Building Custom AI Infrastructure with NVLink Fusion

Rising Complexity of AI Models: From CPUs to the Age of Reasoning

Rack-scale computing: the foundation for tomorrow’s AI infrastructure

Early AI Era

≤100M parameters

Follows Moore’s Law –
Model size doubling every 20 months

Inference runs on CPUs

GPU AI Era

~100M to ~1B parameters

Doubling every 6 months

Inference runs on 1 GPU

Multi-GPU AI Era

~1B to multi-trillion parameters

Commercial AI Driven

Doubling every 10 months

Inference running up to 8 GPUs

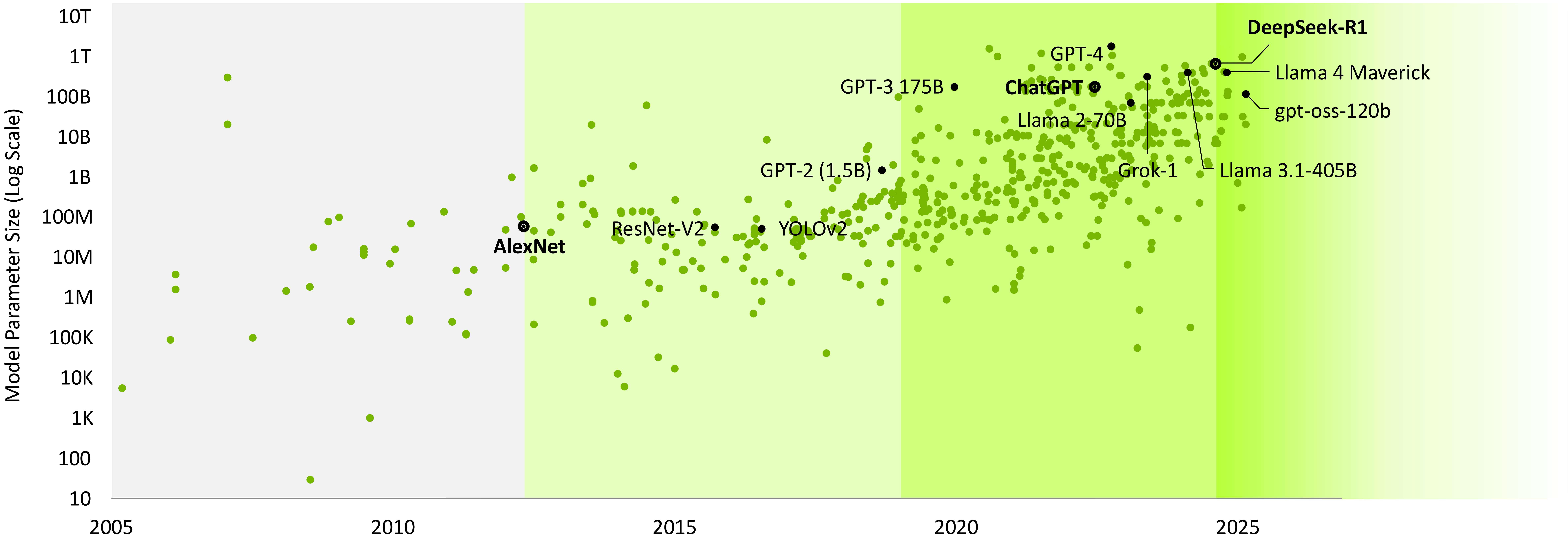
Age of AI Reasoning at Scale

Drastic Increase in Compute for Reasoning

Expansion of Distributed Parallelism Techniques

Large Scale Mixture of Experts

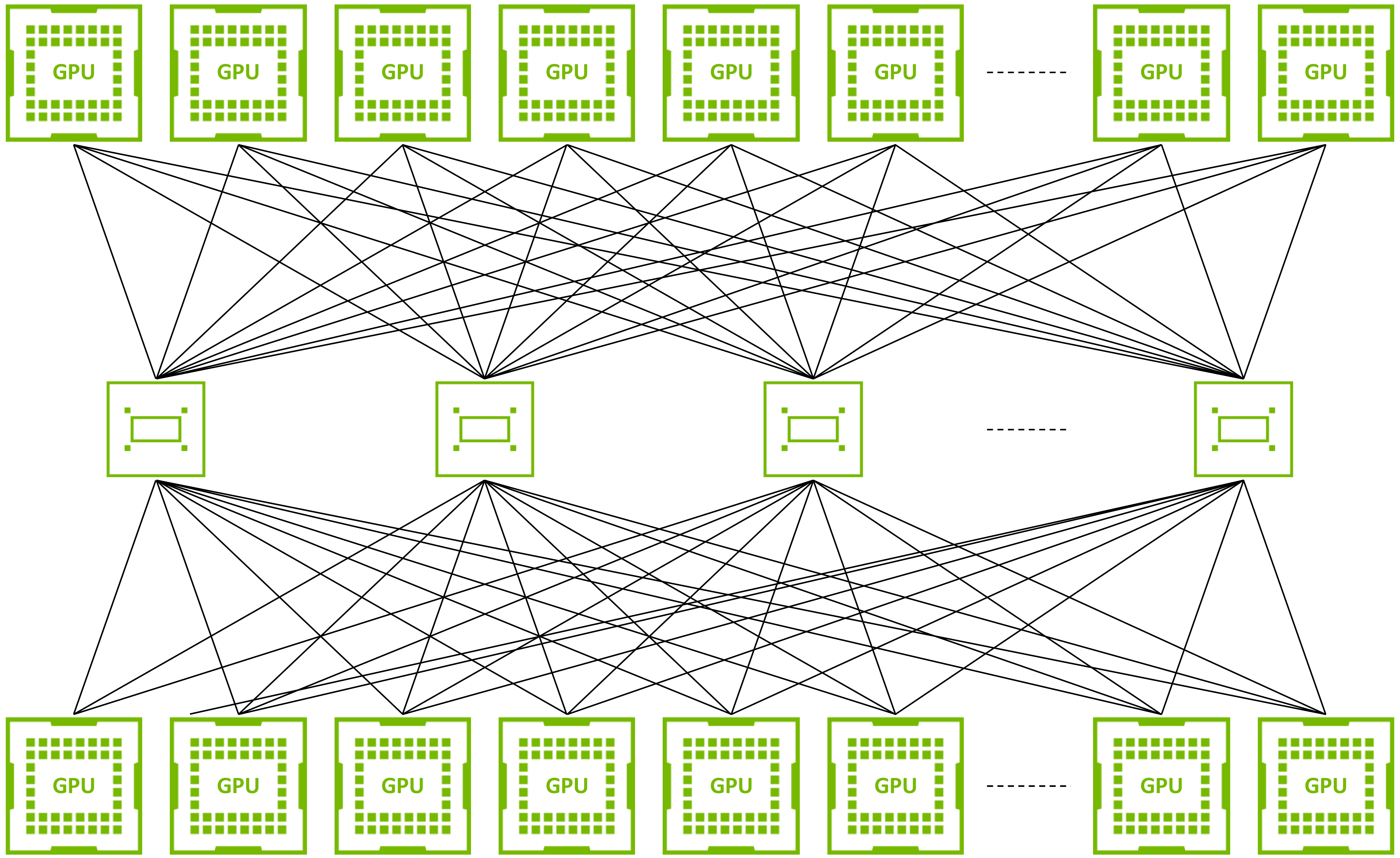
Inference running up to 72 GPUs



