EIDETICOM

Accelerate Everything.

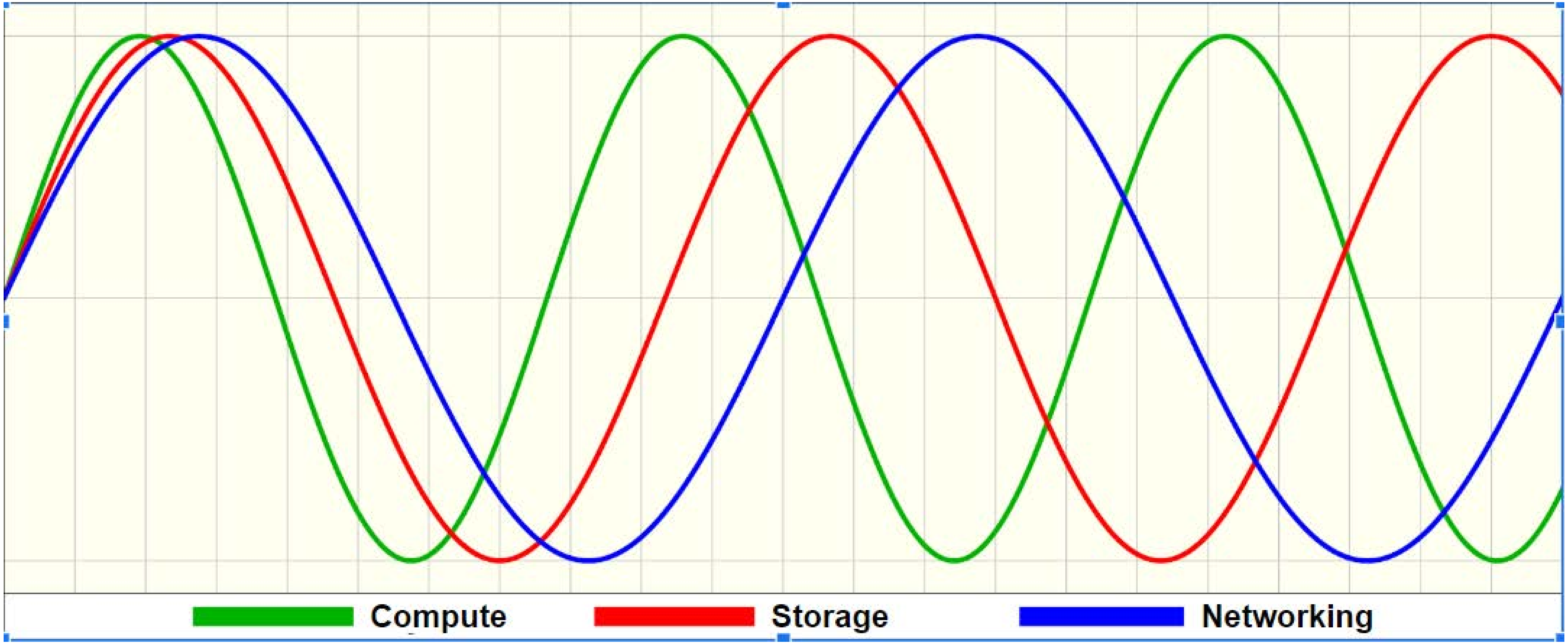
How NVM Express and Computational Storage can make your AI Applications Shine!

Stephen Bates, Chief Technology Officer, MSST 2019

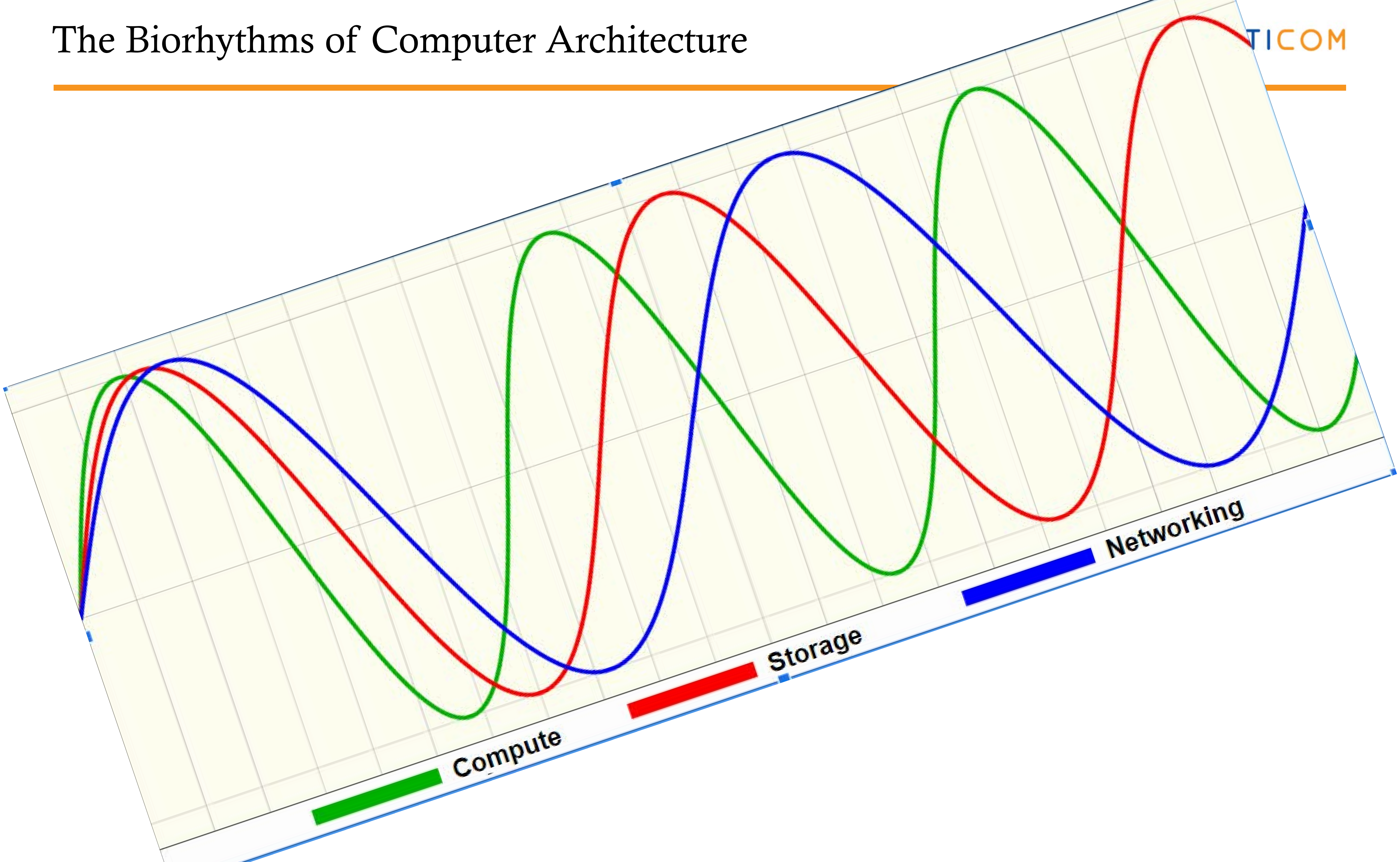
It's all about the software. Until you reach the limits of the hardware. Then it's all about the hardware [1].

[1] some geek, 2017.

The Biorhythms of Computer Architecture



The Biorhythms of Computer Architecture



We’ve Come a Long Way, Baby!

1955: 5MB, 1 million USD

2018: 1TB (1000000MB), 500 USD

We’ve Come a Long Way, Baby!

