
NVIDIA
nvm EXPRESS
PLiOPS
EXTREME DATA PROCESSOR
G2M
RESEARCH
The Need for Speed: NVMe™, NVMe-oF™, and Data Processing Accelerators
Multi-Vendor Webinar
Tuesday Feb 21, 2023

Introduction and Ground Rules

Mike Heumann
Managing Partner

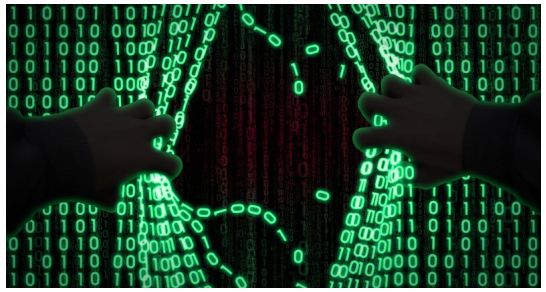
Webinar Agenda

10:00-10:04	Ground Rules and Webinar Topic Introduction (G2M Research)
10:05-10:12	NVIDIA Presentation (Rob Davis, VP of Storage Technology)
10:13-10:20	NVMe Presentation (Peter Onufryk, Intel Fellow)
10:21-10:30	Pliops Presentation (Tony Afshary, VP of Marketing)
10:31-10:36	Panel Discussion Question #1
10:37-10:37	Audience Survey #1
10:38-10:43	Panel Discussion Question #2
10:44-10:44	Audience Survey #2
10:45-10:50	Panel Discussion Question #3
10:51-11:00	Audience Q&A (10 minutes)

The Need for Speed in Today's Workloads

We have very high-speed storage, networking, and in computing (multi-core processors, GPUs, etc.). So why are so many companies focused on storage/networking/processing acceleration?

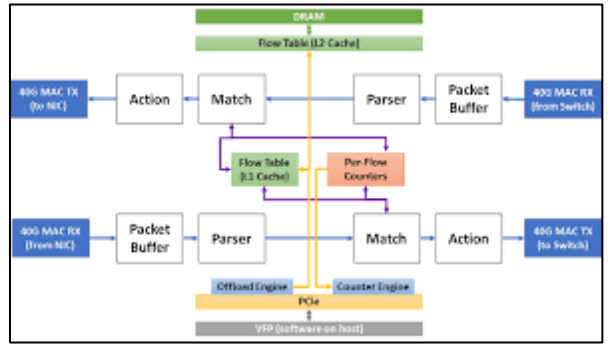
- Higher-level networking protocols and custom protocols for specific workloads require "offloads" to lower CPU utilization and increase application performance
- Advanced storage capabilities such as those offered by NVMe and NVMe-oF can also tax CPUs, reducing cycles available for workloads
- And then there is security, data resilience, and other very real needs that take CPU cycles away from workloads



Acceleration Datapoints

Hyperscalers (AWS, Azure, Google Cloud, etc.) are devoting large amounts of capital and resources to acceleration of various sorts

- Annapurna Labs (now part of AWS) has built several accelerators, including Nitro, ENA, EFA, Neuron, Inferentia ML, and Trainium ML accelerators
- Microsoft Azure also has a long history of accelerating network stacks, neural networks, and storage
- SSD vendors have enabled features such as software-defined flash to accelerate storage workloads
- ***But where can non-hyperscalers go to accelerate their workloads in the same way hyperscalers do?***



