STORAGE DEVELOPER CONFERENCE
SDC 21
BY Developers FOR Developers

Virtual Conference
September 28-29, 2021

BoF Panel

“No-Compromise” NVMeoF/TCP Offload

Presented by Greg Schulz – Server StorageIO™
Greg@StorageIO.com – www.StorageIO.com

BoF Discussion Overview – “No Compromise” NVMeoF/TCP Offload

Many server storage and I/O problems are often the result of host CPU and associated software bottlenecks. Likewise there are different layers and focus areas in the IT data center and cloud data infrastructure stack, from business apps to databases/repositories, to filesystems, data services and data protection, to operating systems, hypervisors, containers, hardware, software among other components.

All applications and their underlying data infrastructure resources and services have some type of Performance, Availabity, Capacity, Economic (PACE) and management demand attributes. Often in the quest to optimize something, somewhere in the data infrastructure stack, one or more PACE attributes are compromised and/or, additional complexity (and cost) increases resulting in loss of effectiveness.

With a continued shift towards software defined storage, networks and data infrastructures, host CPU cycles are in more demand. Boosting application performance, efficiency, and effectiveness of server CPUs including reducing overhead are key priorities for legacy and software defined datacenter environments.

BoF Discussion Agenda & Introductions – NVMeoF/TCP Offload

Opening Remarks

This panel explores the challenges, issues, and benefits of addressing NVMe over TCP deployments without compromise. The session will explore server, storage and I/O workload testing techniques, tools, methodology and approaches to show NVMe over Fabrics including TCP can be accelerated, while freeing up host CPU resources for other software defined workloads.

Introductions

- Greg Schulz – Independent Industry Analyst, Author, Consultant, Founder Server StorageIO™
- Bob Dugan – Director of Engineering at Chelsio Communications

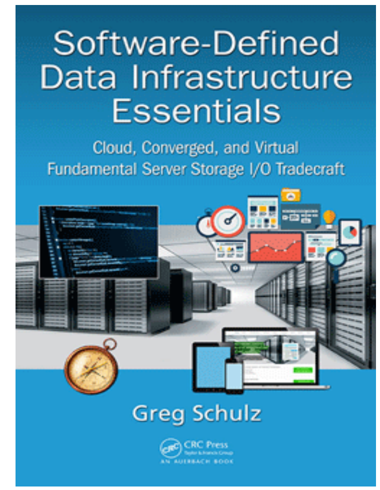
Brief Presentation and Perspectives

- Industry and Data Center Trends "Big Picture, Setting the Stage" – Greg Schulz
- Chelsio Perspectives Brief Presentation – Bob Dugan

Panel and Audience Q&A Discussion, Wrap-up

Industry & Data Center Trends – Greg Schulz StorageIO™

Greg has an Masters Software Engineering from University of St Thomas, worked as the customer in various IT organizations in roles from business applications to systems and data infrastructure. He has worked as a vendor, consulting analyst and author of several books including "Software-Defined Data Infrastructure Essentials" (CRC Press). Greg brings a diverse background with real world perspective across applications, data infrastructures, hardware, software, data protection, Performance and Capacity Planning as well as containers and clouds. Greg is a Microsoft MVP Cloud Data Center Management and previous ten-time VMware vExpert.



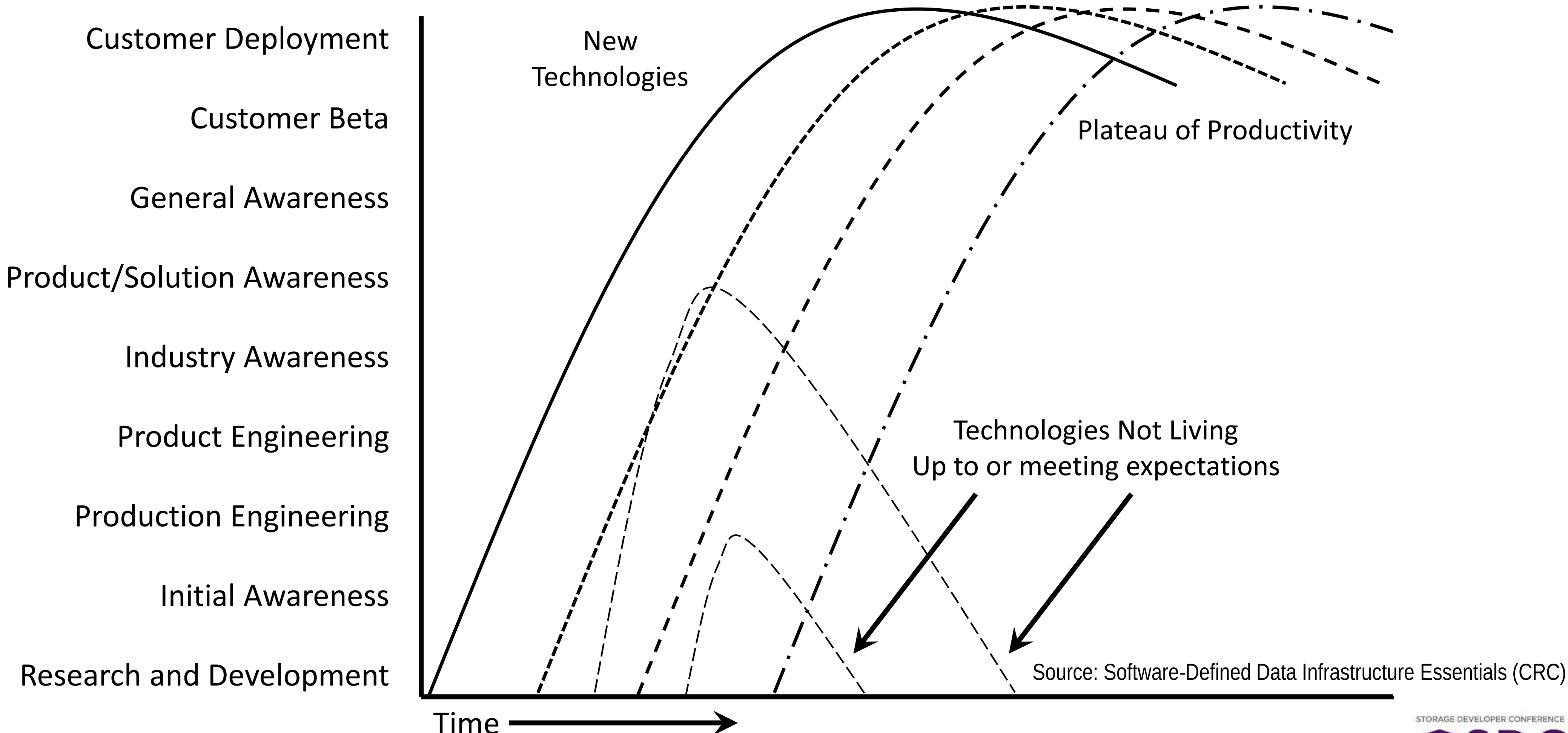
StorageIO.com/book4

@StorageIO

Greg@StorageIO.com

- ✓ Continued shift to software defined data infrastructures (servers, storage, networks)
- ✓ Increased demand for compute resources (CPU, GPU, xPU and other offloads)
- ✓ Expanding focus from resource utilization to effectiveness and productivity
- ✓ NVMe & NVMe over Fabrics (NVMeoF) including TCP challenges and opportunities
- ✓ Many I/O and storage performance problems are software and CPU problems
- ✓ Software needs hardware, hardware needs software, even serverless ;)

Industry and Customer Trends – Adoption and Deployment Timelines



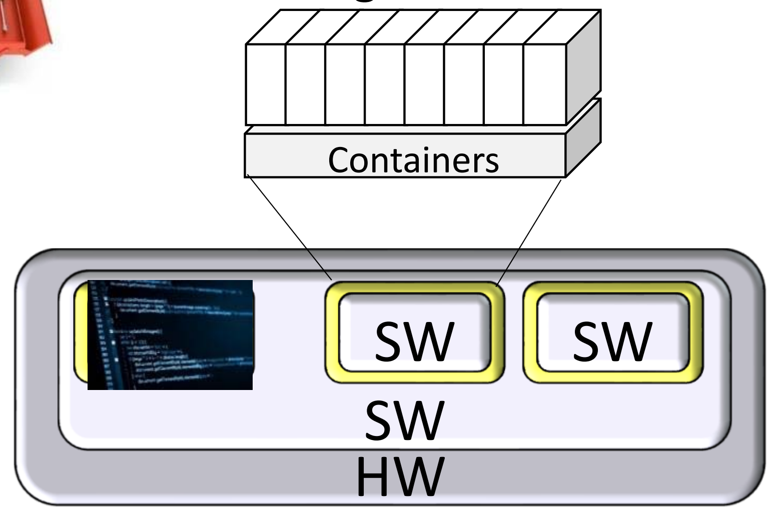
Industry Trends Perspectives – Software Needs Hardware, HW needs SW

"Tin" or "Hardware"
Wrapped SW "Appliance"

What About Management?

