SDC 20

BY Developers FOR Developers

Storage Developer Conference
September 22-23, 2020

High-Performance RoCE/TCP Solutions for End-to-end NVMe-oF Communication

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

Exploiting the full SSD performance in scalable disaggregated architectures is a continuous challenge. NVMe/TCP, released in 2018, enables a broader sharing of distributed storage resources. It complements NVME-oF over RDMA, avoiding performance degradation over distant links and simplifying the deployment. However, this comes at the cost of a heavy networking stack and requires the latest Linux kernels.

In this talk, we will analyze the differences between RoCE and TCP, and show how to eliminate bottlenecks, achieving best-in-class performance for both protocols in an end-to-end NVMe-oF communication. We will demonstrate also how this **solution can be OS agnostic**, ensuring a seamless integration of NVMe-oF in today's datacenter.

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 ”Next Generation Datacenters require composable architecture enablers and programmable intelligence.”

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

See you on Slack!
Jean-François

A bit of History

A bit of History

- 3200
- 5400
- 7200
- 10000
- 15000

What are those numbers ?

A bit of History

- 3200
- 5400
- 7200
- 10000
- 15000

HDD Disk RPM

A bit of History

- 75
- 100
- 300

And those ?

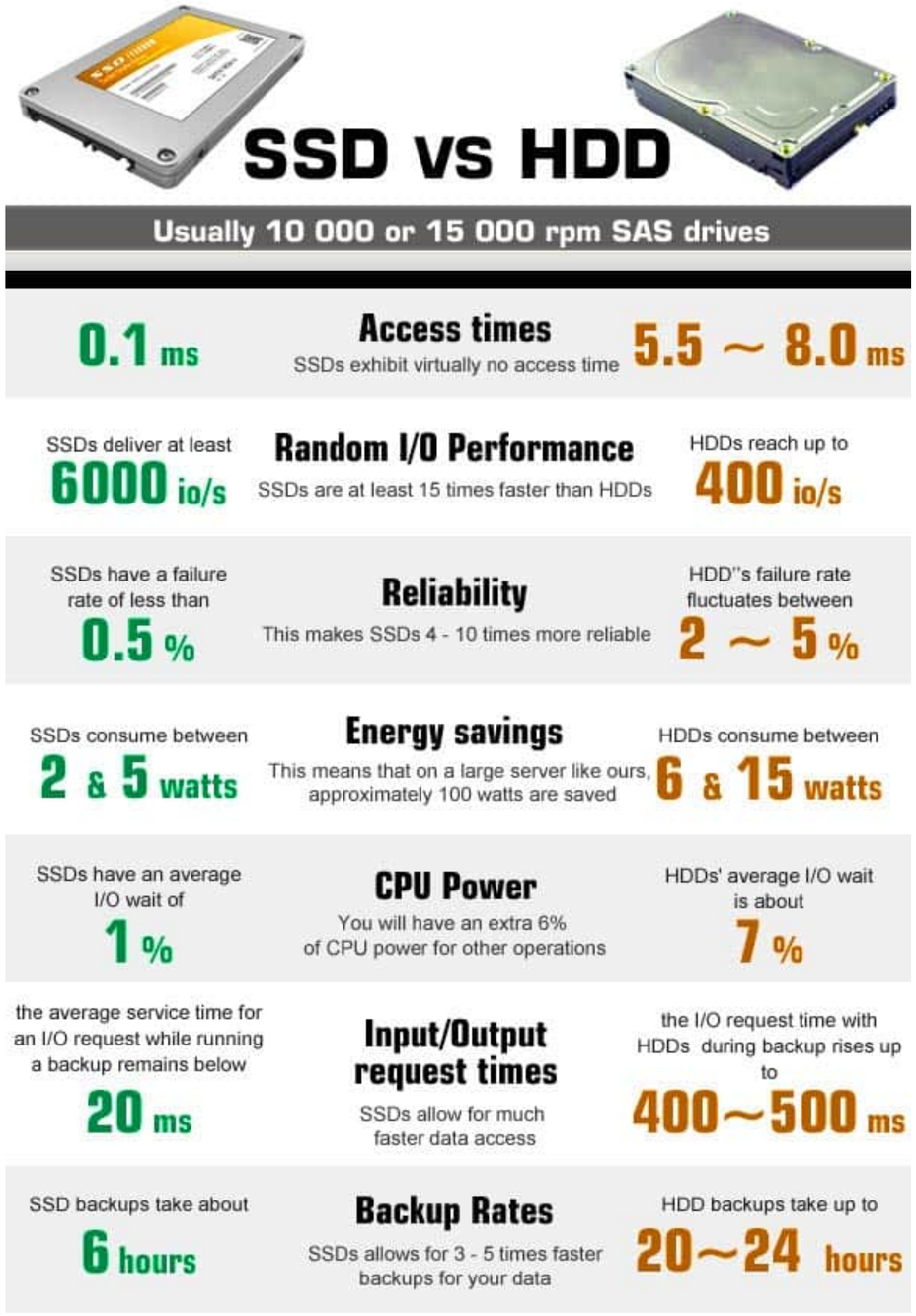
A bit of History

- 75
- 100
- 300

HDD Disk max IOPS

For many years
drives have been the DC bottleneck

SSD vs HDD



Minimum 15x Difference in IOPS !

IO vendors have already revised their drive management

Source
https://wintechlab.com/ssd-vs-hdd-which-is-better-for-you/

More SSD Figures

First Gen SSD – **30 KIOPS**

PCIe Gen 3 SSD – **500 KIOPS**

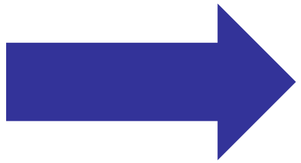
PCIe Gen 4 SSD - **1.4 MIOPS**

This is a **x3000** factor !

