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...

nvm
EXPRESS®

+

TCP/IP

=

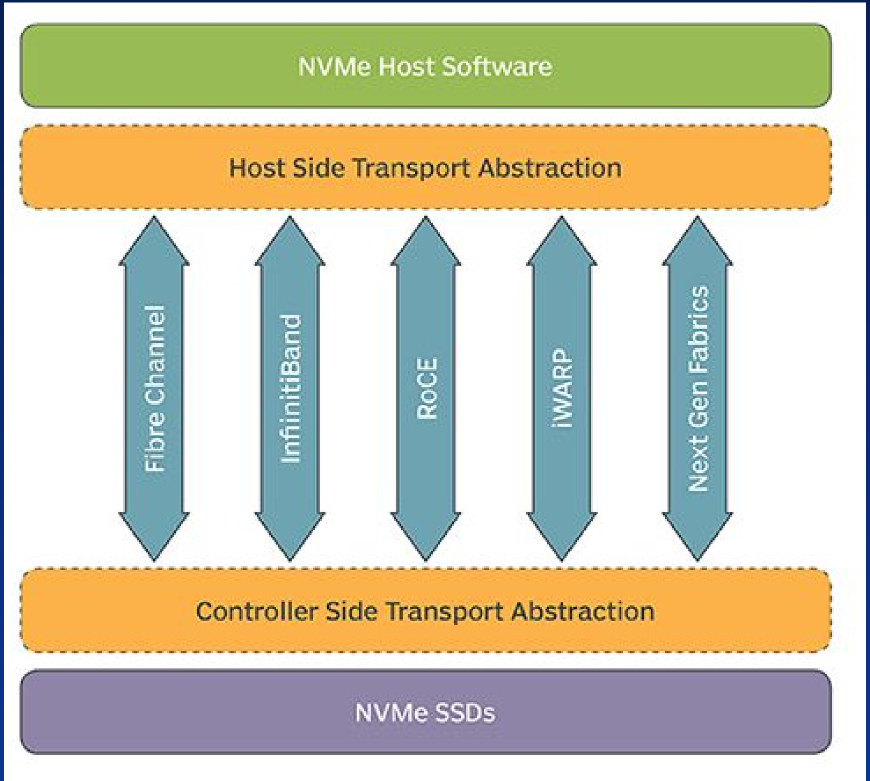
?

