SDC 20

BY Developers FOR Developers

Storage Developer Conference
September 22-23, 2020

Next Generation Datacenters Require Composable Architecture Enablers and Programmable Intelligence

Jean-François MARIE
Chief Solution Architect
jfmarie@kalrayinc.com

KALRAY

Kalray at SDC20

Kalray is well represented this year at SDC with 4 sessions! Please have a look.

- **A NVMe-oF Storage Diode for Classified Data Storage**
  Jean-Baptiste Riaux, Sr Field Application Engineer
- **High-performance RoCE/TCP Solutions for End-to-end NVMe-oF Communication**
  Jean-François Marie, Chief Solution Architect
- **Next Generation Datacenters Require Composable Architecture Enablers and Programmable Intelligence**
  Jean-François Marie, Chief Solution Architect
- **Smart Storage Adapter for Composable Architectures**
  Rémy Gauguey, Sr Software Architect

Abstract

Abstract

For the past years flash drives have started to push performance boundaries. Storage OS based on x86 architectures even with more and more cores have a hard time to scale. Very few architectures can sustain the coming multi-million IOPS workloads expected from next generation Flash drives and memories. Only a multi-dimension scalable architecture can propose an alternative.

At the heart of it, parallel programming and ease of programming are requested. In this talk we will explain why it is important, what are the key components and how you could achieve such a performance. We will use Kalray’s Many Core processor and our SDK as an example to offload storage services such as NVMe-oF.

The Presenter

About the Presenter

Jean-François has more than 30 years of experience in the high tech industry. He started his career dealing with real time systems, before joining Sun Microsystems as a data center architect, then EMC² and finally NetApp in 2006, where he had various roles in a 13-year career. He held various roles, from Chief Technologist and Product Marketing Director for EMEA, to French Expert team manager to handle new technology introduction. He also managed global and regional accounts, alliances and partners.

Jean-François was also an active SNIA member for 10 years and French SNIA President for 2 years. He has a Masters degree in Electronics, specialized in micro-processor design and embedded systems.

On a personal note, he has been a Basket Ball player, a coach and head coach for 25 years.

Ubiquity Exists

I am delighted to run two sessions for this SDC20.

This one, and "High-performance RoCE/TCP solutions for end-to-end NVMe-oF communication."

With the magic of running a virtual event those have been scheduled at the same time today.

See you on Slack!
Jean-François

Why Composable Architectures?

The Data Processing Unit revolution

In the Data-Centric Era

Scale-out Data Center & micro-services based applications

Network traffic explosion
East-West traffic, multi-tenant, overlays...

Data Storage Capacity explosion
Storage spread across servers / disaggregation

Multi-tenant and **security** threat
Cryptography everywhere (storage, network…)

More and more **complex** data processing
AI, analytics …

General purpose CPUs/OS' inefficiencies

~25% of the servers **power** spent in data centric computation
Storage stack, network stack, crypto…

General Purpose CPUs inefficient for data centric computation
But Single threaded user applications

The Data Processing Unit revolution

In the Data-Centric Era

Scale-out Data Center & micro-services based applications

Network traffic explosion
East-West traffic, multi-tenant, overlays...

Data Storage Capacity explosion
Storage spread across servers / disaggregation

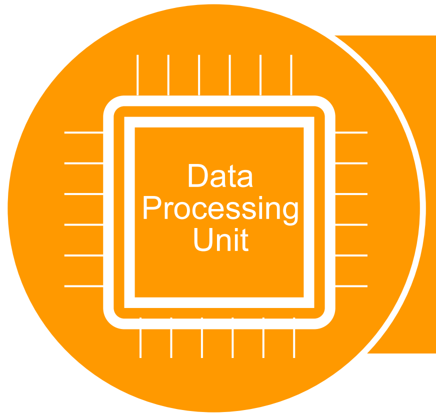
Multi-tenant and **security** threat
Cryptography everywhere (storage, network…)

More and more **complex** data processing
AI, analytics …

General purpose CPUs/OS' inefficiencies

~25% of the servers **power** spent in data centric computation
Storage stack, network stack, crypto…

General Purpose CPUs inefficient for data centric computation
But Single threaded user applications



Need for a new class of processing accelerator for these pre-dominant data-centric processing tasks!

Future Data Center Infrastructure challenges

The Route to Composable Infrastructure

❶ HCI
(Hyper Converged Infrastructure)

- Reduce complexity and hardware sprawl
- Reduce costs
- Increase agility and scalability

❷ Disaggregation

- Larger and larger datasets generated by Containerized applications and VMs
- Large diversity of application workloads

❸ Composable

- Any HW can be plugged into the system and expose new services to the others

Future Data Center Infrastructure challenges

The Route to Composable Infrastructure

❶ HCI
(Hyper Converged Infrastructure)

- Reduce complexity and hardware sprawl
- Reduce costs
- Increase agility and scalability

❷ Disaggregation

- Larger and larger datasets generated by Containerized applications and VMs
- Large diversity of application workloads

❸ Composable

- Any HW can be plugged into the system and expose new services to the others

HCI 2.0 architecture is a solution for HCI/Disaggregation …

Future Data Center Infrastructure challenges

The Route to Composable Infrastructure

❶ HCI
(Hyper Converged Infrastructure)

- Reduce complexity and hardware sprawl
- Reduce costs
- Increase agility and scalability

❷ Disaggregation

- Larger and larger datasets generated by Containerized applications and VMs
- Large diversity of application workloads

❸ Composable

- Any HW can be plugged into the system and expose new services to the others

HCI 2.0 architecture is a solution for HCI/Disaggregation …

… BUT ⚠

- Additional load on HCI cluster CPU by SW disaggregation
- Additional load on HCI cluster interconnect
- Storage Disaggregation is complex and expensive
- Clusters scalability limitation
- **HCI does not enable COMPOSABILITY**

Future Data Center Infrastructure challenges

The Route to Composable Infrastructure

❶ HCI
(Hyper Converged Infrastructure)

- Reduce complexity and hardware sprawl
- Reduce costs
- Increase agility and scalability

❷ Disaggregation

- Larger and larger datasets generated by Containerized applications and VMs
- Large diversity of application workloads

❸ Composable

- Any HW can be plugged into the system and expose new services to the others

HCI 2.0 architecture is a solution for HCI/Disaggregation …

… BUT

- Additional load on HCI cluster CPU by SW disaggregation
- Additional load on HCI cluster interconnect
- Storage Disaggregation is complex and expensive
- Clusters scalability limitation
- **HCI does not enable COMPOSABILITY**

Need a new approach for a true COMPOSABLE infrastructure!

