Flash Memory Summit

AIAP-302-1: Storage for AI Part 1

Agenda: AIAP-302-1: Storage for AI Part 1

- Optimizing Edge AI, Networking and Storage by Combining GPU and DPU Technology
  - John Kim - NVIDIA
- Accelerating the Data Path to the GPU for AI and Beyond
  - Sandeep Joshi - NVIDIA
- Accelerating data services and storage disaggregation thanks to DPUs
  - Tim Lieber – Kalray
- Introductions, question wrangling and other things of limited value
  - Howard Marks – VAST Data
    - @deepstoragenet Howard@VASTdata.com

Flash Memory Summit

Optimizing Edge AI, Networking & Storage

John F. Kim

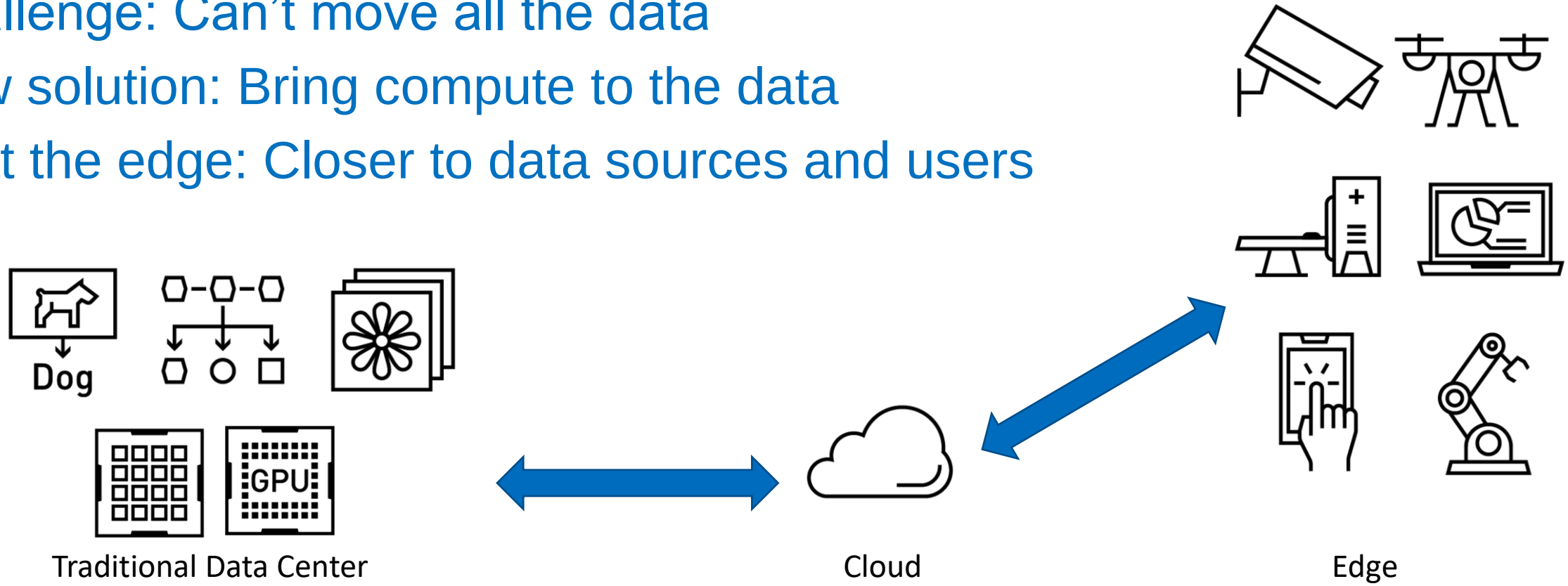
Director of Storage Marketing

NVIDIA

35 YEARS OF NAND FLASH MEMORY INNOVATION

Bring Compute to the Edge

- Traditional: Centralize all data for AI
- Challenge: Can’t move all the data
- New solution: Bring compute to the data
- AI at the edge: Closer to data sources and users

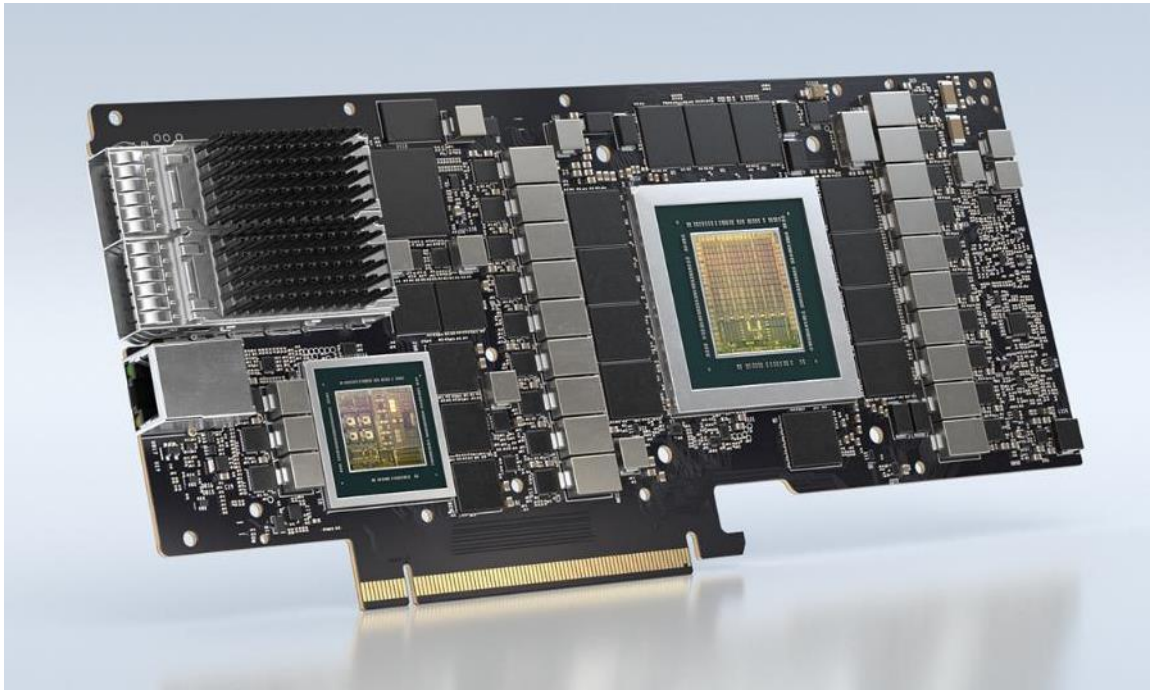


AI at the Edge

- How to run AI at the Edge
  - CPU only
  - CPU+GPU or CPU+FPGA
  - DPU with AI/ML optimizations
  - Computational Storage
  - SOCs with GPU+DPU
- Distributed AI
  - Multiple GPUs in one server—use fast internal fabric
  - Distribute across GPUs in multiple servers

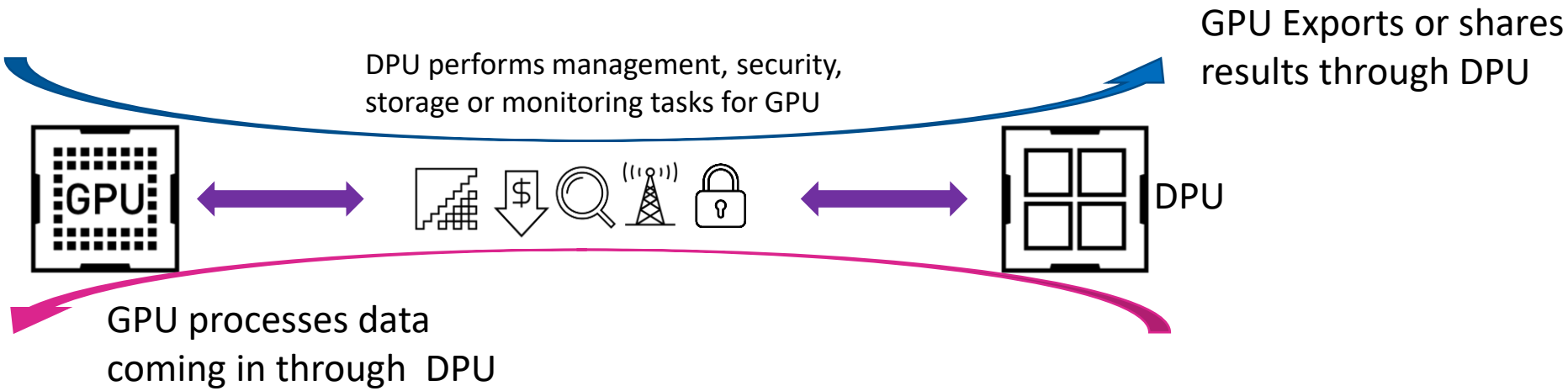
Combine GPU and DPU

- GPU and DPU on one card or one chip
- Eliminate PCIe bottleneck/latency
- Avoid bothering server CPU/DRAM
- Controller: eliminate separate CPU