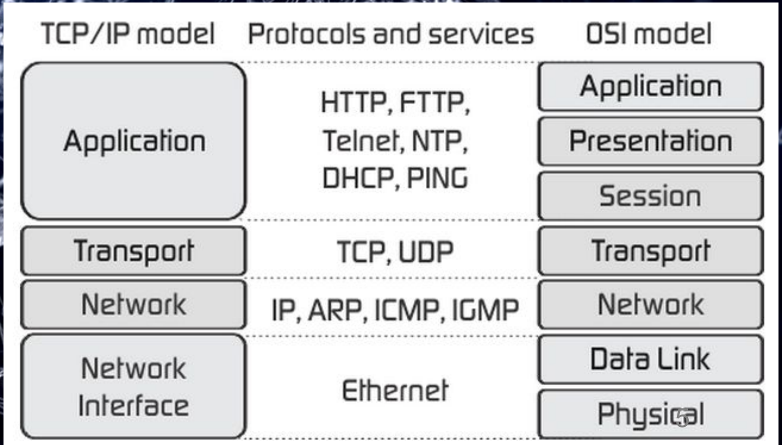
nvm EXPRESS® + TCP/IP

=

NVMe over TCP

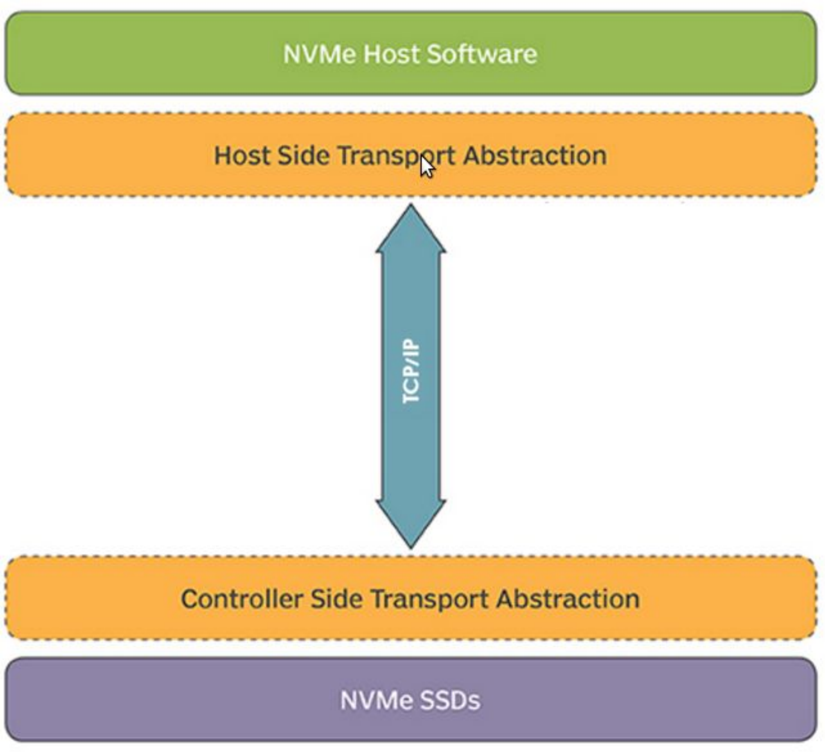
What is NVMe over Fabrics?





TCP/IP model	Protocols and services	OSI model
Application	HTTP, FTTP, Telnet, NTP, DHCP, PING	Application
		Presentation
		Session
Transport	TCP, UDP	Transport
Network	IP, ARP, ICMP, IGMP	Network
Network Interface	Ethernet	Data Link
		Physical