2019: 8TB, 240MB/s, 5ms

2010: 6TB, 2GB/s, 200us

HDDs are cheap (0.02/GB) but slow (throughput and latency). SSDs are expensive (0.20/GB) but fast (throughput and latency). For hot-data (which includes AI) SSDs looking more and more attractive.

AI has broken some things….

ILP tapped out + end of frequency scaling

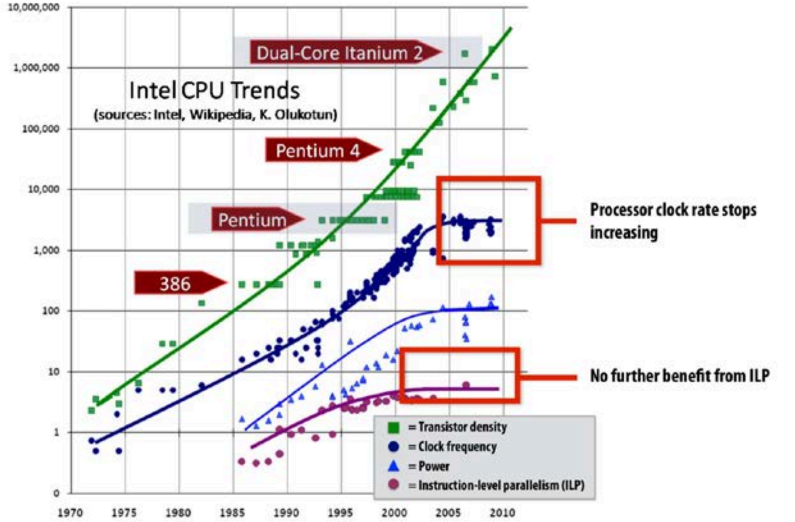
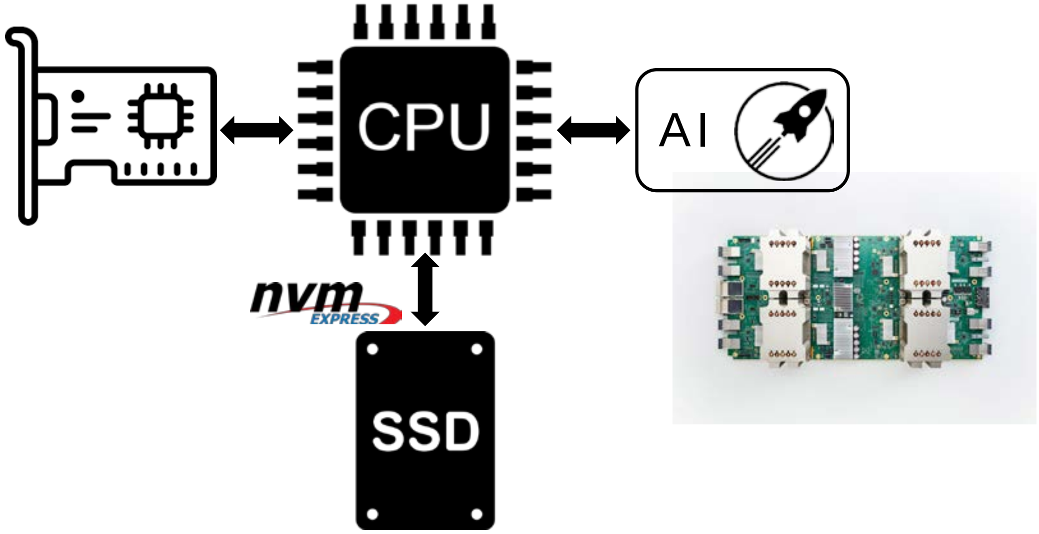


Image credit: "The free Lunch is Over" by Herb Sutter, Dr. Dobbs 2005

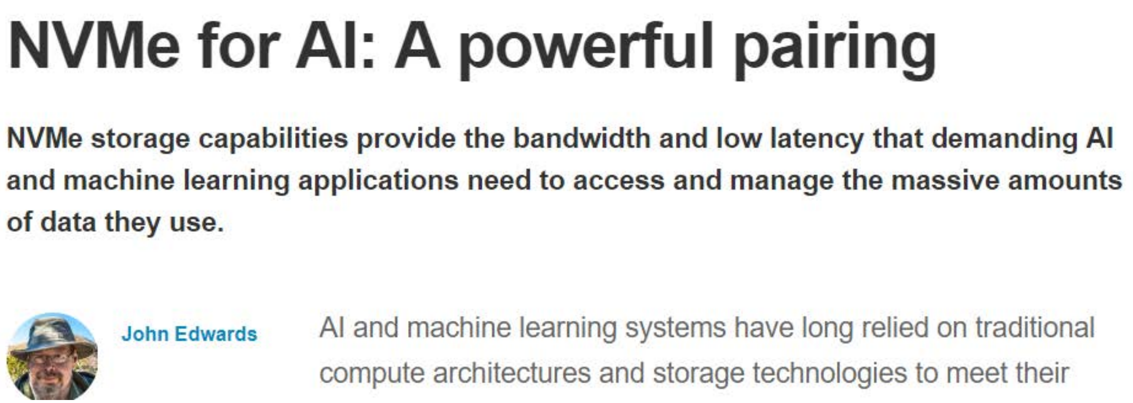
CMU 15-418/618, Spring 2016



AI algorithms suck on CPUs

AI algorithms shine on Accelerators (but data movement sucks)

AI *needs* NVM Express (apparently)

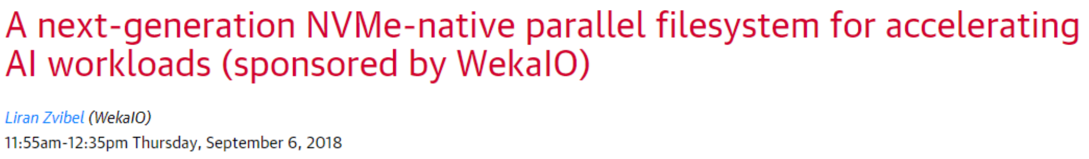


NVMe for AI: A powerful pairing

NVMe storage capabilities provide the bandwidth and low latency that demanding AI and machine learning applications need to access and manage the massive amounts of data they use.

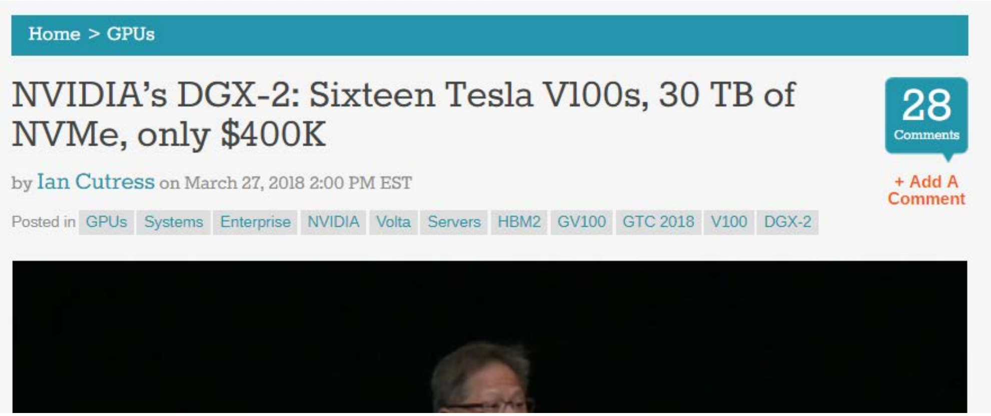
John Edwards

AI and machine learning systems have long relied on traditional compute architectures and storage technologies to meet their



