SDC 20

BY Developers FOR Developers

Storage Developer Conference
September 22-23, 2020

A Data Storage Diode for Classified Sites

Jean-Baptiste RIAUX
Senior Field Application Engineer
jbriaux@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

- Developing a “Storage Diode” by combining specific pieces of storage technologies such as HDF5, multipathing, ACL, user authentication (Kerberos, LDAP...) while leveraging NVMe-oF, is very useful for classified sites requiring remote and secure replication on NVMe SSDs.
- The storage diode is a dedicated storage system with two isolated Read and Write path, with guarantee of the data integrity. Leveraging dual port NVMe drives and the parallelism of advanced processors, this paper reviews how to fully isolate channels at both logical and physical levels, and dedicate write-only and read-only path to storage devices over a NVMe-oF fabric.

  This technique allows restricted/classified computing center to push (write) data to the storage diode, assuring the path to the outside world can be only be accessed in Read-Only.

The Presenter

About the Presenter

Jean-Baptiste Riaux is a Sr Field Application Engineer at Kalray for the Data Center Business Unit.

He specializes in storage and HPC applications, and has a strong experience in deploying NVMe and NVMe-oF storage solutions.

Before joining Kalray, Jean-Baptiste gained an extensive experience of the storage and HPC domains in DDN & Intel.

Context & Requirements

What are HPC Dark Sites ?

- HPC Dark sites: classified datacenters closed from outside world running HPC clusters.
- HPC cluster:
  - Compute cluster sharing a unique parallel filesystem (shared disk FS, distributed FS)
  - Dozens of IO nodes, thousands of clients
  - The core of storage is a parallel filesystem such as Lustre or GPFS (PB of data)
  - GPU farms and specialized nodes for AI workflows

HPC Dark Sites Challenges

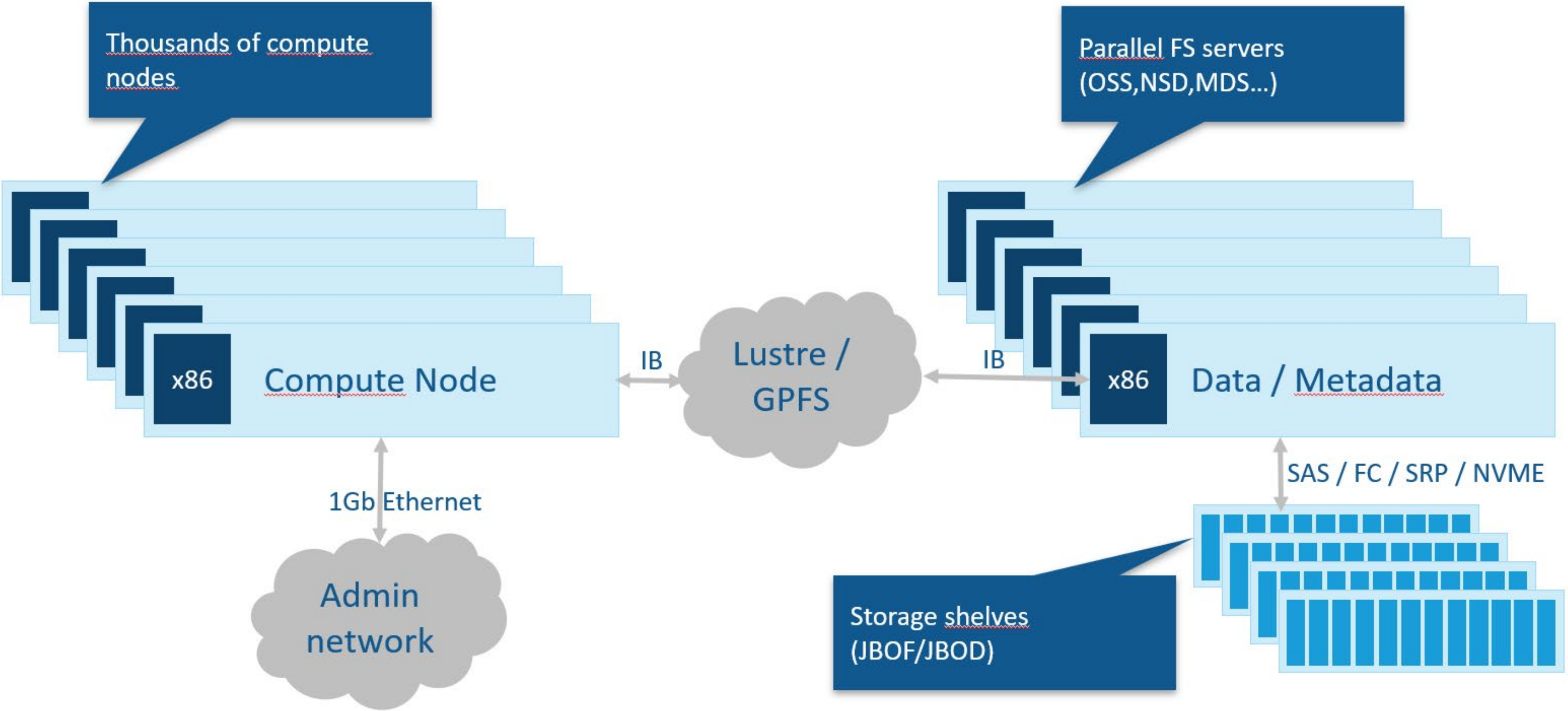
- Dark sites can have high security concerns, such as:
  - No internet access
  - No VPN access (or for a very few restricted system engineers + selected HPC users / researchers)

