G2M RESEARCH

Storage Architectures that Maximize the Performance of HPC Clusters

Multi-Vendor Webinar

Tuesday Feb 23, 2021

Webinar Agenda

9:00-9:05	Ground Rules and Webinar Topic Introduction (G2M Research)
9:06-9:29	Sponsoring Vendor presentations on topic (8 minute each)
9:30-9:36	Key Question 1 (1-minute question; 2 minutes response per vendor)
9:37-9:37	Audience Survey 1 (1 minute)
9:38-9:44	Key Question 2 (1-minute question; 2 minutes response per vendor)
9:45-9:45	Audience Survey 2 (1 minutes)
9:46-9:52	Key Question 3 (1-minute question; 2 minutes response per vendor)
9:53-9:59	Audience Q&A (7 minutes)
9:59-10:00	Wrap-Up

G2M Research Introduction and Ground Rules

Mike Heumann
(Managing Partner, G2M Research)

High-Performance Computing – Expanding Beyond Research and Defense/Aerospace

- HPC was originally utilized by government agencies and research universities
  - Particle Physics
  - Fluid Dynamics
  - Nuclear Weapons Research
  - Aerospace Modeling
- The HPC footprint is now growing into commercial enterprises
  - Product Modeling
  - Business Intelligence
  - Artificial Intelligence/Machine Learning

Differences Between Enterprise and HPC Storage Architectures

- Conventional datacenter storage architectures built to simplify IT imperatives
  - Flexibility/Scalability ➔ Virtualized, migratable storage
  - Data Integrity ➔ Backup, replication architectures
  - Performance, while important, is secondary to the above
- HPC storage architectures are vastly different
  - Support for thousands to hundreds of thousands of processors/GPGPUs from a single storage pool
  - “Project needs” are more or less static for a given project, but may differ from project to project
  - Performant, non-standard architectures are OK for HPC

Challenges of Moving HPC into the Enterprise

- Thinking differently
- Planning differently
- Managing resources differently
- Executing differently

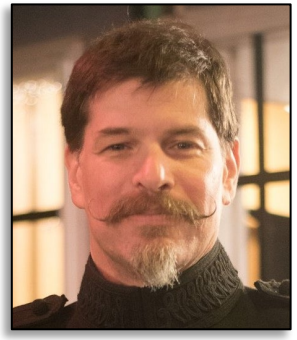
Panelists

KIOXIA

Matt Hallberg
Sr. Product Marketing Manager
Matt.Hallberg@kioxia.com
www.kioxia.com

NVIDIA

Reggie Reynolds
Principal Prod Mktg Mgr, Storage
www.nvidia.com



WEKA

Joel Kaufman
Technical Marketing Manager
www.weka.io

G2M RESEARCH

Mike Heumann
Managing Partner
www.g2minc.com

G2M RESEARCH

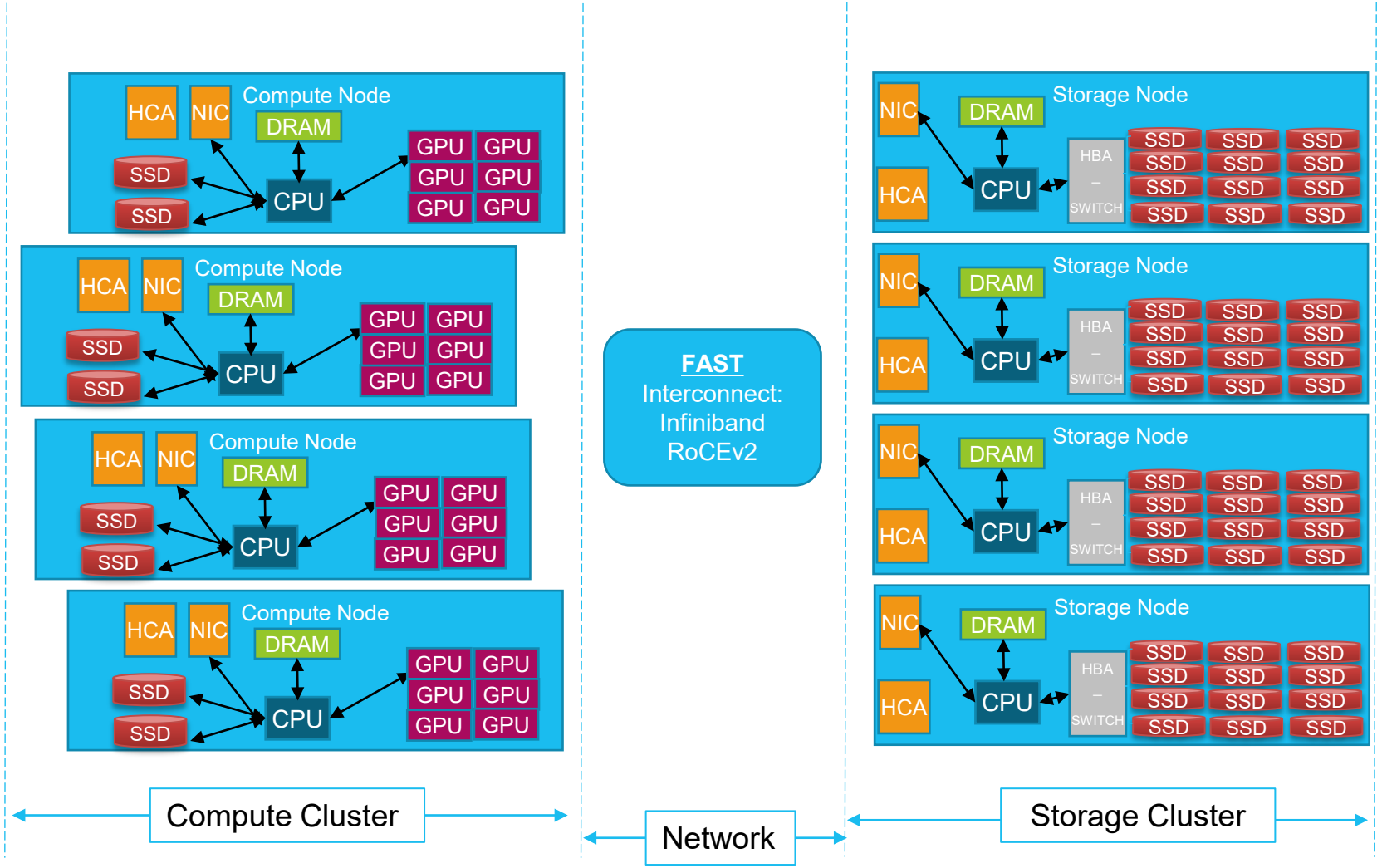
KIOXIA

Kioxia

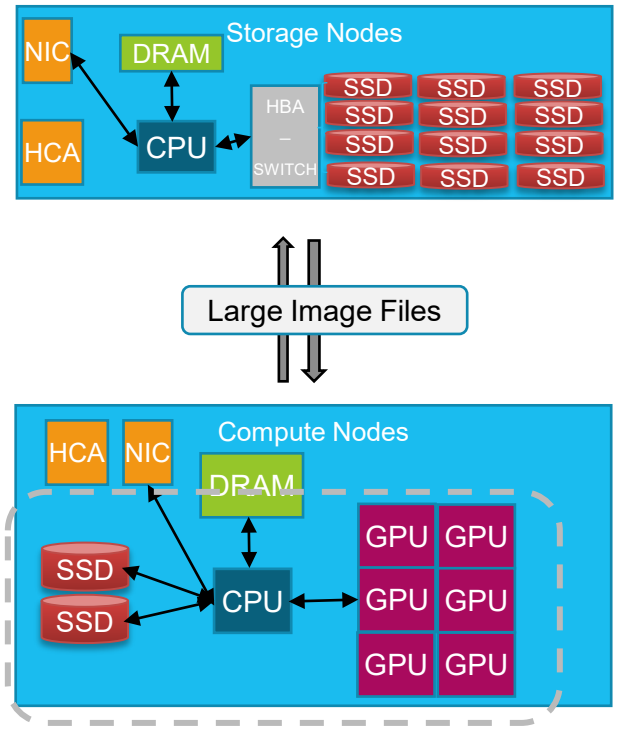
- Matt Hallberg
  Senior Product Marketing Manager
  www.kioxia.com