IO vendors have
to do it again

NVMe / NVMe-oF Standards

A new Transport is Required



1 queue / 256 commands

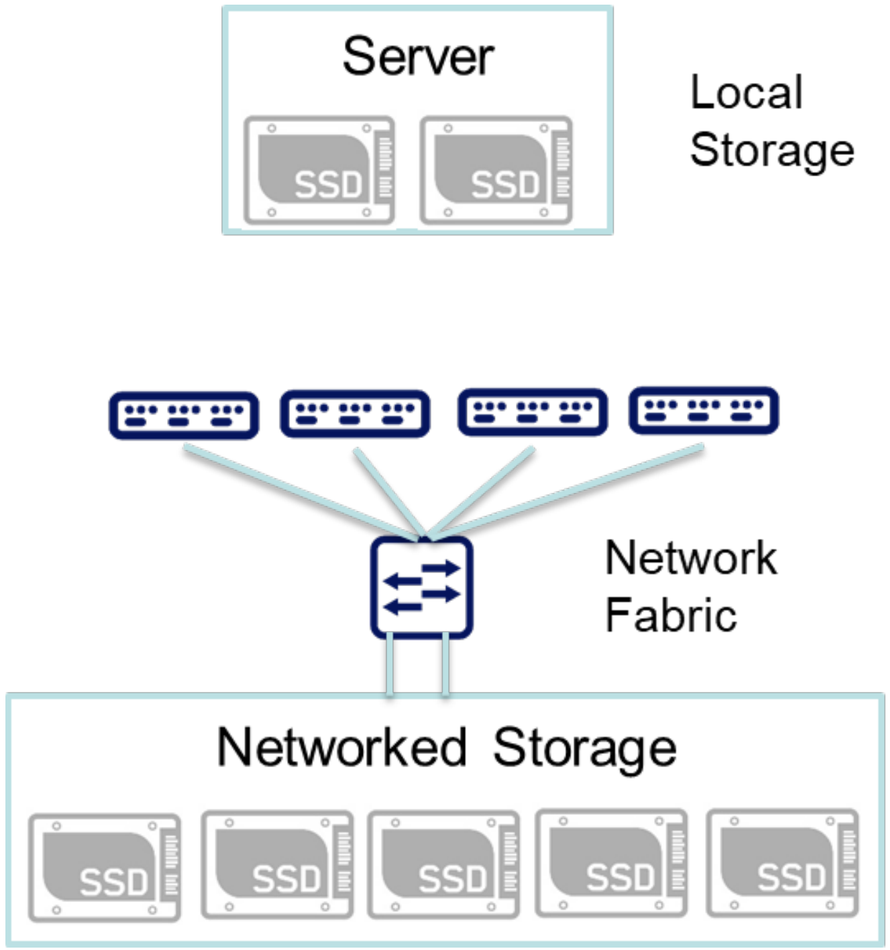
64K queues / 64K commands

Massively Parallel

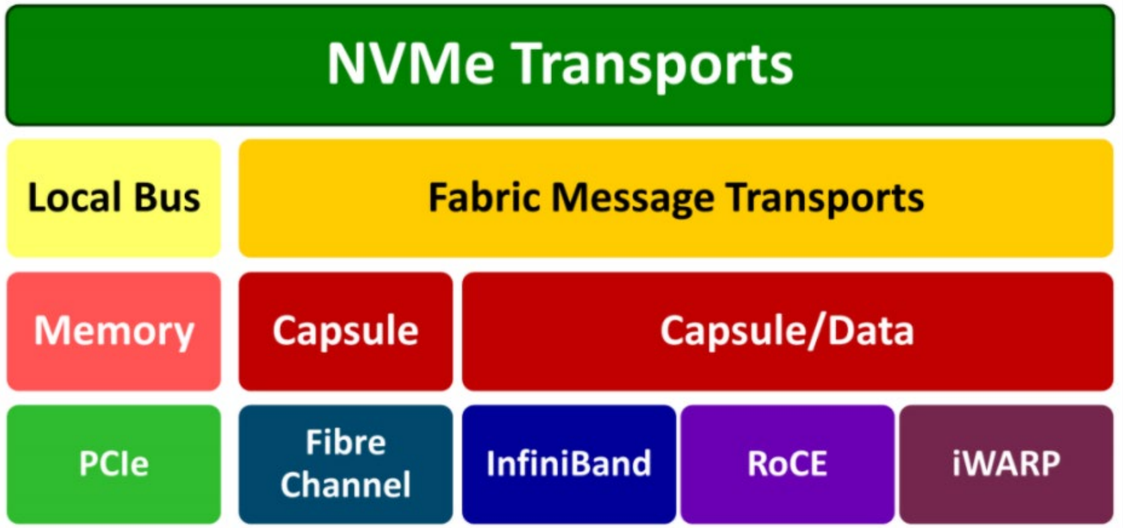
A revolution for any IO stack !

A new Fabric as well

- NVMe™ over PCIe® limited to local use
- Constant desire to network storage
  - Sharing / provisioning
  - Cloud / virtualization / containers
  - Data / workload migration
  - Better efficiency & data protection