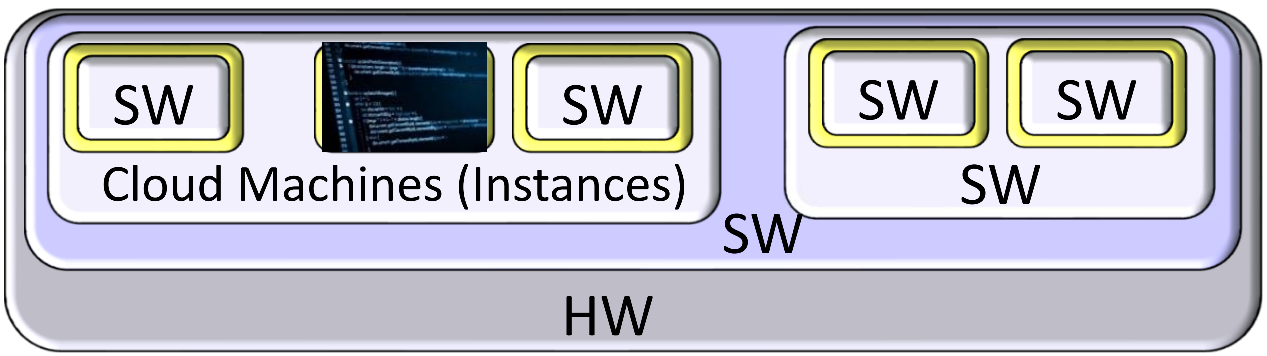
Where Applications (SW) Run
Mode of Deployment

Applications wrapped in Image
Running in a Container



"VM" or "Virtualization"
Wrapped SW (SDDC)

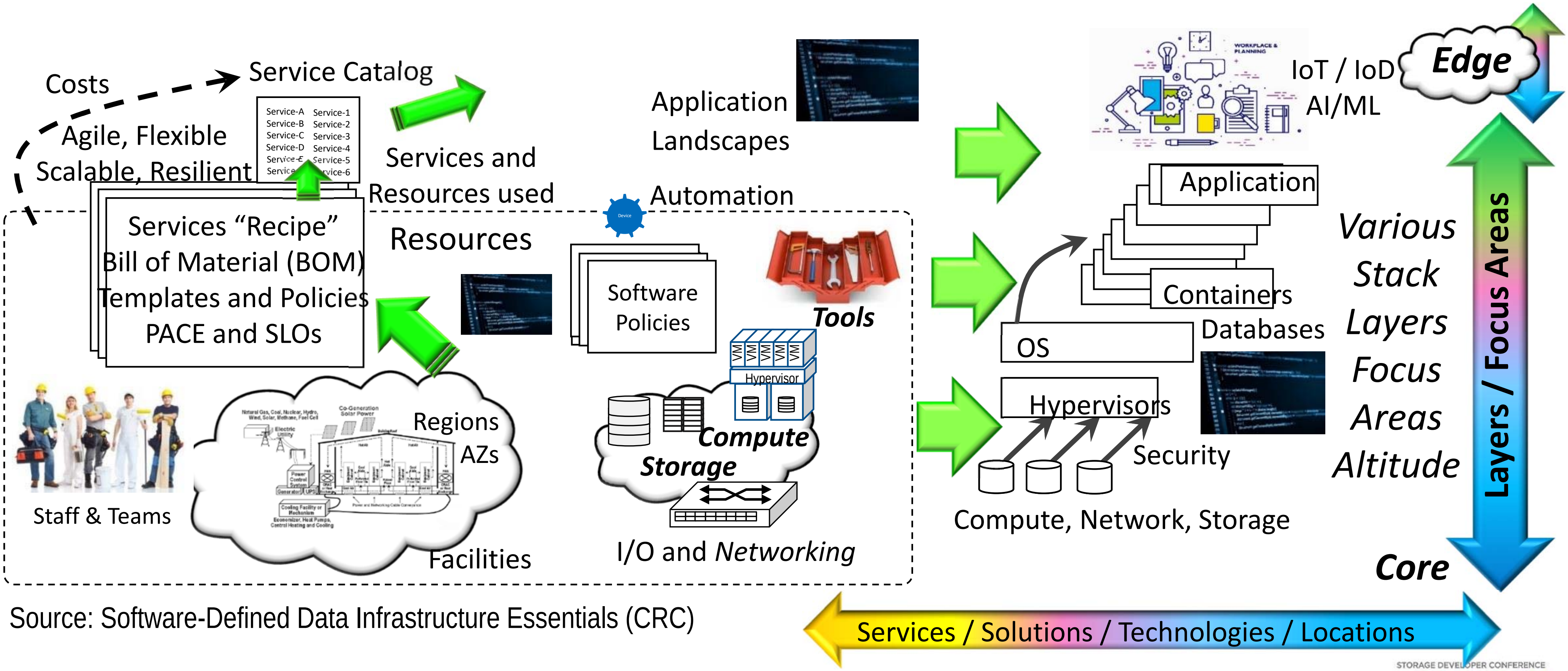
"Cloud" Wrapped SW



Source: Software-Defined Data Infrastructure Essentials (CRC)

OS = Operating System HW = Hardware SW = Software VM = Virtual Machine SDDC = Software Defined Data Center

Data Infrastructures, from on-prem to cloud, core to edge

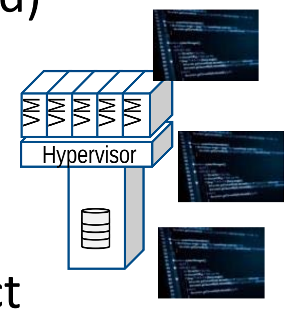


Source: Software-Defined Data Infrastructure Essentials (CRC)

STORAGE DEVELOPER CONFERENCE SDC 21

Industry Trends Perspectives – Key themes and trends

- ✓ Continued adoption of software defined data infrastructures
  - o Software requires hardware, even for serverless ;)
- ✓ Growth occurring at core (cloud & on-prem) and edge (remote & distributed)
  - o Legacy application landscapes and new, merging workloads

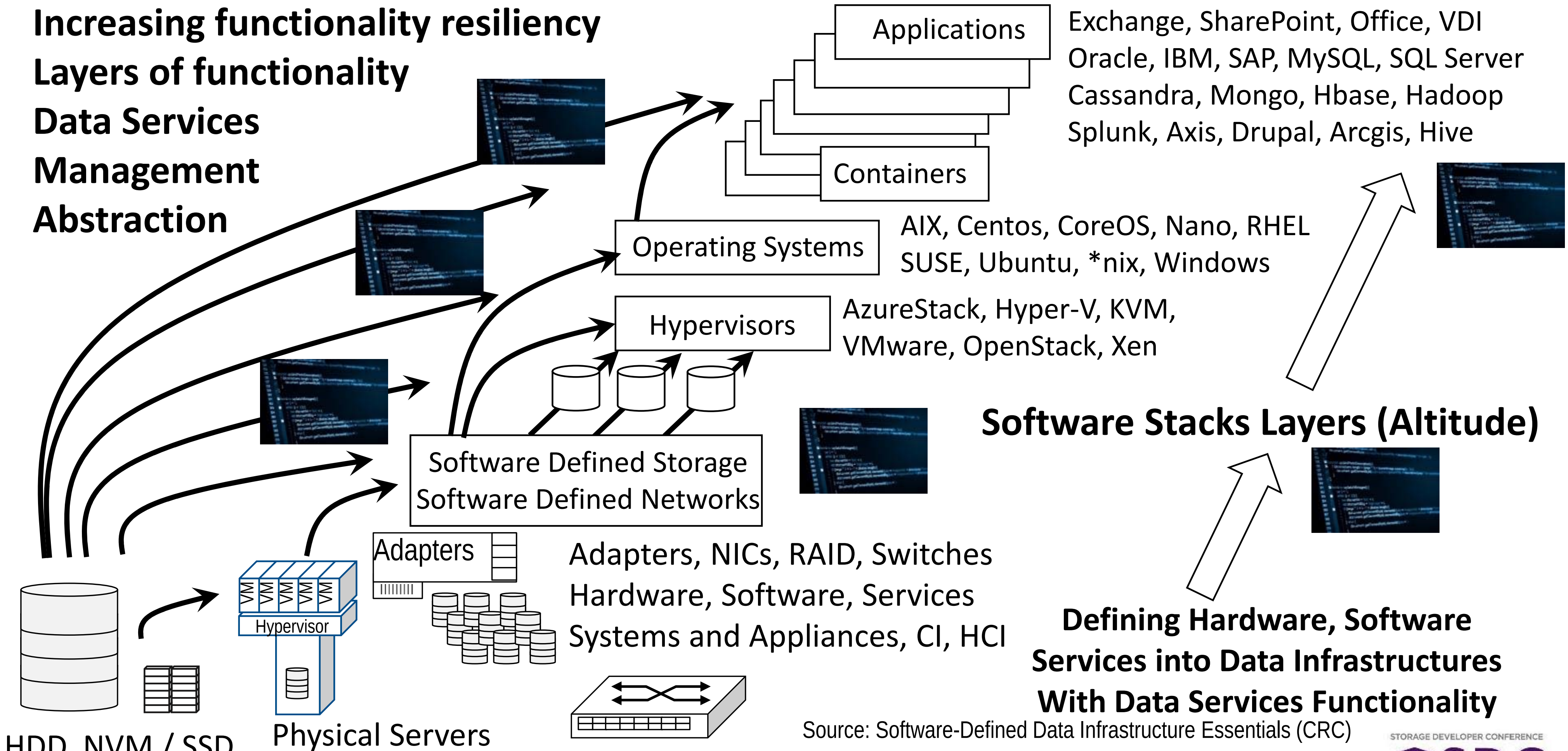


- ✓ Lines are blurring where server, storage, I/O, networking begin and end
  - o In the past, server, storage & I/O were more integrated (e.g. packaging)
- ✓ Small percentage changes on high frequency/volume things have big impact
- ✓ NVMe is the server storage I/O protocol of the future and today
  - o I/O Performance (bandwidth, low latency)
  - o Scalable from laptop to datacenter
  - o Flexibility (various topologies, connectivity options)
  - o Reduced CPU and server I/O overhead

