Understanding NVMe over Fabrics on TCP

NVMF-302A-1

Organizer: Rob Davis, Mellanox
Chair: David Woolf, UNH-IOL

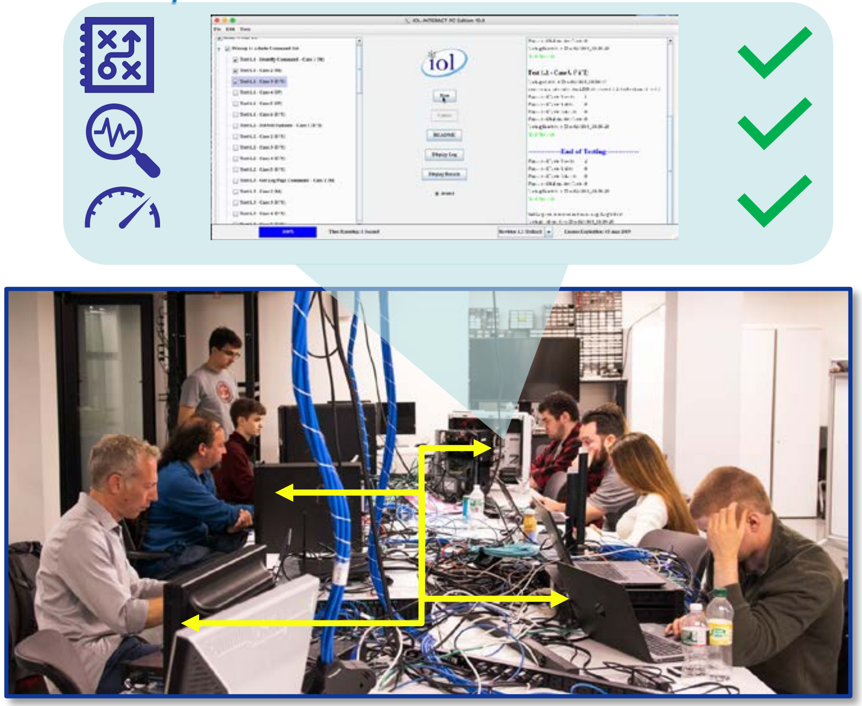
Presenters:
Alex Shpiner, Lightbits Labs
John Kim, Mellanox
Tom Spencer, Xilinx
Ron Renwick, Netronome

Session Agenda

- **2:15 - An NVMe/TCP Software-Defined Platform for Guaranteed QoS**
  - Alex Shpiner, System Architect, Lightbits Labs
- **2:30 - Comparing NVMe-oF on RoCE vs. TCP**
  - John Kim, Director Storage Marketing, Mellanox
- **2:45 Accelerating NVMe over TCP for Disaggregated Storage Applications**
  - Tom Spencer, Senior Director Product Marketing, Xilinx
- **3:00 Using SmartNICS and Buffer Management to Improve NVMe over TCP Performance**
  - Ron Renwick, VP of Products, Netronome
- **3:15 – Q&A**

Flash Memory Summit

A word on Interop…



- NVMe/TCP community has been prioritizing interop.
- Series of plugfest events at UNH-IOL since 2018
- Integrators List for NVMe/TCP
- UNH-IOL Compliance Tools available.

nvm EXPRESS®

NVMe-oF™ Integrator's List v11.0 | NVMe-MI Integrator's List | NVMe Integrator's List

- NVMe Integrator's List Policy v11.0
- NVMe Integrator's List Policy v11.0 Redline

NVMe-oF TCP Devices

Product	Product Type	Software Version/ Kernel Version/ Port Types	Interop Program Revision	Date Listed	Further Info
HPE M-Series SN2100M	Switch	Release 3.8.1112	v11.0	07/02/2019	www.hpe.com
Lightbits SuperSSD	Target	Software: LightOS 1.0.3	v11.0	07/02/2019	www.lightbitslabs.com
SANBlaze VLUN	Target and Initiator	Hardware Version Supermicro X10DRW-i Software Version V8.0-64-dev built on Jun 10 2019 at 14:56:15 Kernel: 4.9.107	v11.0	07/02/2019	www.sanblaze.com
Solarflare ONVMe	Target	Hardware Version XtremeScale X2522 R5 Firmware Version 7.5.0.1008 Software Version 19.2	v11.0	07/02/2019	www.solarflare.com
Toshiba Memory Kumoscale	Target	Software Version 3.11	v11.0	07/02/2019	business.toshiba-memory.com

iol University of New Hampshire InterOperability Laboratory

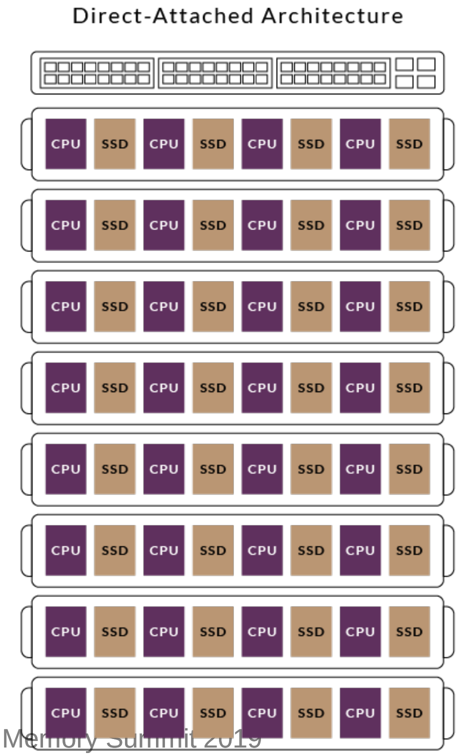
Session Agenda

- **2:15 - An NVMe/TCP Software-Defined Platform for Guaranteed QoS**
  - Alex Shpiner, System Architect, Lightbits Labs
- **2:30 - Comparing NVMe-oF on RoCE vs. TCP**
  - John Kim, Director Storage Marketing, Mellanox
- **2:45 Accelerating NVMe over TCP for Disaggregated Storage Applications**
  - Tom Spencer, Senior Director Product Marketing, Xilinx
- **3:00 Using SmartNICS and Buffer Management to Improve NVMe over TCP Performance**
  - Ron Renwick, VP of Products, Netronome
- **3:15 – Q&A**

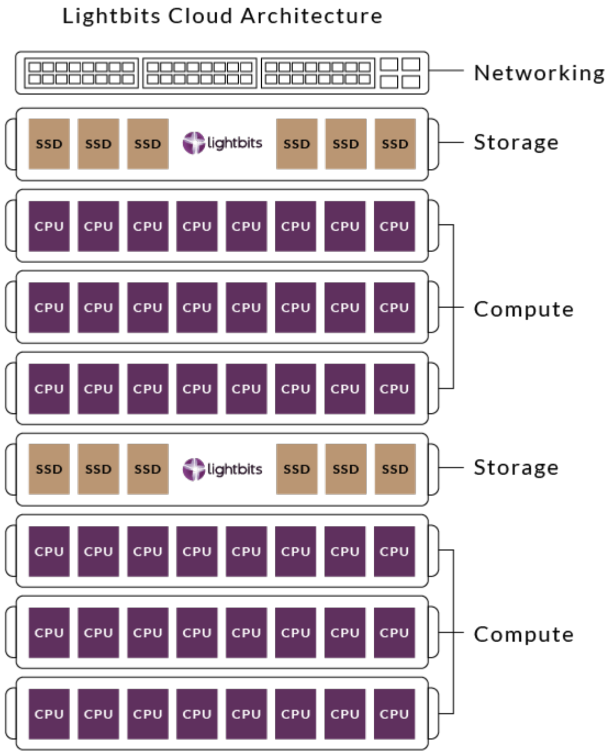
An NVMe/TCP Software-Defined Platform for Guaranteed QoS