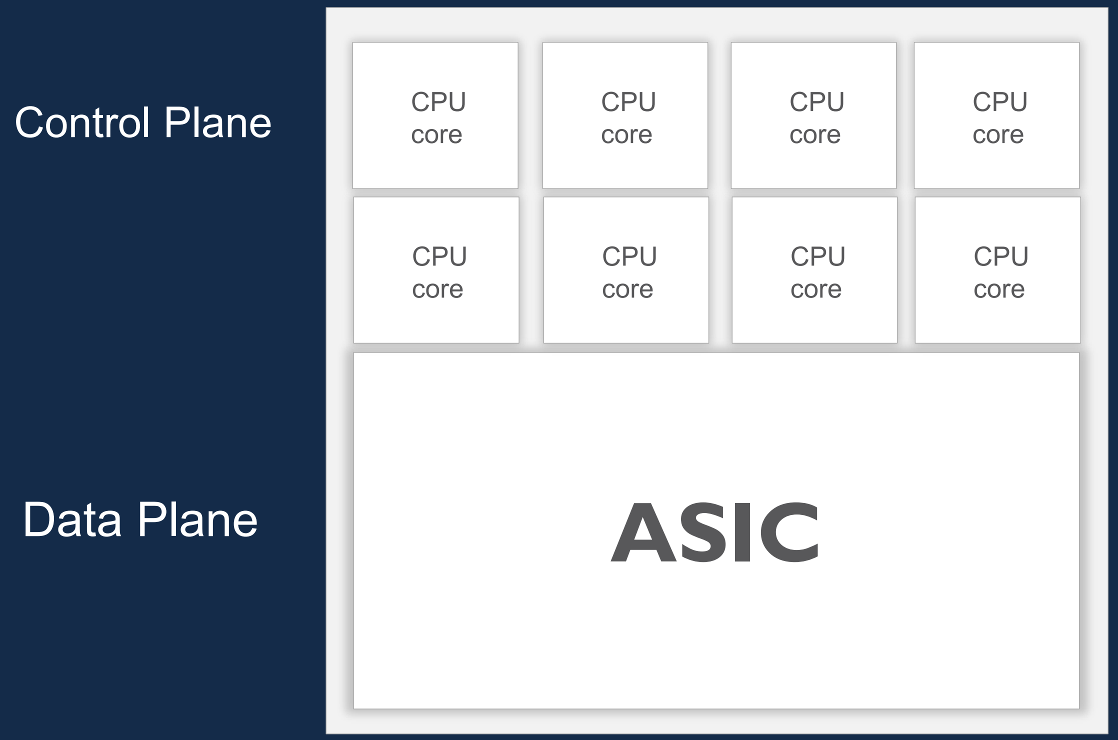
Composable Components

A new type of IO Processor

Coolidge™: The Ultimate I/O Processor

Why Coolidge is a Revolution vs Competition ?

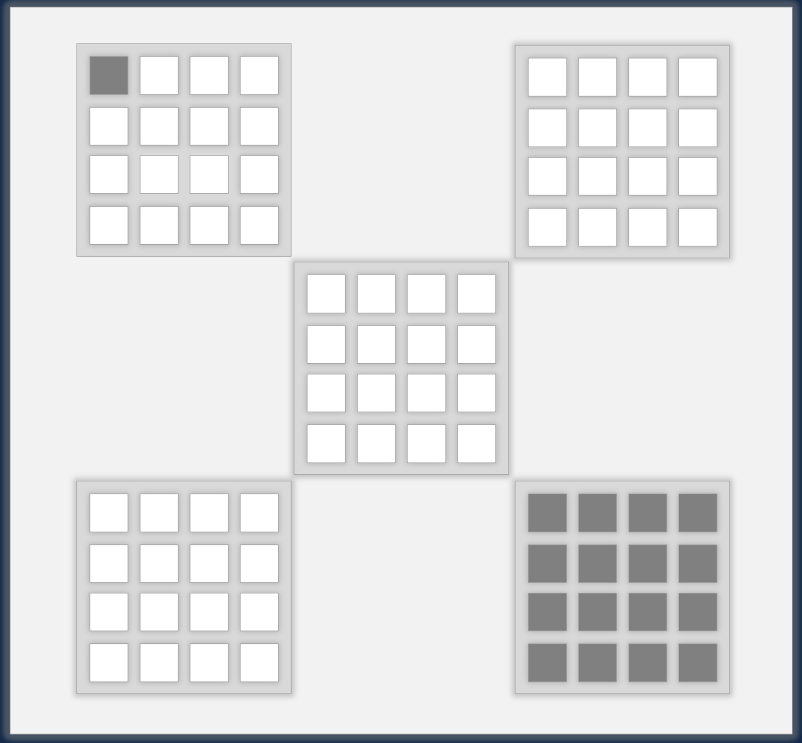
"xPU" Usual Approach



CONS ✗

- A few power hungry RISC CPU cores
- CPU flexibility limited to control plane
- Data plane is "hardwired" –
  No new services / no possible evolution!

Kalray's MPPA®3 Coolidge™



80 highly efficient VLIW independent **CPU** cores, gathered into **5 clusters**, running at **1.2Ghz,** connected to high speed fabrics & high speed interfaces.