- Generated data is not always for internal use only

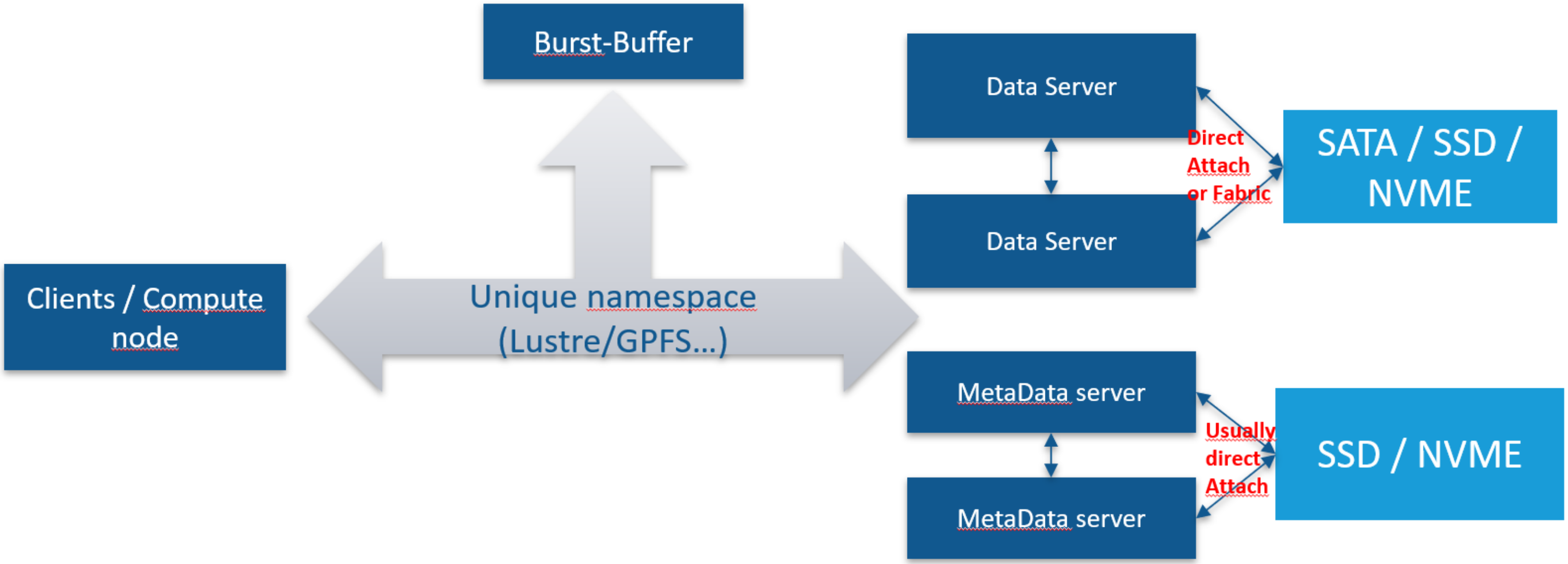
Providing Data Safely to the Outside

- Controlled cold copy on drive(s)
- Special node(s) outside for remote copy using multiple layers of encryption
- Direct optical link to a close and secured datacenter

What is a HPC cluster ?



