Next Generation Low Latency SAN

Rupin Mohan, Hewlett-Packard

Craig Carlson, Qlogic

April 7th,2015

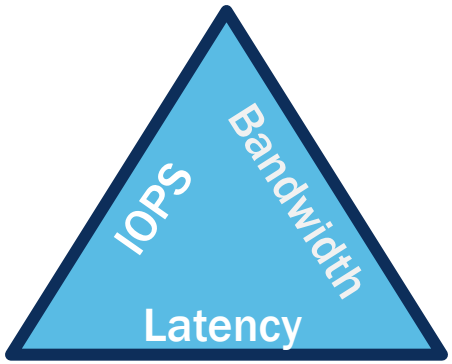
Santa Clara, CA

Agenda

- Introduction
- How do we measure performance?
- State of the union
- Storage Protocol Comparison
- What are the new options?
- A look into the future

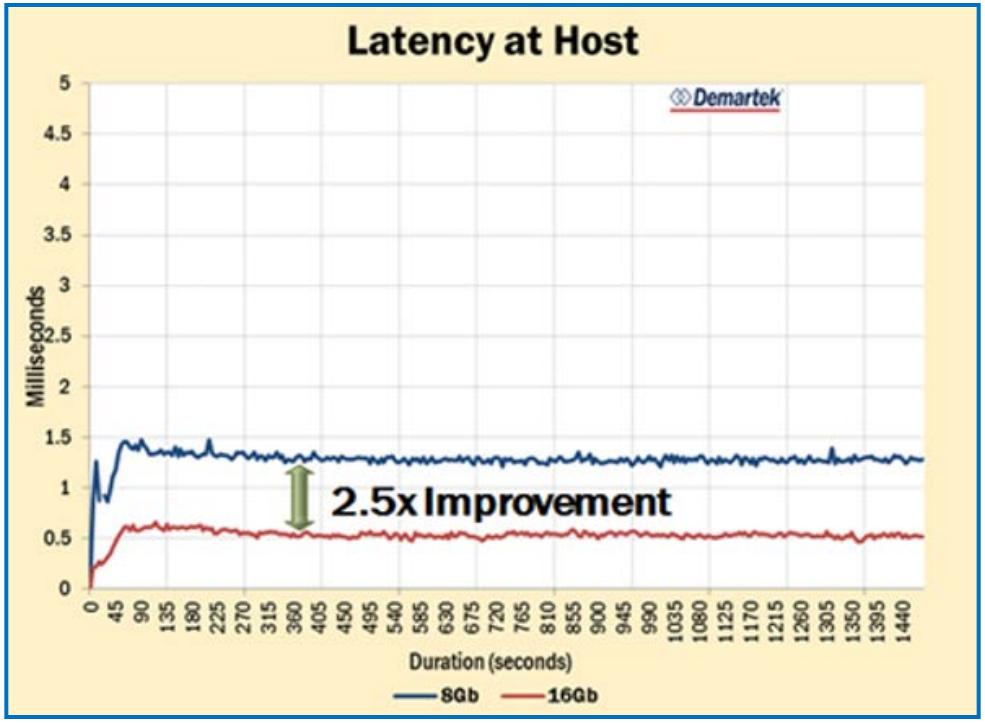
HOW DO WE MEASURE PERFORMANCE?

- **IOPS** – I/O's per second – a measure of the total I/O operations (reads and writes) issued by the application servers.
- **Bandwidth** – a measure of the data transfer rate, or I/O throughput, measured in bytes per second or MegaBytes per second (MBPS).
- **Latency** – a measure of the time taken to complete an I/O request, also known as response time. This is frequently measured in milliseconds (one thousandth of a second). Latency is introduced into the SAN at many points, including the server and HBA, SAN switching, and at the storage target(s) and media.



The application/user experience

State of the Union



- Using All Flash **3PAR 7450**
- End to end **Gen 5** (16Gb) Infrastructure
- Latency at host (end-to-end) around ~.5 ms
- 75% reduction in latency compared to 8Gb FC

What are the new options?

