NVM
PCIe® Networked Flash Storage

Peter Onufryk
Microsemi Corporation
PCI Express (PCIe)

- Specification defined by PCI-SIG
  - www.pcisig.com
- Packet-based protocol over serial links
  - Software compatible with PCI and PCI-X
  - Reliable, in-order packet transfer
- High performance and scalable from consumer to Enterprise
  - Scalable link speed (2.5 GT/s, 5.0 GT/s, 8.0 GT/s, 16 GT/s, and 32 GT/s)
    - Gen5 (32 GT/s) is still being standardized
  - Scalable link width (x1, x2, x4, ..., x32)
- Primary application is as an I/O interconnect
PCIe Characteristics

- Scalable speed
  - Encoding
    - 8b10b: 2.5 GT/s (Gen 1) and 5 GT/s (Gen 2)
    - 128b/130b: 8 GT/s (Gen 3), 16 GT/s (Gen4) and 32 GT/s (Gen5)

- Scalable width: x1, x2, x4, x8, x12, x16, x32

<table>
<thead>
<tr>
<th>Generation</th>
<th>Raw Bit Rate</th>
<th>Bandwidth Per Lane Each Direction</th>
<th>Total x16 Link Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen 1*</td>
<td>2.5 GT/s</td>
<td>~ 250 MB/s</td>
<td>~ 8 GB/s</td>
</tr>
<tr>
<td>Gen 2*</td>
<td>5.0 GT/s</td>
<td>~500 MB/s</td>
<td>~16 GB/s</td>
</tr>
<tr>
<td>Gen 3*</td>
<td>8 GT/s</td>
<td>~ 1 GB/s</td>
<td>~ 32 GB/s</td>
</tr>
<tr>
<td>Gen 4</td>
<td>16 GT/s</td>
<td>~ 2 GB/s</td>
<td>~ 64 GB/s</td>
</tr>
<tr>
<td>Gen 5</td>
<td>32 GT/s</td>
<td>~4 GB/s</td>
<td>~128 GB/s</td>
</tr>
</tbody>
</table>

*Source – PCI-SIG PCI Express 3.0 FAQ
NVM Express™ (NVMe™)

- Two specifications
  1. NVM Express (PCIe)
  2. NVM Express over Fabrics (RDMA and Fibre Channel)
- Architected from the ground up for NVM
  - Simple optimized command set
  - Fixed size 64 B commands and 16 B completions
  - Supports many-core processors without locking
  - No practical limit on the number of outstanding requests
  - Supports out-of-order data deliver

PCIe SSD = NVMe SSD
## Ideal NVM Fabric

<table>
<thead>
<tr>
<th>Property</th>
<th>Ideal Characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>Free</td>
</tr>
<tr>
<td>Complexity</td>
<td>None</td>
</tr>
<tr>
<td>Performance</td>
<td>High</td>
</tr>
<tr>
<td>Power consumption</td>
<td>None</td>
</tr>
<tr>
<td>Standards-based</td>
<td>Yes</td>
</tr>
<tr>
<td>Scalability</td>
<td>Infinite</td>
</tr>
</tbody>
</table>
PCle Fabric

PCle Switch

NVMe Host 1

NVMe Host 2

NVMe Host 3

NVMe Host 4

NVMe SSD

NVMe SSD

NVMe SSD

NVMe SSD

NVMe SSD

NVMe SSD

NVMe SSD

NVMe SSD

NVMe SSD
Non-Transparent Bridging (NTB)
Dynamic Partitioning
NVMe SR-IOV
Multi-Host I/O Sharing
PCle Fabric

- **Storage Functions**
  - Dynamic partitioning (drive-to-host mapping)
  - NVMe shared I/O (shared storage)
  - Ability to share other storage (SAS/SATA)

- **Host-to-Host Communications**
  - RDMA
  - Ethernet emulation

- **Manageability**
  - NVMe controller-to-host mapping
  - PCIe path selection
  - NVMe management

- **Fabric Resilience**
  - Supports link failover
  - Supports fabric manager failover
A high-performance fabric means
- High bandwidth
- Low latency

Increasing bandwidth is easy
- Aggregate parallel links
- Increase link speed (fatter pipe)

