



KubeCon



CloudNativeCon

Europe 2026

#KubeCon #CloudNativeCon

Virtualizing Large Scale GPU Cluster for Sovereign AI : Petasus AI Cloud Journey with Kubernetes

Jian Li, SK Telecom





KubeCon



CloudNativeCon

Europe 2026

Overview

SK Telecom Launches AI Infrastructure



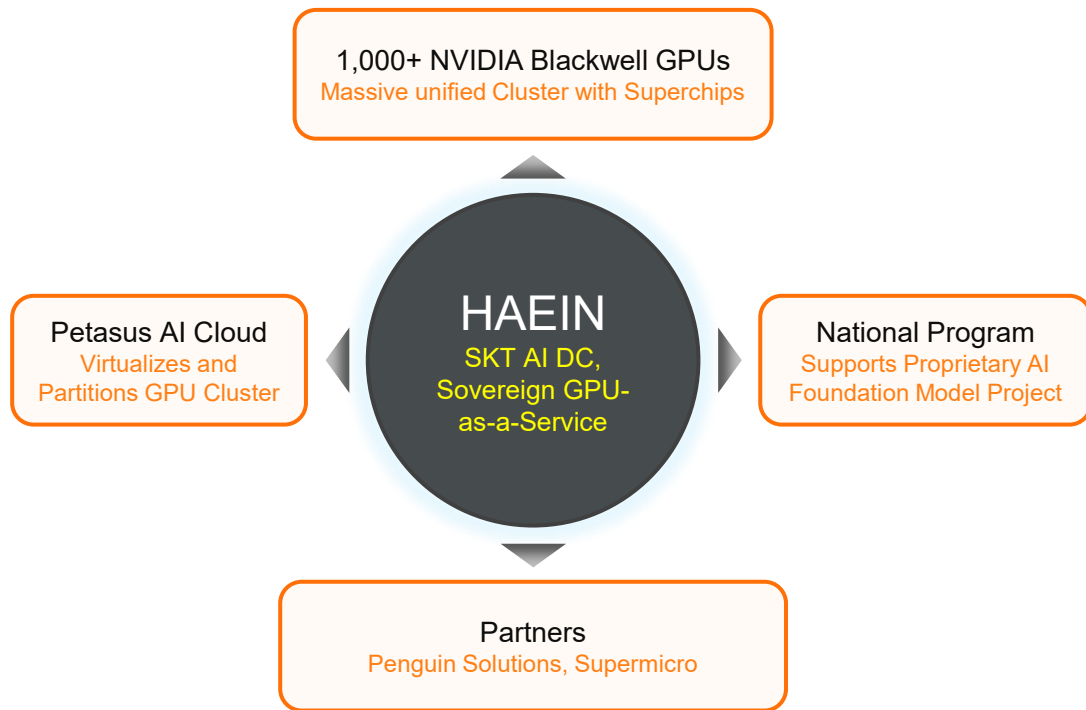
KubeCon



CloudNativeCon

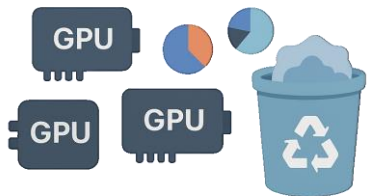
Europe 2026

Korea's sovereign GPU-as-a-Service, with 1,000+ Blackwell GPUs managed by Petasus AI Cloud



Haein Cluster: inspired by Haeinsa Temple – positioned as a core element of Korea's sovereign AI infrastructure

Challenges on Managing AI Infrastructure



GPU Resource Fragmentation & Waste

GPUs are statically allocated to specific teams, leading to low overall utilization and increased TCO

Minimal 8 GPUs →
1+ GPUs



AI Env. Provisioning Delays

Driver and library conflicts slow down experiment cycles and iteration speed

1+ days → 5 mins



Isolation & Security Risks

Increased potential for cross-tenant noise and data leakage

1 tenant → multi tenants



Scalability & Operational Overhead

More manual physical operations and staffing requirements, with delayed failure recovery