View more at http://thenvmeplace.com/

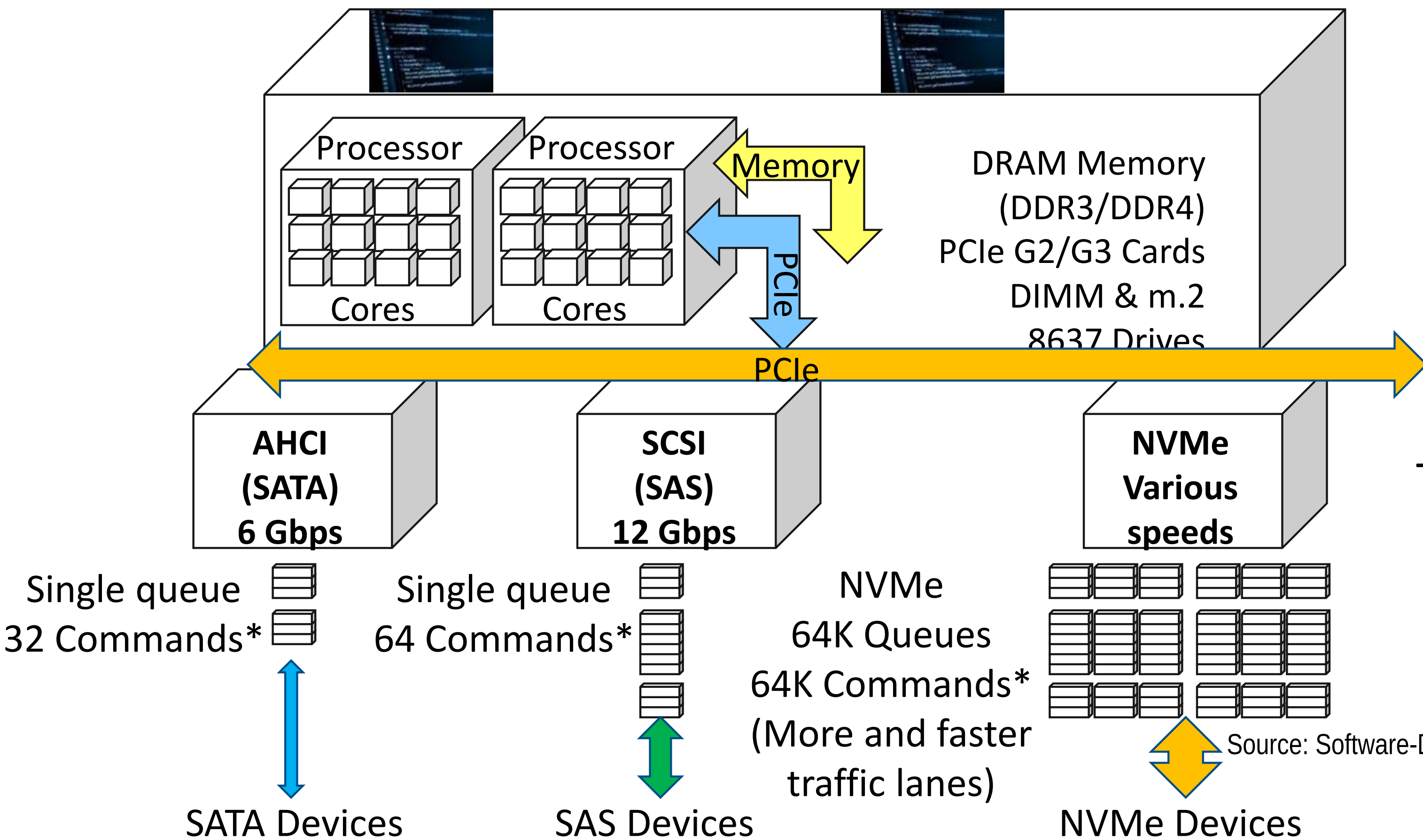
Data Infrastructures – Stacks, Layers, Altitude - Different Focus Areas

Increasing functionality resiliency
Layers of functionality
Data Services
Management
Abstraction



Source: Software-Defined Data Infrastructure Essentials (CRC)

Industry Trends Perspectives – Data Infrastructure Resources

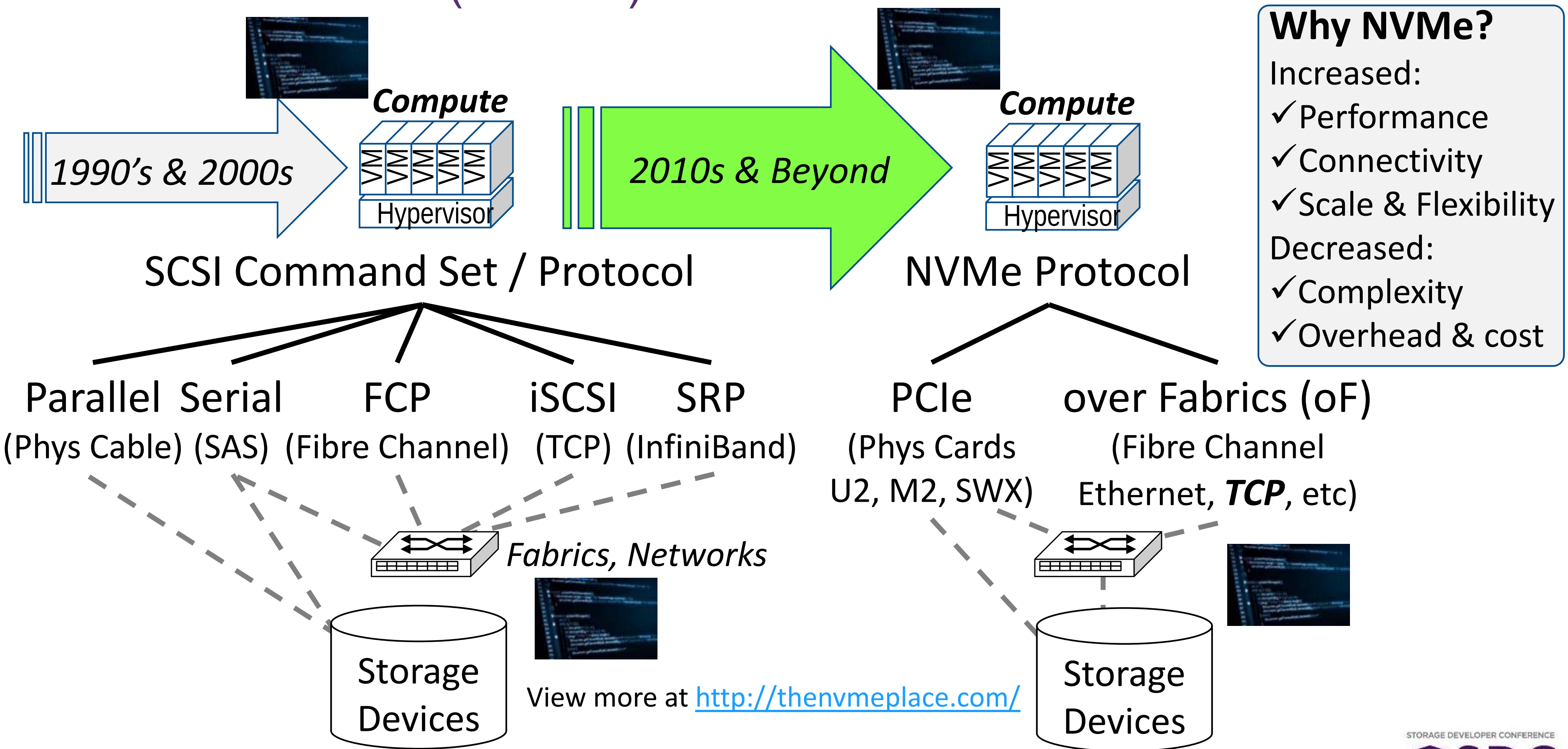


Bluetooth, Ethernet
FC, GPU, IBA
m.2, mSATA
NFC, RoCE, Serial,
Thunderbolt, USB, WiFI

Source: Software-Defined Data Infrastructure Essentials (CRC)

* Protocol Specification, see vendor, device or driver specific implementation notes

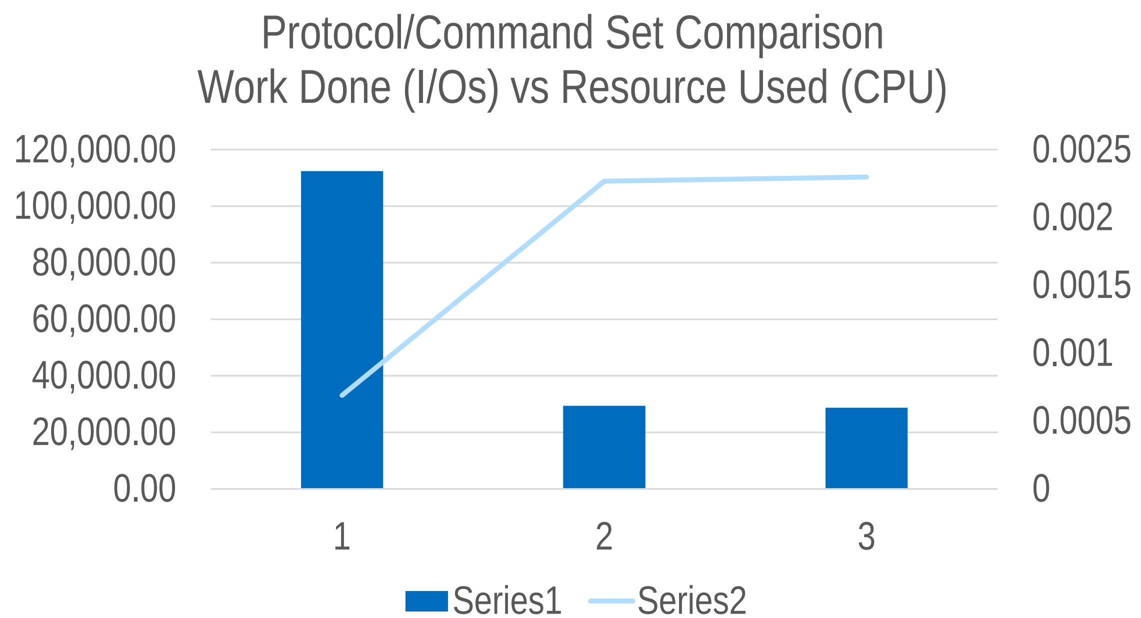
NVMe & NVMeoF (Fabrics) Data Infrastructure Performance Enabler