Panelists

Rob Davis
VP of Storage Technology
www.NVIDIA.com

Peter Onufryk
Intel Fellow
https://nvmexpress.org/

Tony Afshary
VP of Marketing
www.pliops.com

Mike Heumann
Principal Analyst
www.g2minc.com

NVIDIA®

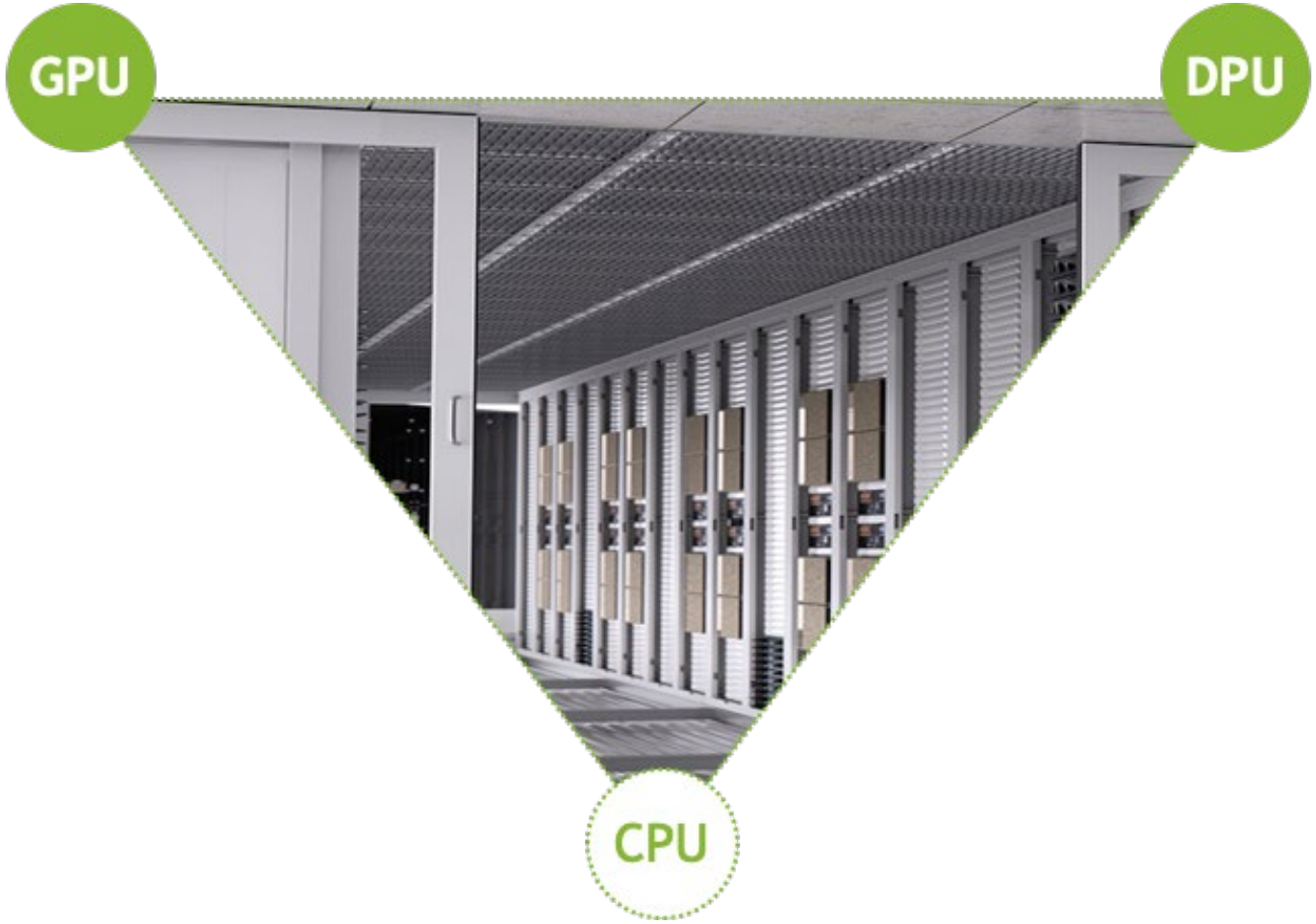
Rob Davis

VP of Storage Technologies

www.NVIDIA.com

DPU – THE NEW BUILDING BLOCK IN A DATA CENTER

The Ultimate NVMe and Data Processing Accelerator

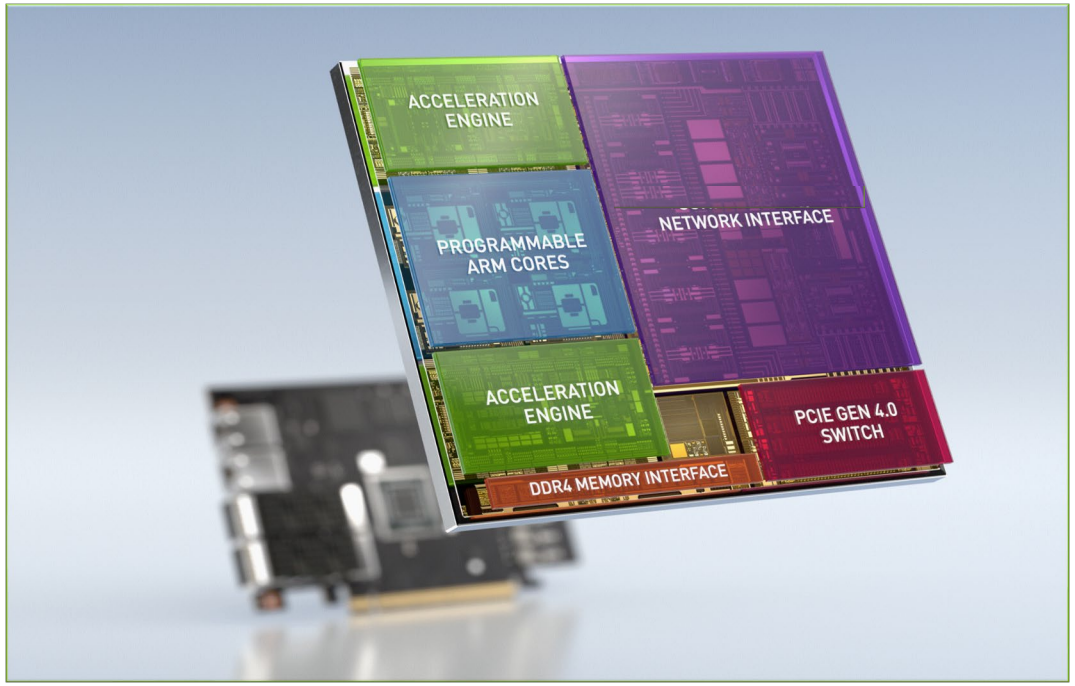


https://blogs.nvidia.com/blog/2020/05/20/whats-a-dpu-data-processing-unit/

A DPU ACCELERATES DATA CENTER WORKLOADS

DPU Puts Data Center Infrastructure on a Chip

NVIDIA BLUEFIELD-2 DPU



Offload | Accelerate | Isolate

Enterprise AI

Cloud Native HPC
Accelerated Computing

Cloud Gaming

Cloud Computing
Bare Metal as a Service

Cybersecurity
AI-Powered

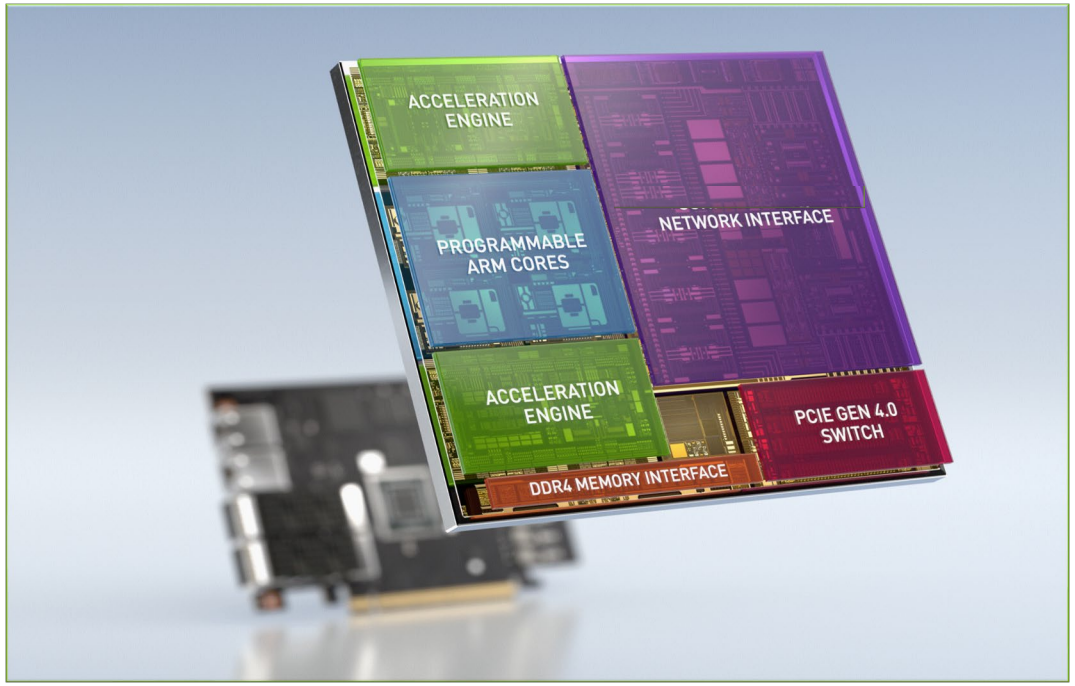
5G
Industrial/Core/Edge

https://www.nvidia.com/en-us/networking/products/data-processing-unit/

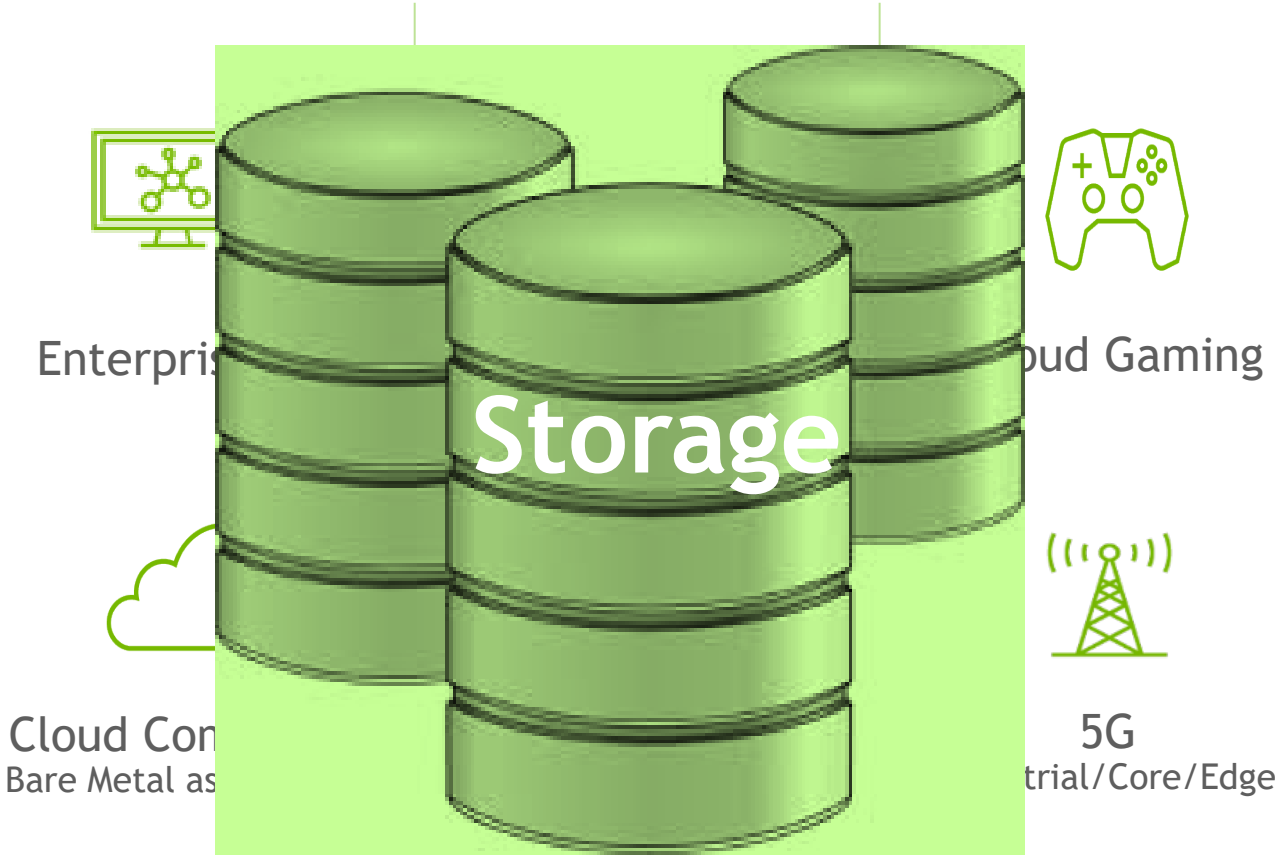
A DPU ACCELERATES DATA CENTER WORKLOADS

DPU Puts Data Center Infrastructure on a Chip

NVIDIA BLUEFIELD-2 DPU



Offload | Accelerate | Isolate



https://www.nvidia.com/en-us/networking/products/data-processing-unit/

THE BLUEFIELD DPU

22 Billion Transistors

16 Core A78 ARM processor

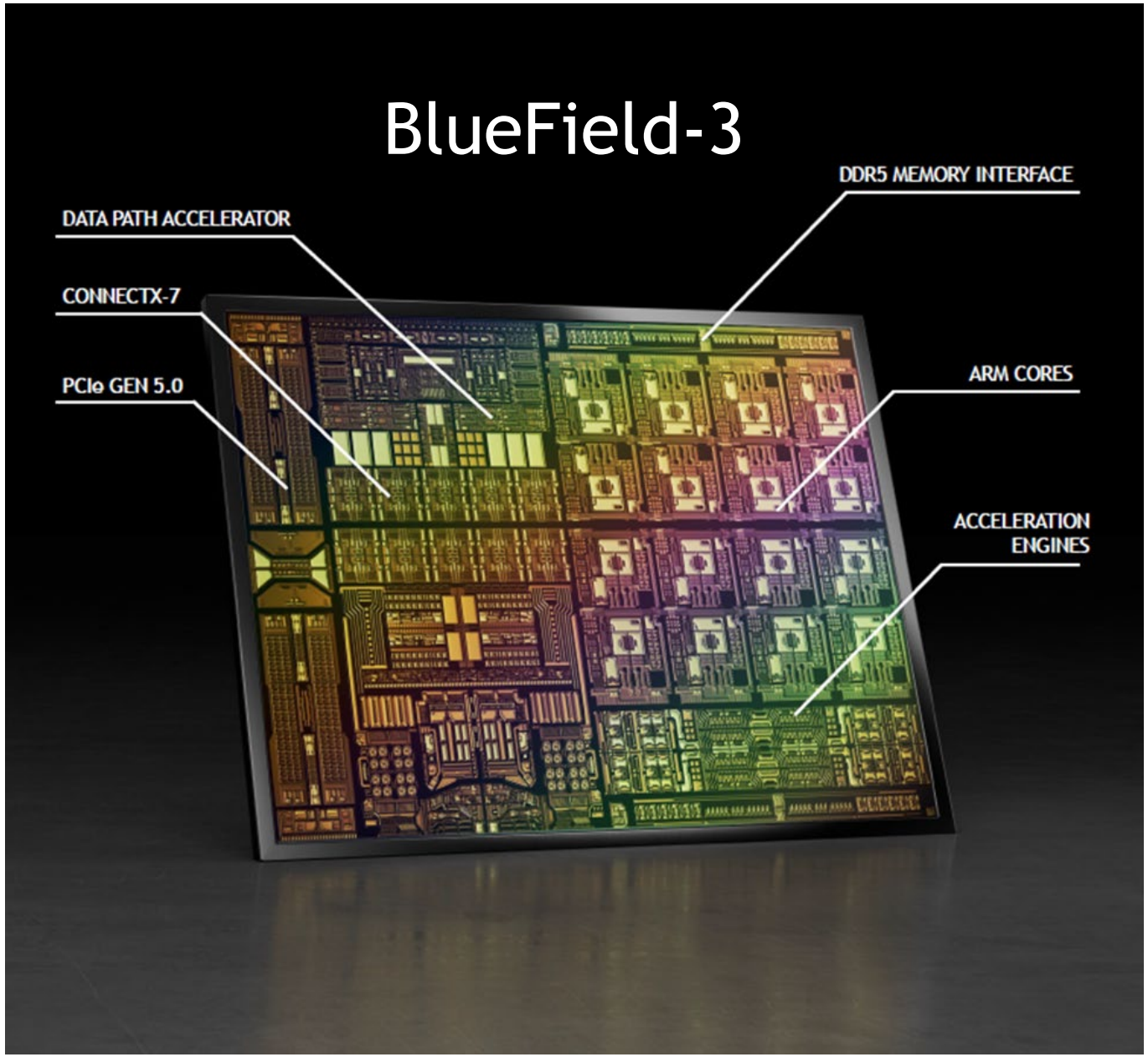
16 Core RISC-V processor

128b DDR-5 memory interfaces

4 up to 400Gb/s Ethernet or IB ports

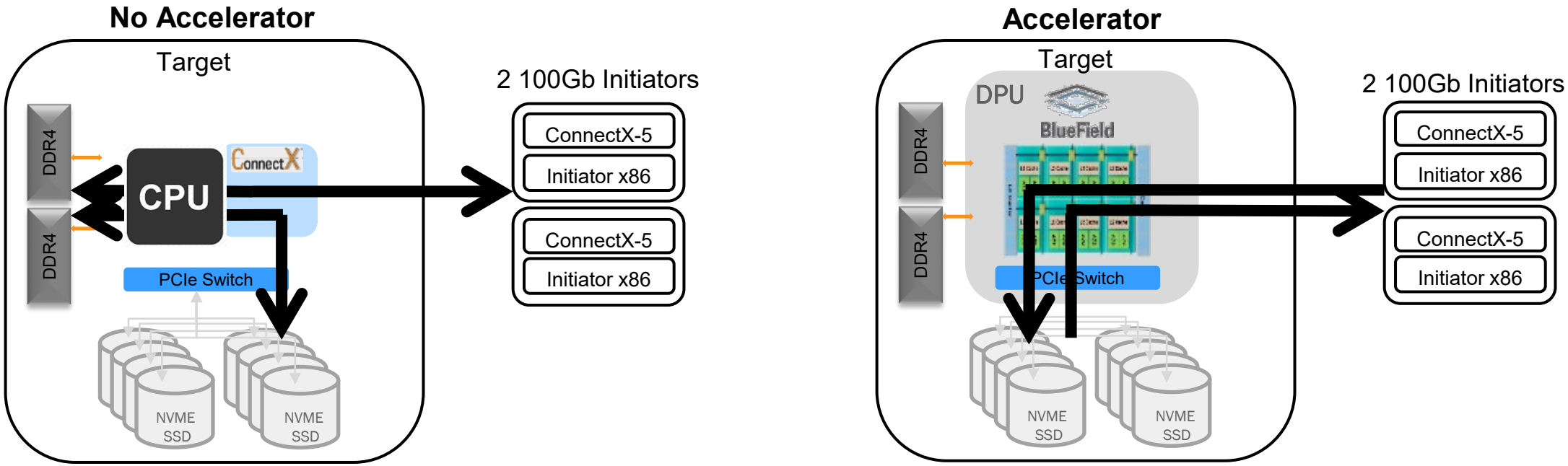
32 port gen 5 PCIe switch

Multiple HW accelerators for Encryption, Decompression, Data Integrity and much more



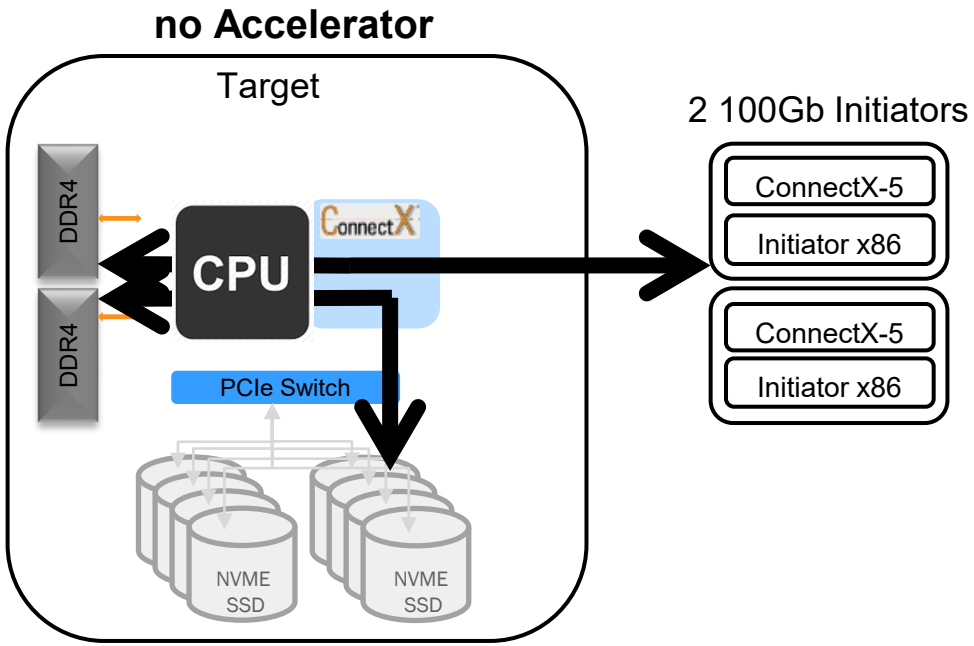
https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/documents/datasheet-nvidia-bluefield-3-dpu.pdf

BLUEFIELD NVME-OF TARGET ACCELERATOR

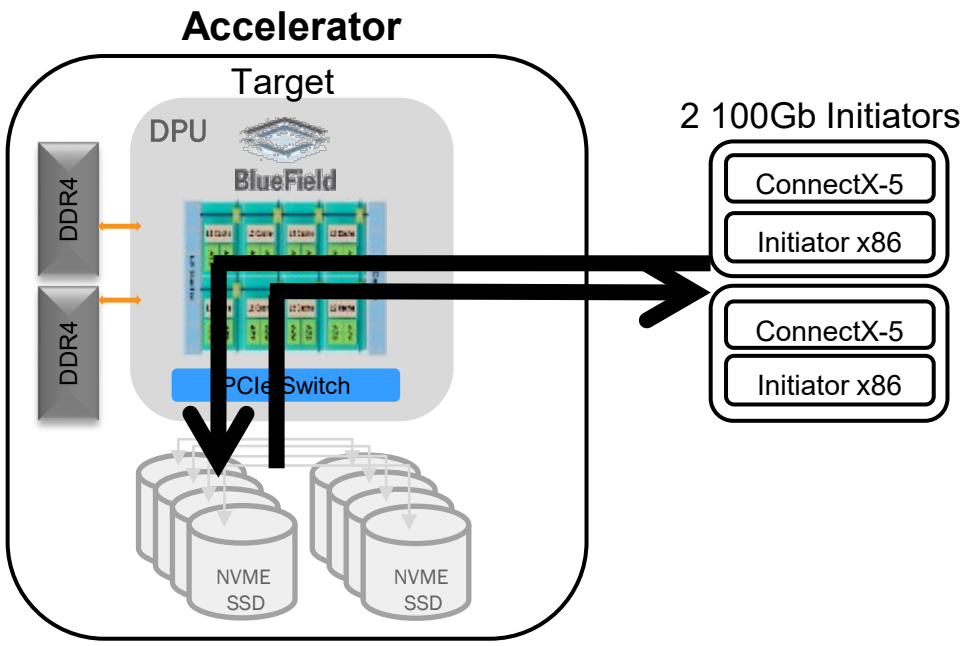


https://docs.nvidia.com/networking/display/MLNXENv451010/NVME-oF+-+NVM+Express+over+Fabrics