PROS ✓

</> Fully programmable
Control Plane / Mgt Plane – Linux – 16 cores
Data Plane - 64 cores

Power efficiency
25W Typ

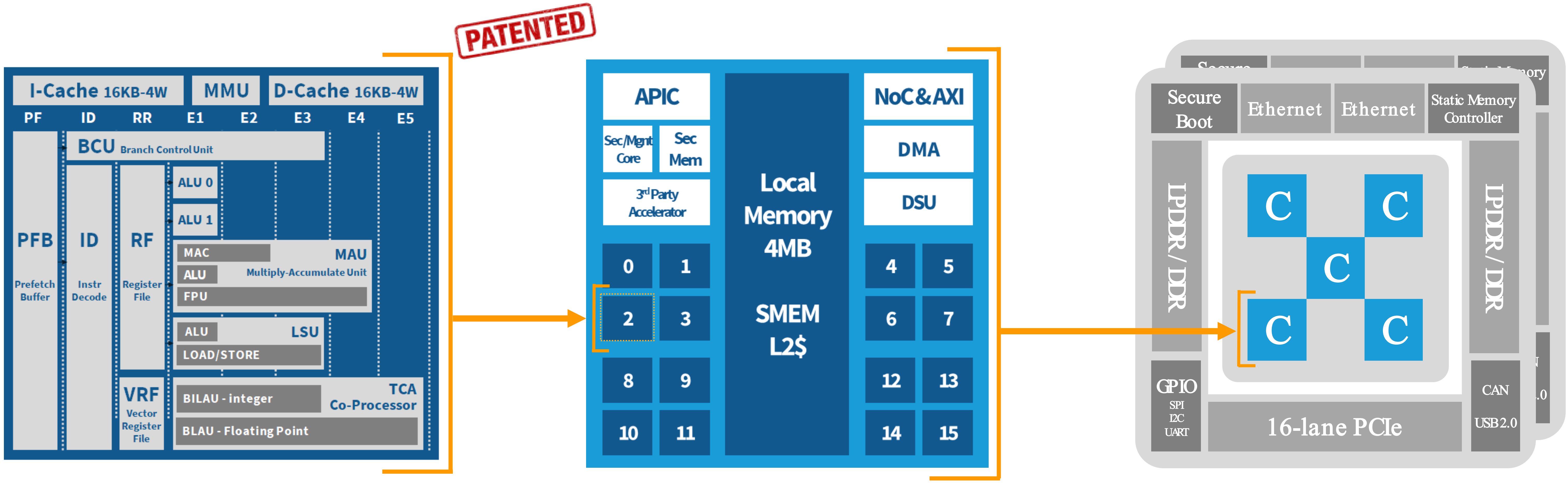
Top Performance Any workload
200KDMIPs, 25TOPS

High Speed I/O
2x100Gbs,PCIGen4,DDR4

Functional Isolation & Safety
Secure Islands, Encrypt/Decrypt, Secure Boot

MPPA® Coolidge™ Architecture

The I/O Processor for Next Gen Intelligent Systems



3RD GENERATION KALRAY CORE

- VLIW 64-bit core
- 6-issue VLIW architecture
- MMU + I&D cache (16KB+16KB)
- 16-bit/32-bit/64-bit IEEE 754-2008 FPU
- Vision/CNN Co-processor (TCA)

CLUSTER

Architecture

- 16 cores
- 1 safety/security dedicated core
- 600 to 1200 MHz

Memory

- L1 cache coherency (configurable)
- 4MB configurable memory (L2 cache)
- 256 bits / bandwidth up to 614GB/s)

MULTI CLUSTER ARCHITECTURE

5 Clusters: 80 cores + 80 co-processors

- Load Balancer / Packet Parser
- 2x100Gbps Ethernet
- PCIGen4
- DDR4 - 3200

AXI Bus + NoC Bus

- L2 refill in DDR and direct access to DDR from clusters
- DMA-based highly efficient data connection

Data Centric Computation

MPPA®3 Coolidge™ is the perfect fit

DATA CENTRIC WORKLOAD CHARACTERISTICS	DATA PROCESSING UNIT REQUIREMENTS	KALRAY'S MPPA®3 COOLIDGE™
High parallelism Many stateless or stateful contexts : TCP/IP, TLS, IPsec sessions , NVMe queues	**Manycore (MIMD) architecture**	✓ **- 80 VLIW cores @ 1.2 GHz** **- 5 Clusters x16 cores**
Short temporal data locality Complex memory hierarchy L1/L2/L3 not well suited	**Large on chip memory (TCM)** - With large bandwidth - Simple and deterministic memory subsystem	
I/O intensive High IOPS and GB/s, low latency	**- Optimized interconnect** High bandwidth, low latency & deterministic on chip **- High speed interfaces**	
Computational intensive Inline AI inference, analytics, crypto, erasure coding…	**- Floating Point Unit** **- AI acceleration** **- Cryptographic acceleration** **- Erasure Coding acceleration**	
Variability and flexibility Programmability / flexibility (C, C++, standard APIs)	**- C / C++ programmable data plane** **- Standard APIs**	

Data Centric Computation

MPPA®3 Coolidge™ is the perfect fit

DATA CENTRIC WORKLOAD CHARACTERISTICS	DATA PROCESSING UNIT REQUIREMENTS	KALRAY'S MPPA®3 COOLIDGE™
High parallelism Many stateless or stateful contexts : TCP/IP, TLS, IPsec sessions , NVMe queues	**Manycore (MIMD) architecture**	✓ **- 80 VLIW cores @ 1.2 GHz** **- 5 Clusters x16 cores**
Short temporal data locality Complex memory hierarchy L1/L2/L3 not well suited	**Large on chip memory (TCM)** - With large bandwidth - Simple and deterministic memory subsystem	✓ **- 20 MB TCM** **- 5 isolated clusters with $L2**
I/O intensive High IOPS and GB/s, low latency	**- Optimized interconnect** High bandwidth, low latency & deterministic on chip **- High speed interfaces**	
Computational intensive Inline AI inference, analytics, crypto, erasure coding…	**- Floating Point Unit** **- AI acceleration** **- Cryptographic acceleration** **- Erasure Coding acceleration**	
Variability and flexibility Programmability / flexibility (C, C++, standard APIs)	**- C / C++ programmable data plane** **- Standard APIs**	