A next-generation NVMe-native parallel filesystem for accelerating AI workloads (sponsored by WekaIO)

Liran Zvibel (WekaIO)
11:55am-12:35pm Thursday, September 6, 2018



Home > GPUs

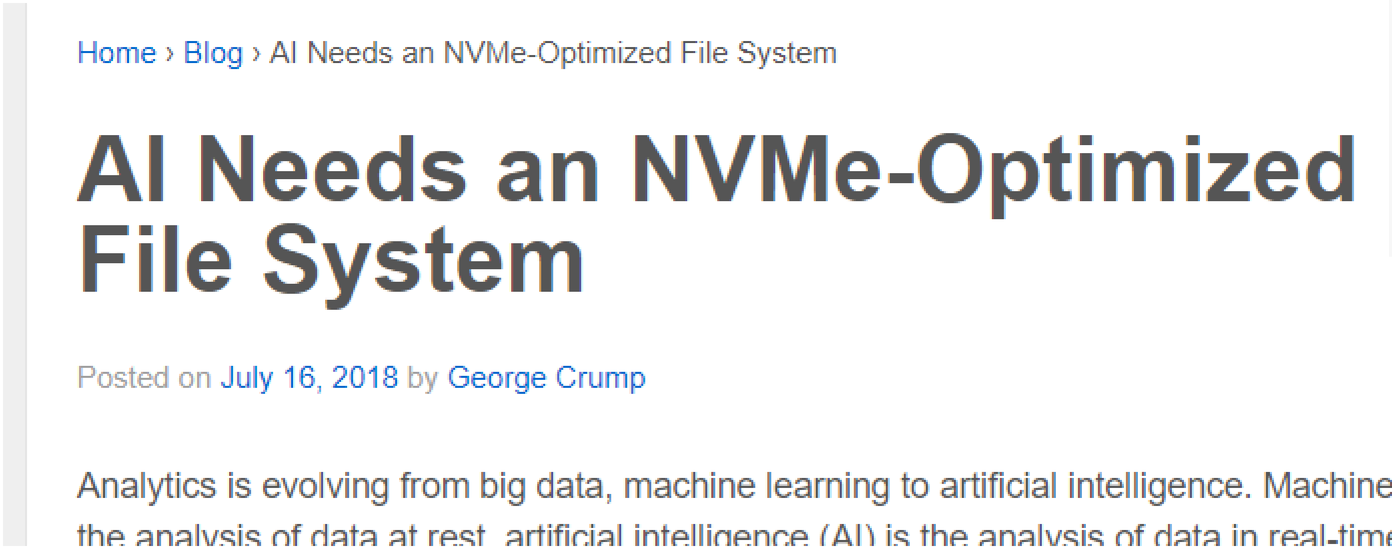
NVIDIA's DGX-2: Sixteen Tesla V100s, 30 TB of NVMe, only $400K

28 Comments

by Ian Cutress on March 27, 2018 2:00 PM EST

+ Add A Comment

Posted in GPUs Systems Enterprise NVIDIA Volta Servers HBM2 GV100 GTC 2018 V100 DGX-2



Home › Blog › AI Needs an NVMe-Optimized File System

AI Needs an NVMe-Optimized File System

Posted on July 16, 2018 by George Crump

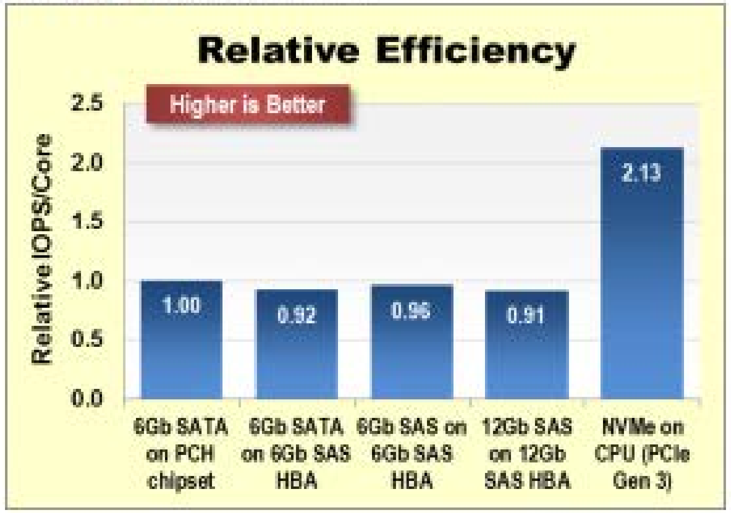
Analytics is evolving from big data, machine learning to artificial intelligence. Machine learning is the analysis of data at rest, artificial intelligence (AI) is the analysis of data in real-time. Machine

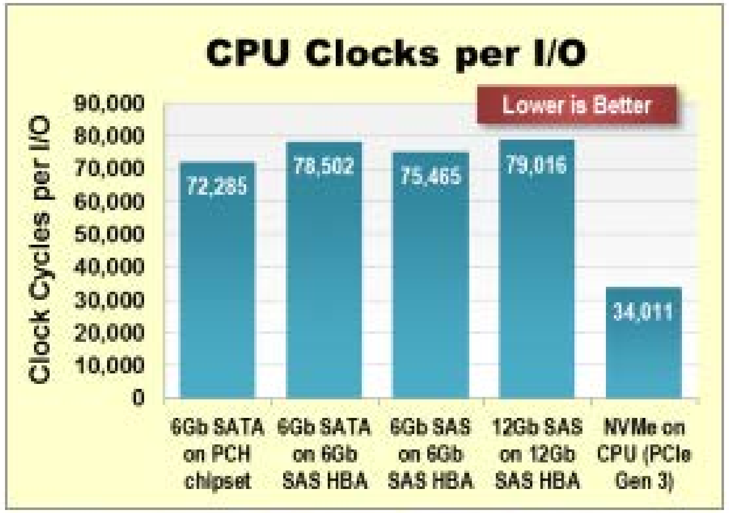
AI *needs* NVM Express (actually)

- **15000 image inferences per second per card (using ResNet50) at 1.3ms latency.**
- **15000 images = 3GB/s (@ 200kB/image)!!**
- **8 cards in a 2U server = 24GB/s (200Gb/s)!**