Overview of HPC Architecture

- **HPC is indeed "built differently"**
  - Nothing is "general purpose", everything hardware component has a specific purpose
    - Network – lowest latency and highest performance possible
    - Compute – the biggest, baddest amount of computations possible
    - Storage – store/prepare the data for the compute nodes, but generally an afterthought…**WHY?**
- **HPC enables many applications**
  - Research
  - Media
  - Resources
  - Manufacturing
  - AI, ML, and Deep Learning



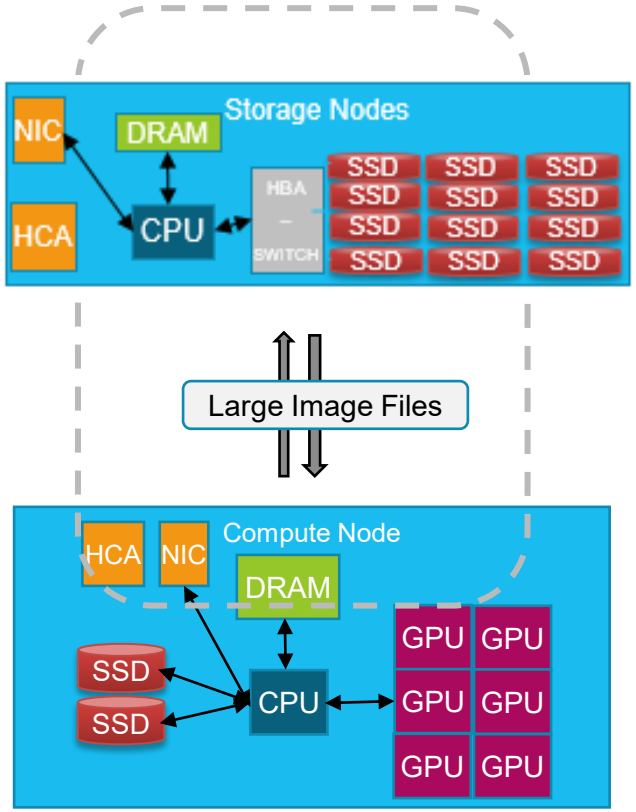
Why Local Storage (Compute Node) Matters



- **The training phase of machine learning is the most-resource intensive set of operations**
  - Datasets are growing at a fast pace, MRIs can each reach multiple TBs, and training sets can be composed of thousands of images
  - Whether you are running on RAM or on local storage, the local storage needs to be able to handle reads and writes in blazing fashion and with little impact to overall latency
  - Moving the data in, moving the data out, and checkpointing all need to be completed quickly to minimize the idleness of the GPUs
- **PCIe® 4.0 SSDs' noticeable benefits with file copying and other I/O tasks versus PCIe 3.0 SSDs**
  - For sequential workloads
    - Up to 7000MB/s on Reads
    - Up to 4200MB/s on Writes
  - For random workloads
    - 1M+ IOPS on Random Reads
    - 70K+ on Random Writes
  - PCIe 4.0 SSDs are also able to take advantage of 3D NAND's higher densities, allowing for up to 30TB in a 2.5" SSD

Cool new technology: GPUDirect® via Magnum I/O from NVIDIA, allows GPUs to talk directly to local storage, bypassing CPU / DRAM!

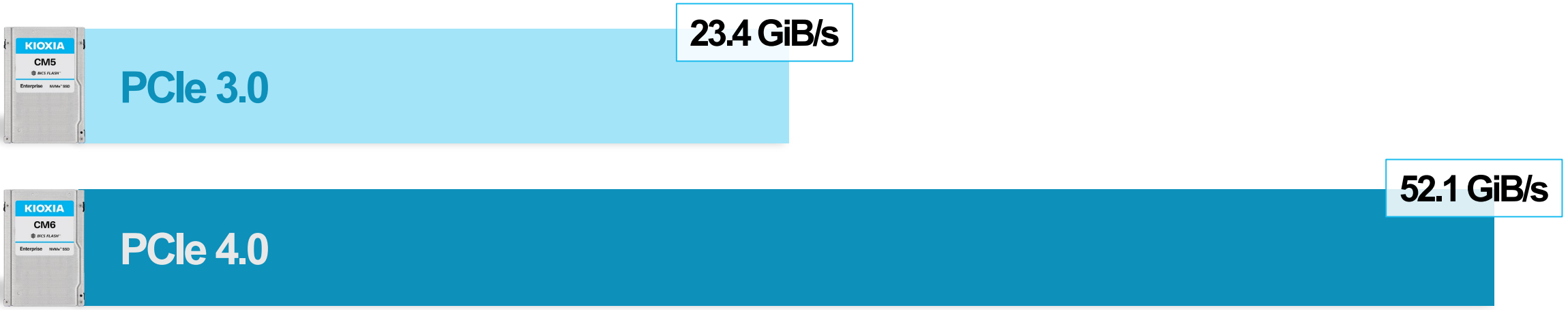
Why Remote Storage (Storage Node) Matters



- Networking speeds, technologies, and topologies have greatly advanced over the past few years
  - 200GbE NICs
  - RoCEv2 / RMDA over Ethernet
  - NVMe® over Fabrics (NVMe-oF™) deployments
  - Etc.
- Data sets are comprised of thousands of files all of which need to be sent and received to and from the compute node to minimize downtime
- The storage behind the NIC(s) should be optimized for sequential performance to send data for processing and receive processed data. The faster the offload can occur, the faster the local storage can send the data to the GPUs
  - NVMe-oF deployments show their strength here by improving performance and reducing latency
  - Using NVMe SSDs for staging “warm” data vs cold data (cheaper storage) optimizes the spend on remote storage
  - GPUDirect® technology can also take advantage of the remote NVMe SSDs

Cool new technology: GPUDirect via Magnum I/O from NVIDIA, allows GPUs to talk directly to local NIC(s), bypassing CPU / DRAM!

QCT / AMD / Broadcom / KIOXIA PCIe® 4.0 SSD Demo @ Microsoft Ignite 2019



Benchmark: Storage I/O for Virtualized App

- 100% sequential 129KiB reads
- More than double the performance over PCIe 3.0 configuration
- Each drive delivered over 7GiB/s!

