OPENFABRICS ALLIANCE

2022 OFA Virtual Workshop

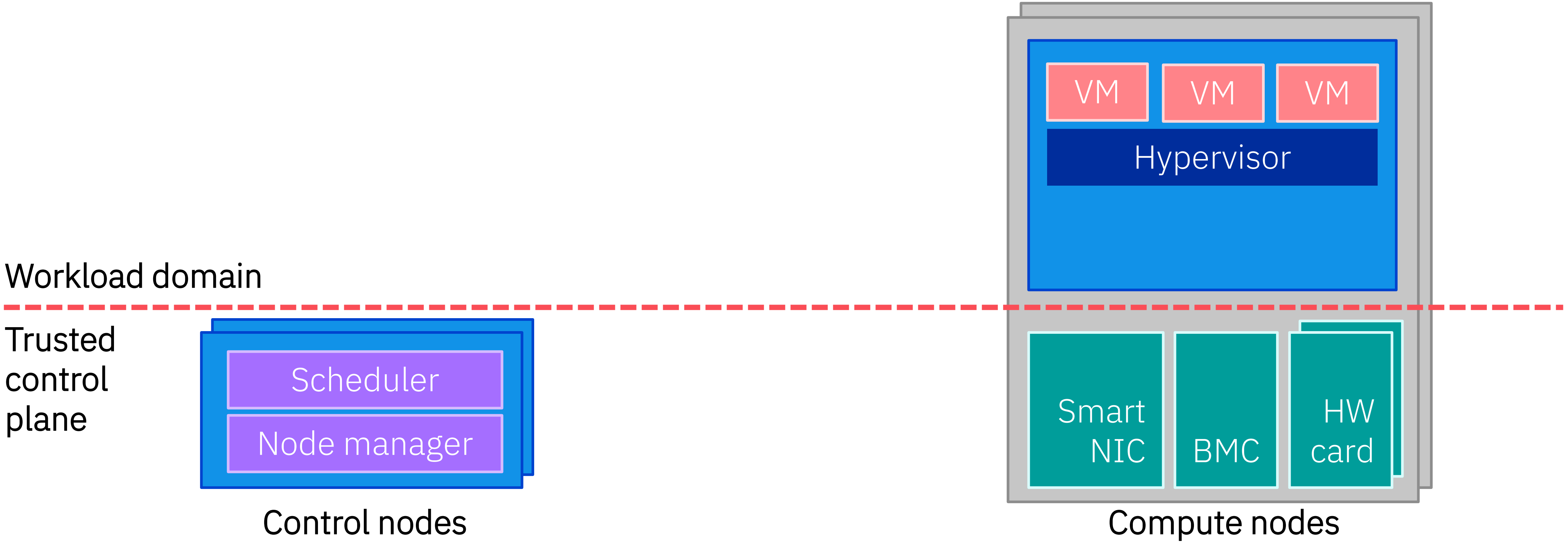
SOFA-STORAGE: CREATING A VENDOR AGNOSTIC FRAMEWORK TO ENABLE SEAMLESS STORAGE OFFLOAD USING SMARTNICS

