G2M
RESEARCH

G2M Research Multi-Vendor Webinar
Best Practices for Deploying NVMe™ Over Fabrics (NVMe-oF™)

November 12, 2019

WD Western Digital®

KIOXIA

Mellanox® TECHNOLOGIES

XILINX®

Excelero

Webinar Agenda

9:00-9:05	Ground Rules and Webinar Topic Introduction (G2M Research)
9:06-9:30	Sponsoring Vendor presentations on topic (5 minute each)
9:31-9:42	Key Question 1 (2-minute question; 2 minutes response per vendor)
9:43-9:44	Audience Survey 1 (2 minutes)
9:45-9:56	Key Question 2 (2-minute question; 2 minutes response per vendor)
9:57-9:58	Audience Survey 2 (2 minutes)
9:59-10:10	Key Question 3 (2-minute question; 2 minutes response per vendor)
10:11-10:18	Audience Q&A (8 minutes)
10:19-10:20	Wrap-Up

© 2019 G2M Communications, Inc. All rights reserved.

G2M
RESEARCH

G2M Research Introduction and Ground Rules

- Mike Heumann
  Managing Partner, G2M Research

Panelists

Rob Davis
VP, Storage
www.mellanox.com

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

Dave Montgomery
Director, Data Center Systems
(www.wdc.com)

Josh Goldenhar
VP Products
www.excelero.com

Joel Dedrick
VP/GM, Networked Storage SW
www.kioxia.com

Host/Emcee:
Mike Heumann
Managing Partner
www.g2minc.com

© 2019 G2M Communications, Inc. All rights reserved.

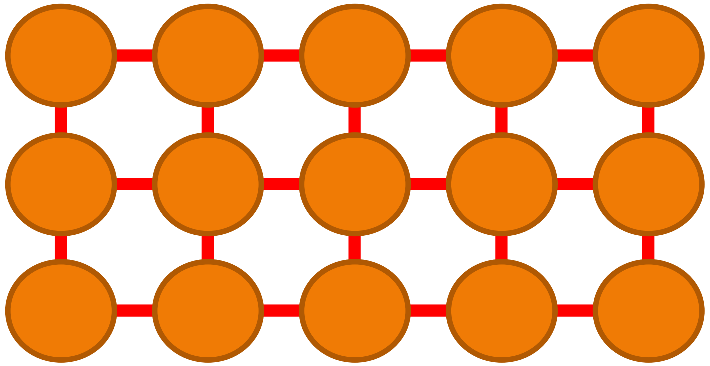
What is NVMe™ over Fabrics (NVMe-oF™)?

- NVMe-oF is a set of transports and protocols that extends NVMe storage access across a variety of datacenter networks
  - Ethernet: NVMe/TCP, NVMe-oRoCE, NVMe-oiWARP
  - Fibre Channel: NVMe-oFC -- InfiniBand: NVMe-oIB
- NVMe-oF provides near-local SSD performance by eliminating the overhead associated with networked SCSI protocols (iSCSI, FCP, etc.)

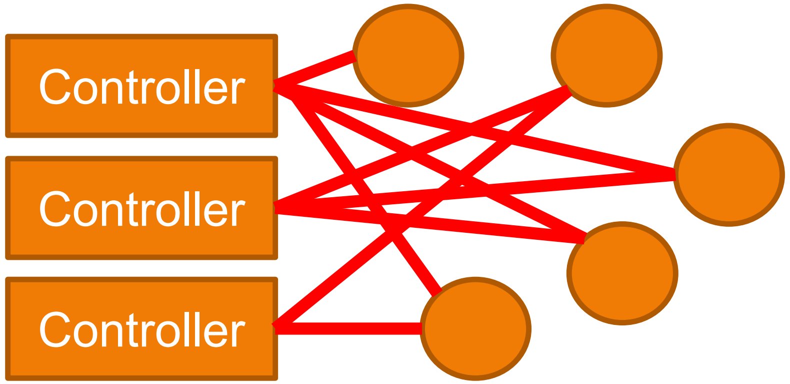
© 2019 G2M Communications, Inc. All rights reserved.

NVMe-oF Use Cases

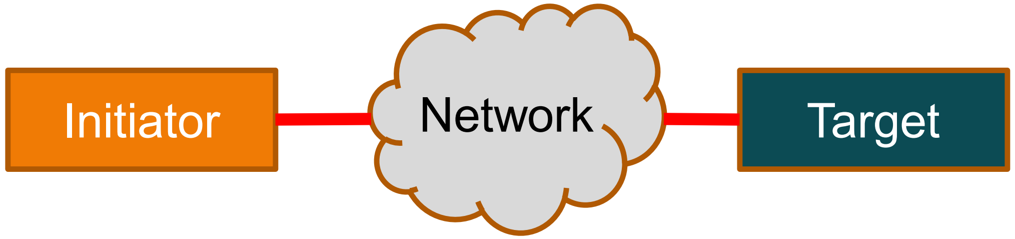
- Scale-Out Flash Storage (SOFS) Use Case
  - Connects servers and storage appliances with NVMe SSDs into a single namespace
  - Provides DAS-like storage performance, but with the ability to manage storage globally



- All-Flash Array (AFA) Back-End Use Case
  - NVMe-oF replaces SAS/SATA “tree” topologies behind network controllers
  - Provides significantly more flexibility (software-defined storage)



- Storage Initiator/Target Use Case
  - Classical connection of storage users (initiators) and storage devices (targets)
  - Provides significantly better performance than SCSI-based protocols



© 2019 G2M Communications, Inc. All rights reserved.

Mellanox

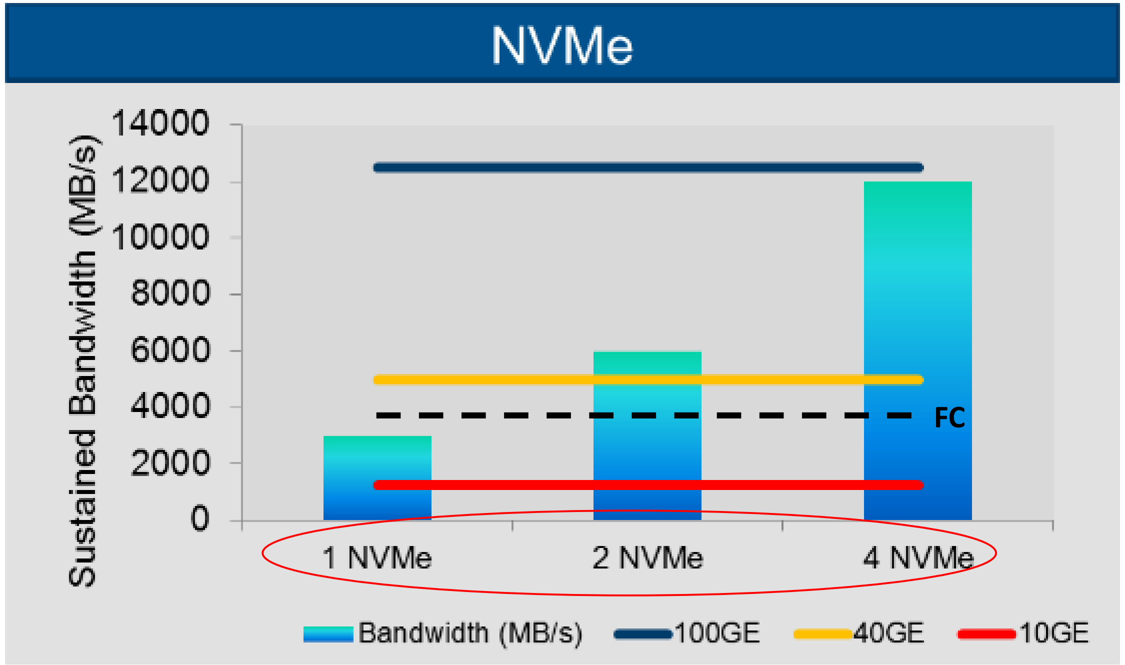
- Rob Davis

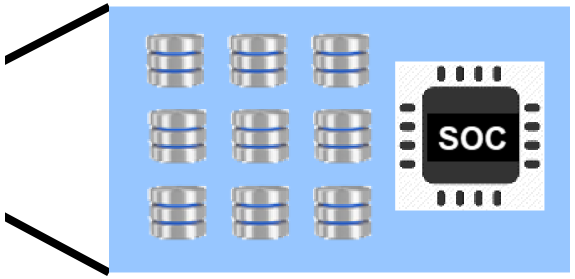
  Vice President,
  Storage Technologies

  www.mellanox.com

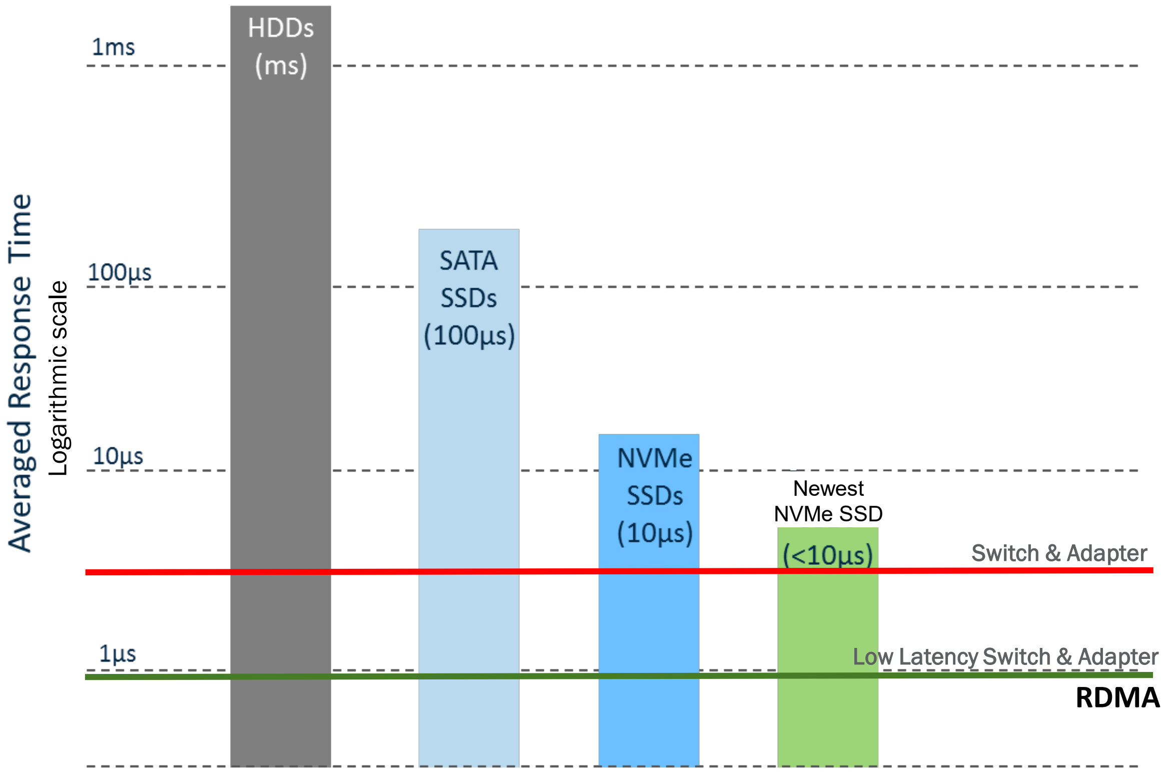
Best Practices Are Use Case and Application Dependent