Data Centric Computation

MPPA®3 Coolidge™ is the perfect fit

DATA CENTRIC WORKLOAD CHARACTERISTICS	DATA PROCESSING UNIT REQUIREMENTS	KALRAY'S MPPA®3 COOLIDGE™
High parallelism Many stateless or stateful contexts : TCP/IP, TLS, IPsec sessions , NVMe queues	**Manycore (MIMD) architecture**	✓ - 80 VLIW cores @ 1.2 GHz - 5 Clusters x16 cores
Short temporal data locality Complex memory hierarchy L1/L2/L3 not well suited	**Large on chip memory (TCM)** - With large bandwidth - Simple and deterministic memory subsystem	✓ - 20 MB TCM - 5 isolated clusters with $L2
I/O intensive High IOPS and GB/s, low latency	**- Optimized interconnect** High bandwidth, low latency & deterministic on chip **- High speed interfaces**	✓ - High perf. NoC - 2x100 Gbps Ethernet - PCIe x16 Gen4 (RC/EP)
Computational intensive Inline AI inference, analytics, crypto, erasure coding…	**- Floating Point Unit** **- AI acceleration** **- Cryptographic acceleration** **- Erasure Coding acceleration**	
Variability and flexibility Programmability / flexibility (C, C++, standard APIs)	**- C / C++ programmable data plane** **- Standard APIs**	

Data Centric Computation

MPPA®3 Coolidge™ is the perfect fit

DATA CENTRIC WORKLOAD CHARACTERISTICS	DATA PROCESSING UNIT REQUIREMENTS	KALRAY'S MPPA®3 COOLIDGE™
High parallelism Many stateless or stateful contexts : TCP/IP, TLS, IPsec sessions , NVMe queues	**Manycore (MIMD) architecture**	✓ **- 80 VLIW cores @ 1.2 GHz** **- 5 Clusters x16 cores**
Short temporal data locality Complex memory hierarchy L1/L2/L3 not well suited	**Large on chip memory (TCM)** - With large bandwidth - Simple and deterministic memory subsystem	✓ **- 20 MB TCM** **- 5 isolated clusters with $L2**
I/O intensive High IOPS and GB/s, low latency	**- Optimized interconnect** High bandwidth, low latency & deterministic on chip **- High speed interfaces**	✓ **- High perf. NoC** **- 2x100 Gbps Ethernet** **- PCIe x16 Gen4 (RC/EP)**
Computational intensive Inline AI inference, analytics, crypto, erasure coding…	**- Floating Point Unit** **- AI acceleration** **- Cryptographic acceleration** **- Erasure Coding acceleration**	✓ **- Up to 1.15TFLOPs (SP)** **- Up to 4.2TFLOPS (half precision)** **- Up to 25 TOPs (8bits) for AI** **- 100Gbps+ Crypto acc.** **- Line rate Reed Solomon**
Variability and flexibility Programmability / flexibility (C, C++, standard APIs)	**- C / C++ programmable data plane** **- Standard APIs**	

Data Centric Computation

MPPA®3 Coolidge™ is the perfect fit

DATA CENTRIC WORKLOAD CHARACTERISTICS	DATA PROCESSING UNIT REQUIREMENTS	KALRAY'S MPPA®3 COOLIDGE™
High parallelism Many stateless or stateful contexts : TCP/IP, TLS, IPsec sessions , NVMe queues	**Manycore (MIMD) architecture**	✓ **- 80 VLIW cores @ 1.2 GHz** **- 5 Clusters x16 cores**
Short temporal data locality Complex memory hierarchy L1/L2/L3 not well suited	**Large on chip memory (TCM)** - With large bandwidth - Simple and deterministic memory subsystem	✓ **- 20 MB TCM** **- 5 isolated clusters with $L2**
I/O intensive High IOPS and GB/s, low latency	**- Optimized interconnect** High bandwidth, low latency & deterministic on chip **- High speed interfaces**	✓ **- High perf. NoC** **- 2x100 Gbps Ethernet** **- PCIe x16 Gen4 (RC/EP)**
Computational intensive Inline AI inference, analytics, crypto, erasure coding…	**- Floating Point Unit** **- AI acceleration** **- Cryptographic acceleration** **- Erasure Coding acceleration**	✓ **- Up to 1.15TFLOPs (SP)** **- Up to 4.2TFLOPS (half precision)** **- Up to 25 TOPs (8bits) for AI** **- 100Gbps+ Crypto acc.** **- Line rate Reed Solomon**
Variability and flexibility Programmability / flexibility (C, C++, standard APIs)	**- C / C++ programmable data plane** **- Standard APIs**	✓ **- Linux, OpenDataPlane** **- SPDK BDEVs, NVMe** **- OpenCL**

Smart IO Adapter

Kalray Smart Storage Adapter

AccessCore ™ Storage framework on MPPA® to deliver Data Services

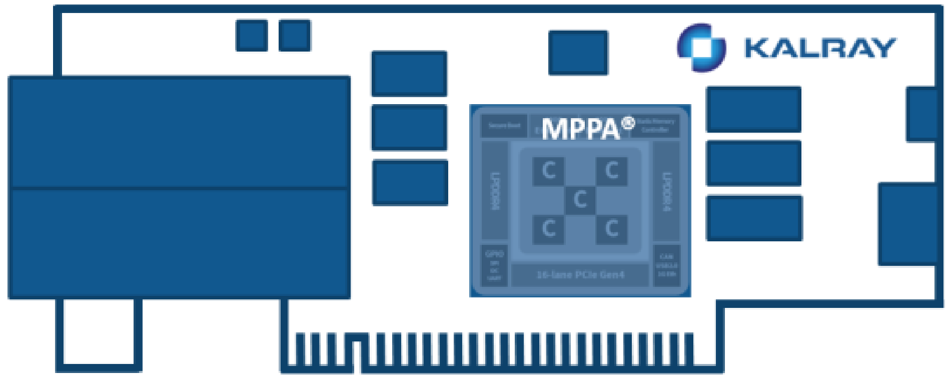
SDK

iol Certified

nvm EXPRESS®

Standard
NVMe-oF
TCP & RoCE

Storage cluster interconnect



Management through GbE

SNIA Swordfish

DMTF Redfish

Ease of integration on a x86 node via SR-IOV
NVMe Emulation

Drives access through PCIe
(RC or P2P)