Raphael Polig, Jonas Pfefferle, Nikolas Ioannou

IBM Research - Zurich

IBM

Terminology



Workload domain

Trusted control plane

Scheduler

Node manager

Control nodes

VM

VM

VM

Hypervisor

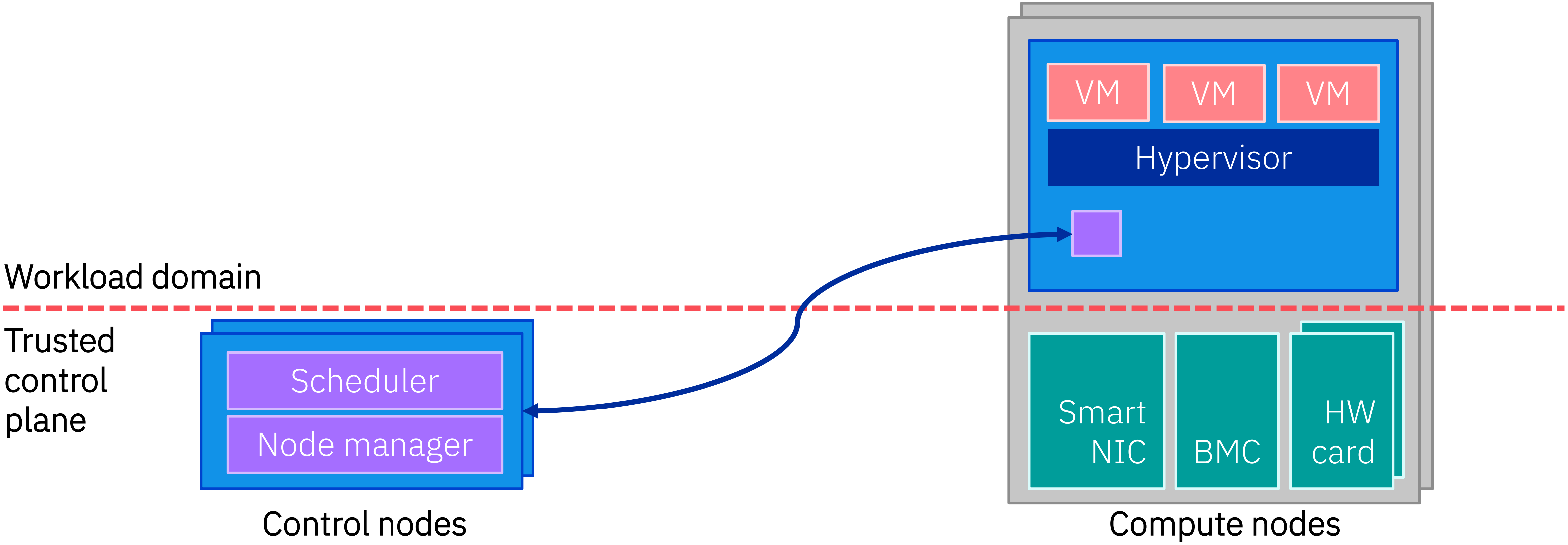
Smart NIC

BMC

HW card

Compute nodes

Motivation



Workload domain

Trusted control plane

Scheduler

Node manager

Control nodes

VM

VM

VM

Hypervisor

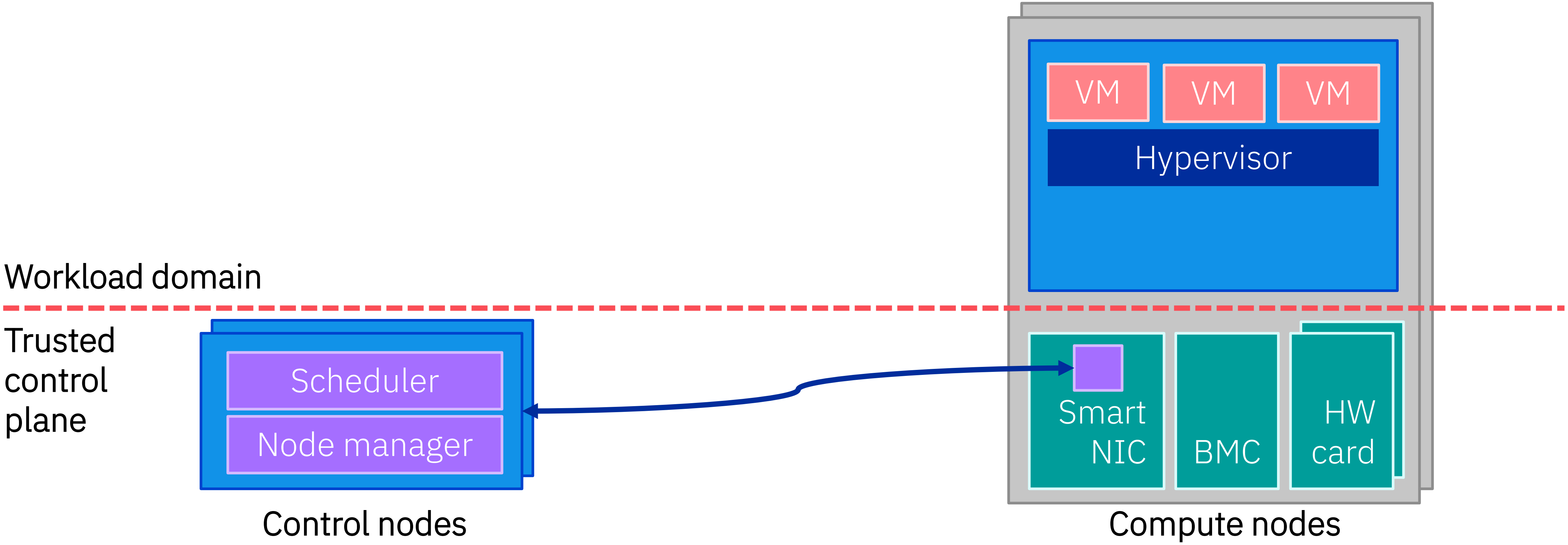
Smart NIC

BMC

HW card

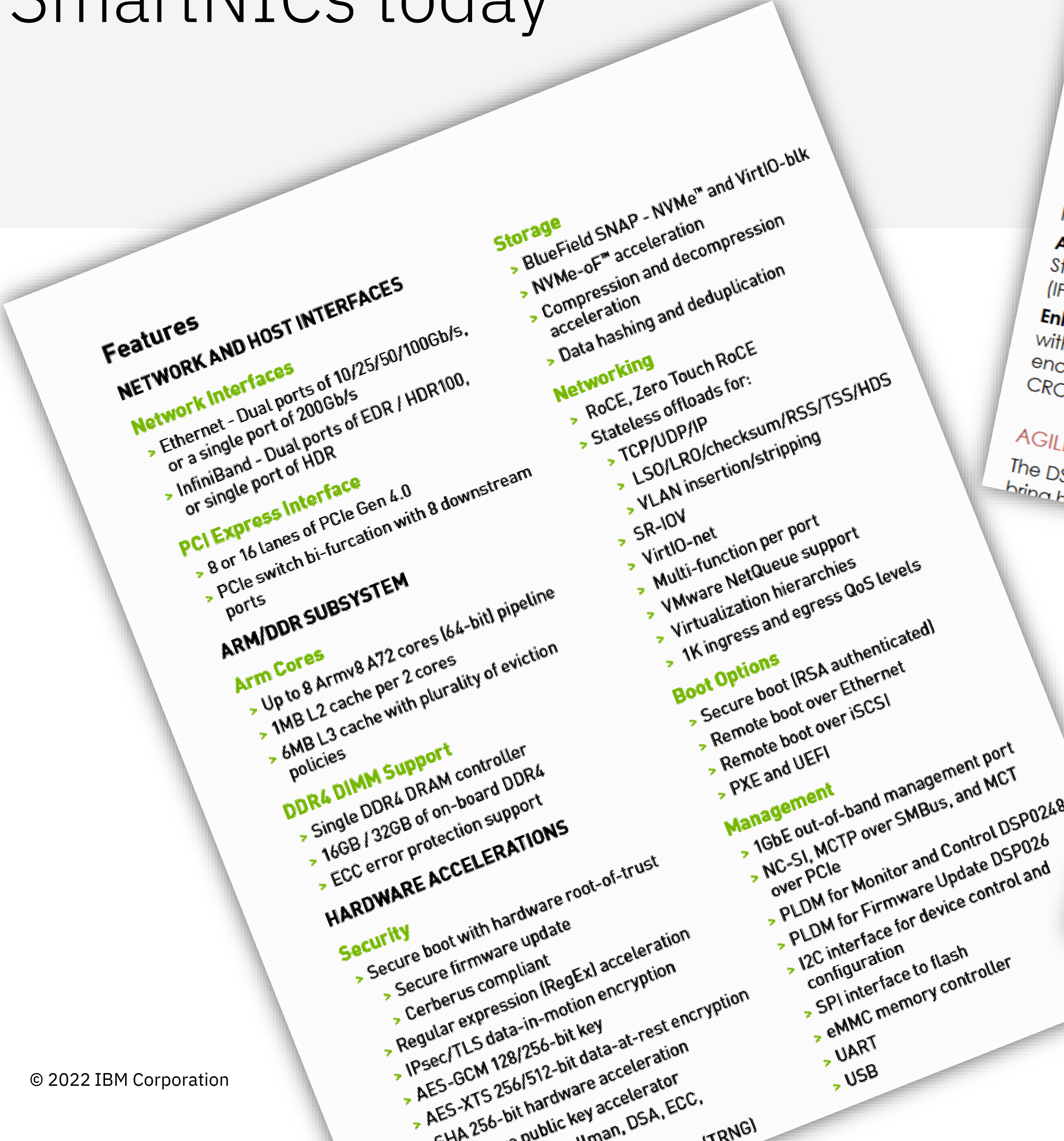
Compute nodes

Motivation



VM
VM
VM
Hypervisor
Workload domain
Trusted control plane
Scheduler
Node manager
Smart NIC
BMC
HW card
Control nodes
Compute nodes

SmartNICs today



Features

NETWORK AND HOST INTERFACES

Network Interfaces
- Ethernet - Dual ports of 10/25/50/100Gb/s, or a single port of 200Gb/s
- InfiniBand - Dual ports of EDR / HDR100, or single port of HDR

PCI Express Interface
- 8 or 16 lanes of PCIe Gen 4.0
- PCIe switch bi-furcation with 8 downstream ports

ARM/DDR SUBSYSTEM

Arm Cores
- Up to 8 Armv8 A72 cores (64-bit) pipeline
- 1MB L2 cache per 2 cores
- 6MB L3 cache with plurality of eviction policies

DDR4 DIMM Support
- Single DDR4 DRAM controller
- 16GB / 32GB of on-board DDR4
- ECC error protection support

HARDWARE ACCELERATIONS

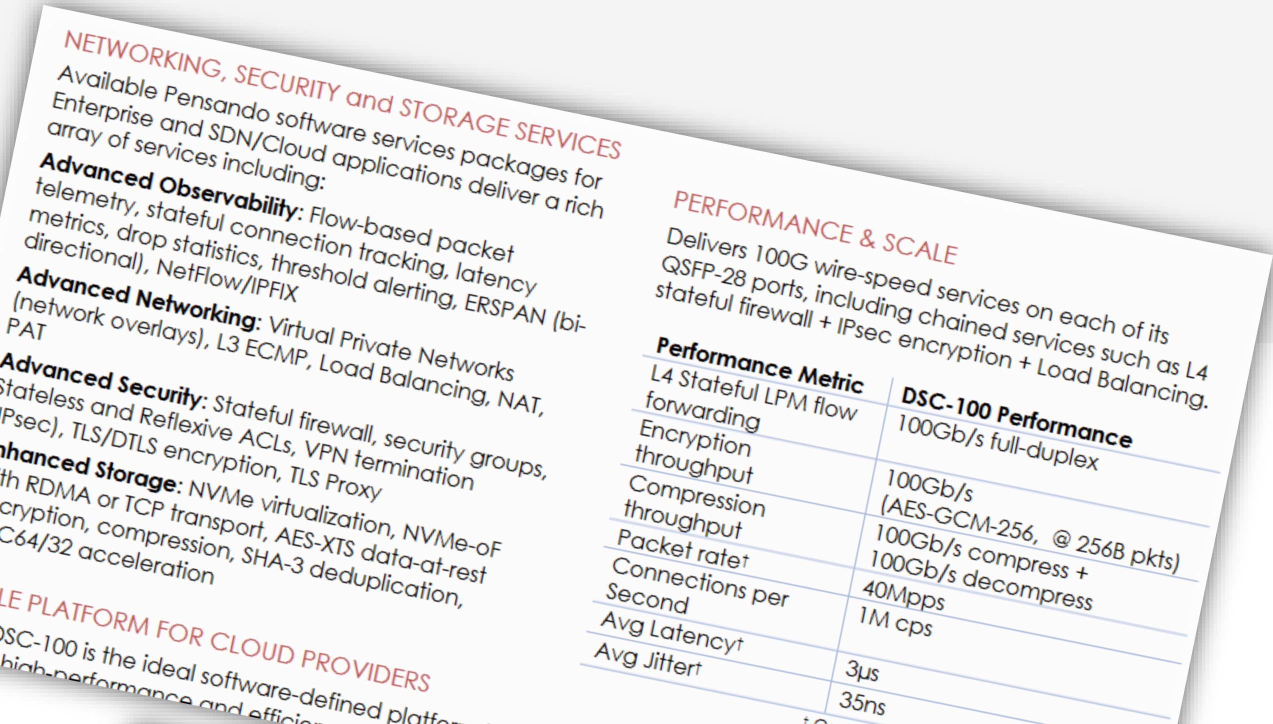
Security
- Secure boot with hardware root-of-trust
- Secure firmware update
- Cerberus compliant
- Regular expression (RegEx) acceleration
- IPsec/TLS data-in-motion encryption
- AES-GCM 128/256-bit key
- AES-XTS 256/512-bit data-at-rest encryption
- SHA 256-bit hardware acceleration
- public key accelerator
- llman, DSA, ECC,
- (TRNG)