- Scale-Out Flash Storage Use Case
  - Provides DAS-like storage performance
  - **Best Practice: 100 or 200GbE**
- Composable Infrastructure, Rack Scale, Compute Storage Disaggregation

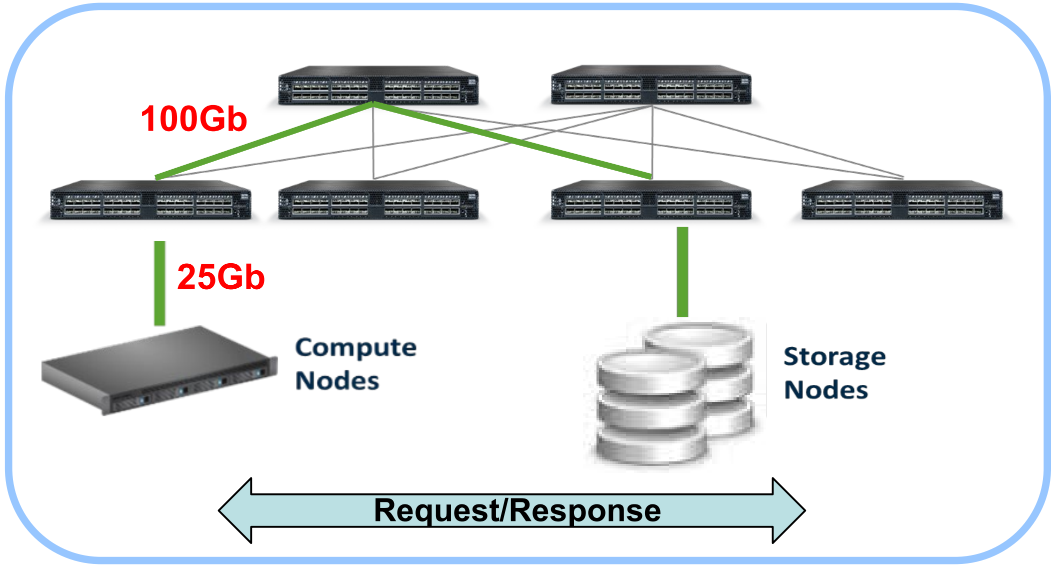




Importance of Network Latency when Comparing DAS to NVMe-oF



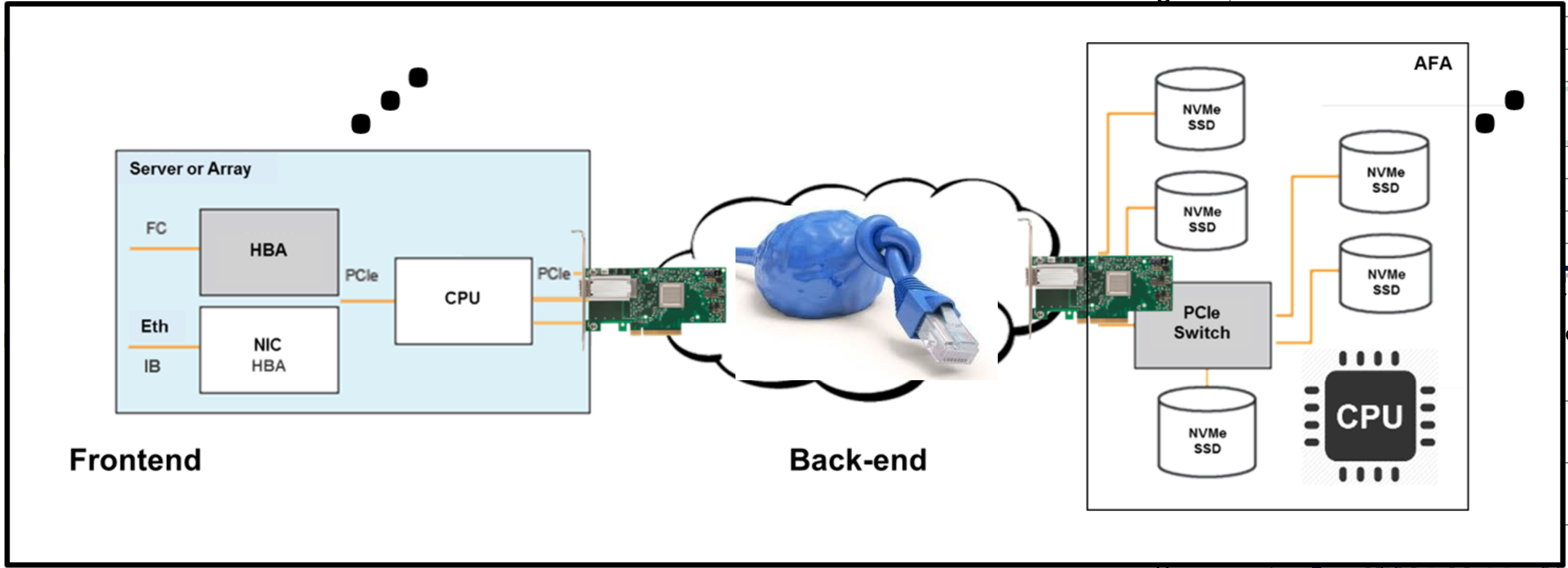
Network hops multiply latency

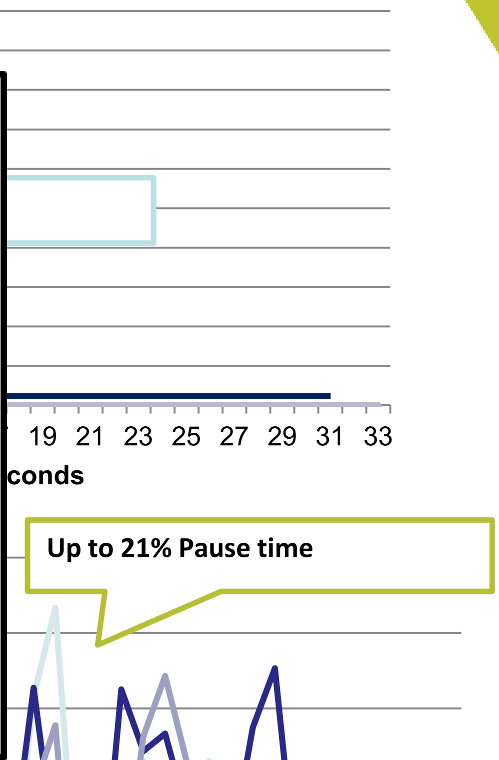


Best Practice: RDMA & Low Latency Switch+Adapter

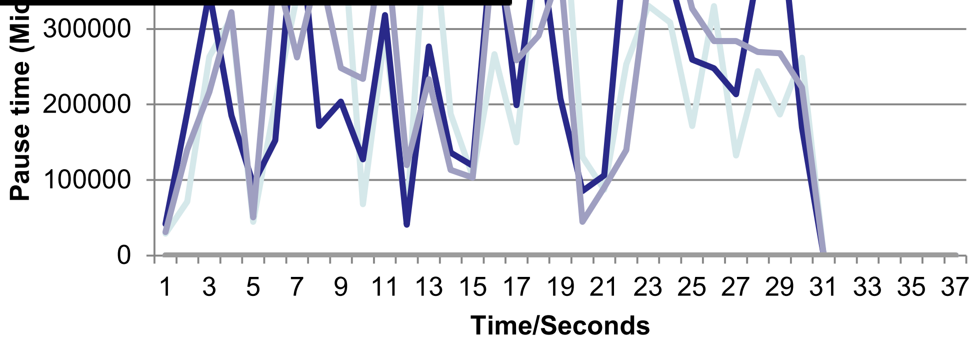
© 2019 Mellanox Technologies | Confidential