- Gen 6 Fibre Channel
- Storage protocols on RDMA
  - iSER
  - SMB Direct
- Transport options for RDMA
  - RoCE
  - iWARP
- NVMe over Fabrics
  - NVMe over RDMA
  - NVMe over FC

Gen 6 Fibre Channel

- 32 Gb/s single lane
- 256 Gb/s multi lane
- How is it lower latency?
  - Lower latency through higher clock rate
  - Possible smaller ASIC geometries

RDMA – Remote Direct Memory Access

- Introduction
  - Accelerated IO delivery model, allowing application software to bypass most layers of software and communicate directly with the hardware
  - Requires new programming model: “verbs” rather than “sockets”
  - Protocol options: Block (iSER) / File (SMB Direct)
  - Transport options: RoCE, iWARP, Infiniband
- RDMA Benefits
  - Low latency, also important is latency jitter
  - High Throughput
  - Zero copy capability, OS / Stack bypass
  - Avoid CPU context switching, interrupt coalescing
- Bulk of the work is done by the Target
  - Read operation, translates to a write by Target
  - Write operation, translates to a read by Target

iSER and SMB Direct over Ethernet

- iSER
  - iSCSI extensions over RDMA
  - Mature protocol
  - Limited OS stacks. Needs to be hardened.
- SMB Direct
  - NFS using Direct I/O
  - Mature protocol
  - Limited OS stacks.

RoCE versus iWARP

RoCE

1. Needs DCB Switching infrastructure
2. Routability being worked on with v2. We need this for a storage network application
3. End to end congestion management is a big issue.
4. Cost: Higher cost solution
5. DCB configuration on switches is cumbersome.

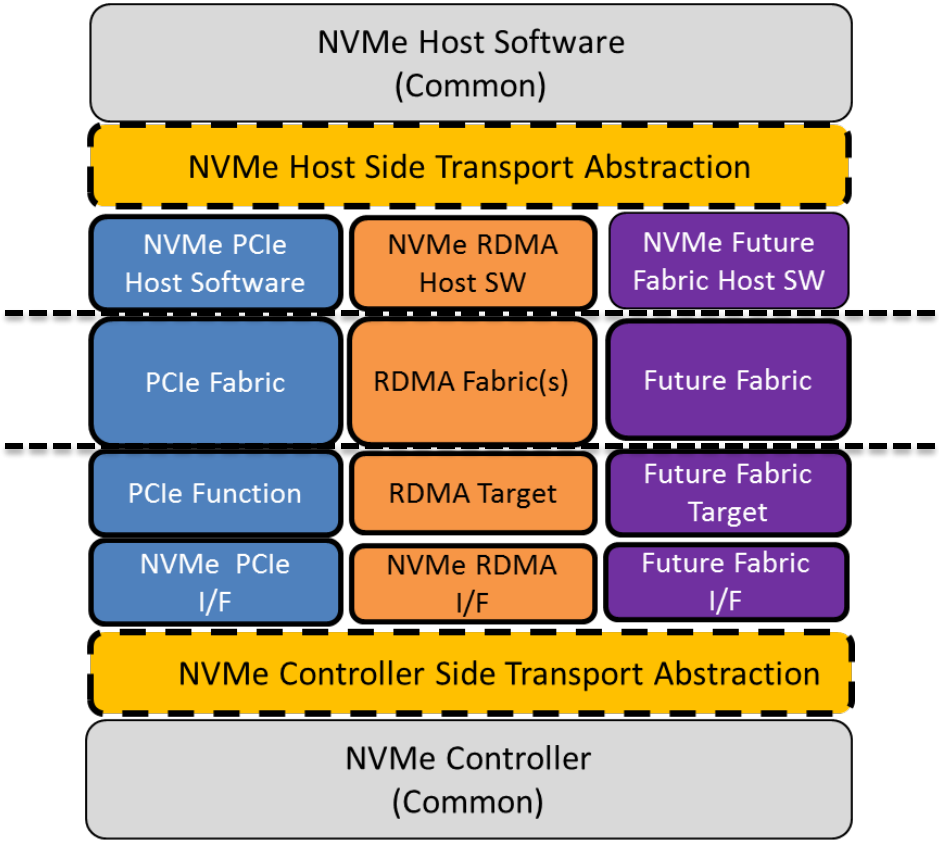
iWARP

1. Does not require DCB Switches
2. Routable as it runs over TCP/IP
3. TCP solves the congestion management issue but adds latency
4. Lower cost solution
5. Switch configuration is simpler

