NVMe over Fabrics

September 20th, 2017

Brandon Hoff, Broadcom

Brandon.Hoff@Broadcom.com
Agenda

- NVMe over Fabrics
  - Update
  - Market
  - Roadmap
  - NVMe-TCP
- The benefits of NVMe over Fabrics
- Transport Independent - Storage Architectures
- Use Cases
The Value of Shared Storage and the ‘need for speed’

- The cost of data-at-rest is no longer the right metric for storage TCO
  - The value of data is based on how fast it can be accessed and processed

- NVMe over Fabrics increases the velocity of data
  - Faster storage access enables cost reduction through consolidation
  - Faster storage access delivers more value from data

- SSDs are going to become much faster
  - 3D Xpoint Memory, 3D NAND, etc.
  - PMEM, Storage Class Memory, etc.
  - … and innovation will continue
NVMe for Fabrics Progress (NVMe-oF)

NVMe-oF Developments (NVM Express)

- **6/6/16** – Fabrics Working group published v1.0
- **6/6/16** – NVMe Fabrics Linux Driver working group published Linux target and driver code for inclusion in Linux kernel.org
- **8/8/16** – NVMe over Fabrics Panel at FMS 2016
- **9/7/17** – NVMe over Fabrics Roadmap released and NVMe-TCP announced at FMS 2017
- **2019** – NVMe-oF.next

Additional Enterprise features being developed
- Boot from SAN
- Multi-pathing for failover & load balancing
NVMe over Fabrics delivers for the External Block Storage Market

| NVMe-oF 1.0 was released in June 2016 and provides support for RDMA and Fibre Channel, plus NVMe-TCP with 1.1 | Only 13% of storage capacity shipped is DAS (inside the server), 87% of the total storage capacity shipped is external storage. |
| All Flash Arrays is a **$6.8B Market** in 2017, growing at a **32% CAGR.** | As NVMe becomes adopted, NVMe-oF will enable applications access to 1000’s of NVMe drives with FC, RoCE, iWARP, IB, or TCP as transport options. |
What Drives AFA Purchases? (Top 5)

From the list below, please, select up to three most important criteria when purchasing/considering AFA

- Reliability: 56%
- Performance: 40%
- Scalability (as measured by effective capacity): 31%
- Performance consistency (in the face of varying I/O workloads): 30%
- Ability to integrate with pre-existing datacenter workflows (APIs, etc.): 27%

Source: IDC, All-Flash Array Adoption
NVMe Roadmap

NVMe-oF

1.2 and prior: Building features baseline

NVMe 1.2 Nov’14
- Namespace Management
- Controller Memory Buffer
- Host Memory Buffer
- Active/Idle Power and RTD3

NVMe-oF 1.0 May’16
- Transport and protocol
- RDMA binding

Fabric: Beyond PCIe

Base features: enterprise, client, mobile, and cards

NVMe 1.2.1 May’16
- Virtualization Enhancements
- Directives/Streams
- Sanitize
- Boot Partitions

NVMe 1.3 May’17
- IO Determinism
- Async. Namespace Access
- Persistent Memory Region

NVMe (next)*
- Enhanced Discovery
- Authentication
- TCP Transport

Inflection Point
- Hyper-scale
- Multipath
- New Fabrics

NVMe-MI

NVMe-MI 1.0 Nov’15
- Out-of-band management
- Device discovery
- Health & temp monitoring
- Firmware Update

Establish Management standard

NVMe-MI 1.0a Apr’17
- Errata fixes

NVMe-MI 1.1*
- SES Based Enclosure Mgmt
- In-band NVMe-MI
- Storage Device Enhancements

NVMe (next)*
- Subject to change

Specification Release

Planned Specification Release

2014 Q1 Q2 Q3 Q4
2015 Q1 Q2 Q3 Q4
2016 Q1 Q2 Q3 Q4
2017 Q1 Q2 Q3 Q4
2018 Q1 Q2 Q3 Q4
2019 Q1 Q2 Q3 Q4

* Subject to change
NVMe-oF coming to a network near you

- NVMe-oF V1.0 enabled efficient end-2-end NVMe over RDMA and Fibre Channel networks
  - RDMA because of its high efficiency and similar architecture characteristics
  - FC because of its reliable credit based flow control and delivery mechanism
- What about existing IP network infrastructures?
NVMe-TCP Transport

• Enables the use of NVMe-oF over existing Datacenter IP networks
• Supports all of the NVMe-oF and NVMe Architecture features
• Layered over standard IETF TCP transport to allow software-only and/or hardware (accelerated/offloaded) implementations
Simplicity of NVMe over Fabrics

