



Networking NVMe-based Flash with TCP/IP Using the Protocol Everyone Knows

Muli Ben-Yehuda(*)
Lightbits Labs

(*) team effort with contributions from Lightbits, Facebook, Intel, Solareflare, NVMe TWG...

Santa Clara, CA
August 2017



+



=

?



+

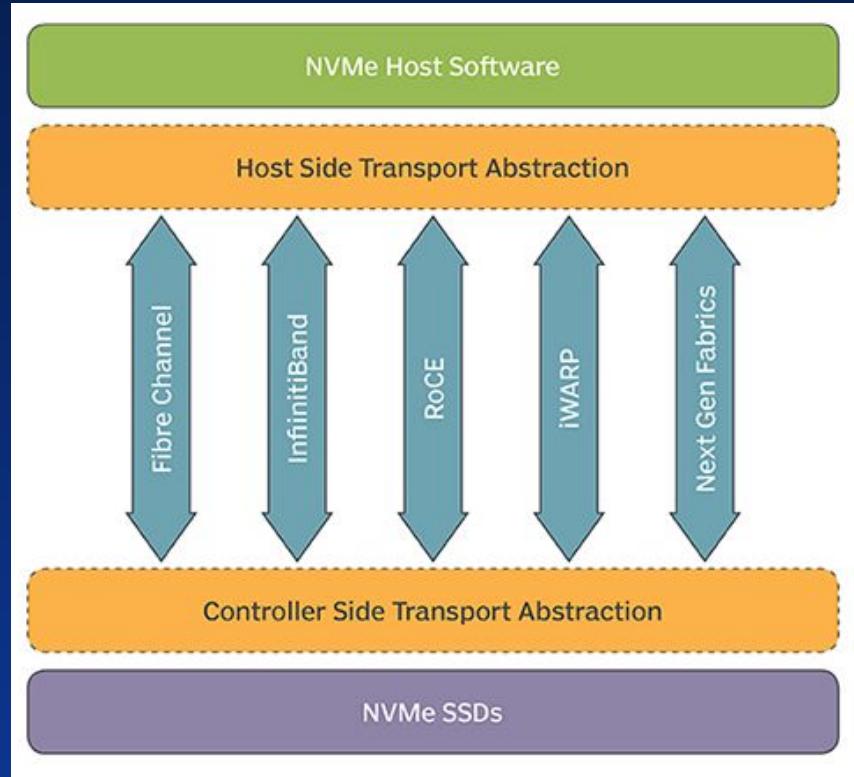


=

NVMe over TCP



What is NVMe over Fabrics?