Microsoft Windows Server 2019, Azure Stack HCI 2 node cluster

- Build 177763.775 Data Center Edition
- Hyper-V and Storage Spaces Direct
- Two-way mirror volumes
- VM Fleet with diskspd 2.0.21a
- FIO

QCT D43K-1U Server(2)

- 2 x AMD Epyc 7742
- 512GiB DRAM
- 4x KIOXIA CM5 PCIe 3.0, 7.68T SSD
- 4x KIOXIA CM6 PCIe 4.0, 7.68T SSD
- Broadcom NetXtreme™-P2100G 200Gbps NIC

There's more to come: Ethernet Bunch of Flash (EBOF) over NVMe-oF™

- **There is a new type of storage appliance on the horizon: Ethernet Bunch of Flash (EBOF)**
  - Utilizes the already proven performance gains and latency reductions of RoCEv2 / RDMA and NVMe-oF
  - Ethernet SSDs are attached to the network, offering up to 2 ports of 25Gbps Ethernet and RoCEv2 RDMA connections
  - No need for CPU or DRAM, allowing for direct access to storage
    - Avoids the "bounce buffer" of data needing to cycle through CPU and DRAM during offload to storage
  - Highly reduced system costs and cooling benefits
    - No CPU, DRAM, NIC, HBAs, etc.
    - Only cooling SSDs and components… 24 SSDs @ 18W = 432W!
  - EBOF could work with SDS for GPUDirect® / Magnum IO™ technology to allows direct GPU access to NIC(s)

- **Current proof of concept EBOF system highlights**
  - 2U Chassis with up to 24 drives
    - Can route for high availability (single drive going to 2 switches)
  - 600Gb/s storage throughput per 100Gb/s embedded network switch
  - High performance: 830K IOPS per drive, 20M+ IOPS per 24 bay system
  - Press Release: https://business.kioxia.com/en-us/news/2020/ssd-20200922-2.html

KIOXIA CM6 Series Enterprise NVMe SSDs

- Enterprise PCIe® 4.0, NVMe™ 1.4 SSDs
- Form factors: 2.5-inch, 15mm Z-height
- Proprietary KIOXIA architecture: controller, firmware and BiCS FLASH™ 96-layer 3D TLC memory
- SFF-TA-1001 conformant (U.3) works with Tri-mode controllers and backplanes
- Dual-port design for high availability applications
- 6th generation die failure recovery and double parity protection
- High performance with lower power consumption
- Power loss protection (PLP) and end-to-end data protection
- Suited for 24x7 enterprise workloads
- Data security options: SIE, SED, FIPS 140-2
- Six power mode settings
- Available now

			CM6 (Mixed-Use)					CM6 (Read-Intensive)					
Endurance		DWPD	3					1					
User Capacity*		GB	800	1600	3200	6400	12800	960	1920	3840	7680	15360	30720
Sequential Read	128KB(QD32)	MB/s	6900	6900	6900	6900	6900	6900	6900	6900	6900	6900	6850
Sequential Write	128KB(QD32)	MB/s	1400	2800	4200	4000	4000	1400	2800	4200	4000	4000	4000
Random Read	4KB(QD256)	KIOPS	800	1300	1400	1400	1400	800	1200	1400	1300	1400	900
Random Write	4KB(QD32)	KIOPS	100	215	350	325	330	50	100	170	170	170	70

* KIOXIA Corporation definition of capacity: 1 GB = 1,000,000,000 (10^9) bytes (see end of presentation for full capacity disclaimer).

Note: Specifications are subject to change

KIOXIA CD6 Series Data Center NVMe SSDs

- Data Center PCIe® 4.0, NVMe™ 1.4 SSDs
- Form factors: 2.5-inch, 15mm Z-height
- Proprietary KIOXIA architecture: controller, firmware and BiCS FLASH™ 96-layer 3D TLC memory
- SFF-TA-1001 conformant (U.3) works with Tri-mode controllers and backplanes
- Single-port design, optimized for data center class workloads
- 6th generation die failure recovery and double parity protection
- Consistent performance and reliability in demanding 24x7 environments
- Designed for high density storage deployments
- Power loss protection (PLP) and end-to-end data correction
- Data security options: SIE, SED, FIPS 140-2
- Five power mode settings
- Available Now

			CD6 (Mixed-Use)					CD6 (Read-Intensive)				
Endurance		DWPD	3					1				
User Capacity*		GB	800	1600	3200	6400	12800	960	1920	3840	7680	15360
Sequential Read	128KB(QD32)	MB/s	5800	5800	6200	6200	5500	5800	5800	6200	6200	5500
Sequential Write	128KB(QD32)	MB/s	1300	1150	2350	4000	4000	1300	1150	2350	4000	4000
Random Read	4KB(QD256)	KIOPS	700	700	1000	1000	750	700	700	1000	1000	750
Random Write	4KB(QD32)	KIOPS	90	85	160	250	110	30	30	60	85	30

* KIOXIA Corporation definition of capacity: 1 GB = 1,000,000,000 (10^9) bytes (see end of presentation for full capacity disclaimer).

Note: Specifications are subject to change

KIOXIA

Definition of capacity: KIOXIA defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes and a terabyte (TB) as 1,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1GB = 2^30 = 1,073,741,824 bytes and therefore shows less storage capacity. Available storage capacity (including examples of various media files) will vary based on file size, formatting, settings, software and operating system, such as Microsoft Operating System and/or pre-installed software applications, or media content. Actual formatted capacity may vary.

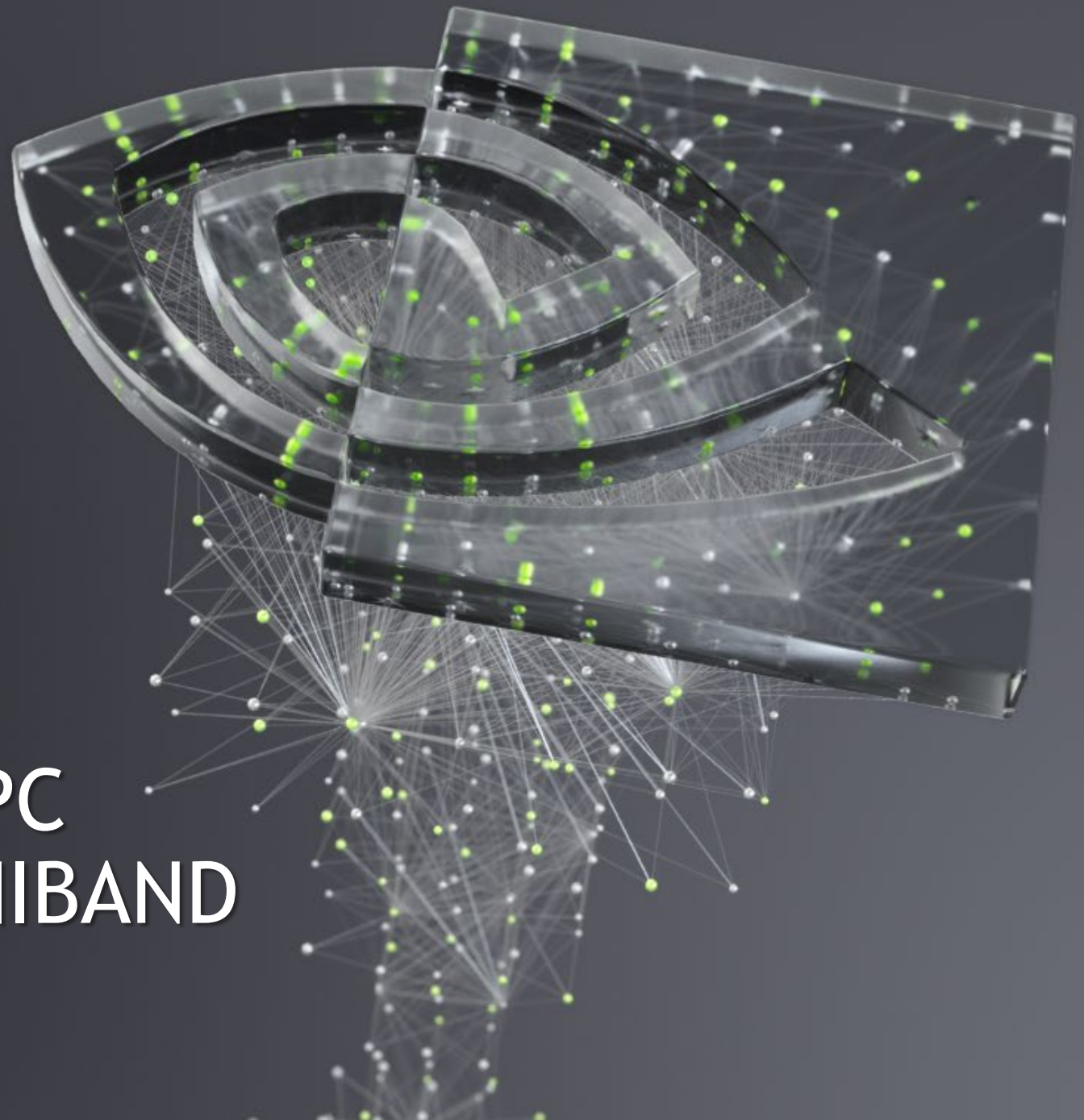
All company names, product names and service names may be trademarks of their respective companies.

