KALRAY
THE POWER OF MORE

APPLYING DPU TECHNOLOGY TO DISAGGREGATE AND ACCELERATE FLASH TECHNOLOGY

Tim Lieber,
Lead Solutions Architect

Flash Memory Summit 2022

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

DATA PROCESSING UNIT

A NEW CLASS OF PROGRAMMABLE PROCESSOR

Specialized in running datacenter infrastructure services

CPU, GPU … DPU

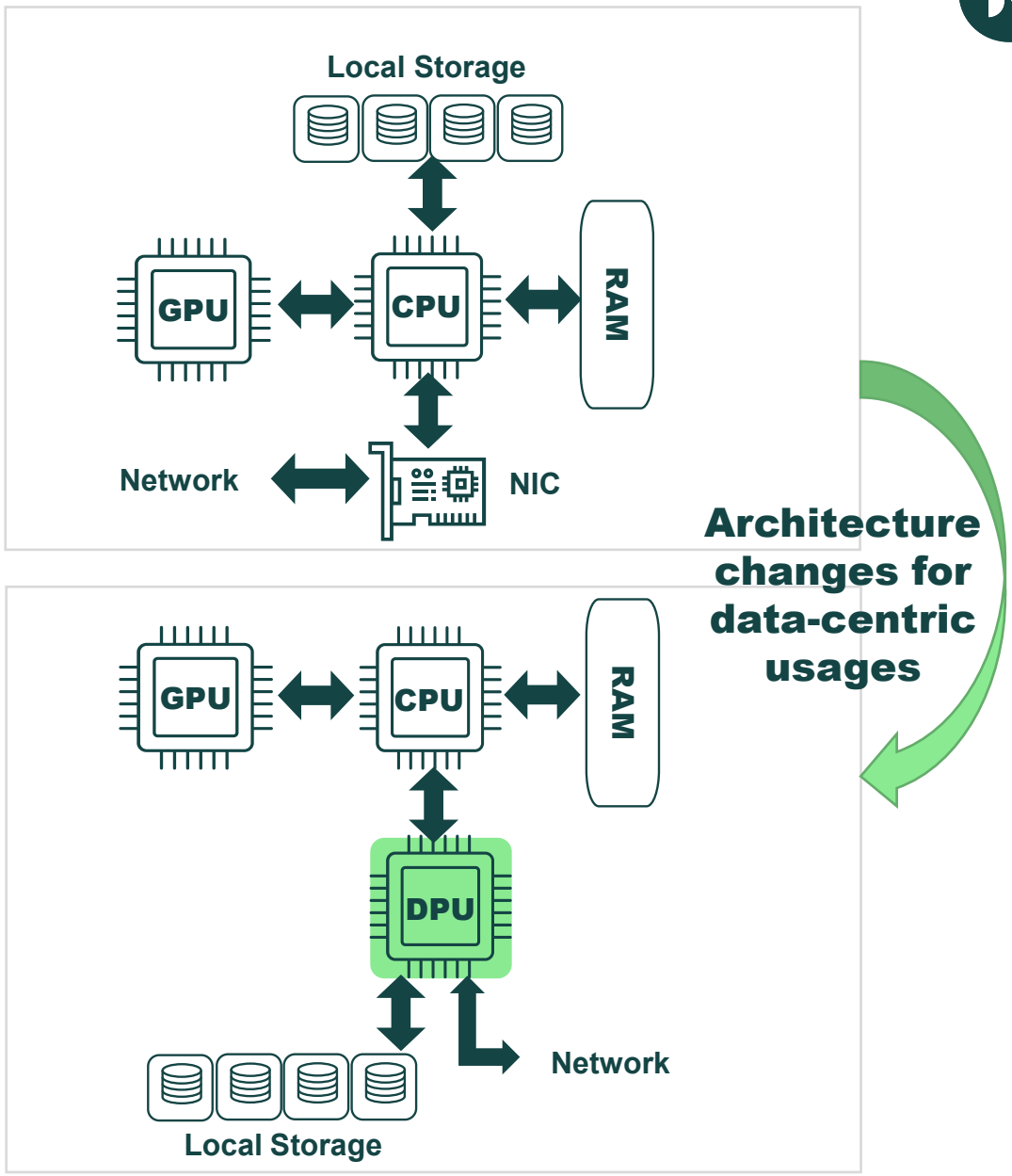
The 3rd socket in data centers alongside CPUs and GPUs

DPU at the core of new server architecture.