View more at http://thenvmeplace.com/

Industry Trends Perspectives – Protocol Productivity vs Resource Used

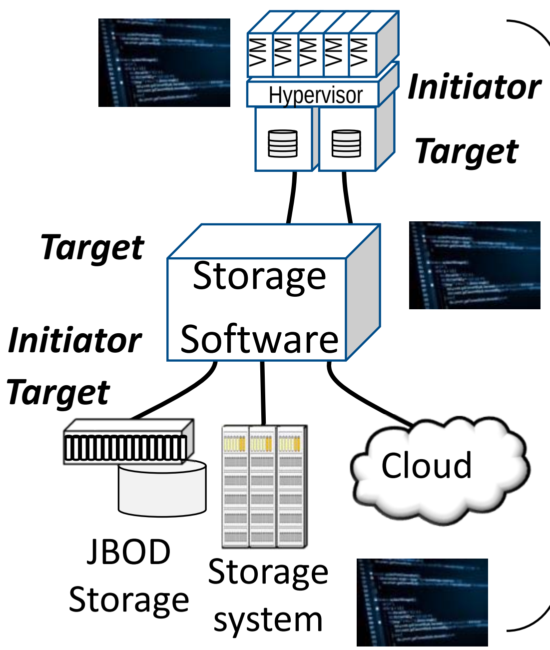
		8 KB I/O	1 MB I/O
NAND flash SSD NVMe PCIe AiC		100% Ran. Read	100% Ran. Read
	IOPs	112,353.6	1,336.94
	Bandwidth	877.76	1,336.94
	Resp.	1.30	191.27
	CPU/IOP	0.000689	0.009798
12 Gb SAS	IOPs	29,373.5	416.68
	Bandwidth	229.48	416.68
	Resp.	4.56	614.22
	CPU/IOP	0.002267	0.01416
6 Gb SATA	IOPs	28,677.12	356.06
	Bandwidth	224.04	356.06
	Resp.	4.67	718.81
	CPU/IOP	0.002298	0.015166



Source: Software Defined Data Infrastructure Essentials (CRC)
View more at StorageIO.com/book4.html

View more at https://storageioblog.com/server-and-storage-io-benchmark-resources/

Industry Trends – Storage System Fundamentals – Different Packaging



Data Services and Storage Functionality

Access and interface
Metadata and monitor
Data services and PACE
Cache and tiering
Reporting and analytics
Device, end-point and namespace
Management

Various Software and Applications

Block	File	Object
iSCSI, FC, NBD SAS, NVMe	POSIX, NFS, pNFS CIFS/SMB, HDFS, FUSE	S3, Swift, Blobs, ODBC Key Value, API

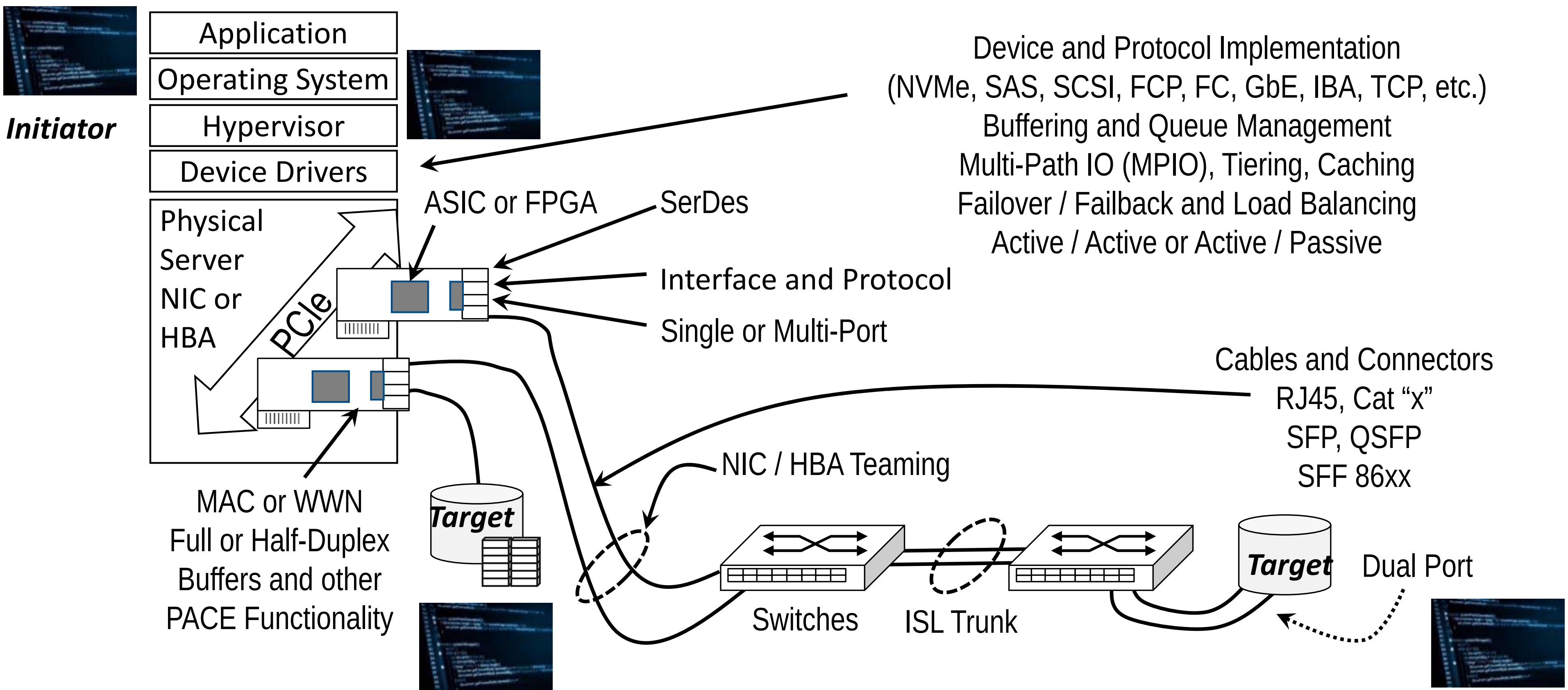
Storage System Functionality

PACE, Name space and end-pint management
Optional data services, tiering, cache, gateway
Access, device emulation, protocol implementation
Management, monitoring, security
In-band, Out-of-band, fast-path control path

Block	File	Object and API
SAS, SATA PCIe, NVMe NBD, FC, iSCSI	SMB/CIFS HDFS	S3, Swift, Blobs Public, Private Cloud

Source: Software-Defined Data Infrastructure Essentials (CRC)

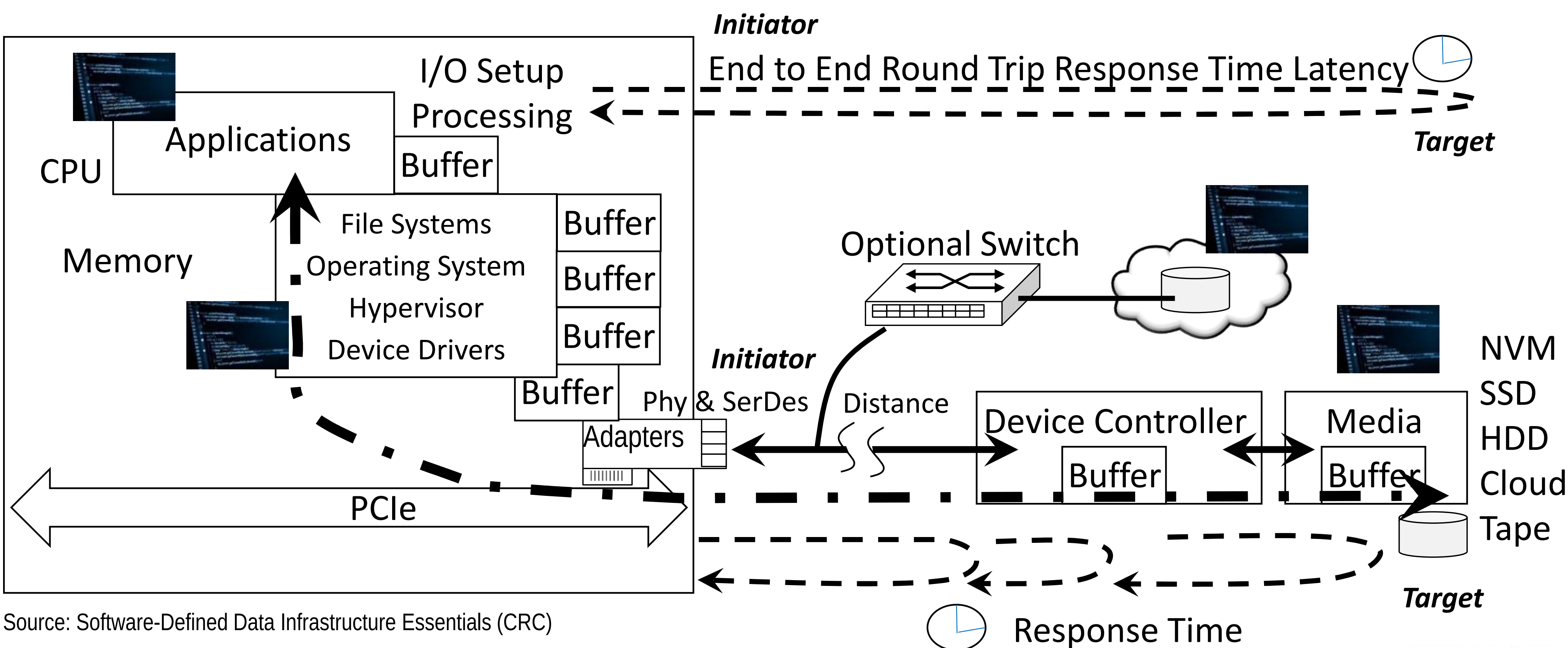
Industry Trends – Measuring Server Storage I/O – Points of Interests