NVMe over TCP/IP in a nutshell



NVMe Host Software

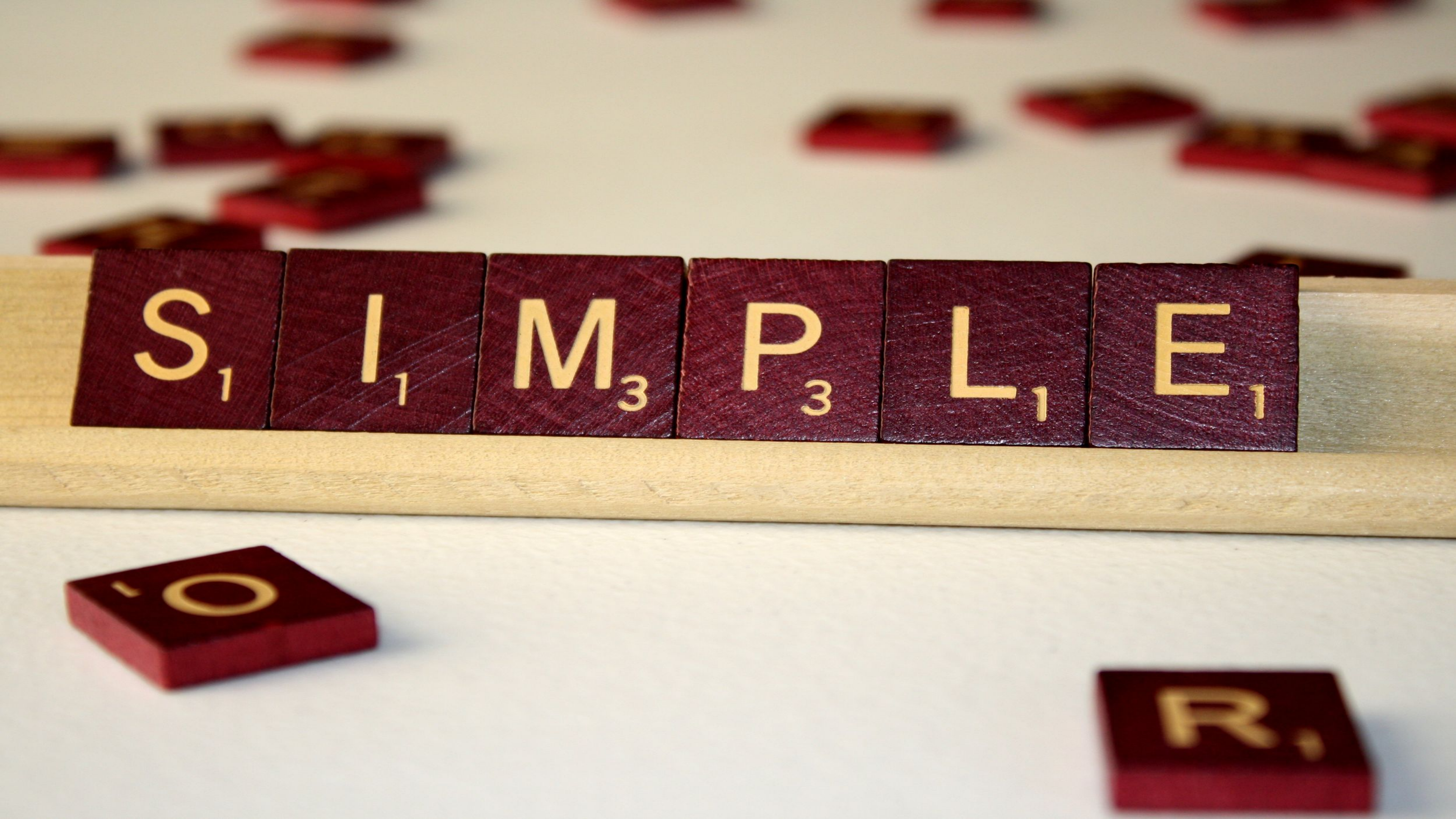
Host Side Transport Abstraction