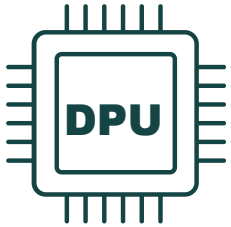
ACCELERATION

Accelerates software-defined datacenter infrastructure services
… and more !

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



KEY FEATURES OF A DPU WHICH MAKE IT BENEFICIAL TO DATA PROCESSING



FULLY PROGRAMMABLE

Management plane, control plane and data plane

PCIe

HIGH PERFORMANCE PCIe INTERFACE

SR-IOV for virtualization support

Network

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

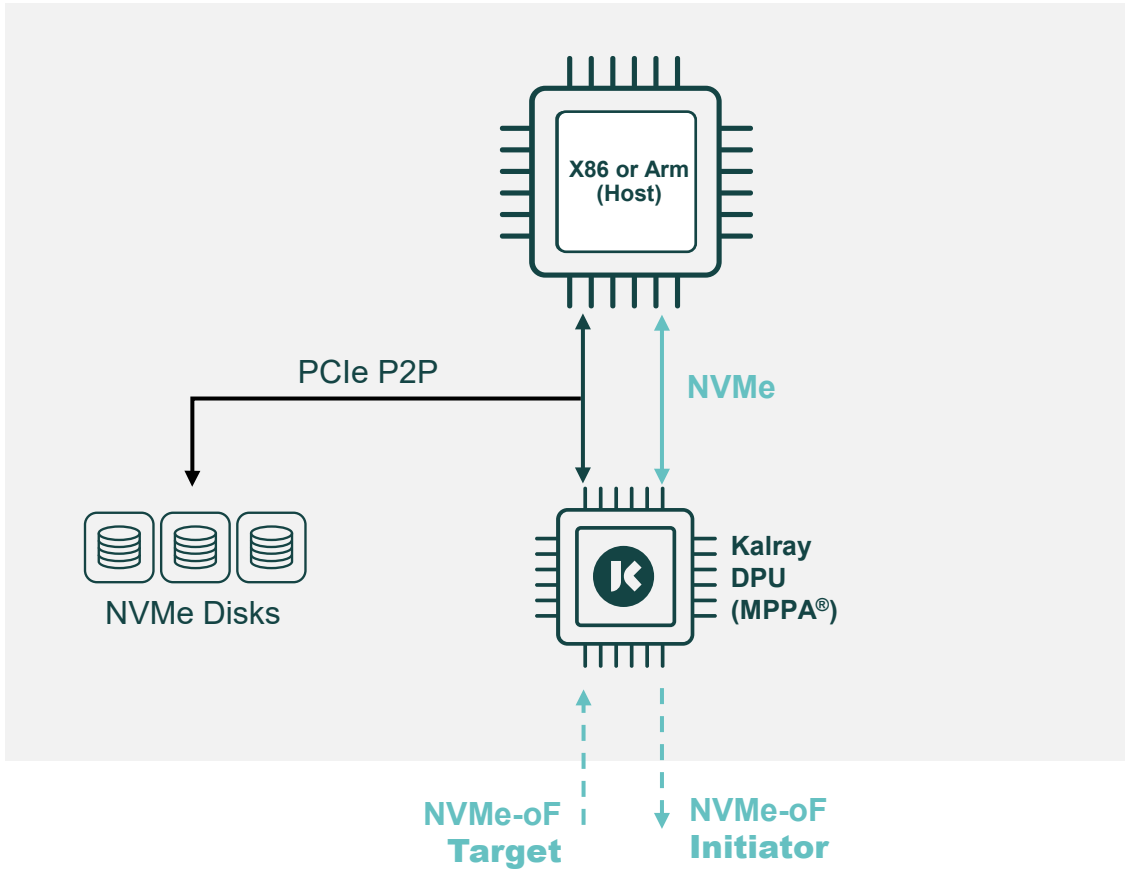
SECURITY

