NVMe/TCP is the Best Way to Disaggregate Flash

Muli Ben-Yehuda
Co-Founder & CTO
A little bit about Lightbits Labs and me

- Lightbits is a well-funded stealth mode startup with offices in Israel and San Jose, CA
- Doing cool things with NVMe and NVMe-oF
- Inventors of NVMe/TCP

- Me: Lightbits Labs co-founder & CTO
- Operating systems, hypervisor, clouds, high-speed networking and storage
Disaggre-what?
From direct-attached to a disaggregated cloud

- Maximize utilization
- Reduce TCO
- Easy to maintain & scale
- Better user experience
- Support more users
NVMe over Fabrics
NVMe over Fabrics

- PCIe is a network (transport)
- Can we do NVMe over other transports?
- Retain NVMe efficiency and performance over network fabrics
- Eliminate unnecessary protocol translations
- Enable low-latency and high IOPS remote NVMe storage solutions
Spot the missing protocol...

- NVMe Host Software
- Host Side Transport Abstraction
- Controller Side Transport Abstraction
- NVMe SSDs
- Fibre Channel
- InfiniBand
- RoCE
- iWARP
- Next Gen Fabrics
Enter TCP/IP
<table>
<thead>
<tr>
<th>TCP/IP model</th>
<th>Protocols and services</th>
<th>OSI model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application</td>
<td>HTTP, FTP, Telnet, NTP, DHCP, PING</td>
<td>Application</td>
</tr>
<tr>
<td>Transport</td>
<td>TCP, UDP</td>
<td>Session</td>
</tr>
<tr>
<td>Network</td>
<td>IP, ARP, ICMP, IGMP</td>
<td>Transport</td>
</tr>
<tr>
<td>Network Interface</td>
<td>Ethernet</td>
<td>Network</td>
</tr>
</tbody>
</table>

Santa Clara, CA
August 2017
ubiquitous (adjective)
1. Being everywhere at once: omnipresent.
YES to INNOVATION!

NO to CHANGE!

networking
Efficiency

Innovation
Vision
Investment
Development
Analysis
Process
Strategy
Management
Teamwork
Marketing
Research
Partner
ESSENCE

meaning, definition, explanation...
NVMe/TCP in a nutshell

- A TCP/IP transport binding for NVMe over Fabrics
- NVMe-OF Commands sent over standard TCP/IP sockets
- Each NVMe queue pair mapped to a TCP connection
- TCP provides a reliable transport layer for NVMe queueing model
Is this the real life?
Is this just fantasy?
NVMe over TCP Standardization

Standardizing TCP/IP transport binding, adding to NVMe-oF spec alongside RDMA & FC

- Expect standard ratification in 2H 2018
- Supports remote NVMe SSDs with minimal additional latency compared to local SSDs
- Same NVMe model: sub-systems, controllers namespaces, admin queues, data queues

**Lightbits is leading new TCP/IP transport**
- Developed pre-standard client available to NVMe members
- Tested preliminary implementation with several customers & partners
- Key contributor to standard and Linux upstreaming process
Comparing DAS vs. NVMe/TCP

IOPS, average and 99.99% latencies
Comparing DAS vs. NVMe/TCP

IOPS, average and 99.99% latencies

Come see the demo live at Intel booth #745!
NVMe/TCP unleashes disaggregation at DAS performance

- Hyperscale data center deployment
- Multiple Application Server Live Demonstration
- Low latency and High performance as DAS
- Standard TCP/IP Network Infrastructure
- No modification in Application server Software

- High IOPS for serving many Application Servers
- Thin Provisioned Storage
- In-line Hardware Accelerated Data Reduction at line rate speed
- Optane™ ready - Supports variety of SSD technologies
High Performance, Thinly Provisioned, Data Reduction

DAS PERFORMANCE
- Low Latency on NVMe/TCP
- Scalable IOPS for serving many application servers
- In-line acceleration

THIN PROVISIONING
- Provision per use only
- Buy only what you need
- Consume flash at the time of writes

DATA REDUCTION
- Maximize Flash Utilization
- No performance compromise

Lightbox Storage utilization (5.12TB used)

x2.90 cost saving
The Demo Dashboard

Lightbits Labs NVMe/TCP Disaggregation at DAS Performance

**Average Read & Write Latency**
- Read Average Latency: 296 µs
  - LB vs DAS (Similar Performance)
  - Lightbits Advantage: 0.99x
- Write Average Latency: 57 µs
  - LB vs DAS (Lightbits Advantage)
  - Lightbits Advantage: 4.19x

**99.99% Read & Write "Tail" Latency**
- Read 99.9% Latency: 2.384 ms
  - LB vs DAS (Lightbits Advantage)
  - Lightbits Advantage: 3.18x
- Write 99.9% Latency: 171 µs
  - LB vs DAS (Lightbits Advantage)
  - Lightbits Advantage: 10.53x

**Data Reduction:** 293%

<table>
<thead>
<tr>
<th>Data Services</th>
<th>Total Logical Capacity</th>
<th>Logical Capacity: Client 1</th>
<th>Logical Capacity: Client 2</th>
<th>Logical Capacity: Client 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15 TiB</td>
<td>5 TiB</td>
<td>5 TiB</td>
<td>5 TiB</td>
</tr>
<tr>
<td></td>
<td>Total Physical Used</td>
<td>5.12 TiB</td>
<td>Physical Used: Client 1</td>
<td>1.2 TiB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Physical Used: Client 2</td>
<td>0.721 TiB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Physical Used: Client 3</td>
<td>3.2 TiB</td>
</tr>
</tbody>
</table>
NVMe/TCP Standard

Lightbits NVMe/TCP disaggregation party

Lightbits NVMe/TCP technology enables hyperscalers to move from inefficient direct-attached storage to a shared flash model where compute and storage are scaled independently.

NVMe/TCP is an open standard that enables flash disaggregation without compromising performance and without requiring any changes to compute clients or networking infrastructure.
Thank you!