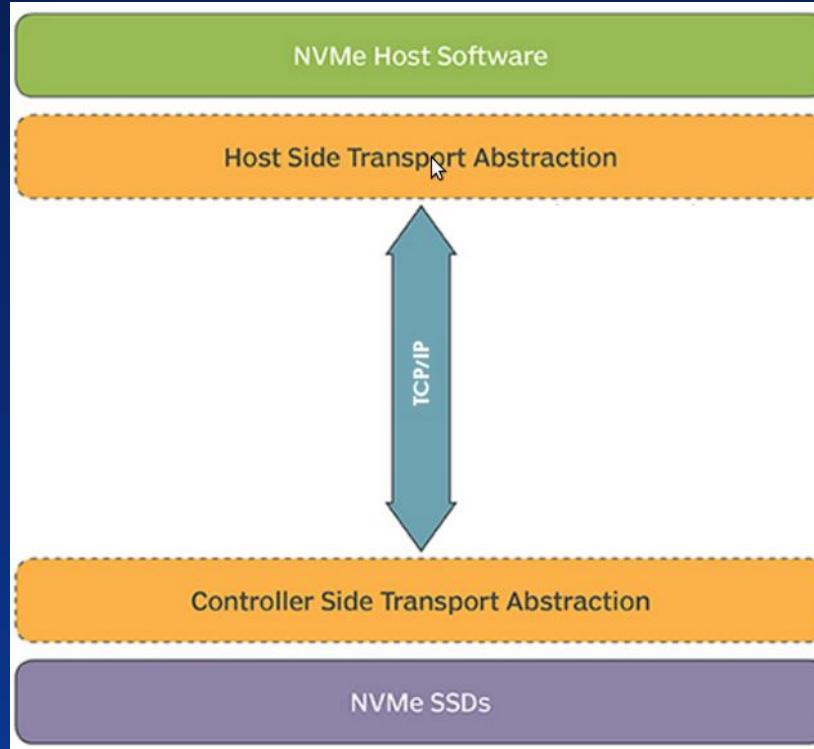




Santa Clara, CA
August 2017



NVMe over TCP/IP in a nutshell



Why?

S₁

I₁

M₃

P₃

L₁

E₁

Ó

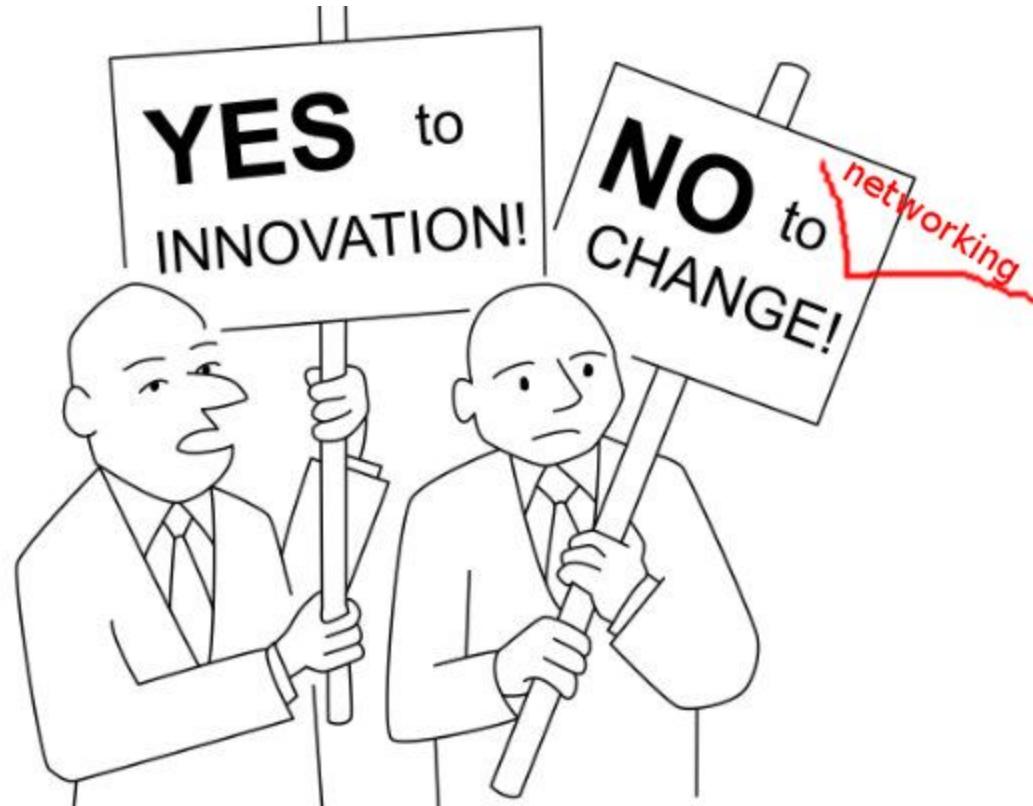
R,

ubiquitous (adjective)