Alex Shpiner
System Architect, Lightbits Labs

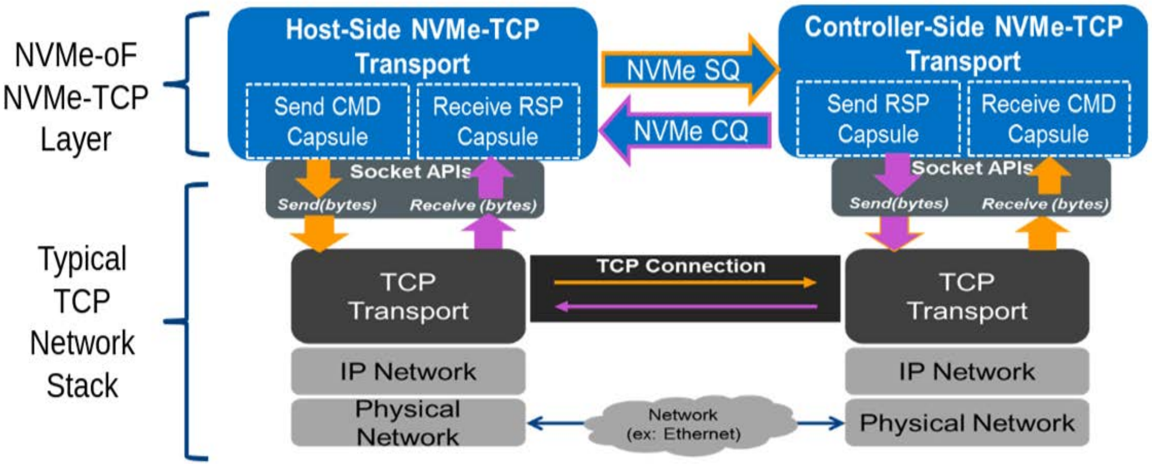
NVMeoF: from direct-attached storage to a disaggregated cloud



- Efficient scalability
- Maximal utilization - support more users
- Easy maintenance and operation



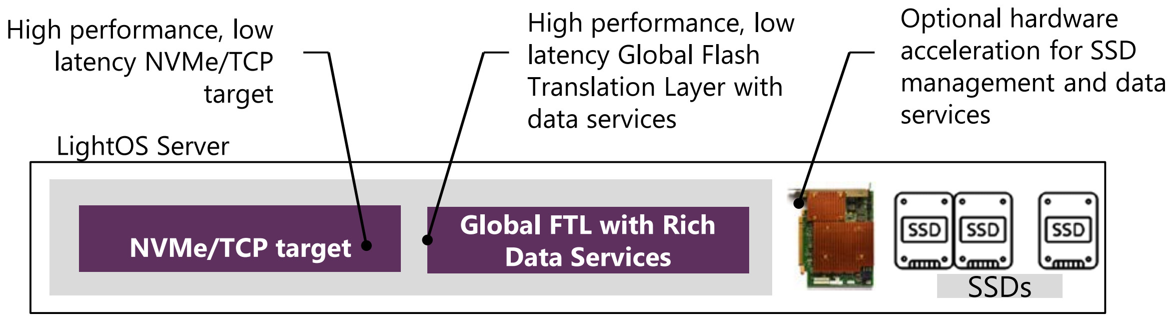
NVMe/TCP in a nutshell



- TCP is the transport layer below NVMe layer.
- NVMe commands are sent over standard TCP/IP sockets
- Each NVMe queue pair mapped to a TCP connection
- TCP provides a reliable transport layer and congestion control.

LightOS™ - NVMe/TCP target with data services

- First commercially-available NVMe/TCP open storage platform
- Software-Defined Storage
- Runs on standard servers, with commodity SSDs.
- Based on standard networks without proprietary client software
- High throughput, consistent low latency, data services, QoS
- 100Gbps streaming compression/decompression and erasure coding
- Thin provisioning
- Storage server clustering (multi-server data protection)



Flash Memory Summit

Multi tenant storage challenges

Problem: Unpredictable performance or behaviour of the application (service)

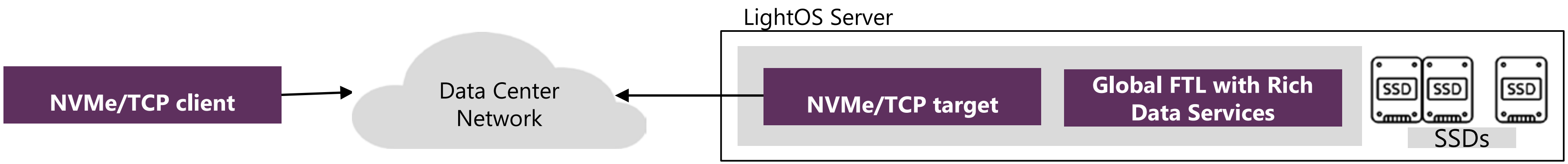
- Noisy neighbours
- Impact of writes on performance of reads.
- Write imbalance across SSDs.
- No performance (throughput, latency, etc.) guarantees per tenant.

Naive solution: overprovision resources so there are always spare IOPs.

But this is expensive...

Better solution: QoS (Quality of Service)

LightOS end-to-end QoS value proposition



Reads and writes separation

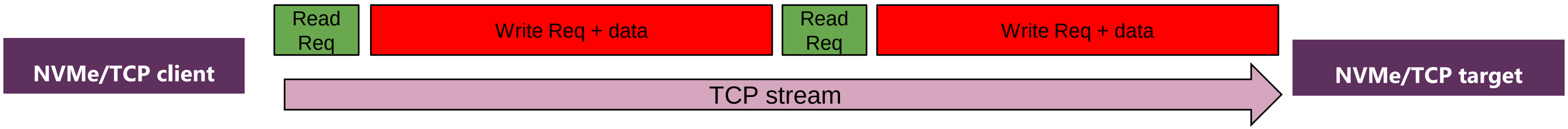
Guaranteed Service: Per-tenant IO capping

Balance write IOs and endurance across SSDs

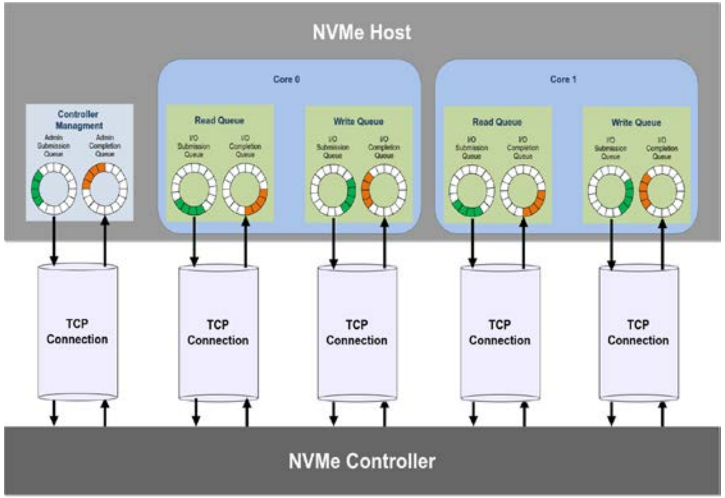
Read/Write Separation

<u>Problem:</u> Head-of-queue (HOQ) blocking: read latency is affected by presence of writes.