What is a HPC filesystem ?



Limitations

- Limit as much as possible the accesses required for an outside copy
- Constraints:
  - Can’t be interconnected to the parallel filesystem
  - Can’t have access to PFS namespace
  - Can’t be connected to the same admin network
  - Can’t be a file based access protocol (network based)

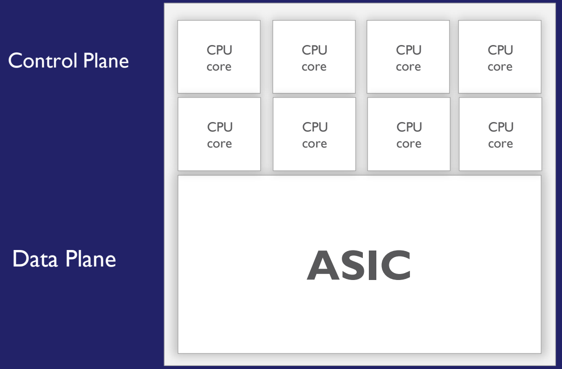
➕ Block-based access solution

How can Kalray Help?

COOLIDGE™: the ULTIMATE I/O Processor

Why Coolidge is a Revolution vs Competition ?

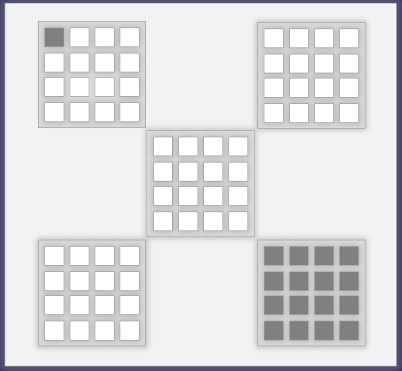
"xPU" Usual Approach



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

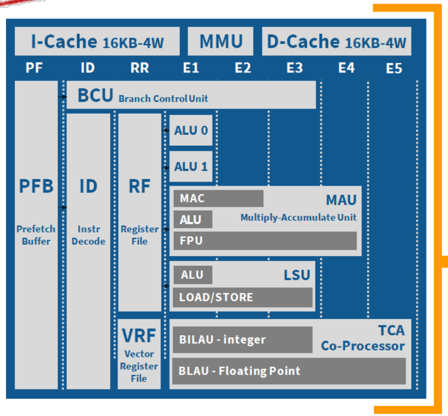
High Speed I/O
2x100Gbs,PCIGen4,DDR4

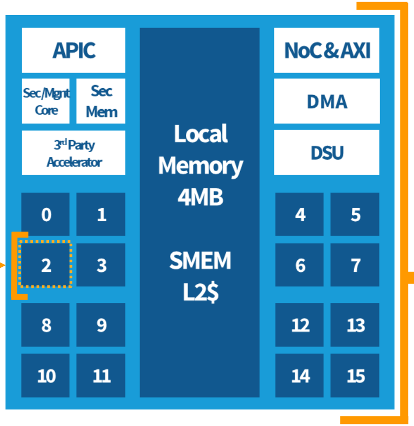
Functional Isolation & Safety
Secure Islands, Encrypt/Decrypt, Secure Boot