Importance of Congestion Control for All-Flash Array (AFA) Back-End Use Case



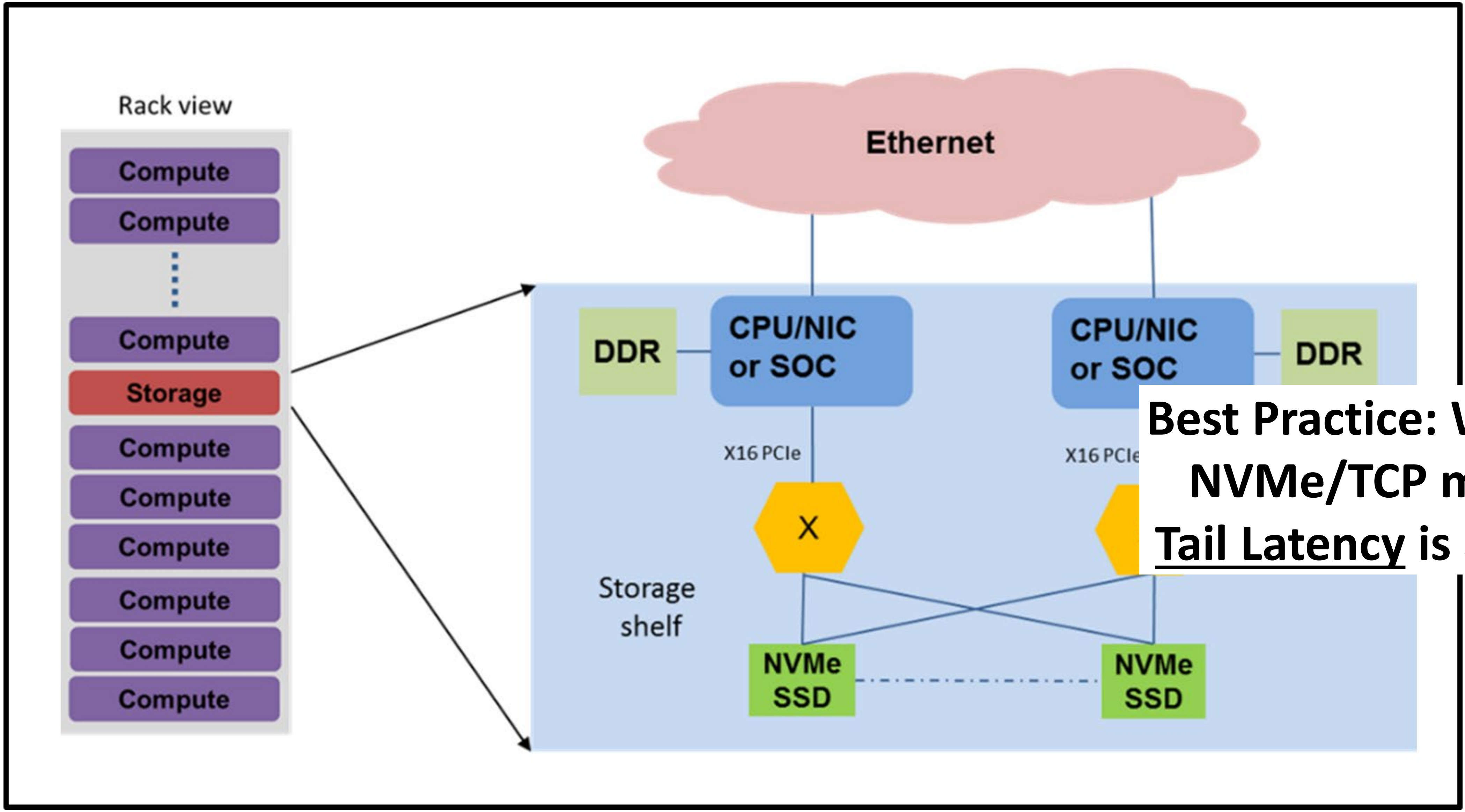


Up to 21% Pause time



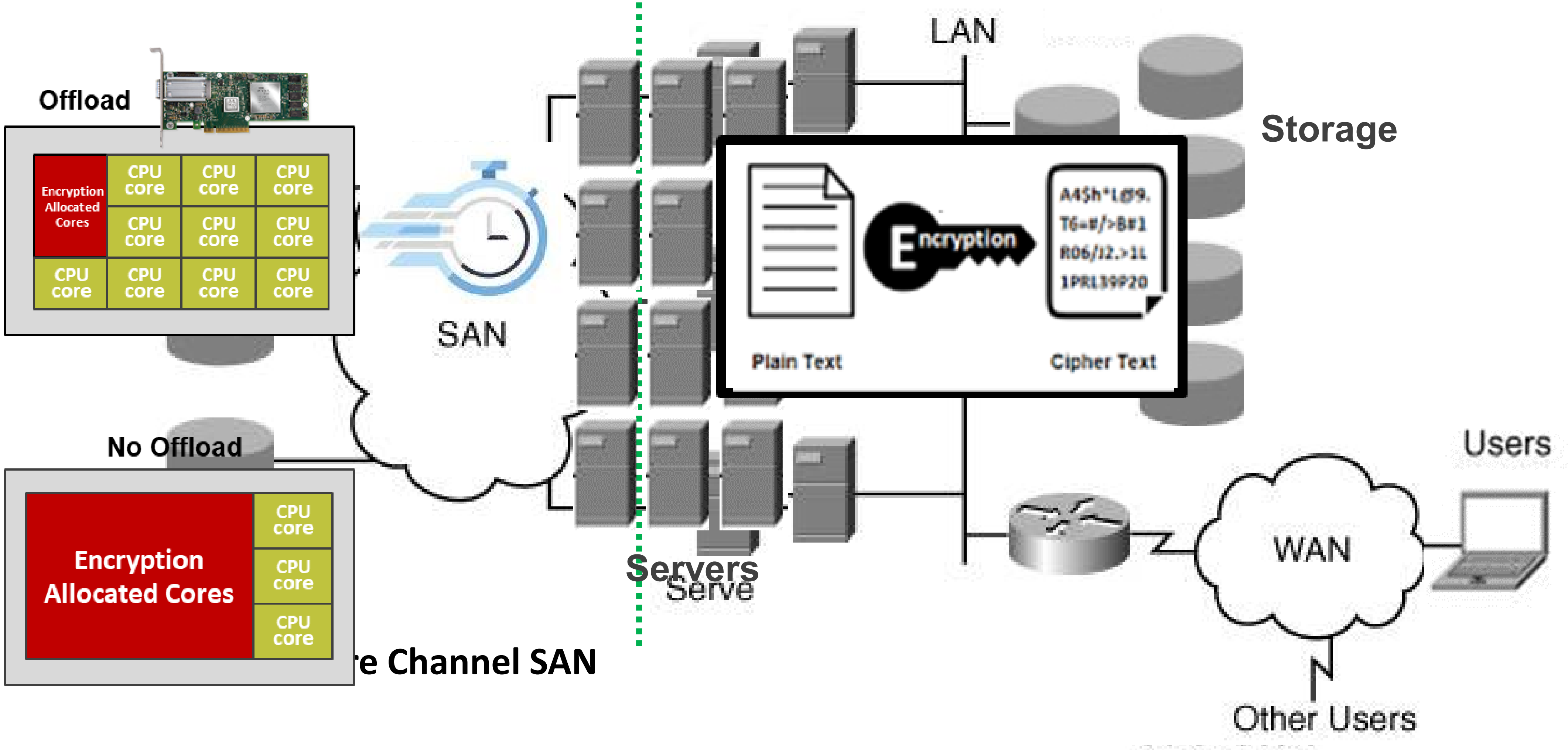
Best Practice: Switches & Adapters with Great Congestion Control

Classic Storage Initiator/Target Use Case



Best Practice: When using NVMe/TCP make sure <u>Tail Latency</u> is acceptable

NVMe-oF Security



Offload

Encryption Allocated Cores	CPU core	CPU core	CPU core
	CPU core	CPU core	CPU core
CPU core	CPU core	CPU core	CPU core

No Offload

Encryption Allocated Cores	CPU core
	CPU core
	CPU core

Best Practice: Use Encryption offloads to maintain NVMe-oF Performance

Xilinx

- Tom Spencer

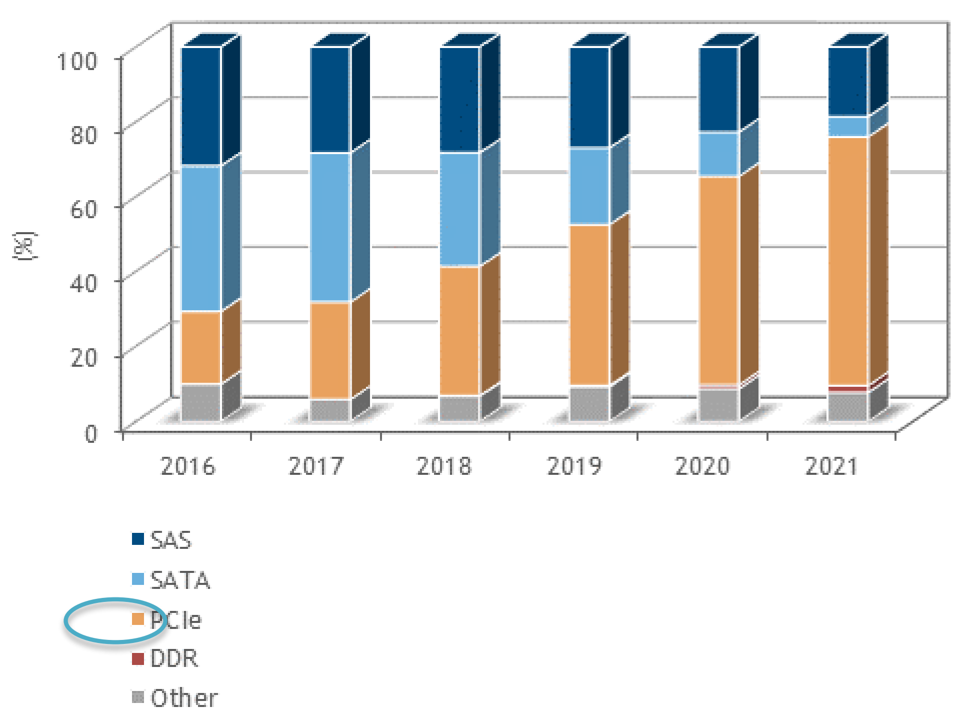
Sr. Director, Product Marketing

www.xilinx.com

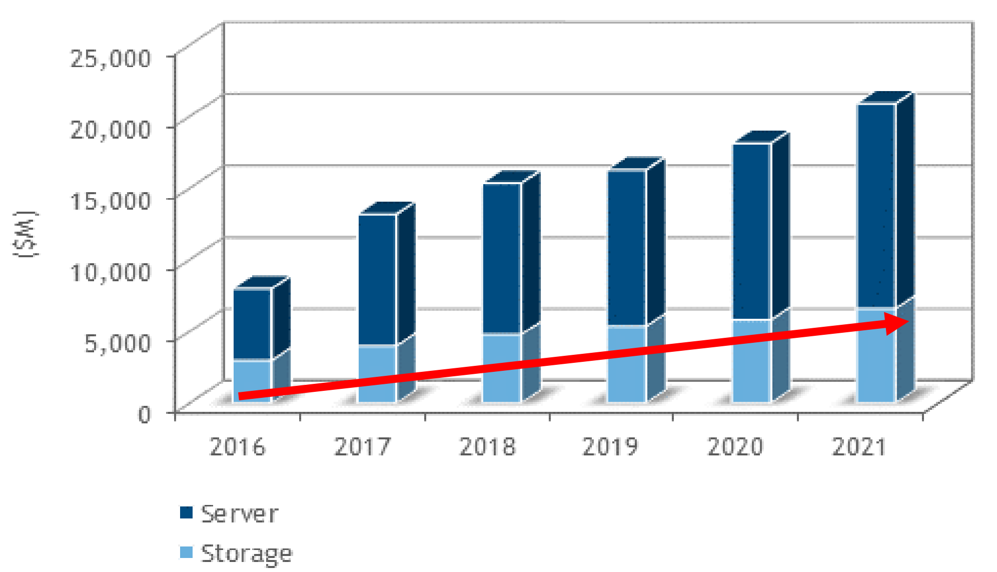
Industry Flash SSD Storage Market Dynamics

XILINX

IDC: WW Enterprise SSD Shipment share by Interface, 2016–2021



IDC: Worldwide Enterprise SSD Revenue by Location, 2016–2021



Flash In Storage Targets Growing

IDC: WW Enterprise SSD Shipments by Interface, 2017–2021 (000)

	2017	2018	2019	2020	2021	2016–2021 CAGR (%)
SAS	3,280	3,927	4,306	4,187	3,870	**7.9**
SATA	15,889	14,178	11,788	8,418	4,952	**-16**
PCIe	4,019	8,817	14,732	21,605	28,512	**69**