Root of trust, secure boot, secure firmware upgrades

ARCHITECTURE #1 - USING DPU AS COMPANION TO CPU

Storage Accelerator and Adapter

Use Case #1
In a Server

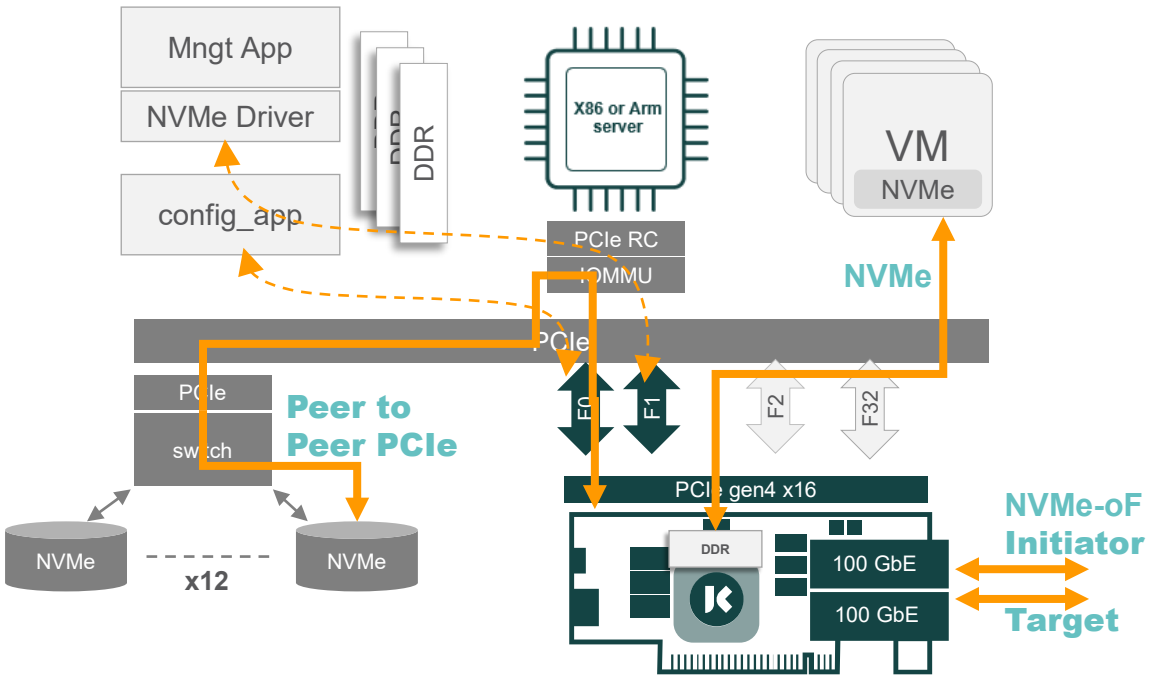


Compute Nodes/HCI

- 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

Use Case #1
In a Server



K200-LP™ with Kalray DPU (MPPA®)

Data Path

- **Local NVMe unbinded** 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
- **mgnt_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 Passthrough) exposing NVMe devices