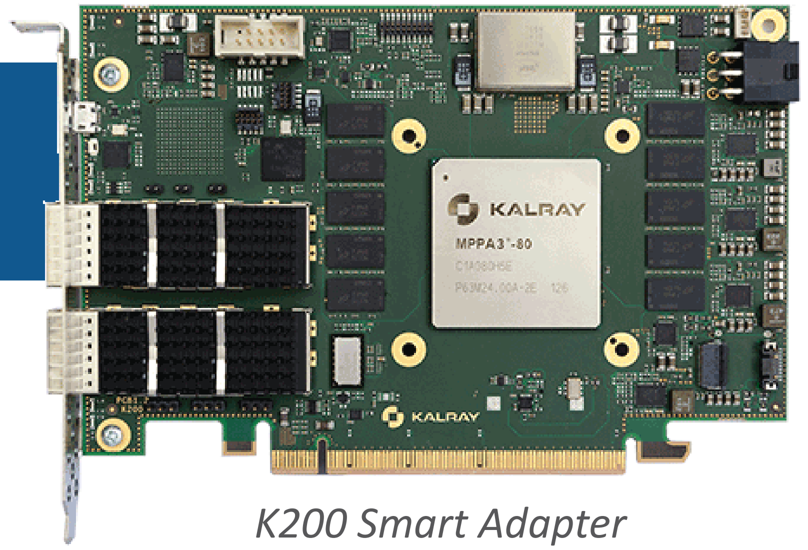
nvm EXPRESS®

Serial Attached SCSI

Kalray Smart Storage Adapter Solution

K200/ K200-LP & ACS SDK

K200 & K200-LP
manufactured by wistron



K200 Smart Adapter

2 Form Factors
- FHHL (Full Height) - K200 - Single Slot
- HHHL (Low Profile) - K200-LP
  Single or Double Slots

Manycore Architecture
- 80 VLIW cores @ 1.2 Ghz
- 5 Clusters x16 cores

High Speed Ethernet
- 2x100GbE / 8x25 GbE

Certified NVMe-oF Stack
- NVMe-oF 1.1 (Target, Intiator)
- RoCE v1/v2, TCP

Advanced SSD interface
- PCIe-Gen4
- NVMe 1.1 to 1.4 SSDs
  No need for CMB
- Dual port SSD support

2 Modes
- Stand-alone
- Host CPU co-processor
  / "host-agnostic" support

Agnostic Host Support
- NVMe Driver

DDR-3200
- 8GB to 32GB

H/W Accelerators
- Encryption / Decryption
- Hashing (SHA-256, SHA-3)
- Erasure Coding

Low Power
- 35W (single slot)
- 65W (double slot)

Key figures (per card)
- Random R/W RoCE: **4-6 MIOPS**
- Random R/W TCP: **2-4 MIOPS**
- Sequential R/W (RoCE&TCP): **25GB/s**
- Latency (RoCE/TCP): **10 /30 usec**

SDK AccessCore®
Open Software & Tools

Open Software Environment
- Linux / SPDK Control Plane (16 Cores)
- Fully Programmable Data Plane (64 Cores)
- Storage, Network and Compute Services
  (AI,DSP,NVMe,NVMe-oF,ROCE,TCP, RAID, de-dup,..)

Agnostic Host Support
- NVMe Driver

+ Extra compute available
- @ 3MIOPS, 50% cores available !
- Storage Services (RAID, de-dedup ...)
- AI
- Analytics ...

An Opened Storage Stack

AccessCore® for Storage & Networking

ACS4.x architecture highlights

SDK AccessCore®
Open Software & Tools

PROGRAMMABILITY

- Full programmability on data, control & management planes
  - Control & Management plane : Linux (typical : 1 Cluster - 16 cores)
  - Data plane : Cluster OS (light POSIX OS) (typical: 1 to 4 Clusters – 16 to 64 cores)