TCP/IP

Controller Side Transport Abstraction

NVMe SSDs

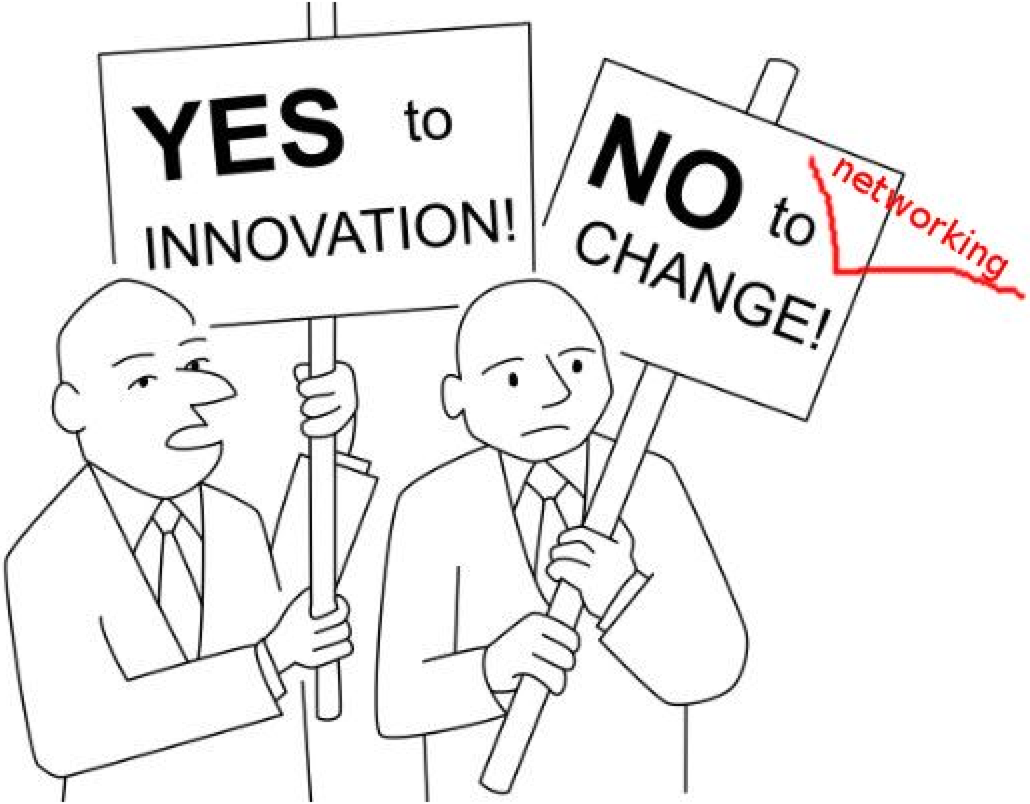
Why?