BLUEFIELD NVME-OF TARGET ACCELERATOR PERFORMANCE



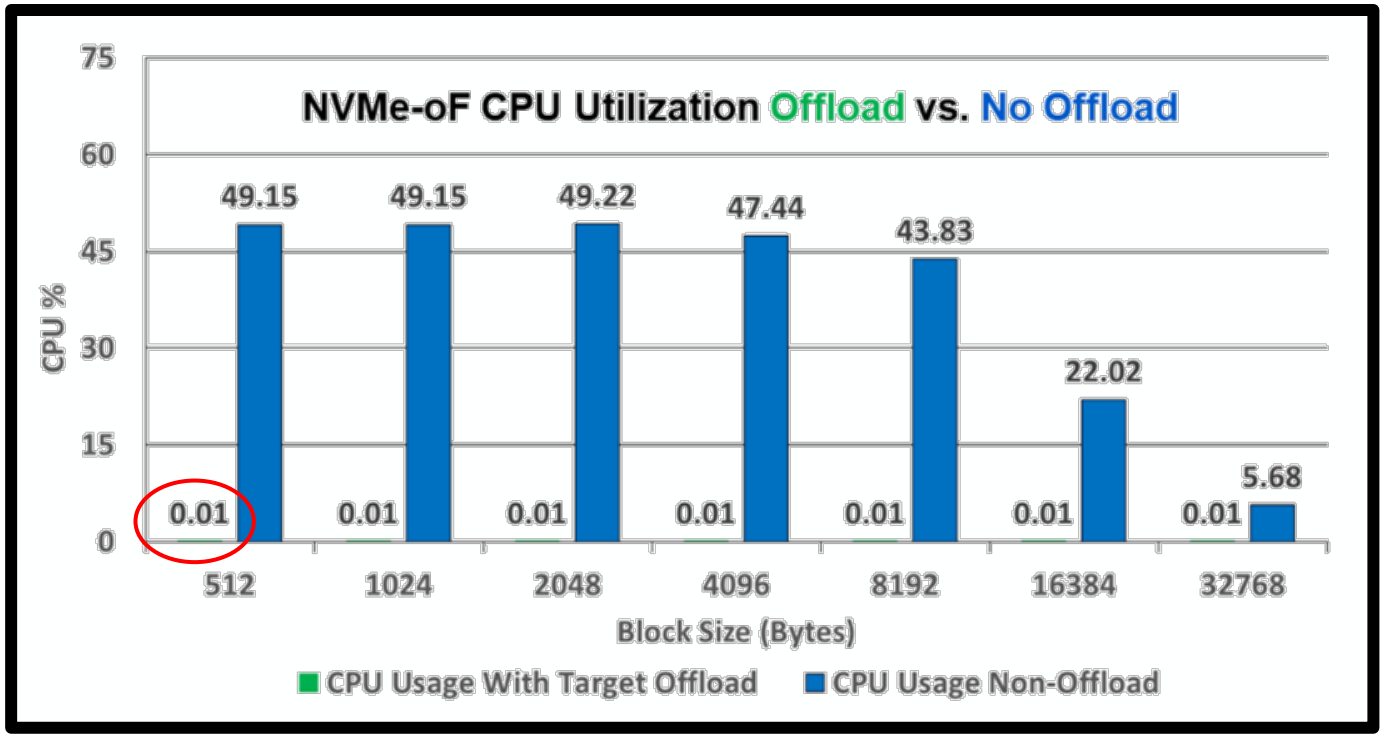
- 6M IOPs, 512B block size
- 2M IOPs, 4K block side
- ~15 usec latency (not including SSD)

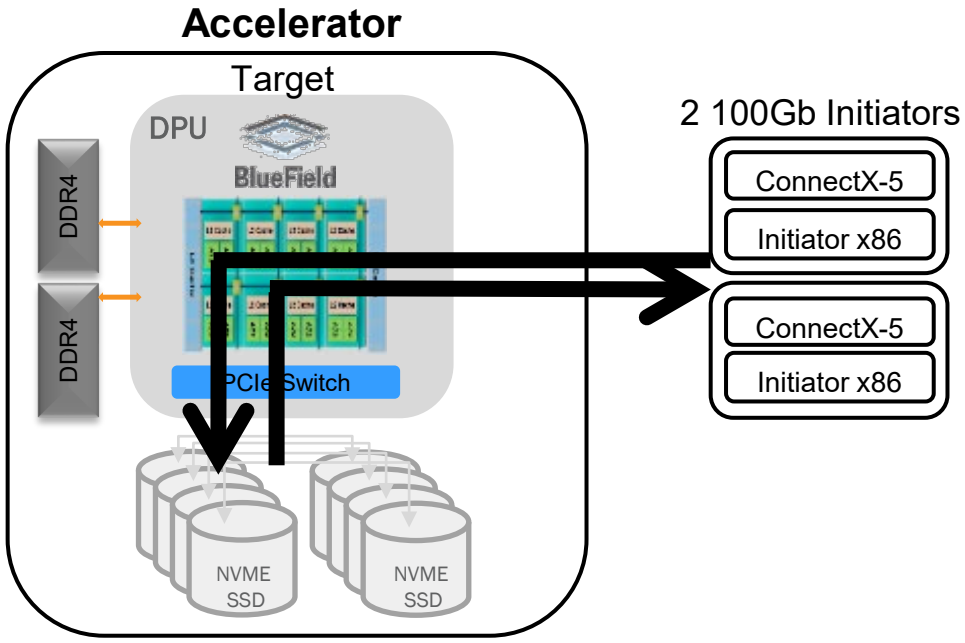


- 8M IOPs, 512B block size
- 5M IOPs, 4K block side
- ~5 usec latency (not including SSD)

https://docs.nvidia.com/networking/display/MLNXENv451010/NVME-oF+-+NVM+Express+over+Fabrics

BLUEFIELD NVME-OF TARGET ACCELERATOR PERFORMANCE





- 6M IOPs, 512B block size
- 2M IOPs, 4K block side
- ~15 usec latency (not including SSD)

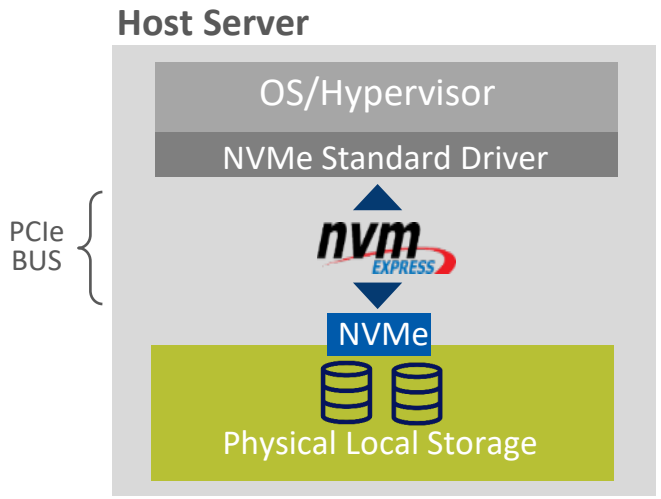
- 8M IOPs, 512B block size
- 5M IOPs, 4K block side
- ~5 usec latency (not including SSD)

https://docs.nvidia.com/networking/display/MLNXENv451010/NVME-oF+-+NVM+Express+over+Fabrics

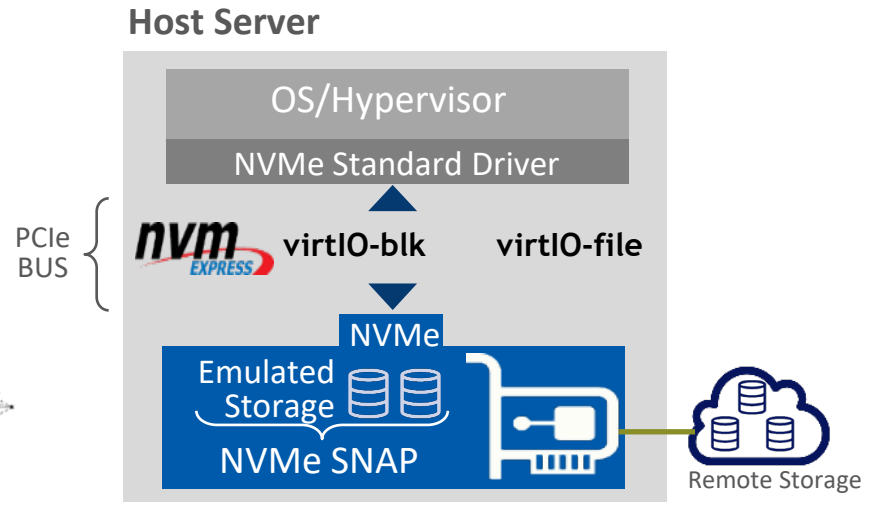
BLUEFIELD SNAP (NVME EMULATION) ACCELERATOR

NVMe-oF Initiator Acceleration

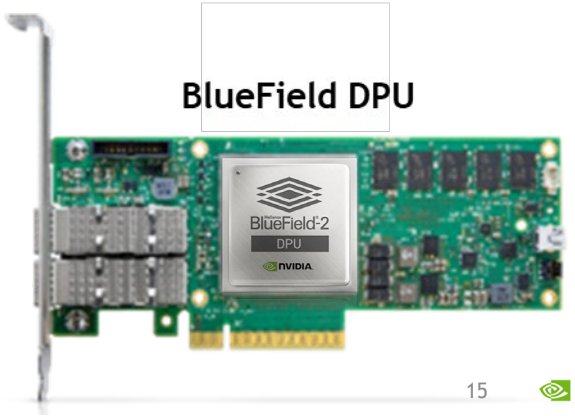
Physical Local NVMe Storage



NVMe SNAP Drive Emulation



Local Physical Storage to Hardware Emulated Storage

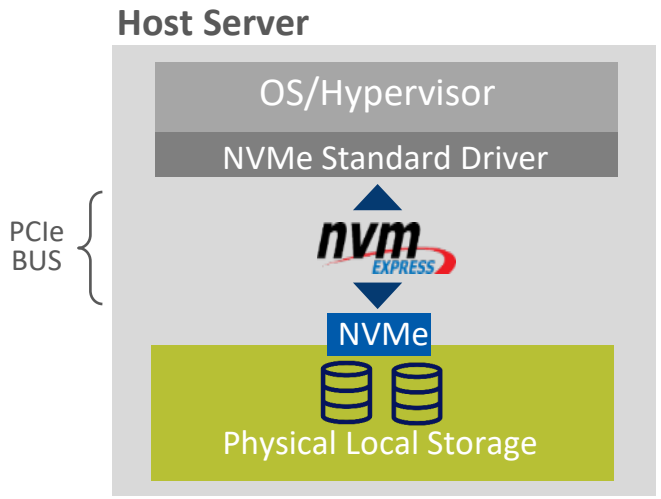


https://nvidianews.nvidia.com/news/mellanox-introduces-breakthrough-nvme-snapTM-technology-to-simplify-composable-storage

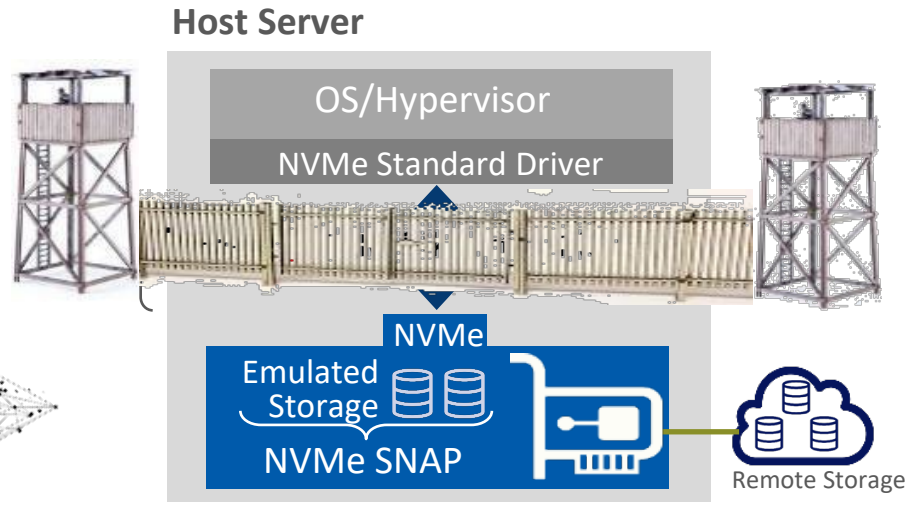
BLUEFIELD SNAP (NVME EMULATION) ACCELERATOR

NVMe-oF Initiator Acceleration

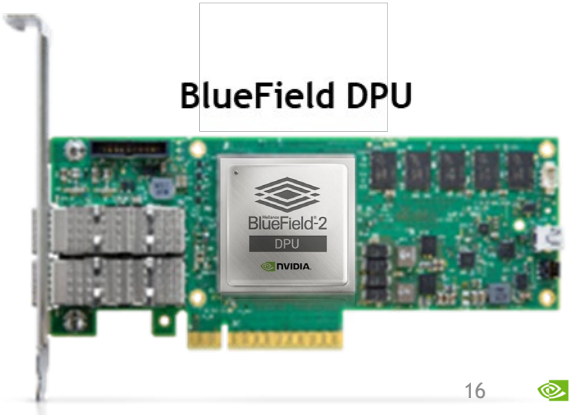
Physical Local NVMe Storage



NVMe SNAP Drive Emulation



Local Physical Storage to Hardware Emulated Storage

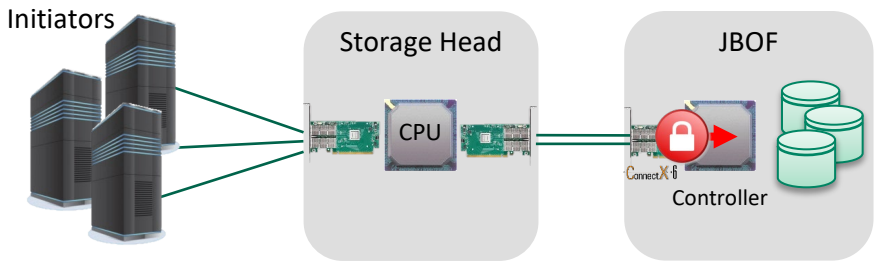


https://nvidianews.nvidia.com/news/mellanox-introduces-breakthrough-nvme-snapTM-technology-to-simplify-composable-storage

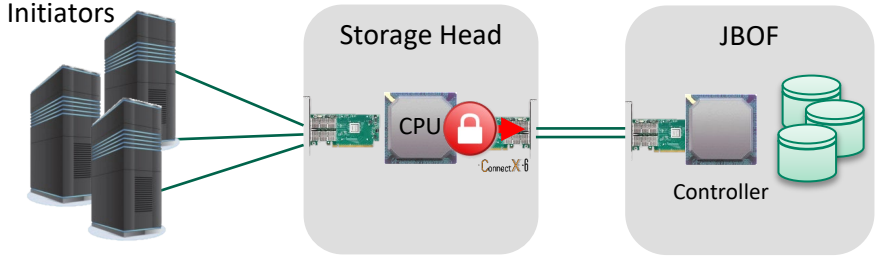
BLUEFIELD DPU ENCRYPTION ACCELERATORS

Secure Data Path and Data at Rest

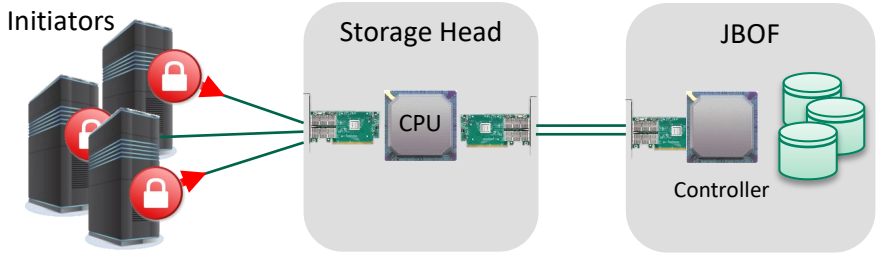
- At the Target JBOF
  - Secure data at rest on drives with AES



- At the Front-End Controller
  - Secure data on drives and in-flight to JBOFs and AFAs with AES, TLS, MACsec, and IPsec