EFFICIENCY

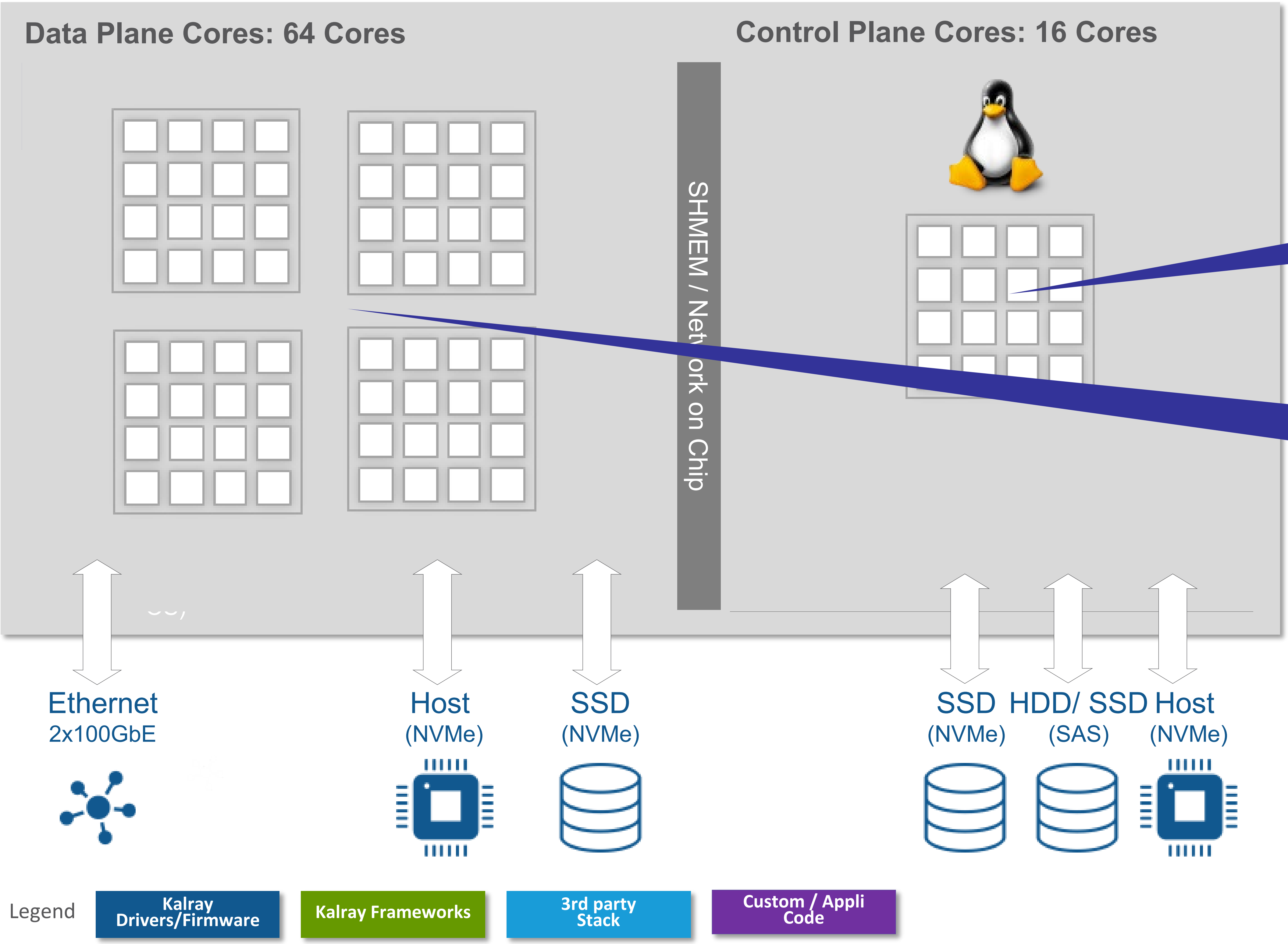
- Run to completion full dataplane
  - From network functions to NVMe stack on light OS cores
- True inline processing
  - No need for x86 pre/post processing

STANDARDIZED

- Hardware interfaces
  - NVMe emulation
- Software APIs & tool chain
  - Linux APIs: SPDK, virtio, ibverbs …
  - Data plane APIs: sockets, SPDK nvme lib, SPDK BDEV, ODP
  - Librairies : ISA-L, Buildroot, binutils