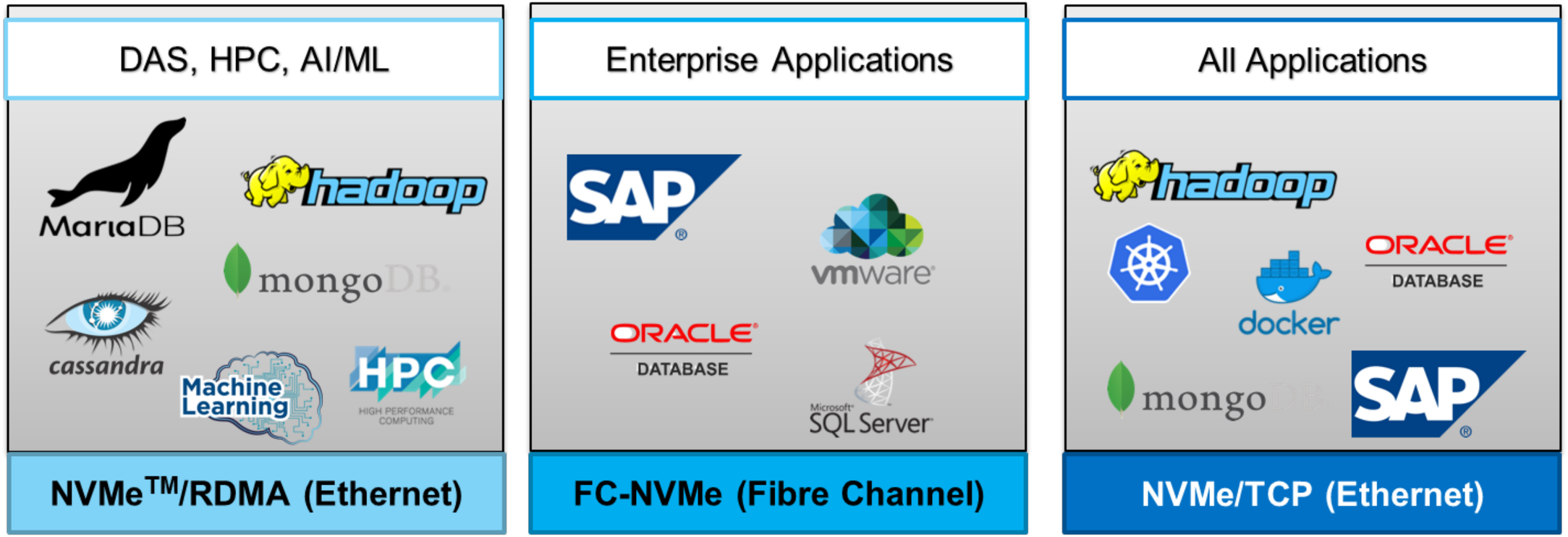
Different Types of NVMe Transport



- PCIe
  - Great for direct attached NVMe™ SSDs
  - Does not scale well to large topologies
- FC and RDMA (Infiniband, RoCE, iWARP)
  - Provides a high degree of scalability - requires special networks and hardware
  - Provides performance (throughput and latency) comparable to direct attached NVMe SSDs
- TCP