Images are for illustration purposes only.

G2M RESEARCH

NVIDIA

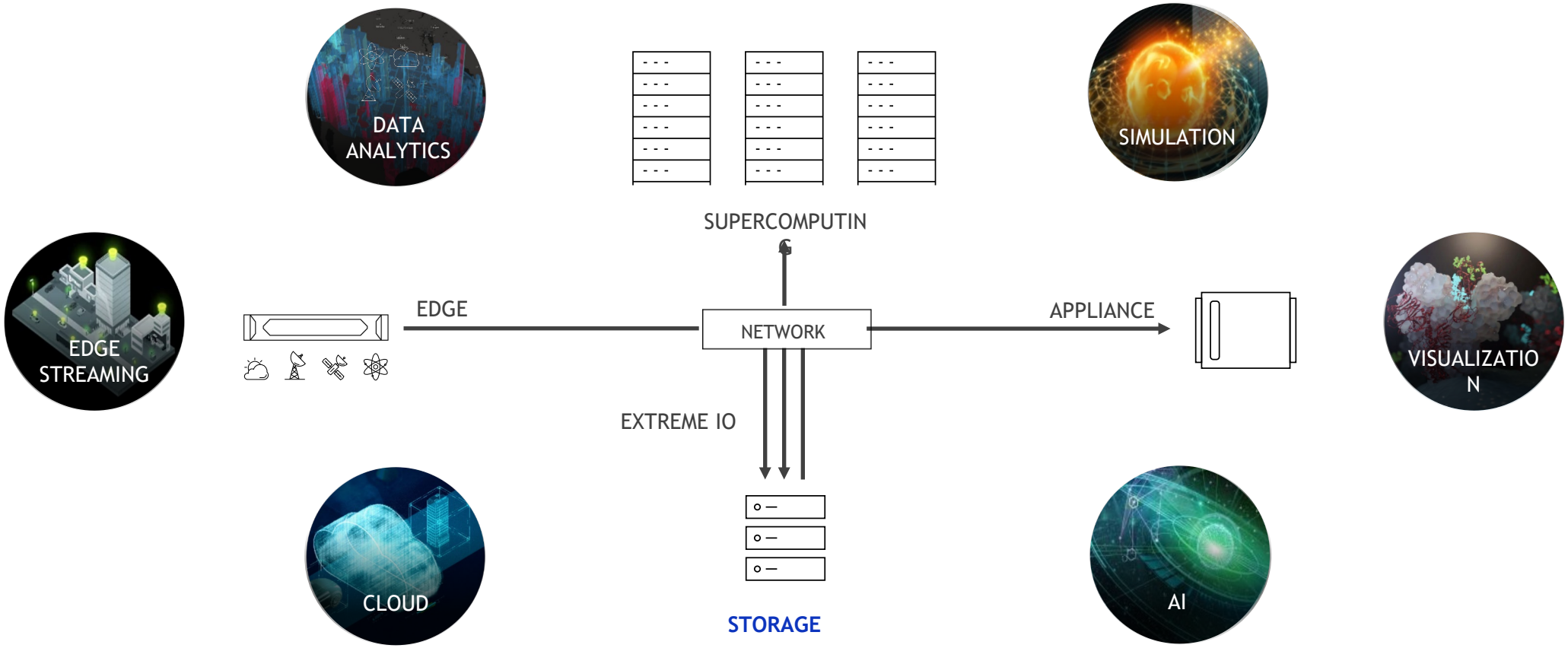
- Reggie Reynolds
  Principal Product Marketing Manager, Storage
  www.nvidia.com



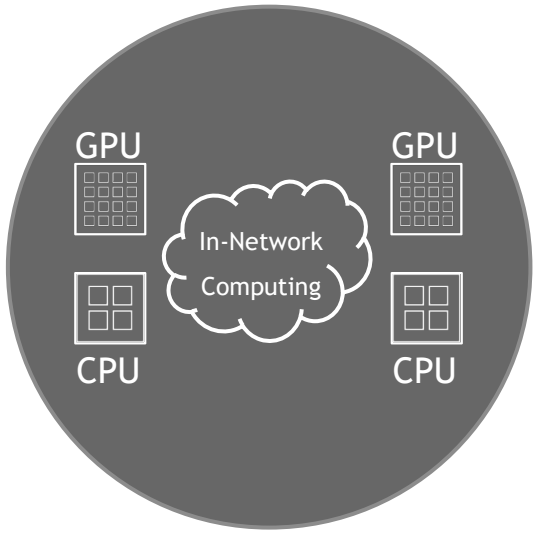
NVIDIA

MAXIMIZING THE PERFORMANCE OF HPC STORAGE WITH INFINIBAND

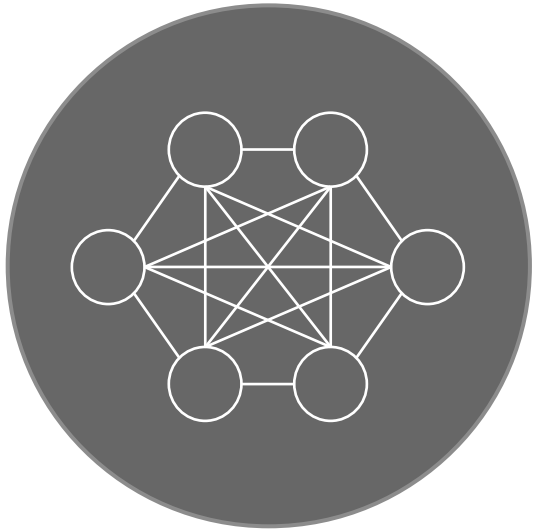
EXPANDING UNIVERSE OF SCIENTIFIC COMPUTING