Several hours →
< 1 hour

Virtualizing AI Infrastructure

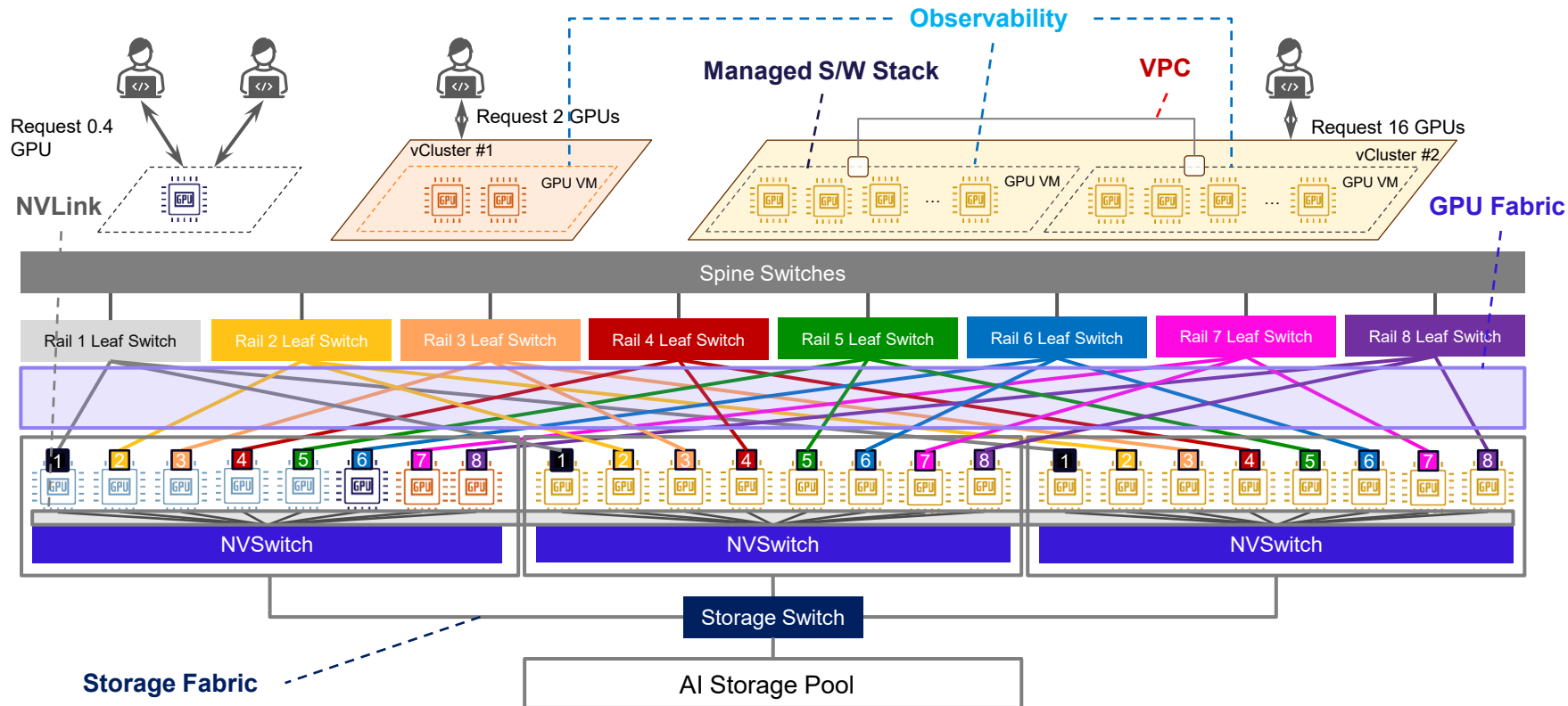


KubeCon



CloudNativeCon

Europe 2026

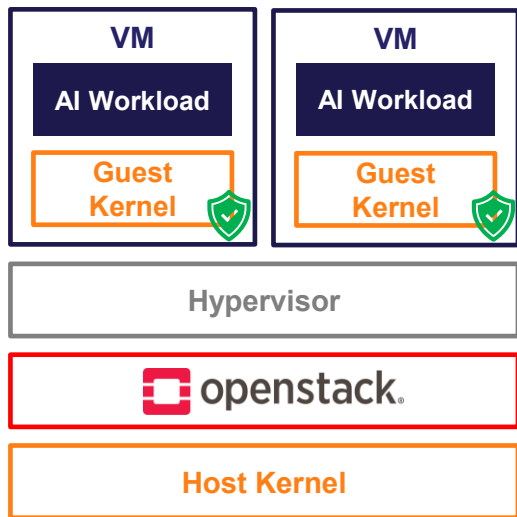


Evolution of Virtualization Technologies

1) Performance; 2) Multi-tenancy; 3) Management; 4) Security; 5) Maintenance

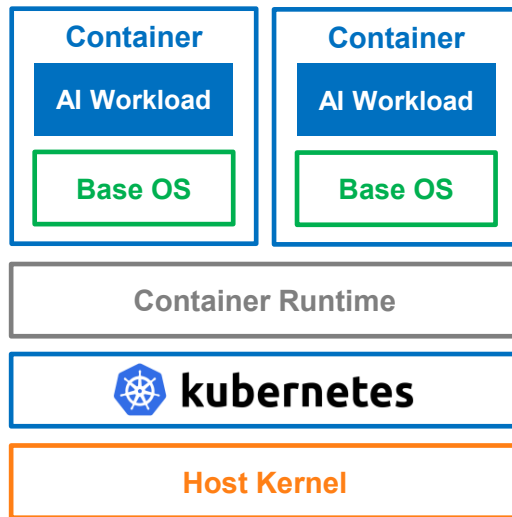
Virtual Machine

Secure but