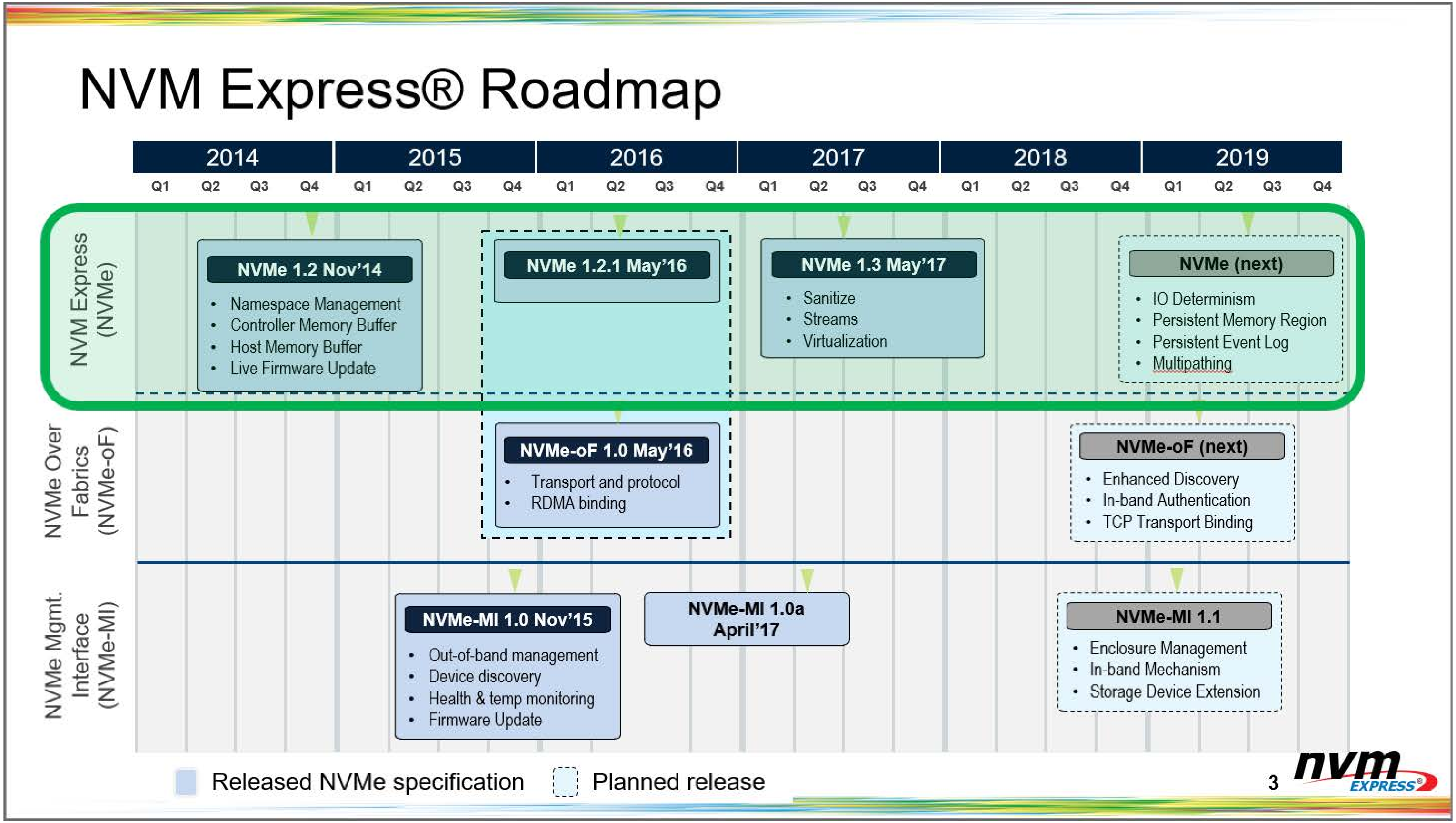
What is NVM Express?

- Initially a protocol that sat on PCIe and used to talk to Non-Volatile Memory (NVM)
  - Avoided vendor specific solutions for PCIe attached flash.
  - Inherently parallel, which is good for NAND and for multi-core CPUs
    - Out of order execution of commands on a queue.
    - Many queues allowed.
  - Lightweight (less CPU instructions per IO compared to SCSI)
  - Leveraged PCIe which comes directly out of every CPU worth caring about.





What is NVM Express?



NVM Express® Roadmap

NVMe 1.2 Nov'14

- Namespace Management
- Controller Memory Buffer
- Host Memory Buffer
- Live Firmware Update

NVMe 1.2.1 May'16

NVMe 1.3 May'17

- Sanitize
- Streams
- Virtualization

NVMe (next)

- IO Determinism
- Persistent Memory Region
- Persistent Event Log
- Multipathing

NVMe-oF 1.0 May'16

- Transport and protocol
- RDMA binding

NVMe-oF (next)

- Enhanced Discovery
- In-band Authentication
- TCP Transport Binding

NVMe-MI 1.0 Nov'15

- Out-of-band management
- Device discovery
- Health & temp monitoring
- Firmware Update