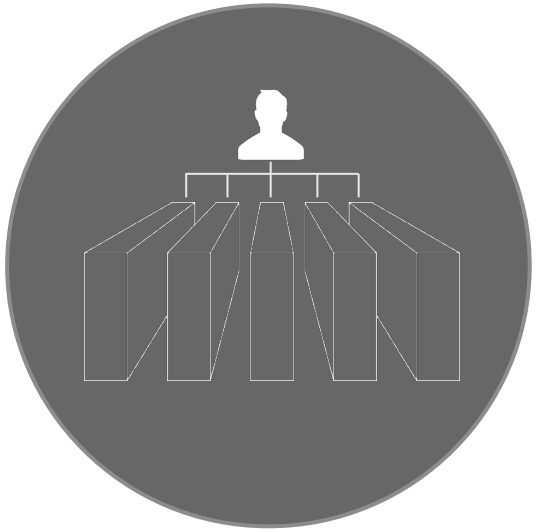
NVIDIA MELLANOX INFINIBAND TECHNOLOGY



In-Network Computing



Architected to Scale



Centralized Management

Standard

NVIDIA MELLANOX INFINIBAND INFRASTRUCTURE

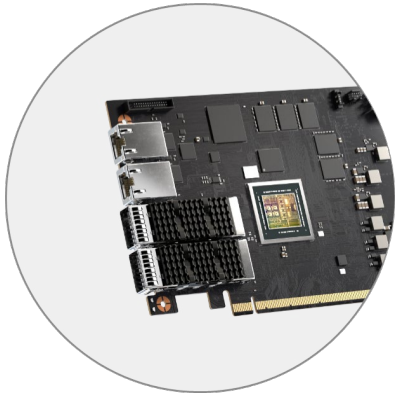
In-Network Computing Accelerated Network for Supercomputing

METROX-2 LONG-HAUL

SKYWAY GATEWAY

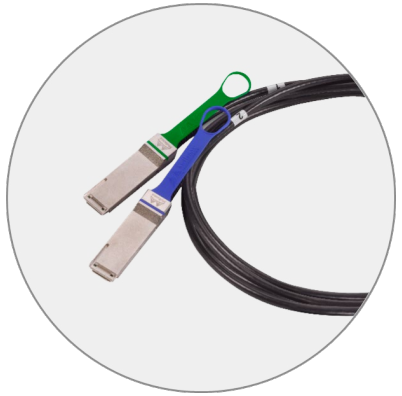
UFM CYBER-AI

CONNECTX ADAPTER



BLUEFIELD DPU

QUANTUM SWITCH



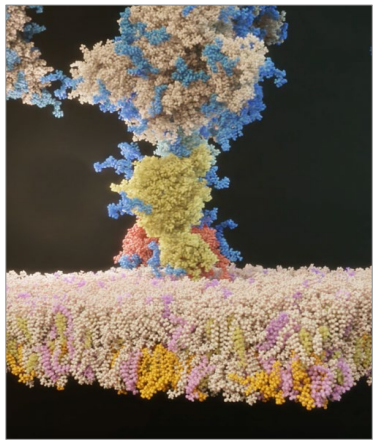
LINKX

DRIVING THE MOST SUCCESSFUL AI-OUTCOMES

Training, inference and data science on the best AI platform

Full Stack

GPU Compute
InfiniBand Network
Management
Accelerated Stacks for AI and HPC



Out of Box One-Click Modular Design

Faster Deployment
Simplified, Verified
Scalable
Field Proven

Breakthrough Performance

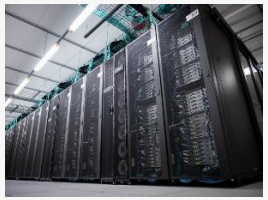
In-Network Computing
SHARP AI Technology
GPUDirect
HDR 200G InfiniBand
Lowest Latency
One Network for All

Seamless Integration

InfiniBand Clusters
Bridge via Skyway to
External Ethernet

NVIDIA MELLANOX HDR 200G INFINIBAND ACCELERATES NEXT GENERATION SUPERCOMPUTERS (EXAMPLES)