ARCHITECTURE #1 - DATA SERVICE ACCELERATION

Inline or Look-aside data processing

Use Case #1
In a Server

Inline Data Processing

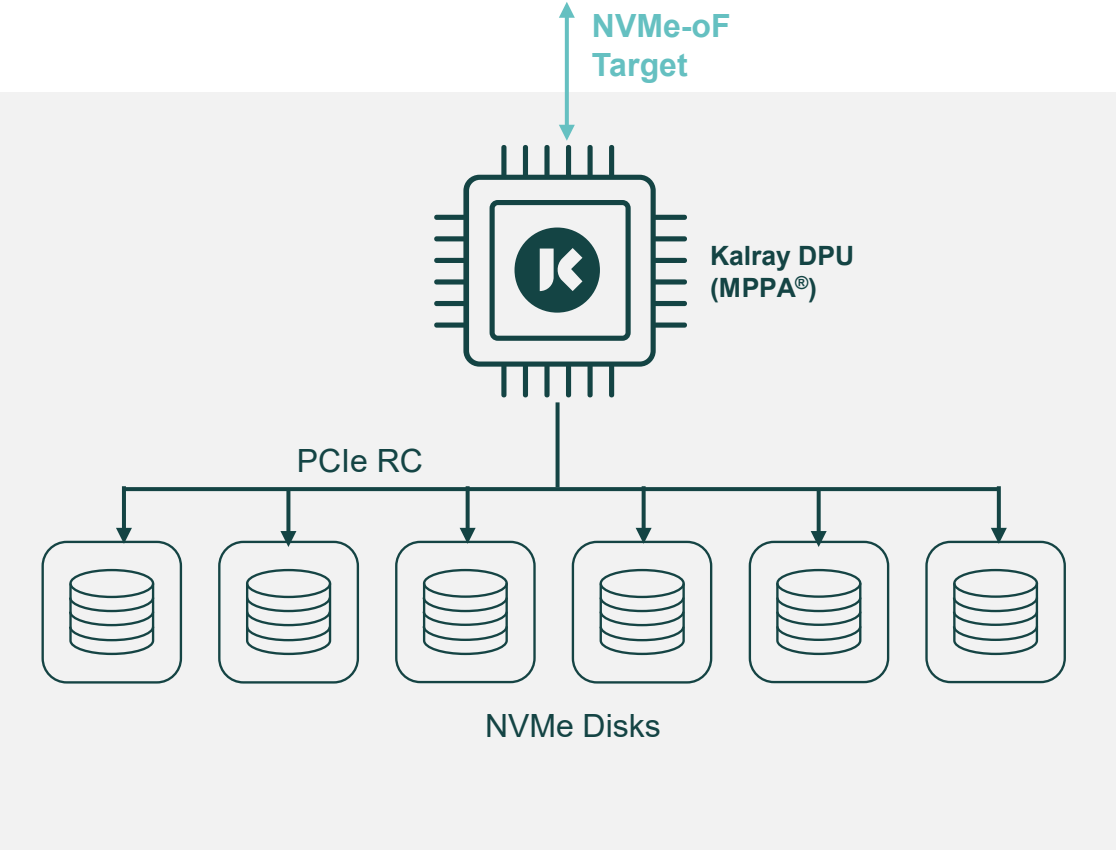
- Storage blocks processed in the storage path (local or remote)
- Interface to host is a ‘virtual’ NVMe volume
- Physical Backend Devices can be seamlessly local NVMe or remote (NVMe-oF)
- Typical Data Services :
  - Data Reduction : zero-detect, dedup, compression
  - Logical volume with thin-provisioning, 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 Backet Storage device
- Pseudo NVMe namespaces exposed to host with dedicated processing capabilities
- Typical Data Services :
  - Raw block processing : Crypto, Compression, EC
  - Non block processing (file/object) : AI, computer vision, NLP etc.