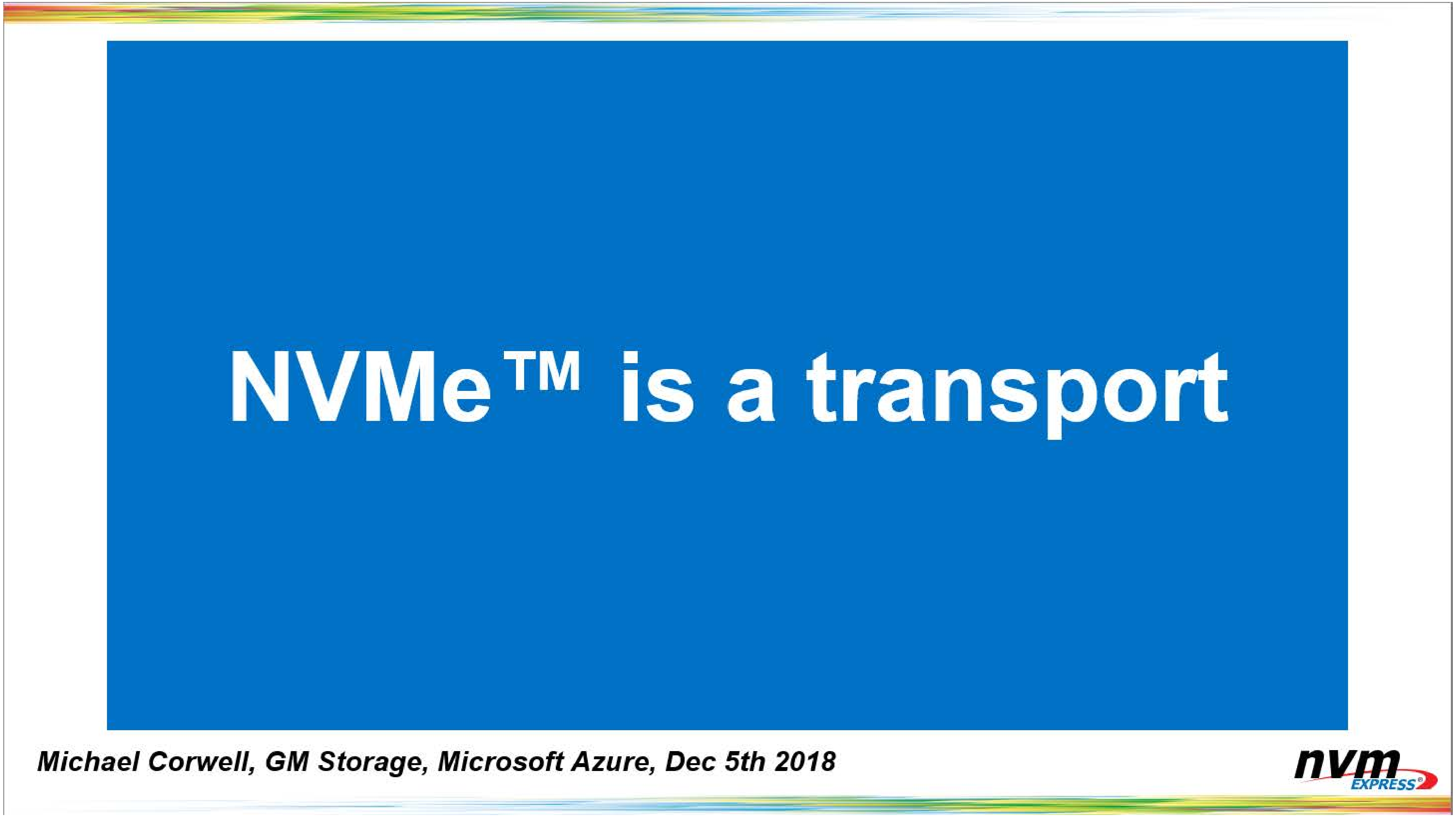
NVMe-MI 1.0a April'17

NVMe-MI 1.1

- Enclosure Management
- In-band Mechanism
- Storage Device Extension

▨ Released NVMe specification ⬚ Planned release

3



NVMe™ is a transport

Michael Corwell, GM Storage, Microsoft Azure, Dec 5th 2018

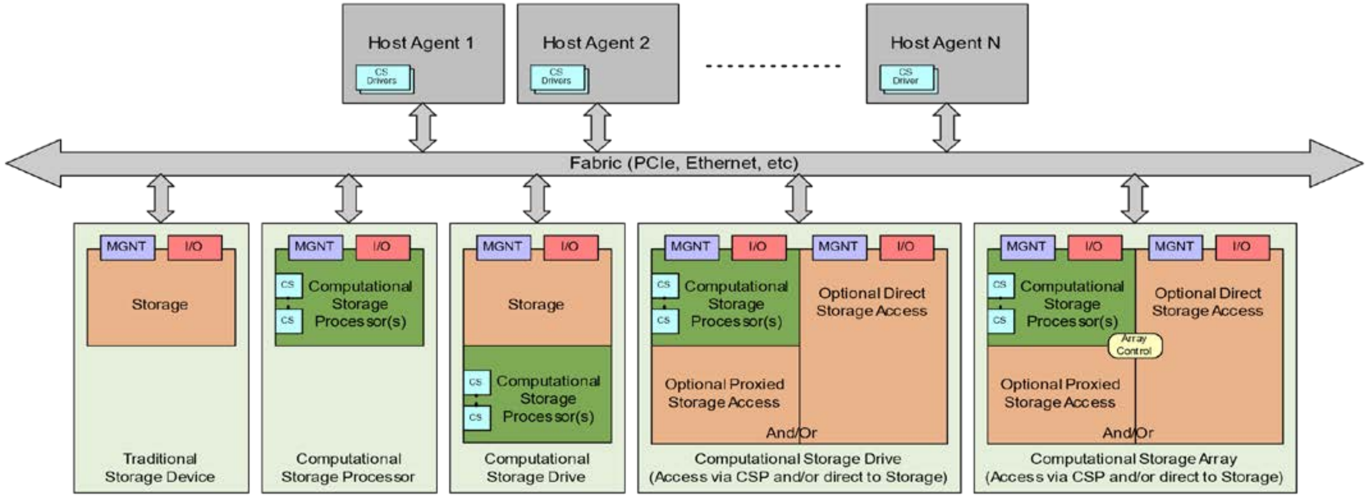
One Driver to Rule Them All?!

- NVMe™ has been incredibly successful as a storage protocol.
- Also being used for networking (NVMe-oF™ and things like AWS Nitro and Mellanox’s Sexy NVMe Accelerator Platform (SNAP)).
- Why not extend NVMe to compute and make it ***the one driver to rule them all?***

A New Product Category

SNIA COMPUTATIONAL STORAGE

- Computational Storage Device (CSx)



Host Agent 1
CS Drivers
Host Agent 2
CS Drivers
Host Agent N
CS Driver
Fabric (PCIe, Ethernet, etc)
MGNT
I/O
Storage
Traditional Storage Device
CS
Computational Storage Processor(s)
Computational Storage Processor
Computational Storage Drive
Optional Direct Storage Access
Optional Proxied Storage Access
And/Or
Computational Storage Drive (Access via CSP and/or direct to Storage)
Array Control
Computational Storage Array (Access via CSP and/or direct to Storage)

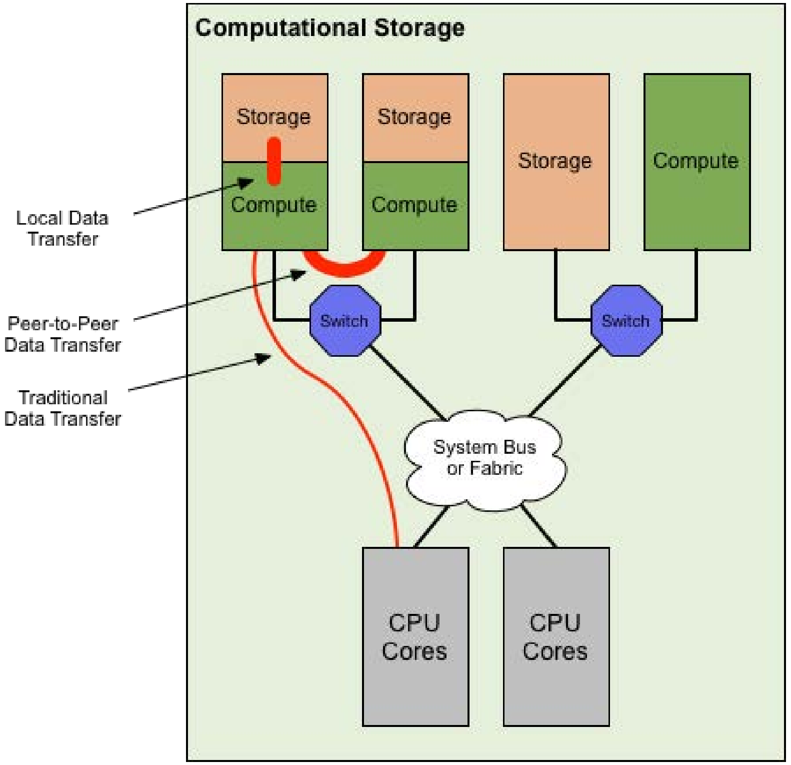
- Computational Storage Drive (CSD)
- Computational Storage Processor (CSP)
- Computational Storage Array (CSA)

What is Computational Storage?
SNIA has Defined the Following

Computational Storage Drive (CSD): A component that provides persistent data storage and computational services

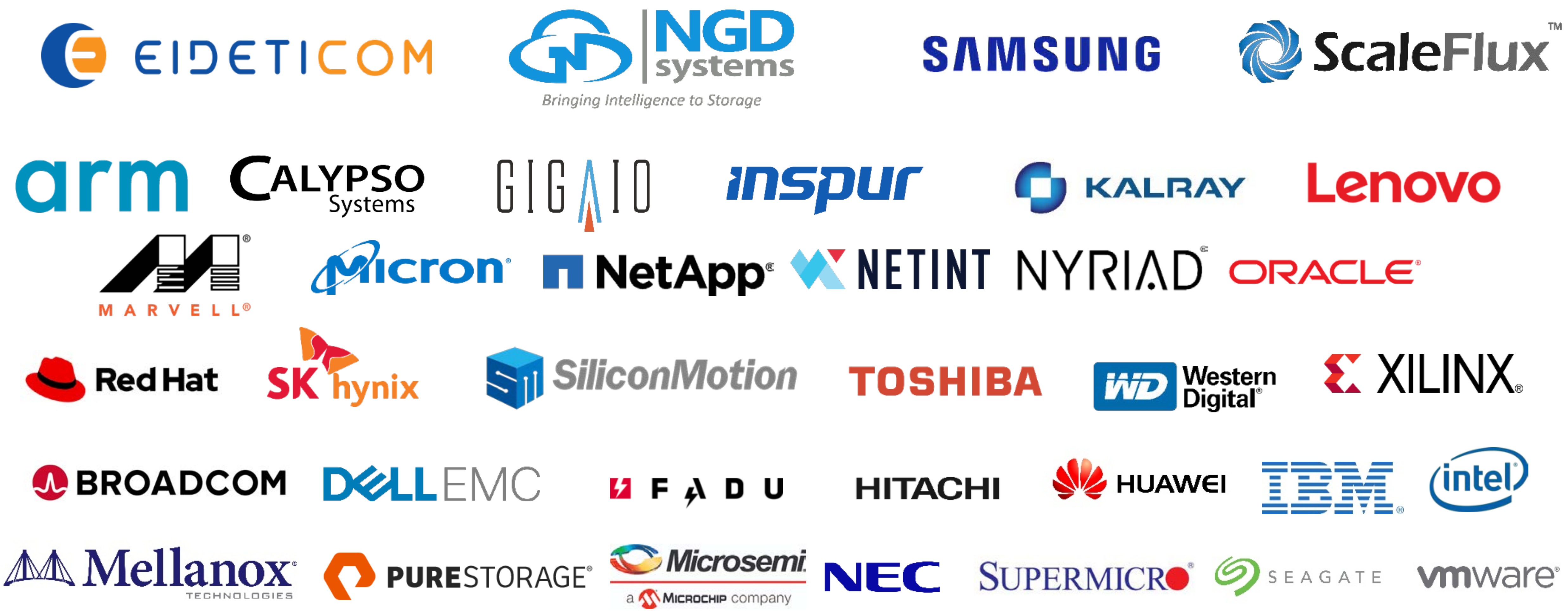
Computational Storage Processor (CSP): A component that provides computational services to a storage system without providing persistent storage