PCIe Accelerating

© Copyright 2019 Xilinx

Faster Storage Needs Faster Networking Infrastructure

XILINX

2018

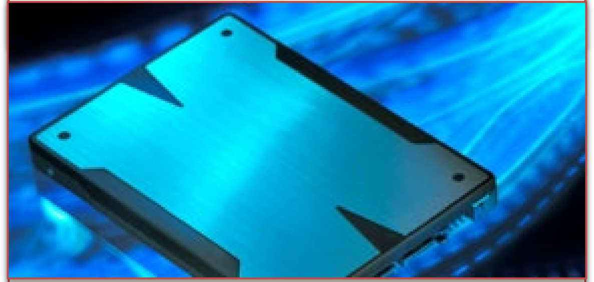
2020

SAS/SATA HDDs
PCIe Gen2

- 6G SAS/SATA
- 100s IOPS
- 100 MB/sec BW
- 10 ms Latency

1GbE

SAS/SATA SSDs
PCIe Gen3



- 12G SAS/SATA
- 300K IOPS
- 800 MB/sec BW
- 80 – 100 us Latency

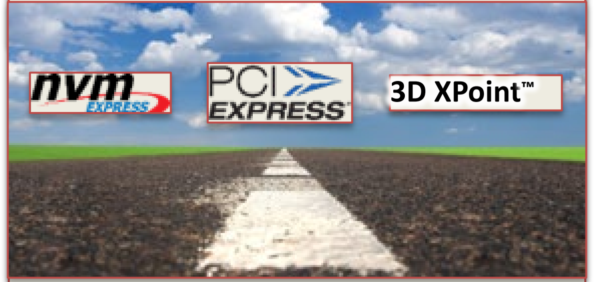
10GbE

NVMe SSDs
NVMe over Fabric
PCIe Gen3

- NVMe PCIe Gen3 x 4
- 1.5M IOPS
- 3,500 MB/sec
- 20 – 40 us Latency

10-25GbE

NVMe, 3D XPoint
NVMe over Fabric
PCIe Gen4



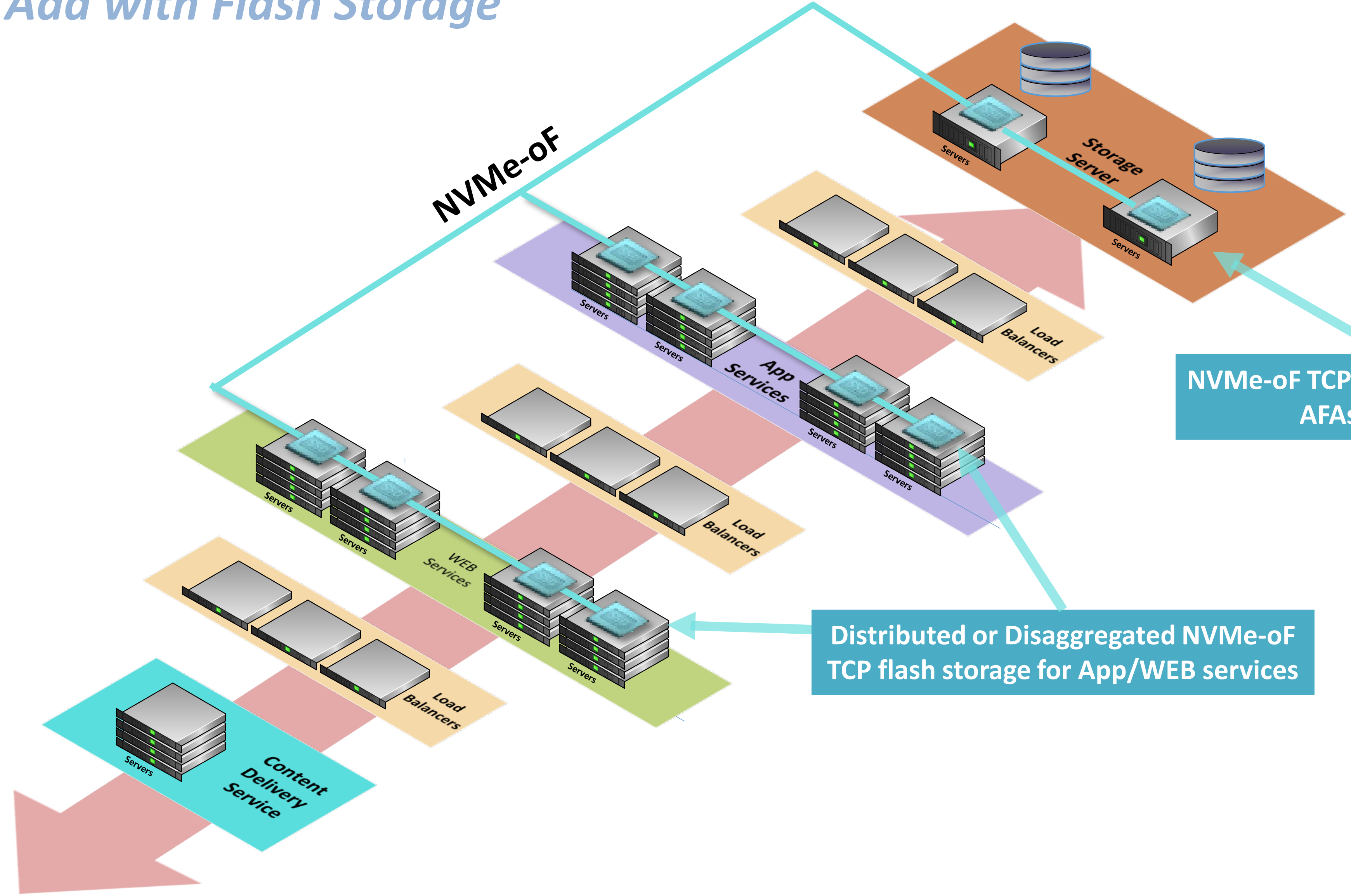
- NVMe PCIe Gen4 x 4
- 3M+ IOPS
- >5,000 MB/sec
- <10 us Latency

25-50-100GbE

© Copyright 2019 Xilinx

Modern Cloud Infrastructure

Xilinx Value Add with Flash Storage

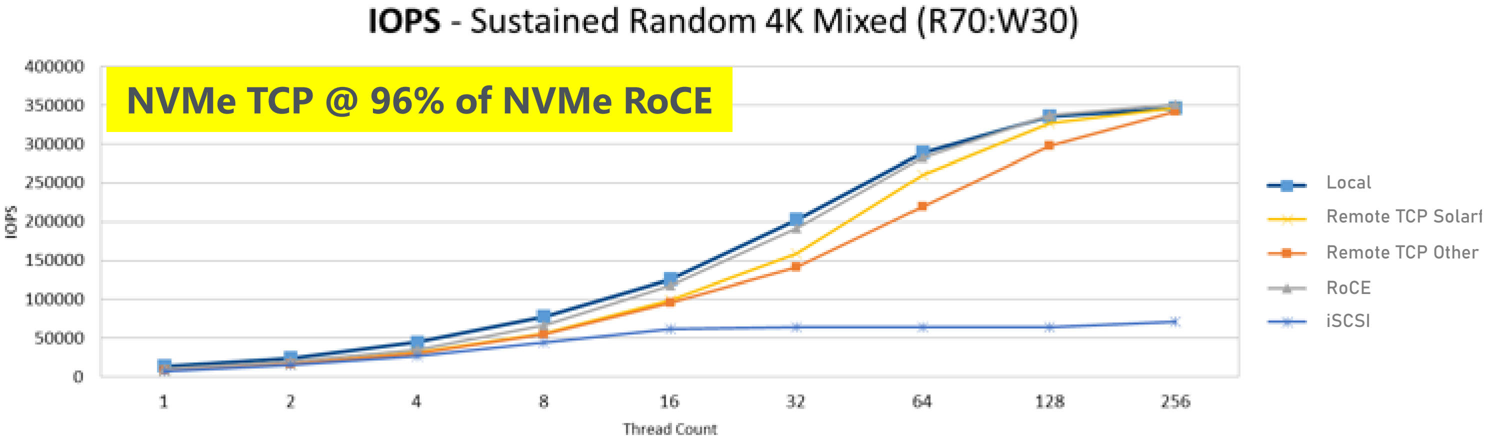


NVMe-oF TCP for backend storage AFAs and JBoFs

Distributed or Disaggregated NVMe-oF TCP flash storage for App/WEB services

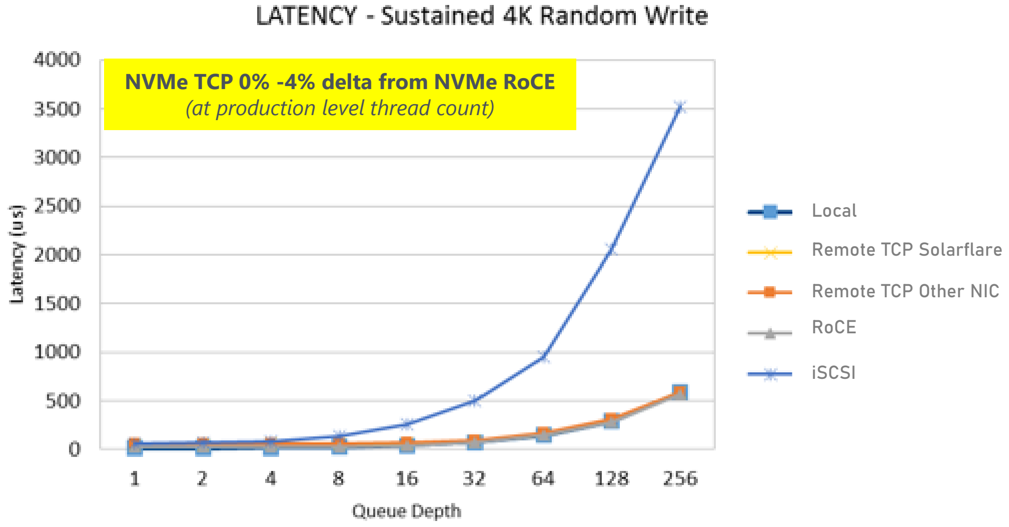
© Copyright 2019 Xilinx

Latest Kernel NVMe-oF TCP Benchmarking



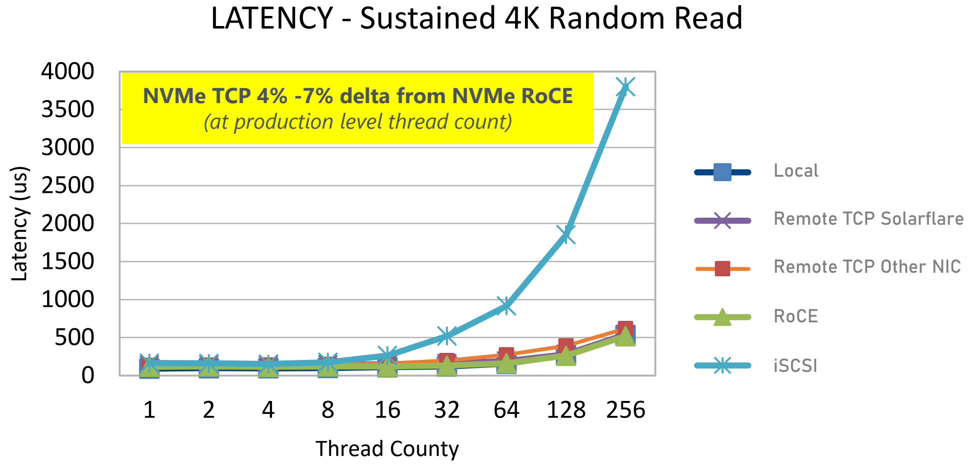
IOPS - Sustained Random 4K Mixed (R70:W30)

NVMe TCP @ 96% of NVMe RoCE



LATENCY - Sustained 4K Random Write

NVMe TCP 0% -4% delta from NVMe RoCE
(at production level thread count)



LATENCY - Sustained 4K Random Read

NVMe TCP 4% -7% delta from NVMe RoCE
(at production level thread count)