Poor Performance, Inextensible



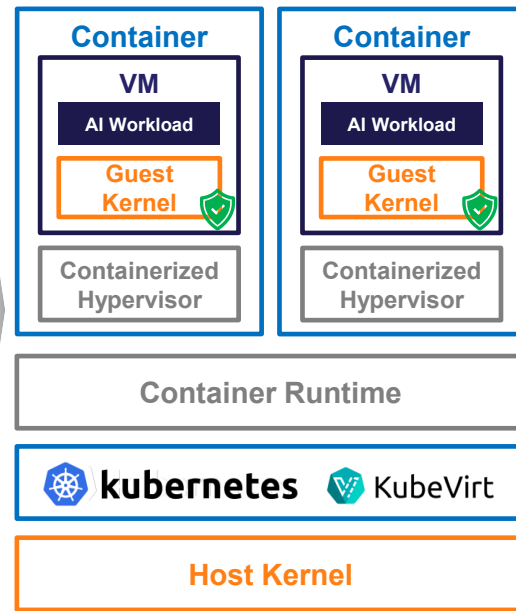
Container

Flexible, Extensible, Highly-performant
but Insecure



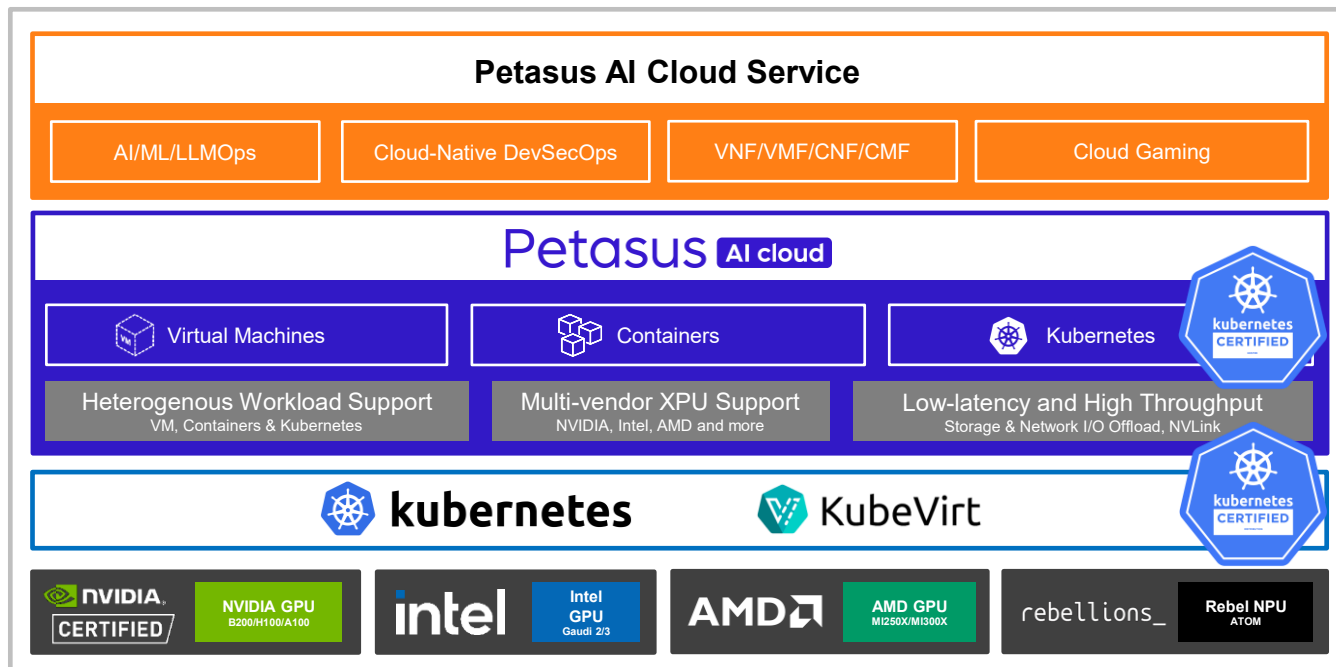
Containerized Virtual Machine

Flexible, Extensible, Secure, but Poor Performance



Petaus AI Cloud Architecture

Focuses on providing GPUaaS with an emphasis on high performance and cost efficiency