Storage
- BlueField SNAP - NVMe™ and VirtIO-blk
- NVMe-oF™ acceleration
- Compression and decompression acceleration
- Data hashing and deduplication

Networking
- RoCE, Zero Touch RoCE
- Stateless offloads for:
  - TCP/UDP/IP
  - LSO/LRO/checksum/RSS/TSS/HDS
  - VLAN insertion/stripping
- SR-IOV
- VirtIO-net
- Multi-function per port
- VMware NetQueue support
- Virtualization hierarchies
- 1K ingress and egress QoS levels

Boot Options
- Secure boot (RSA authenticated)
- Remote boot over Ethernet
- Remote boot over iSCSI
- PXE and UEFI

Management
- 1GbE out-of-band management port
- NC-SI, MCTP over SMBus, and MCT over PCIe
- PLDM for Monitor and Control DSP0248
- PLDM for Firmware Update DSP026
- I2C interface for device control and configuration
- SPI interface to flash
- eMMC memory controller
- UART
- USB



NETWORKING, SECURITY and STORAGE SERVICES

Available Pensando software services packages for Enterprise and SDN/Cloud applications deliver a rich array of services including:

Advanced Observability: Flow-based packet telemetry, stateful connection tracking, latency metrics, drop statistics, threshold alerting, ERSPAN (bi-directional), NetFlow/IPFIX

Advanced Networking: Virtual Private Networks (network overlays), L3 ECMP, Load Balancing, NAT, PAT

Advanced Security: Stateful firewall, security groups, Stateless and Reflexive ACLs, VPN termination (IPsec), TLS/DTLS encryption, TLS Proxy

Enhanced Storage: NVMe virtualization, NVMe-oF with RDMA or TCP transport, AES-XTS data-at-rest encryption, compression, SHA-3 deduplication, CRC64/32 acceleration

AGILE PLATFORM FOR CLOUD PROVIDERS

The DSC-100 is the ideal software-defined platform to bring high-performance and efficiency to the cloud

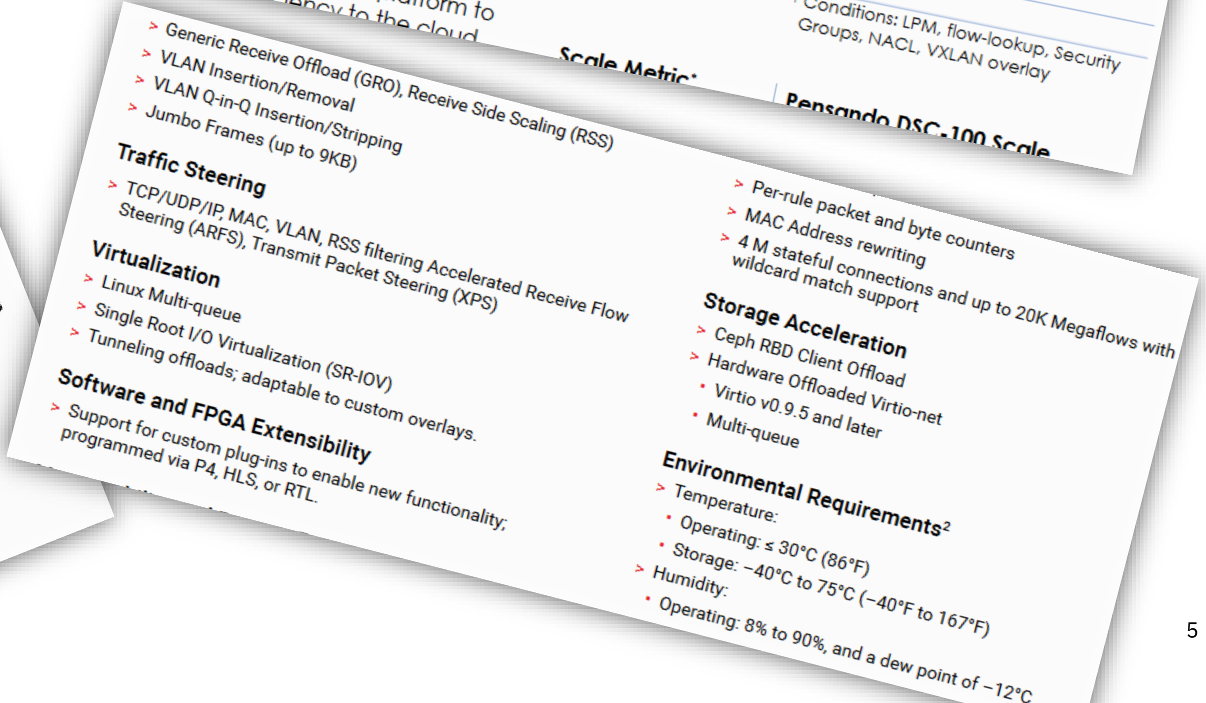
PERFORMANCE & SCALE

Delivers 100G wire-speed services on each of its QSFP-28 ports, including chained services such as L4 stateful firewall + IPsec encryption + Load Balancing.

Performance Metric	DSC-100 Performance
L4 Stateful LPM flow forwarding	100Gb/s full-duplex
Encryption throughput	100Gb/s (AES-GCM-256, @ 256B pkts)
Compression throughput	100Gb/s compress + 100Gb/s decompress
Packet rate†	40Mpps
Connections per Second	1M cps
Avg Latency†	3µs
Avg Jitter†	35ns

† Conditions: LPM, flow-lookup, Security Groups, NACL, VXLAN overlay

Scale Metric*	Pensando DSC-100 Scale



- Generic Receive Offload (GRO), Receive Side Scaling (RSS)
- VLAN Insertion/Removal
- VLAN Q-in-Q Insertion/Stripping
- Jumbo Frames (up to 9KB)

Traffic Steering
- TCP/UDP/IP, MAC, VLAN, RSS filtering Accelerated Receive Flow Steering (ARFS), Transmit Packet Steering (XPS)

Virtualization
- Linux Multi-queue
- Single Root I/O Virtualization (SR-IOV)
- Tunneling offloads; adaptable to custom overlays.

Software and FPGA Extensibility
- Support for custom plug-ins to enable new functionality; programmed via P4, HLS, or RTL.

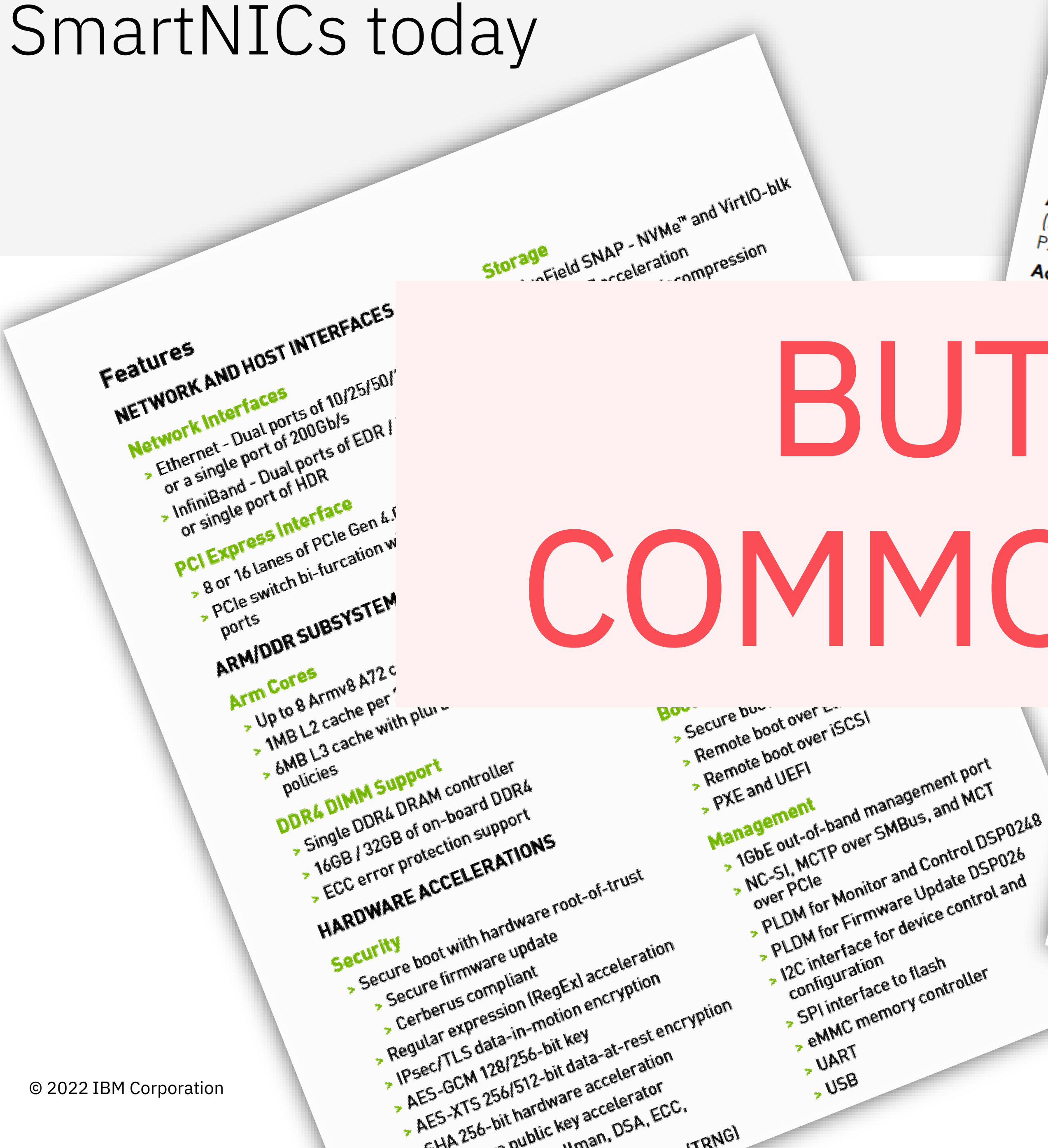
- Per-rule packet and byte counters
- MAC Address rewriting
- 4 M stateful connections and up to 20K Megaflows with wildcard match support