- Read requests (few bytes) can be placed behind large write request (eg. 1MB)
- Read requests will not be processed before write request is consumed by application from the network

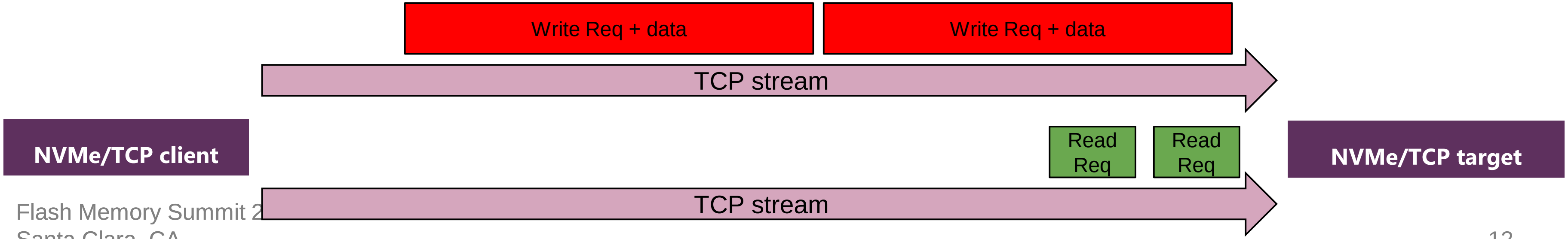


Read/Write Separation



<u>Solution:</u> Read/Write Separation

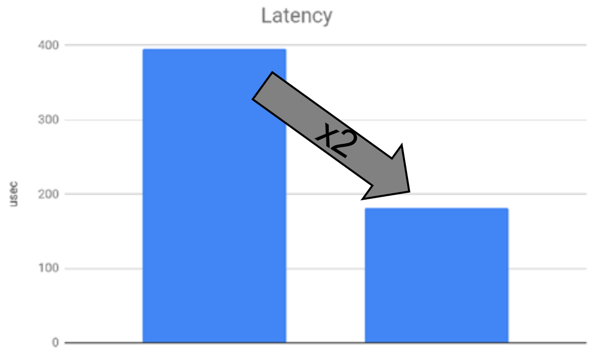
- Client side dedicated read queues and dedicated write queues (TCP connections)
- Target side dedicated NIC queues for read connections and write connections



Read/Write Separation: Lab Results

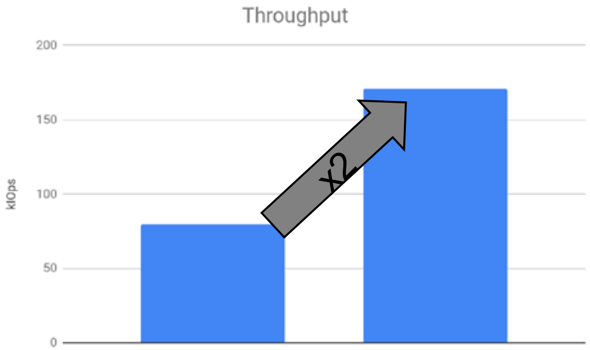
Test the impact of Large Write I/O on Read Latency

- 32 Readers issuing synchronous 4KB Read I/O
- 1 Writer that issues 256KB Writes, QD=16



without R/W queue separation

with R/W queue separation



without R/W queue separation

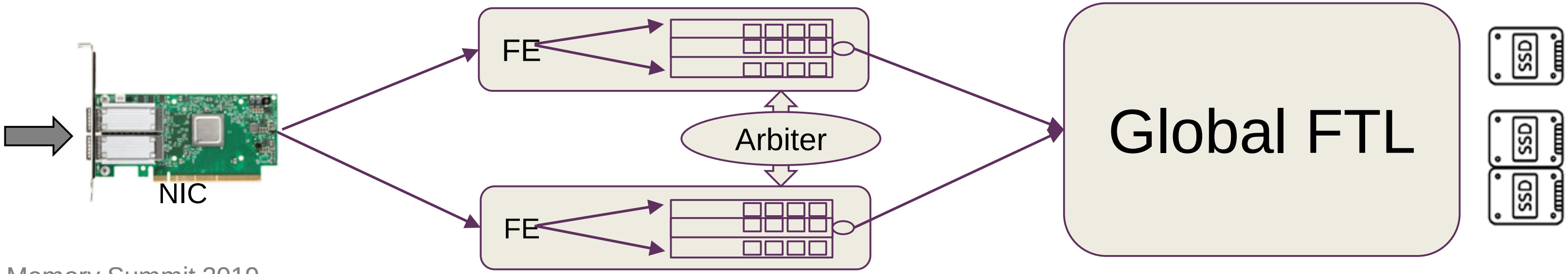
with R/W queue separation

IO Capping

Problem: noisy neighbours

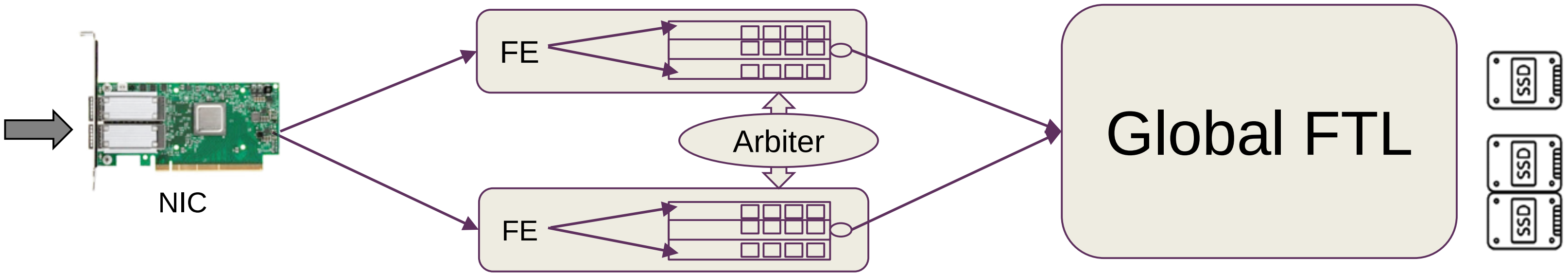
Solution: IO capping per tenant

- Multi-queues system
- Arbiter - coordination between queues
  - and between parallel front-end cores



IO Capping: Multi-queues System

- Arriving requests are separated to queues by: (tenant, {write | read}).
- I/O capping per queue:
  - Queues are served (requests submitted to GFTL) according to quota allocated by the arbiter.
  - Spare quota is spread equally among the queues (incl. best-effort queues).