Functional Data Flow

- GPUs ingest data through DPU
- GPU AI results exported/shared via DPU
- GPUs collaborate over network – distributed AI
- GPU processes data coming from DPU
- DPU provides security, management, storage, monitoring



Edge Use Cases

- Video processing
- 5G
- Cybersecurity
- Retail / Healthcare / Manufacturing -- if need AI cluster

Any intersection of AI/ML processing with high-speed data movement.
Ideally with 1:1 relationship between GPU and DPU.

Flash Memory Summit

Accelerating the Data Path to the GPU for AI and Beyond

Sandeep Joshi, NVIDIA

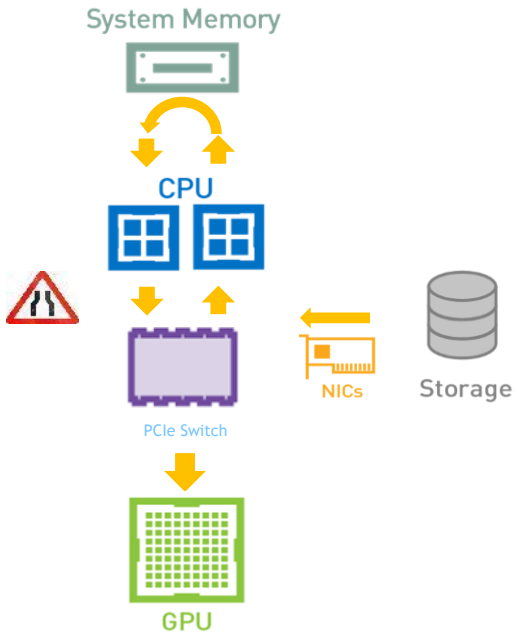
35 YEARS OF NAND FLASH MEMORY INNOVATION

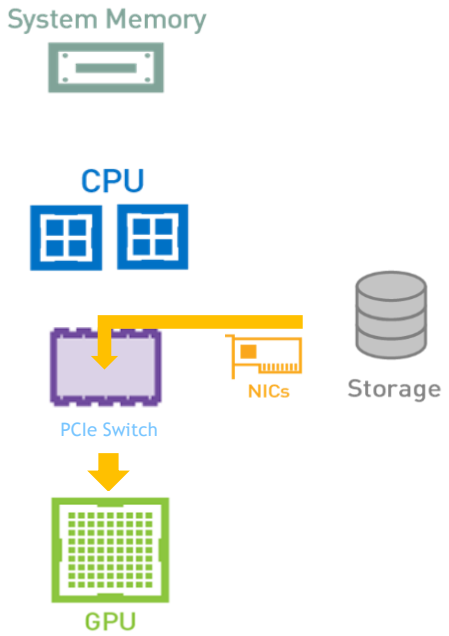
NVIDIA GPUDirect™ Storage Overview

A Direct Path Between Storage and GPU Memory

Limits IO Performance

- 2 DMA operations
- Needs bounce buffers
- Competes for CPU memory bandwidth
- Competes for CPU cycles
- Halves Bandwidth (CPU PCI Switch)





Excellent IO Performance

- Single DMA operation
- No bounce buffer
- No competition for CPU memory bandwidth
- No dependence on CPU PCI Switch bandwidth

Application Speedups!

Framework/Setups	Speedup/ benefit	Workload
PyTorch, DALI relative to standard NumPy-only baseline	7.2x	Inference
Verizon multi-camera video processing with TCP	10x	Video + ML processing
RAPIDS cuDF	3-6x	VCF read for genotyping data
cuCIM	11x	reading a non-compressed/multi-resolution TIFF file
Merlin with NVTabular	1.4x	Click-through-rate logs (CTR)
KvikIO/Zarr	2.9x	weather and climate
User-defined functions in HDF5 by IBM	5x	Blur convolution filter, MSAVI Index

GPUDirect™ Storage Support Matrix

Product Info

Component	Supported Version
HW platform	DGX, EGX
GPUs	Data center GPUs with Compute Capability > 6
CUDA toolkit	>= 11.4
Linux Distros	Ubuntu >= 18.04, RHEL >= 8.3
File systems	Ext4, xfs, NFS(NetApp, VAST, Isilon, Pavilion), Lustre, Spectrum Scale, BeeGFS

Feature	Description
IOMMU	off/passthrough
Power9 or Arm	Not supported
Virtualization (vSphere)	vGPU with NFS
Language bindings	C, C++, Python

GPUDirect™ Storage Ecosystem

Flash Memory Summit

ALL GA PARTNERS

ddn · DELL EMC · Hewlett Packard Enterprise · HITACHI Inspire the Next · IBM · KIOXIA · LIQID · Micron · NetApp · PAVILION · SAMSUNG · ScaleFlux · VAST · WEKA

FILE SYSTEM PARTNERS

Partner company	Partner Product	GDS Version	Date
NetApp	ONTAP 9.10.1	1.0 and higher	Jan-22
NetApp ThinkParQ System Fabrics Works	BeeGFS Tech Preview	1.1 and higher	Mar-22
IBM	Spectrum Scale 5.1.2	1.1 and higher	Nov-21
DDN	EXAScaler 6.0*	1.1 and higher	Nov-21
VAST	Universal Storage 4.1	1.1 and higher	Nov-21
WekaIO	WekaFS 3.13	1.0	Jun-21
DellEMC	PowerScale 9.2	1.0	Oct-21
Hitachi Vantara	HCSF	1.0	Oct-21

*Open source Lustre 2.15 supports GDS

Medical Imaging – Deep dive

Use Case I: NVIDIA Clara Holoscan - Microscopy Simulation

- The Lightsheet microscopy (@Advanced Bioimaging Center at UC Berkeley)
  - Streams 3TB of data in an hour
  - To process and visualize the streaming microscopy data
  - **Result** - To automatically detect rare biological events in real time
- See: https://www.youtube.com/watch?v=rXG27G3bWzY

Use Case II: NVIDIA Clara cuCIM

What is cuCIM?

- IO for Multi-Resolution Images in Digital Pathology
- Pre & Post Processing for Digital Pathology Images

Why GPU?

- Compute intensive

Why GDS?