I. Being everywhere at once:
omnipresent.









Innovation



Vision



Investment



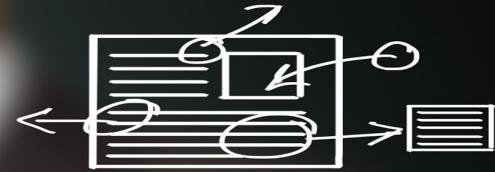
Development



Analysis



Process



Research

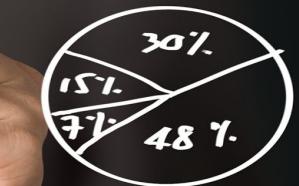
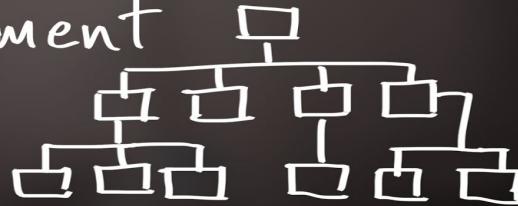
Efficiency

Strategy



Teamwork

Management



Marketing



Partner



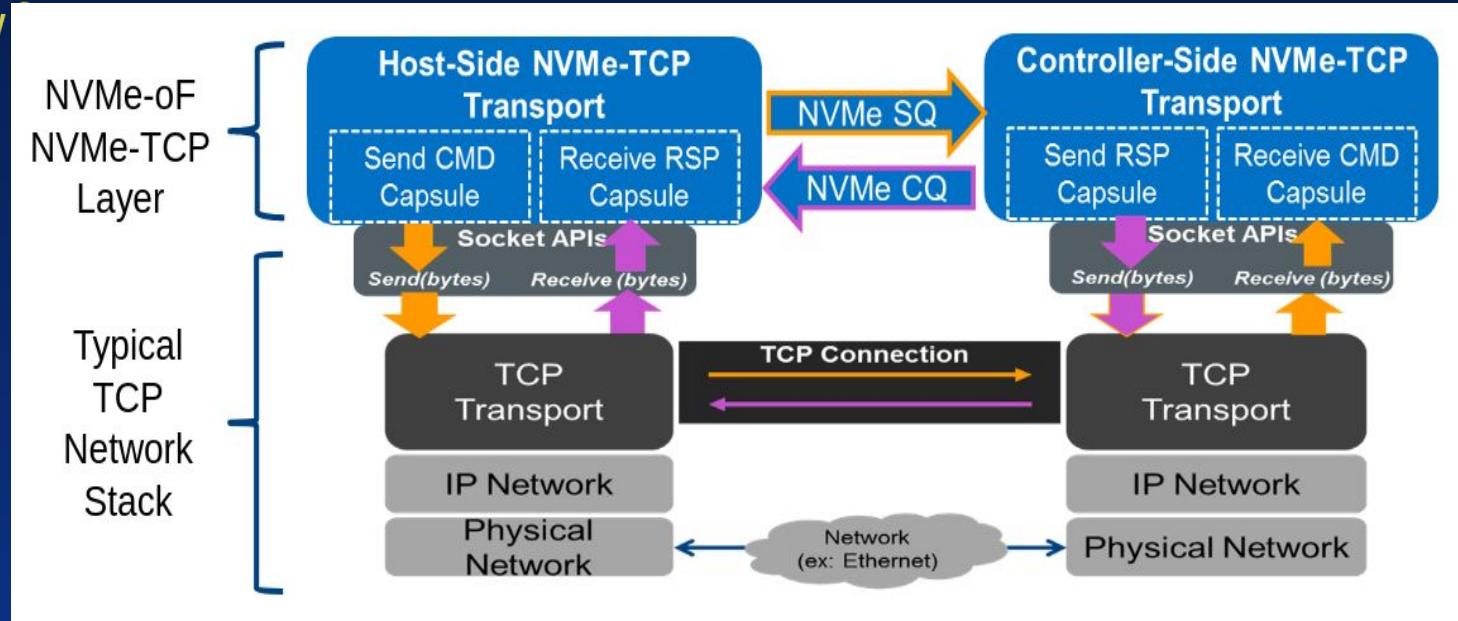
ESSENCE

meaning, definition, explanation...