Storage Acceleration
- Ceph RBD Client Offload
- Hardware Offloaded Virtio-net
  - Virtio v0.9.5 and later
  - Multi-queue

Environmental Requirements[2]
- Temperature:
  - Operating: ≤ 30°C (86°F)
  - Storage: −40°C to 75°C (−40°F to 167°F)
- Humidity:
  - Operating: 8% to 90%, and a dew point of −12°C

SmartNICs today

BUT NO COMMON APIs



Storage

Field SNAP - NVMe™ and VirtIO-blk
acceleration
compression

Features

NETWORK AND HOST INTERFACES

Network Interfaces
- Ethernet - Dual ports of 10/25/50/ or a single port of 200Gb/s
- InfiniBand - Dual ports of EDR / or single port of HDR

PCI Express Interface
- 8 or 16 lanes of PCIe Gen 4.
- PCIe switch bi-furcation w ports

ARM/DDR SUBSYSTEM

Arm Cores
- Up to 8 Armv8 A72 c
- 1MB L2 cache per
- 6MB L3 cache with plu policies

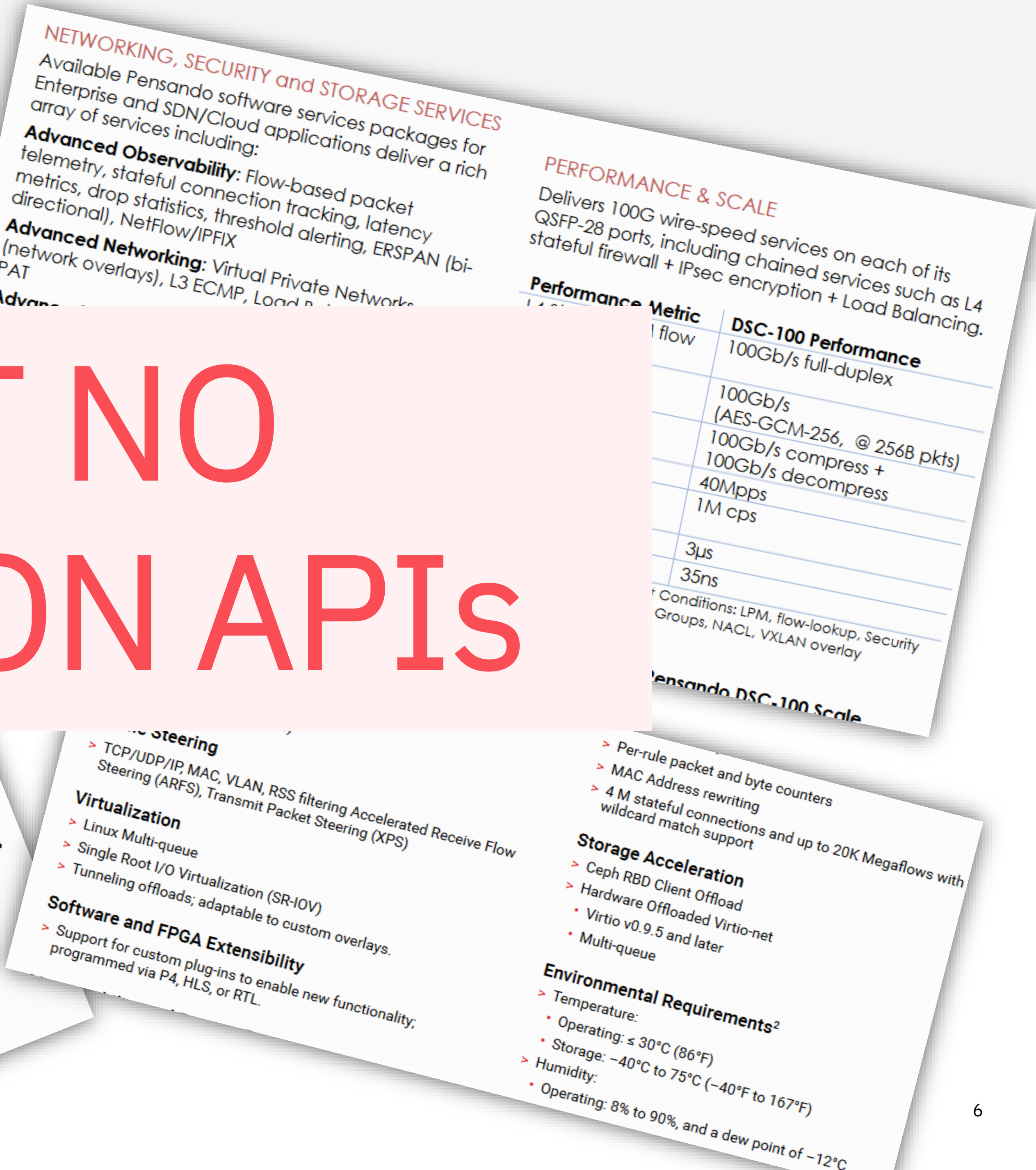
DDR4 DIMM Support
- Single DDR4 DRAM controller
- 16GB / 32GB of on-board DDR4
- ECC error protection support

HARDWARE ACCELERATIONS

Security
- Secure boot with hardware root-of-trust
- Secure firmware update
- Cerberus compliant
- Regular expression (RegEx) acceleration
- IPsec/TLS data-in-motion encryption
- AES-GCM 128/256-bit key
- AES-XTS 256/512-bit data-at-rest encryption
- SHA 256-bit hardware acceleration
- public key accelerator
- man, DSA, ECC,
- TRNG)

Boo
- Secure boo
- Remote boot over E
- Remote boot over iSCSI
- PXE and UEFI

Management
- 1GbE out-of-band management port
- NC-SI, MCTP over SMBus, and MCT over PCIe
- PLDM for Monitor and Control DSP0248
- PLDM for Firmware Update DSP026
- I2C interface for device control and configuration
- SPI interface to flash
- eMMC memory controller
- UART
- USB



NETWORKING, SECURITY and STORAGE SERVICES

Available Pensando software services packages for Enterprise and SDN/Cloud applications deliver a rich array of services including:

Advanced Observability: Flow-based packet telemetry, stateful connection tracking, latency metrics, drop statistics, threshold alerting, ERSPAN (bi-directional), NetFlow/IPFIX

Advanced Networking: Virtual Private Networks (network overlays), L3 ECMP, Load B PAT

Adva

PERFORMANCE & SCALE

Delivers 100G wire-speed services on each of its QSFP-28 ports, including chained services such as L4 stateful firewall + IPsec encryption + Load Balancing.

Performance Metric	DSC-100 Performance
flow	100Gb/s full-duplex
	100Gb/s (AES-GCM-256, @ 256B pkts)
	100Gb/s compress + 100Gb/s decompress
	40Mpps
	1M cps
	3µs
	35ns

* Conditions: LPM, flow-lookup, Security Groups, NACL, VXLAN overlay

ensando DSC-100 Scale

Steering
- TCP/UDP/IP, MAC, VLAN, RSS filtering Accelerated Receive Flow Steering (ARFS), Transmit Packet Steering (XPS)

Virtualization
- Linux Multi-queue
- Single Root I/O Virtualization (SR-IOV)
- Tunneling offloads; adaptable to custom overlays.