2024 Winner
SK Telecom for Cloud Radar
(Telco's AI-Powered Cloud Business Platform for Telcos & Enterprises)

2025 Winner
SK Telecom for Petaus Cloud

2026 Winner
SK Telecom for SKT Haein GPU Cluster



KubeCon



CloudNativeCon

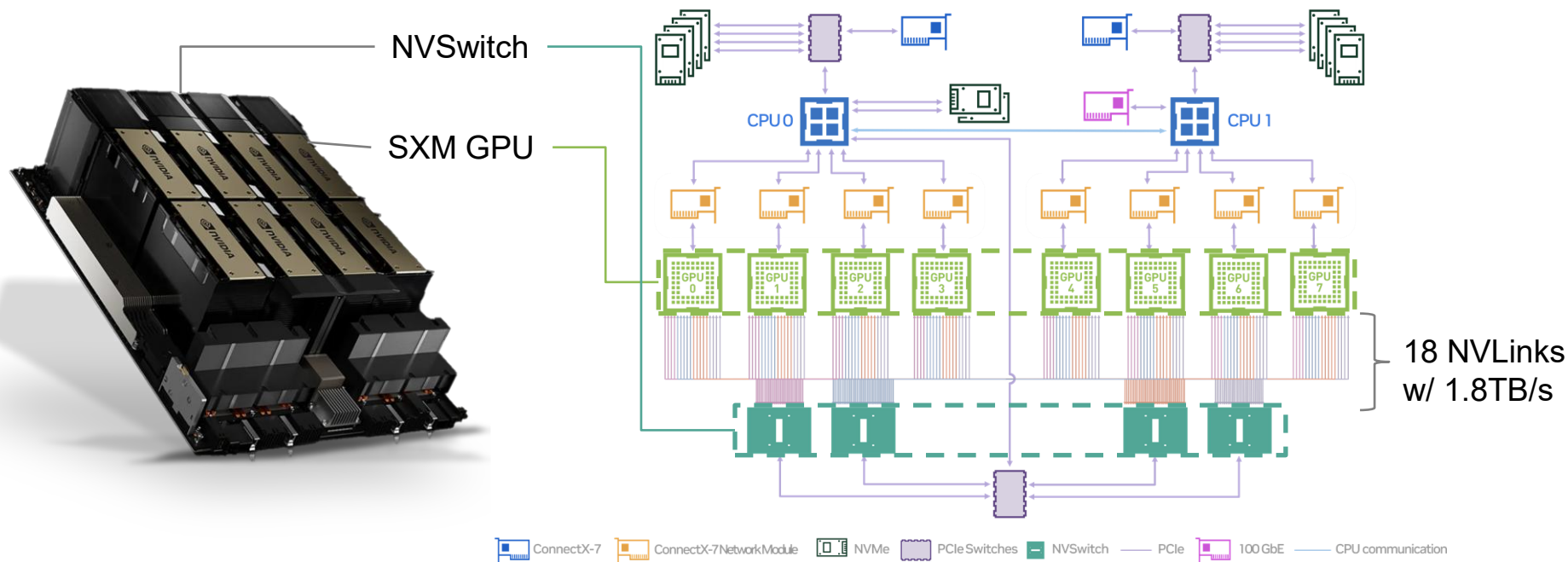
Europe 2026



NVLink and GDR Virtualization



NVLink enables fast communication among GPUs for enhanced performance



GPUDirect RDMA (GDR)