S1 I1 M3 P3 L1 E1
O
R1

ubiquitous (adjective)

1. Being everywhere at once: omnipresent.


YES to
INNOVATION!
NO to
networking
CHANGE!

GT R

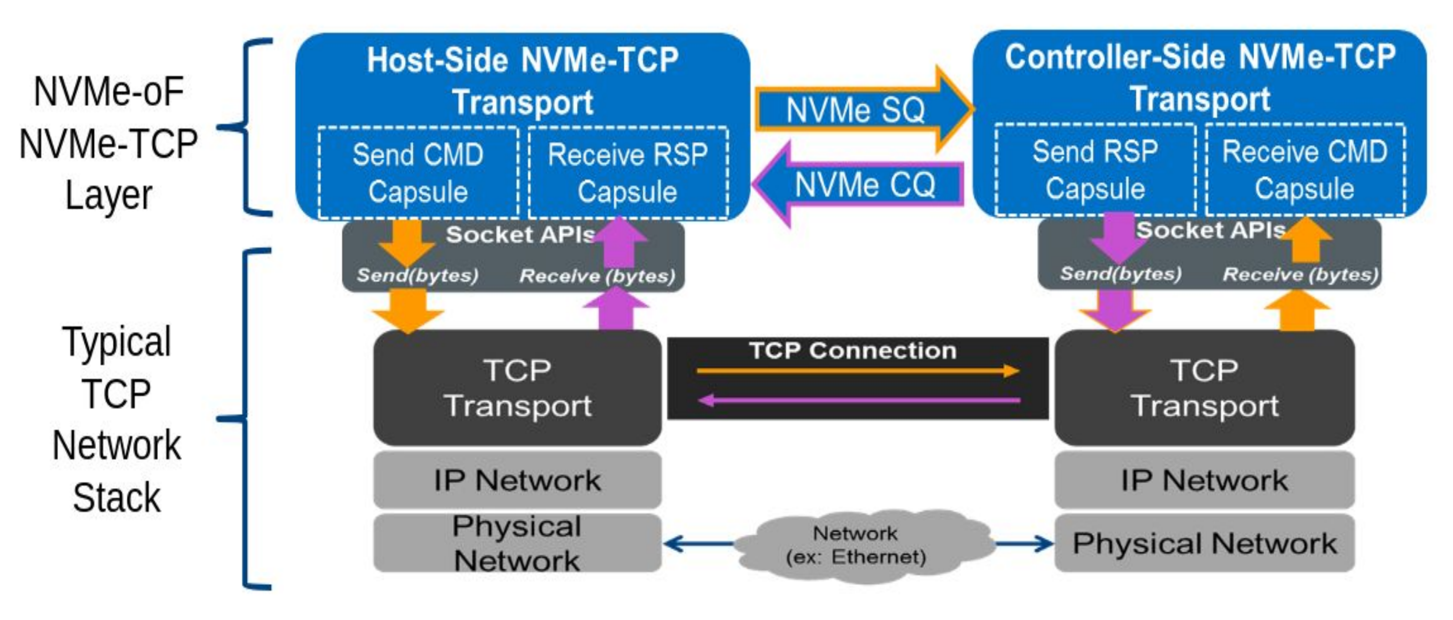
Innovation
Process
Research
Efficiency
Vision
Investment
Strategy
30%
15%
7%
48 %
?
Marketing
Development
Teamwork
Analysis
Management
Partner