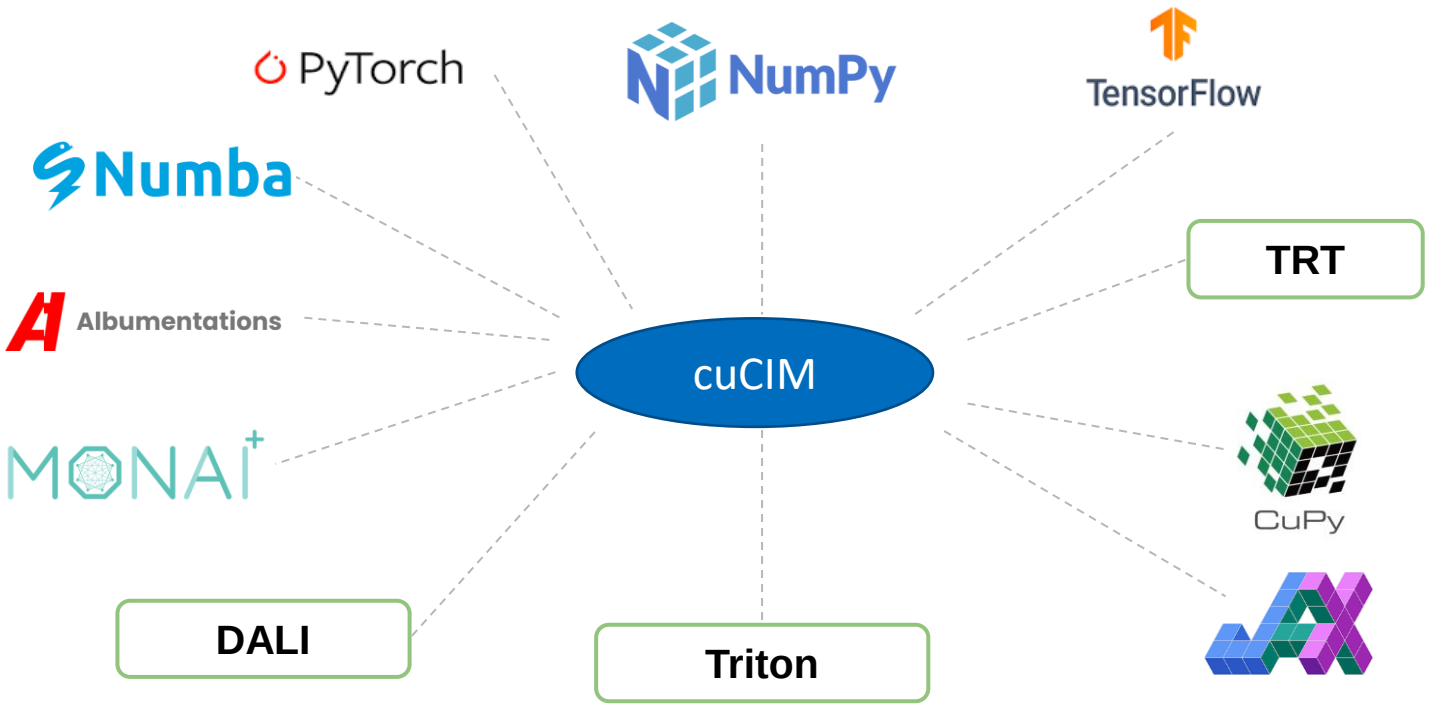
- CPU and Memory intensive

How?

- Readers/Writers for Tiff File/Zarr

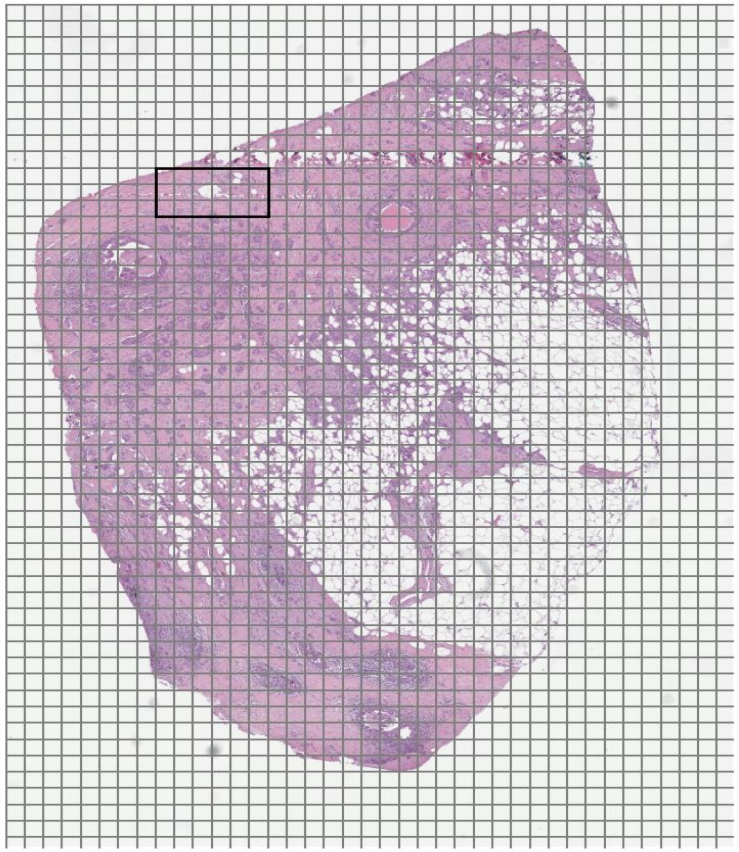
cuCIM documentation
GitHub - rapidsai/cucim
Multi-dimensional image procesing

Integration with Other Tools

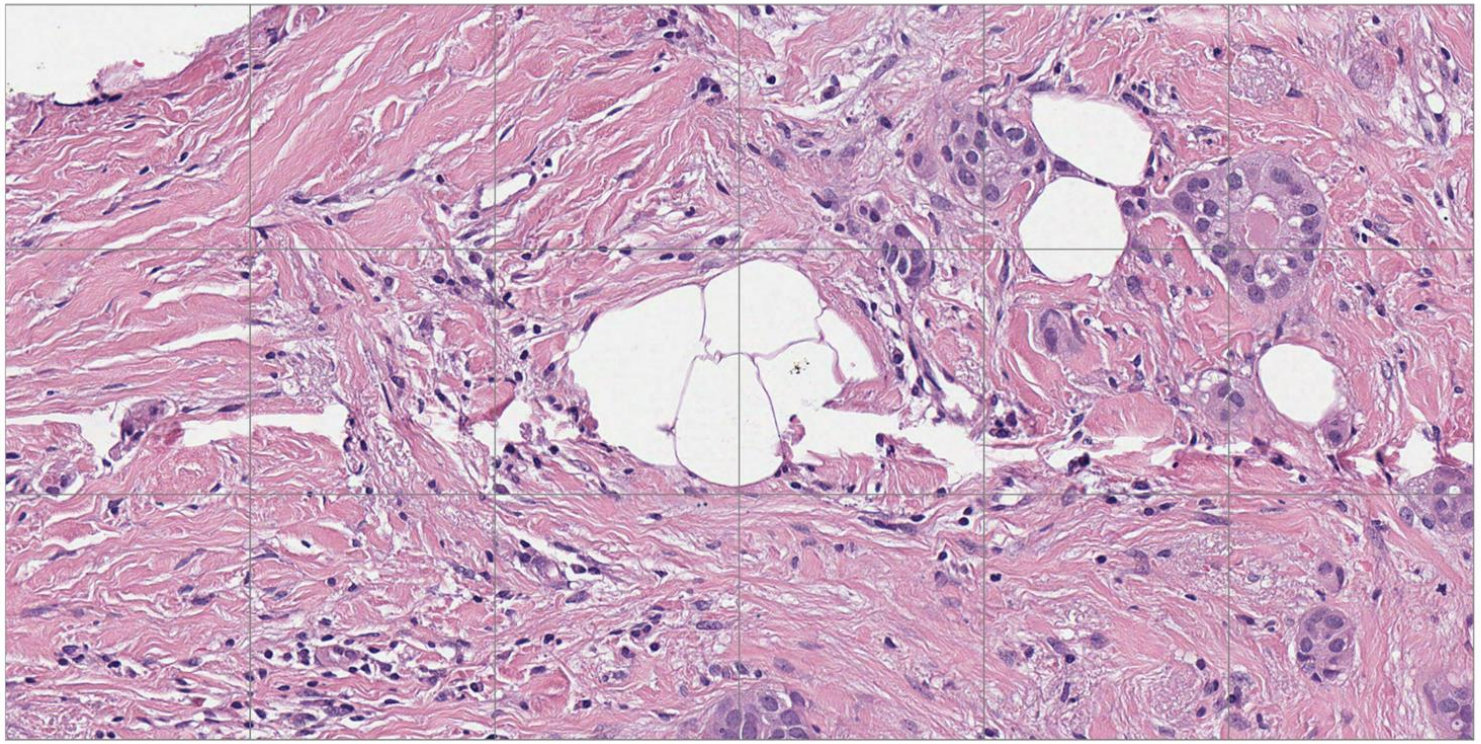


Data Overview

The highest resolution layer contains 2028 RGB tiles of size (512, 512).