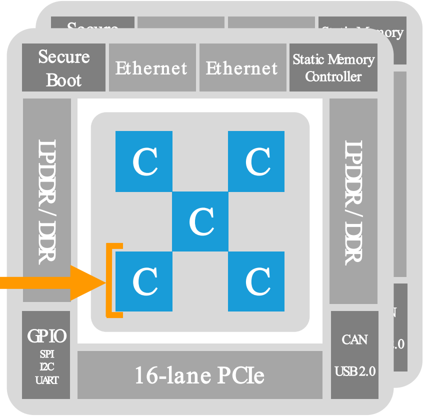
MPPA® COOLIDGE Architecture

The I/O Processor for Next Gen Intelligent Systems

PATENTED







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

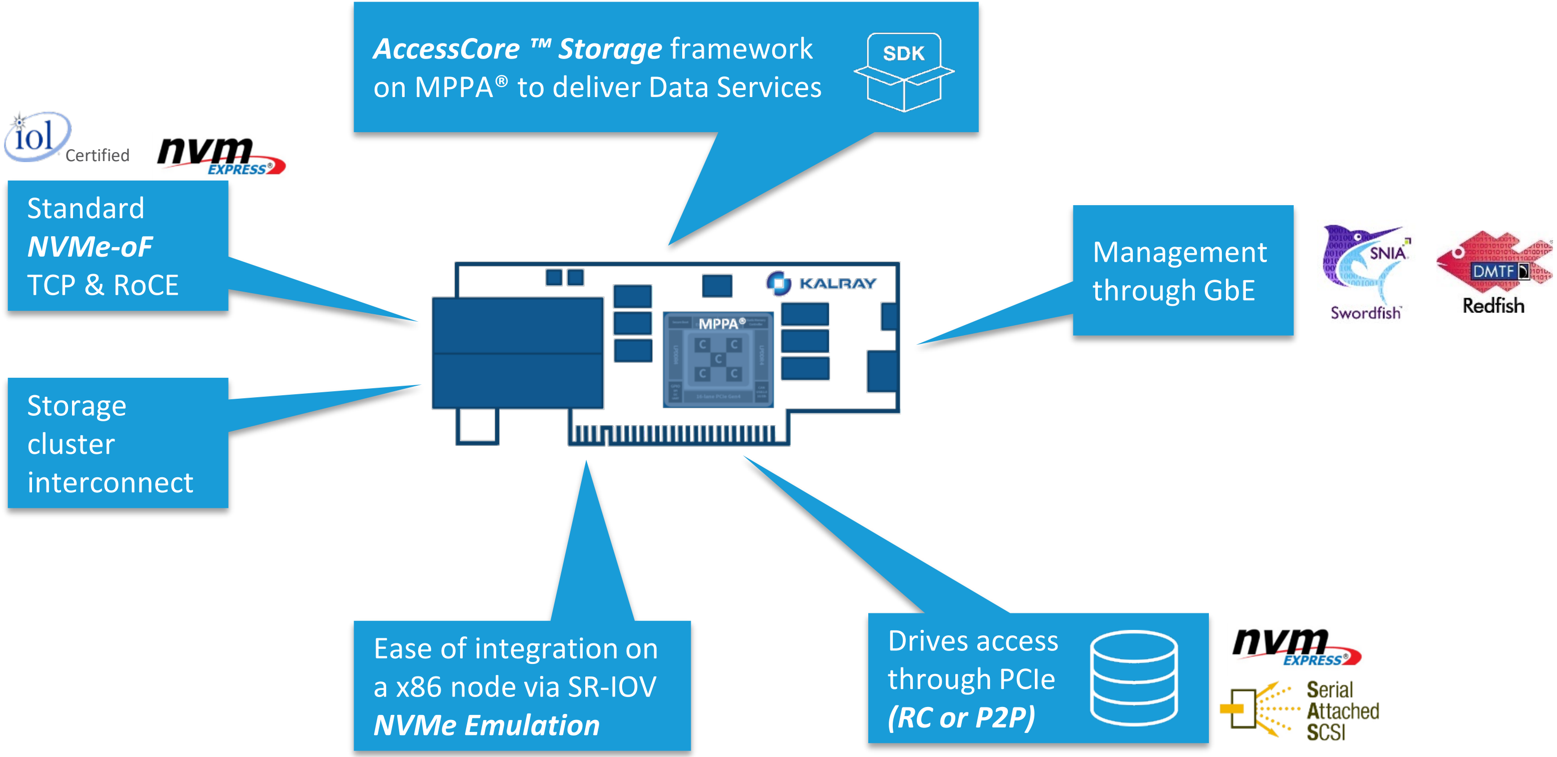
5 Clusters: 80 cores + 80 co-processors

- 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

Kalray Smart Storage Adapter



AccessCore ™ Storage framework on MPPA® to deliver Data Services

Standard ***NVMe-oF*** TCP & RoCE

Storage cluster interconnect

Management through GbE

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

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

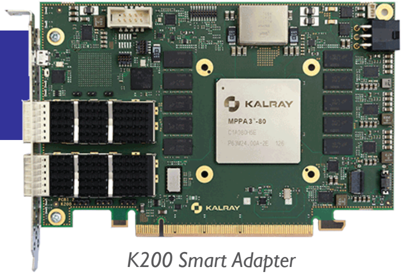
Kalray Smart Storage Adapter Solution

K200 & K200-LP

manufactured by wistron

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



K200 Smart Adapter

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®
SDK

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 ...