Australian National University



9 PFlops
3K HDR Nodes
Dragonfly+
Topology

JAMSTEC 国立研究開発法人 海洋研究開発機構
JAPAN AGENCY FOR MARINE-EARTH SCIENCE AND TECHNOLOGY

19.5 PetaFLOPS
2.5K HDR Nodes
Dragonfly+ and Fat
Tree

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

16 PFlops
3K HDR Nodes
Dragonfly+
Topology

ECMWF
EUROPEAN CENTRE FOR MEDIUM RANGE WEATHER FORECASTS

8K HDR Nodes
Dragonfly+
Topology

35.5 PFlops
2K HDR Nodes
Fat-Tree Topology

H L R S
High-Performance Computing Center | Stuttgart

19.3 PFlops
5.6K HDR Nodes
Dragonfly+
Topology

NVIDIA

63.5 PFlops
4.5K HDR Nodes
Fat-Tree Topology

Microsoft Azure

HPC/AI Cloud
HDR InfiniBand

NSF
SDSC
SAN DIEGO SUPERCOMPUTER CENTER
PITTSBURGH SUPERCOMPUTING CENTER
PURDUE UNIVERSITY

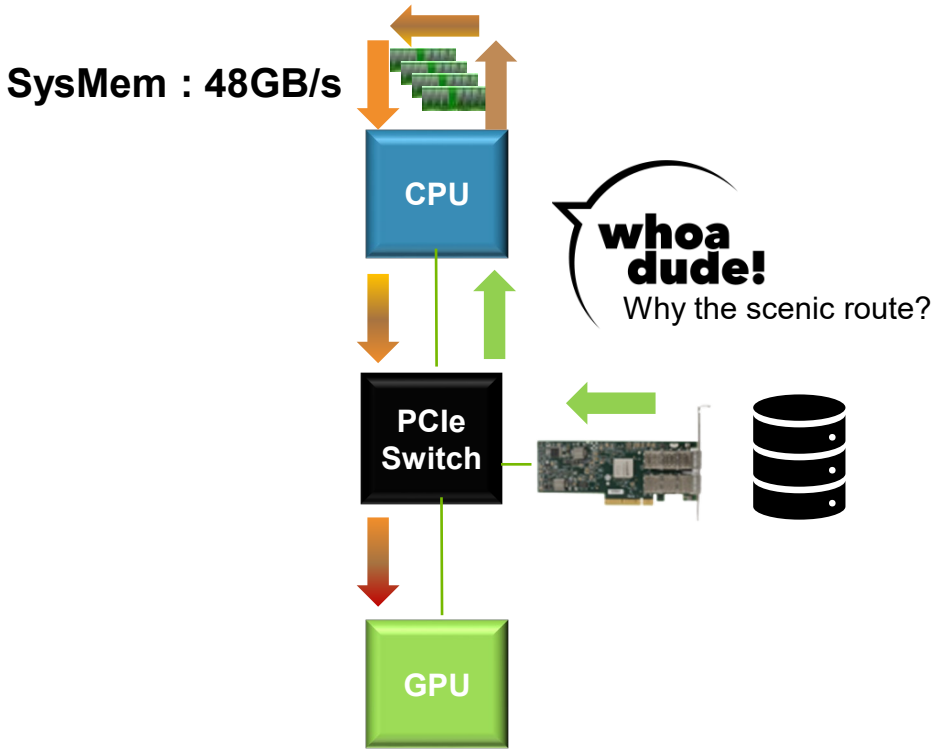
HDR InfiniBand
Supercomputers

TACC
TEXAS ADVANCED COMPUTING CENTER

23.5 PFlops
8K HDR Nodes
Fat-Tree Topology

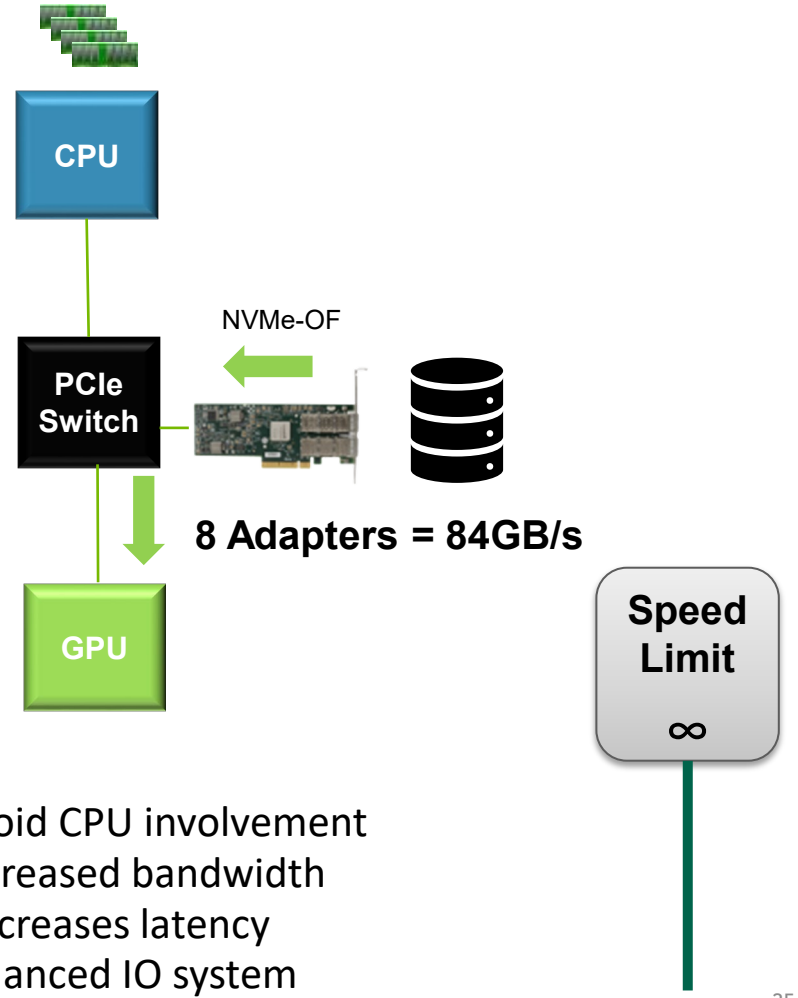
NVIDIA GPUDirect Storage (Magnum IO)

CPU-Centric (Onload)



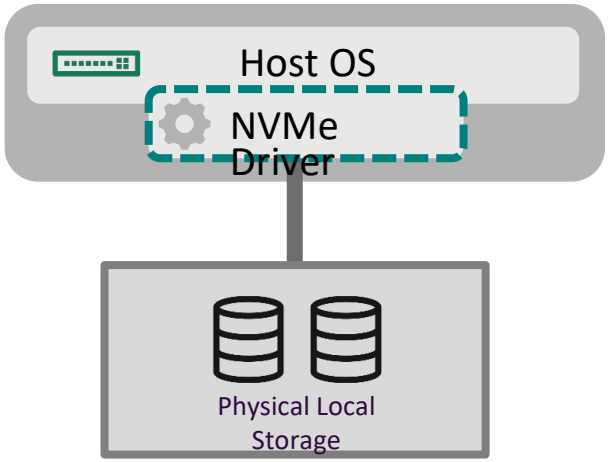
- Copying into system memory “bounce buffer”
- Bandwidth limited
- CPU must be involved
- cudaMemcpy overhead

Data-Centric (Offload)



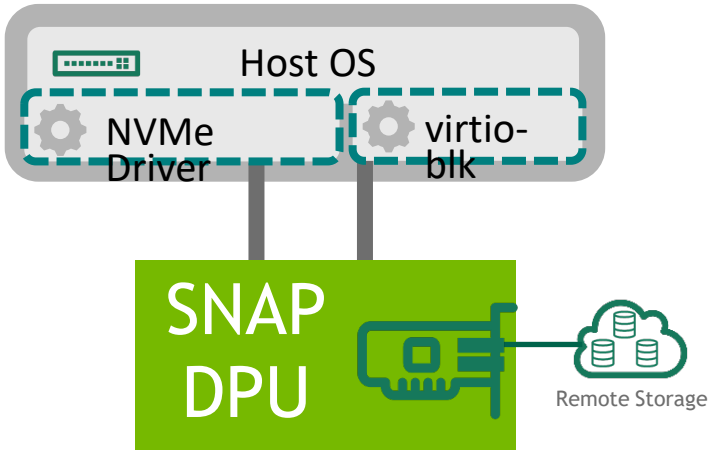
- ✓Avoid CPU involvement
- ✓Increased bandwidth
- ✓Decreases latency
- ✓Balanced IO system

SNAP: LOCAL STORAGE TO EMULATED STORAGE



Physical Local NVMe Storage

- ✓ Serving bare-metal and hypervisor/VMs
- ✗ Bound by physical SSDs capacity
- ✗ Under-utilized storage
- ✗ Scalability on demand
- ✗ Over-provisioning bound to compute node



DPU SNAP Drive Emulation

- ✓ Serving bare-metal and hypervisor/VMs
- ✓ Over-provisioning, scaled to rack/cluster
- ✓ Saving OPEX and CAPEX
- ✓ OS-agnostic using inbox standard driver
- ✓ Supports all network transport types – NVMe-oF, iSCSI, iSER and even proprietary
- ✓ Accelerated data path for VMs
- ✓ Live-migration with virtio-blk

SUMMARY

InfiniBand is the network technology of choice for demanding storage performance and low latency requirements at scale both now and far into the future in HPC and AI.

NVIDIA provides a complete end-to-end Mellanox InfiniBand solution that can seamlessly connect to Ethernet network via the Skyway gateway.

BlueField DPU provides a multitude of accelerators and offloads along with a programmable platform. Combining the DPU with InfiniBand provides storage with the opportunity to deliver extremely flexible high performing solutions.

G2M RESEARCH

WEKA

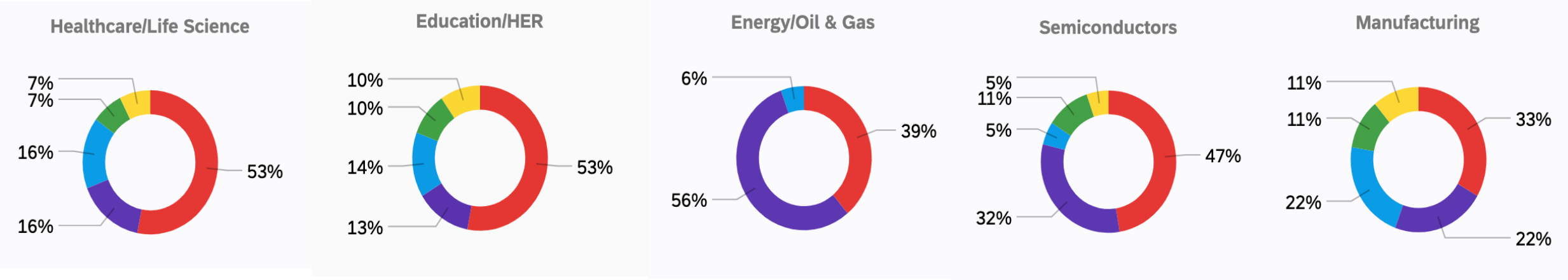
Weka

- Joel Kaufman
  Technical Marketing Manager
  www.weka.io

The Data Scale Problem

- **TL;DR: Big data needs big compute, big networking, big storage to achieve big outcomes.**
  - As compute gets faster, the ability to distribute performance by fanning out becomes problematic
  - Data Is always growing, and the rate of growth in HPC data is increasing
  - In order to be competitive, the expectations for speed of outcomes has increased
- **Weka solves many of these issues with storage that accelerates time-to-value for HPC environments.**