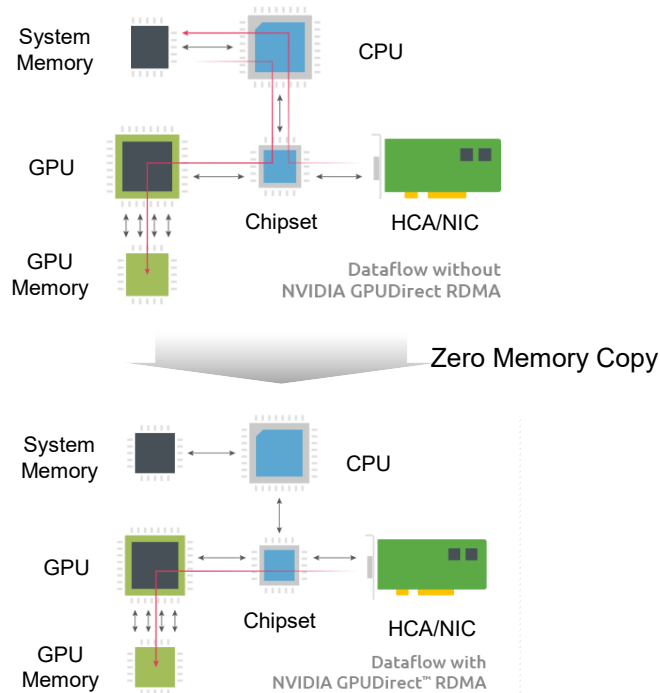
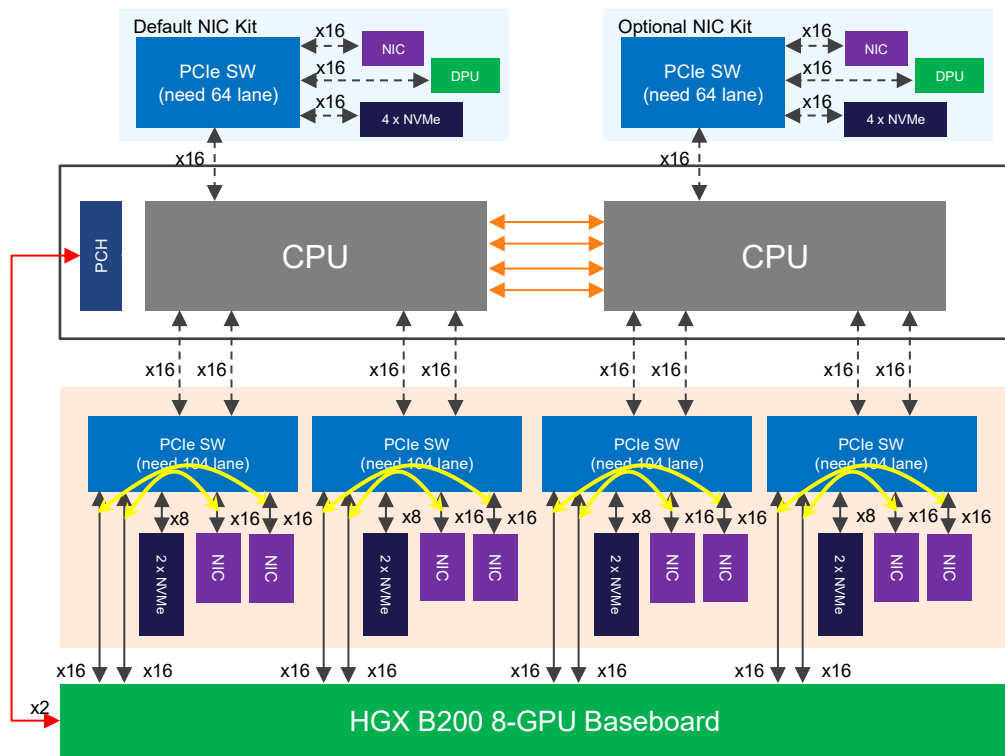


KubeCon



CloudNativeCon

Europe 2026



NUMA Aware PCIe Topo Arrangement



KubeCon

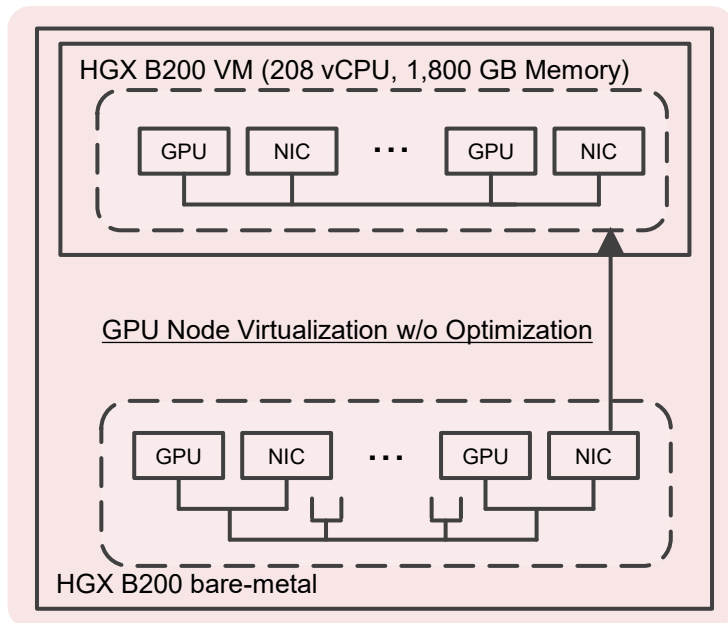


CloudNativeCon

Europe 2026

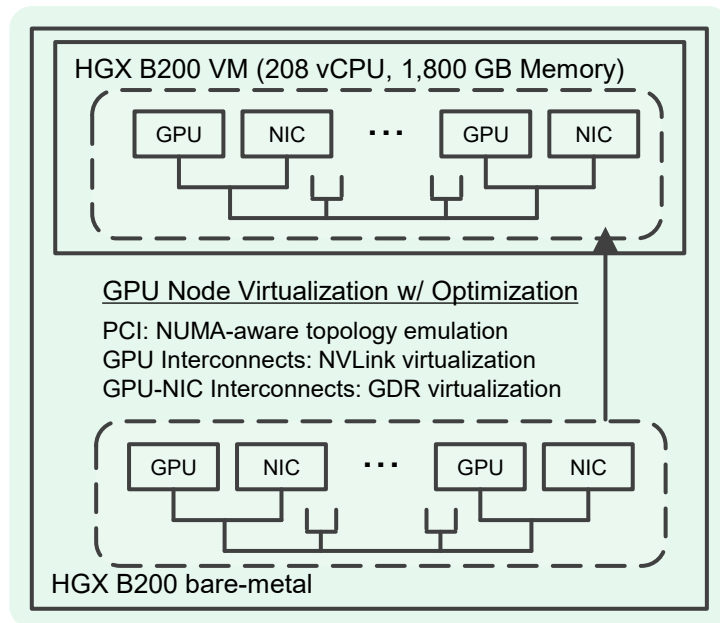
Strictly aligns with the HGX PCIe topology of bare-metal to realize **GDR** and **GDS** for VM