Computational Storage Array (CSA): A collection of computational storage drives, computational storage processors and/or storage devices, combined with a body of control software



40+ Participating Companies
128+ Individual Members

SNIA COMPUTATIONAL STORAGE



Why NVMe™?

- Accelerators require:
  - Low latency
  - High throughput
  - Low CPU overhead
  - Multicore awareness
  - Management at scale
  - QoS awareness

- NVMe provides:
  - Low latency
  - High throughput
  - Low CPU overhead
  - Multicore awareness
  - Management at scale
  - QoS awareness

Real question is “Why not NVMe?”

NoLoad® CSP – Platforms

NoLoad® CSP U.2

- Standard U.2 SSD form-factor: Utilizing SFF-8639 connector.

NoLoad® CSP Alveo

- Standard GPU form-factor: x16 PCIe
- Deployed on Xilinx Alveo U200, 250 or U280

PCIe Gen4

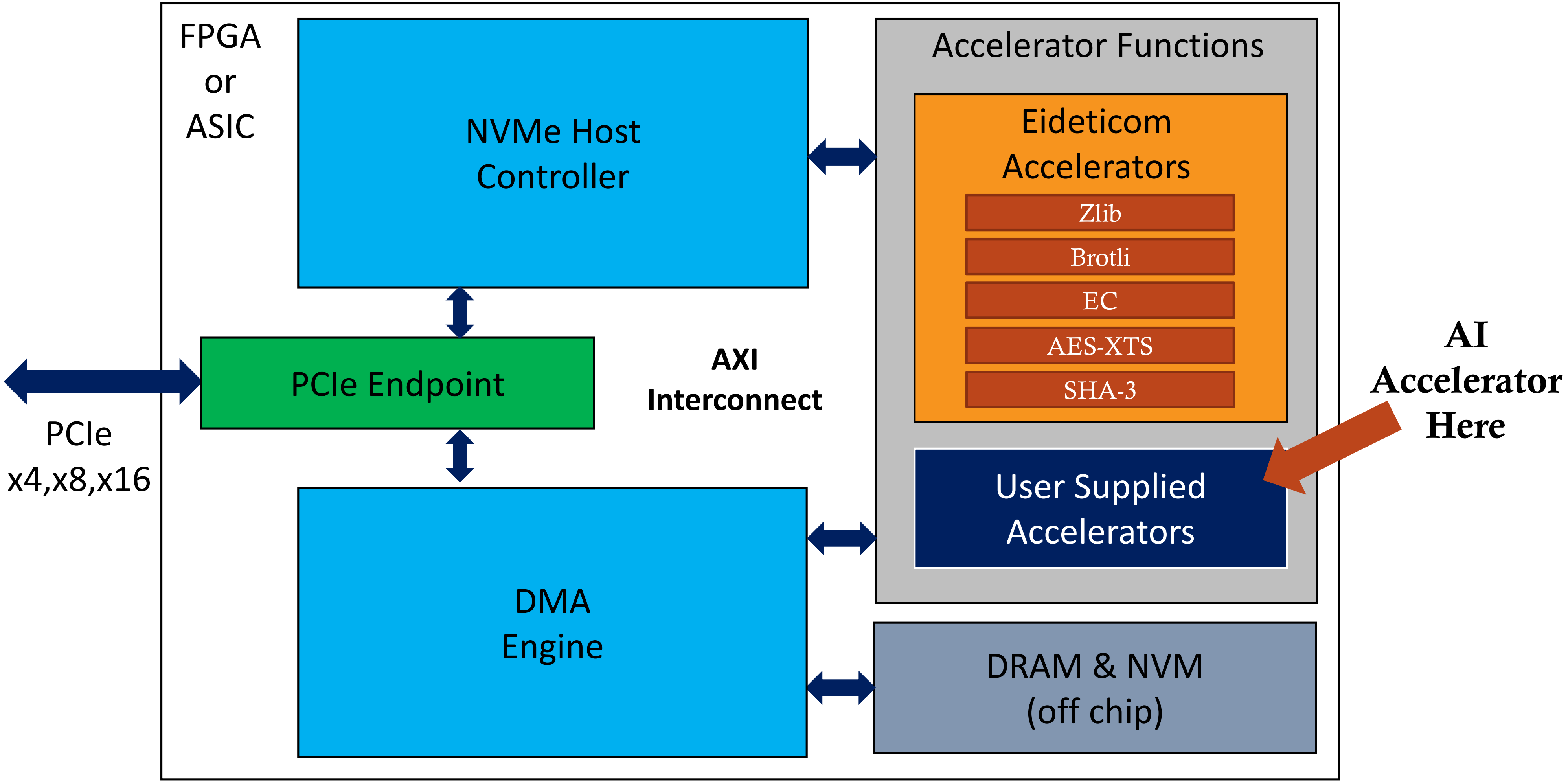
- 16GB/s of data ingestion/egestion.

Eideticom NoLoad IP:

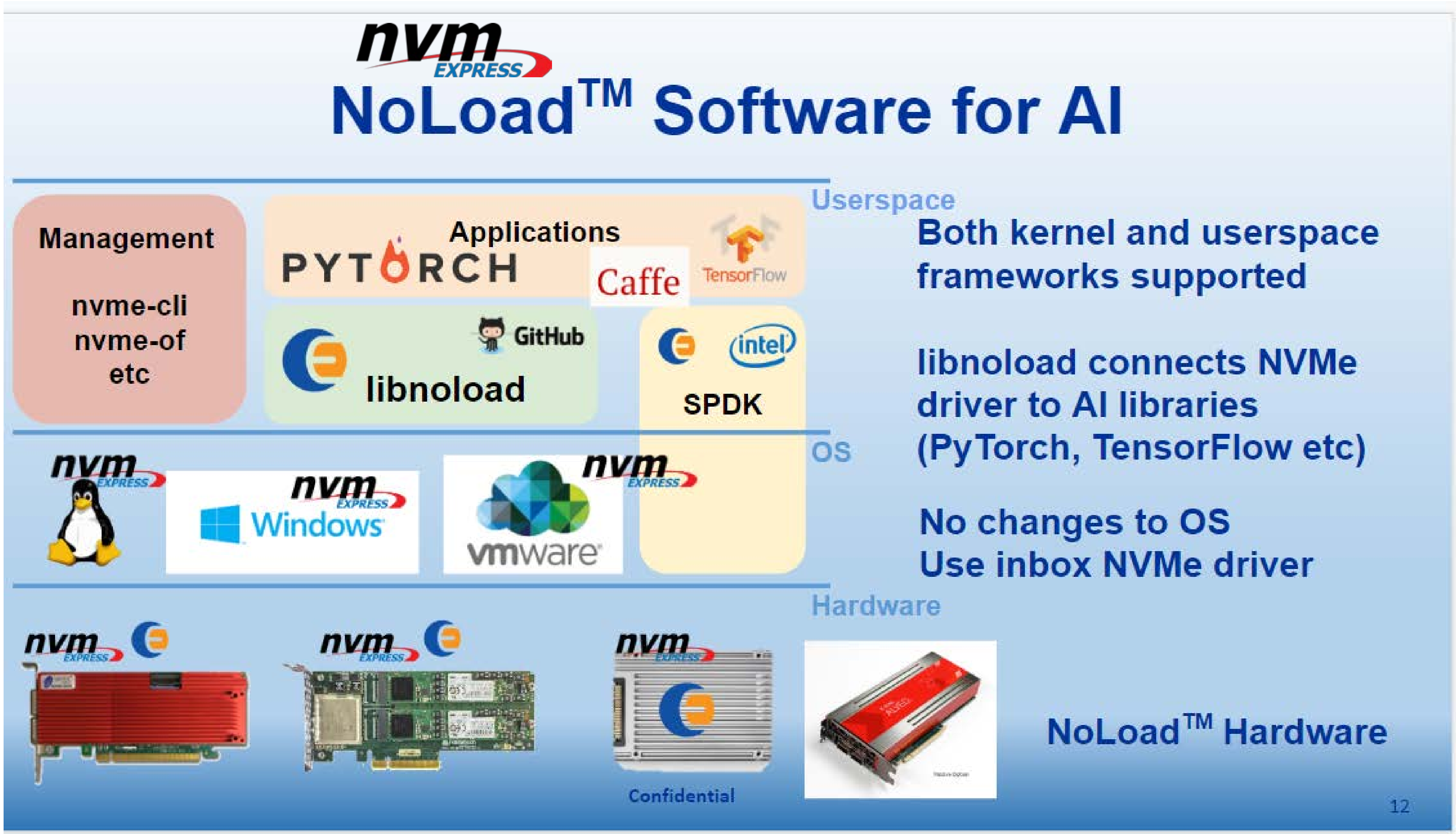
- NVM Express end-point
- Storage and Compute Accelerators
- NVMe SGL support
- CMB and P2P support

Available Now

NoLoad® CSP - Architecture



NVMe Computational Storage for AI



NoLoad^{TM} Software for AI

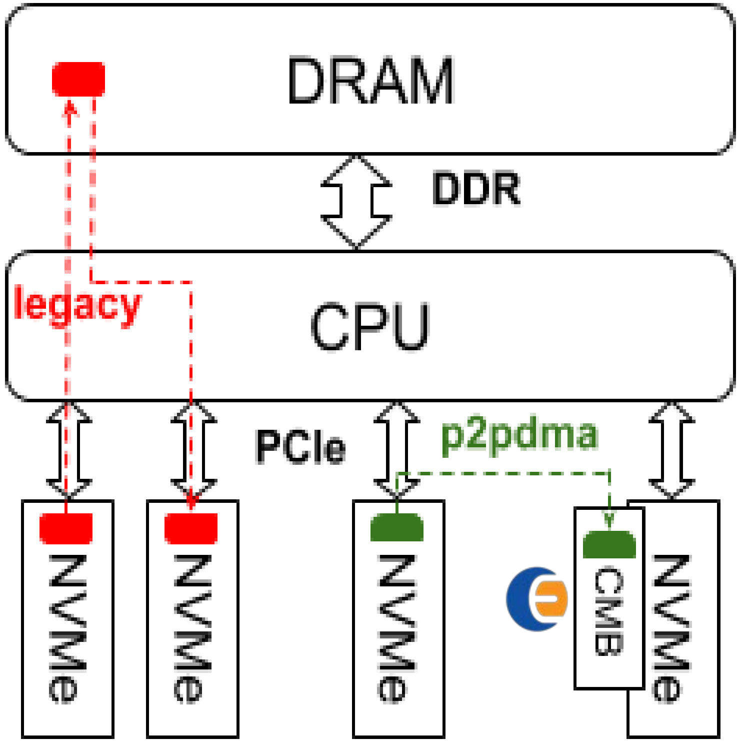
Both kernel and userspace frameworks supported

libnoload connects NVMe driver to AI libraries (PyTorch, TensorFlow etc)

No changes to OS
Use inbox NVMe driver

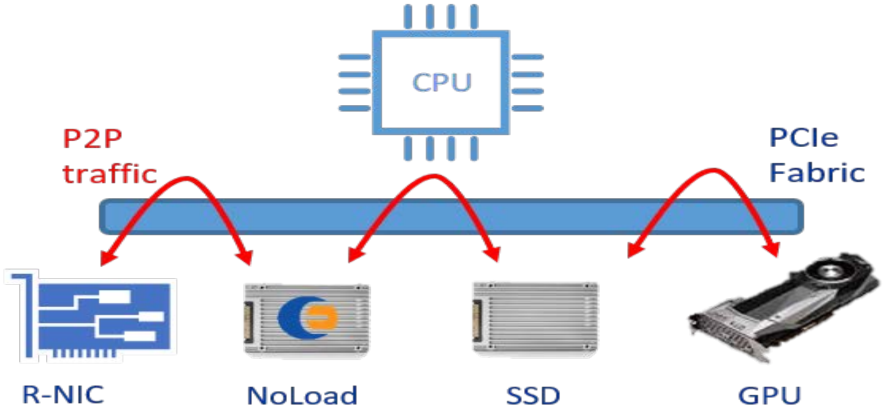
NoLoad^{TM} Hardware

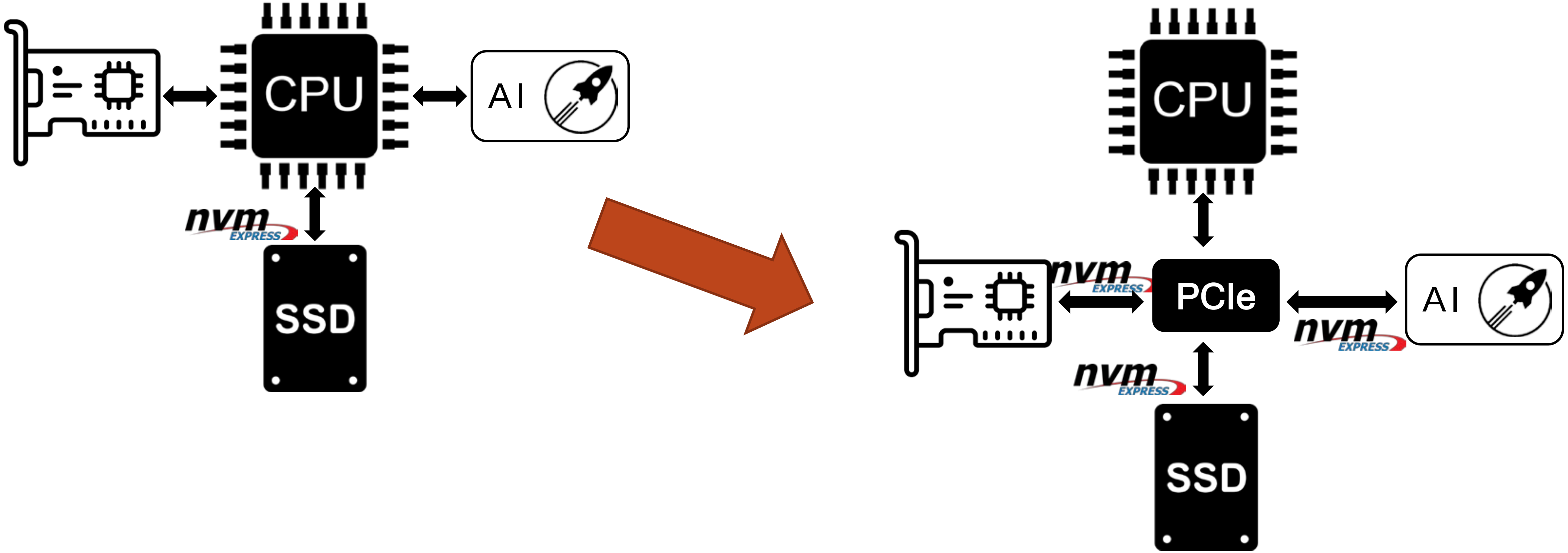
NoLoad® CSP – Peer to Peer (P2P)