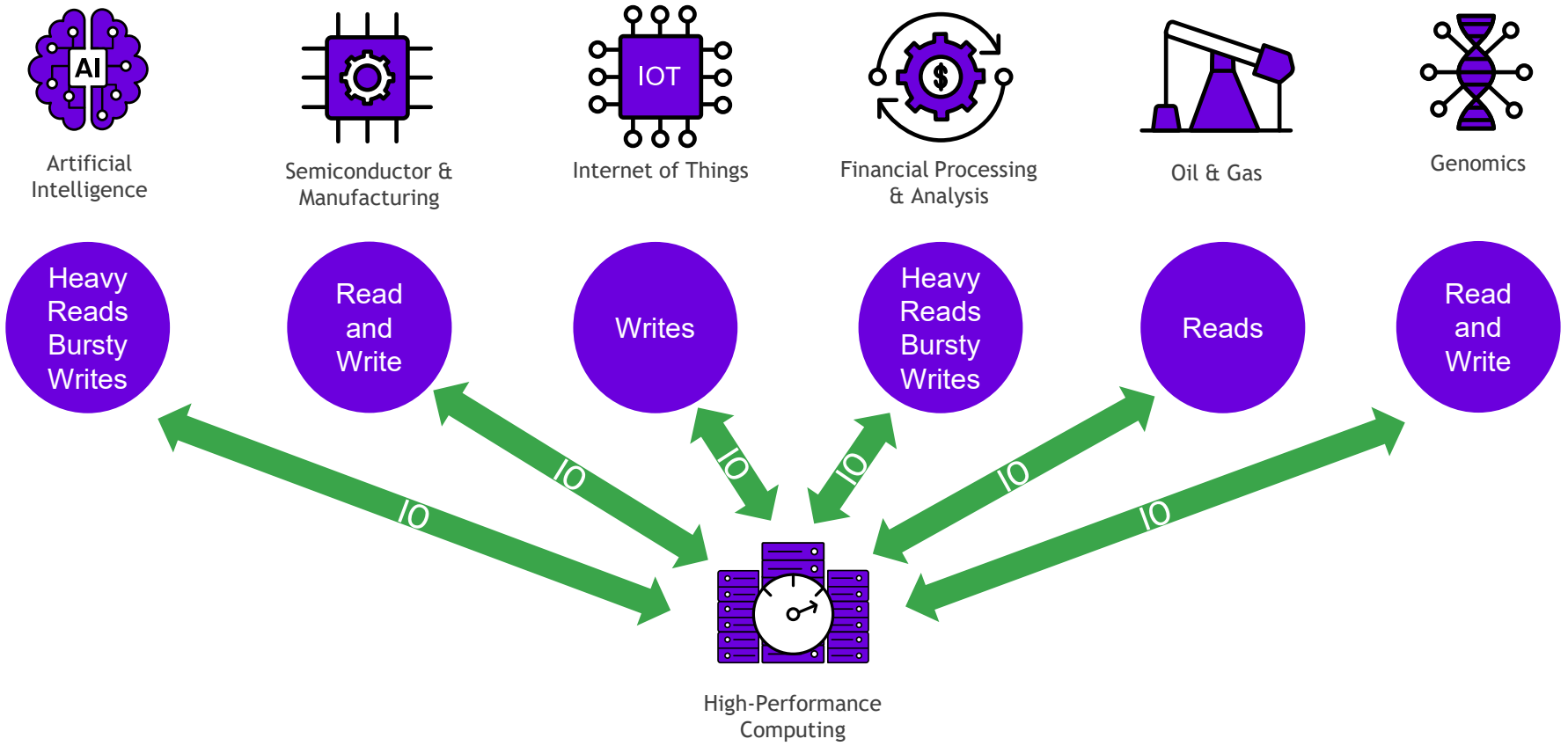
Primary Sources of Data for Workloads by Vertical



Here are some interesting findings about data sources:

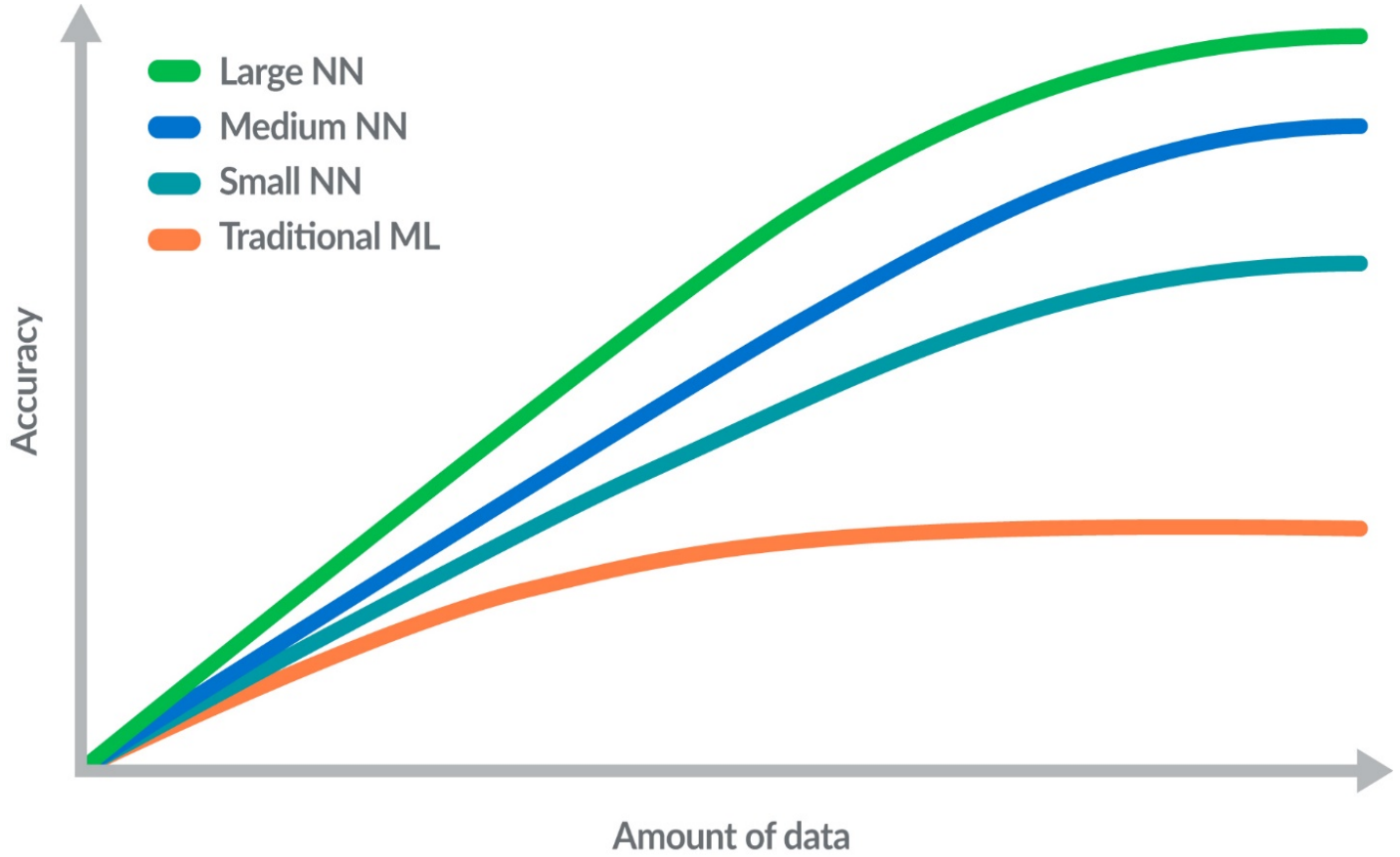
- Healthcare and higher education depended significantly on self- generated data.
- The Oil and Gas and Semiconductor segments have higher amounts of data sourced from sensors and used for monitoring.
- Finance leverages external data sources, such as LexisNexis and Reuters, Bloomberg, and real-time stock exchange data.