AS-IS



**PCIe
Topology
Arrangement**

TO-BE



<https://github.com/kubevirt/enhancements/blob/main/veps/sig-compute/115-pcie-numa-topology-awareness/pcie-numa-topology-awareness.md>

Dev. and Vol. Assignment with KubeVirt

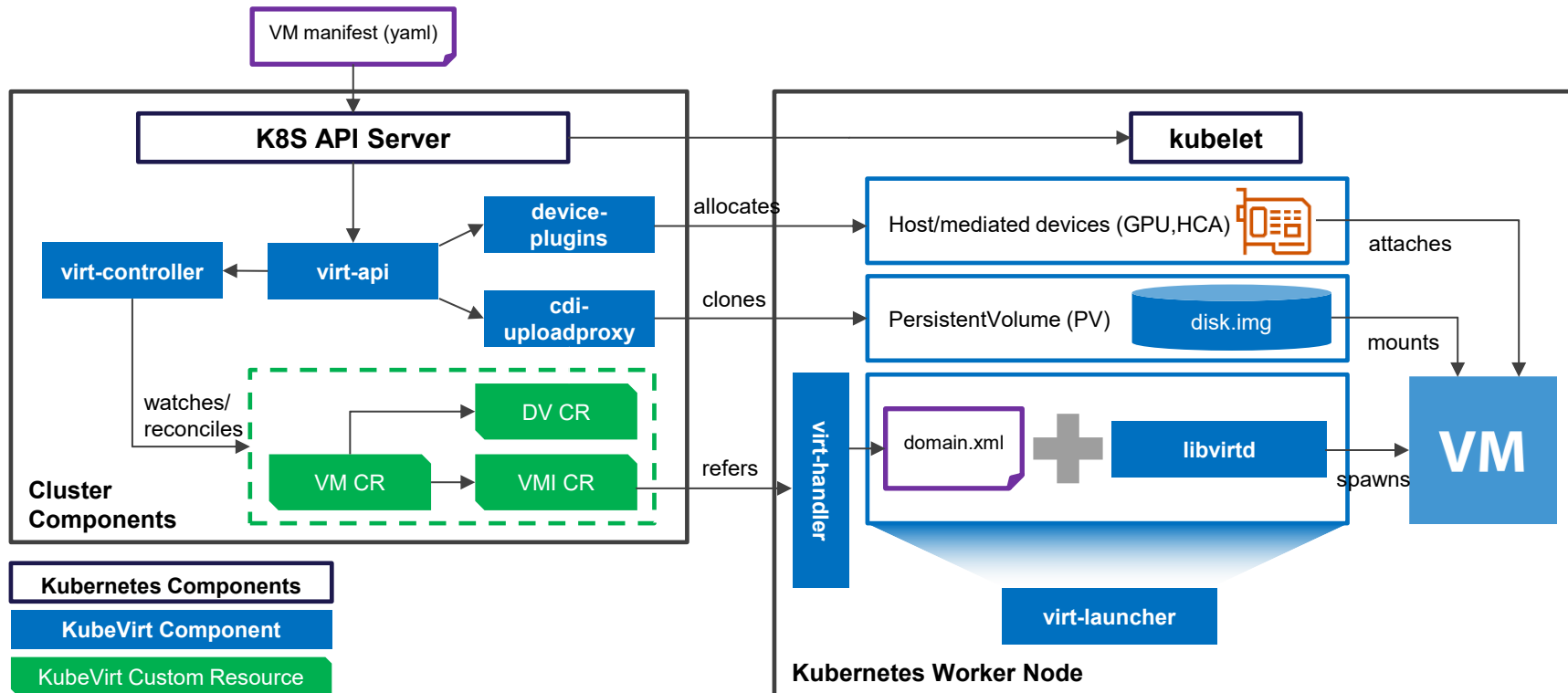


KubeCon



CloudNativeCon

Europe 2026



Virtualized GDR and NVLink!