Software and FPGA Extensibility
- Support for custom plug-ins to enable new functionality; programmed via P4, HLS, or RTL.

- Per-rule packet and byte counters
- MAC Address rewriting
- 4 M stateful connections and up to 20K Megaflows with wildcard match support

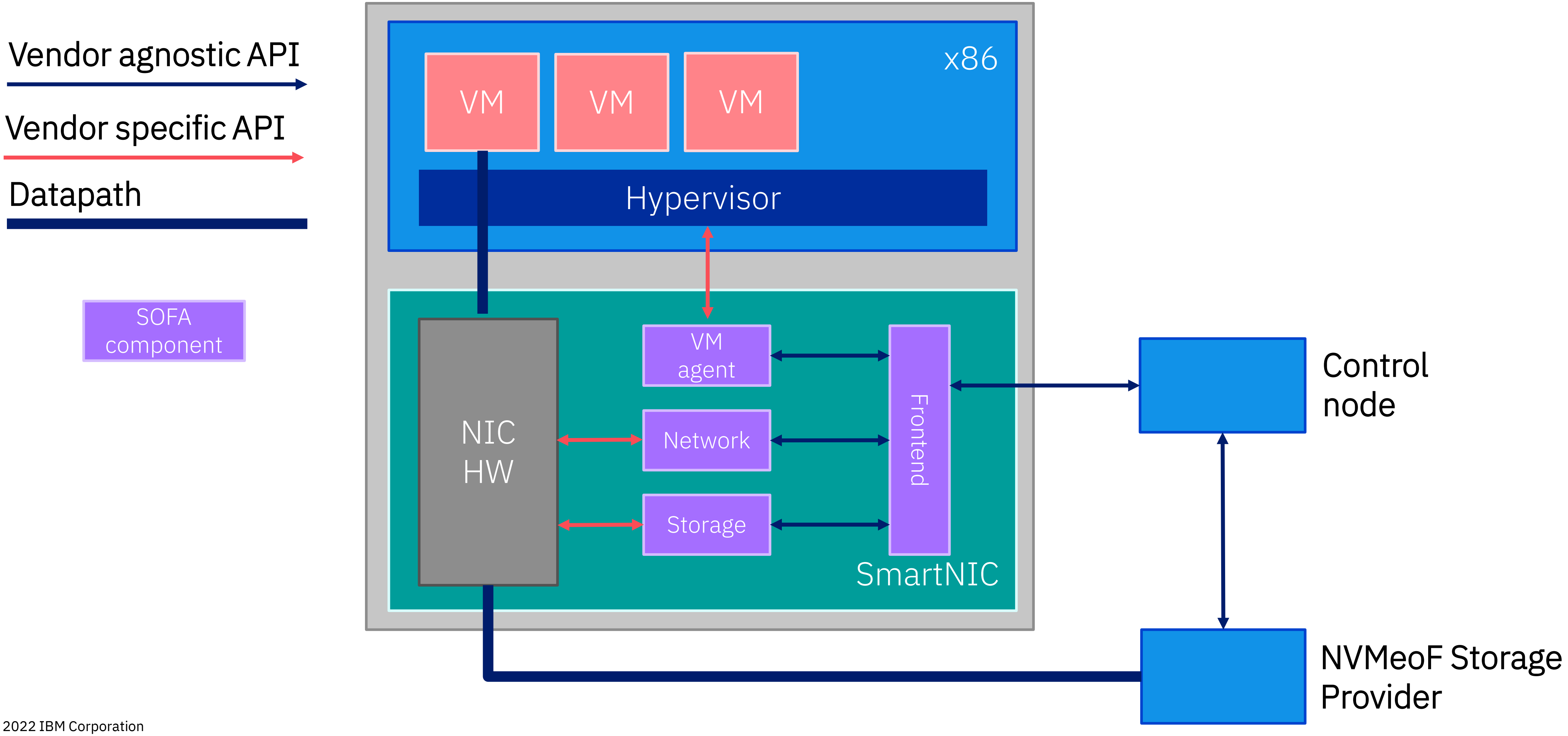
Storage Acceleration
- Ceph RBD Client Offload
- Hardware Offloaded Virtio-net
  - Virtio v0.9.5 and later
  - Multi-queue

Environmental Requirements[2]
- Temperature:
  - Operating: ≤ 30°C (86°F)
  - Storage: –40°C to 75°C (–40°F to 167°F)
- Humidity:
  - Operating: 8% to 90%, and a dew point of –12°C

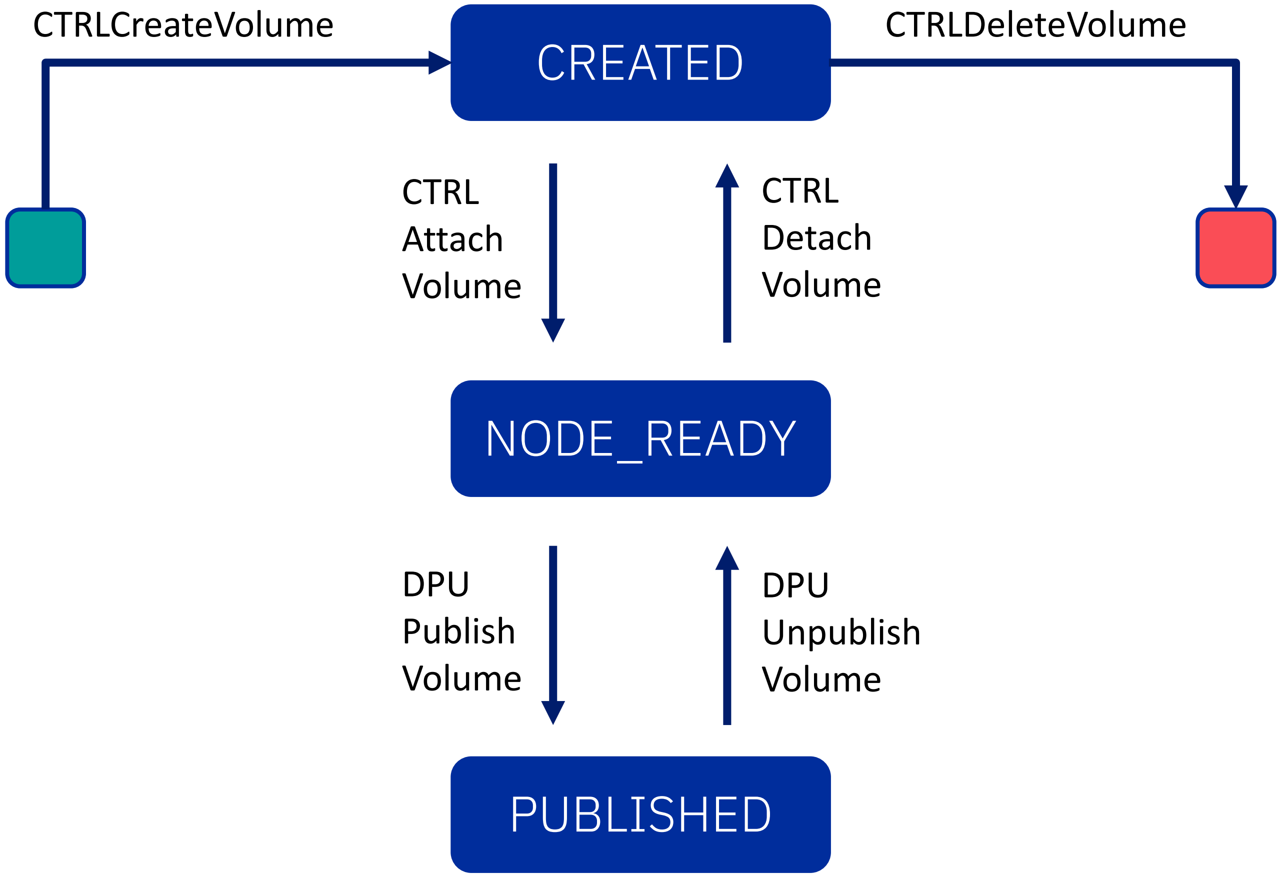
Mission statement

"Define a single API
to leverage SmartNICs from different vendors
to improve security and performance
in a Cloud environment."