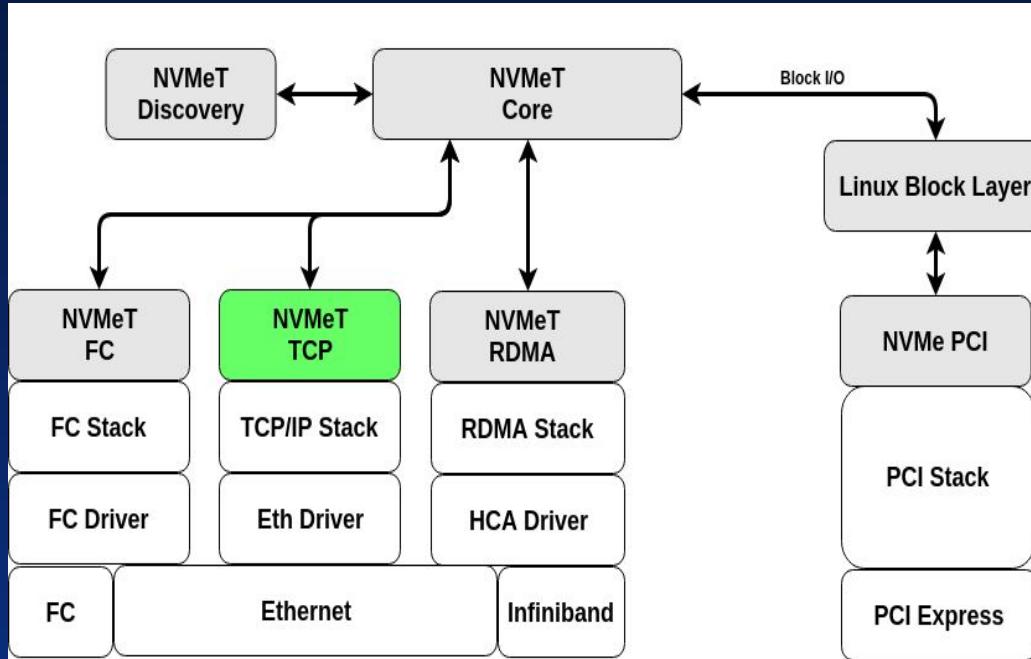
Flash Memory

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

**STATUS
SIADS**



- NVMe Technical Working Group is working on standardizing TCP/IP transport bindings for NVMe
- TCP/IP transport bindings will be added to the spec alongside RDMA & FC
- Key contributors are Lightbits, Intel & Facebook, with lots of contributions from Mellanox, Sun, others
- NVMe/TCP reference Linux host & target implementations based on Lightbits pre-standard code are available to NVMe/TCP TWG contributors and will be upstreamed to coincide with the spec
 - Contributions welcome!