ARCHITECTURE #2 - USING DPU AS STAND ALONE

Use Case #2
Stand Alone
Mode



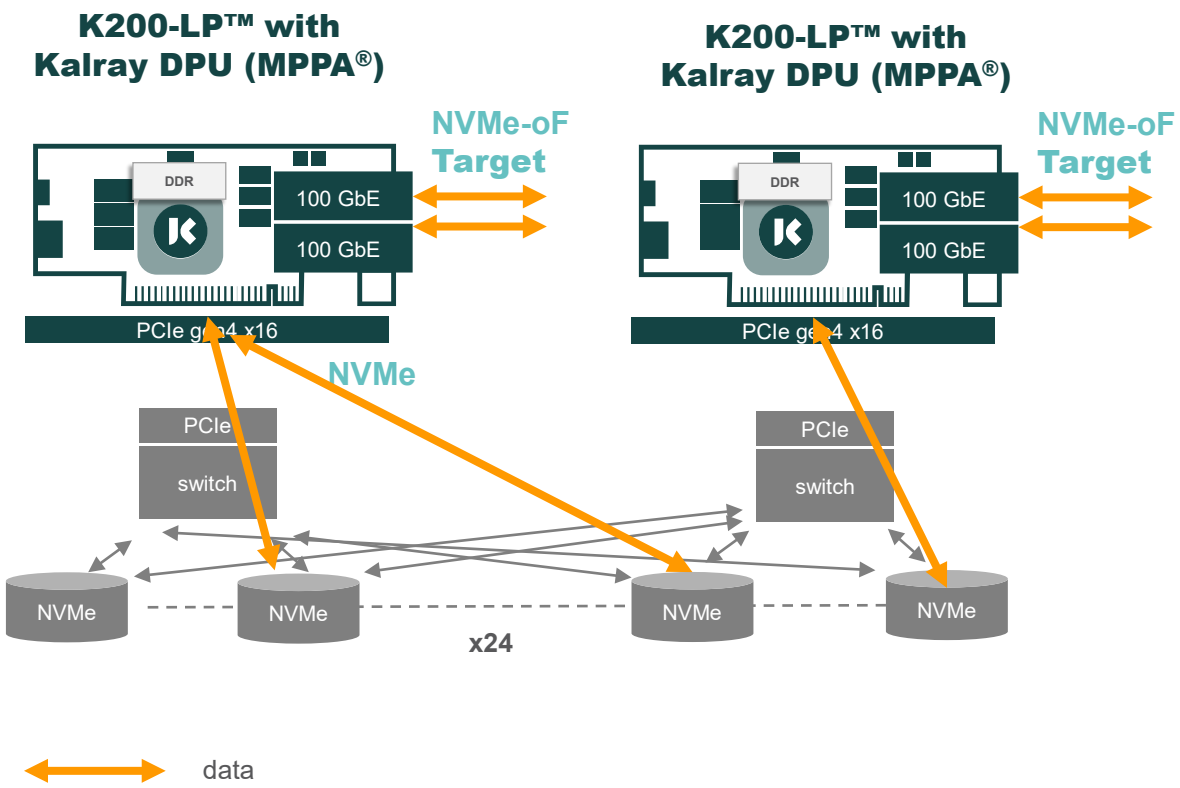
Stand Alone Storage Appliance

- 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:
  - Passthru
  - LVM
  - Data protection
  - Data reduction
  - Data encryption