Source: Software-Defined Data Infrastructure Essentials (CRC)

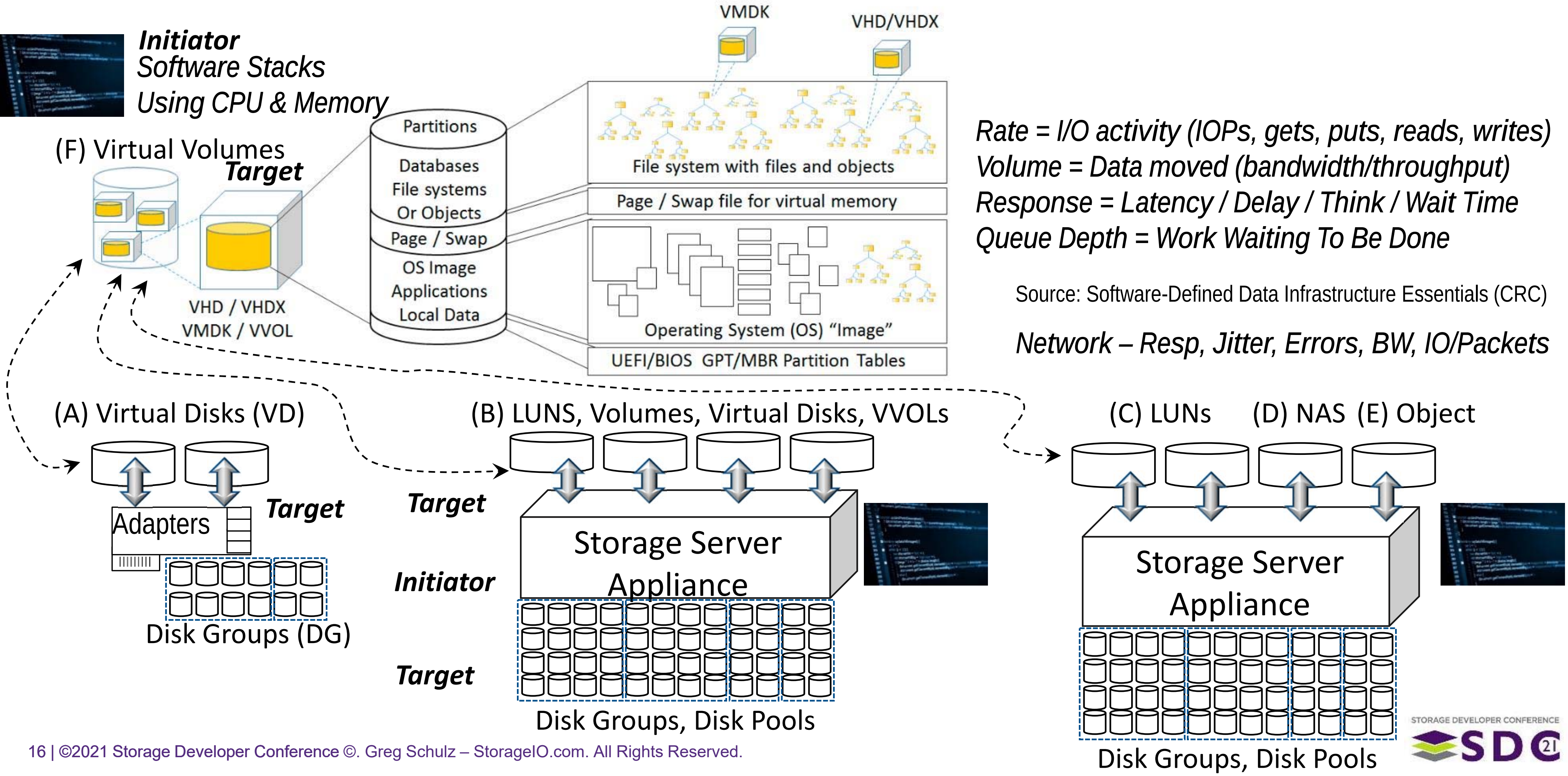
STORAGE DEVELOPER CONFERENCE SDC 21

Industry Trends – Measuring Server Storage I/O – Points of Interests



Source: Software-Defined Data Infrastructure Essentials (CRC)

Industry Trends – Measuring Server Storage I/O – Points of Interests



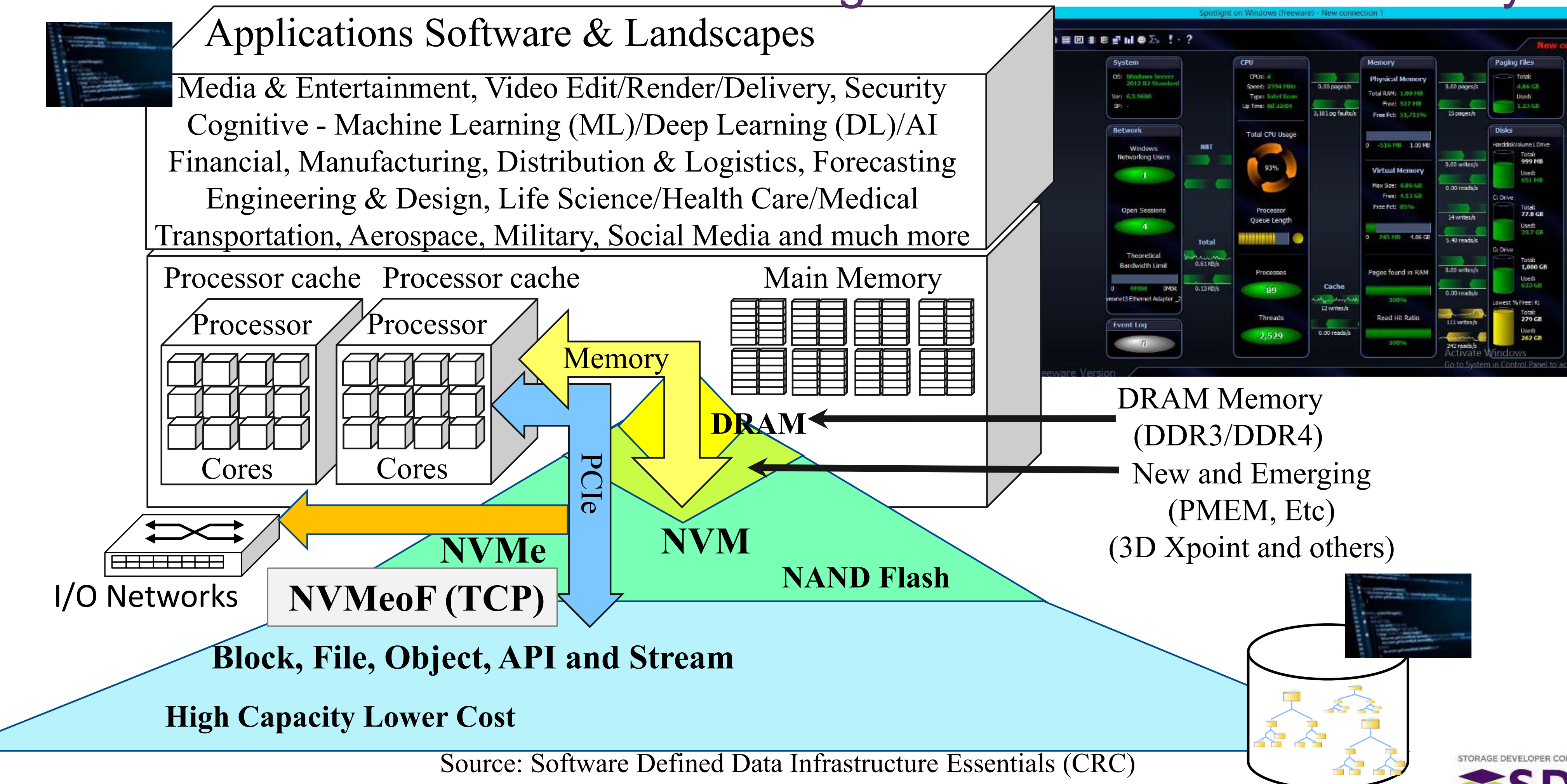
Rate = I/O activity (IOPs, gets, puts, reads, writes)
Volume = Data moved (bandwidth/throughput)
Response = Latency / Delay / Think / Wait Time
Queue Depth = Work Waiting To Be Done

Source: Software-Defined Data Infrastructure Essentials (CRC)

Network – Resp, Jitter, Errors, BW, IO/Packets

STORAGE DEVELOPER CONFERENCE SDC 21

Data Infrastructures – Maintain insight across various IT stack layers



Source: Software Defined Data Infrastructure Essentials (CRC)

Industry Trends Perspectives – Key themes and trends

- ✓ Software placing more demand on compute capabilities
  - o Host server CPU, GPU, xPU, ToEs and Offloads
- ✓ Many storage I/O problems are tied to host server CPU and software bottlenecks
  - o Fast applications and software need fast servers (CPU), memory, I/O and storage
- ✓ Expanding focus from utilization towards resource effectiveness and productivity
  - o High server CPU utilization may not be good if high system/kernel overhead
- ✓ TOEs, GPUs and other off loads are a key enablers for data infrastructures
  - o Get more useful, productive work done, boost effectiveness of resources
- ✓ Maintain situational awareness up and down “the stack” at different layers
  - o Avoid flying blind, leverage metrics that matter up from different stack layers
  - o Leverage compound metrics that show bigger picture, such as CPU used per IOP

STORAGE DEVELOPER CONFERENCE

SDC 21

BY Developers FOR Developers

Virtual Conference
September 28-29, 2021

BoF Panel

"No-Compromise" NVMeoF/TCP Offload

Presented by Greg Schulz – Server StorageIO™

Greg@StorageIO.com – www.StorageIO.com

STORAGE DEVELOPER CONFERENCE

SDC 21

BY Developers FOR Developers

Virtual Conference
September 28-29, 2021

BOF Session “No-compromise” NVMe/TCP Deployment using Server Storage I/O Offload

Part 2: Boosting Performance While Reducing Costs

Presented by Bob Dugan

Chelsio Communications, Inc.