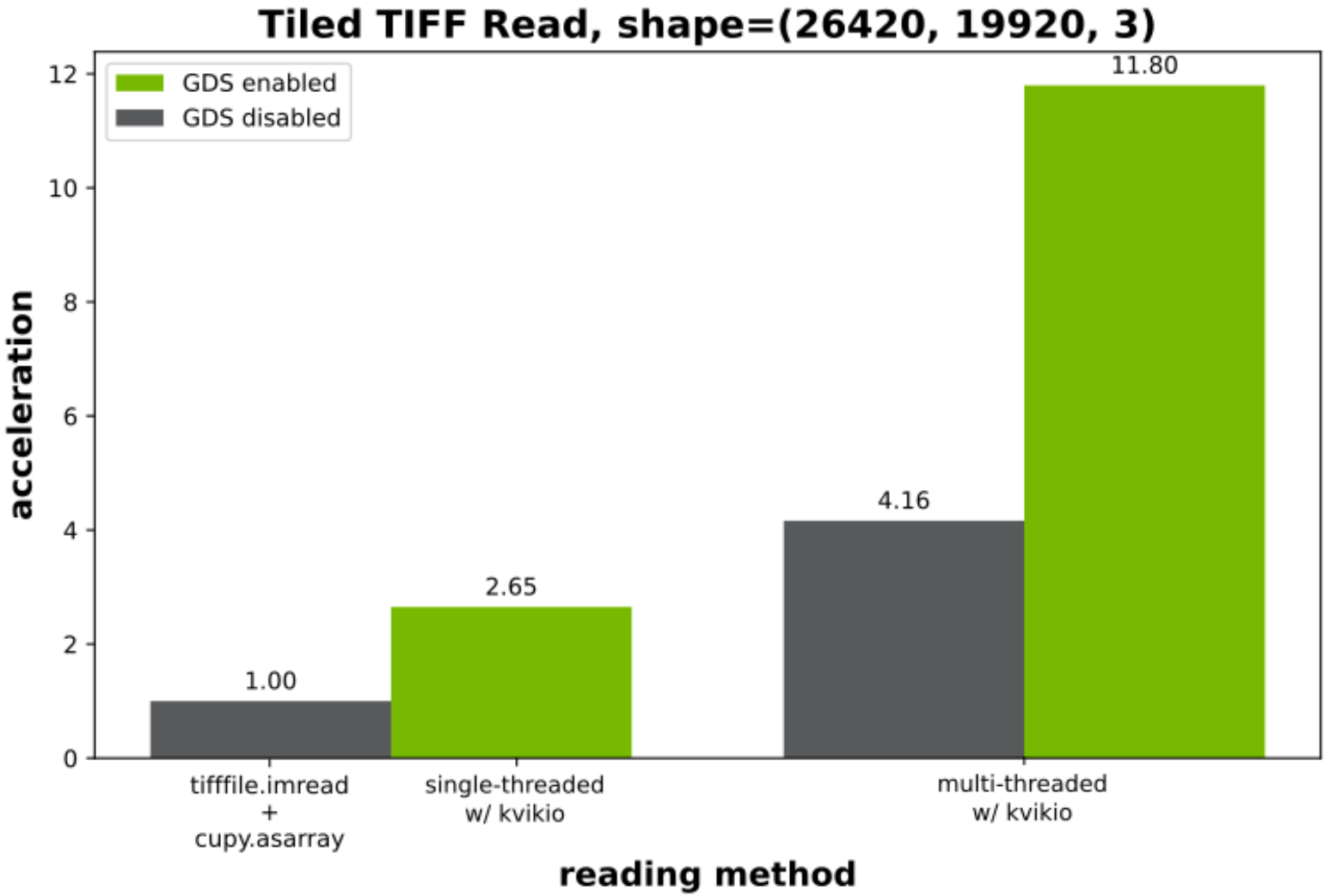


High resolution view of a region-of-interest (ROI) (corresponding to the black rectangle on the left). Tile boundaries are overlaid in gray. Reference: Veta, Mitko, et al.

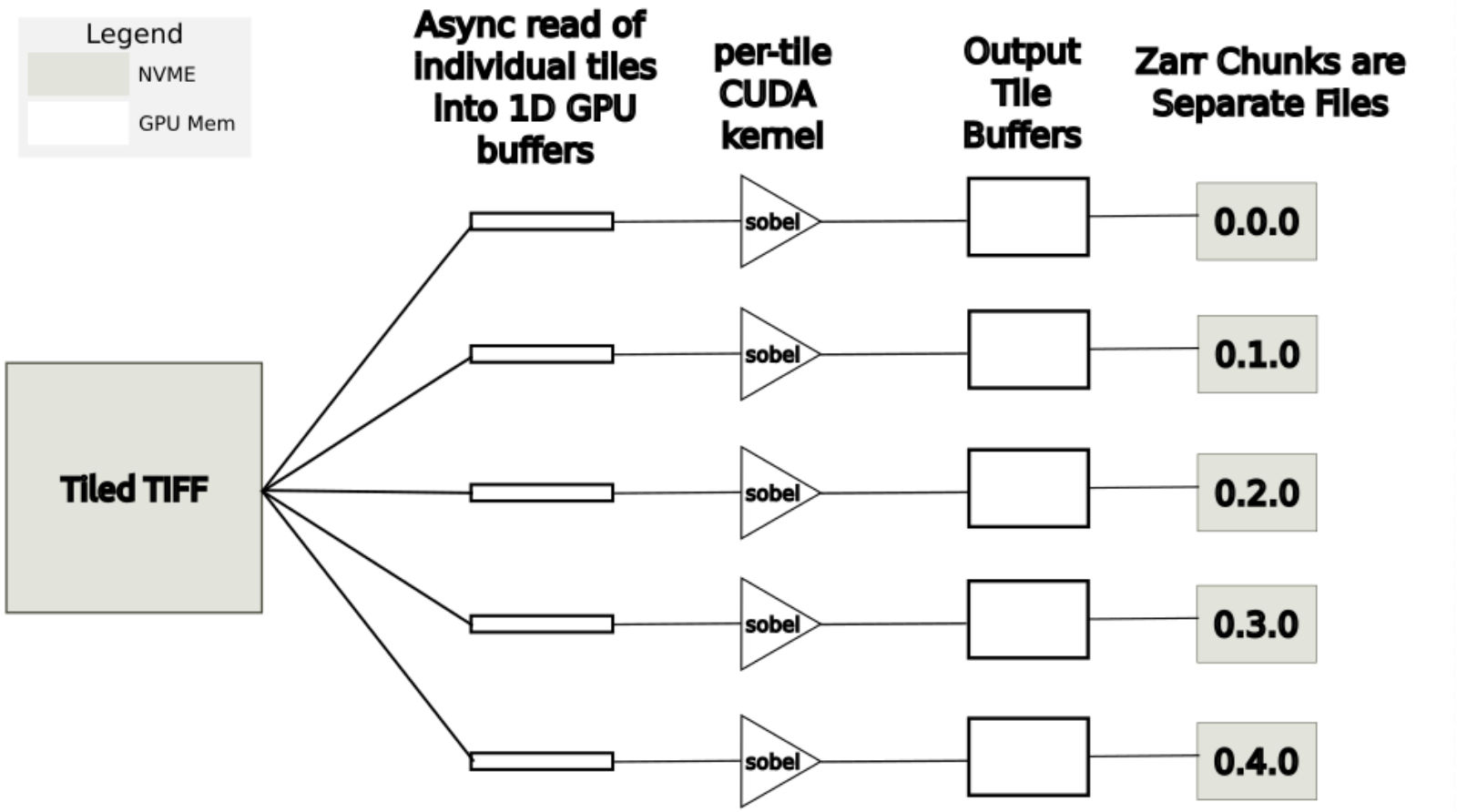


Veta, Mitko, et al. "Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge." Medical image analysis 54 (2019): 111-121.