NVMe Transports

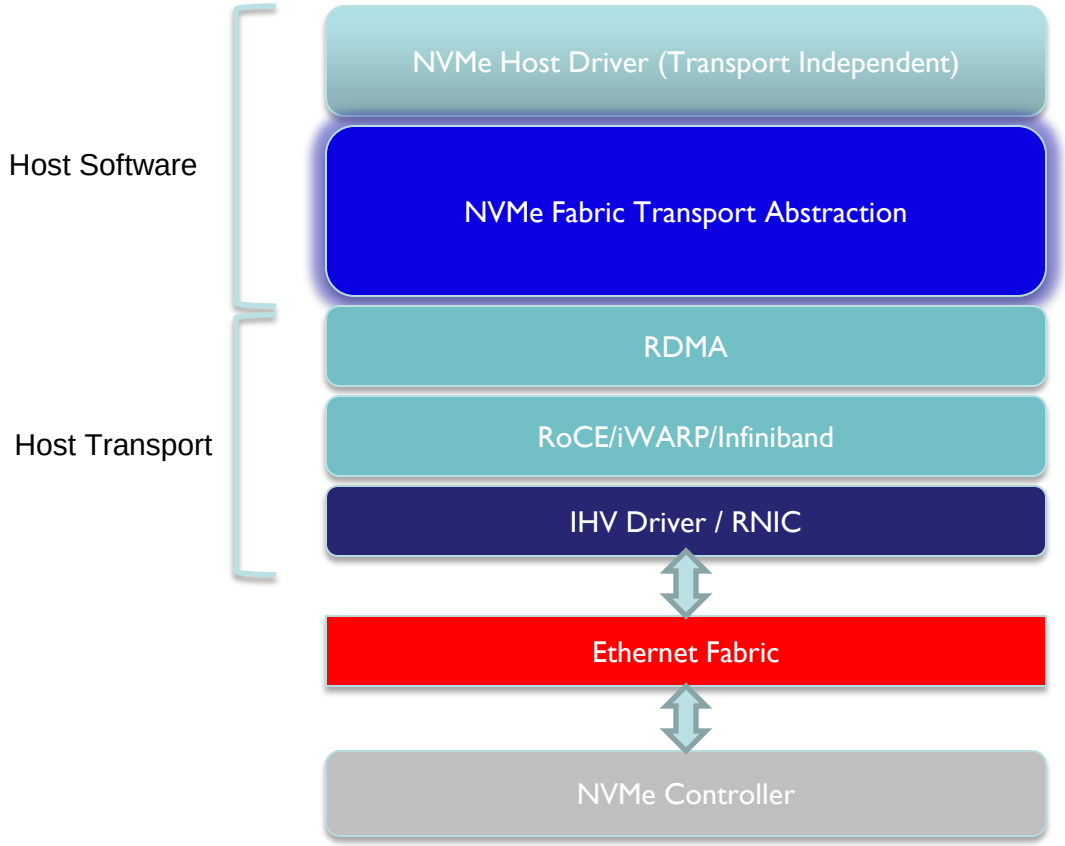
- Two new fabric transport projects for NVMe
  1. NVMe over Fabrics – Being defined in a subgroup of the NVM Express group
     - NVMe over RDMA
  2. NVMe over FC (FC-NVMe) – New T11 project to define an NVMe over Fibre Channel Protocol

NVMe over Fabrics



- Being defined by a Technical sub-group of the NVM Express group
  - Defined as upper level protocol on top of OFED RDMA verbs
  - Fabric agnostic
    - Supports RDMA fabrics – Ethernet (iWARP, RoCE), Infiniband
    - Support for other fabrics – FC
  - Complies to the NVMe programming model

NVMe over Fabrics host software



- Need to define a standard for
  - Naming of Endpoints
  - Discovery of Targets
  - Command Flow Control
  - Error Recovery
  - Authentication

FC-NVMe definition

- T11 is working on a mapping of NVMe over Fabrics on Fibre Channel
  - New T11 Project FC-NVMe
- Use FCP semantics for NVMe over Fabrics transport
  - Use existing accelerated paths
  - FCP defines the exchange protocol – Host interface is still NVMe
  - FCP exchanges similar to RDMA exchanges
- Map NVMe over Fabrics discovery mechanisms into FC
  - Reuse FC mechanisms where it makes sense
    - PRLI
    - FC Name Server
  - Use native FC mechanisms

Key Takeaways

- Fibre Channel is still very low latency
- iSER and SMB Direct are in product development right now
- Gen 6 Fibre Channel is in product development, FC is also working on lowering latency
- NVMe over Fabrics is under standards development now
- Data center networks are key to application latency

Thank You

- Rupin Mohan
  - Rupin.mohan@hp.com
  - +1-774-245-2947
- Craig Carlson
  - Craig.Carlson@qlogic.com
  - +1-612-860-7878