AccessCore®

A fully flexible software environment

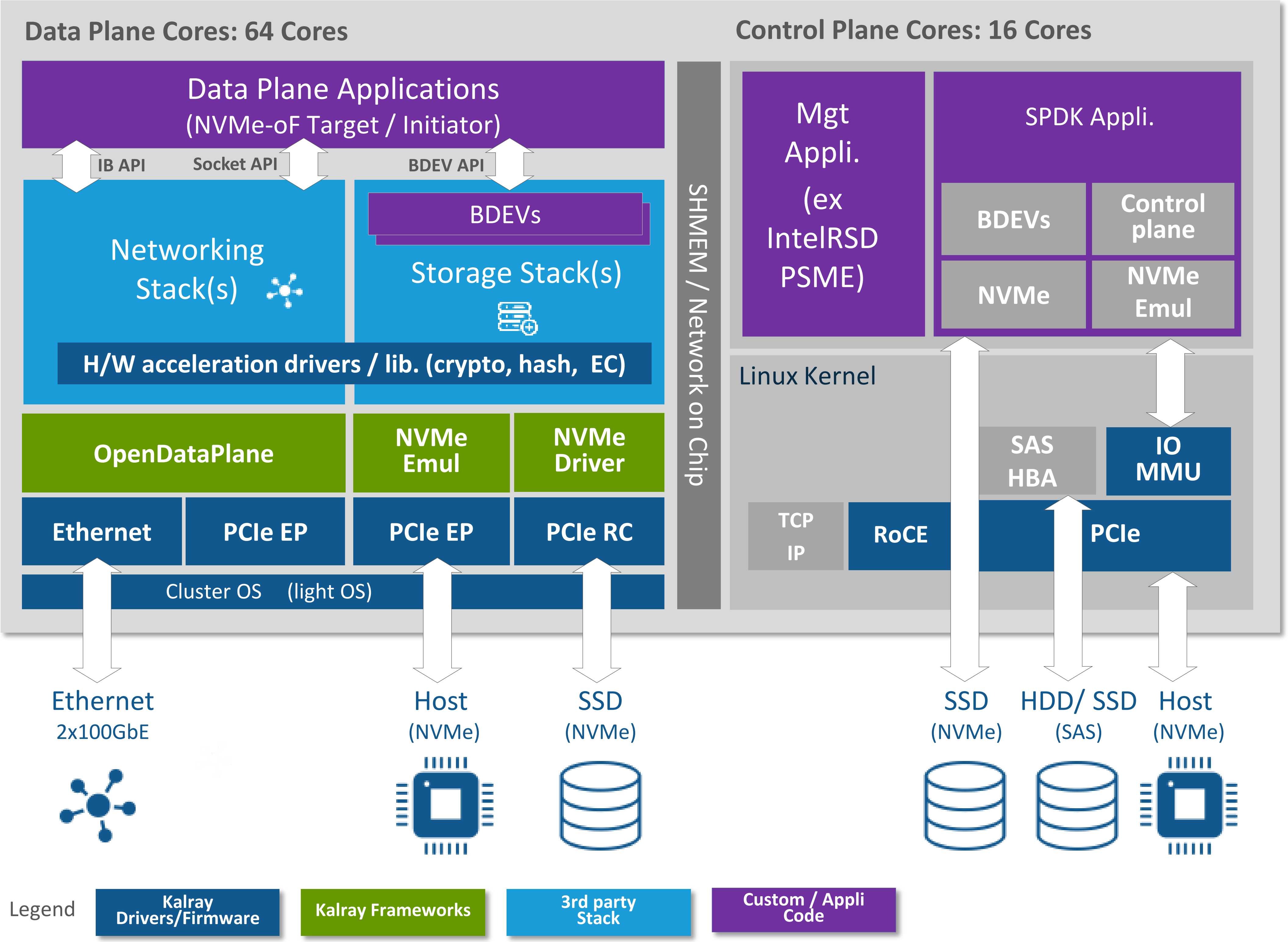


Linux on control plane cluster (16 cores)

ClusterOS on 4 data plane clusters (64 cores)

AccessCore®

A fully flexible software environment



- A complete & modular software framework
- Based on an optimized SPDK for both data plane **AND** control plane
- Open to partners

