How NVMe can support Remote Access via RDMA

NVMe and RDMA – A Perfect Match?
Yaniv Romem, CTO, Excelero
The NVMe interface at the hardware level yields an immediate 7X BW improvement and a greater than 3X reduction in latency vs. SATA (AHCI)

...and 10X better random performance!!
Why NVMe?

- NVMe (NVM express)
  - A streamlined protocol for accessing Non-Volatile memories via PCIe
  - Justification = Take advantage of:
    - Increased IOPs • 100,000s Typical
    - Increased Throughput • Over 3 GB/s
    - Reduced Latency • ~100/20μs Random R/W
    - Lower CPU Overhead • >1M IOPs per CPU
RDMA

- **Remote Direct Memory Access**
  - Enables a network adapter to access remote memory directly
  - Justification = Utilize new memories remotely:

- **Increased Remote Accesses**
  - Millions of Remote Memory Accesses/s

- **Increased Throughput**
  - 100Gb/s has arrived

- **Reduced Latency**
  - As little as 1μs

- **Lower CPU Overhead**
  - >>1M Memory Accesses per CPU
Similarities, Coincidence?

Both Work via Completion Queues

NVMe

- Increased IOPs
  - ~100,000s Typical
- Increased Throughput
  - Over 3 GB/s
- Reduced Latency
  - ~100/20μs Random R/W
- Lower CPU Overhead
  - >1M IOPs per CPU

RDMA

- Increased Remote Accesses
  - Millions of Remote Memory Accesses/s
- Increased Throughput
  - 100Gb/s has arrived
- Reduced Latency
  - As little as 1μs
- Lower CPU Overhead
  - >>1M Memory Accesses per CPU

Flash Memory Summit 2015
Santa Clara, CA
Fabric NVMe & RDMA

- NVMe-over-Fabrics
  - As a protocol, fabric agnostic, but:
    - Why would you want to add 200μs-1000μs latency to a 20us device???
- NVMe interface is layered above the networking
  - NVMe layer = control layer
  - RDMA = data layer
- Established RDMA Fabrics allow for choices in storage topology
Fabric NVMe Quirks/Challenges

- NVMe over Fabric dictates
  - 1-to-1 NVMe namespace over a link
  - Within the target, no implementation definitions

- Challenges
  - Volume definition & management
  - Device Aggregation, extension and redundancy
  - Centralized management
  - Storage service from a single target
    - CPU can become a bottleneck if it is doing control translation and / or data manipulation
  - Storage service from a cooperative group of targets
    - CPU can still be a bottleneck
    - Inter-communication between targets puts additional stress on network
Converged Infrastructure
- Each Server has 1 or more NVMe SSDs
- All SSDs are treated as one storage pool
- RDMA peer-to-peer makes all storage high performance
Fabric NVMe – RDMA Topology 2

- Centralized Model
  - Optional Redundancy
  - Serviceability
  - RDMA can make centralized, remote storage as fast as local

Centralization of storage services and the use of RDMA (Remote Direct Memory Access) can enhance performance and flexibility in a data center environment. RDMA enables low-latency, high-bandwidth data transfer directly between memory of two machines, bypassing the operating system and the interprocess communication mechanisms. This is particularly beneficial in scenarios requiring high-speed access to distributed storage, such as in cloud computing and large-scale data analytics. The diagram illustrates how RDMA NICs (Network Interface Controllers) facilitate these transactions, connecting clients with central servers in a topology optimized for efficiency and reliability.
Putting It All Together…

Local vs. Remote Flash, 4K Reads and Writes, Single Queue Depth

- **Local Reads**
  - 1 NVMe Disk: 80μs
  - 2 NVMe Disks: 81μs

- **Remote Reads**
  - 1 NVMe Disk: 85μs
  - 2 NVMe Disks: 87μs

- **Local Writes**
  - 1 Remote Target: 13μs
  - 2 Remote Targets: 14μs

- **Remote Writes**
  - 1 Remote Target: 19μs
  - 2 Remote Targets: 17μs

**Ratio of Latency, Remote vs. Local, 4K Reads in the 99.9th Percentile**

<table>
<thead>
<tr>
<th>IO Load</th>
<th>Remote Latency</th>
<th>Local Latency</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>105.6%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>102.7%</td>
<td>101.3%</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>98.6%</td>
<td>98.5%</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>101.3%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>98.6%</td>
<td>98.5%</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>100%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>105.6%</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

Remote ≈ Local Latency Across all IO loads!

**Number of Outstanding IOs (NumThreads * QueueDepth)**

- Ratio Remote Latency vs. Local Latency
Thank You

Yaniv Romem, CTO, Excelero
info@excelero.com
http://www.excelero.com

“Free Your Flash!”