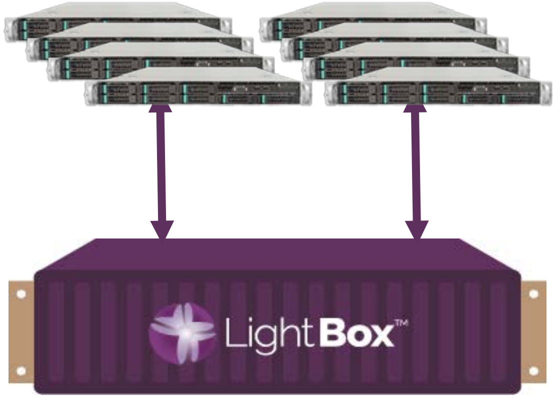
- SLO-driven volume allocation (SLO - service level objective)
  - New volumes are not allocated if combined SLO is not achieved by system capabilities.

IO Capping: Lab Results

- Scenario
  - Two tenants sending read requests of 4KB from 4 clients each.
    - A: queue depth **8**
    - B: queue depth **128**

Tenant A:
QD=**8**

Tenant B:
QD=**128**



Throughput

Each client gets **fair** share of BW

Each client gets **unfair** share of BW proportional to its queue depth

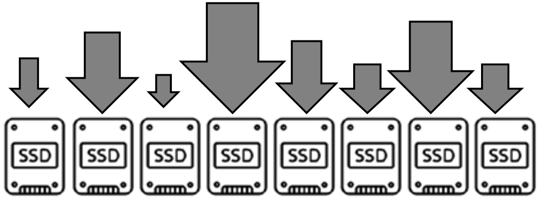
IO capping **enabled**

IO capping **disabled**

Balance IOs and SSD Endurance

Problem: Writes are not balanced across all SSDs

- Write amplification and garbage collection activity is different across SSDs
- Endurance of each SSD is different
- Read latency varies depending on which SSD is used to handle the read request



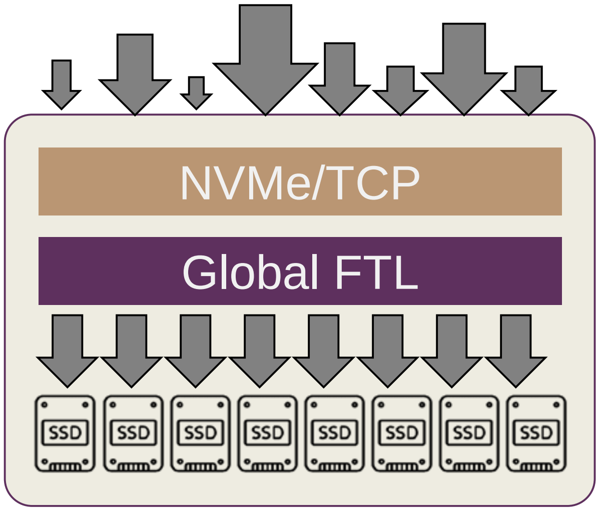
Balance IOs and SSD Endurance

Solution: Writes are distributed evenly across all SSDs as they come

- Append only, no write-in-place
- SW-controlled garbage collection

Result:

- Same endurance for all SSDs
- Write amplification is balanced
- Read latency is predictable
  - Each SSD is serving the same write activity when a read arrives



Summary

- LightOS is a first commercial high-performance NVMe/TCP target with data services.
- QoS is integral part of the system that copes with multi-tenant storage challenges.
- Read-write separation provides low read latency by avoiding head-of-line blocking.
- Per tenant IO capping provides guaranteed and isolated performance for every tenant.
- Global FTL balances writes uniformly across SSDs for endurance and predictable read latency.

Visit our partner booth #848 - International Computer Concepts to see a demonstration of LightOS NVMe/TCP

Contact information:

www.lightbitslabs.com

alex@lightbitslabs.com

Thank You

Session Agenda

- **2:15 - An NVMe/TCP Software-Defined Platform for Guaranteed QoS**
  - Alex Shpiner, System Architect, Lightbits Labs
- **2:30 - Comparing NVMe-oF on RoCE vs. TCP**
  - John Kim, Director Storage Marketing, Mellanox
- **2:45 Accelerating NVMe over TCP for Disaggregated Storage Applications**
  - Tom Spencer, Senior Director Product Marketing, Xilinx
- **3:00 Using SmartNICS and Buffer Management to Improve NVMe over TCP Performance**
  - Ron Renwick, VP of Products, Netronome
- **3:15 – Q&A**

Which Fabric for NVMe-oF

- InfiniBand for HPC, AI/ML
- FC for enterprise SAN (if you have it)
- Ethernet for everything else
- Assume going with NVMe-oF on Ethernet

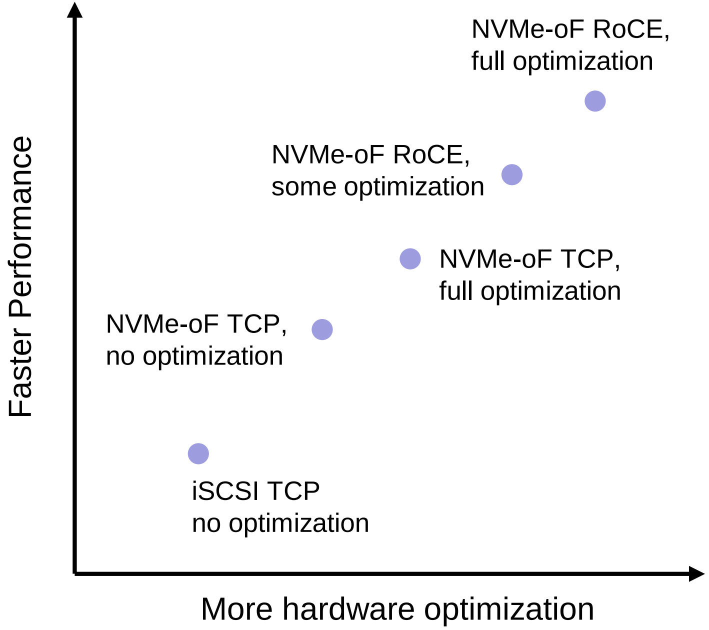
- **RoCE or TCP?**

Decision Criteria

- Performance
- Adapter support
  - Interop, Offloads, Cost, Availability
  - Software stack maturity
- Switch/network changes

Performance vs. Optimization

- More hardware assist
  - Best performance
  - Specific adapters required
  - Specific switch settings
- Zero hardware optimization
  - Slowest performance
  - Runs on any hardware
  - No switch setting changes



Comparing the Options

Factors	TCP plain	TCP optimized	RoCE some optimization	RoCE fully optimized
NIC choices	Any	Several	Several	Several
NIC Interop	Any-to-any	Depends	Several	Several
Switch choices	Any	Many	Many	Several
Switch setting changes	None	Minor	Minor	Major

NVMe-oF over TCP Questions

- What are your performance requirements?
  - Tolerance for latency?
- Can you deploy special NICs selectively?
  - Is same NIC required on both ends?
- Will you make any switch changes?
  - Can you deploy switch changes selectively?