Tifffile.read vs cuFileRead

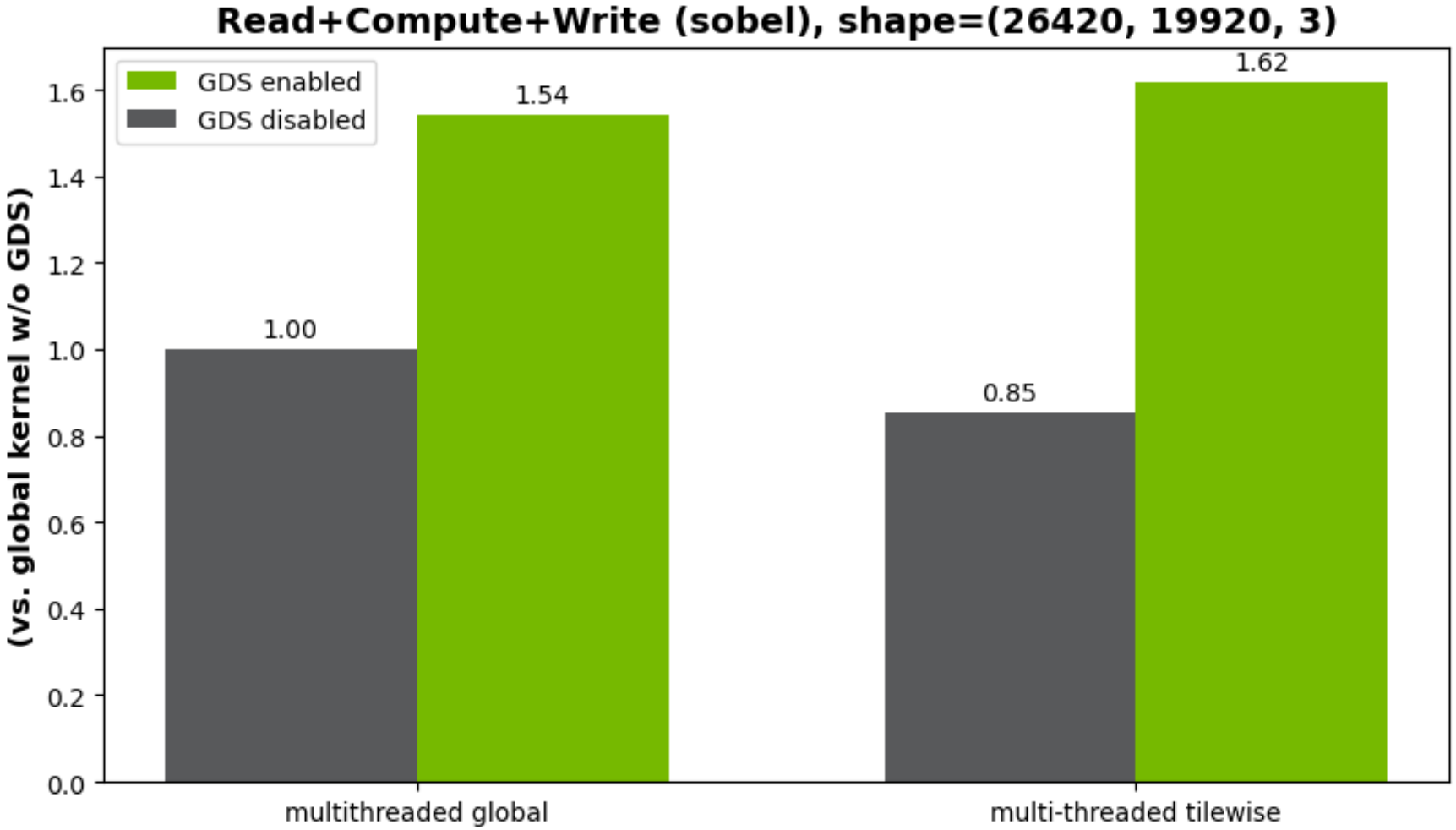


End-to-End flow (Read + Compute + Write)



Much lower memory usage: only a small subset of tiles in memory at any one time. (e.g. ~700 MB peak memory use)

End-to-End flow (Read + Compute + Write)



Technology Directions

- Batch APIs: reduce overheads on a mix of reads and writes
  - Available as of CUDA 11.6
  - Large # of larger transfers (>512KB): 20x fewer threads for similar read perf
- Hopper GPUs (Gen 5) + CX7: 400 Gbs = 50 GB/s per NIC
- Grace (leading-edge Arm CPU) + Hopper: Certify on this new ISA target

Call to action

Flash Memory Summit

Check it out

- Documentation - https://docs.nvidia.com/gpudirect-storage/index.html

Try it

- Get from CUDA - https://developer.nvidia.com/cuda-downloads
- Check out repo - https://github.com/NVIDIA/MagnumIO/tree/main/gds

Bring use cases

Flash Memory Summit

Applying DPU technology to Disaggregate and Accelerate Flash Technology

Tim Lieber
Lead Solutions Architect - Kalray

35 YEARS OF NAND FLASH MEMORY INNOVATION

Abstract

ACCELERATING DATA SERVICES AND STORAGE DISAGGREGATION THANKS TO DPUs

NVMe protocol is state-of-the-art technology that enhances the performance benefits of flash-based storage by removing performance bottlenecks. To fully exploit the performances of NVMe devices, storage nodes must dedicate significant portions of compute resources towards storage functions.

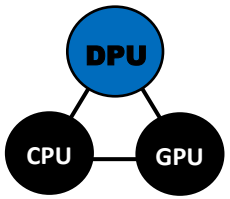
This is especially true when storage services such as LVM, data protection, data reduction or data cryptography are employed. Performance of both local and NVMeoF based disaggregated storage can be adversely affected by system bottlenecks which reduces the expected benefits to TCO of modern NVMe architectures. This is amplified further in a virtualized environment, where hypervisors must offer storage virtualization and disaggregation to Virtual Machines.

In this presentation we will detail the benefits of using DPUs: demonstrating how DPU-based acceleration cards like the Kalray Smart Storage Accelerator PCIe card can seamlessly offload storage services as well as storage disaggregation by exposing many NVMe controllers on the PCIe bus while taking control of local or remote SSDs and share concrete outcomes for the most demanding workflows in domains such as AI, HPC and High-End Media Production.

