

PCIe

NVMe SSD

PCIe

NVMe SSD

PCIe

NVMe SSD

PCIe

NVMe SSD

Front End Fabric

Back End Fabric

Linux Host

Windows Host

NVMe Optimized Host

SCSI

iSCSI

SMB

FC

iSCSI

iSER

SRP

FC

NVMe

Over Fabric

NVMe

Bridge

NVMe Subsystem

NVMe Subsystem

NVMe Subsystem

Copyright © 2017 Tata Consultancy Services Limited
### Matrix for best suitable fabric

<table>
<thead>
<tr>
<th>Fabric</th>
<th>Vendors</th>
<th>Transport</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infiniband</td>
<td>Mellanox</td>
<td>Infiniband</td>
<td>Lowest latency</td>
<td>Highest Cost</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>High security</td>
<td>Low Volume</td>
</tr>
<tr>
<td>iWARP</td>
<td>Chelsio</td>
<td>TCP/IP</td>
<td>Cheap</td>
<td>Highest Latency</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Not Scalable</td>
</tr>
<tr>
<td>RoCE/Routable</td>
<td>Mellanox</td>
<td>Converged Ethernet</td>
<td>Datacenter - Preferred</td>
<td>Non-Legacy Equipment</td>
</tr>
<tr>
<td>RoCE(V2)</td>
<td>Avago</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cavium</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FC</td>
<td>Cisco</td>
<td>Fiber Chanel</td>
<td>Full Compatible with SCSI and FC</td>
<td>Upgraded switch and HBA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>High security</td>
<td></td>
</tr>
</tbody>
</table>
### NVMe Over Fabric supported Products – Sample List

<table>
<thead>
<tr>
<th>Arrays</th>
<th>Adapters</th>
<th>Reference Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supermicro</td>
<td>Mellanox supports RoCE</td>
<td>Seagate</td>
</tr>
<tr>
<td>Mangstor</td>
<td>Chelsio supports iWARP</td>
<td>WD</td>
</tr>
<tr>
<td>E8 Storage</td>
<td>Qlogic supports iWARP and RoCE</td>
<td>Toshiba</td>
</tr>
<tr>
<td>Pavillion Data</td>
<td></td>
<td>Micron</td>
</tr>
<tr>
<td>Excelero</td>
<td></td>
<td>Kingston</td>
</tr>
<tr>
<td>Aperion</td>
<td></td>
<td>Samsung</td>
</tr>
</tbody>
</table>
"Flash is technology of choice for Storage, NVMe is protocol for Flash, Storage network is the new bottleneck where NVMe Over Fabric is the solution “

Moving on NVMeF Solution require:

- NVMe Over FC Requirement
  - Fibre Channel Gen 5 and Gen 6 switches supported, full compatibility with SCSI & NVMe over FC
  - Generation 6 HBA’s with new devices drivers required to support NVMe over fabrics, concurrently along with SCSI

- NVMe Over RDMA Requirement
  - iWARP requires iWARP specific RDMA NIC’s and device drivers
  - InfiniBand requires both IB HBA and IB switches
  - RoCE requires DCB Ethernet switches, along with driver support in NICs

Plenty of demos for NVMe over Fabric supported products has been done in Flash Memory Summit and Intel Developers Forum in 2016 where preferred choice of fabric was RoCE solution.
Q&A
mail us @ sanjeev24.k@tcs.com
Thank You