- At the Initiator (Server)
  - Secure data across entire storage data path
  - Owner of the data controls the keys



https://docs.nvidia.com/networking/display/BlueFieldDPUOSLatest/IPsec+Functionality

EXAMPLE OF BLUEFIELD DPU STORAGE ACCELERATORS

From 40GB/s to 64GB/s Performance at Half the Size and Half the Power

VAST DF-5615,
VAST DF-5630

VAST DF-3015

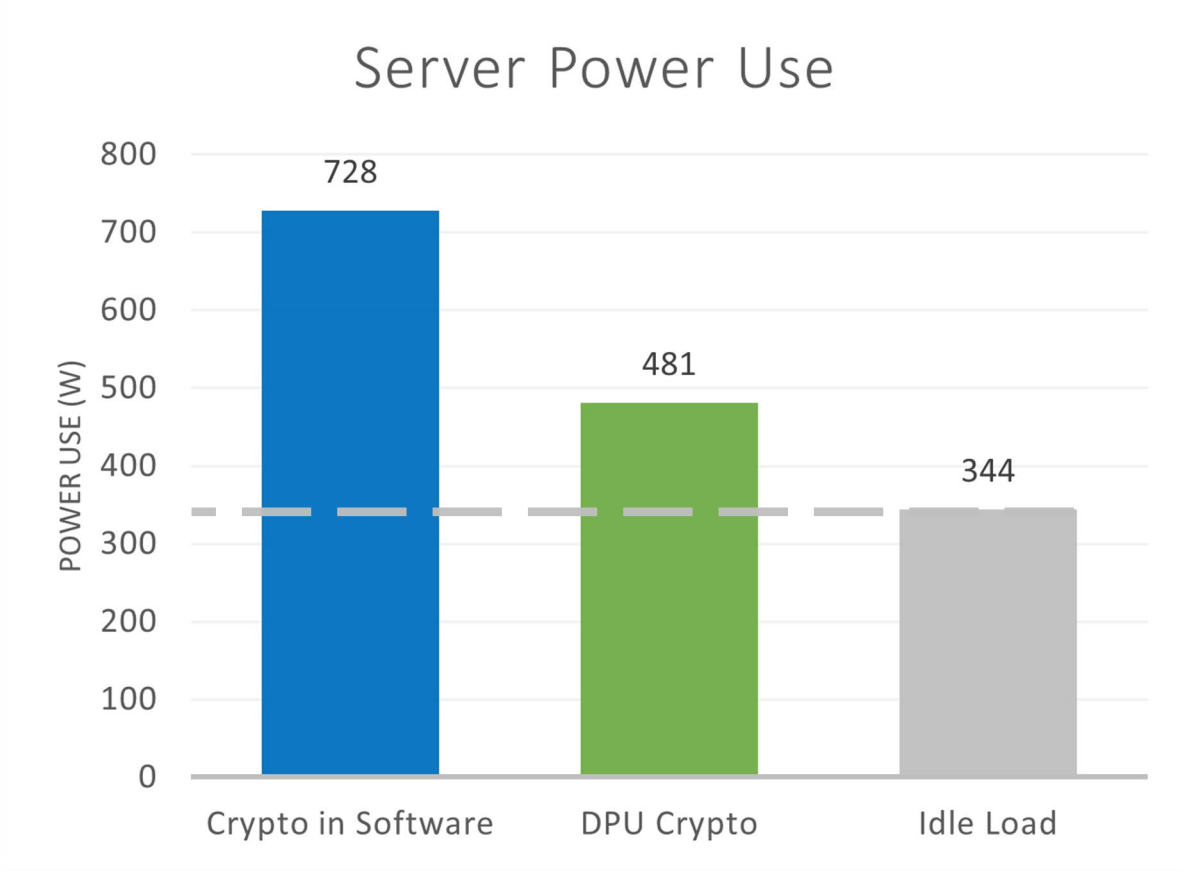
- x86 CPU based two Active/Active I/O Modules
- Power Consumption – 1200W Avg/ 1500W Max
- 2U Rackmount, 85lb

- BlueField DPU based two Active/Active I/O Modules
- Power Consumption – 500W Avg/ 1000W Max
- 1U Rackmount, 50lb

EXAMPLE OF DPU STORAGE SECURITY ACCELERATORS

Encrypting IPSec traffic at 100GbE line rate w BlueField

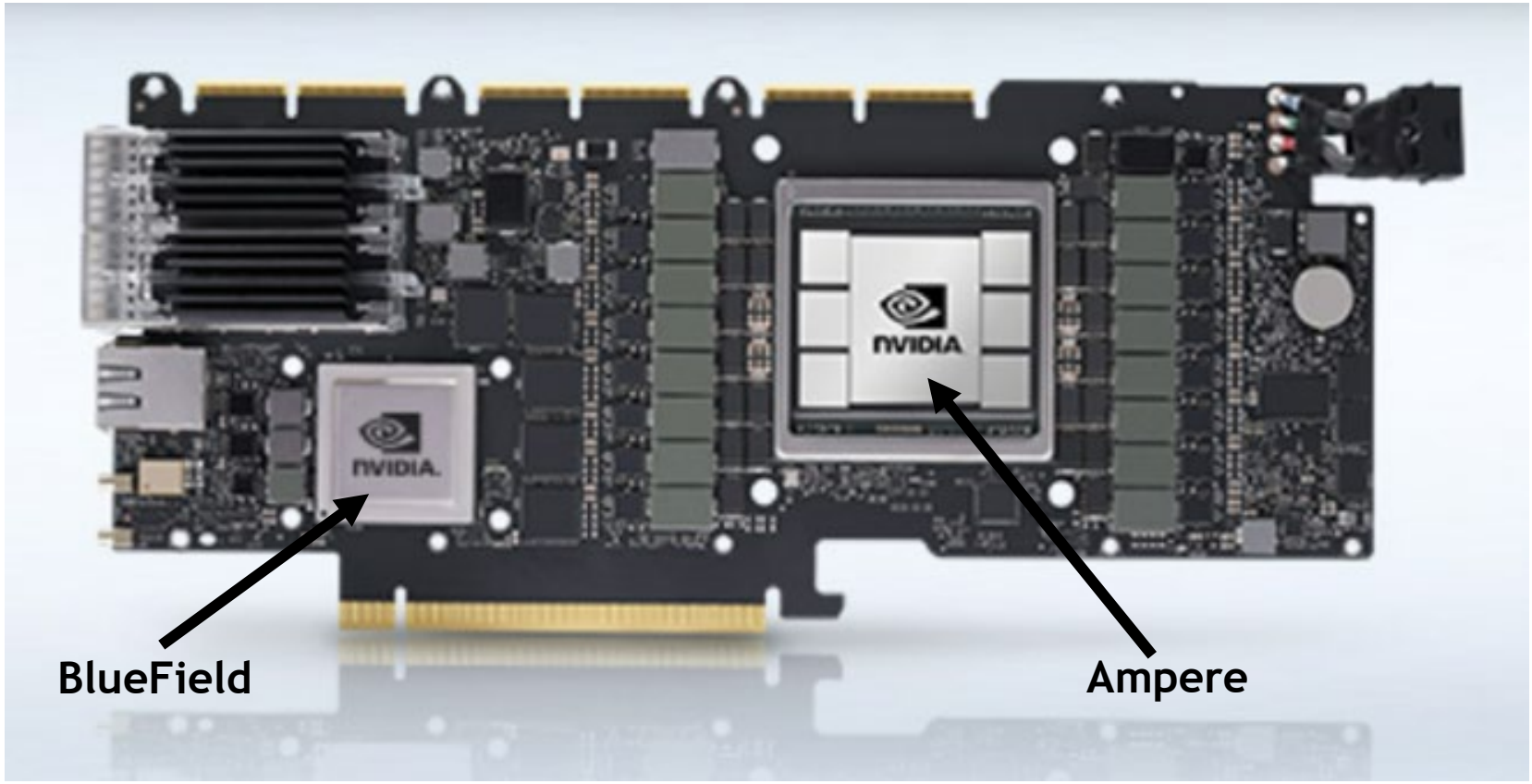
IPsec offload at 100% load	Power Use per server
IPSec crypto in software	728W
IPSec crypto in DPU	481W (34% savings)
3-year savings for **10,000** hosts	**$8.7M** (at $0.15/kWh)



Server with 2x Intel Xeon Platinum 8380 (Ice Lake) CPUs @2.30GHz (80 physical/160 HT cores total), RHEL 8.3
DPU is BlueField-2 VPI card with 2x100GbE/EDR ports and 16GB DRAM, Ubuntu 20.04. IPSec at 100GbE line rate

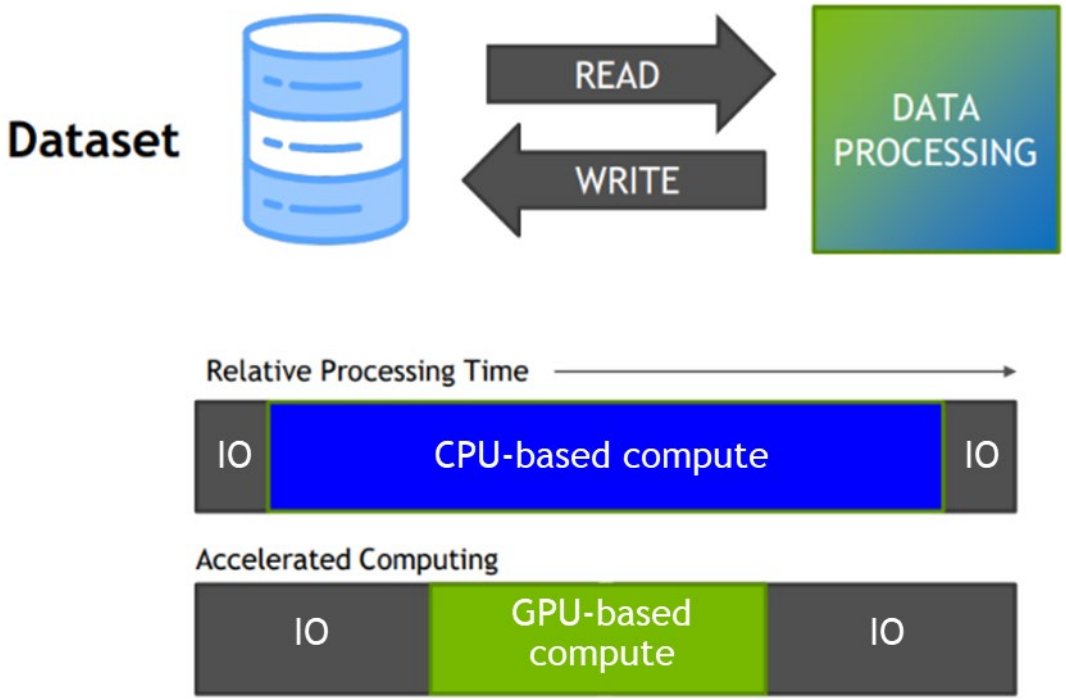
CONVERGED ACCELERATOR FOR STORAGE SOLUTIONS