© Copyright 2019 Xilinx

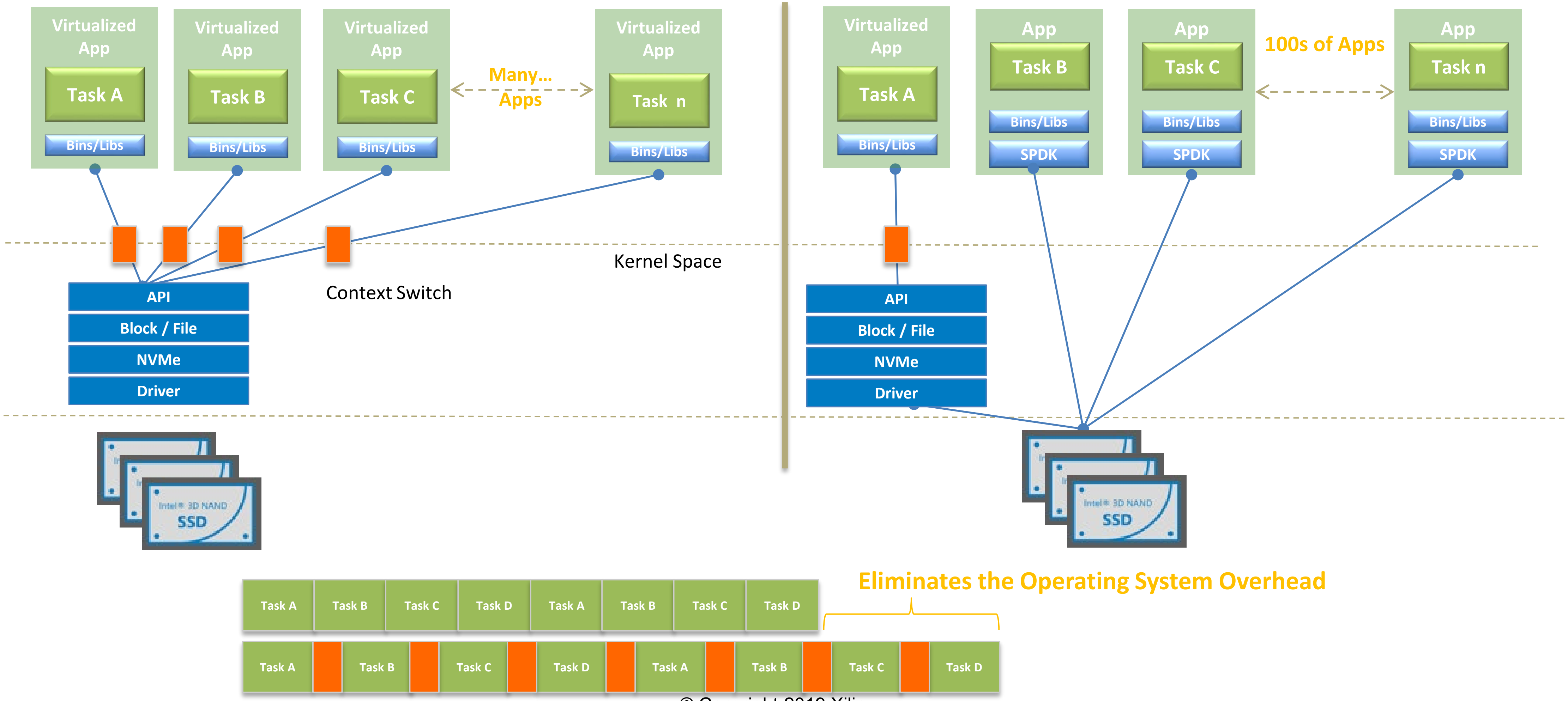
Inside The Server Bottleneck

The Problem: OS Context Switching

Common problem that occurs when many micro services are sharing storage devices is operating system context switching, buffer copying, the constant suspending and resuming of processes which kills application performance.

The Solution: User Space Block Storage (SPDK)

SPDK moves Block/NVMe layer in to user space eliminating context switching, buffer copying and blocking. Significantly removing overhead reducing latency and increasing scalability.

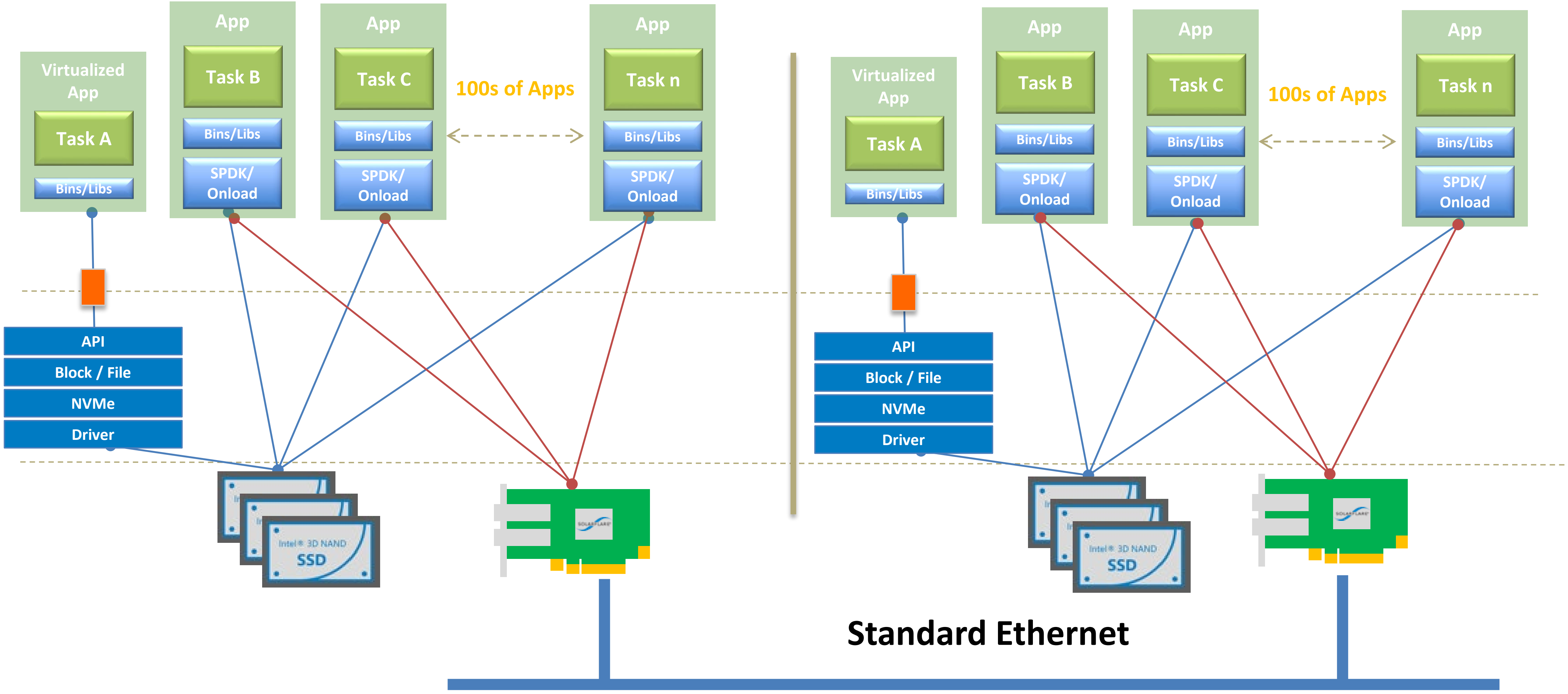


© Copyright 2019 Xilinx

The Networking Bottleneck – SPDK+Onload® Acceleration

The Solution: User space Block Interface (SPDK) and TCP/IP (Onload) Stack

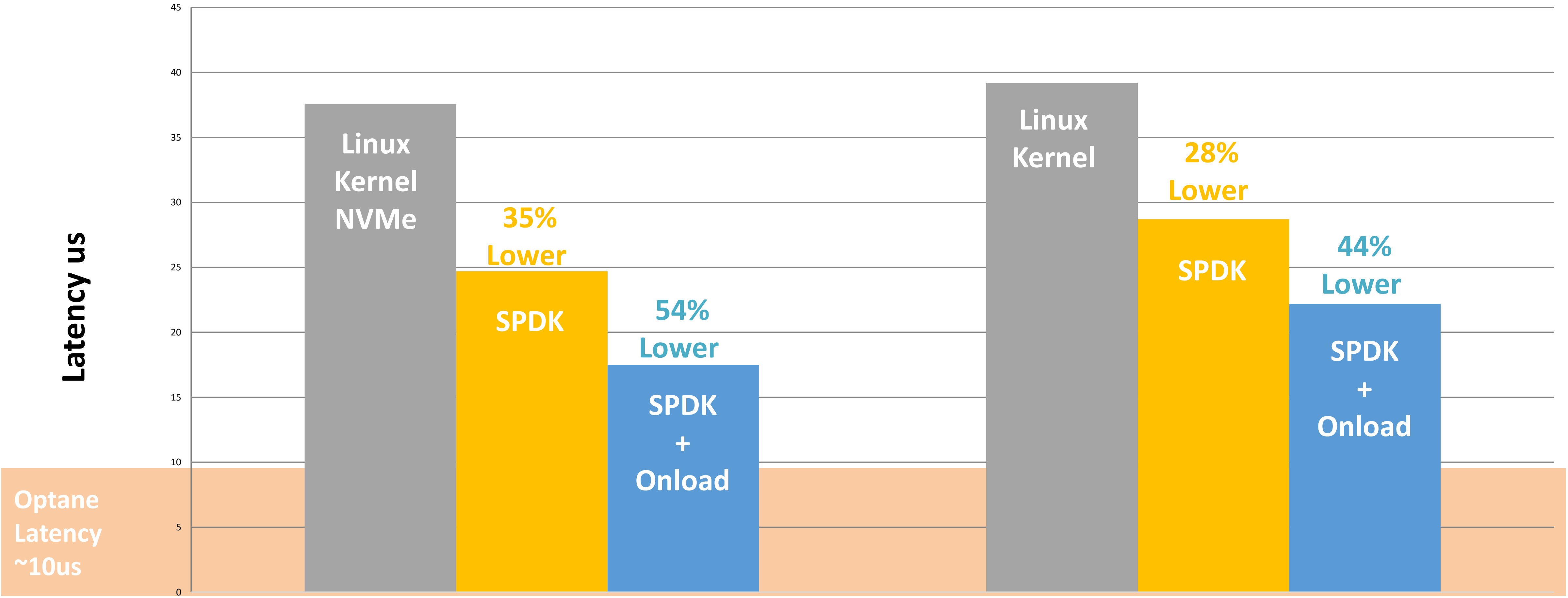
By bypassing Operating System for both NVMe and NCMe-oF TCP eliminates Bottleneck



© Copyright 2019 Xilinx

User Space Delivers Performance

Intel® Optane™ SSD and NVMe over TCP



© Copyright 2019 Xilinx

Summary

- TCP will likely serve the larger market for NVMe-oF transport in Cloud data centers
- TCP is ubiquitous
- TCP allows NVMe-oF to be deployed in legacy infrastructure
- Get up to 40% boost in lower latency and higher IOPS with User Level Networking (kernel bypass)
- Xilinx has a long term strategy for complete NVMe-oF solutions
- Xilinx already working with multiple eco-system partners. Let us know how we can work together!

NVMe-oF TCP: Collaborate with Xilinx TODAY!