SOFA: Secure Offload Framework for the Cloud



Volume lifecycle

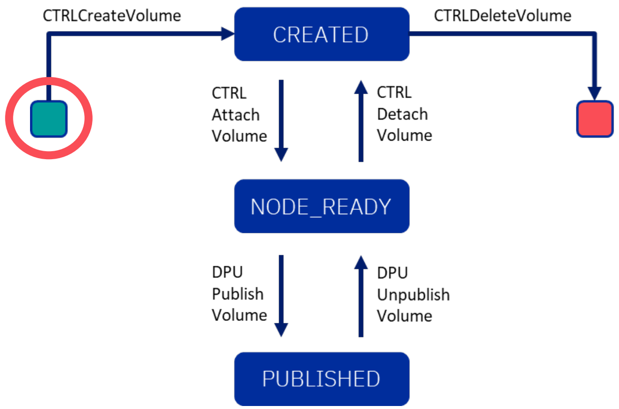


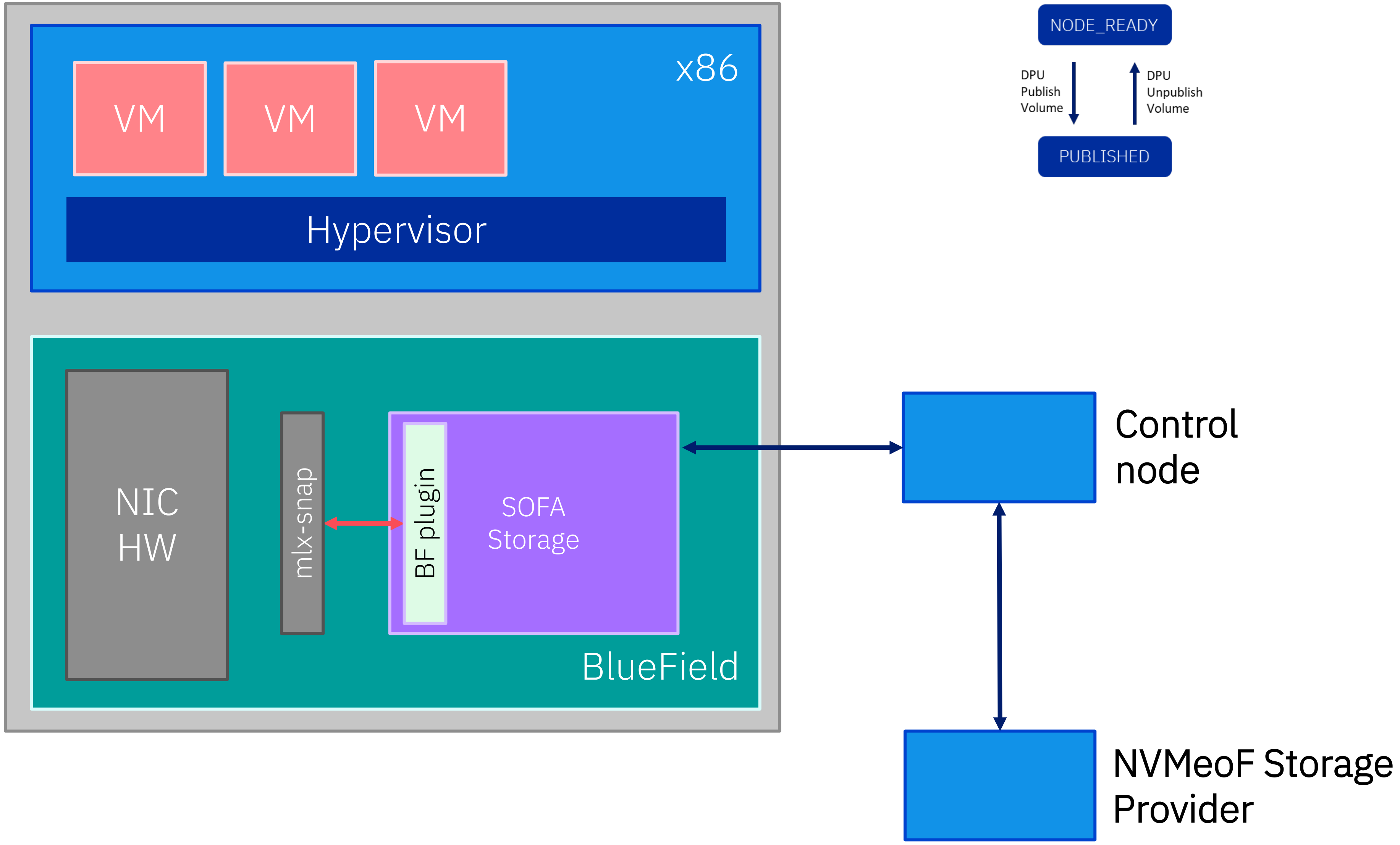
SOFA-Storage PoC

Vendor agnostic API →

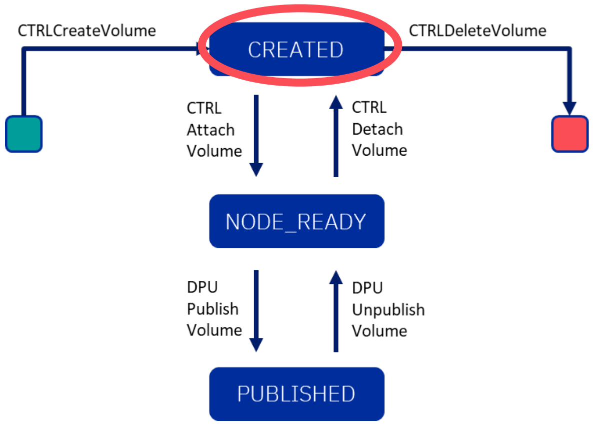
Vendor specific API →

Datapath





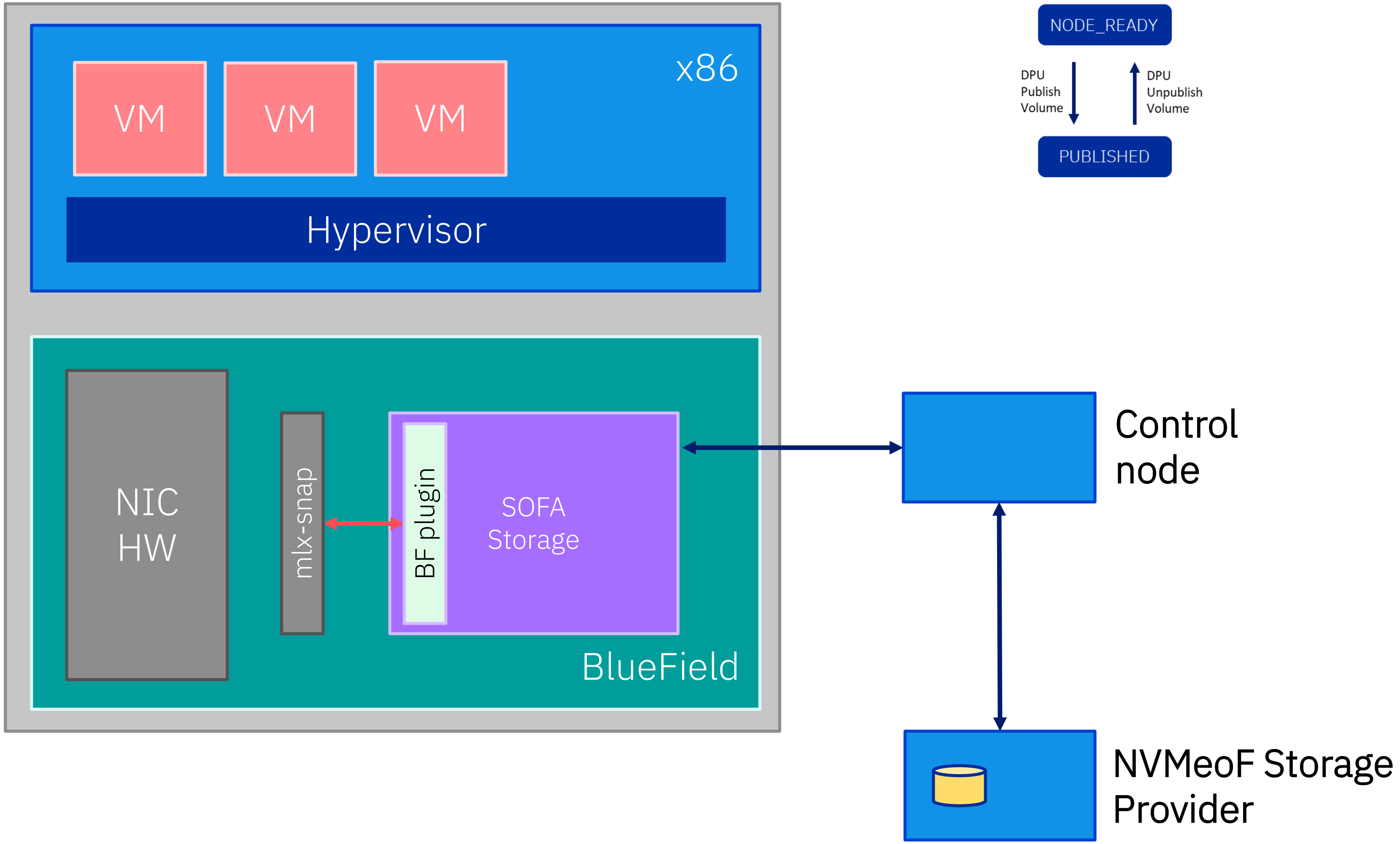
SOFA-Storage PoC



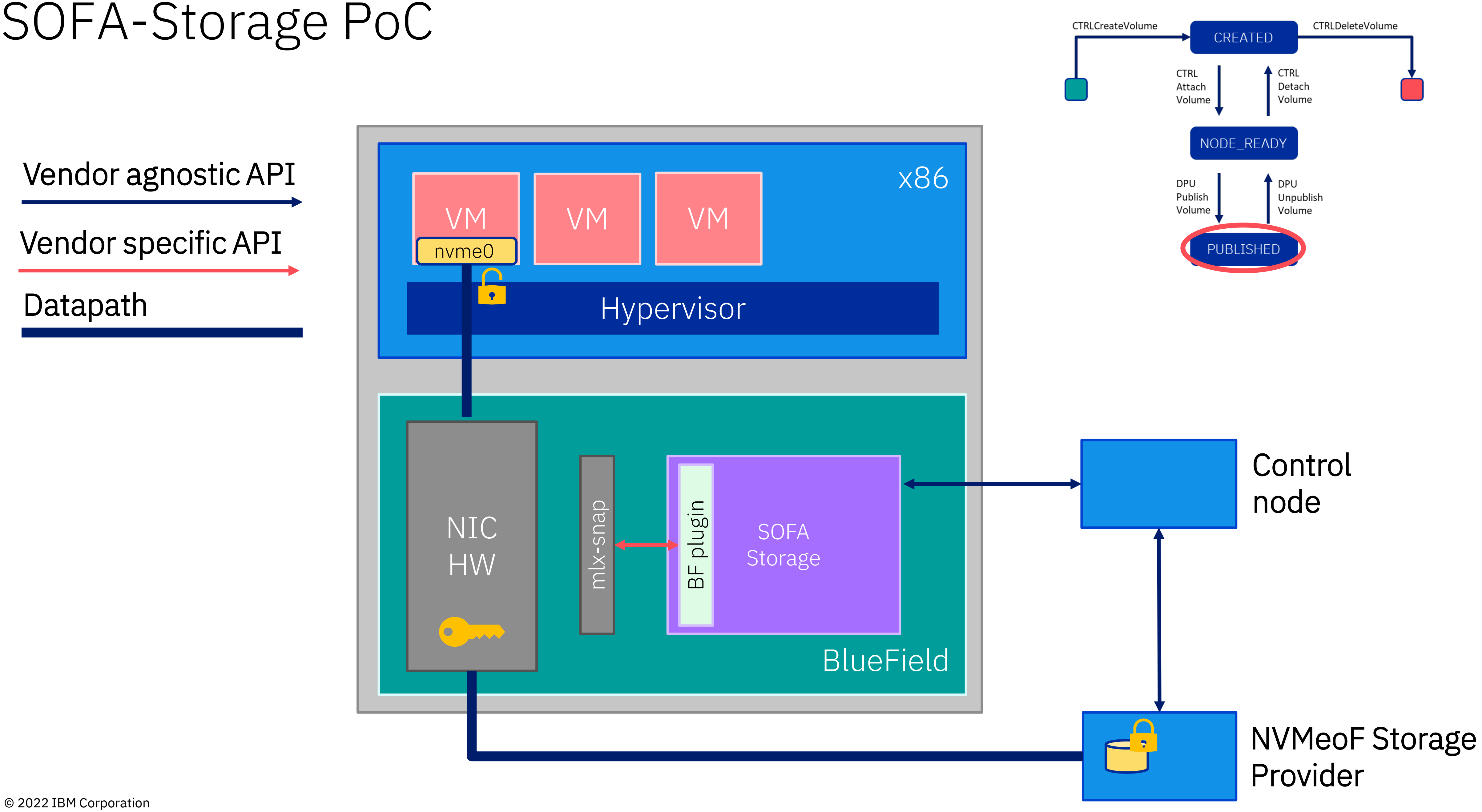
Vendor agnostic API

Vendor specific API

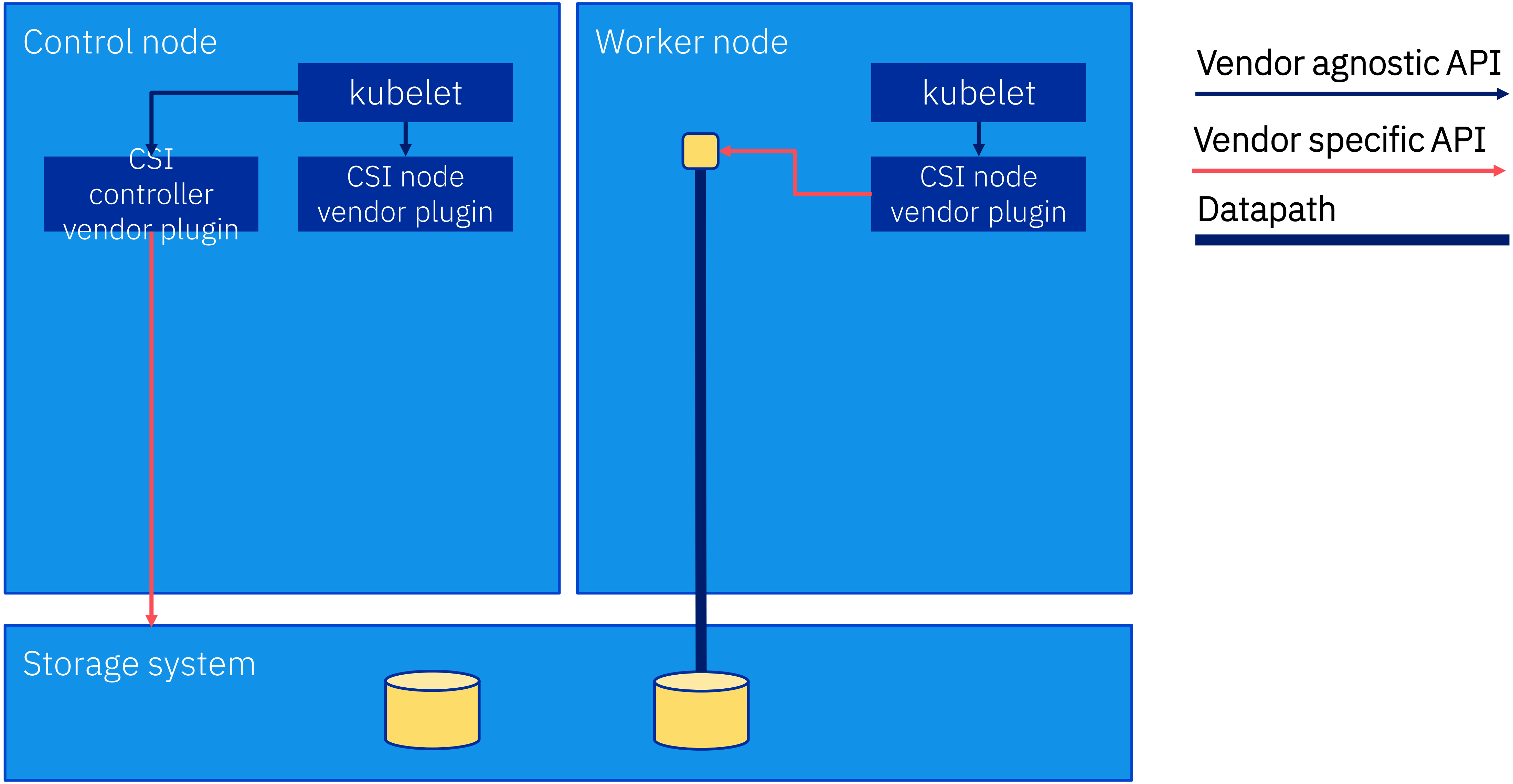