VM provisioning with Virtualized NVLinks with GDR enabled on DGX/HGX B200

```
ubuntu@truss-001:~$ nvidia-smi topo -m
```

	GPU0	GPU1	GPU2	GPU3	GPU4	GPU5	GPU6	GPU7	NIC0	NIC1	NIC2	NIC3	NIC4	NIC5	NIC6	NIC7	CPU Affinity	NUMA Affinity	GPU NUMA ID
GPU0	X	NV18	NV18	NV18	NV18	NV18	NV18	NV18	PIX	18-lane NVLink	NODE	NODE	SYS	SYS	SYS	SYS	80-159	1	N/A
GPU1	NV18	X	NV18	NV18	NV18	NV18	NV18	NV18	NODE	full-mesh	NODE	NODE	SYS	SYS	SYS	SYS	80-159	1	N/A
GPU2	NV18	NV18	X	NV18	NV18	NV18	NV18	NV18	NODE	NODE	PIX	NODE	SYS	SYS	SYS	SYS	80-159	1	N/A
GPU3	NV18	NV18	NV18	X	NV18	NV18	NV18	NV18	NODE	NODE	NODE	PIX	SYS	SYS	SYS	SYS	80-159	1	N/A
GPU4	NV18	NV18	NV18	NV18	X	NV18	NV18	NV18	SYS	SYS	SYS	SYS	PIX	NODE	NODE	NODE	0-79	0	N/A
GPU5	NV18	NV18	NV18	NV18	NV18	X	NV18	NV18	SYS	SYS	SYS	SYS	NODE	PIX	NODE	NODE	0-79	0	N/A
GPU6	NV18	NV18	NV18	NV18	NV18	NV18	X	NV18	SYS	SYS	SYS	SYS	NODE	NODE	PIX	NODE	0-79	0	N/A
GPU7	NV18	NV18	NV18	NV18	NV18	NV18	NV18	X	SYS	SYS	SYS	SYS	NODE	NODE	NODE	PIX	0-79	0	N/A
NIC0	PIX	NODE	NODE	NODE	SYS	SYS	SYS	SYS	X	NODE	NODE	NODE	SYS	SYS	SYS	SYS			
NIC1	NODE	PIX	NODE	NODE	SYS	SYS	SYS	SYS	NODE	X	NODE	NODE	SYS	SYS	SYS	SYS			
NIC2	NODE	NODE	PIX	NODE	SYS	SYS	SYS	SYS	NODE	NODE	X	NODE	SYS	SYS	SYS	SYS			
NIC3	NODE	NODE	NODE	PIX	SYS	SYS	SYS	SYS	NODE	NODE	NODE	X	SYS	SYS	SYS	SYS			
NIC4	SYS	SYS	SYS	SYS	PIX	NODE	NODE	NODE	SYS	SYS	SYS	SYS	X	NODE	NODE	NODE			
NIC5	SYS	SYS	SYS	SYS	SYS	NODE	PIX	NODE	SYS	SYS	SYS	SYS	NODE	X	NODE	NODE			
NIC6	SYS	SYS	SYS	SYS	NODE	NODE	PIX	NODE	SYS	SYS	SYS	SYS	NODE	NODE	X	NODE			
NIC7	SYS	SYS	SYS	SYS	NODE	NODE	NODE	PIX	SYS	SYS	SYS	SYS	NODE	NODE	NODE	X			

Legend:

- X = Self
- SYS = Connection traversing PCIe as well as the SMP interconnect between NUMA nodes (e.g., QPI/UPI)
- NODE = Connection traversing PCIe as well as the interconnect between PCIe Host Bridges within a NUMA node
- PHB = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
- PIX = Connection traversing multiple PCIe bridges (without traversing the PCIe Host Bridge)
- PIX = Connection traversing at most a single PCIe bridge
- NV# = Connection traversing a bonded set of # NVLinks