ESSENCE

meaning, definition, explanation...

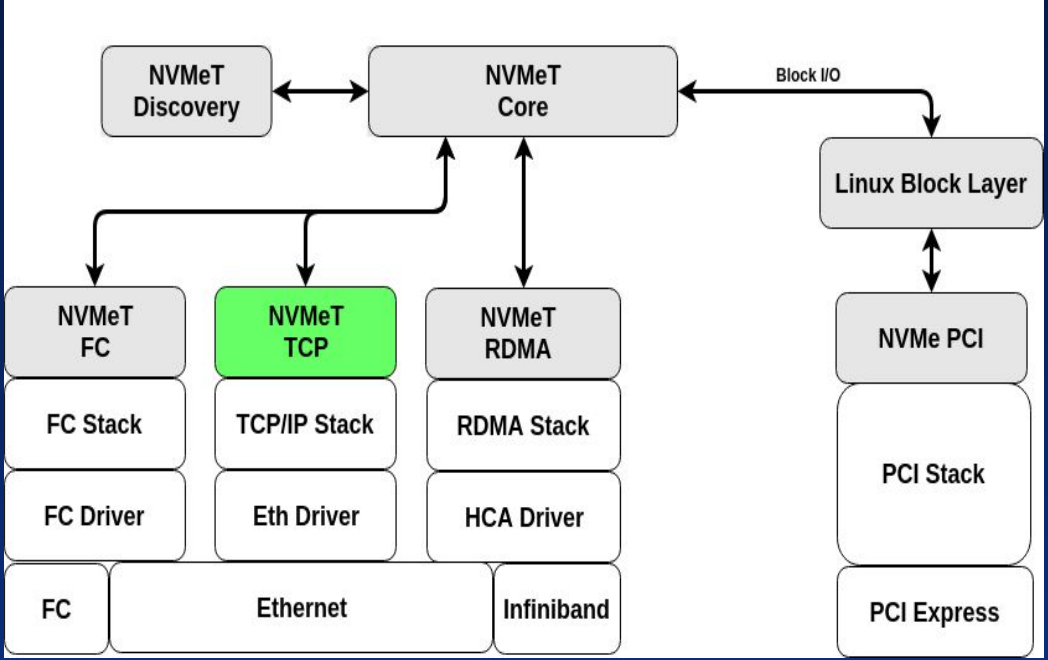
Flash Memory

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



- 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

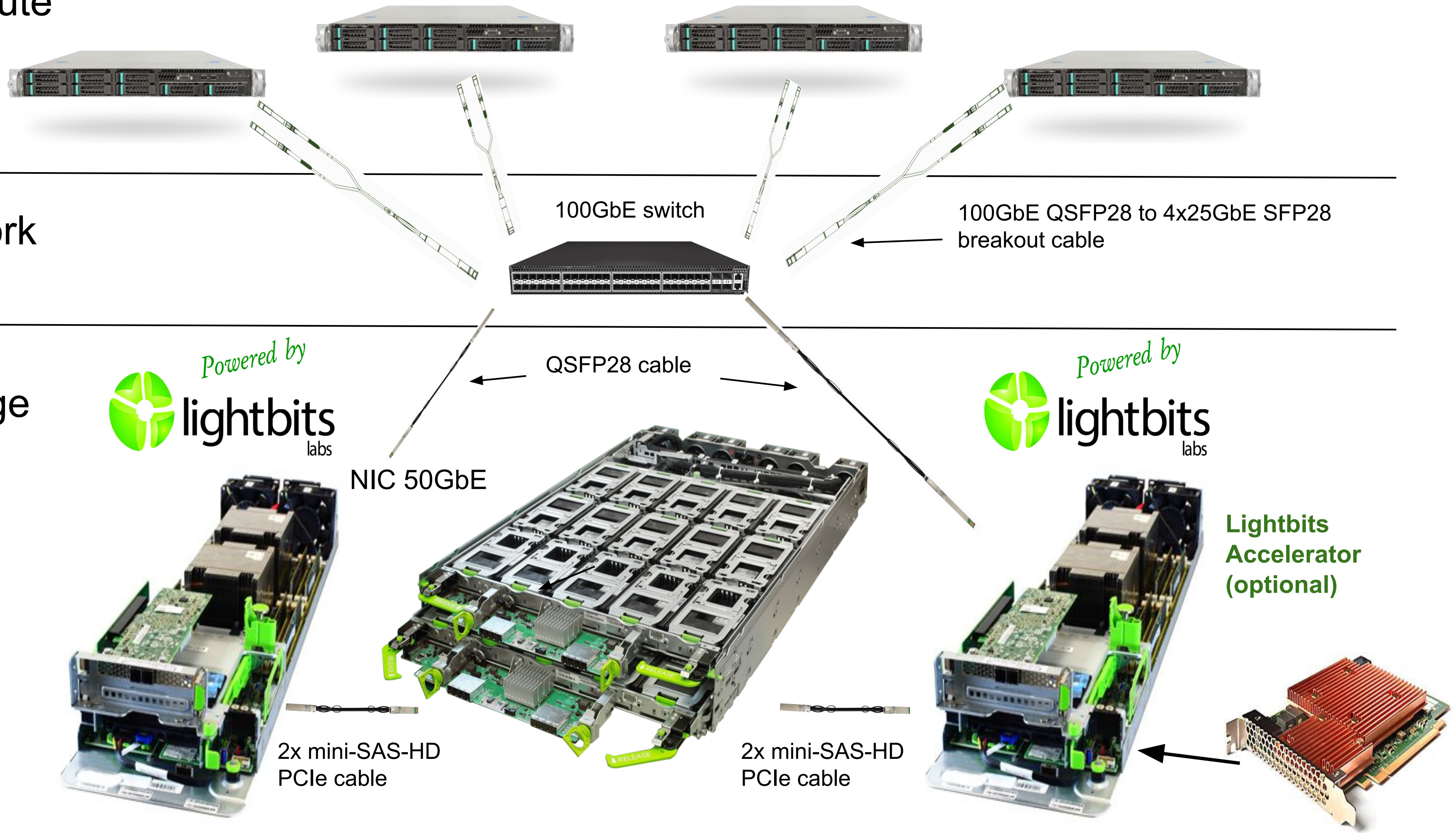
Pre-Standard

Pre-Alpha

Compute

Network

Storage



100GbE switch

100GbE QSFP28 to 4x25GbE SFP28 breakout cable

QSFP28 cable

NIC 50GbE

Powered by lightbits labs

Powered by lightbits labs

Lightbits Accelerator (optional)