NVMe-oF over RoCE Questions

- Do you need the RoCE performance boost?
  - Can you mix TCP and RoCE?
- Do your servers/NICs already support RoCE
  - Will other applications need RoCE?
- Would you make switch changes anyway?
  - To optimize other storage traffic

Thank You

Session Agenda

- **2:15 - An NVMe/TCP Software-Defined Platform for Guaranteed QoS**
  - Alex Shpiner, System Architect, Lightbits Labs
- **2:30 - Comparing NVMe-oF on RoCE vs. TCP**
  - John Kim, Director Storage Marketing, Mellanox
- **2:45 Accelerating NVMe over TCP for Disaggregated Storage Applications**
  - Tom Spencer, Senior Director Product Marketing, Xilinx
- **3:00 Using SmartNICS and Buffer Management to Improve NVMe over TCP Performance**
  - Ron Renwick, VP of Products, Netronome
- **3:15 – Q&A**

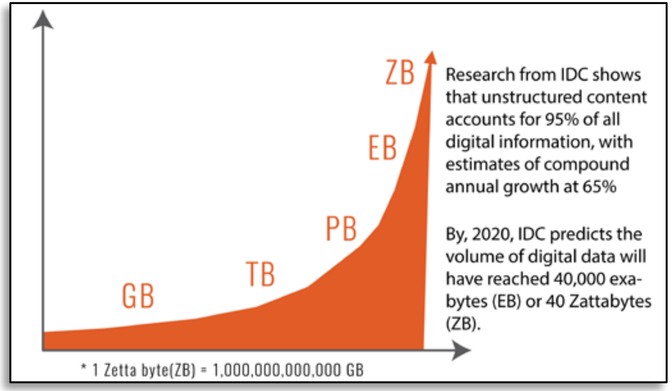
Accelerating NVMe/TCP for Disaggregated Storage Applications

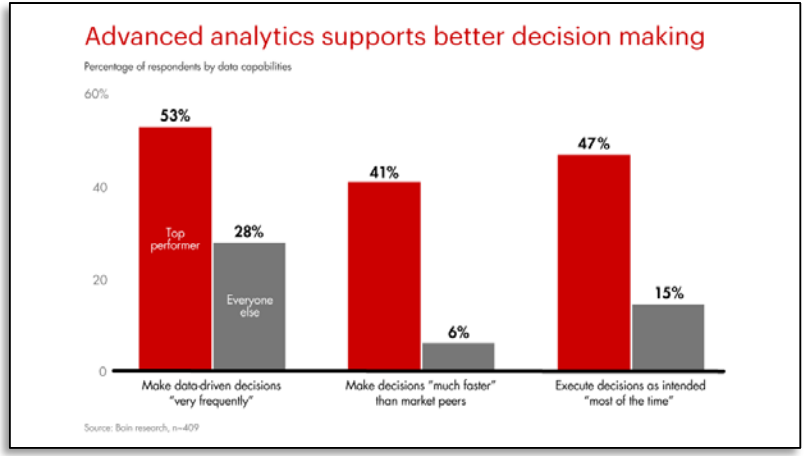
Tom Spencer

Big Data Requires Fast I/O Solutions

Flash Memory Summit

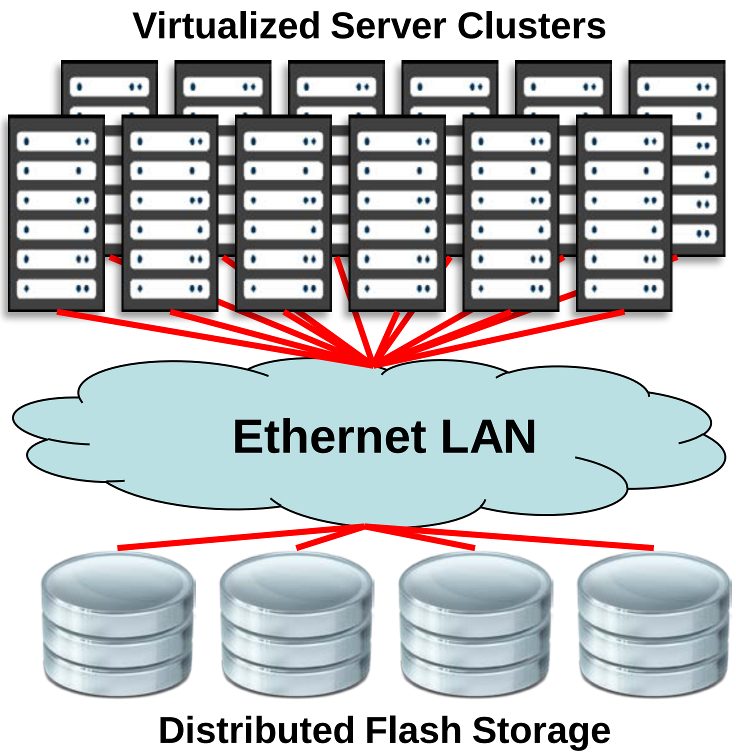
- The size of datasets is growing exponentially
- Rapid access to this data is critical for many use cases
  - Real-time analytics
  - Artificial Intelligence
  - Machine/Deep Learning
  - Business Intelligence





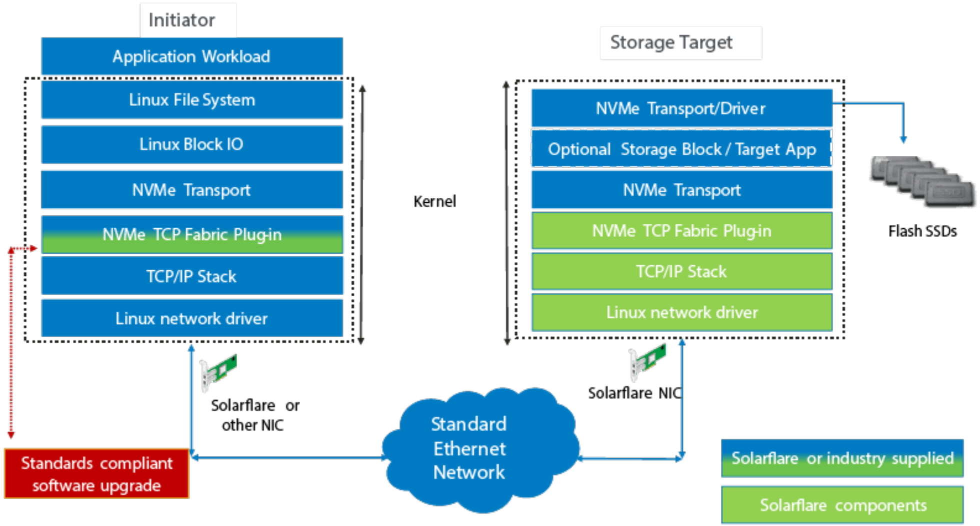
Typical Big Data Deployment

- Clusters with lots of highly virtualized servers
- Connection via Ethernet
- Widespread use of “flash area networks”
- Dynamically scalable