PXN in Rail-optimized Virtual Environment



KubeCon

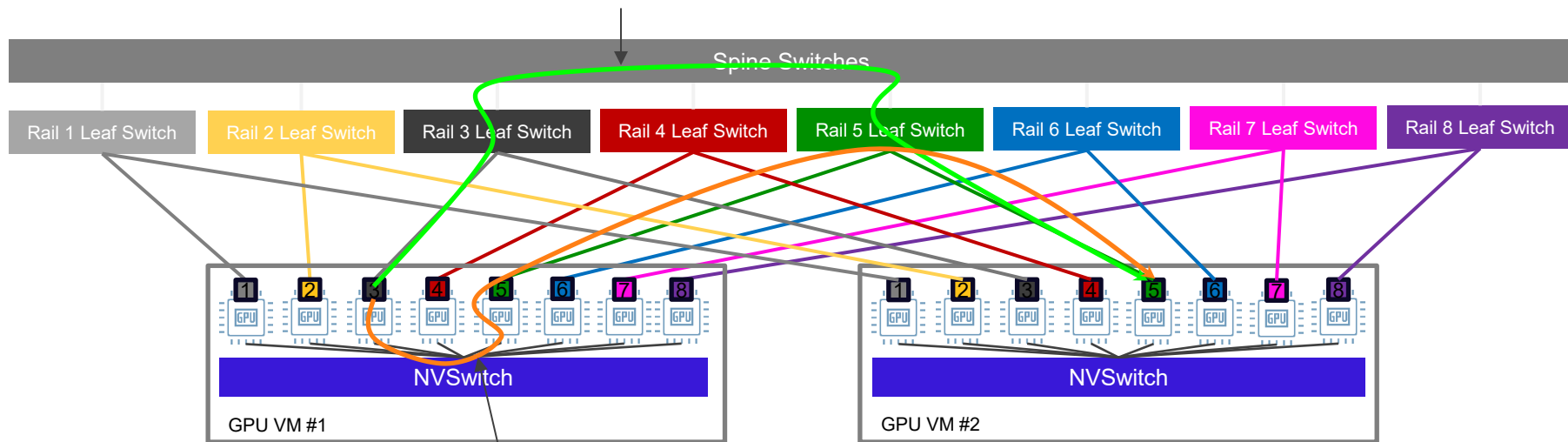


CloudNativeCon

Europe 2026

PXN combines PCIe (GDR) and NVLink to reduce traffic flow and optimizes network traffic

Slow path with potential spine switch congestion, 3 hops required at least



Fast path with no congestion, only 1 hop required

→ Message path without PXN
→ Message path with PXN

GPU Fabric Isolation



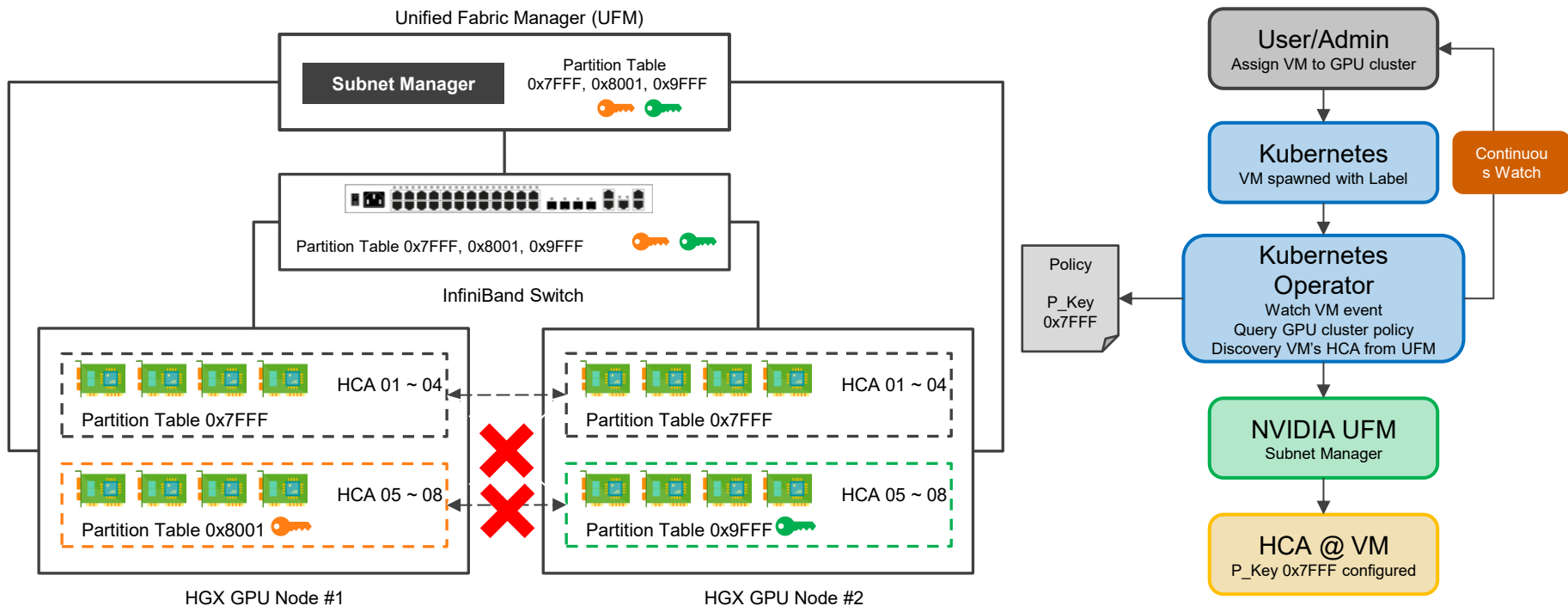
KubeCon



CloudNativeCon

Europe 2026

Full InfiniBand(IB) network isolation using Partition Key (P_Key)
Full remote GPU to GPU communication with **GDR** and **GDS** enabled





KubeCon



CloudNativeCon

Europe 2026

Thank you