What Is DPU Technology

A NEW CLASS OF PROGRAMMABLE PROCESSOR

Specialized in running data center infrastructure services



CPU, GPU … DPU

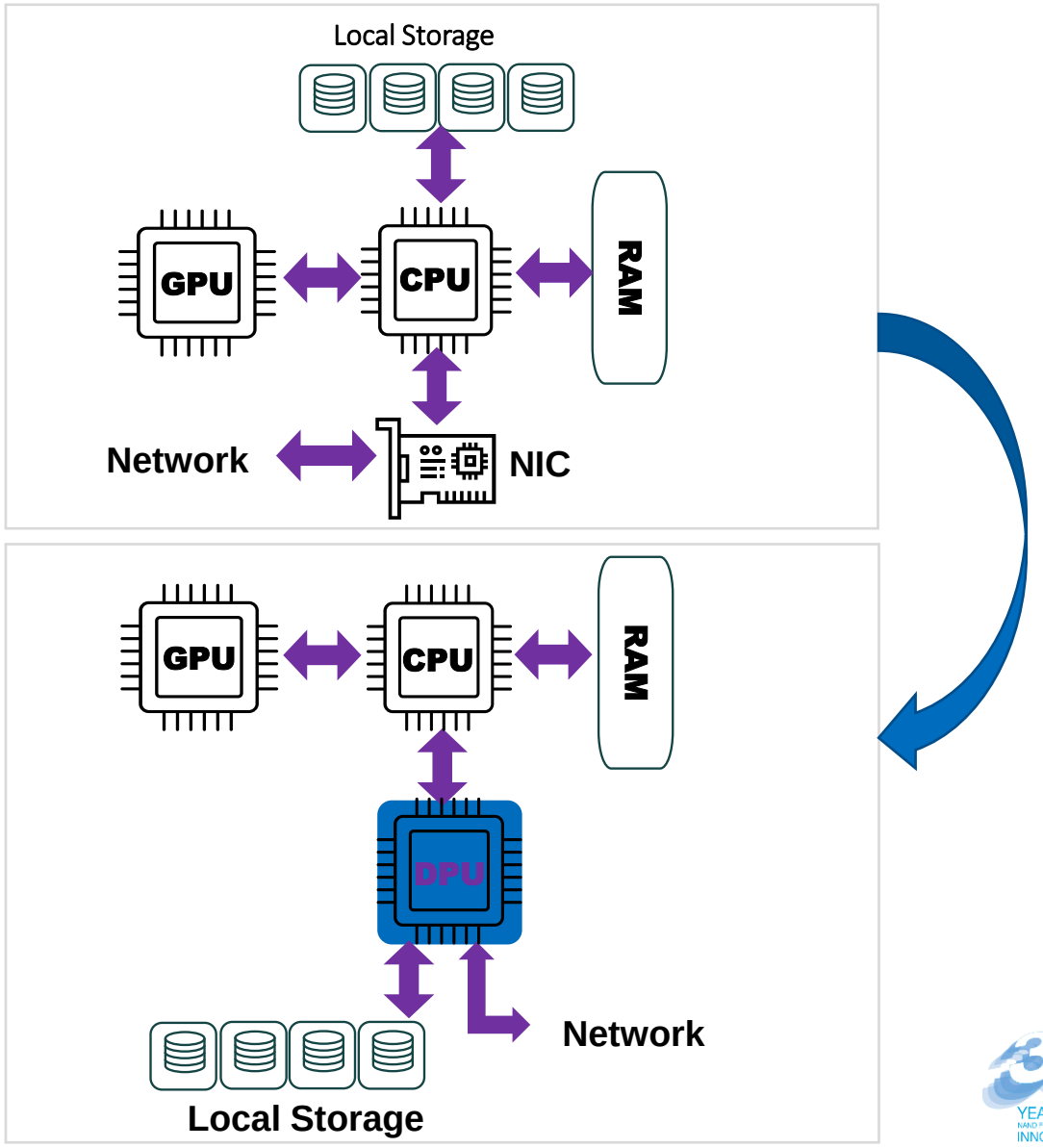
The 3rd socket in data centers alongside CPUs and GPUs

DPU at the core of new server architecture

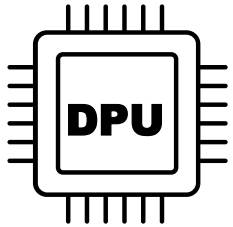
ACCELERATION

Accelerates software-defined data center infrastructure services … and more !

- **Networking**: NFV, vSwitch, NAT, …
- **Storage**: NVMe-oF, compression, deduplication, encryption, …
- **Security**: Firewall, encryption, Ipsec, …



Key Features of a DPU



FULLY PROGRAMMABLE

- Control plane
- Data plane

HIGH PERFORMANCE PCIe INTERFACE

- SR-IOV for virtualization support
- Root complex and peer-to-peer support

HIGH PERFORMANCE NETWORK INTERFACES

- Packet parsing / matching / dispatching
- RDMA support
- TCP acceleration (RSS, LRO, checksums, …)

TIGHTLY COUPLED INLINE ACCELERATORS

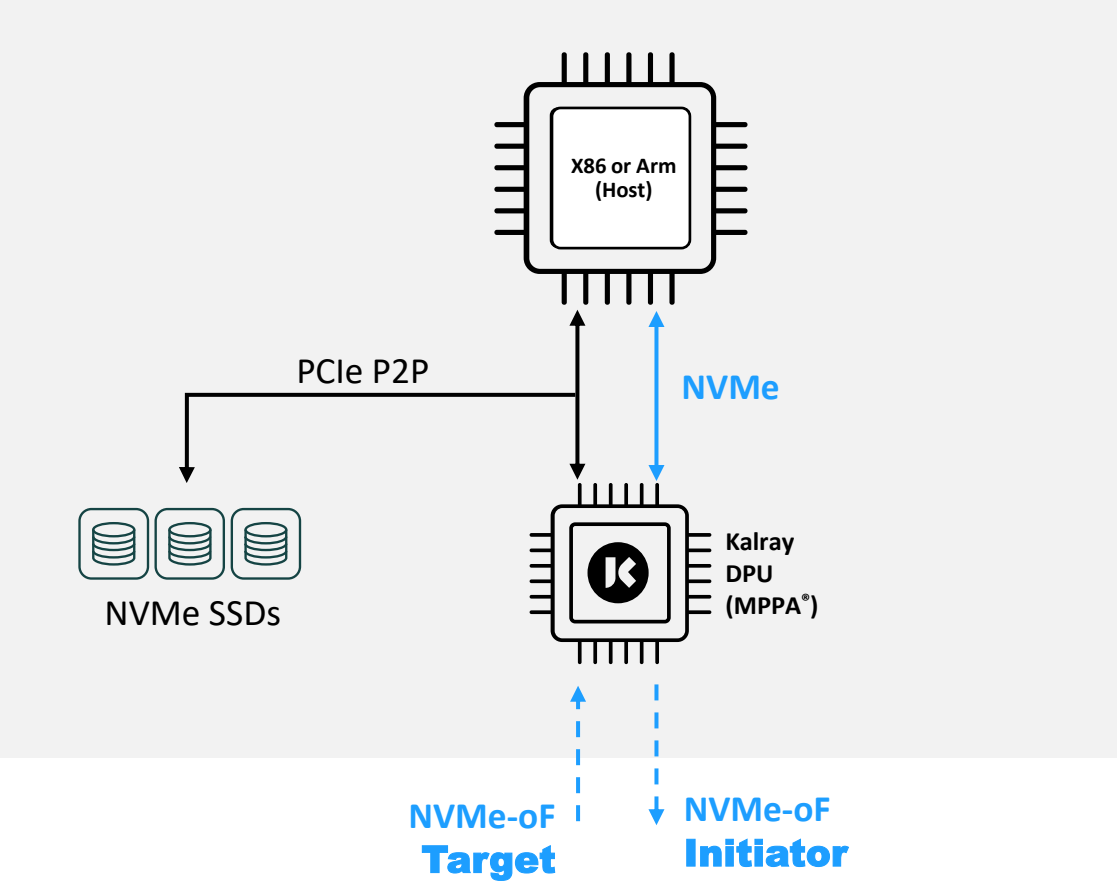
- Crypto accelerators (IPsec, TLS)
- Compression (storage)
- Erasure Coding (KV hashing)