AccessCore® for Storage & Networking

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

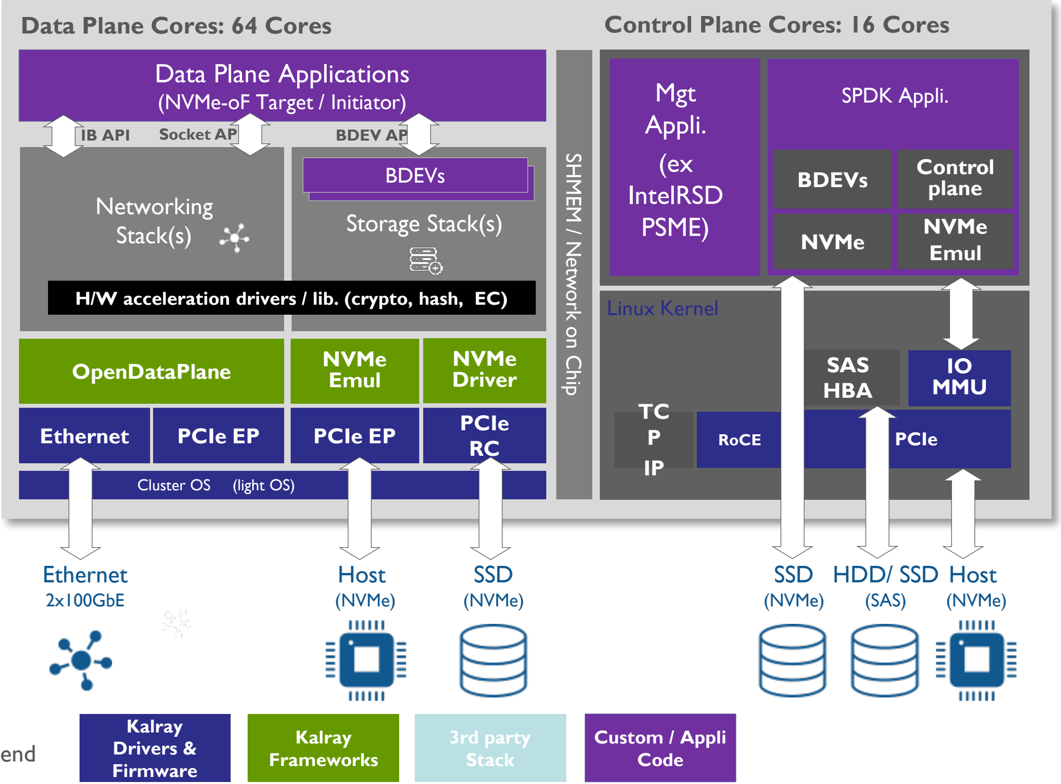
- Run to completion full dataplane
  - From network functions to NVMe stack on light OS cores
- True inline processing
  - No need for x86 pre/post processing

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

SDK AccessCore®
Open Software & Tools

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

SDK AccessCore® Open Software & Tools

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 Cards
  - 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