I/O Size and Patterns Vary Throughout HPC Workloads



Data sets vary from millions of very large files to billions of tiny files

Data Sets Are Huge and Need Performance



Source: Andrew Ng - Accuracy of models grows with the amount of training data

”It is not who has the best algorithm that wins. It’s who has the most data.”
Andrew Ng

Requires a massively scalable system

GPUs Have “Densified” Compute into a Single Server Creating a Huge Data Bottleneck

CPU-Only Servers

- 100’s of servers with CPUs
- 100’s of low bandwidth network connections
- No one server was particularly demanding on storage

100x More Compute
40x more network

GPU Accelerated Server

Current NAS solutions cannot feed these machines with enough data

Highest Performance to GPU Servers

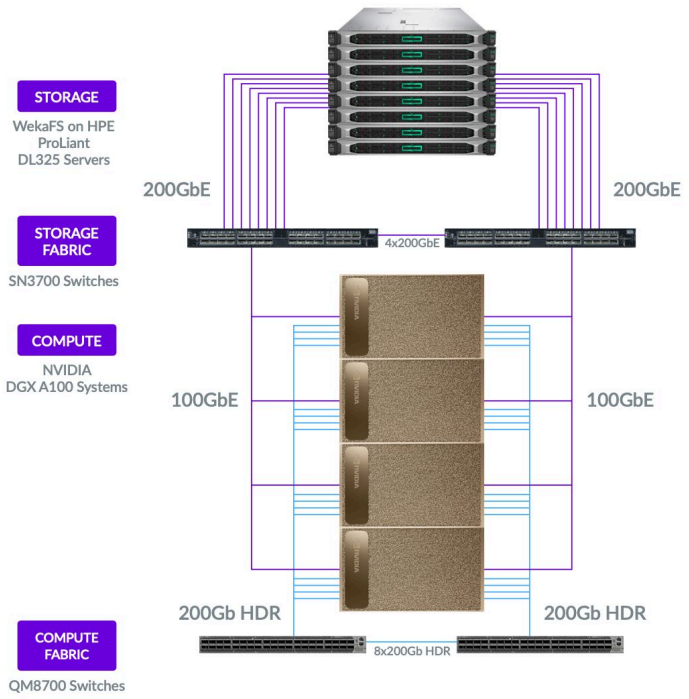
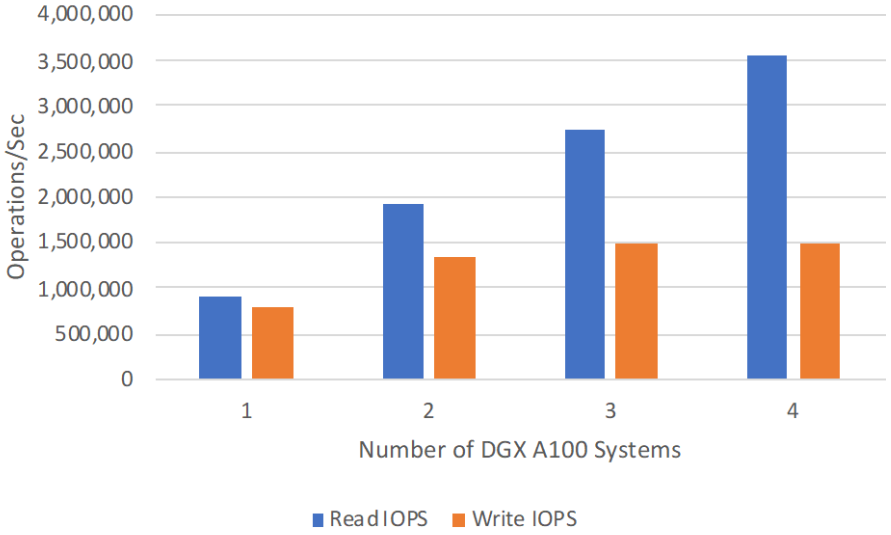
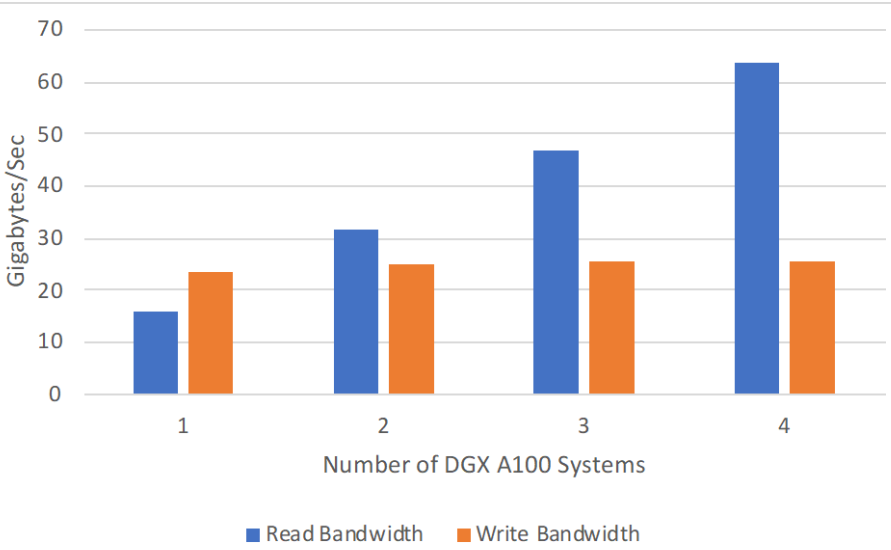


Figure 2 – Weka AI validated architecture with DGX A100 systems





FIO Operations

- 3.5 Million IOPs

FIO Bandwidth

- 64 GB/sec
- Network limited: we see 150GB/sec+ with full networking.

G2M RESEARCH

Panel Questions and Audience Surveys

Panel Question # 1

- How does storage for HPC applications differ from more traditional storage applications?
  - Kioxia
  - NVIDIA
  - Weka

Audience Survey Question #1

- To what extent is your organization utilizing HPC concepts and architectures (check one):
  - We use HPC concepts and architectures widely in our business/organization: 27%
  - We have several problem domains which utilize HPC concepts: 10%
  - We have a few problem domains which utilize HPC concepts : 12%
  - We are studying the use of HPC concepts and architectures for new workloads, but have not yet implemented any HPC solutions: 22%
  - We do not plan on utilizing HPC concepts or architectures in our organization in the near future: 10%
  - Don’t know: 20%

Panel Question #2

- HPC is becoming more widespread outside of its historical home of universities, federal agencies, and aerospace. As more traditional industries utilize HPC, what key lessons do they need to keep in mind related to storage and storage infrastructure?
  - NVIDIA
  - Weka
  - Kioxia

Audience Survey Question #2

- When looking at implementing HPC storage solutions, who do you look to for guidance on hardware and software choices (check all that apply):
  - My current IT server vendor(s): 11%
  - My current IT storage vendor(s): 19%
  - My current IT integrators: 8%
  - HPC hardware and software vendors: 38%
  - HPC application providers/integrators: 22%
  - Other: 19%

Panel Question # 3

- Under what conditions does it make sense for organizations to integrate their HPC storage solution with their general-purpose storage infrastructure?
  - Weka
  - Kioxia
  - NVIDIA

Audience Q&A

G2M RESEARCH

Thank You For Attending

G2M

RESEARCH