- P2P avoids CPU's memory subsystem
- NoLoad implements a high performance NVMe CMB which can be used for P2P DMAs.

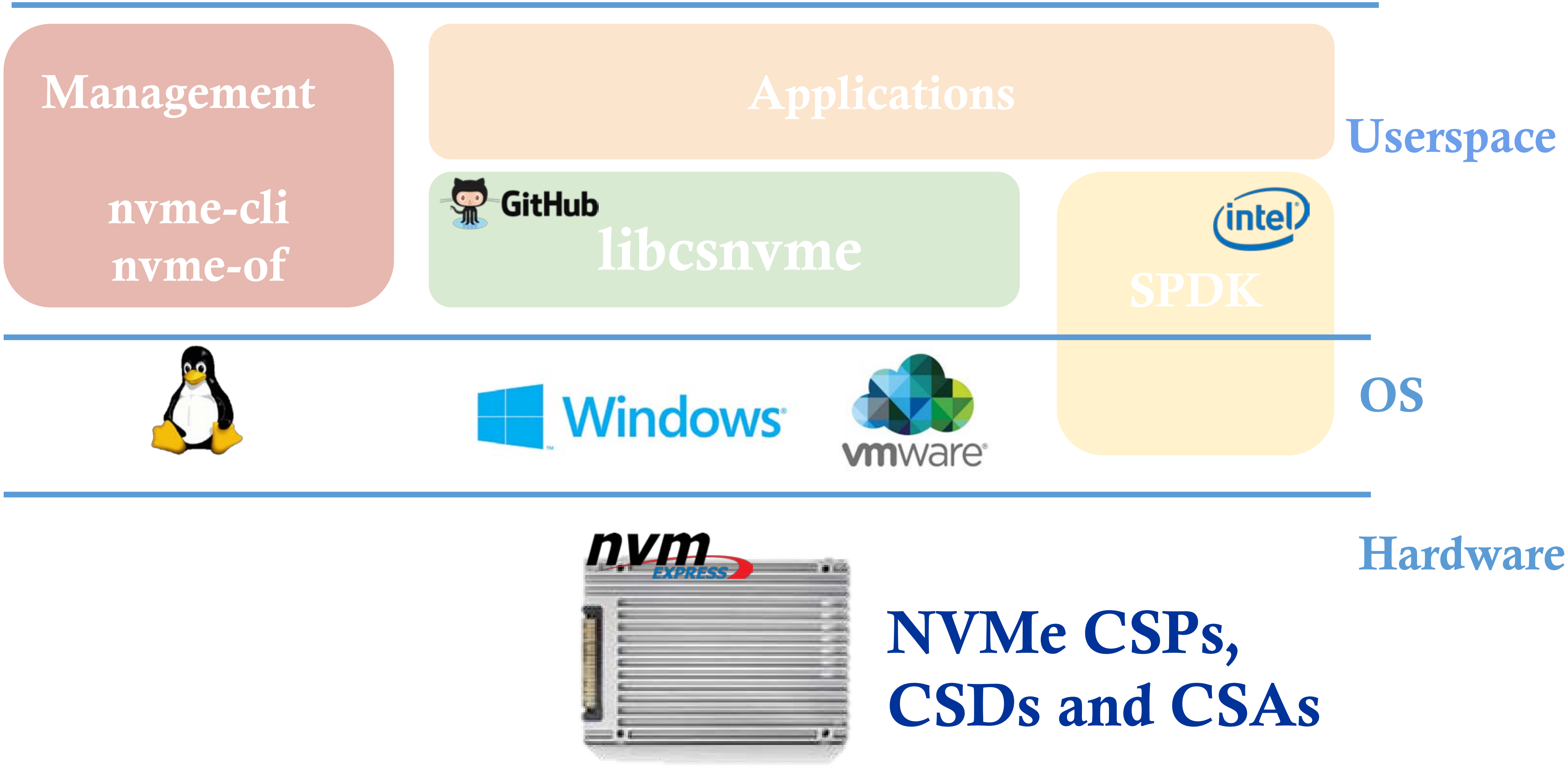
- PCIe End-Points (EPs) are getting faster e.g. NVMe SSDs, RDMA NICs & GPGPUs
- Bounce buffering all IO data through system memory is a waste of system resources
- NoLoad P2P allows PCIe EPs to DMA to each other whilst under host CPU control
- CPU/OS still responsible for security, error handling etc.
- 99.99% of DMA traffic now goes direct between EPs





- Reduced data movement
- CPU offload of processing
- CPU offload of DMA traffic
- Standards based drivers and software
- Vendor neutral
- Management

NVMe™ for Computation: Software



Management

nvme-cli
nvme-of

Applications

GitHub
libcsnvme

intel
SPDK

Userspace

Windows

vmware

OS

Hardware

nvm EXPRESS

NVMe CSPs,
CSDs and CSAs

NVMe Computational Storage for AI

Advantages of NVMe Computational Storage for AI

- Offloads inference to CSx which improves cost, power and space efficiency.
- p2pdma decouples data-plane from CPU further improving cost and power savings.
- CSx is a NVMe controller and can be **managed at scale using NVMe-MI**
- NoLoad Inference engine is a NVMe namespace and **can be shared over Ethernet** using NVMe-oF. Inference disaggregation reduces cost, saves space and saves power.
- **No proprietary drivers**. In time NVMe Computational Storage standard will allow for vendor-neutral approach.
- SNIA developing a **vendor-neutral interface** so hardware from different vendors is supported by the same SW stack.

NVMe Computational Storage for RocksDB

EIDETICOM

RocksDB Acceleration using NoLoad® CSP

Eideticom's NoLoad® CSP deployed on Alveo

- 6x more transactions per sec
- 2.5x more efficient
- 4x reduced NAND costs
- Improved QoS

Similar results achievable for AI inference.

Conclusions

NVM Express for AI

- Efficient and parallel PCIe protocol for talking to PCIe-attached NVM.
- Performance metrics aligned with needs of AI (GB/s, sub ms latency).
- Need filesystems to catch-up!

Computational Storage for AI

- Offload key AI tasks to accelerators in a standards based way.
- Move the computation closer to the data to reduce data movement and improve efficiency.
- Standardize accelerator interfaces and management via NVMe.

EIDETICOM

Eideticom HQ
3553 31st NW,
Calgary, AB,
Canada T2L 2K7

Eideticom (Bay Area)
168 South Park,
San Francisco, CA 94107
USA

www.eideticom.com

Contact: sales@eideticom.com

Thank You!