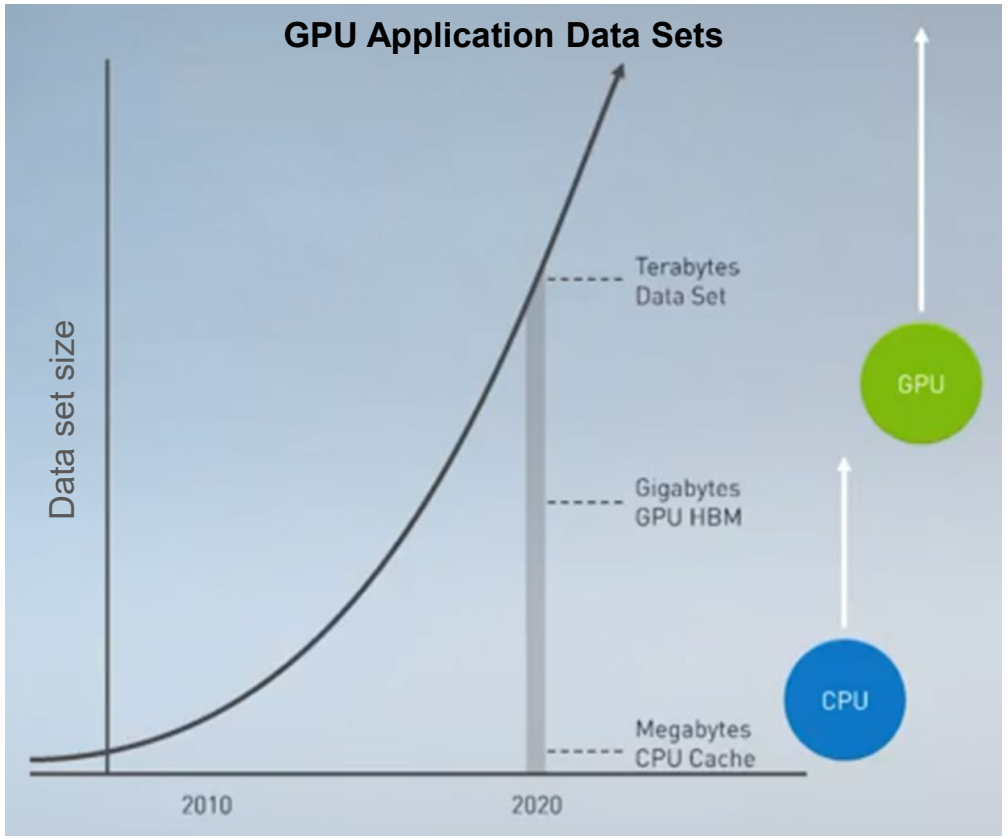
Combined DPU and GPU together



https://www.nvidia.com/en-us/data-center/products/egx-converged-accelerator/

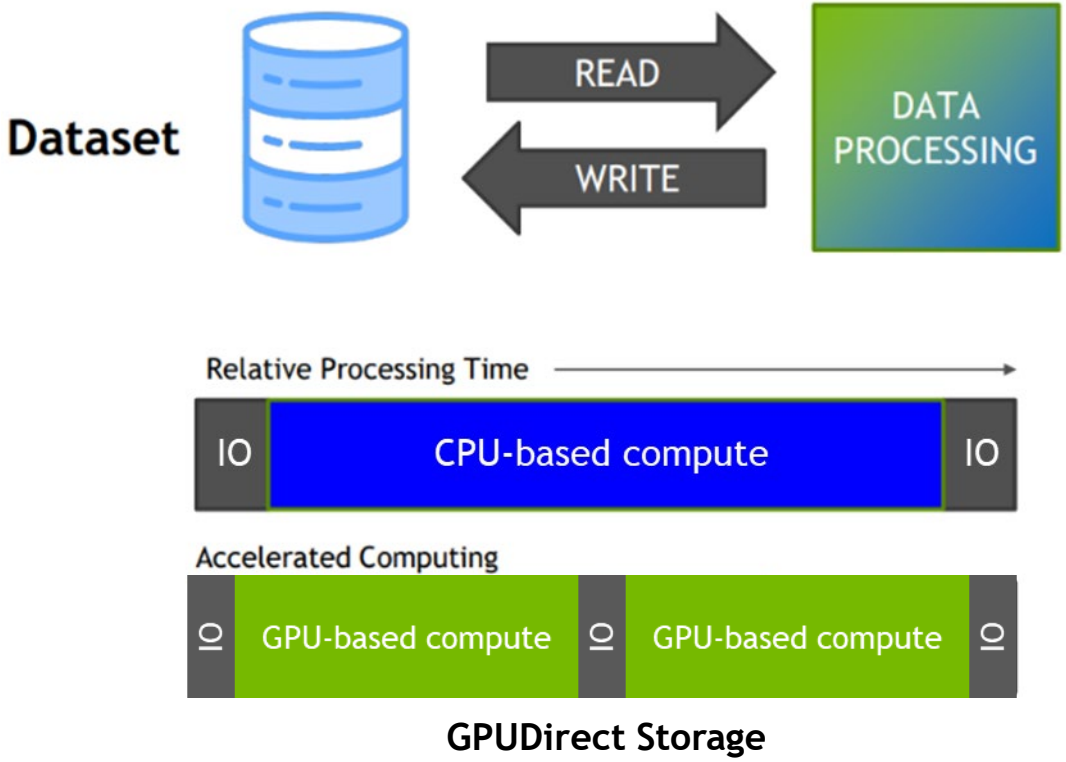
STORAGE IO CHALLENGES FOR GPU STORAGE



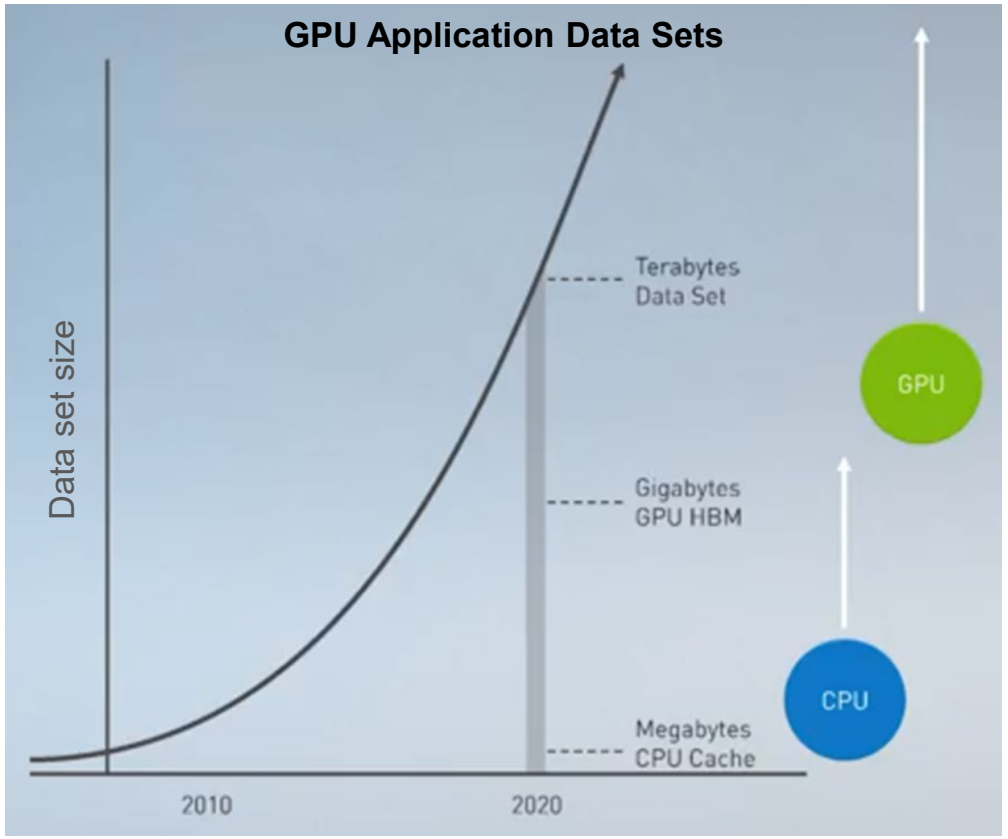


GPUDIRECT STORAGE (GDS)

BlueField RDMA Accelerators Enable Remote GDS

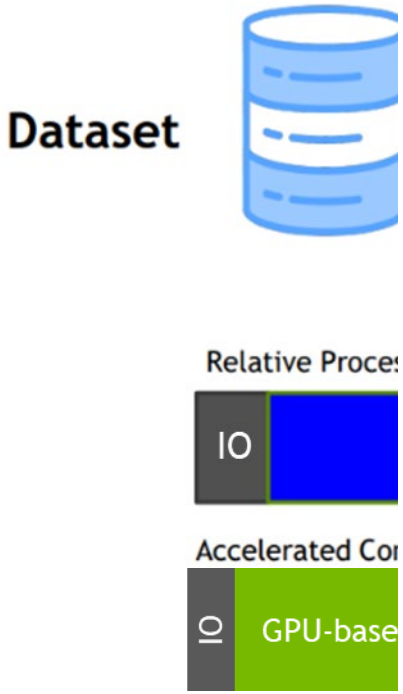


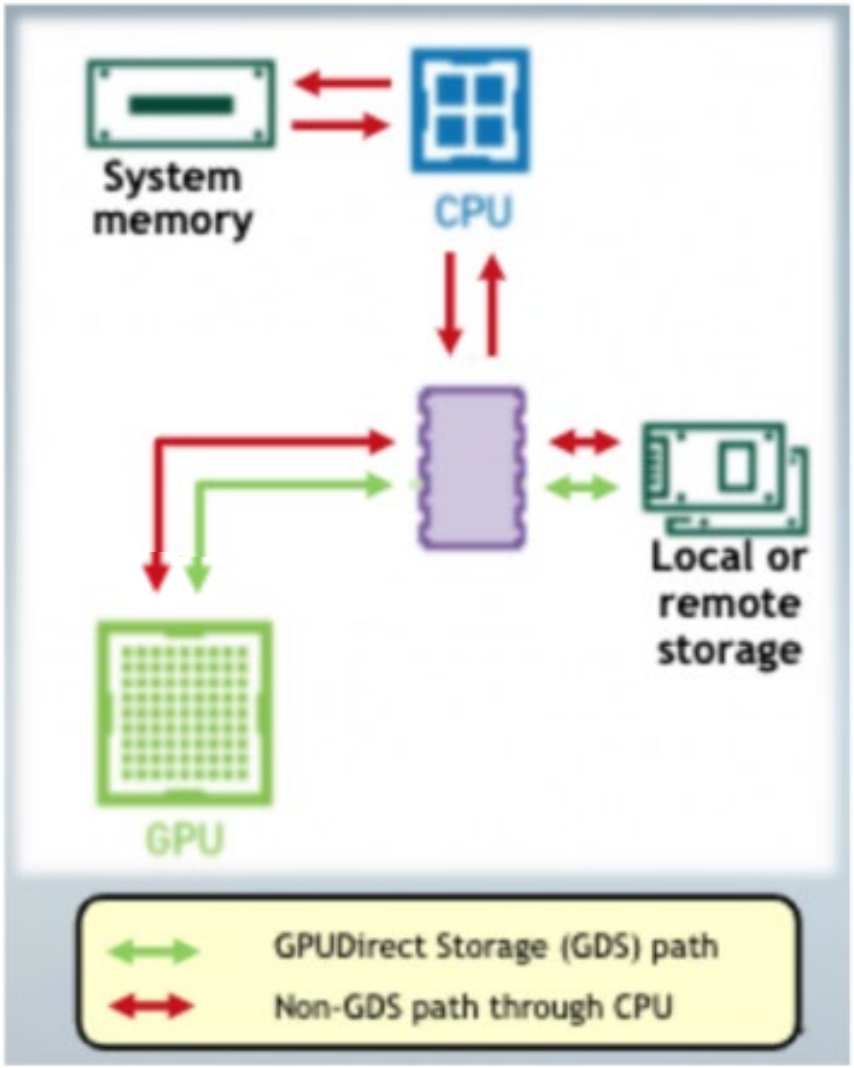
GPUDirect Storage

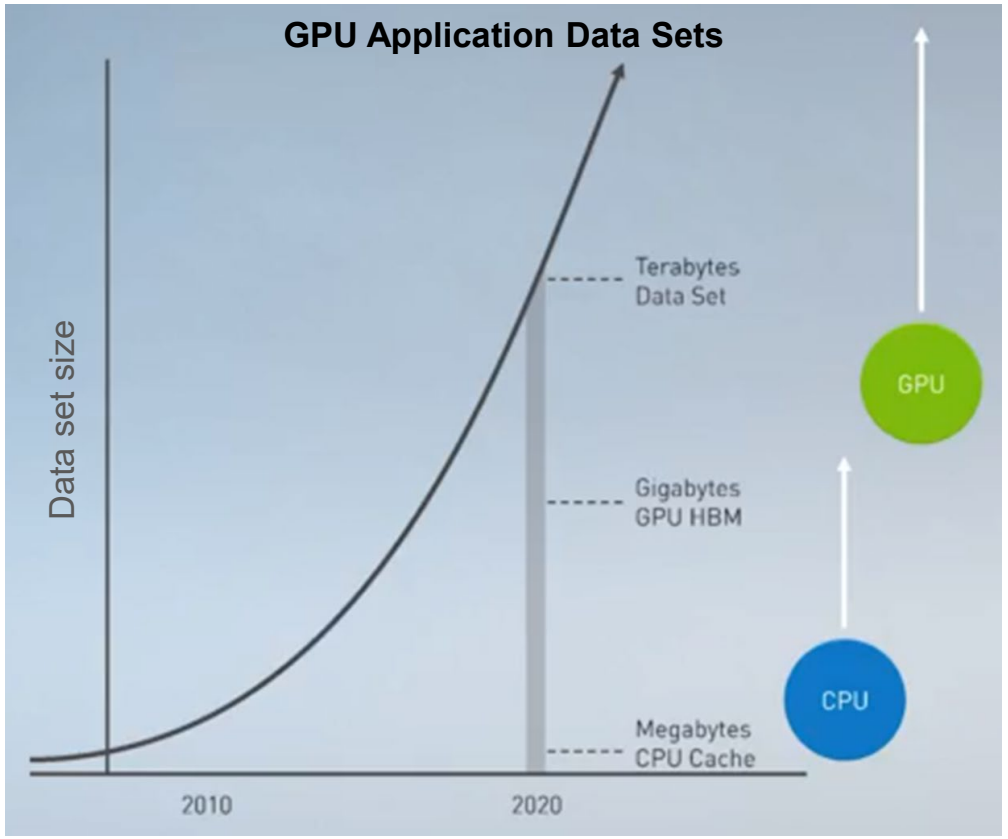


https://developer.nvidia.com/blog/gpudirect-storage/

GPUDIRECT STORAGE (GDS)