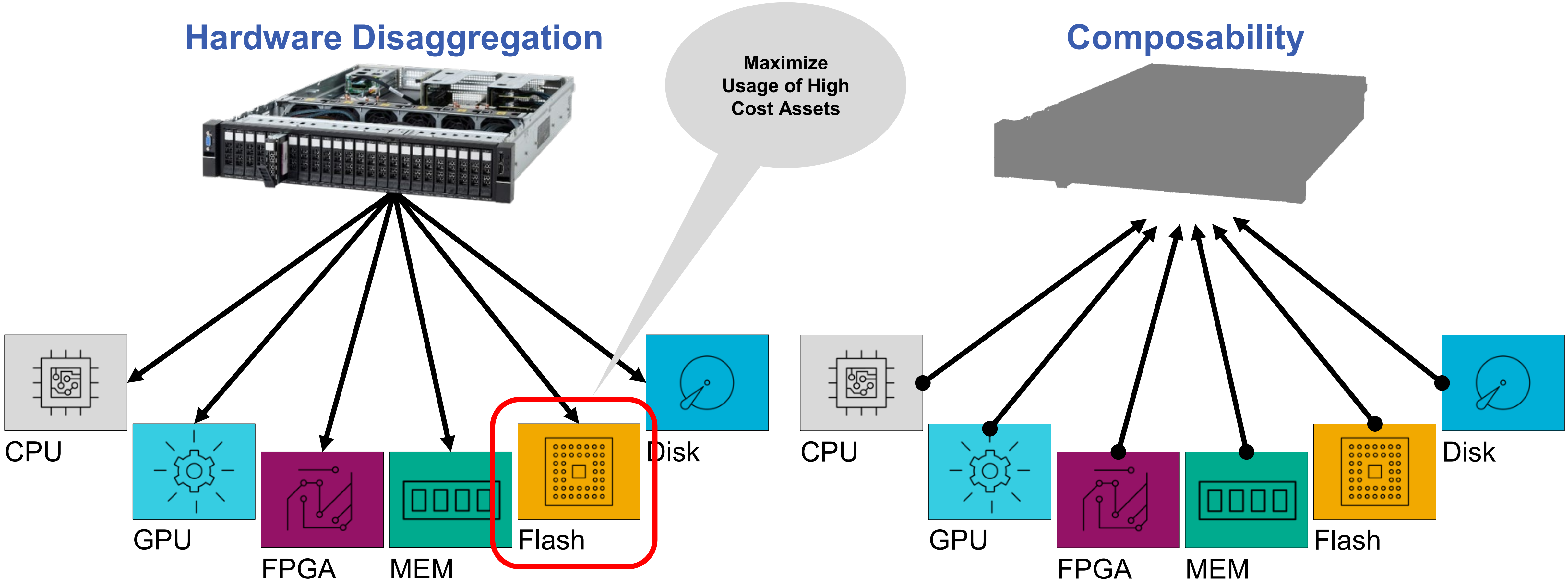
© Copyright 2019 Xilinx

Western Digital

- Dave Montgomery
  Director, Data Center Systems
  www.wdc.com

NVMe-oF™ - The On-Ramp to Composability

Provides the advantages of Hardware Composed Infrastructure with no vendor lock-in



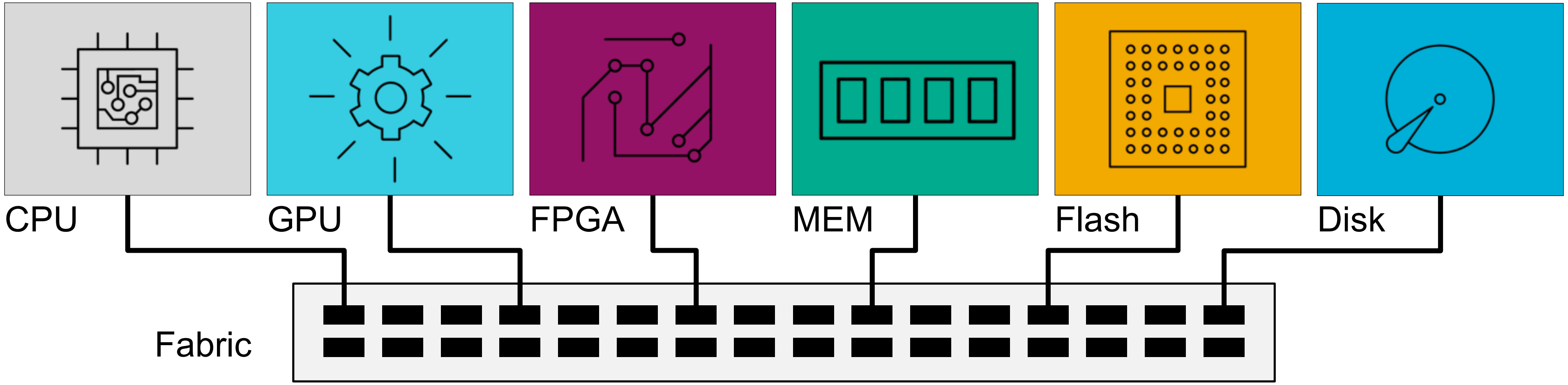
Disaggregate hardware components from the server so they can be efficiently pooled

Orchestrate virtual systems that can be optimally sized to the task

© 2019 G2M Communications, Inc.. All rights reserved.

Western Digital's Vision - Open Composability

Fabric Attached Devices

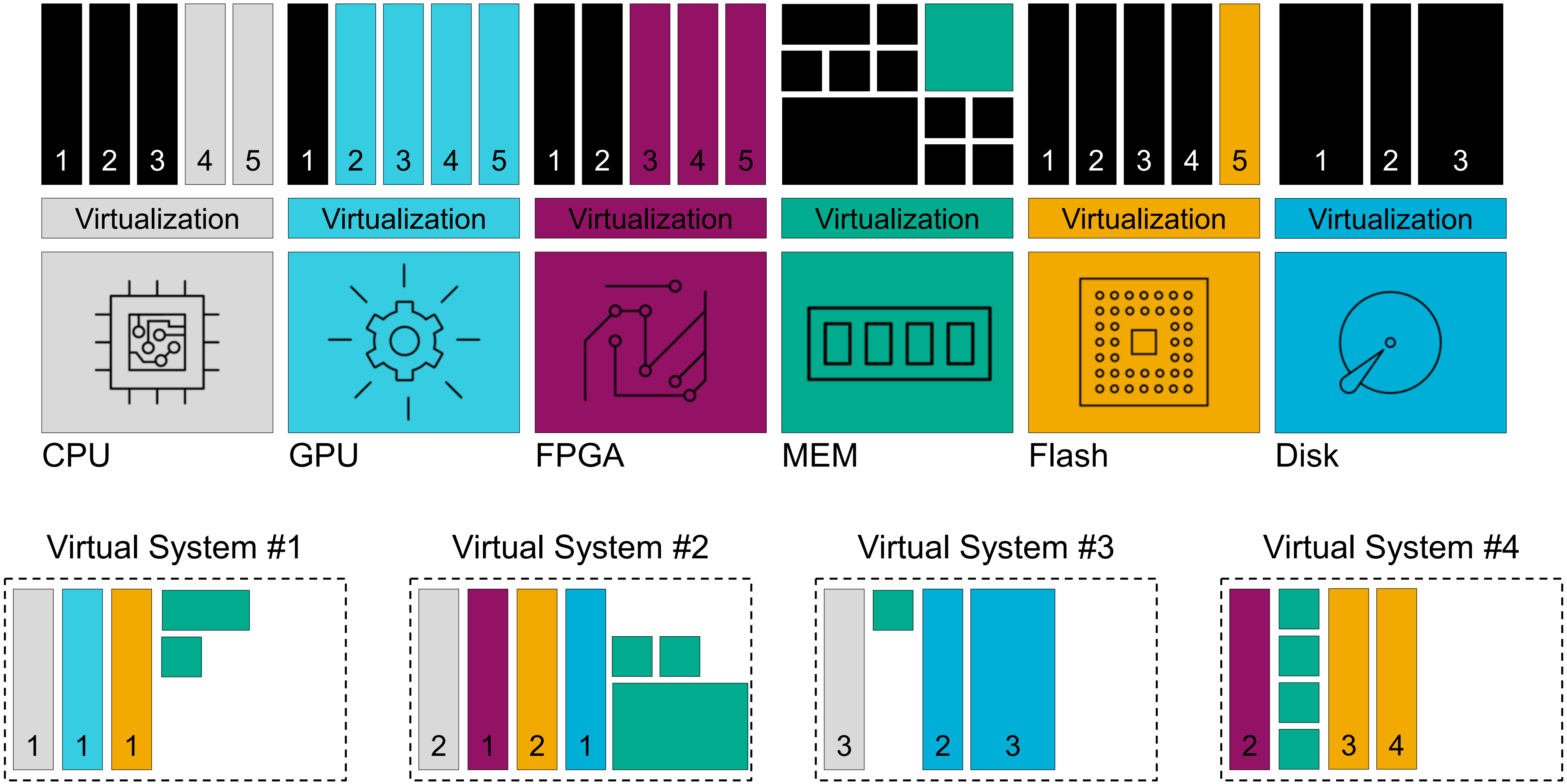


- No physical systems – Only virtual systems – Disaggregated hardware procured from separate suppliers
- Each device provides a resource that is offered over the fabric
- No established hierarchy – CPU doesn't 'own' the GPU or the Memory
- All devices are peers on the network & they communicate with each other

© 2019 G2M Communications, Inc.. All rights reserved.

Example – Composing Resources for Various Workloads

Orchestrate Virtual Systems Using Fabric Attached Devices



1 2 3 4 5 | 1 2 3 4 5 | 1 2 3 4 5 | | 1 2 3 4 5 | 1 2 3

Virtualization | Virtualization | Virtualization | Virtualization | Virtualization | Virtualization

CPU | GPU | FPGA | MEM | Flash | Disk

Virtual System #1: 1 1 1

Virtual System #2: 2 1 2 1

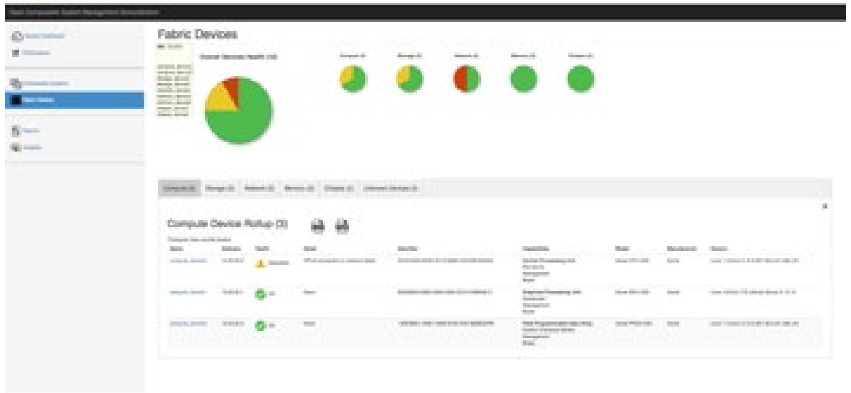
Virtual System #3: 3 2 3

Virtual System #4: 2 3 4

© 2019 G2M Communications, Inc.. All rights reserved.

Making Composable Infrastructure a Reality using NVMe-oF

- Product
  - F3100 NVMe-oF flash storage device
  - E3000 enclosure
- Architecture
  - OpenFlex^{TM} HW architecture standardized in SNIA SFF
    - SFF-TA-1013, SFF-TA-1014, SFF-TA-1015
  - Open Composable API contributed to opencompute.org Storage Project
  - REST based commands to discover and compose virtual systems



© 2019 G2M Communications, Inc.. All rights reserved.

NVMe-oF™ Fabric Devices

OpenFlex™ F3100 Fabric Device and E3000 Enclosure

Dual-port, high-performance, low-latency fabric-attached SSD

Self-virtualized device with up to 256 namespaces for dynamic provisioning

3U enclosure with 10 dual-port slots offering up to 614 TB

Multiple storage tiers over the same wire – Flash and Disk accessed via NVMe-oF

NVMe™-over-Fabrics | Infrastructure Disaggregation | Software Composable

© 2019 G2M Communications, Inc.. All rights reserved.