PERFORMANCE EVALUATION



PERFORMANCE EVALUATION

Pre-Standard

Pre-Alpha

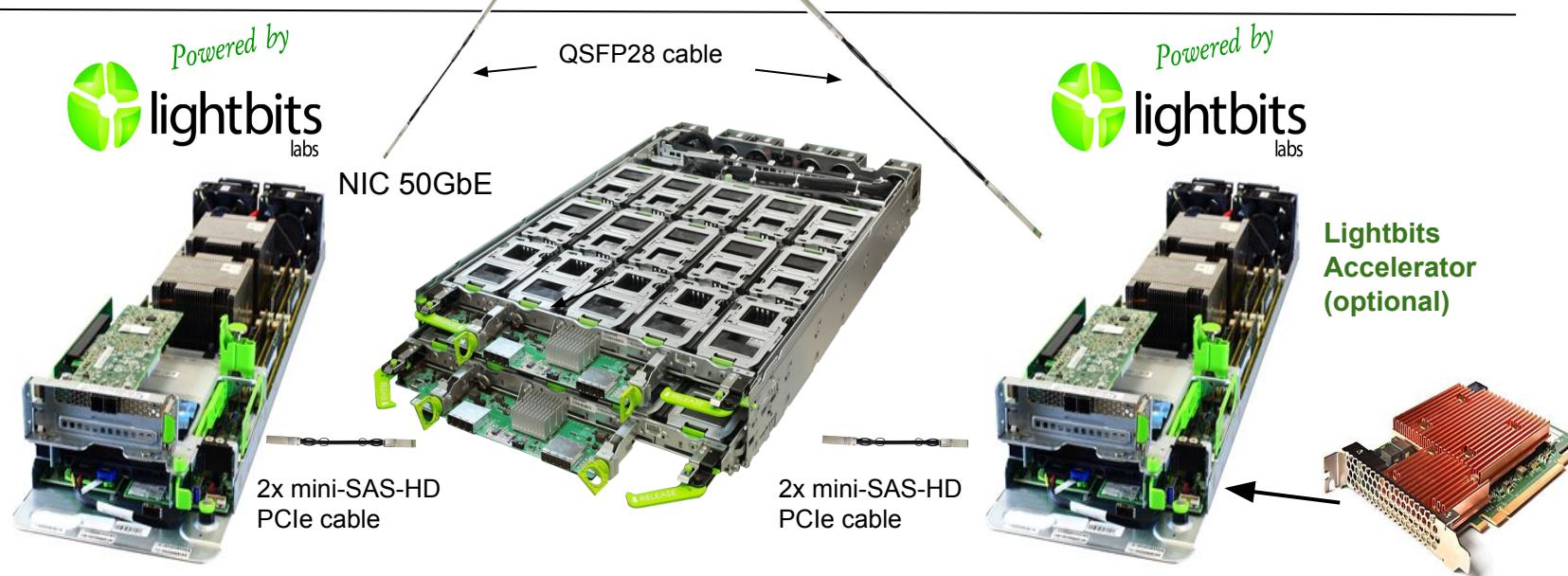
Compute



Network



Storage



Compute



Network

Come see
NVMe/TCP in action!

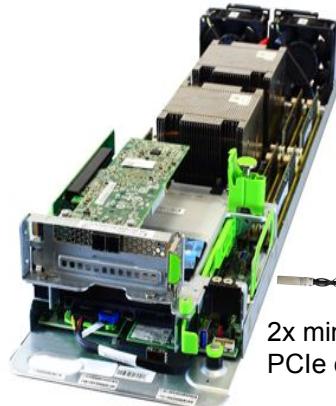
Storage



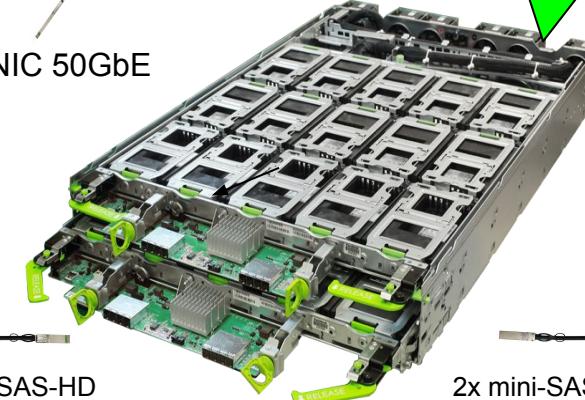
QSFP28 cable

NIC 50GbE

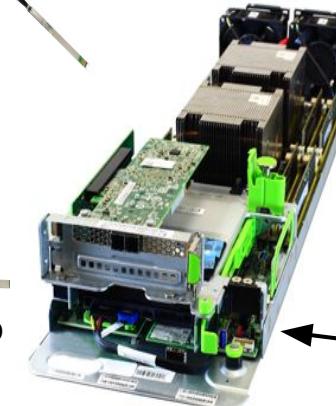
Powered by



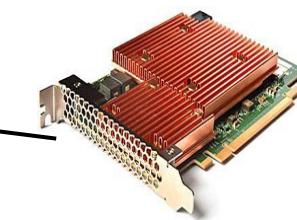
2x mini-SAS-HD
PCIe cable



2x mini-SAS-HD
PCIe cable



Lightbits
Accelerator
(optional)