Global Design Overview

Design overview

Dark-site node

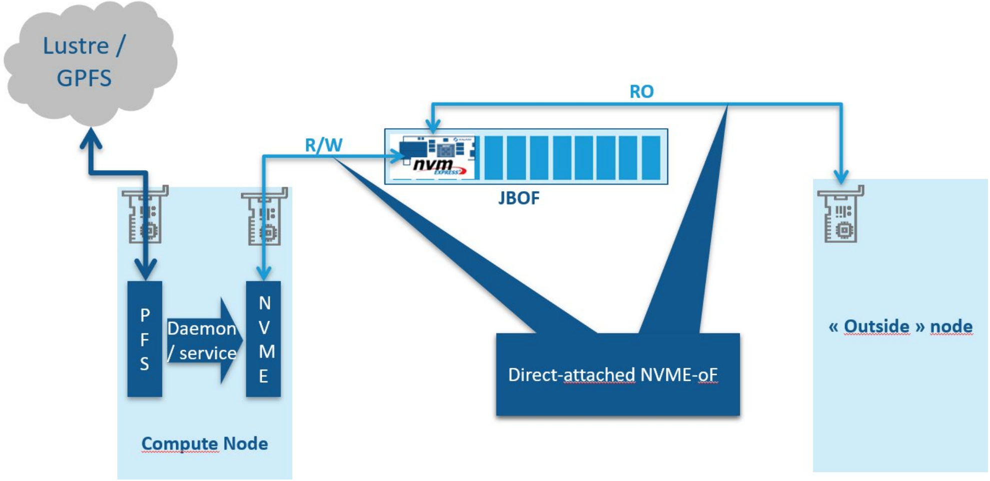
Lustre/GPFS client with a second high-speed interface (40/100G)

The second interface is used for NVME-oF

Periodically writes data from PFS to NVME devices

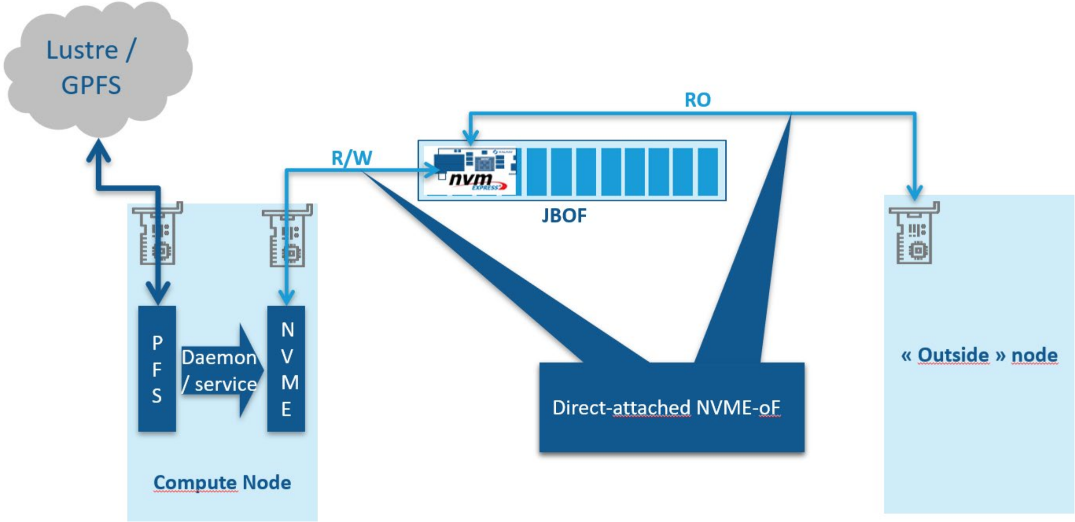
Outside node

Single high speed interface used for NVME-oF
Direct-attached NVMe-oF for security