Agenda

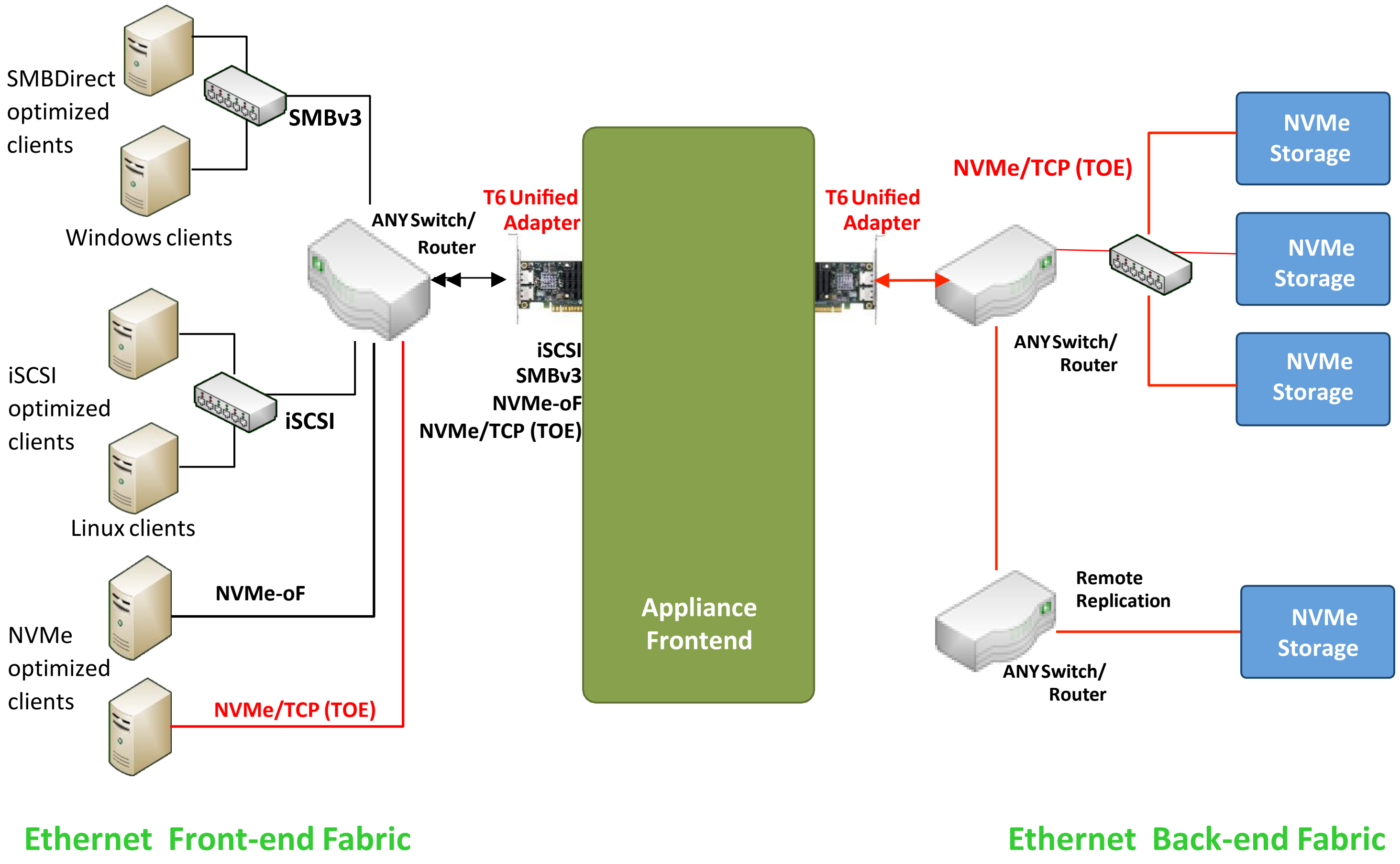
- NVMe/TCP using TCP/IP Offload
  - NVMe/TCP using TOE – Highlights
  - NVMe/Ethernet Fabric with TCP
  - Host-Based TCP/IP vs TCP Offload Engine (TOE)
  - NVMe/TCP (TOE) Layering – Closer View
- Performance Benchmarks
  - NVMe/TCP (TOE) – BW & IOPs Test Configuration
  - Kernel NVMe/TCP (TOE) – CPU Savings
  - NVMe/TCP (TOE) – Target Bandwidth & IOPs
  - TOE Jitter Handling
  - NVMe/TCP (TOE) – Latency Test Configuration
  - NVMe/TCP Latency Measurement Comparison
- Testing NVMe/TCP (TOE)
  - No Compromise Testing
- Conclusions
- Q&A and General Discussion

NVMe/TCP using TCP/IP Offload

NVMe/TCP using TOE – Highlights

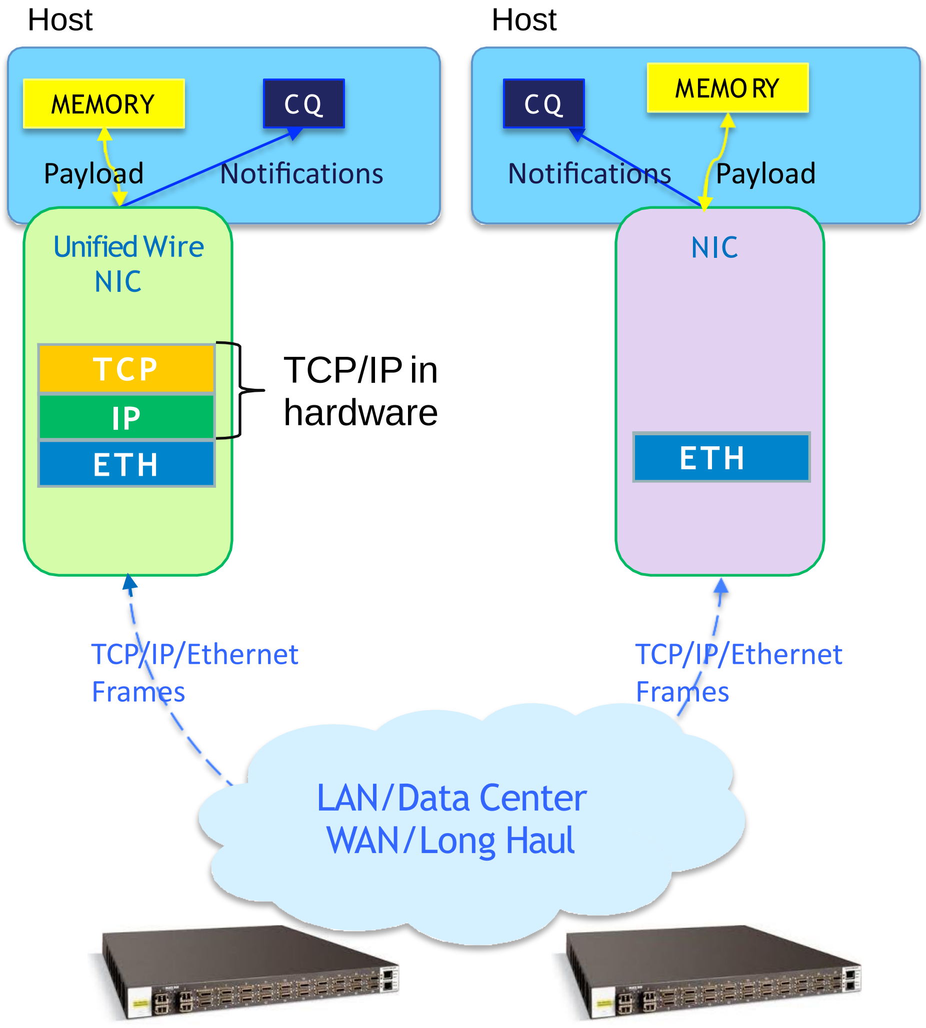
- Extends NVMe over fabrics using TCP/IP at large scale
  - TOE allows scaling more effectively
  - Free ups CPU from network system overhead
  - Reduces congestion on the network
- NVMe/TCP using TOE first proof point
  - Chelsio 100GbE TOE
    - 8.85 µs delta latency between remote and local storage
    - 2.9 Million IOPs at 4K I/O size
    - Reduced host CPU by up to 50% vs host-based TCP/IP

NVMe/Ethernet Fabric with TCP



T6 supports iSCSI, SMBDirect, NVMe-oF, & NVMe/TCP offload simultaneously

Host-Based TCP/IP vs TCP Offload Engine (TOE)