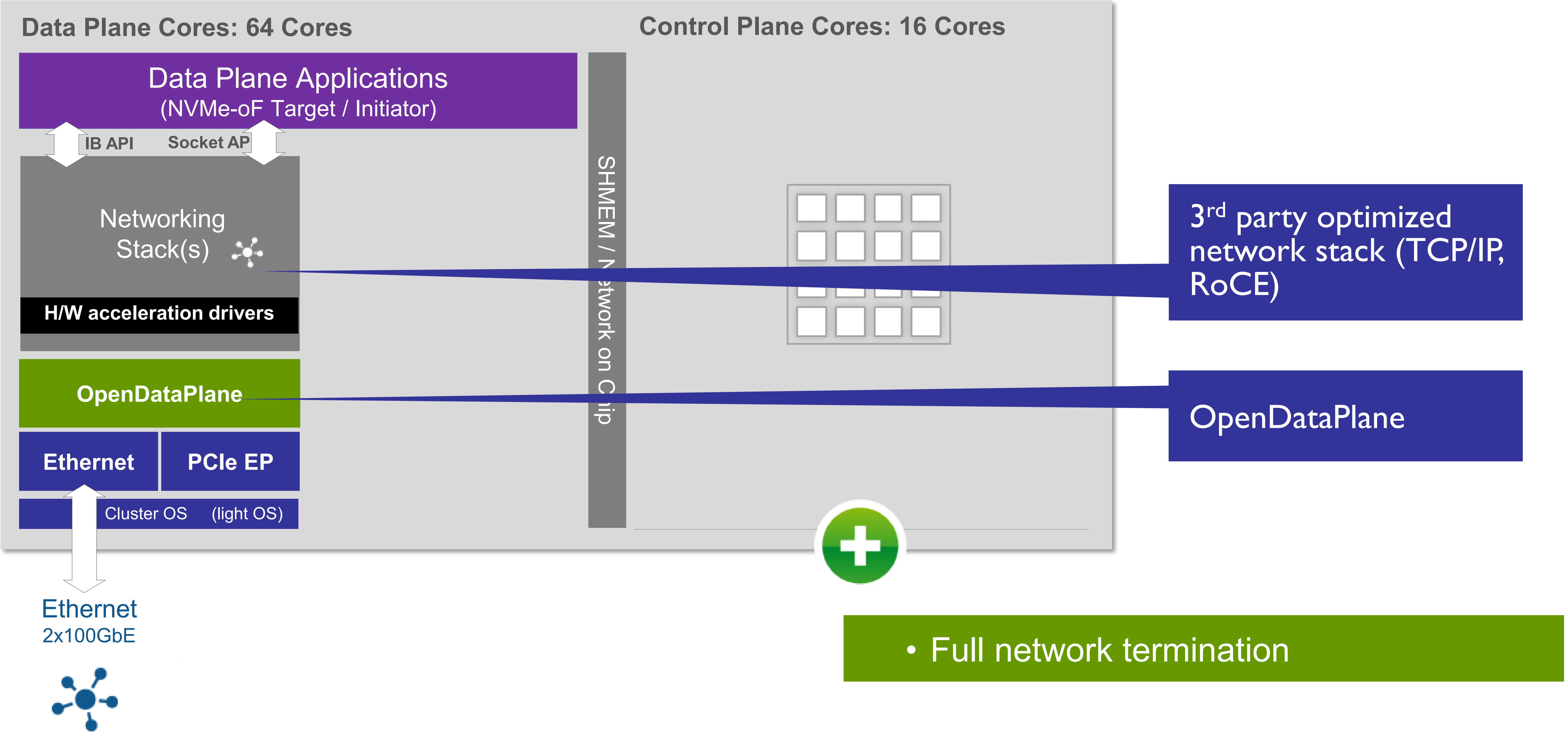
Offloads

AccessCore® Open Software & Tools

SDC 20 KALRAY

AccessCore®

NVMe-oF offloading



- Full network termination

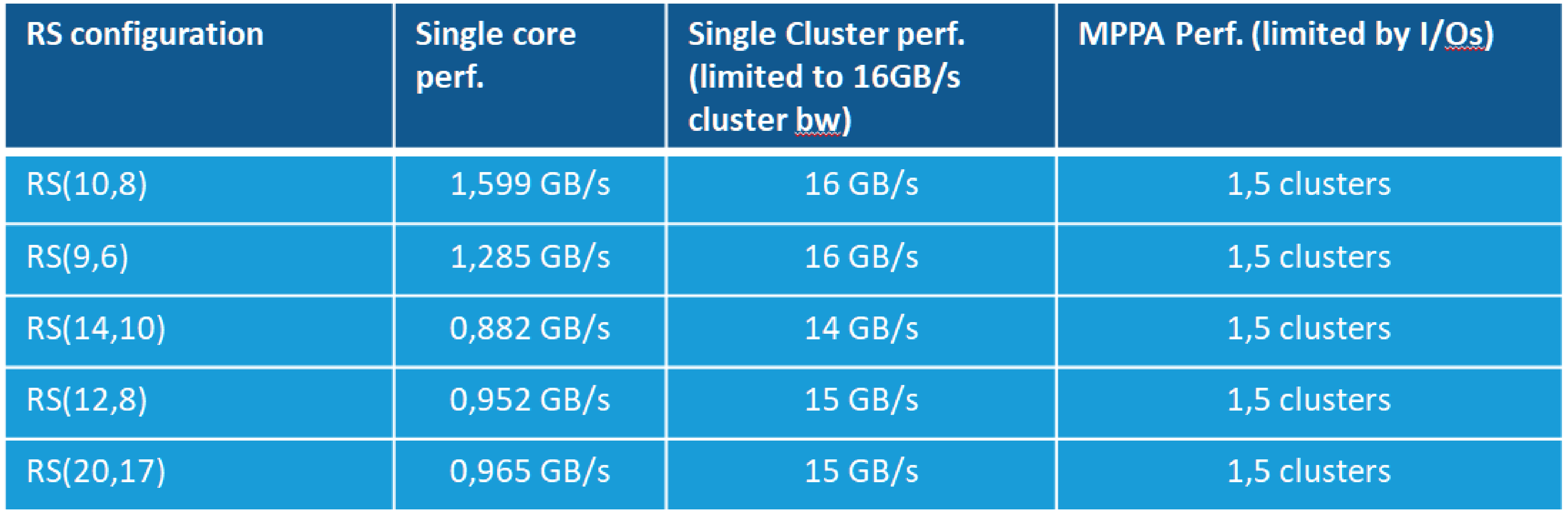
Legend				
	Kalray Drivers & Firmware	Kalray Frameworks	3rd party Stack	Custom / Appli Code

SDK AccessCore®
Open Software & Tools

AccessCore®

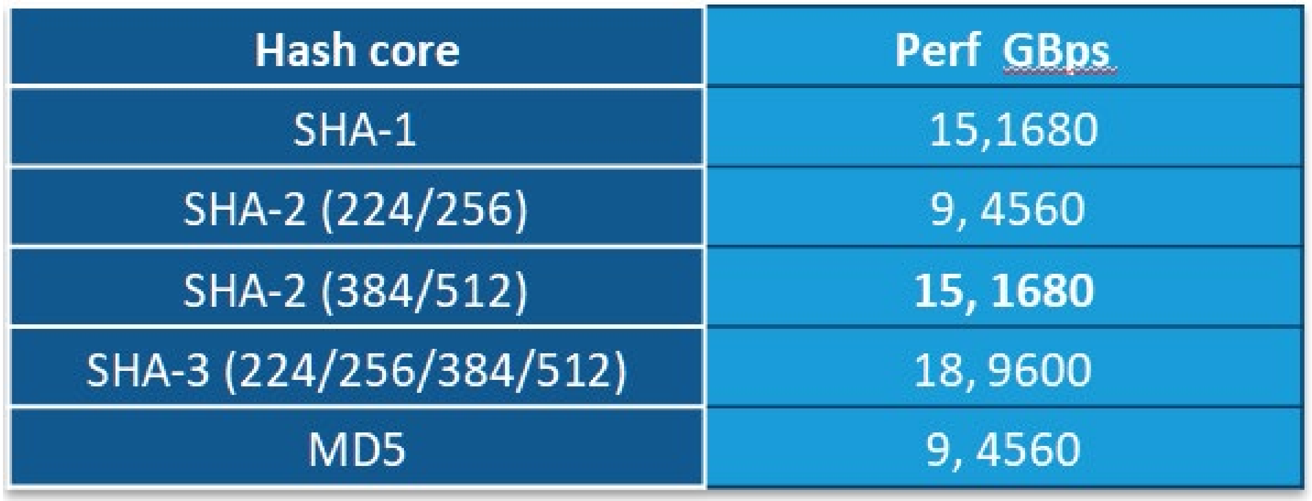
Other offloads

- Intel ISA-L compatible library
- Kalray (patent pending) optimized code based specific Bit Matrix Multiplication instructions



RS configuration	Single core perf.	Single Cluster perf. (limited to 16GB/s cluster bw)	MPPA Perf. (limited by I/Os)
RS(10,8)	1,599 GB/s	16 GB/s	1,5 clusters
RS(9,6)	1,285 GB/s	16 GB/s	1,5 clusters
RS(14,10)	0,882 GB/s	14 GB/s	1,5 clusters
RS(12,8)	0,952 GB/s	15 GB/s	1,5 clusters
RS(20,17)	0,965 GB/s	15 GB/s	1,5 clusters

- Inline or look-aside object/block hashing acceleration



Hash core	Perf GBps
SHA-1	15,1680
SHA-2 (224/256)	9, 4560
SHA-2 (384/512)	**15, 1680**
SHA-3 (224/256/384/512)	18, 9600
MD5	9, 4560

In Summary

Toward a true & efficient composable disaggregated Infrastructure

HIGHER PERFORMANCE

- Leverage Kalray cards performance and exploit full NVMe SSD capabilities
- Offload x86 from heavy storage stacks

LOWER COST

- Switch to a true **C**omposable **D**isaggregated **I**nfrastructure with commodity components
- Optimize HCI nodes efficiency

FULLY FLEXIBLE

- Fully programmable data plane
- Data Plane additional storage services based on SPDK framework (EC, caching…)

FUTURE PROOF

- Leverage standard NVMe-oF protocols
- Compliant with other NVMe-oF appliances
- Ease of in-the-field update

Please take a moment to rate this session.

Your feedback matters to us.