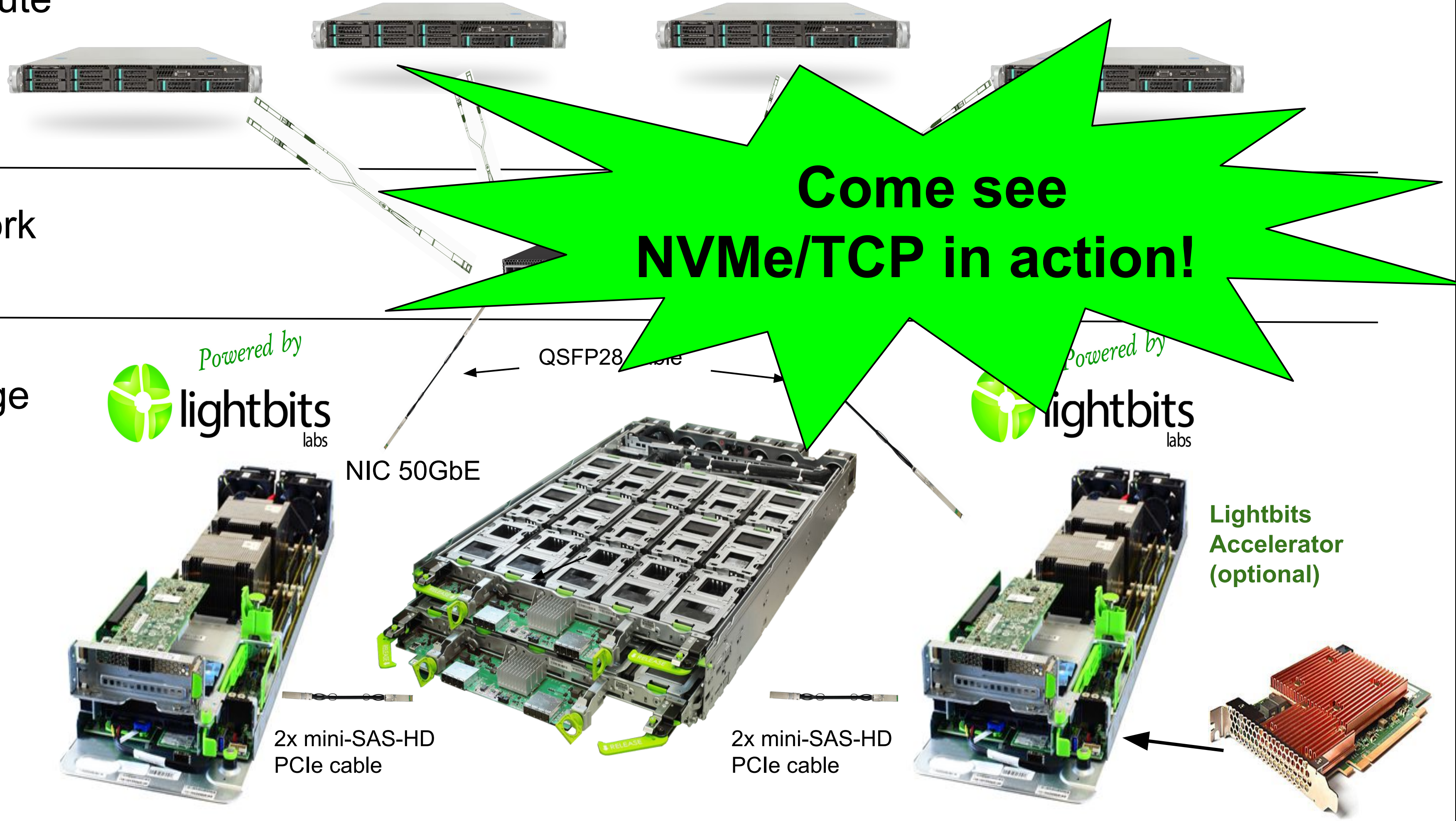
2x mini-SAS-HD PCIe cable

2x mini-SAS-HD PCIe cable

Compute

Network

Storage



Come see NVMe/TCP in action!

Powered by lightbits labs

QSFP28 ble

NIC 50GbE

2x mini-SAS-HD PCIe cable

2x mini-SAS-HD PCIe cable

Lightbits Accelerator (optional)

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