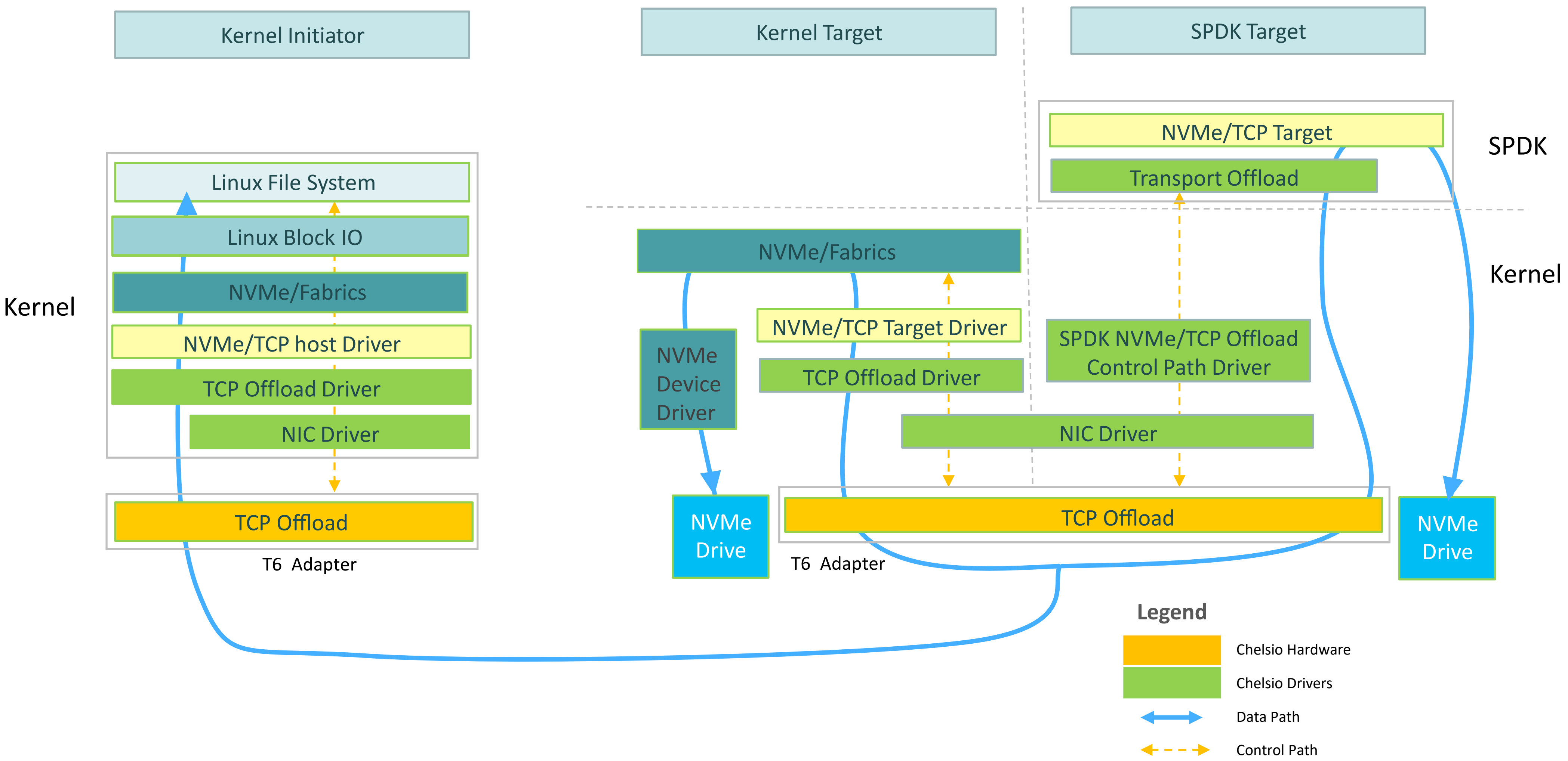


Any Switches/Routers

- TCP/IP Open, transparent and mature
  - Runs on Host CPU or TOE
  - IETF Standard (1981)
  - Plug-and-play
  - Built-in reliability congestion control, flow control
  - Natively routable
- Cost effective
  - Regular switches
  - Same network appliances
  - Works with DCB but NOT required
  - No need for lossless configuration
- No network restrictions
  - Architecture, scale, distance, RTT, link speeds
- Hardware performance
  - Exception processing in HW
  - Ultra low latency
  - High packet rate and bandwidth

STORAGE DEVELOPER CONFERENCE SDC 21

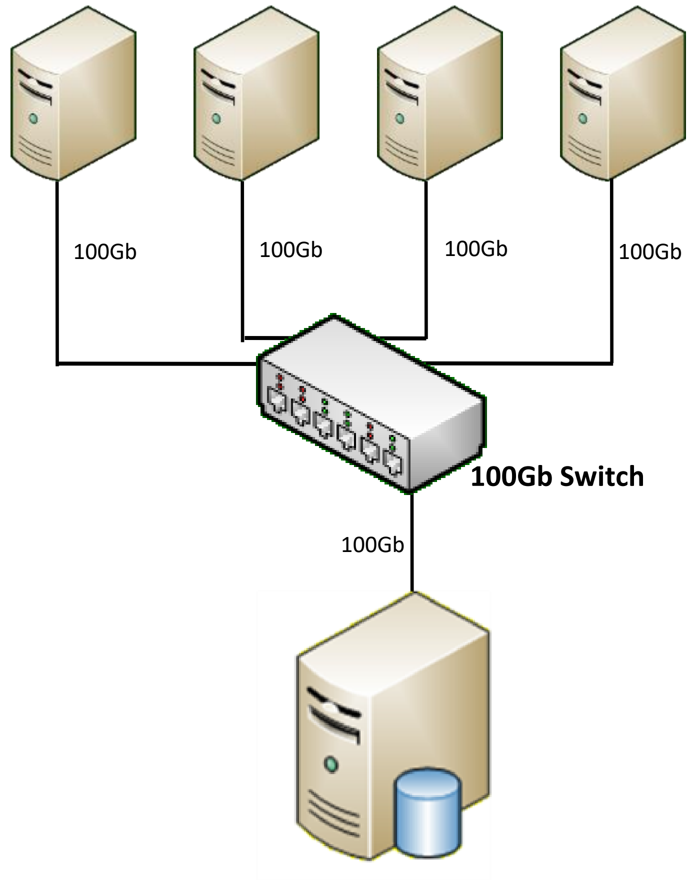
NVMe/TCP (TOE) Layering – Closer View



STORAGE DEVELOPER CONFERENCE SDC21

Performance Benchmarks

NVMe/TCP (TOE) – BW & IOPs Test Configuration

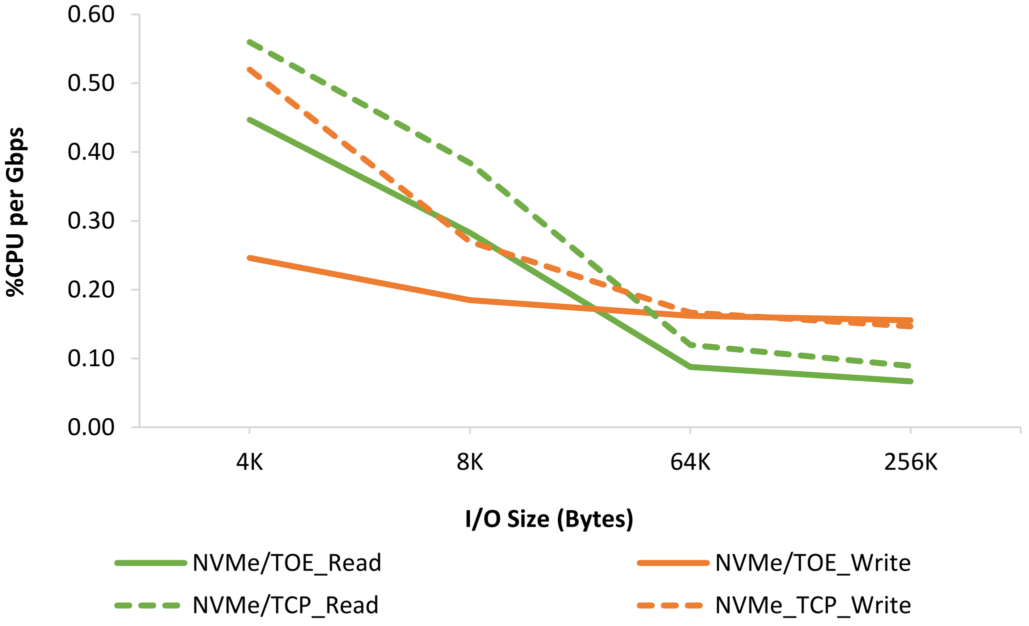


- Hosts
  - 1 Intel Xeon CPU E5-1620 v4
    - 8 cores (HT enabled) @ 3.50GHz
  - 32GB of RAM
  - Chelsio T62100-CR (2 x 100Gbps)
  - RHEL 8.3 (5.4.143 kernel)
- Target
  - 2 Intel Xeon CPU E5-2687W v4
    - 48 cores (HT enabled) @ 3.00GHz
  - 128GB of RAM
  - Null Block Devices
  - Chelsio T62100-CR (2 x 100Gbps)
  - RHEL 8.3 (5.4.143 kernel)

Note: Benchmarks details presented are in the Chelsio White Paper "100G Kernel and User Space NVMe/TCP Using Chelsio Offload": https://www.chelsio.com/wp-content/uploads/resources/t6-100g-nvmetcp-offload.pdf