- NVMe-oF delivers a new level of performance for today’s business-critical applications
- NVMe-oF is, by design, is transport agnostic:
  - Application developers can write to a single block storage stack and access NVMe over Fibre Channel, TCP, or RDMA networks
- Data is DMA’d in and out of the adapters to maximize performance
  - Zero copy is available today for Fibre Channel and RDMA protocols for improved performance and the new TLDK (new) for TCP can provide zero copy for TCP
NVMe over Fabrics – Storage Architectures

**End to NVMe and NVMe-oF Solutions**

- Front-side Fabric
- NVMe-oF
- Storage Controllers
- Back-side Fabric
- NVMe-oF
- NVMe JBOF/EBOF

**Traditional SAN**

- Front-side Fabric
- NVMe-oF
- Enterprise Arrays

**Server SAN/Disaggregated Storage**

- Front-side Fabric
- NVMe-oF
- e.g. Rows of servers with ~20 disks per unit

**Rack Scale/Scaleout/HyperScale**

- Front-side Fabric
- NVMe-oF
- Blocks of Storage
- Blocks of Compute
Evolving the Fabric for a New Generation of Enterprise Applications

Use Case 1: Faster Flash

Accelerated Enterprise Data Fabric

Faster Flash

- Lower latency for NVMe-oF versus legacy SCSI
- 25GE/32GFC 100GE/128GFC Bandwidth to support 32G (PCIe) and faster NVMe drives
- Lower CPU Utilization
- Scales to 100’s and 1000’s of drives making more capacity available to applications
Benefits of NVMe over Fabrics for SQL Workloads

**Test Setup**
- Standard Servers
- SLES 12 SP3

**Results**
- **Performance**
  - NVMe-oF: 48,354
  - SCSI: 28,342
  - **2x improvement**

- **% CPU utilization**
  - NVMe-oF: 78
  - SCSI: 81

**Diagram Notes**
- PostgreSQL
- XFS FS
- NVMe-oF Controller
- Intel P3608 NVMe Drive
- NVMe-oF Controller
- NVMe-oF or SCSI Initiator
- NVMe-oF or SCSI Target

**Legend**
- NVMe-oF
- SCSI
- DB Transactions/sec

**Source**
- Pgbench: Read only Transaction workload, source Broadcom
NVMe over Fabrics Concurrent IO Use Case

Use Case 2: Staged Analytics on Real-World Data Sets

- Database app maintains high value database on high SLA legacy array
- Data mining app requires super low latency reference image of DB
- Regularly Snapshot DB in legacy array
- Use Data Analytics server to copy snapshot to Ultra-low latency NVMe-oF copy
- Run Data Analytics application using low latency NVMe-oF reference copy
Integrating New Storage Technologies into the Datacenter

• Storage Class Memory (SCM) as a Cache
  – Adding Flash into a DIM to support low latency storage for NVMe over Fabrics

• Data Management for In-Memory Databases
  – Adding data management for business-critical data
  – Replication, deduplication, Tiering data off, etc.

“Storage-class memory technologies could alter price/performance and density of current storage and compute platforms while bringing both benefits and uncertainties to I&O leaders responsible for IT infrastructure planning,”

Gartner
Creating the Fabric for a New Generation of Enterprise Applications

Use Case 3: **Storage Class Memory as a storage array cache**

**New Media as Cache**

- NVMe over Fabrics

**SCM as Cache**

- Improved performance: 4x more IOPs
  - Leverage 32G NVMe
  - Benefit from low latency media
- NVMe over Fabrics fits well
- Highest performance, highest bandwidth, lowest latency, storage target available
  - Removes PCIe latency in storage for caching/fast storage
NVMe over Fabrics
Storage Class Memory (SCM) As a Cache

Enterprise Data Management everywhere

- New Storage Tier: In-Memory DBs Databases
- NVMe over Fabrics enables
  - Snapshots
  - Data Tiering (e.g. key for HANA)
  - Availability
  - Workload Migration
- Eliminates another silo in the datacenter
- Eliminates stranded storage

In Memory Database
NVMe over Fabrics

Creating the Fabric for a New Generation of Enterprise Applications

Use Case 4: Data Management for in-memory databases
Resources from NVM Express and FMS

NVMe over Fabrics 2016 Webinar:

NVMe over Fabrics 2017 FMS Presentation(s):

NVMe over Fabrics Standard:
http://www.nvmexpress.org/resources/specifications/

NetApp’s Presentation on SCM and PMEM from FMS:
https://www.youtube.com/watch?v=1fQlJusiVdE