SECURITY

- Root of trust, secure boot, secure firmware upgrades

Architecture #1 – Companion Mode

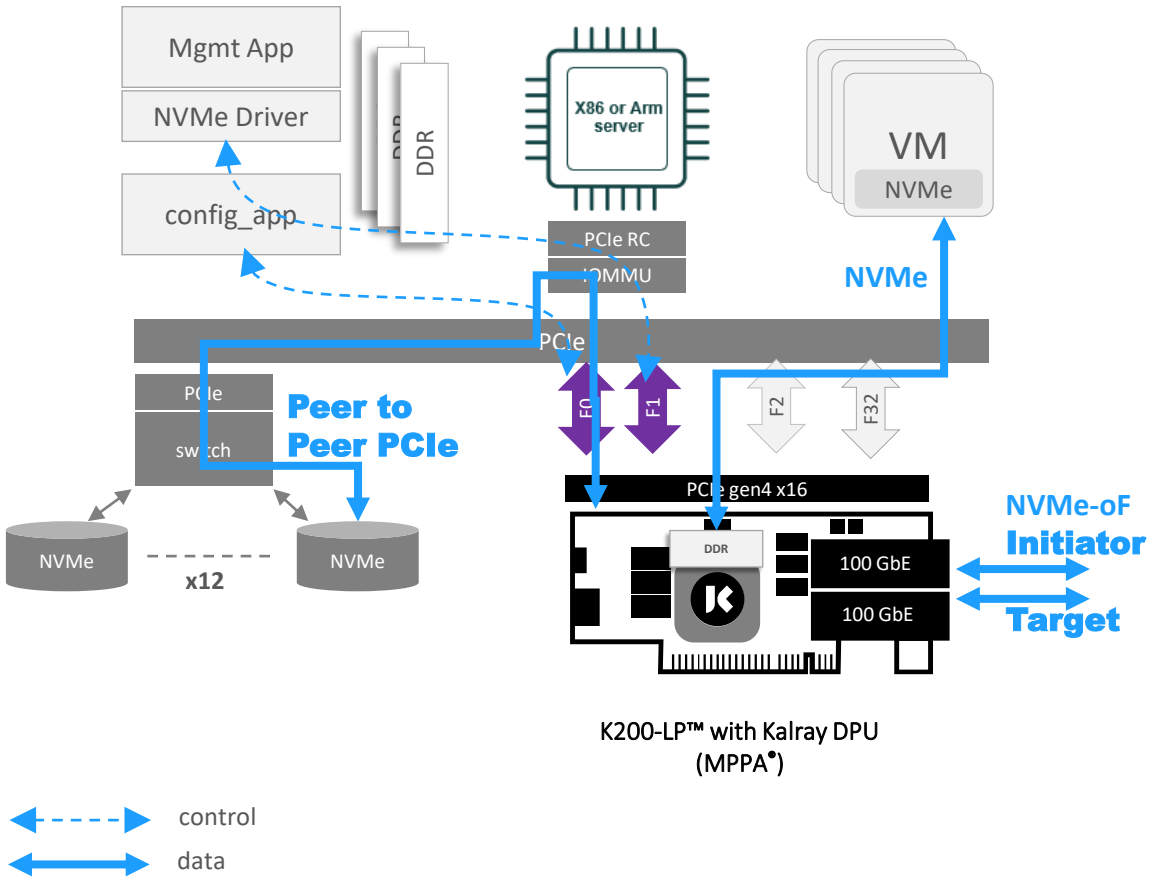
Storage Accelerator and Adapter



Compute Nodes/HCI

- DPU presents itself as PCIe NVMe devices (SR-IOV) to Host
- DPU takes ownership of local SSDs via PCIe peer-to-peer in a transparent fashion
- DPU offload storage data services
- DPU device can be used as NVMe to NVMe-oF storage adapter
- DPU can act as NVMe-oF target for storage disaggregation
- DPU can act as NVMe-oF initiator for distributed services

Architecture #1 - Data Path



K200-LP™ with Kalray DPU (MPPA®)

Data Path

- **Local NVMe unbounded** from host NVMe drivers
- **NVMe emulation**: No DPU specific host device driver needed
- **Config_app**: user space application in charge of
  - Setup x86 IOMMU to map DPU memory
  - Remoting K200 PCIe config space accesses
- **Mgmt_app** (optional):
  - Small application using legacy nvme driver to send custom vendor commands for DPU configuration (logical volumes, storage services…)
- **VMs**:
  - Virtual Machines having direct access to PCIe VFs (PCIe pass-through) exposing NVMe devices

Architecture #1 - Data Service Acceleration

Decision: In-line or Look-aside processing

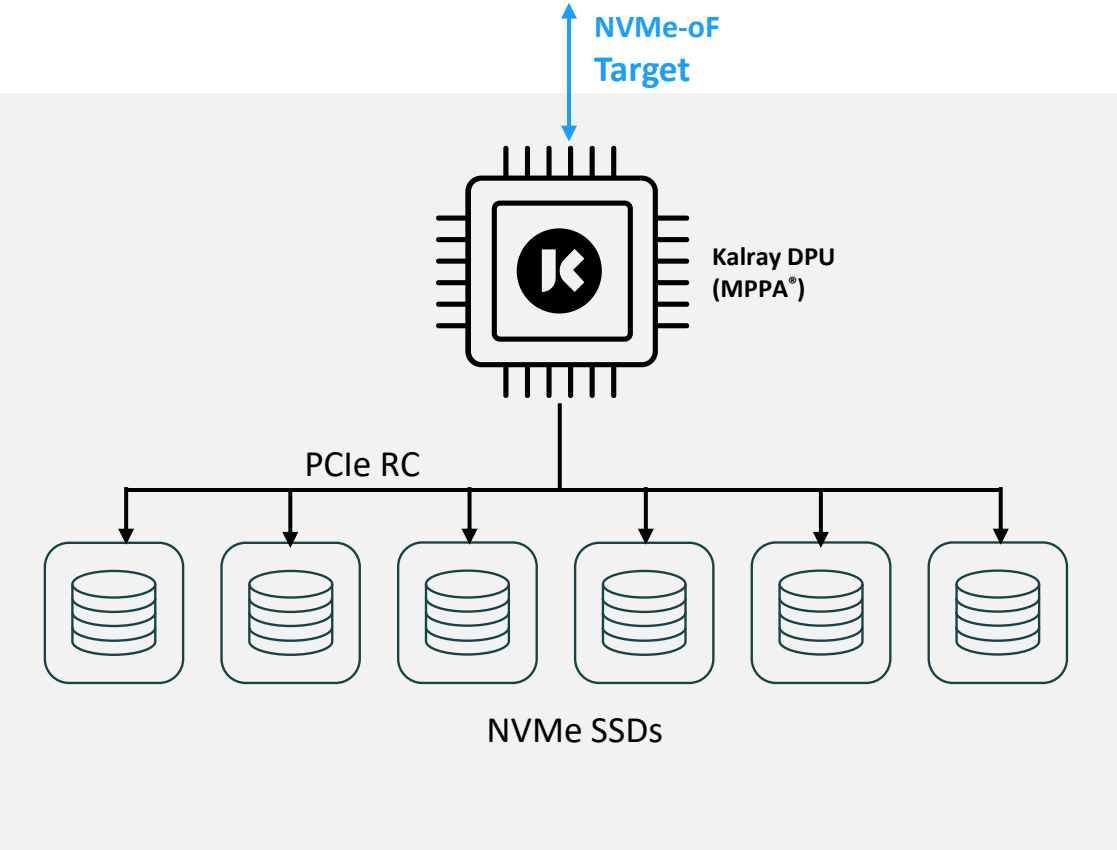
Inline Data Processing

- Storage blocks processed in the storage path (local or remote)
- Interface to host is a “virtual” NVMe volume
- Physical backend devices can be seamlessly local NVMe or remote (NVMe-oF)
- Typical Data Services:
  - data reduction: zero-detect, dedup, compression
  - logical volume with thin-provisioning, snapshot and clones
  - data protection: RAID10, RAID6, distributed EC
  - encryption/Decryption
  - key-Value to block APIs translation/acceleration

Look aside Data Processing

- Storage blocks processed by DPU, but host read them back for further processing (Object, Filesystem ...)
- No physical backend storage device
- Pseudo NVMe namespaces exposed to host with dedicated processing capabilities
- Typical Data Services:
  - computational storage function processor
  - non block processing (file/object): AI, computer vision, NLP etc.
  - raw block processing: Crypto, Compression, EC