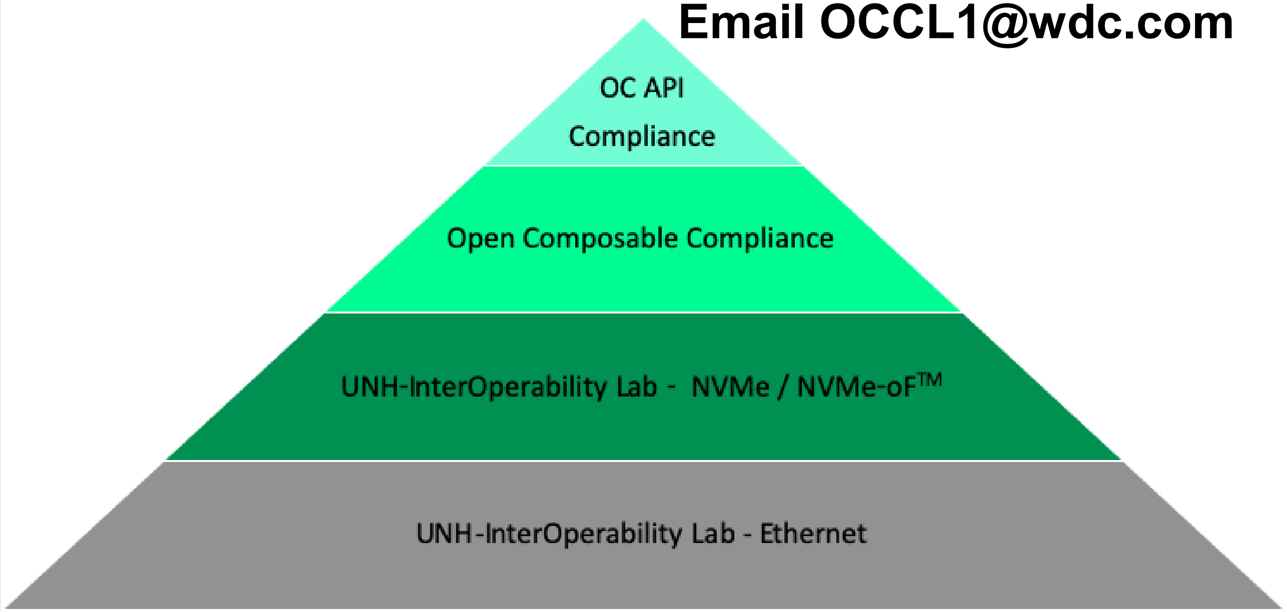
Getting Started - Open Composable Compliance Lab

Objectives

- Composable Infrastructure is in the early adoption stage – multivendor interoperability will accelerate adoption
- Testing also available via the UNH IOL lab
- Get started sharing flash storage with Western Digital now
- OCCL is a center of excellence for NVMe-oF^{TM}

TM

Email OCCL1@wdc.com



OC API Compliance

Open Composable Compliance

UNH-InterOperability Lab - NVMe / NVMe-oF^{TM}

UNH-InterOperability Lab - Ethernet

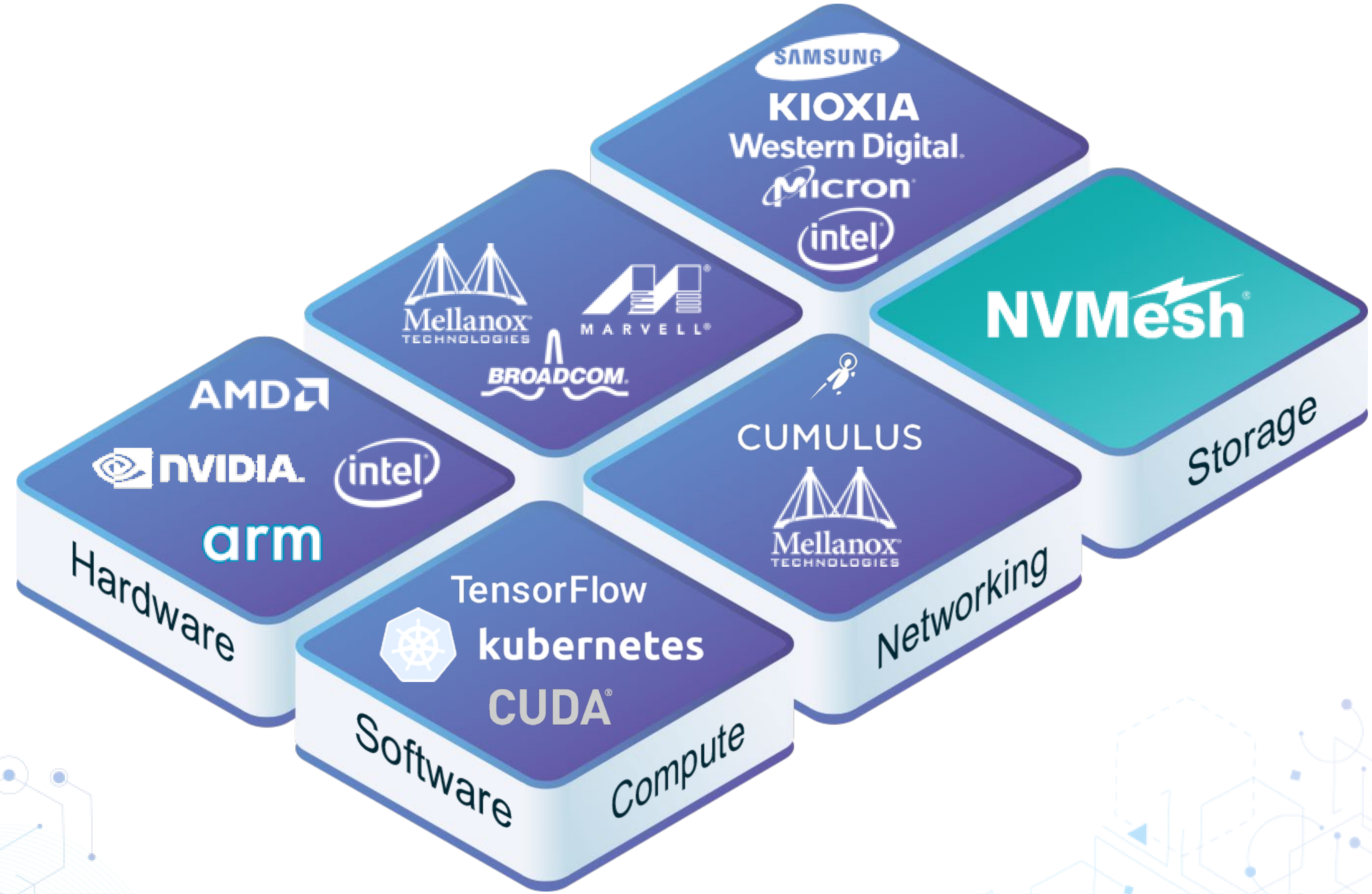
University of New Hampshire InterOperability Laboratory

© 2019 G2M Communications, Inc.. All rights reserved.

Excelero

- Josh Goldenhar
  VP, Products
  www.excelero.com

Technologies enabling the high-performance data center



NVMe Flash is the new storage standard

- **Unprecedented performance:**
  - Up to 2.5M IOPs, 9+GB/s per drive
  - Ultra-low latency (8-20μs)
- **Game changer for data-intensive workloads:**
  - Mission-Critical Databases
  - Analytical Processing
  - AI and Machine Learning

Copyright Excelero 2019

NVMe delivers phenomenal performance, but…

IOPs and Bandwidth Utilization

- Applications struggle to utilize local NVMe performance beyond 3-4 drives
- Stranded IOPS and/or bandwidth = poor ROI

Sharing is the Logical Answer, with local latency

- Physical disaggregation is often operationally desirable
- 24 Drive servers are common and readily available

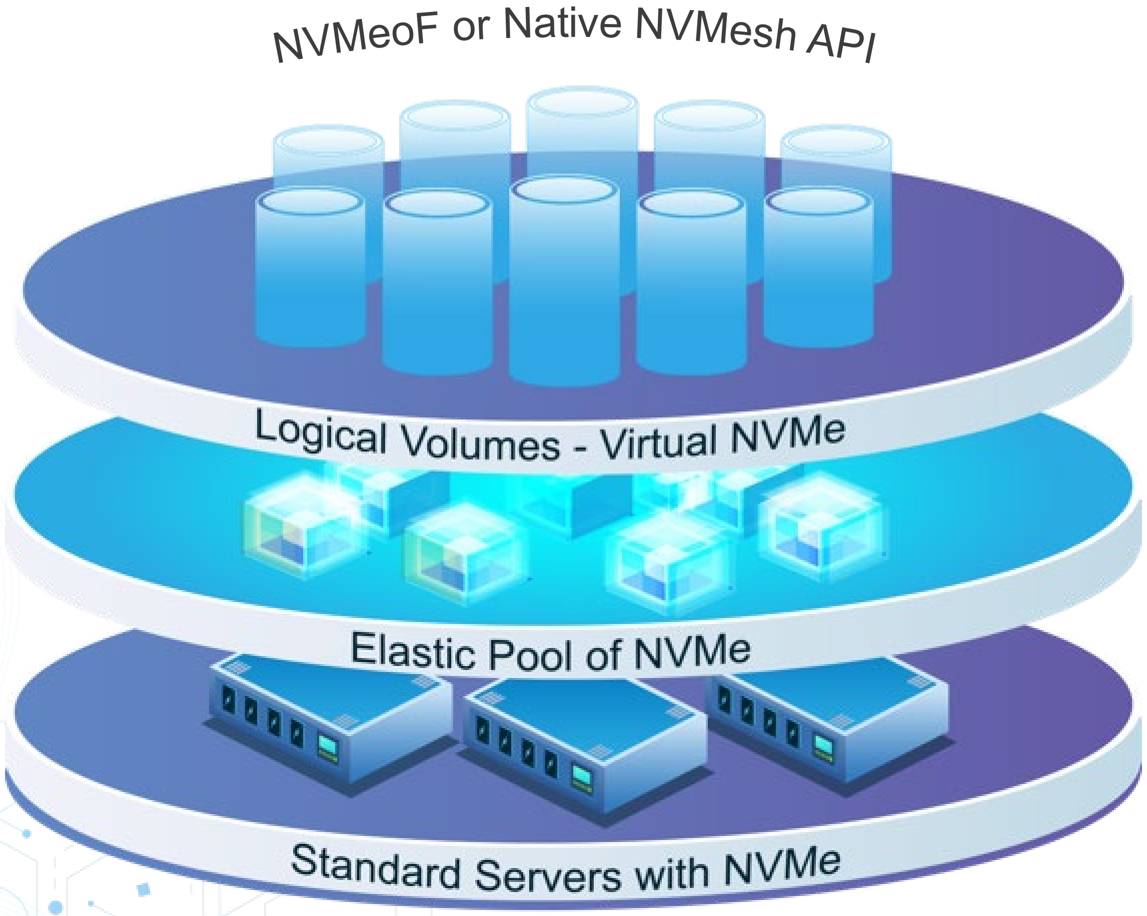
Data Protection Desired

- NVMe performs, but by itself offers no data protection
- Local data protection does not protect against server failures

Some NVMe-over-fabrics solutions offer controller based data protection, but limit IOPs, bandwidth and sacrifice latency

Copyright Excelero 2019

Elastic NVMe: access data anywhere at local speeds


NVMeoF or Native NVMesh API
Logical Volumes - Virtual NVMe
Elastic Pool of NVMe
Standard Servers with NVMe

NVMesh virtualizes NVMe flash

Deploy NVMe at data center scale

Feed GPU’s with local speeds & enable GPU virtualization

Maximize GPU and NVMe utility and ROI

Copyright Excelero 2019

Kioxia

KIOXIA

- Joel Dedrick
  VP/GM, Networked Storage Software
  www.kioxia.com

KUMOSCALE™

Networked NVMe® Flash at SCALE

HOW TO DO IT WRONG

© 2019 KIOXIA America, Inc. All Rights Reserved.