  - Uses generic TCP networks
  - Scalable allowing large scale deployments and operation over long distances

Use Cases by Fabric

No one size fits all!



Performance at the cost of complexity

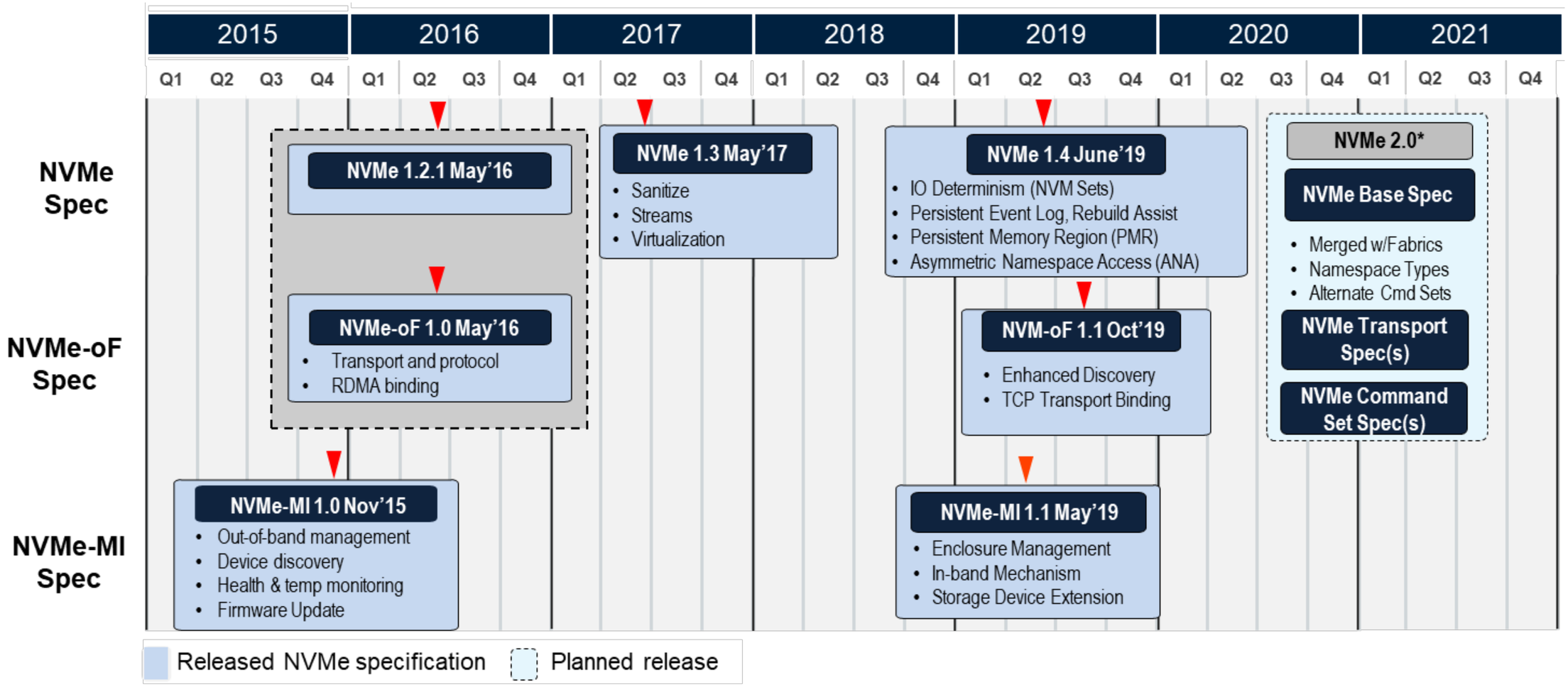
Leverage existing infrastructure. Reliability is key

Simplicity is key. Balance of performance and cost

Logos are indicative of workload characteristics only.