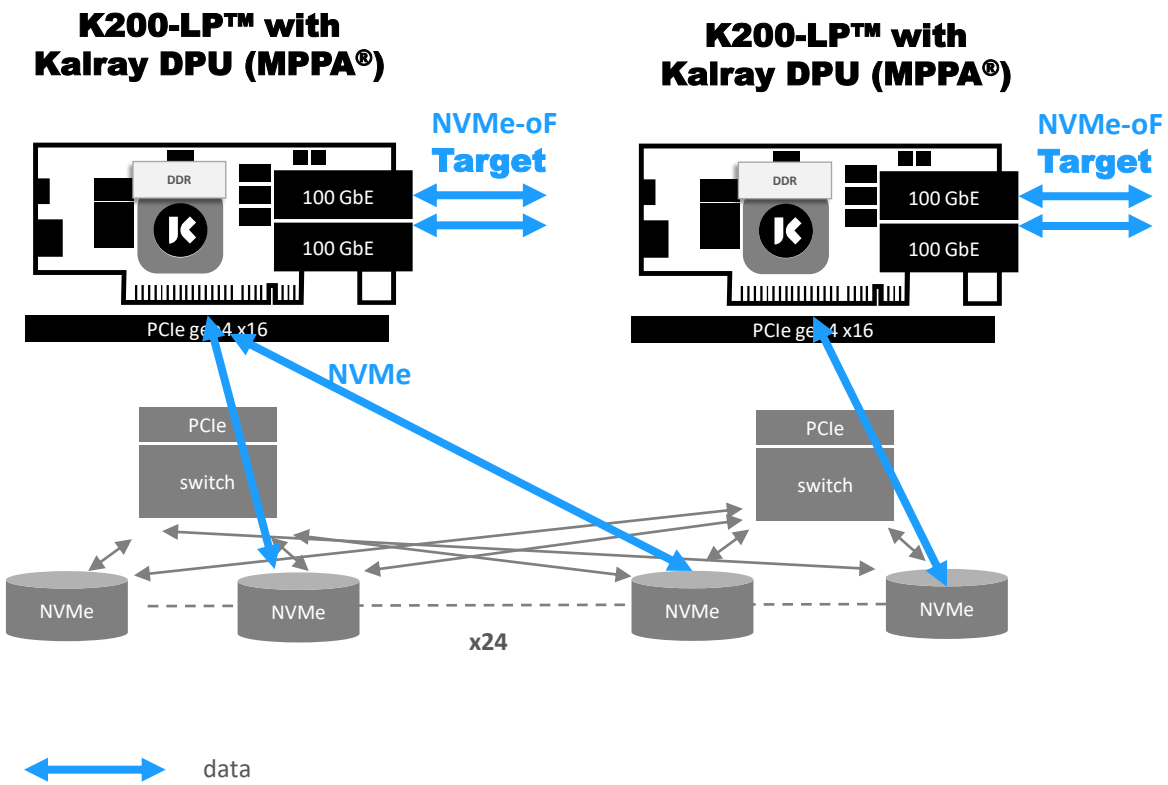
Architecture #2 – Standalone Mode



Stand Alone Storage Appliance

- DPU acts as standalone NVMe-oF target controller in storage node
- No x86 host attached needed
- Exposes local SSDs via NVMe-oF with data services:
  - pass-through mode
  - LVM
  - data protection
  - data reduction
  - data encryption

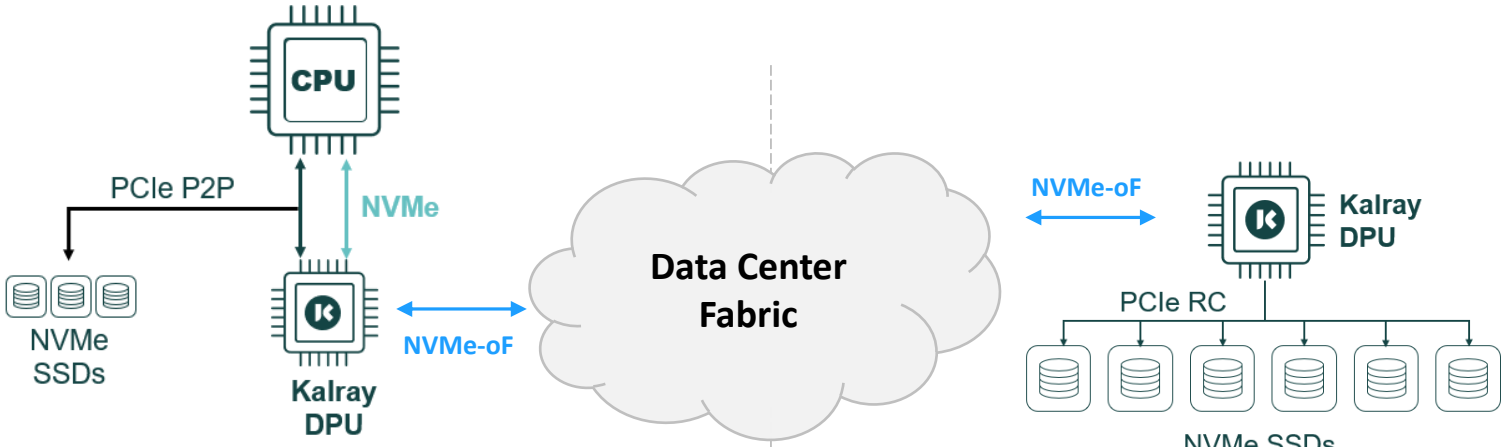
Architecture #2 - Data Path



High Availability Storage Appliance

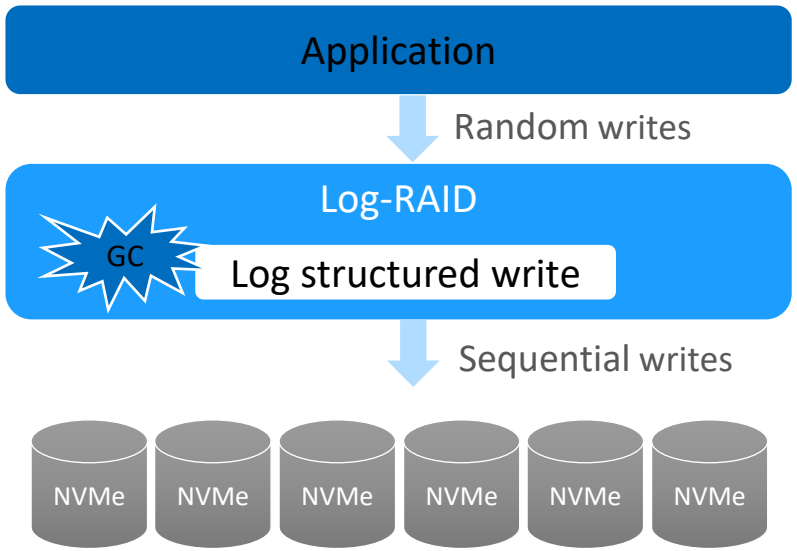
- 24x dual ported NVMe SSD controlled by 2 DPUs (active/passive mode)
- Storage volumes (NVMe pass-through or virtual volumes) exposed via NVMe-oF in both NVMe-TCP or RDMA mode
- Optional data services (RAID10, RAID6, compression) between NVMe-oF volumes and NVMe SSDs
- Fail-over mechanism to ensure data services continuity

DPUs on Both Ends of the Wire



Data Reduction	Compression/Dedup	
Data Protection		Erasure Encoding/RAID
Security	Encryption	
Data Analysis	AI/CSF/Video Processing/ DB Queries	AI/CSF/Video Processing/ DB Queries

Special Use Case-Offloading QLC ZNS Support



- Virtual “back-end” store exposed as contiguous write-log made of “segments” (several Mbytes each)
- Storage controller maintains a Page Mapping Table (PMT) for Virtual Block Address to Physical ones plus a local persistent cache (fast NVMe / NVRAM)
- Writes always occur on new blocks:
  - allows sequential writes
  - always full striping -> no RMW, no Write Hole
  - adapted to ZNS
  - reduced GC load on SSD
- Garbage collector retrieves free segments by freeing overwritten blocks (PMT modification)
- Large PMT shall be persistent: on demand paging from fast NVMe

Conclusions, Predictions, Observations

Flash Memory Summit

1 **DPU CARDS WILL REPLACE NICS** given the cost similarity and Value add of DPU over NIC

2 **CUSTOMIZABLE!**

Each DC formulation must determine how DPUs best fit into their topography

3 **DPU ASSISTED CPUs** overcome the disparity between CPUs and NVMe devices

4 **CPUs AREN'T GOING AWAY!**

KALRAY's Flashbox - Winner of a Most Innovative Flash HIGH SPEED Memory Technology Award

Kalray is a leader in DPU technology with our **3rd generation DPU** named MPPA® (Massively Parallel Processor Array)

- **80 cores -** with hardware acceleration and special data coprocessors
- **High Speed Interfaces -** x16 PCIe Gen 4, DDR4 memory and 2x 100Gbe Ethernet ports
- **Access Core Software -** programmable control and data plane
- Only 30W

Booth #940 Please visit us if you haven't stopped by already


Questions

Q&A

Flash Memory Summit

Please take the Session Survey
Thank you!

Tuesday, August 2

Wednesday, August 3

Thursday, August 4

35 YEARS OF NAND FLASH MEMORY INNOVATION