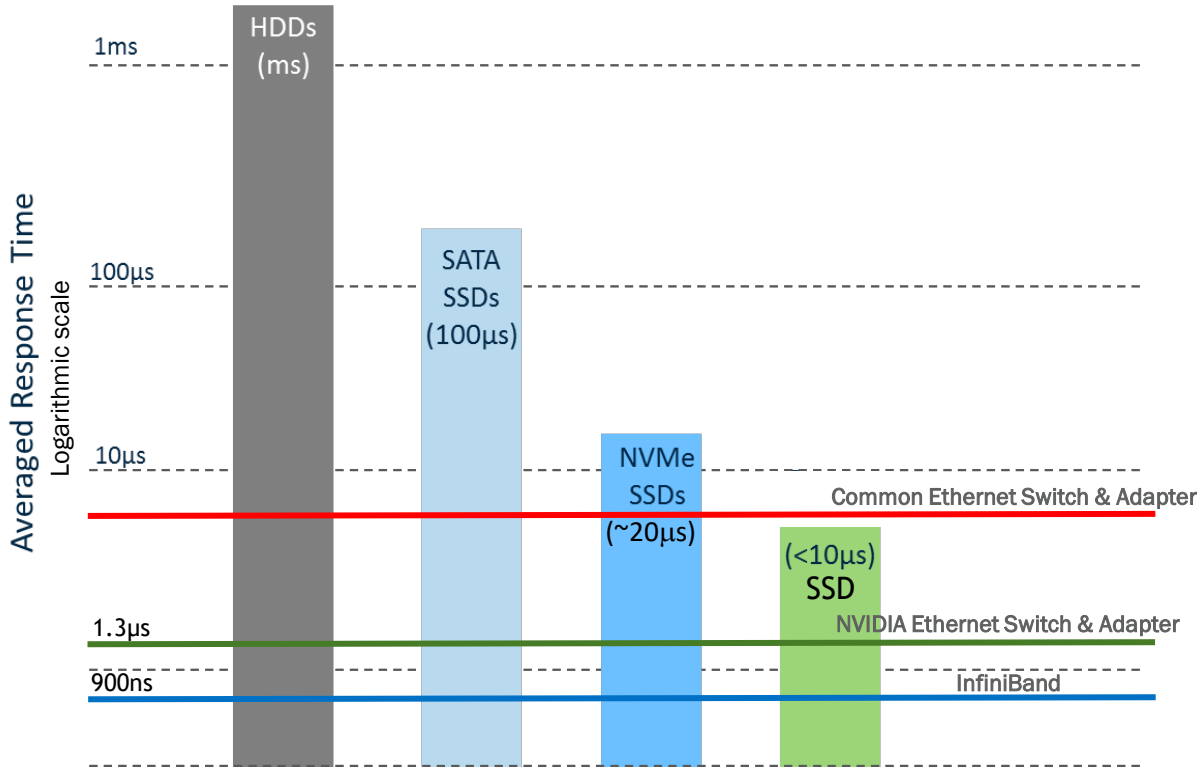




https://developer.nvidia.com/blog/gpudirect-storage/

IMPORTANCE OF LATENCY WHEN NETWORKING HIGH PERFORMANCE NVME BASED STORAGE

Network Switch Accelerators to Reduce Latency

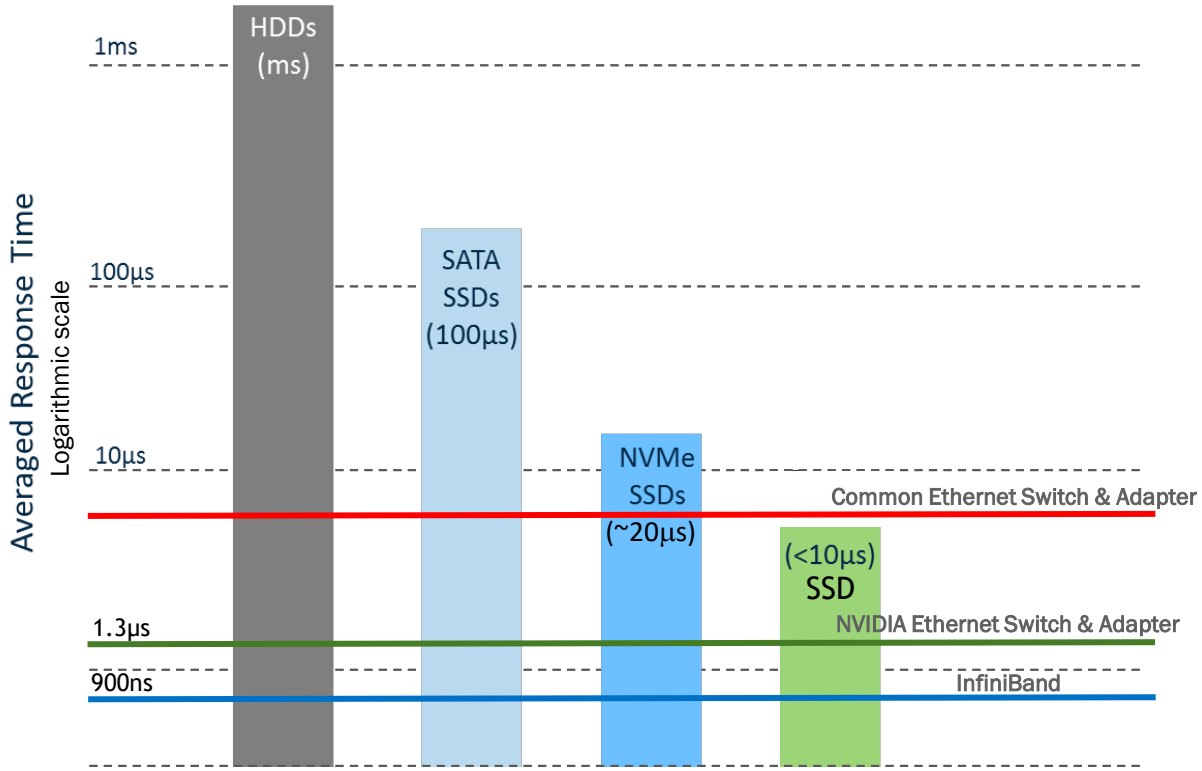


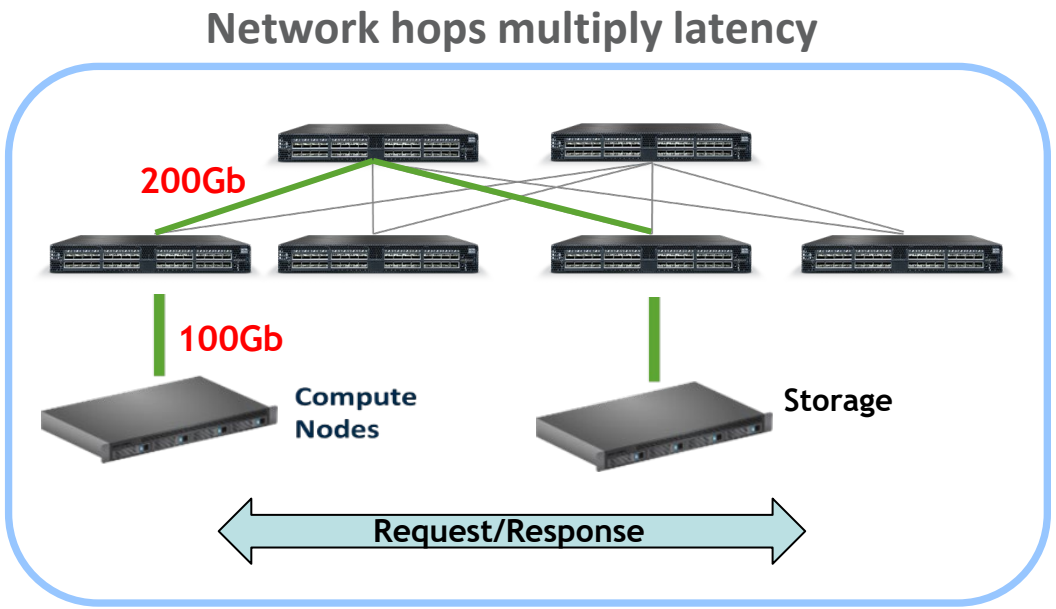
https://www.nvidia.com/en-us/networking/ethernet-switching/

https://www.nvidia.com/en-us/networking/infiniband-switching/

IMPORTANCE OF LATENCY WHEN NETWORKING HIGH PERFORMANCE NVME BASED STORAGE

Network Switch Accelerators to Reduce Latency





https://www.nvidia.com/en-us/networking/ethernet-switching/

https://www.nvidia.com/en-us/networking/infiniband-switching/

nvm™ EXPRESS

Peter Onufryk

Intel Fellow

http://nvme.org

NVMe® Technology Overview

G2M RESEARCH

Architected from the Ground Up for High Performance

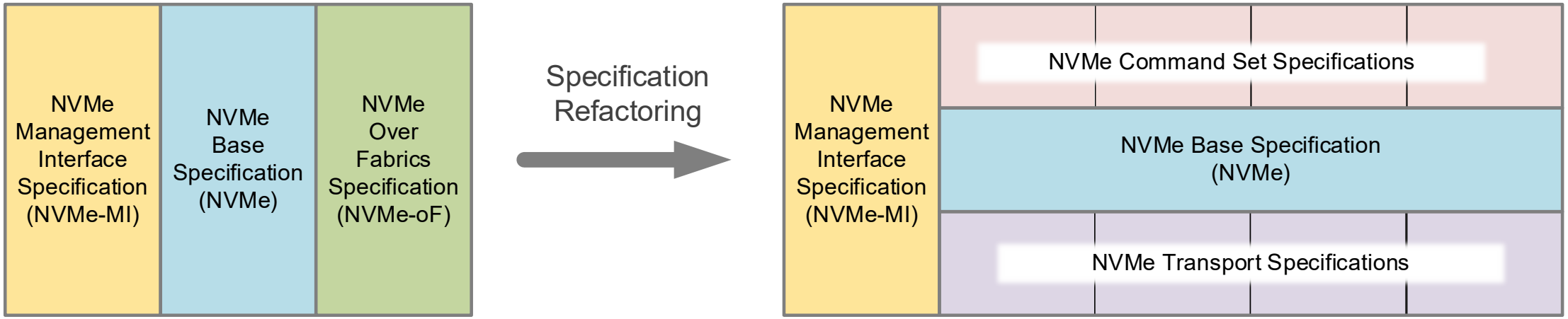
- **No practical limit on number of outstanding commands**
  - Up to 64K-1 queues each with up to 64K-1 outstanding commands
- **Supports many-core processors without locking**
  - Each core my allocated its own queues and interrupts
- **Efficient doorbell mechanism eliminates need for register reads in I/O path**
- **Fixed size 64B commands and 16B completions along with streamlined command sets enable fast and efficient command decode and execution**
- **Out of order command processing and data delivery**

NVMe Transports

2011	**PCIe**
2015	**NVM-oF™** NVMe/RDMA NVMe/FC
2018	**NVM-oF** NVMe/TCP

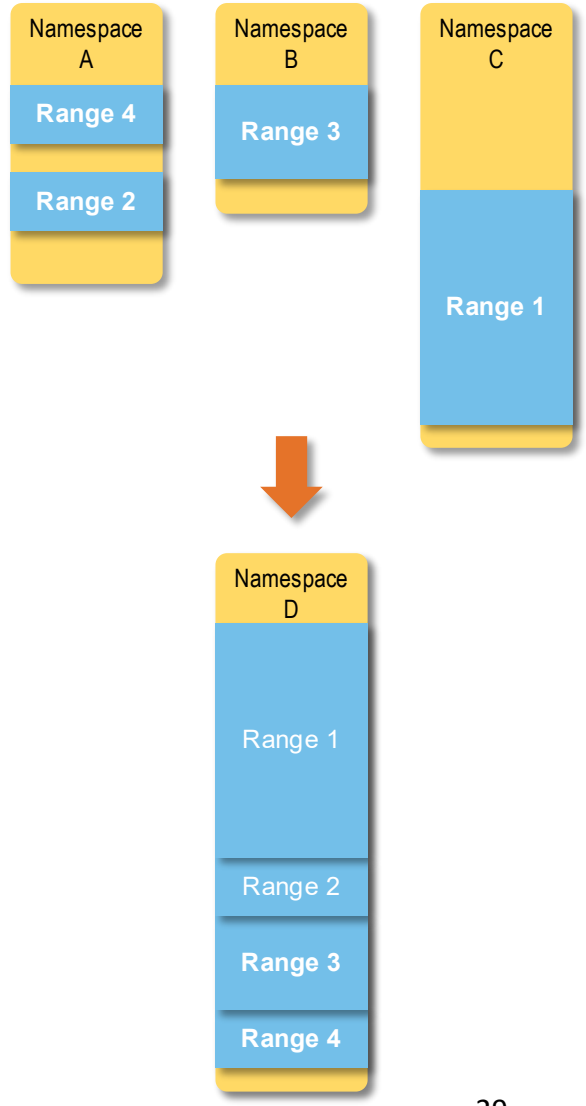
NVMe® 2.0 Family of Specifications

- Why Refactor?
  - Ease development of NVMe-based technology
  - Enable rapid innovation while minimizing impact to broadly deployed solutions
  - Create extensible spec infrastructure that enables the next phase of growth for NVMe technology



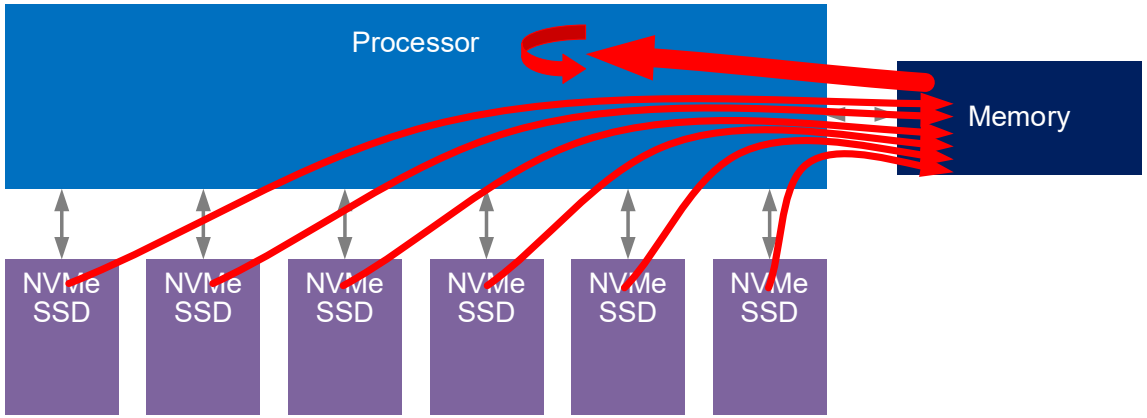
Cross Namespace Copy

- Copy command enhancement to copy data across namespaces
  - “Original” Copy Command
    - One or more source logical blocks ranges in a namespace to a single contiguous destination logical block range in the same destination namespace
  - “Enhanced” Copy Command
    - One or more source logical blocks in one **or more namespaces** to a single consecutive destination logical block range in a **destination** namespace
- Copy command does not reformat data
  - Logical block data and metadata format must be the same
  - End-to-End Data Protection type and size must be the same
  - Logical Block Storage Tag Mask and Storage Tag Size must be the same

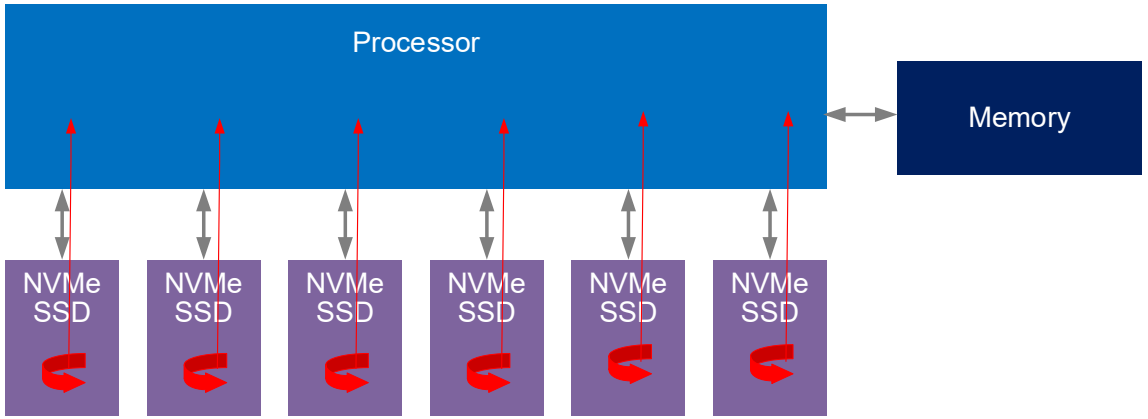


The Promise of Computational Storage

- **Higher performance and reduced latency** due to multiple SSDs operating in parallel
- **Reduced power** due to less data movement
- **Higher performance and reduced latency** due to elimination of processor I/O and memory bottlenecks