ARCHITECTURE #2 - DATA PATH

High Availability

Use Case #2 Stand Alone Mode

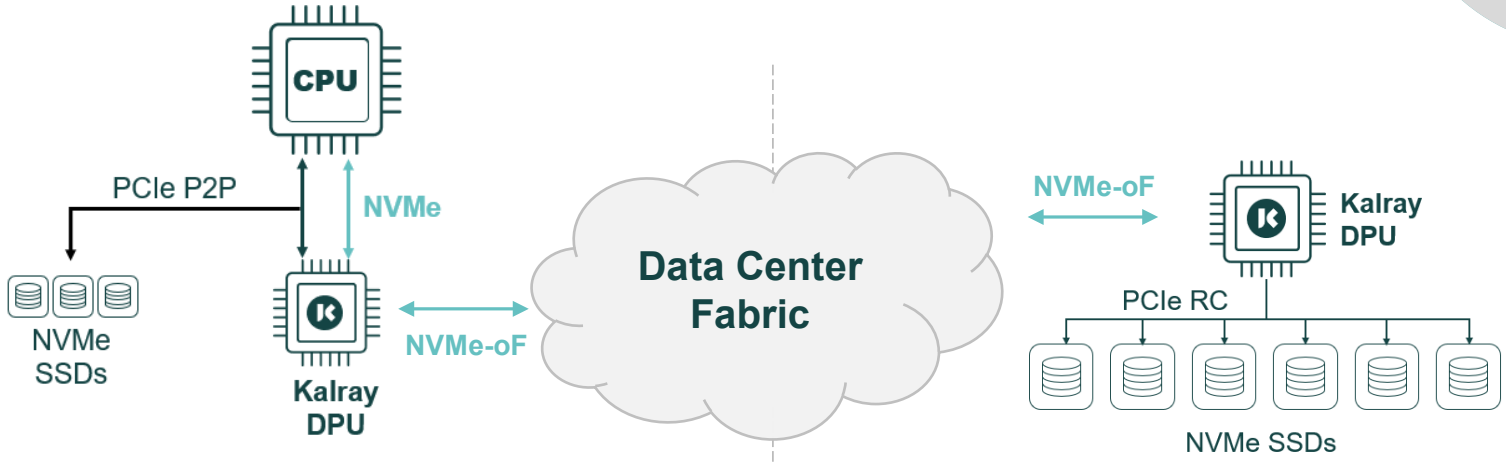


High Availability Storage Appliance

- 24x dual ported NVMe SSD controlled by 2 DPUs (active / passive mode)
- Storage volumes (NVMe passthrough 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

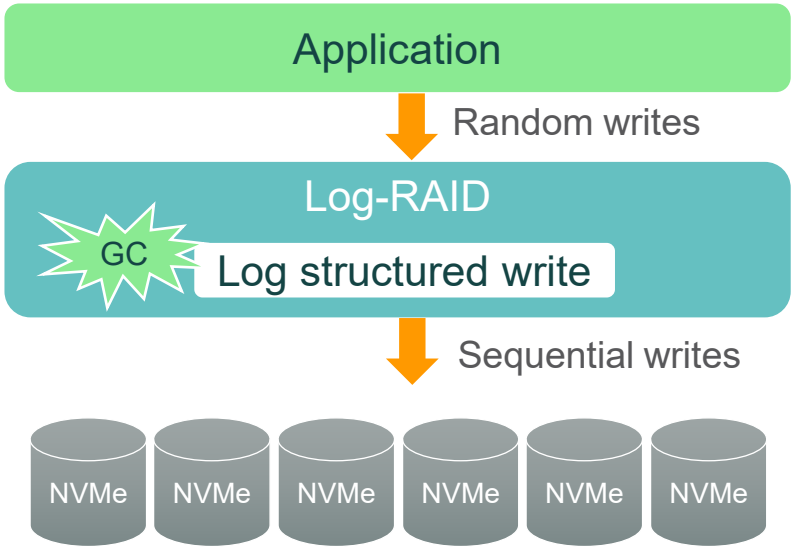
Use Case #3
Distributed Data Services



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 OF OFFLOADING ZNS BASED QLC SUPPORT

From Random to Sequential Writes



- 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 retrieve free segment by freeing overwritten blocks (PMT modification)
- Large PMT shall be persistent : on demand paging from fast NVMe

CONCLUSIONS, PREDICTIONS, OBSERVATIONS

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

2. **CUSTOMIZABLE!**
   Each DC formulation can 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 CANDIDATE FOR THE 3 MOST INNOVATIVE HIGH SPEED MEMORY TECHNOLOGY AWARDS!

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

- **80 cores** specialized in data processing with shared cache memory and special coprocessors and hardware accelerators to efficiently process data in all the many forms of today's modern data.
- **x16 PCIe Gen 4**, **DDR4 memory and 2x 100Gbe Ethernet ports**. We offer a software stack that has built-in data and control paths, is rich in data services that can offload a CPU and is highly programmable through the Kalray SDK.
- **Only 30W** nominal power.

THANK YOU

KALRAY

THE POWER OF MORE

www.kalrayinc.com

DISCLAIMER

Kalray makes no guarantee about the accuracy of the information contained in this document. It is intended for information purposes only and shall not be incorporated into any contract. It is not a commitment to deliver any material, code or functionality, and should not be relied upon in making purchasing decisions. The development, release and timing of any features or functionality described for Kalray products remains at the sole discretion of Kalray.

KALRAY

THE POWER OF MORE

www.kalrayinc.com