Design overview

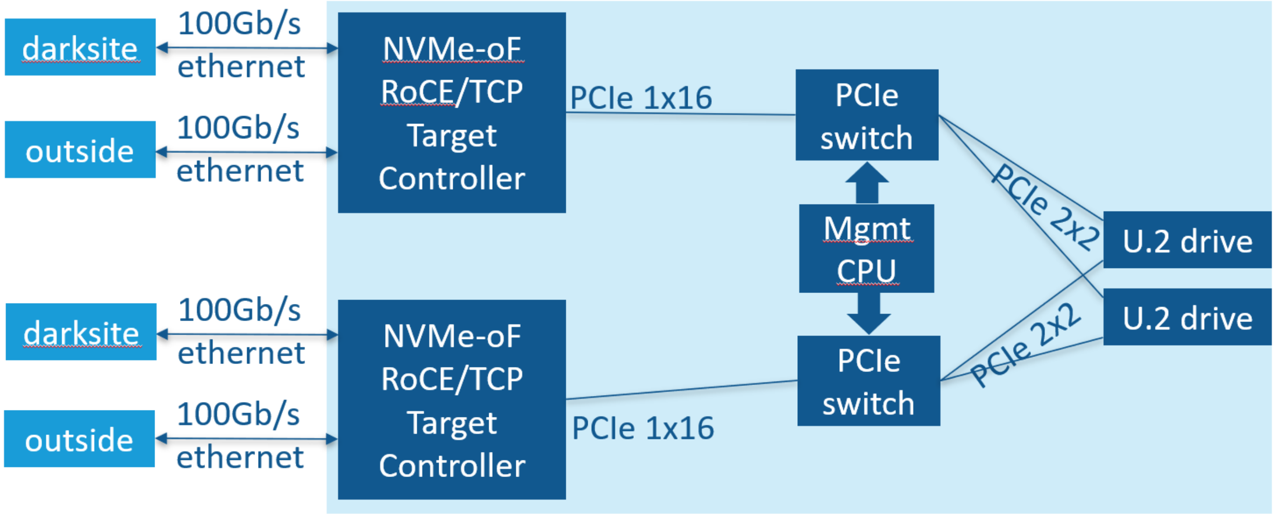
- Looks easy but two problems remain:
  - What about inside the FBOF ?
  - With block level access only, still need to ensure file locking at filesystem level.
- Inside the FBOF: separation at PCI level on dual ported drives
  - Using PCI switches for either flow separation or resiliency



FBOF Parameters

FBOF PCI Configuration

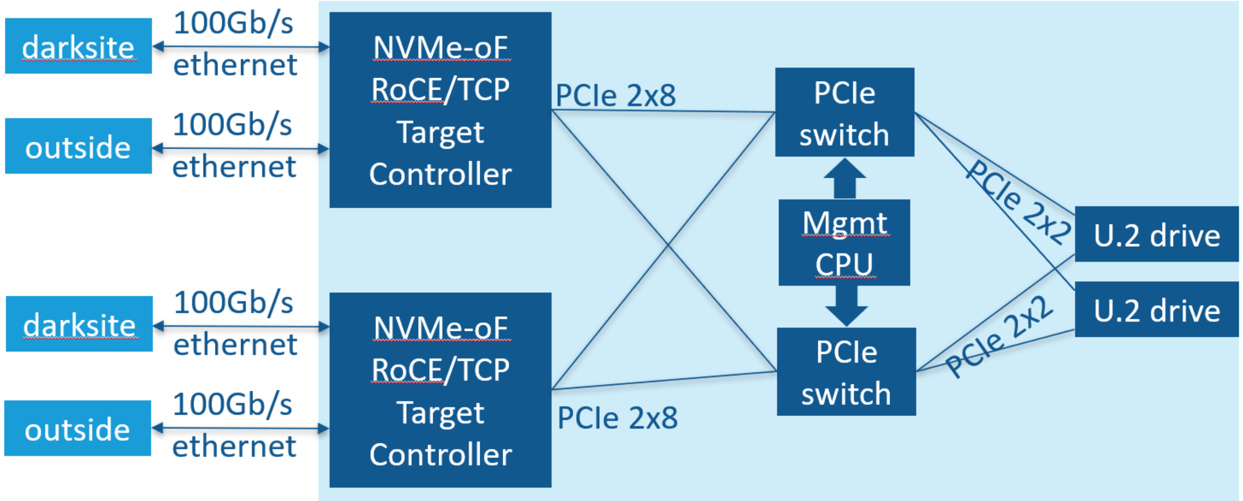
- U.2 dual ported drives to ensure flow separation between the two physical paths
- At PCI level, multiple configurations/usages can be done depending on resiliency vs isolation



Custom security with K200

- Target controller PCI bifurcation can be also defined between 2x8, 4x4 and 1x16 PCI lanes