Flash Memory Summit

IOPs

Random 4K Read: 70%

Random 4K Write: 30%

3.2M IOPs(*)

QD = 32

(*) Alpha target: 5M IOPs



Average & Tail Latencies

Random Read (μs)			Random Write (μs)		
Average	99%	99.9%	Average	99%	99.9%
120	167	212	47	71	95

QD = 1





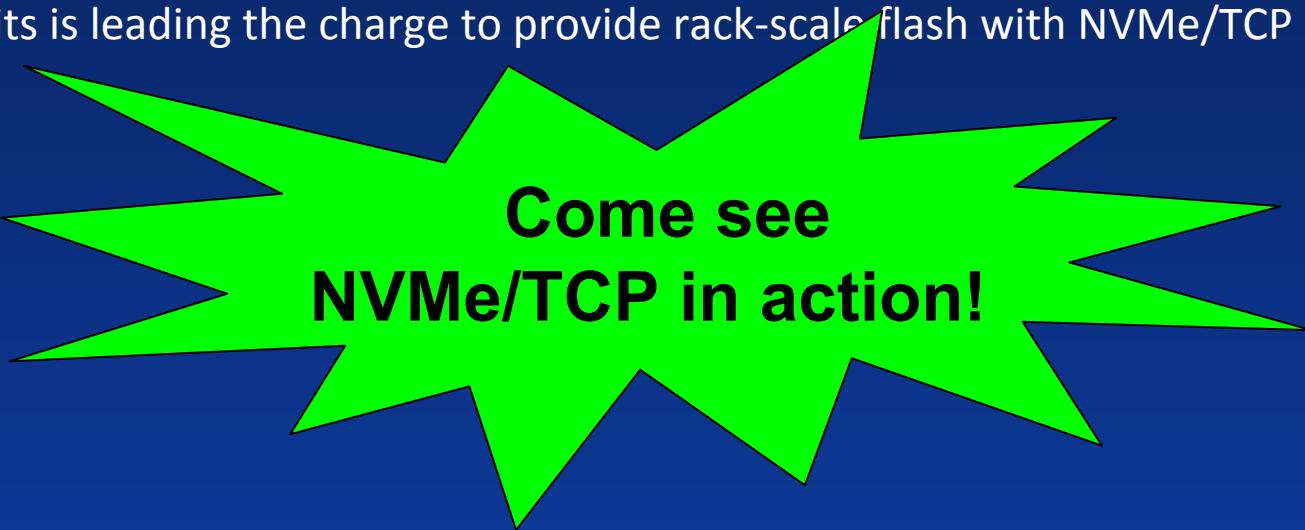
Potential Issues with TCP/IP

- Absolute latency is higher than RDMA?
- There could be head-of-line blocking leading to increased latency?
- Delayed acks could increase latency?
- Incast could be an issue?
- Network congestion could be an issue?
- Lack of hardware acceleration?



Summary and Conclusions

- NVMe over TCP/IP is here to stay
 - Simple, ubiquitous, and fast!
- Complements -- not replaces -- NVMe over RDMA/FC
- Spec and Linux implementation coming soon
- Lightbits is leading the charge to provide rack-scale flash with NVMe/TCP



**Come see
NVMe/TCP in action!**

LIGHT UP YOUR CLOUD



lightbits
labs

THANK YOU

For more NVMe/TCP goodness:
<http://www.lightbitslabs.com>