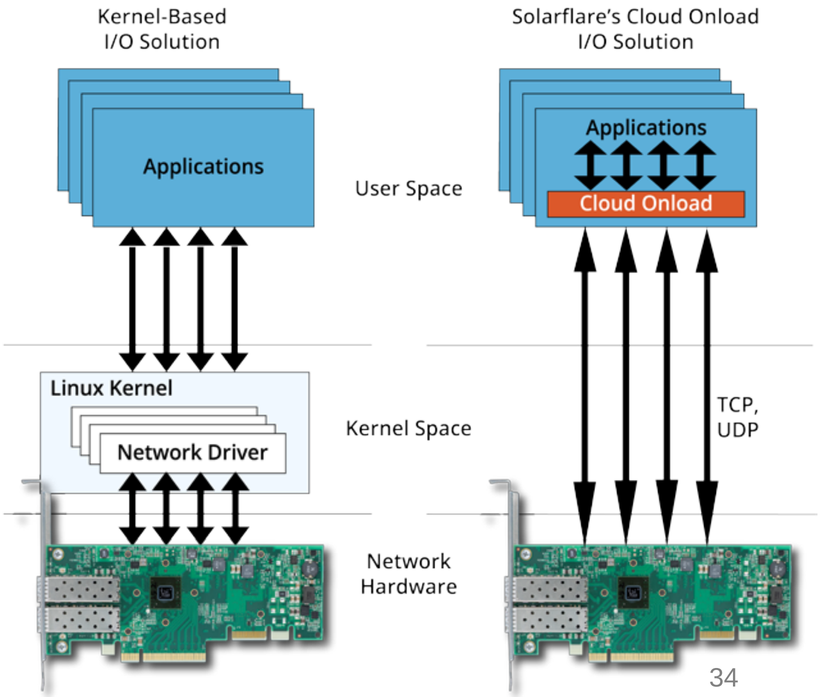
NVMe/TCP: Enabling Disaggregated Flash Storage Architectures

- NVMe/TCP was ratified in 2018 by NVM Express
- NVMe/TCP simplifies flash storage deployments
  - No “stranded servers”
  - No application modification
- Brings local flash performance to storage networks



User Space I/O: Further Acceleration of Big Data Applications

- Kernel-based drivers SLOW DOWN Big Data
- User space (kernel bypass) I/O solutions overcome this issue
    - No context switching
    - No memory copies
- User space I/O increases bandwidth while decreasing CPU utilization
    - Improved CapEx and OpEx
    - Better solution scalability

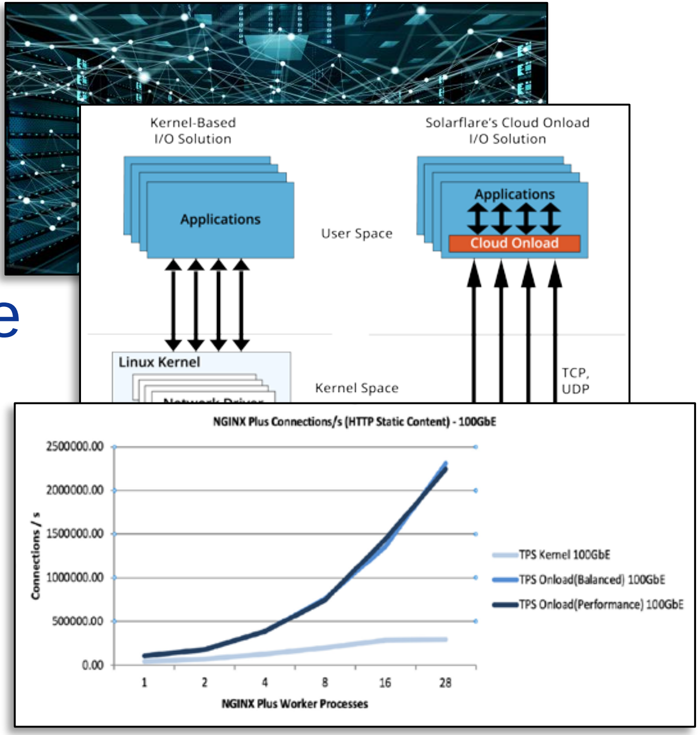


NVMe/TCP: Typical Disaggregated Storage Use Cases

- Artificial Intelligence/ Machine Learning
- Databases
- Container-Based Computing
- Real-Time Analytics
- High-Resolution Video Post-Production

Summary: NVMe/TCP + User Space Equals High Performance I/O

- Big Data requires high performance
- NVMe/TCP enables disaggregated flash storage network deployments
- Kernel space I/O slows down storage networks (even NVMe-oF networks)
- User space NVMe/TCP provides the performance Big Data needs



Thank You!

Tom Spencer
Sr. Director, Product Marketing
www.solarflare.com

Session Agenda

- **2:15 - An NVMe/TCP Software-Defined Platform for Guaranteed QoS**
  - Alex Shpiner, System Architect, Lightbits Labs
- **2:30 - Comparing NVMe-oF on RoCE vs. TCP**
  - John Kim, Director Storage Marketing, Mellanox
- **2:45 Accelerating NVMe over TCP for Disaggregated Storage Applications**
  - Tom Spencer, Senior Director Product Marketing, Xilinx
- **3:00 Using SmartNICS and Buffer Management to Improve NVMe over TCP Performance**
  - Ron Renwick, VP of Products, Netronome
- **3:15 – Q&A**

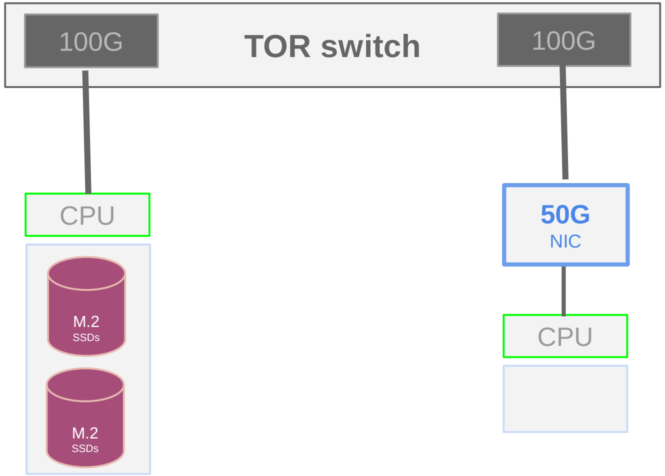
Using SmartNICs and Buffer Management to improve NVMe over TCP Performance