KIOXIA

What is KumoScale?

KUMOSCALE™

A Software product

- Implements a fast, networked block storage service
- Disaggregation based on NVMe-oF™ (NVM Express® over Fabrics) standard

Target: mid/large-scale, “on-prem clouds”

- Service providers
- SaaS services delivered via smartphone
- Marketplaces, clearinghouses (travel, tickets, stock trades)
- Massively multi-player gaming

Cloud-focused architecture

- Integrates with (not replaces) management infrastructure
- Focused on speed, very low cost

“kumo”		
雲	くも	**Cloud**
蜘蛛	くも	**Spider**

KumoScale storage software enables NVMe Flash as a service

KIOXIA

NVM Express and NVMe are registered trademarks, and NVMe-oF is a trademark of NVM Express, Inc.

Scale

KUMOSCALE™

“At Scale”

- 50,000 – 500,000 nodes
- Multiple data centers, multiple zones

Scale brings unique challenges

- Automation is mandatory; “human in the loop” doesn’t work
- Things break
- Neighbors are not just noisy; also unpredictable, and possibly malicious

NVMe-oF™ at Scale -- How to do it Wrong (1): *Neglect sophisticated flash media management*

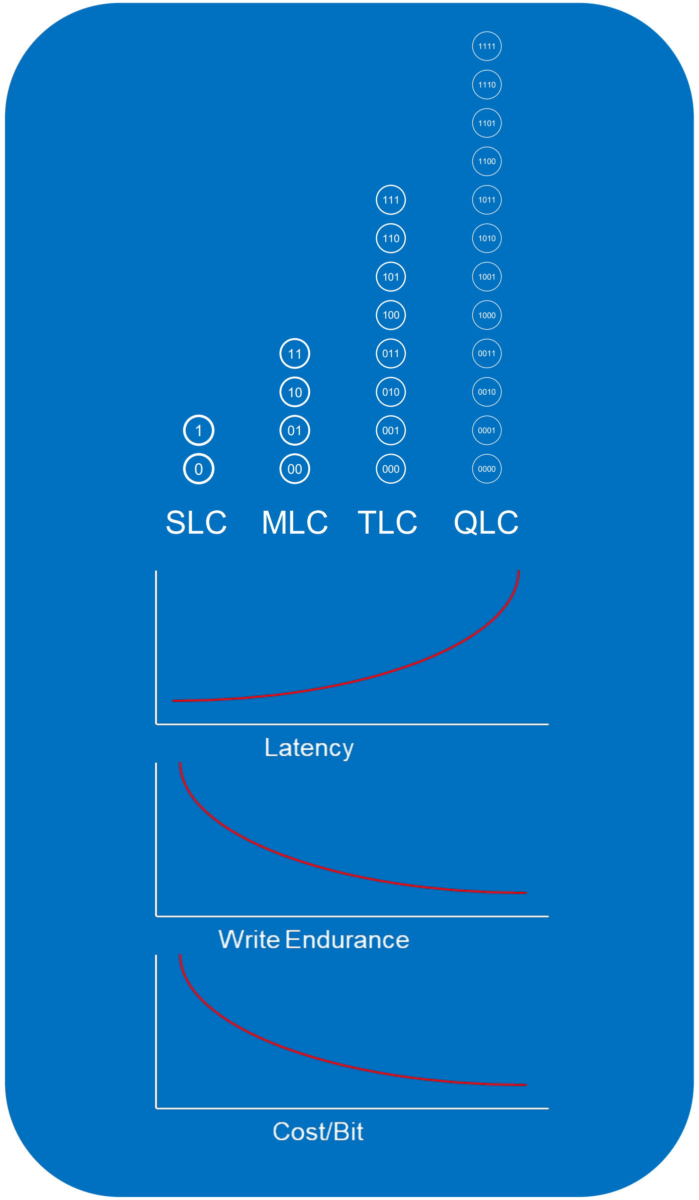
Flash will soon come in a variety of performance/cost options

- “Low Latency” (many-plane, SLC) NAND
- TLC, of course.
- QLC, even PLC (5 bits per cell)

Enormous range of performance, endurance, cost

- ➢ Putting the wrong application on QLC could wear it out in months.
- ➢ Putting the wrong applications on high performance flash could triple your storage cost

Automatic, closed-loop workload: media matching will be mandatory



NVM Express and NVMe are registered trademarks, and NVMe-oF is a trademark of NVM Express, Inc.

NVMe-oF™ at Scale -- How to do it Wrong (2): *Neglect Zero-Touch Deployment*

KUMOSCALE™

When you install 1000 storage nodes, each one must:

- Boot/install/upgrade over the network with no unique per-node configuration
- Participate in internetworking protocols, learn/advertise its place in the topology
- Integrate seamlessly with *existing* logging, telemetry, provisioning, orchestration (whatever they are)

KIOXIA

NVM Express and NVMe are registered trademarks, and NVMe-oF is a trademark of NVM Express, Inc.

NVMe-oF™ at Scale -- How to do it Wrong (3): *Neglect Topology Awareness*

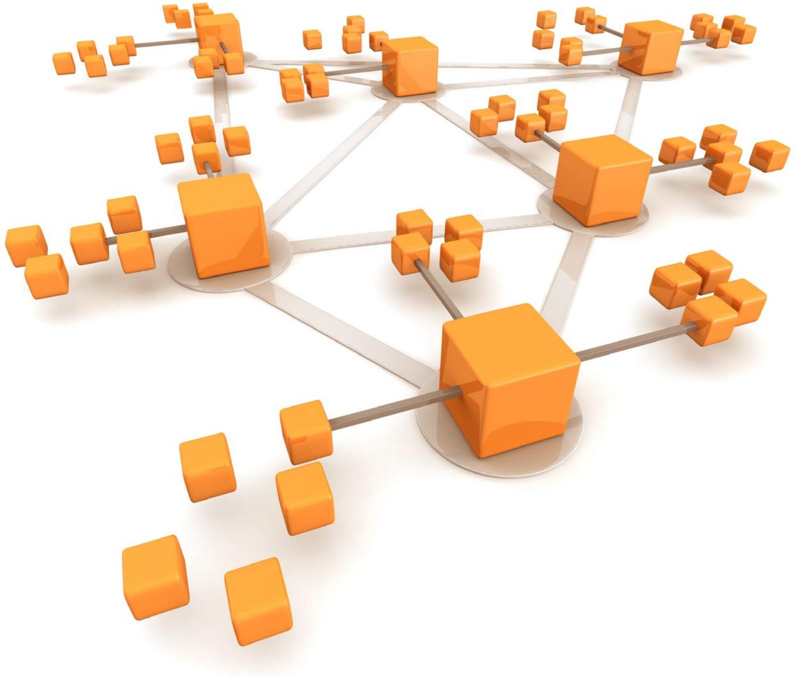
KUMOSCALE™

At scale, things break.

- Protecting against SSD failure is just table stakes.
- How do you handle a rack partition? A row of racks? An AZ?

Awareness of failure zone topology is a must

- How do you ensure that all of your redundancy isn't in the same failure domain?
- How do you prevent overreacting to planned/maintenance network partitions?



KIOXIA

NVM Express and NVMe are registered trademarks, and NVMe-oF is a trademark of NVM Express, Inc.

NVMe-oF™ at Scale -- How to do it Wrong (4):
Fail to Disaggregate

KUMOSCALE™

SSD Innovator's Dilemma:

- "Sweet spot" capacity is too large
- NVMe performance is too high
  - ➢ Very few client nodes can make good use of a 16TB, million IOPS drive.

Obviously, we need to share them. But how?

- "Direct attached" (SOFS) has hidden risks:
  - Ingress storms happen. Client NIC is a bottleneck you can't plan for.
  - When I/O streams are blended, regression to the mean works *for* you
- Can't pay as you grow. Have to buy all your flash up front.
- Law of Direct-Attached Flash: *No matter what size drive you chose, it's wrong.*

NVM Express and NVMe are registered trademarks, and NVMe-oF is a trademark of NVM Express, Inc.

NVMe-oF™ at Scale -- How to do it Wrong (5): *Underestimate the Power of Orchestration*

KUMOSCALE™

Orchestration framework "war" was over before it started.

- Solutions lacking seamless, deep integration with Kubernetes will not age well…

Static workload assignment, storage provisioning via scriptware is a "dead-man walking."

- Economic power of workload blending too large to ignore

NVM Express and NVMe are registered trademarks, and NVMe-oF is a trademark of NVM Express, Inc.

NVMe-oF™ at Scale -- How to do it Right

KUMOSCALE™

- Choose a disaggregated solution designed from the ground up for large scale deployment.
- Choose a vendor who understands flash media profoundly and has stamina
- “Trust, but Verify”

NVM Express and NVMe are registered trademarks, and NVMe-oF is a trademark of NVM Express, Inc.

► Panel Discussion

Panel Question #1

- When it comes to planning NVMe-oF deployments, what are the top three factors that companies should consider?
  - Mellanox
  - Xilinx
  - Western Digital
  - Excelero
  - Kioxia

© 2019 G2M Communications, Inc. All rights reserved.

Audience Survey Question #1

- What experience does your organization have with NVMe-oF? (check all that apply):
  - Explored information on its use (conferences, articles, etc.): %
  - Talked to NVMe-oF vendors (NW adapters, storage, software, etc.): %
  - Defined potential NVMe-oF projects: %
  - Started one or more proof-of-concept evaluations: %
  - Budgeted for actual production NVMe-oF deployments: %
  - Deployed NVMe-oF in production: %

© 2019 G2M Communications, Inc. All rights reserved.

Panel Question #2

- NVMe-oF has significantly better performance than SCSI-based storage networking protocols. When “retrofitting” existing storage networks with NVMe-oF, what are some of the bottlenecks that might be encountered that could impact these performance gains?
  - Western Digital
  - Excelero
  - Kioxia
  - Mellanox
  - Xilinx

© 2019 G2M Communications, Inc. All rights reserved.

Audience Survey Question #2

- Which classes of NVMe-oF use cases have your organization evaluated? (check all that apply):
  - Scale-Out Flash Storage deployments (servers and/or storage appliances with local storage in a single common namespace): %
  - Deployment of all-flash arrays with NVMe-oF back-ends: %
  - Deploying NVMe-oF into existing or new networked storage configurations: %
  - Other use cases (converged infrastructure, etc.) %

© 2019 G2M Communications, Inc. All rights reserved.

Panel Question #3

- 2019 marked the year that every large storage array vendor supported NVMe-oF across their product line. What other "enablers" are needed for NVMe-oF to be widely utilized in production storage networks?
  - Kioxia
  - Mellanox
  - Xilinx
  - Western Digital
  - Excelero

© 2019 G2M Communications, Inc. All rights reserved.

Audience Q&A

© 2019 G2M Communications, Inc. All rights reserved.

Thank You For Attending

G2M

RESEARCH