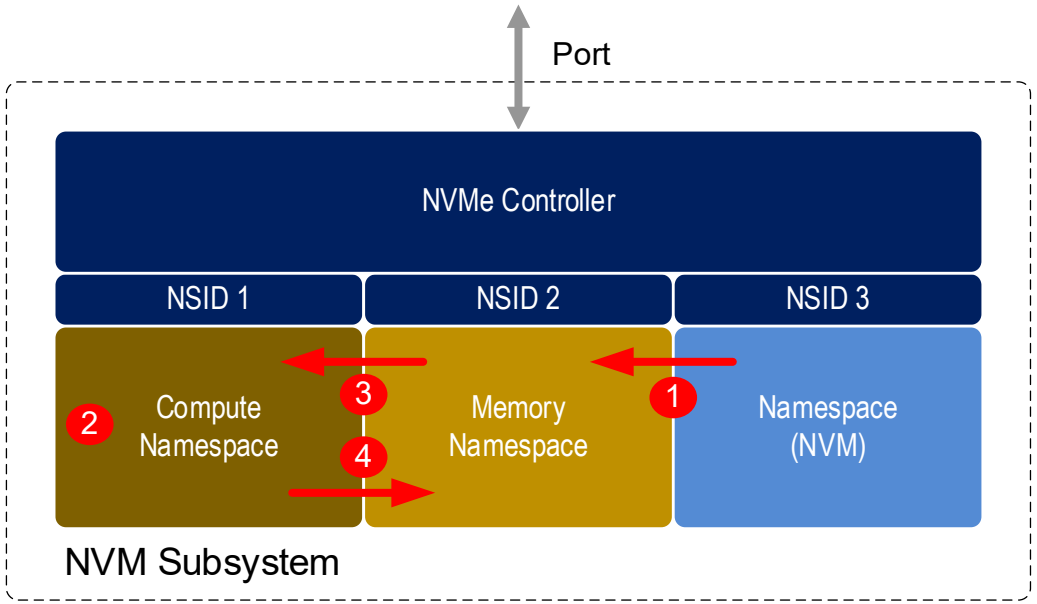
Data Processing in Main Processor



Data Processing in NVMe SSDs (Computational Storage)

Computational Programs

- Standardized framework for computational storage
- New command set for operating on Compute Namespaces
  - Fixed function programs
  - Downloadable programs
- Memory Namespaces and Subsystem Local Memory command set
  - Required for computational programs but is new general NVMe architectural element
  - Mechanism to copy to and from any other type of NVM namespace to memory namespace

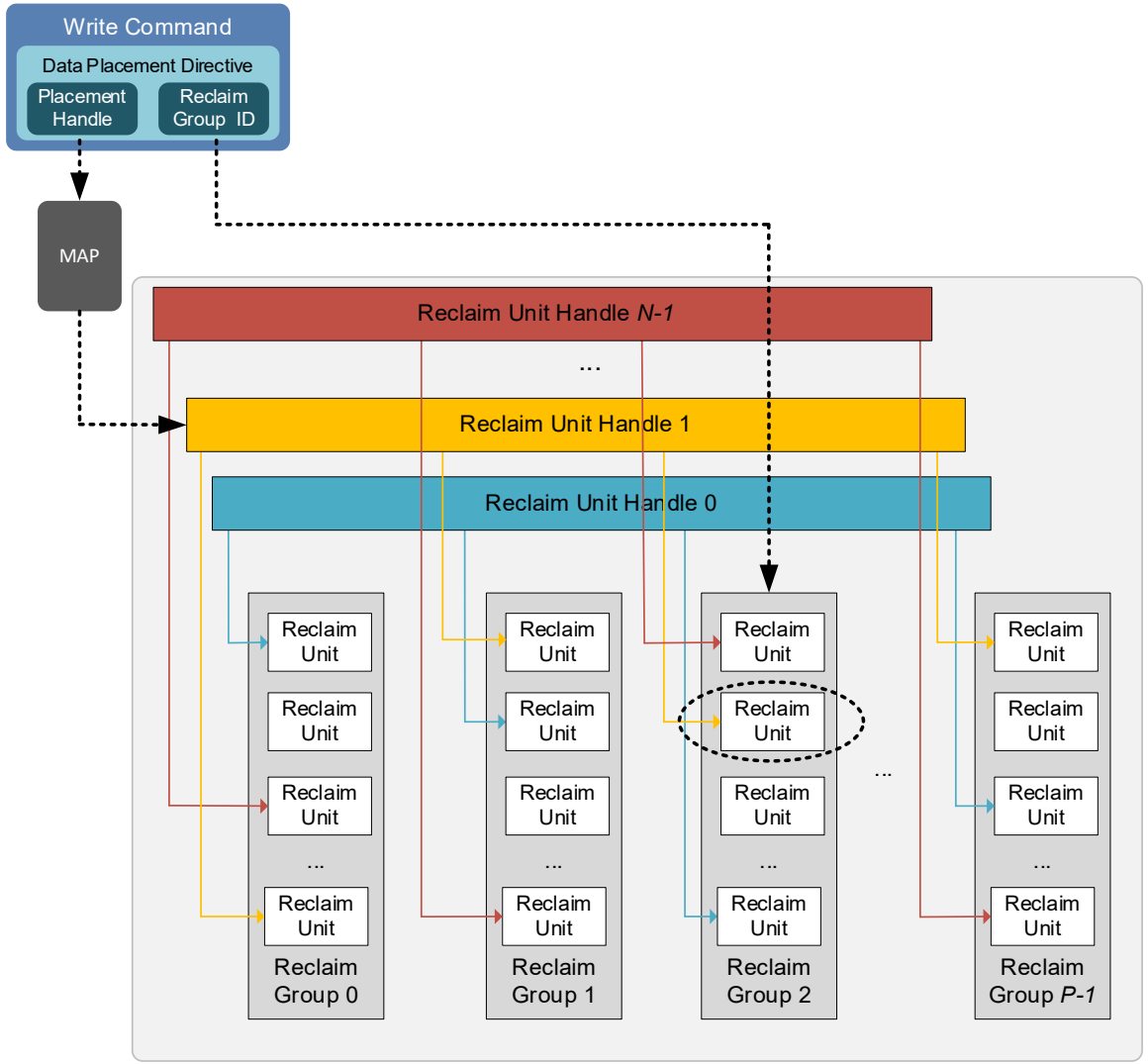


Example Operation:

1. Read data from NVM namespace into memory namespace
2. Execute program associated with computational namespace
3. Program reads data from memory namespace
4. Program stores result into memory namespace

Flexible Data Placement

- Enhancement to the NVM Command Set to enable host guided data placement
- Reclaim Unit (RU) is a unit of NVM storage that may be independently read, written, and erased
- A Reclaim Groups (RG) is an independent collection one or more RUs
  - Limited interference between RGs
  - Each RG has one or more Reclaim Unit Handles (RUH) that each point to an RU
- Data Placement Directive allows host to specify RG and RU of where to place written data



PLiOPS

EXTREME DATA PROCESSOR

Tony Afshary

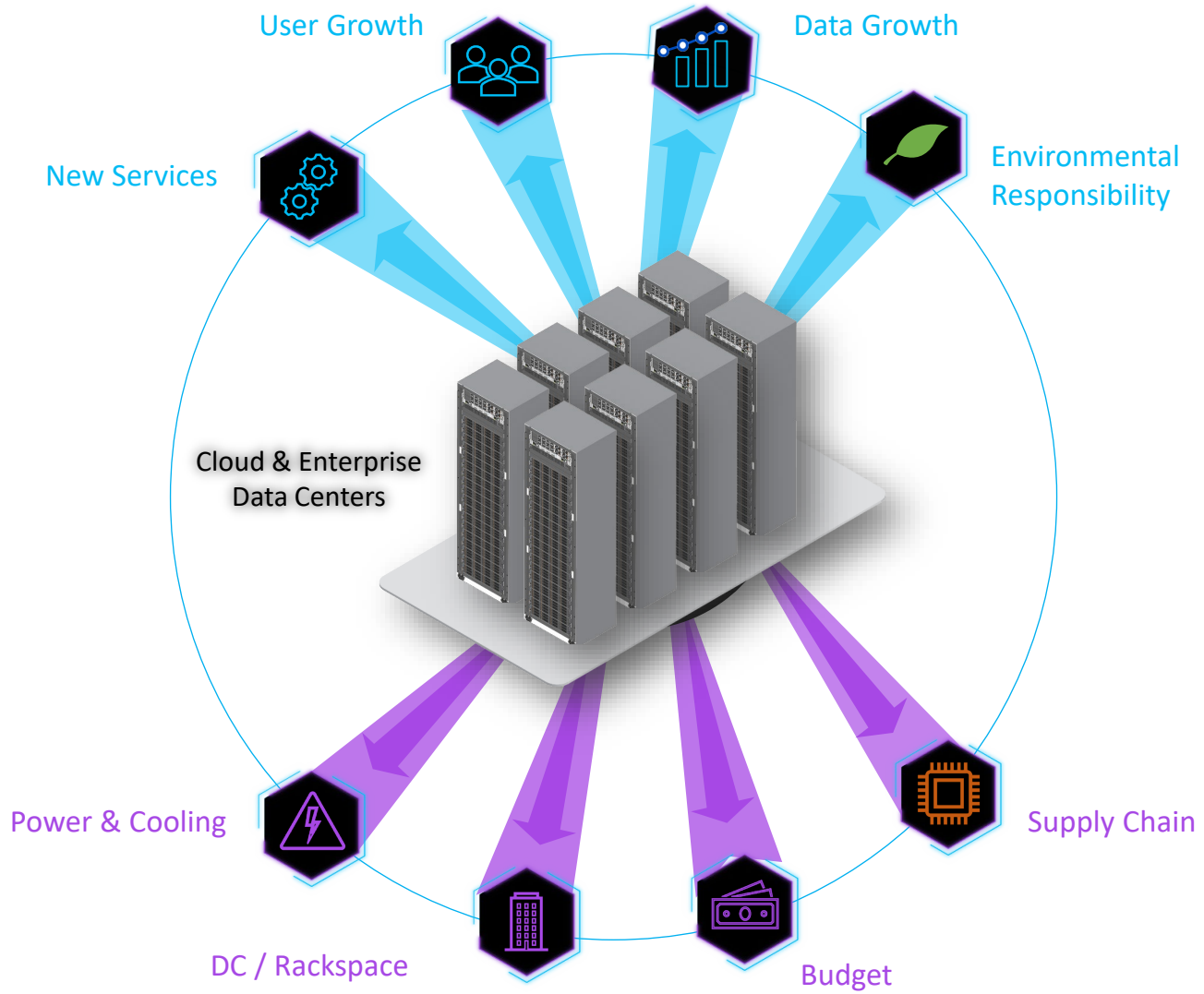
VP of Marketing

www.pliops.com

Expanding Demands and Constraints

Business Demands

Data Center Constraints



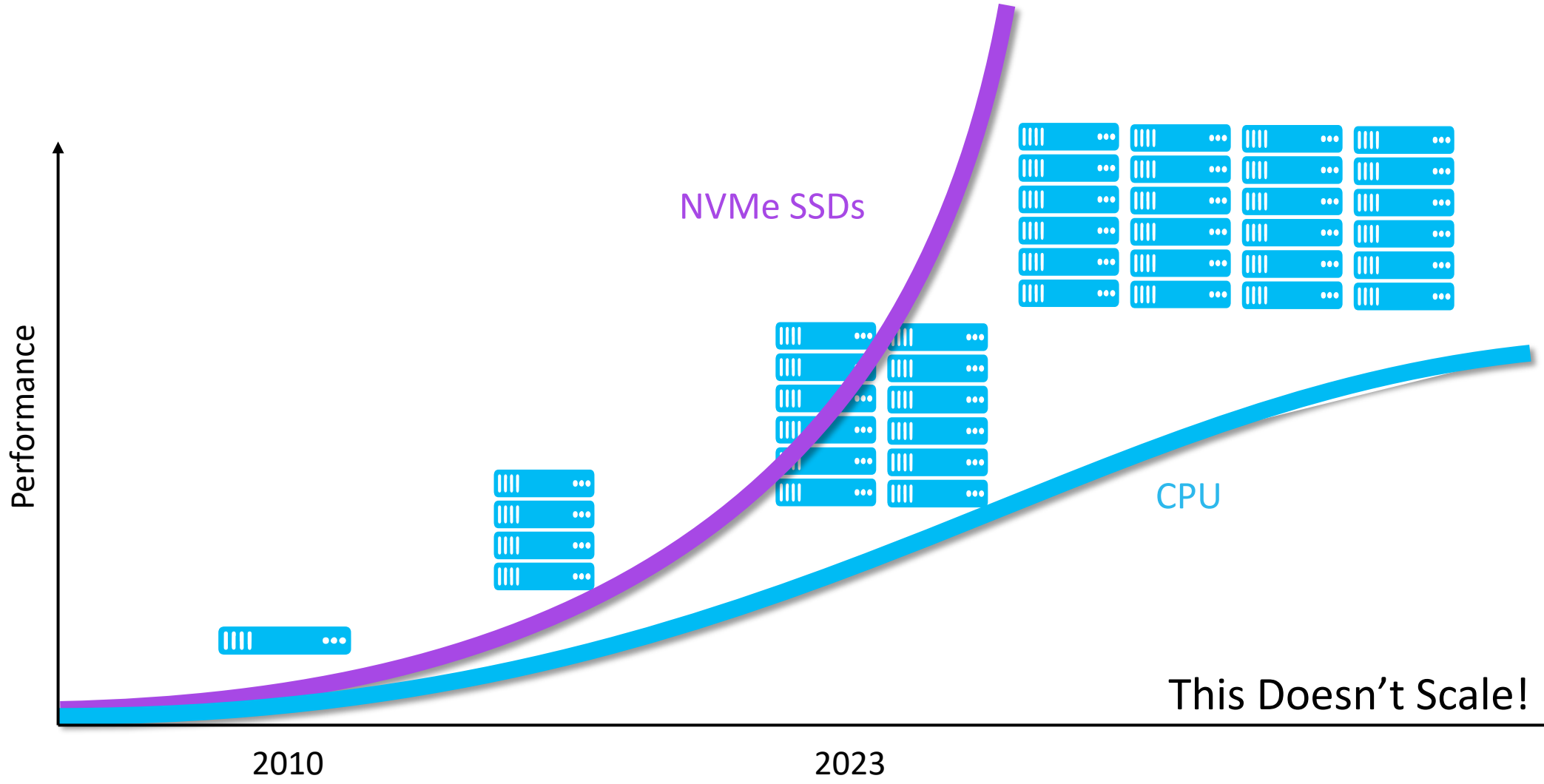
Current solutions don't adequately address

Supply constraints make it worse

Need to rethink data center architecture

PLiOPS
EXTREME DATA PROCESSOR

The Data-Compute Bottleneck



PLiOPS
EXTREME DATA PROCESSOR

Challenges With Broad SSD Adoption

Server Architectures Not Balanced

SSDs' 1000x increase in performance over HDD has not been matched by server advances