Ron Renwick

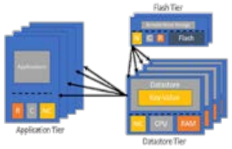
Netronome

Customer Use Case: Disaggregated NVMe Storage

Head Node
- SSD Storage tier
- Not a primary application tier

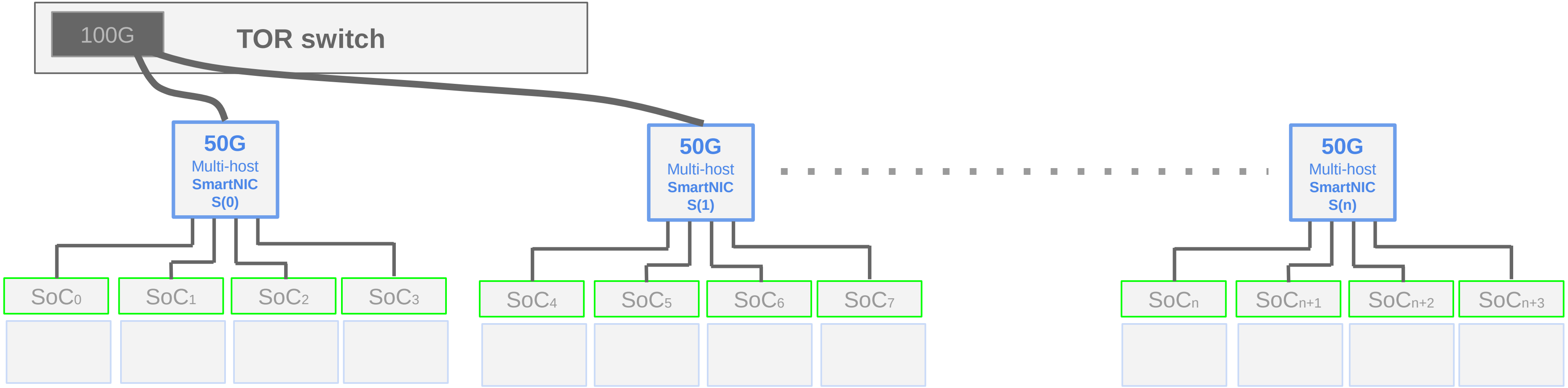


Client Node
- Compute/Application tier
- No Storage present
- Must access Head Node for application data

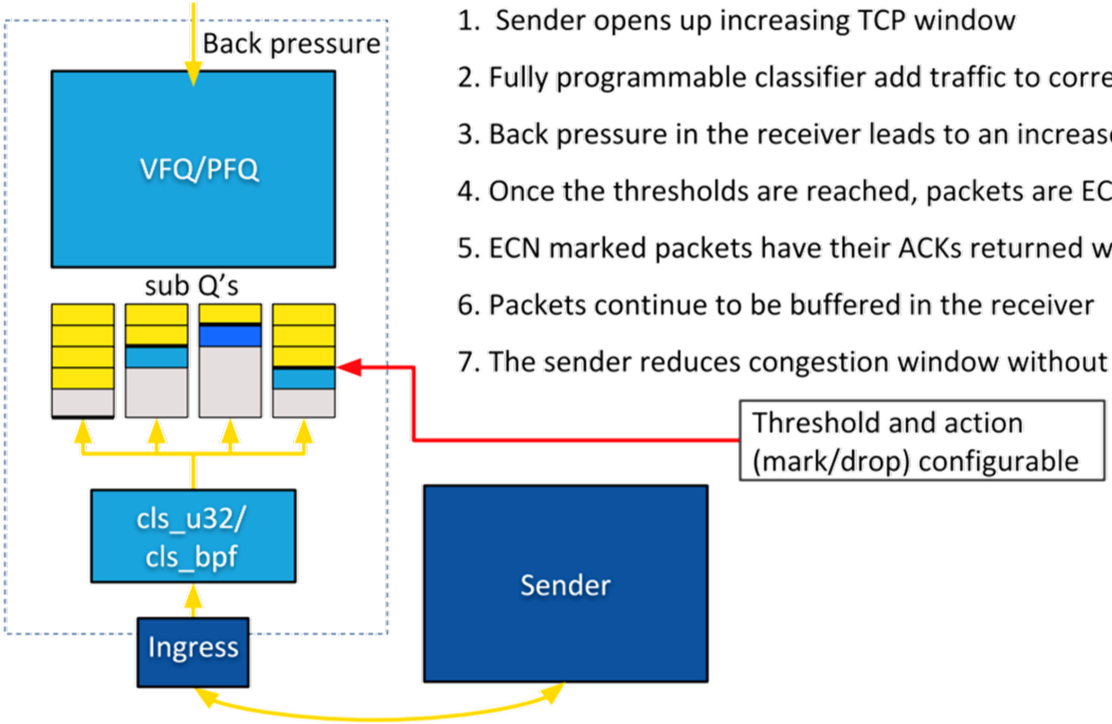


Test Bed: OCP Yosemite servers

- Single NIC creates impedance mismatch and increase tail latency for NVMe storage access
  - 50GbE Ingress into each sled
  - 4x PCIe Gen3x4 to each compute node
- Need alternative solution to provide improved storage access

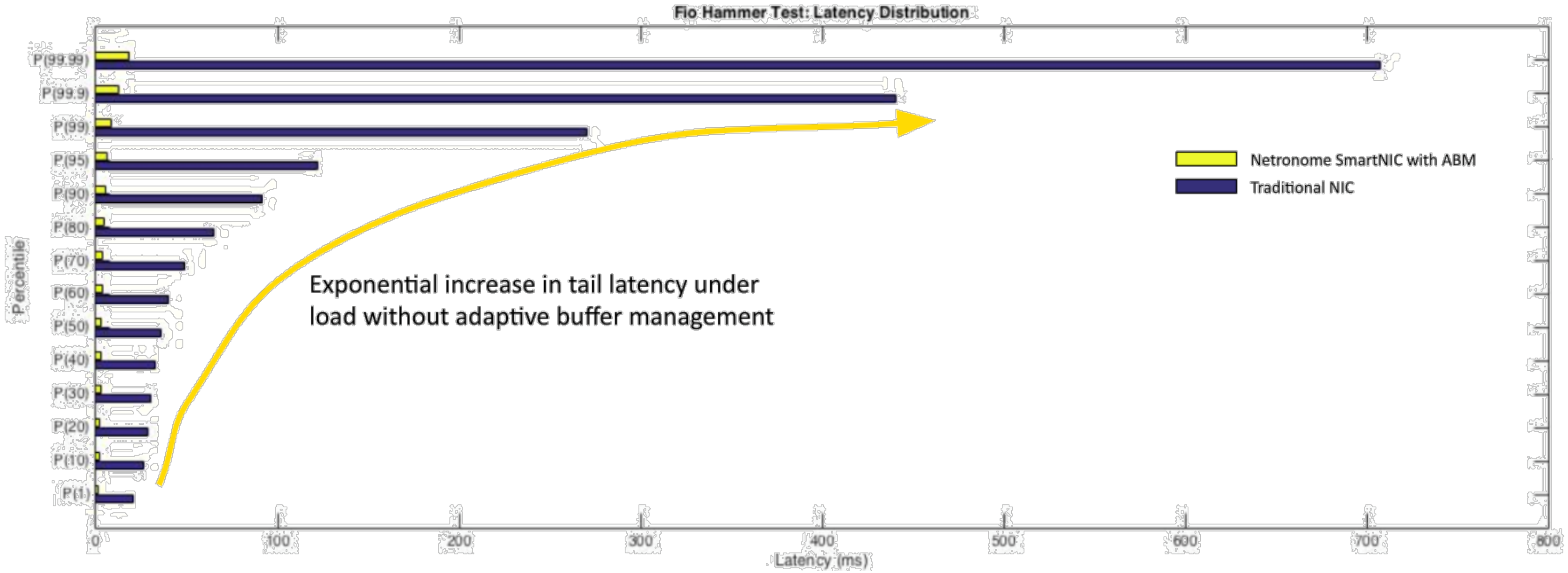


Buffer Management Architecture



1. Sender opens up increasing TCP window
2. Fully programmable classifier add traffic to correct queues
3. Back pressure in the receiver leads to an increase in buffer queue utilization
4. Once the thresholds are reached, packets are ECN marked (or dropped)
5. ECN marked packets have their ACKs returned with an ECN-Echo flag
6. Packets continue to be buffered in the receiver
7. The sender reduces congestion window without packets being dropped