Source: Flash Memory Summit, nvm EXPRESS®

NVM Express™ Technology Specification Roadmap



NVMe is Pushing Boundaries

IO Pressure on Drives

Compute Nodes / apps

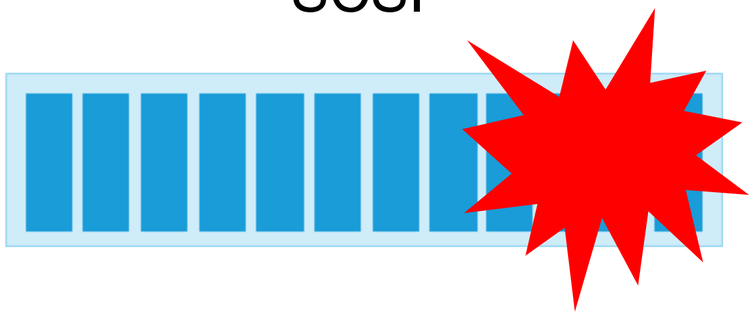
Latency in mS

FC/SCSI – Infiniband
Ethernet : iSCSI - NAS

Storage Nodes

Latency in mS

SCSI



HDD - 300 IOPS
Latency in 10s mS

Storage Arrays
implement cache
to hide poor drive latencies

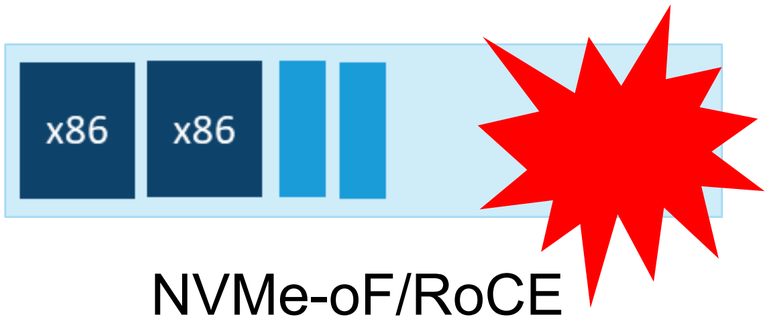
Pressure is on drives

IO Pressure on Storage OS

Compute Nodes / apps

Latency in mS

FC/SCSI-NvMe – Infiniband
Ethernet : NVMe-oF - iSCSI – NAS



Storage Nodes

Latency in mS

NVMe-oF/RoCE

SSD – 30K to 500 KIOPS

Latency in 100s uS

Storage Arrays reinvent IO stack to support SSD

Pressure is on Storage Controllers

IO Pressure up to the apps

Compute Nodes / apps
Latency in mS

FC/SCSI-NvMe – Infiniband
Ethernet : NVMe-oF – NAS

Storage Nodes
Latency in mS

NVMeoF/RoCE

SSD – 1.5 MIOPS
Latency in 10s uS

IO stack needs to
be redesigned
= road to composable

Pressure is
on the full IO stack

TCP or RoCE

Comparison Summary

	TCP	RoCE
Standard driver / NIC	✓	X
Direct connect	✓	✓
Latency as local NVMe SSD	X	✓
Minimize CPU load	X	✓
iSCSI replacement	✓	X
Infiniband replacement	X	✓
FC alternative	✓	✓

TCP is best for capacity and general purpose apps.
RoCE is best for high performance and low latency purposes.

Building the Future

How to be OS Independent?

- Using standards such as NVMe is a great start !
- NVMe Emulation is a key enabler
  - ⇒ Any card is viewed as a NVMe device
  - ⇒ Highway to independence

Kalray Smart Storage Adapter

AccessCore ™ Storage framework on MPPA® to deliver Data Services

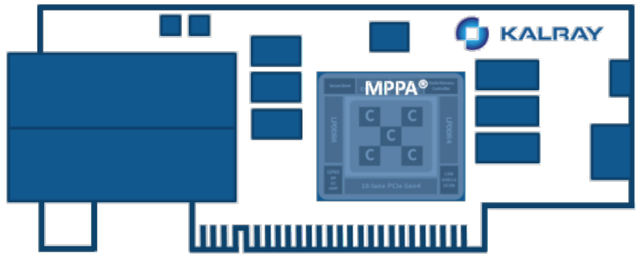
SDK

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***

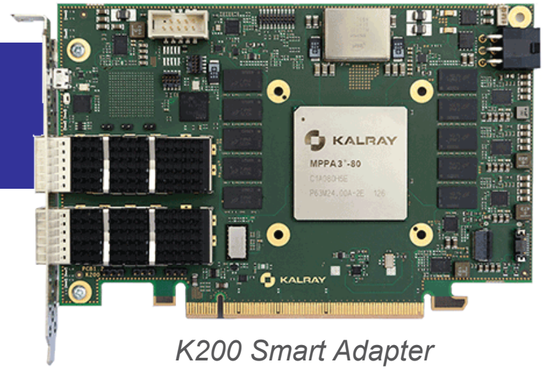
Drives access through PCIe ***(RC or P2P)***