- Using our own board (Kalray's K200) allows us to add custom security



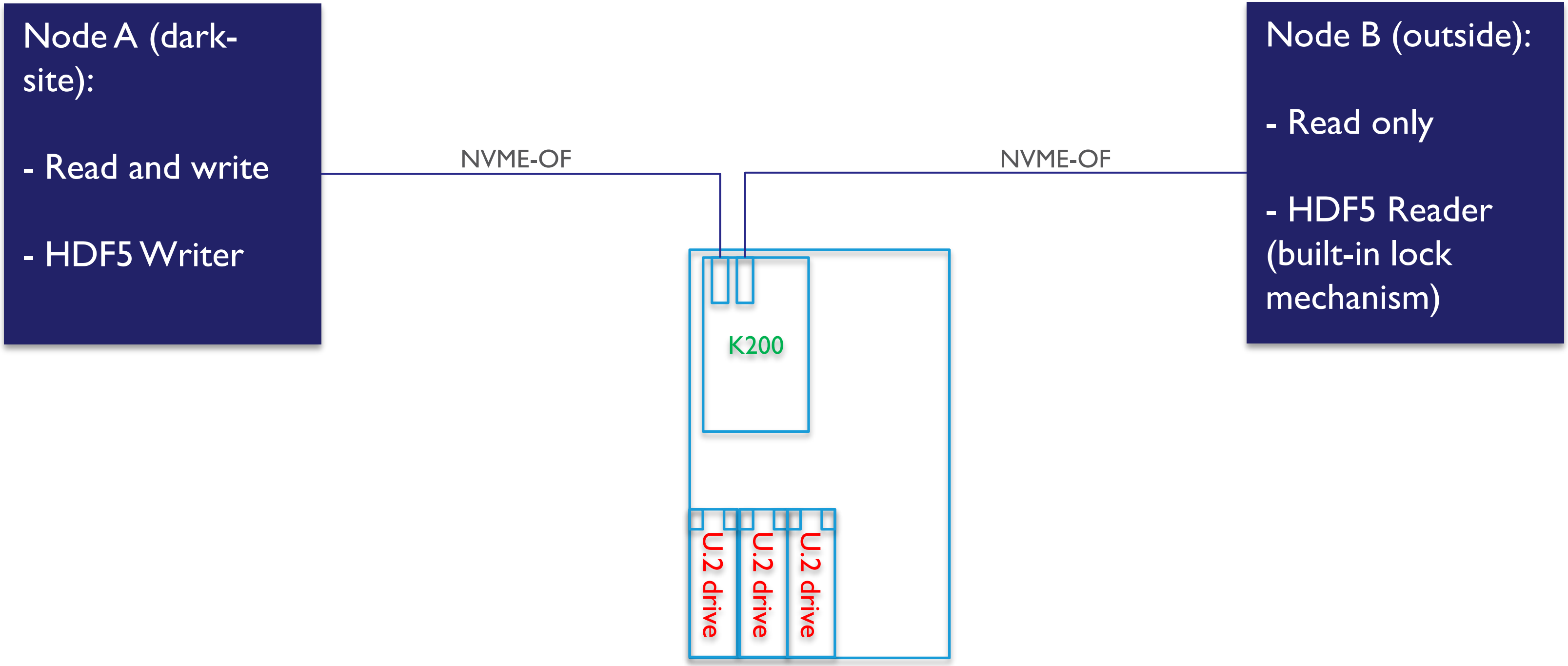
The Provided Solution

Design Options

- Two nodes connected to a NVME target controller, in direct connect
- One node as read and write
- One node as read only
- Guaranteed separated data paths
- Block level protocol (NVME)

- Need for a filesystem using internal locking at file level or containers
  - HDF5 has been chosen

Architecture View



Node A (dark-site):

- Read and write

- HDF5 Writer

NVME-OF

NVME-OF

Node B (outside):

- Read only

- HDF5 Reader (built-in lock mechanism)

Key Elements

- Ideal for WORM workflows (Write Once Read Many)
- Benchmarking is necessary to find a good balance between:
  - Writers and number of drives (x86 saturation vs nvme saturation)
  - Drive IOPS in write and read (specialized read drives or not)
  - Number of readers to a single target

Key Benefits

Key benefits

SECURITY FOCUSED

- Guaranteed read-only access from outside world
- PFS namespace hidden
- Restricted NVME access

SCALABLE

- Easy to scale, possibility to start small and add:
  - K200 cards
  - U.2 drives
  - Writers nodes
  - Reader nodes
  - Full diode

FUTURE PROOF

- As MPPA® processor is fully programmable
- Ease to update
- Add custom code (security / feature)

Q&A on slack

Thank you!