Reducing latency is hard
- Transfer latency is typically a small component of overall latency
- Other sources of latency:
  - Software (drivers)
  - Complex protocols
  - Protocol translation
  - Fabric switches/hops
Latency

- Media Access Time
  - Hard drive: Milliseconds
  - NAND flash: Microseconds
  - Next-generation NVM: Nanoseconds
The PCIe Advantage

Host
PCIe
HBA/HCA/NIC
Network Switch
PCIe
Network Interface
PCIe
NVMe SSD

Other Flash Storage Networks

Host
PCIe
PCIe Switch
< 200 ns
PCIe
NVMe SSD

PCIe Fabric
The PCIe Latency Advantage

Latency data from Z. Guz et al., “NVMe-over-Fabrics Performance Characterization and the Path to Low-Overhead Flash Disaggregation” in SYSTOR ’17
## PCIe Fabric Characteristics

<table>
<thead>
<tr>
<th>Property</th>
<th>Ideal Characteristic</th>
<th>PCIe Fabric</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>Free</td>
<td>Low</td>
<td>• PCIe built into virtually all hosts and NVMe drives</td>
</tr>
<tr>
<td>Complexity</td>
<td>None</td>
<td>Medium Complexity</td>
<td>• Builds on existing NVMe ecosystem with no changes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• PCIe fabrics are an emerging technology</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Requires PCIe SR-IOV drives for low-latency shared storage</td>
</tr>
<tr>
<td>Performance</td>
<td>High</td>
<td>High</td>
<td>• High bandwidth</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• The absolute lowest latency</td>
</tr>
<tr>
<td>Power consumption</td>
<td>None</td>
<td>Low</td>
<td>• No protocol translation</td>
</tr>
<tr>
<td>Standards based</td>
<td>Yes</td>
<td>Yes</td>
<td>• Works with standard hosts and standard NVMe SSDs</td>
</tr>
<tr>
<td>Scalability</td>
<td>Infinite</td>
<td>Limited</td>
<td>• PCIe hierarchy domain limited to 256 bus numbers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• PCIe has limited reach (cables)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• PCIe fabrics have limited scalability (less than 256 SSDs and 128 hosts)</td>
</tr>
</tbody>
</table>
Persistent Memory and Next Gen. NVM

**Traditional Memory**
- Volatile
- Byte addressable
- Memory load/store operations
- Memory bus

**Traditional Storage**
- Non-volatile (persistent)
- Block, file, or object addressable
- I/O operations
- Storage interconnect

**Next Generation NVM**
- Non-volatile (persistent)
- Byte, block, file, or object addressable
- Memory load/store operations and I/O operations

*Examples:* phase-change memory (PCM), resistive RAM (RRAM), spin-transfer-torque magnetic RAM (STT-MRAM), ferroelectric RAM (fRAM)
NVMe and Memory Operations

- **Controller Memory Buffer (CMB)**
  - PCI memory space exposed to host (byte addressable)
  - May be used to store commands and data
  - Contents do not persist across power cycles and resets

- **Persistent Memory Region (PMR)**
  - PCI memory space exposed to host (byte addressable)
  - May be used to store data
  - Content persist across power cycles and resets
NVMe together with a PCIe fabric allows direct network-to-storage and accelerator-to-storage communications.

Example:
1. Data transferred from network to NVMe CMB
2. NVMe block write operation imitated from CMB to NVM
   … sometime later …
3. NVMe block read operation initiated from NVM to CMB
4. GPU/accelerator transfers data from NVMe CMB for processing
Putting It All Together

- **NVMe Storage Functions**
  - Dynamic partitioning (drive-to-host mapping)
  - NVMe shared I/O (shared storage)
- Direct accelerator-to-NVMe and network-to-NVMe transfers
- Byte-addressable persistent memory
Summary

- PCIe fabrics build on the existing PCIe and NVMe ecosystem
  - Work with standard NVMe SSDs, OS drivers, and PCIe infrastructure
- PCIe fabrics support both byte-addressable memory and traditional storage operations
- PCIe fabrics are well-suited for applications that require low cost, the absolute lowest latency, and limited scalability
  - NVMe SSD sharing inside a rack and small clusters
- PCIe fabrics are not well-suited for long-reach applications or where a high degree of scalability is required
  - NVM Express over Fabrics (NVMe-oF™) is well-suited for these applications