Using ABM for Latency Improvement

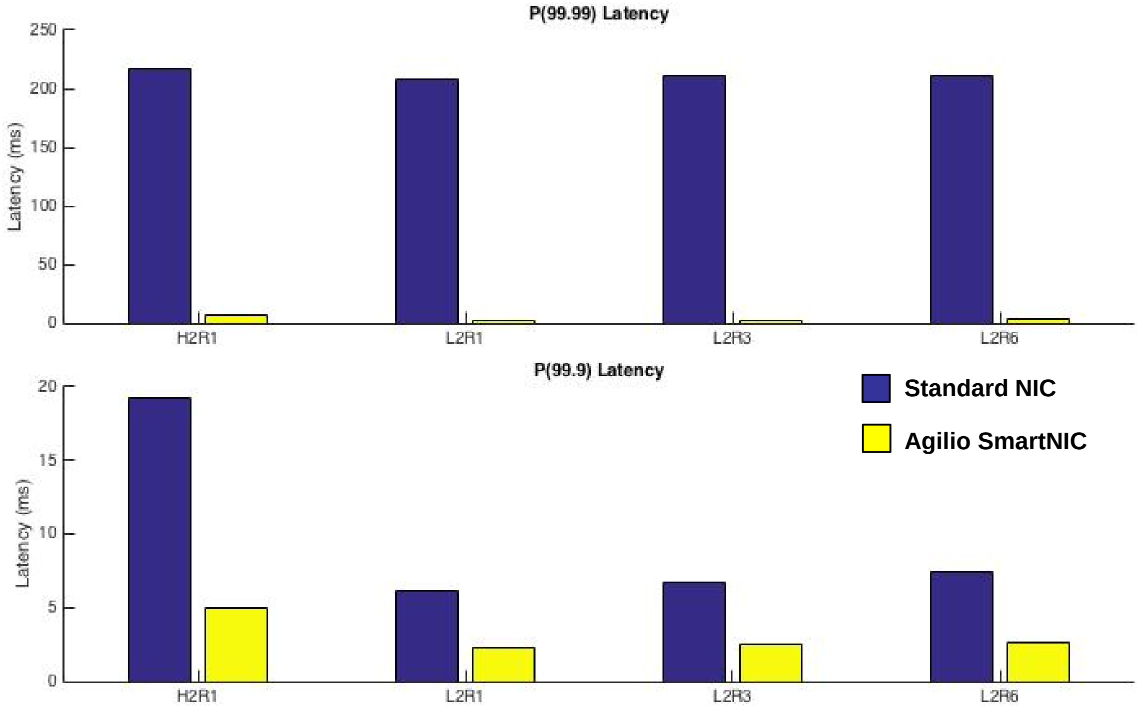


- Adaptive buffer management improvement relative to traditional TCP + CUBIC network congestion management
- Buffer management alleviates heavy loading imposed on PCIe link from a congested network
- ECN threshold ensures buffer is more efficiently used before imposing transmission backoff

Latency Reduction over TCP

33-70X Latency Reduction

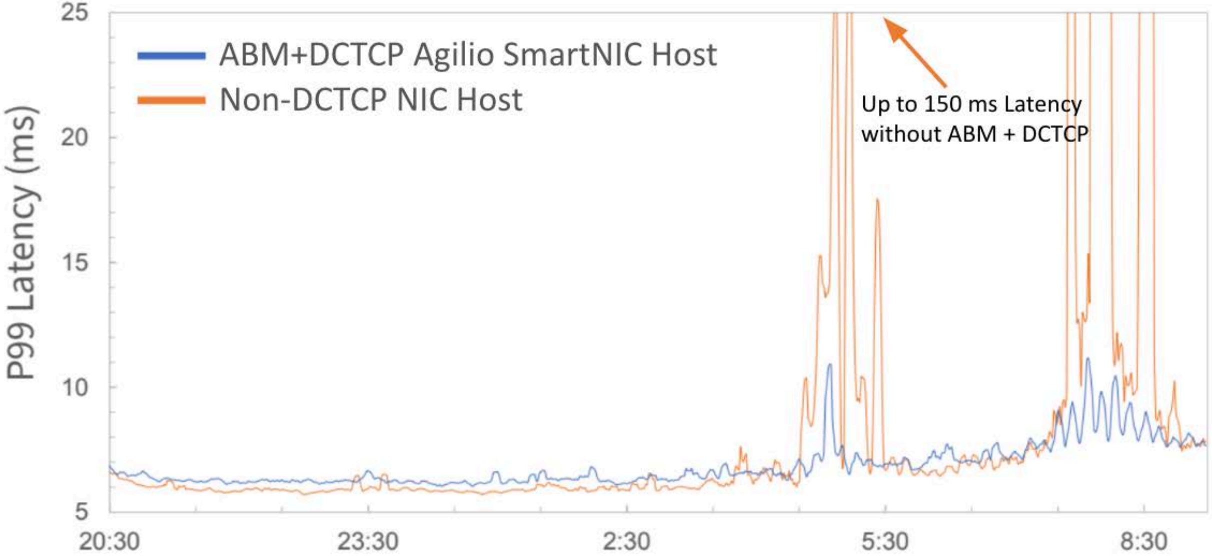
2.7-3.8X Latency Reduction



Allows 2 type 1 servers (Twin Lake) to replace a type 6 server (Leopard/Tioga Pass) within a disaggregated flash architecture.

Saves ~100W (33% of total power) per replacement.

Latency Improvements for Bursty Traffic w/ ABM + DCTCP



- Adaptive buffer management together with pre-emptive TCP congestion control (DCTCP) protocol reduces the negative effect heavy loading has on network latency
- Results collected in a production datacenter hosting customer VMs with real world workload traffic profiles
- Non-DCTCP NIC spiked up to 150 ms latency under load without buffer management

UsingNVMEoTCP (w/ ABM)

- Using ABM w/ TCP can improve NVMe
  - Reduced tail latency across server nodes
  - Eliminate Packet drops/retransmits
- Leverages standard Linux ECN and DCTCP
  - Not vendor specific

Thanks!

Q/A - discussion

Thanks!