STORAGE DEVELOPER CONFERENCE SDC 21

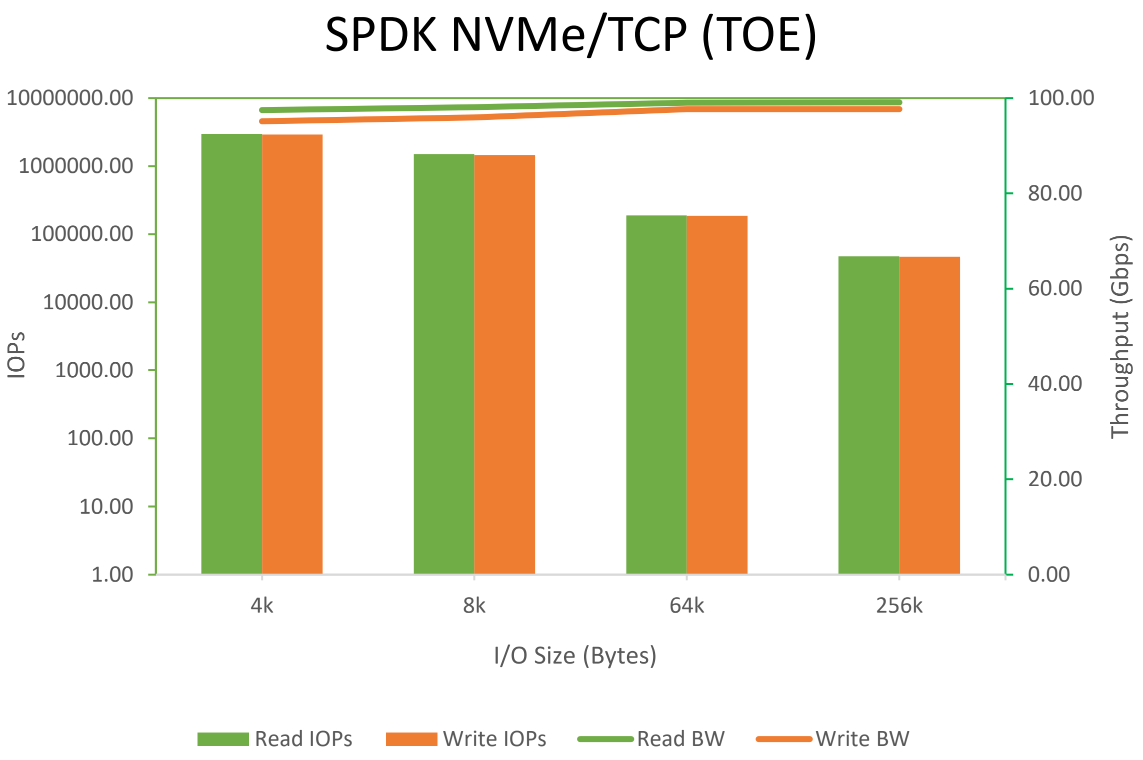
Kernel NVMe/TCP (TOE) – CPU Savings

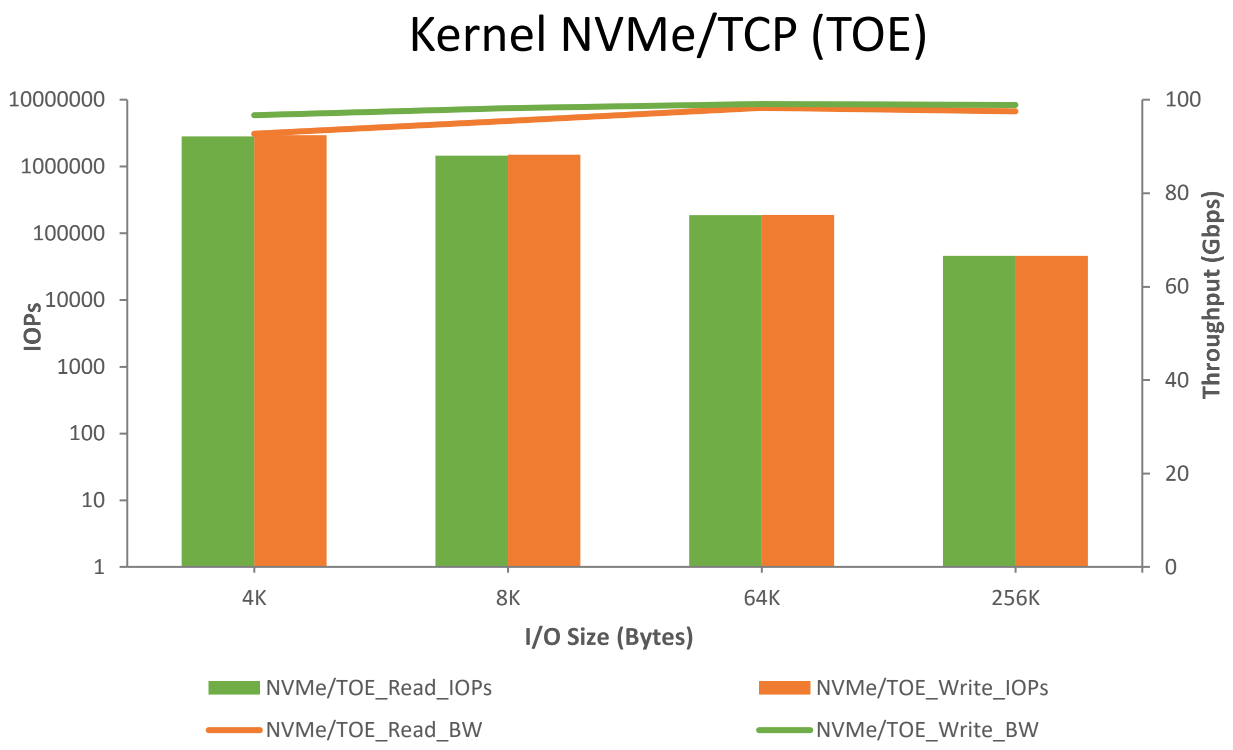


Summary

- Up to 50% CPU savings with Chelsio TOE compared to Host-Based TCP/IP

NVMe/TCP (TOE) – Target Bandwidth & IOPs

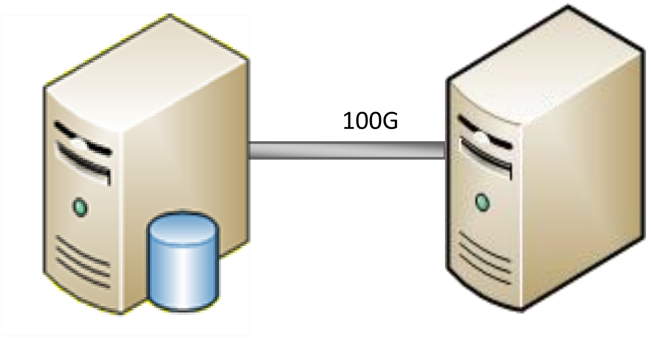




Summary – Kernel & SPDK based NVMe/TCP (TOE)

- Line Rate throughput of 99 Gbps
- 2.9 Million IOPs at 4K I/O size

NVMe/TCP (TOE) – Latency Test Configuration



- Host
  - 1 Intel Xeon CPU E5-1620 v4
    - 4 cores (HT disabled) @ 3.50GHz
  - 32GB of RAM
  - Chelsio T62100-CR (2 x 100Gbps)
  - RHEL 8.3 (5.4.143 kernel)
- Target
  - 2 Intel Xeon CPU E5-2687W v4
    - 24 cores (HT disabled) @ 3.00GHz
  - 128GB of RAM
  - 1 Micron 9100 MAX 2.4TB PCIe NVMe SSD
  - Chelsio T62100-CR (2 x 100Gbps)
  - RHEL 8.3 (5.4.143 kernel)

TOE Jitter Handling

	Average Latency (µsec)	Standard Deviation
NIC <-> NIC	4615	5240
TOE <-> TOE	4253	2160

- Reducing jitter is critical for reducing overhead
  - High jitter = high storage I/O delays
  - Reducing jitter allows for less retransmissions, dropped packets, etc.
- Latency measured in a traffic congested environment.
- TOE handles jitter exceptionally well
  - Standard Deviation 60% lower than NIC
  - Average Latency 8% lower than NIC

NVMe/TCP Latency Measurement Comparison

With & Without Offload / Kernel Space & SPDK

Target <-> host	Read			Write		
	Local	Remote	Delta	Local	Remote	Delta
Kernel TCP <-> Kernel TCP	109.15	130.61	21.46	24.43	44.65	20.22
Kernel TOE <-> Kernel TCP	109.15	126.18	17.03	24.43	42.67	18.24
Kernel TOE <-> Kernel TOE	109.15	124.95	15.8	24.43	40.84	16.41
SPDK NIC <-> SPDK NIC	105.31	126.87	21.57	20.08	39.9	19.1
SPDK TOE <-> SPDK NIC	105.31	114	8.69	20.08	29.65	8.85
SPDK iWARP <-> SPDK iWARP	105.31	110.88	5.57	20.08	27.4	6.6

Summary

- NVMe/TCP (TOE) latencies with T6 approach those of RDMA & are very close to those of local disk
- On Reads (slightly less on Writes):
  - Kernel based TOE gives ~25% latency improvements over Host-Based TCP/IP
  - SPDK based TOE gives ~60% latency improvements over Host-Based TCP/IP
- Small changes in performance in large scale networks with high frequency and/or large volumes of traffic have a big impact!

Testing NVMe/TCP (TOE)

No Compromise Testing

- Robust testing covering functional, conformance, interoperability and stress provides stable protocol offload for NVMe/TCP
  - Tools include fio, iozone, Dbench, SPDK fio plugin tools
  - Disks include RAMdisk, SSDs, NVMe disks
  - Adding UNH IOL test suite InterACT for conformance & plugfests for interoperability
- Performance
  - No compromise – delivering workload performance while maintaining interoperability
  - Performance measurement using fio providing
    - Bandwidth, IOPs, Latency, Jitter
    - Most importantly CPU Usage .. obtained with mpstat
- Chelsio's TOE tested in usual networking scenarios which include
  - Netperf, Iperf, Netpipe, Sockperf tools
  - Applications: nfs, scp, ssh, rsh, cifs, http
  - Network related: MTU, VLANs, IP Alias, Bonding, Nagle, Pause, congestion algos

Conclusions

- TOE is ideal for NVMe/TCP workloads & environments by
  - Allowing more host CPU cycles for application software stacks
  - Reducing host CPU costs
- It does this by
  - Reducing host CPU overhead to support growth and/or do more with less
  - Boosting server storage I/O performance (IOPs, response time, throughput)
  - Maintaining Interoperability & Plug-and-Play (no DCB required)
  - Leveraging high performance, low latency SSDs
  - Enabling remote storage performance similar to local storage

Q&A and General Discussion

Contact info

Greg Schulz: greg@unlimitedio.com

Bob Dugan: bobdugan@chelsio.com

STORAGE DEVELOPER CONFERENCE SDC 21

Please take a moment to rate this session.

STORAGE DEVELOPER CONFERENCE SDC 21