Datapath



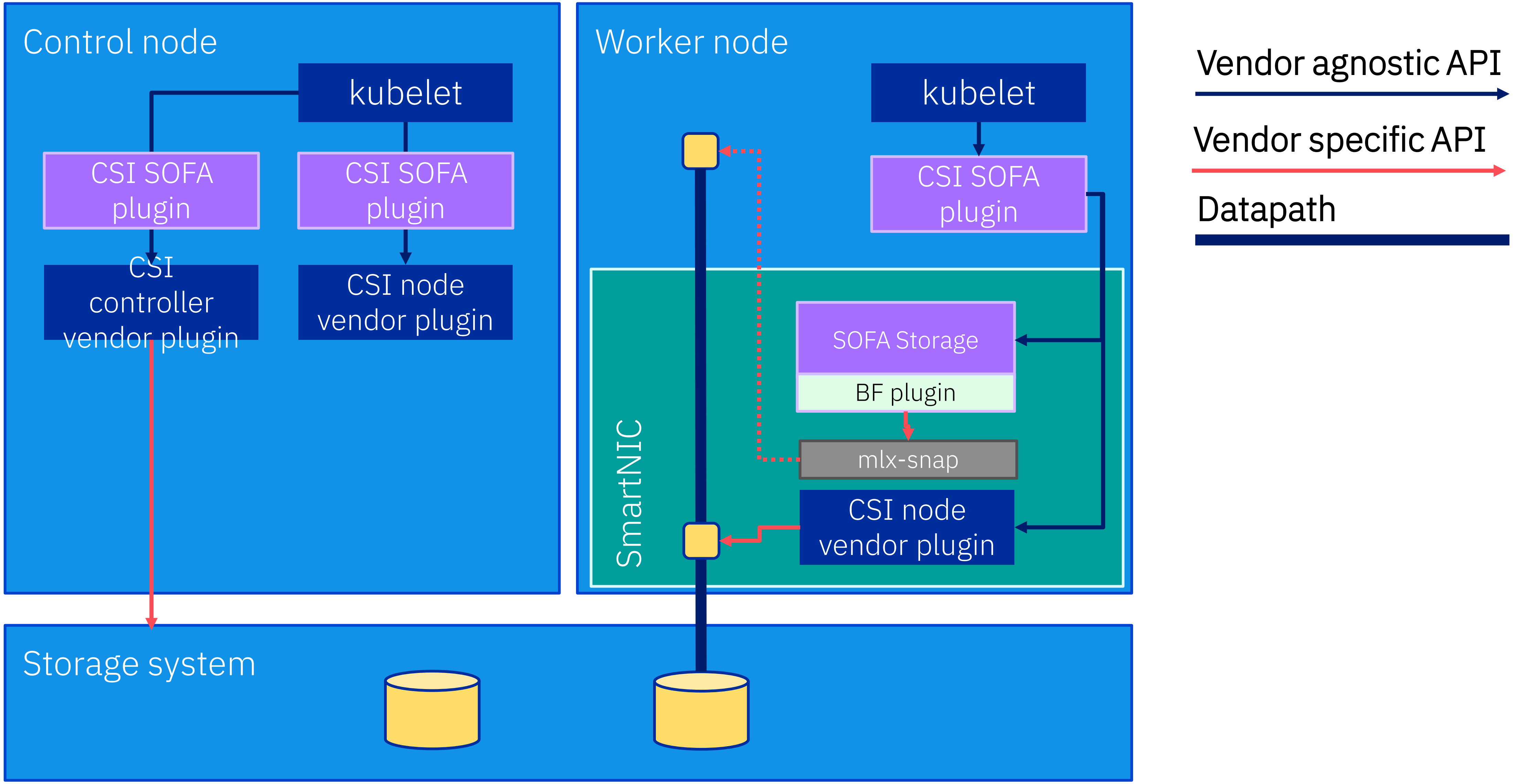
SOFA-Storage PoC



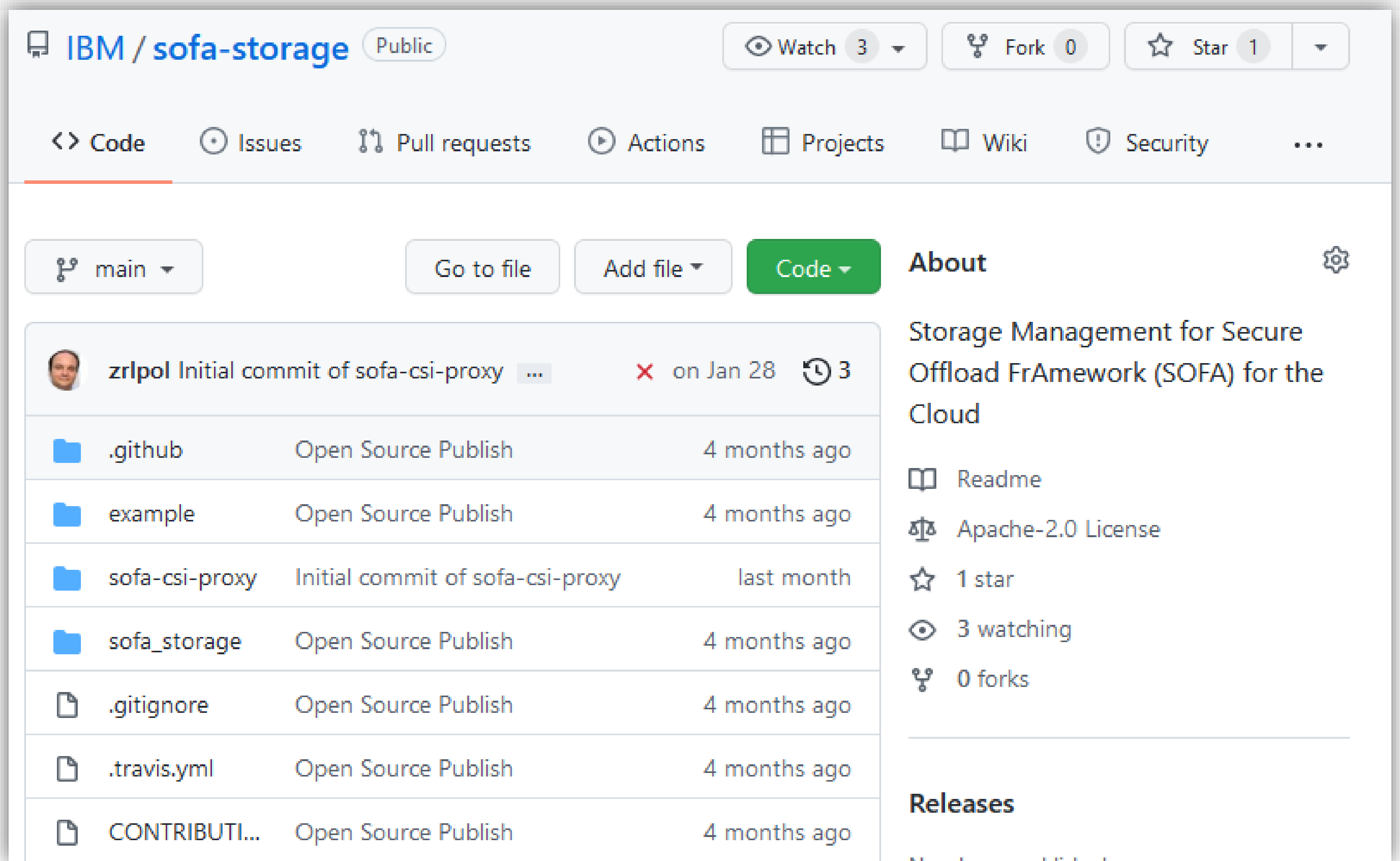
Kubernetes Container Storage Interface (CSI)



SOFA-Storage CSI Proxy



Open Source @ github.com/ibm/sofa-storage



zrlpol Initial commit of sofa-csi-proxy ... ✕ on Jan 28 3

.github	Open Source Publish	4 months ago
example	Open Source Publish	4 months ago
sofa-csi-proxy	Initial commit of sofa-csi-proxy	last month
sofa_storage	Open Source Publish	4 months ago
.gitignore	Open Source Publish	4 months ago
.travis.yml	Open Source Publish	4 months ago
CONTRIBUTI...	Open Source Publish	4 months ago

About

Storage Management for Secure Offload FrAmework (SOFA) for the Cloud

- Readme
- Apache-2.0 License
- 1 star
- 3 watching
- 0 forks

Releases

2022 OFA Virtual Workshop

THANK YOU

Raphael Polig, Jonas Pfefferle, Nikolas Ioannou

IBM Research - Zurich