nvm EXPRESS®

Serial Attached SCSI

Kalray Smart Storage Adapter Solution

K200 & K200-LP
manufactured by wistron



K200 Smart Adapter

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

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

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

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**

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 ...

Example of NVMe-oF (RoCE/TCP) JBOF

Hyper Optimized JBOF (no x86)

- JBOF Chassis :
  - Stand-alone
  - 2U – 1200W Redundant
  - 24 U.2 NVMe SSDs
  - 6xPCIe Gen3 x16
- Kalray Smart Controller Carc
  - 2 to 6 Cards
- BMC chip – AST2500 (ASpeed)
- 1Gbps management interface

wistron

NVMe SSDs	Redundant Power
System Cooling FANs	PCIe Card Cages 12

Lymma JBOF Reference Platform
White Label NVMe-oF (RoCE/TCP) JBOF

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

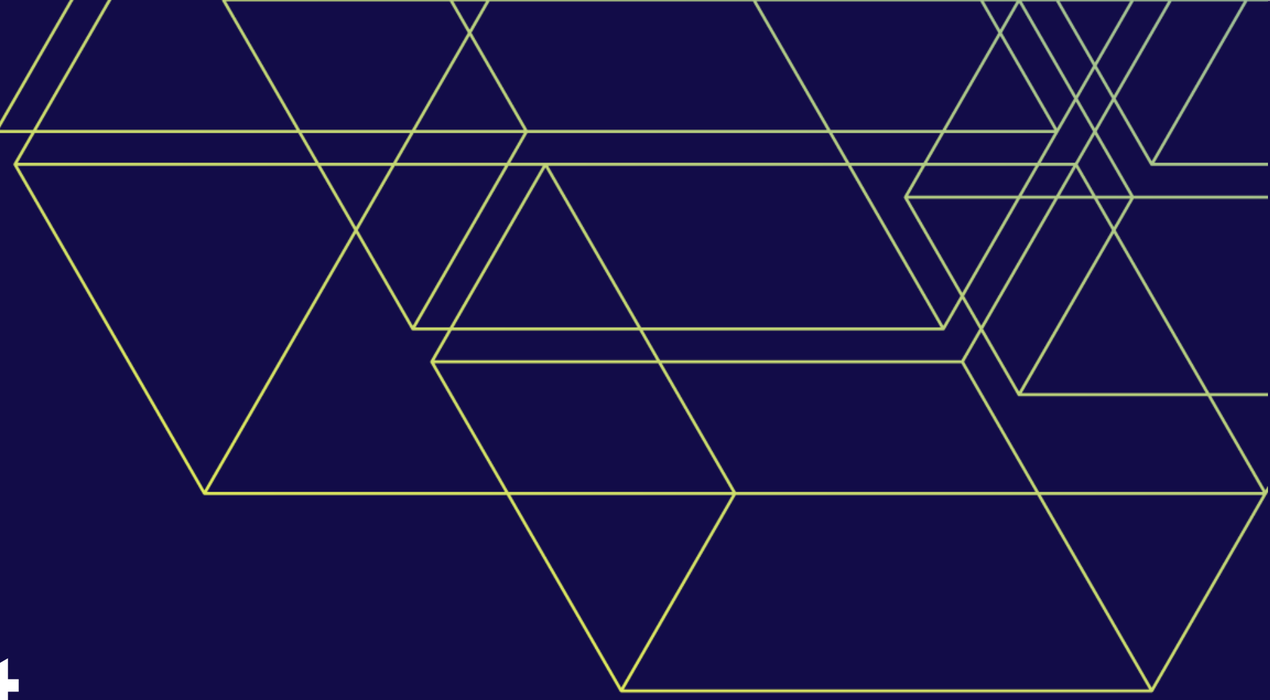
- Switch to a true **C**omposable **D**isaggregated **I**nfrastructure with commodity components
- Optimize HCI nodes efficiency

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.