Amplified Data

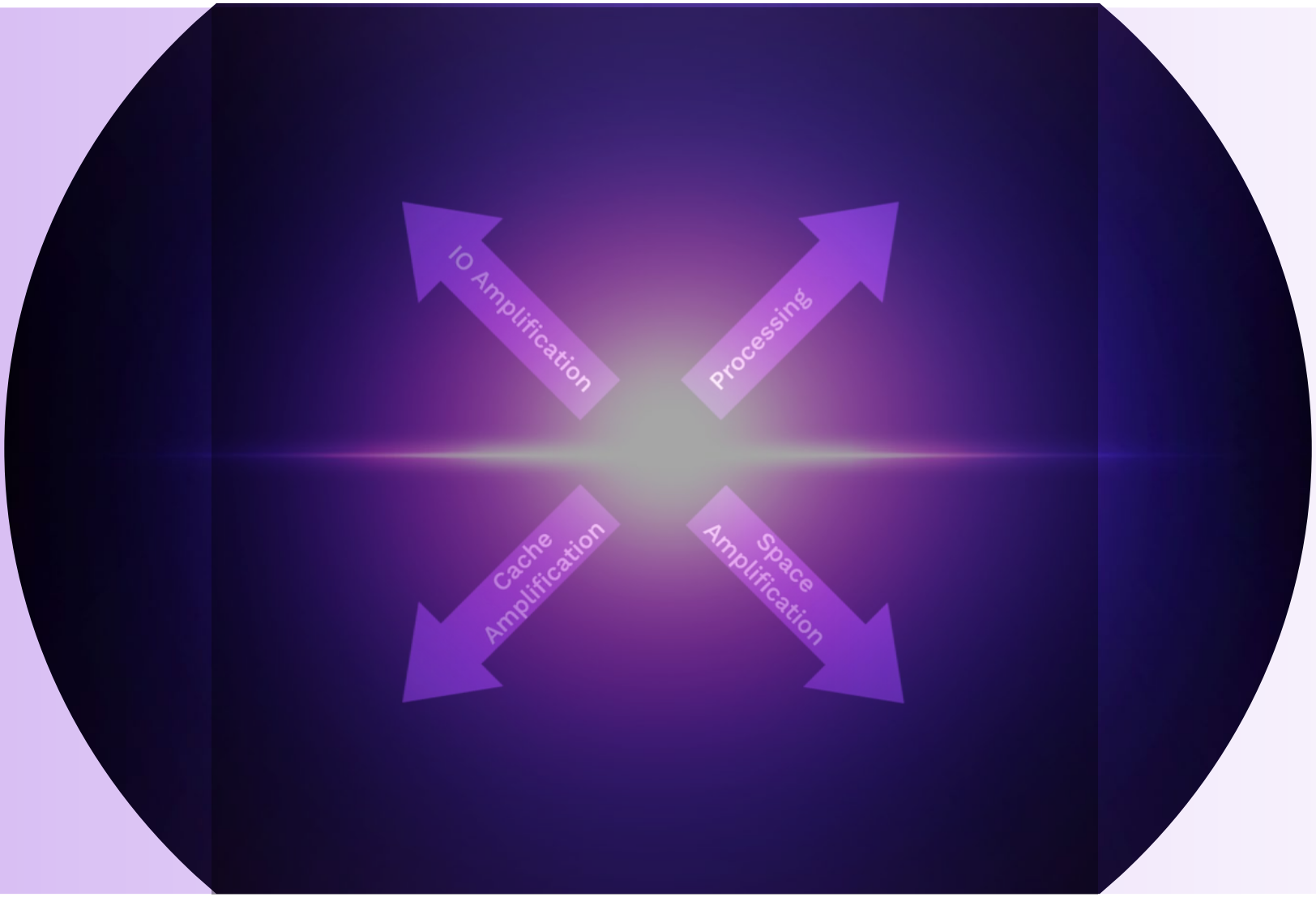
Software that uses SSDs amplifies reads and writes up to 100x, stored data up to 6x, crushing storage and network efficiency

System Reliability Compromised

Traditional RAID is rarely used with NVMe SSDs due to the huge performance penalty, requires costly workarounds

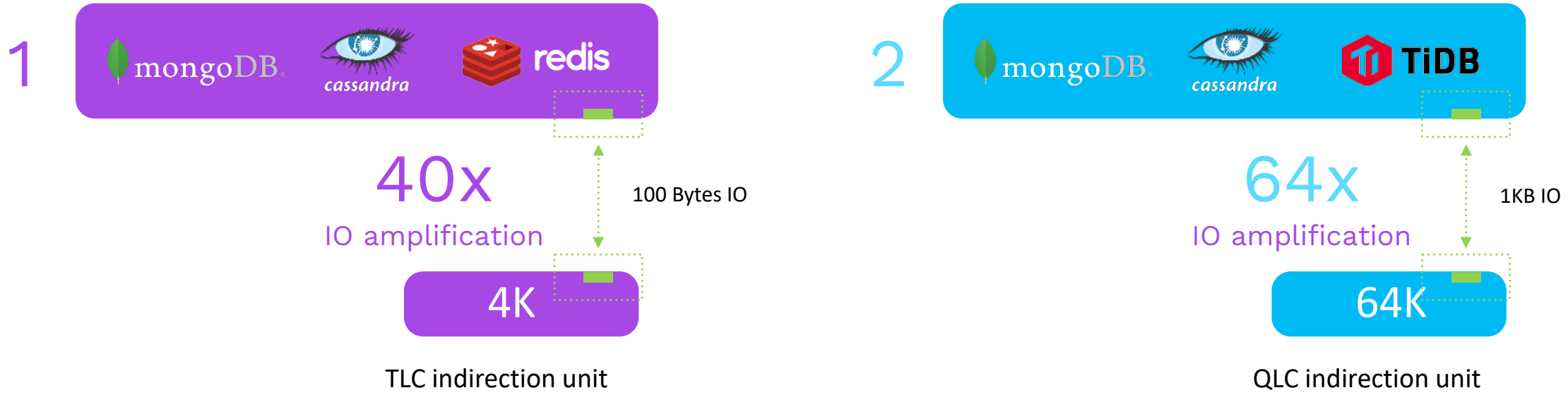
PLiOPS
EXTREME DATA PROCESSOR

Optimization Points and Tradeoff Relationships



PLiOPS
EXTREME DATA PROCESSOR

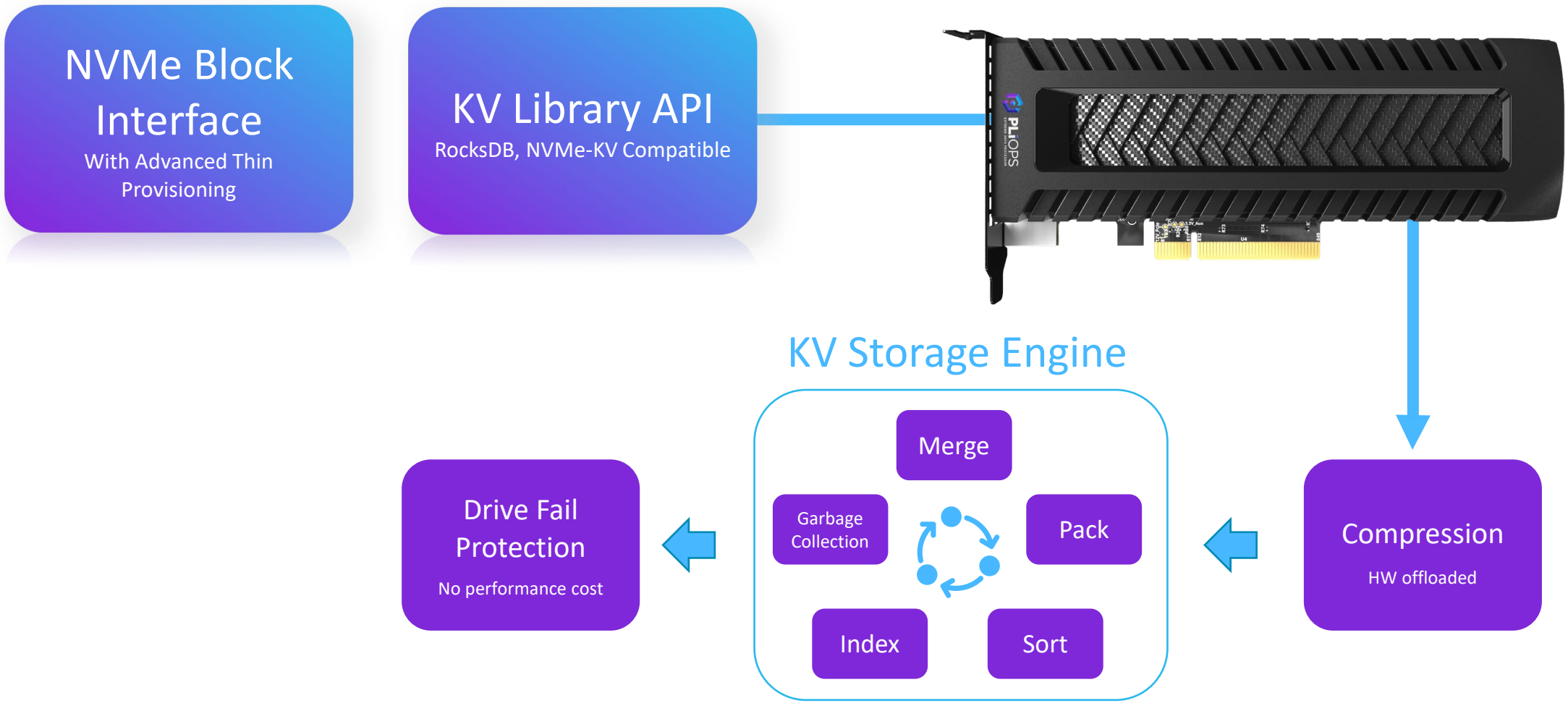
Application IO Amplification Challenge



- Impacts Network, Storage, SSD, CPU – Must Overprovision for This Extra Data Transfer and Processing
- Improving IO Amplification consumes CPU, cache or storage space.

PLiOPS
EXTREME DATA PROCESSOR

XDP Architecture



PLiOPS
EXTREME DATA PROCESSOR

XDP-*AccelKV*

World's 1st HW Key-Value Accelerator for Real Time Analytics and ML Training

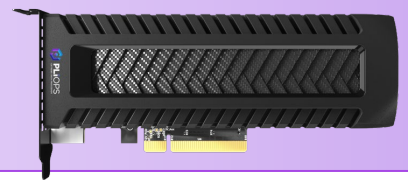
XDP-*AccelDB*

Best-In-Class Universal Database & SDS Accelerator.

XDP-*RAIDplus*

Best-In-Class Data Integrity and RAID+ Solution for NVMe and NVMeoF

Pliops XDP Data Services Platform



PLIOPS
EXTREME DATA PROCESSOR

XDP-*RAIDplus*

Best-In-Class Data Integrity and RAID+ Solution for NVMe and NVMeoF

Full NVMe Performance

Ultra-Fast SSD Rebuild

DWPD

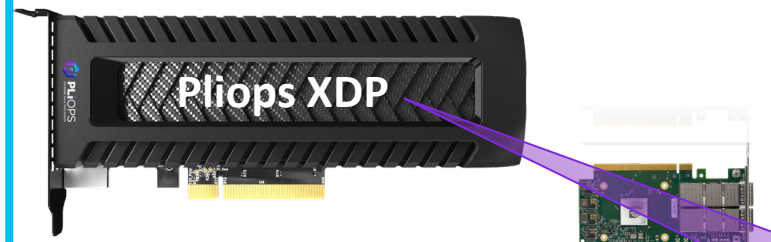
Enhanced SSD Endurance

Multiple Single Drive Failures

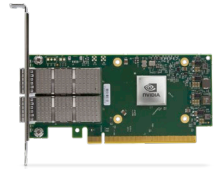
Fully PFAIL Protected

PLiOPS
EXTREME DATA PROCESSOR

200GbE NVMe-oF disaggregated storage deployment w/ Pliops



200GbE Switch



8x NVMe SSDs

TYAN
Transport SX
TS65A-B803

Drive Failure Protection (XDP-RAIDplus), Capacity Expansion, Extended Drive Life

TYAN
Transport HX
TN83-B8251

NVMe-oF Client

NVMe-oF Target

XDP-*AccelDB*

Best-In-Class Universal Database & SDS Accelerator.

Pliops XDP-AccelDB Data Service is the Best-In-Class Database Accelerator for SQL applications such as MySQL, MariaDB & PostgreSQL. It is also able to accelerate NoSQL applications including MongoDB and Software Defined Storage solutions such as Ceph.

3-5x
Higher Throughput

8x
Latency Reduction

30%
Better CPU Utilization

6X
Better Capacity Expansion

PLiOPS
EXTREME DATA PROCESSOR

XDP-*AccelKV*

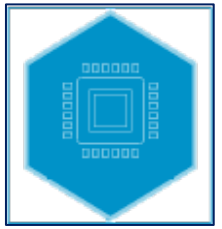
World's 1st HW Key-Value Accelerator for Real Time Analytics and ML Training

Pliops *XDP-AccelKV* Data Service is the Best-In-Class Key value accelerator solution for storage engines such as RocksDB, WiredTiger and other NoSQL storage engines. Being a native hardware key value accelerator, it can provide an order of magnitude higher performance than software-only solutions.

Pliops has introduced a RocksDB binary compatible library called XDP-Rocks which offloads the software functionality to hardware. It also has some key value enhancements such as key-value separation and LSM Bypass to achieve extreme performance acceleration and scalability

20x
Higher Throughput

100x
Latency Reduction



10x
Better CPU Utilization

6X
Better Capacity Expansion

PLiOPS
EXTREME DATA PROCESSOR

Panel Questions and Audience Surveys

Audience Survey Question #1

Are acceleration technologies something that your company is currently exploring (pick all that apply):

- We definitely see data processing acceleration, storage acceleration, and networking acceleration as being critical technologies: 46%
- We would most likely take advantage of acceleration directly in our organic data centers: 8%
- We are interested in utilizing cloud-based resources for acceleration: 15%
- We have teams actively exploring how acceleration can help us: 31%
- Sounds good, but not sure how it would apply to our workloads: 15%
- Don’t know/no opinion: 23%

Panel Question #1

What are the greatest challenges that companies are likely to encounter when exploiting and deploying acceleration, and what are some tips/best practices that they can utilize to avoid these issues?

- Peter Onufryk – NVMe
- Tony Afshary – Pliops
- Rob Davis - NVIDIA

Audience Survey Question #2

What types of workloads in your organization are candidates for acceleration (pick all that apply):

- Artificial Intelligence and Machine Learning: 25%
- Database Apps (SQL, NoSQL, unstructured search): 67%
- Business Intelligence: 17%
- Engineering and Scientific: 17%
- Simulation: 0%
- Other: 8%
- No opinion/don’t know: 17%

Panel Question #2

Do any of you see CXL (particularly memory pooling and memory sharing), or other similar “forward-looking” technologies as being important for storage, networking, or data processing acceleration, and how?

- Tony Afshary - Pliops
- Rob Davis – NVIDIA
- Peter Onufryk – NVMe

Audience Q&A

G2M COMMUNICATIONS

Effective Marketing & Communications with Quantifiable Results