Data Source: Epoch (2025) – with major processing by Our World In Data

Requirements of Scale-Up Fabric for Rack Scale Computing

Scale-up fabric is a compute fabric



Very large models

Tensor parallelism

Expert parallelism

Large contexts

Domain size

Peer-compute memory access

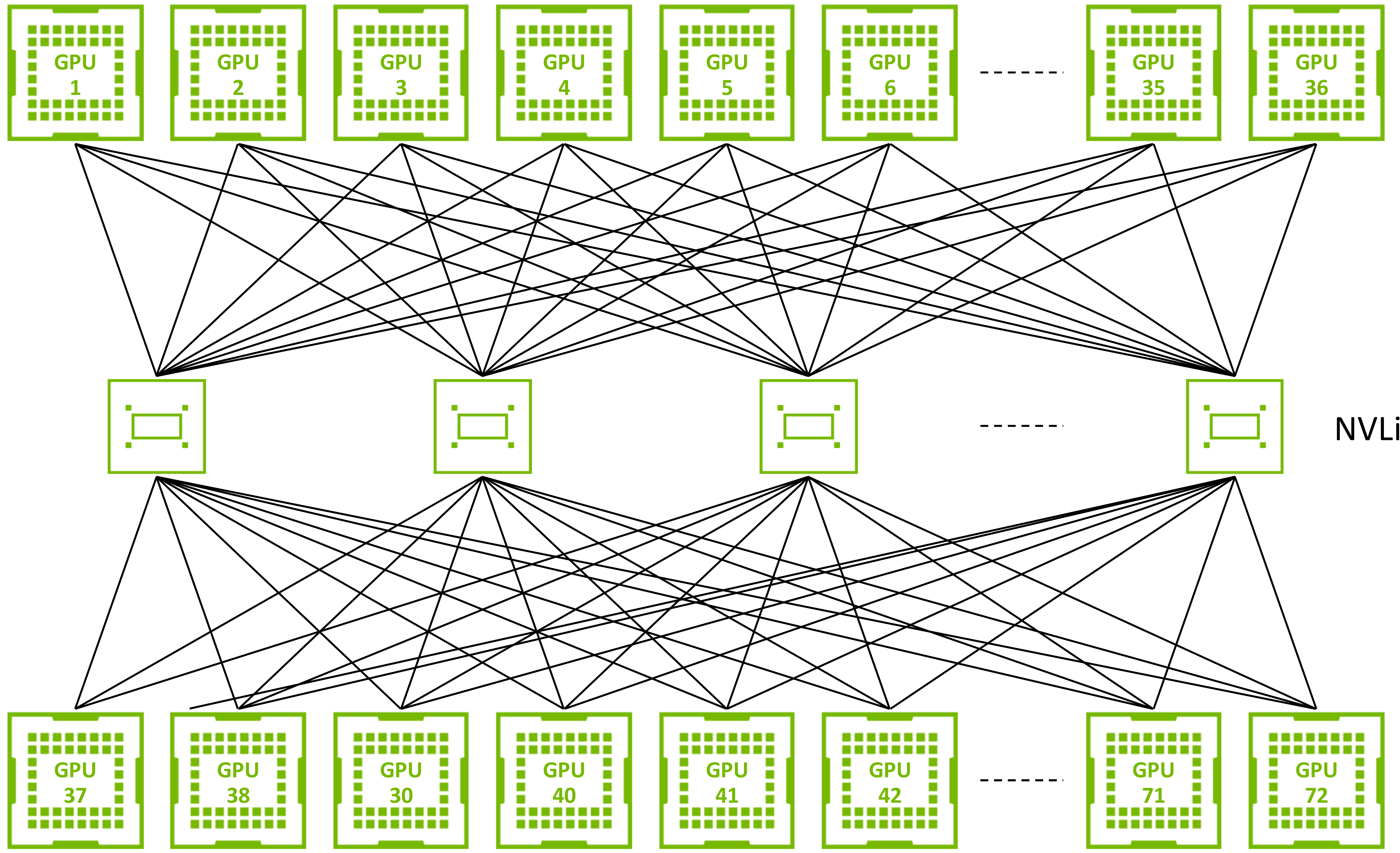
High bandwidth, low latency

Collective offloads

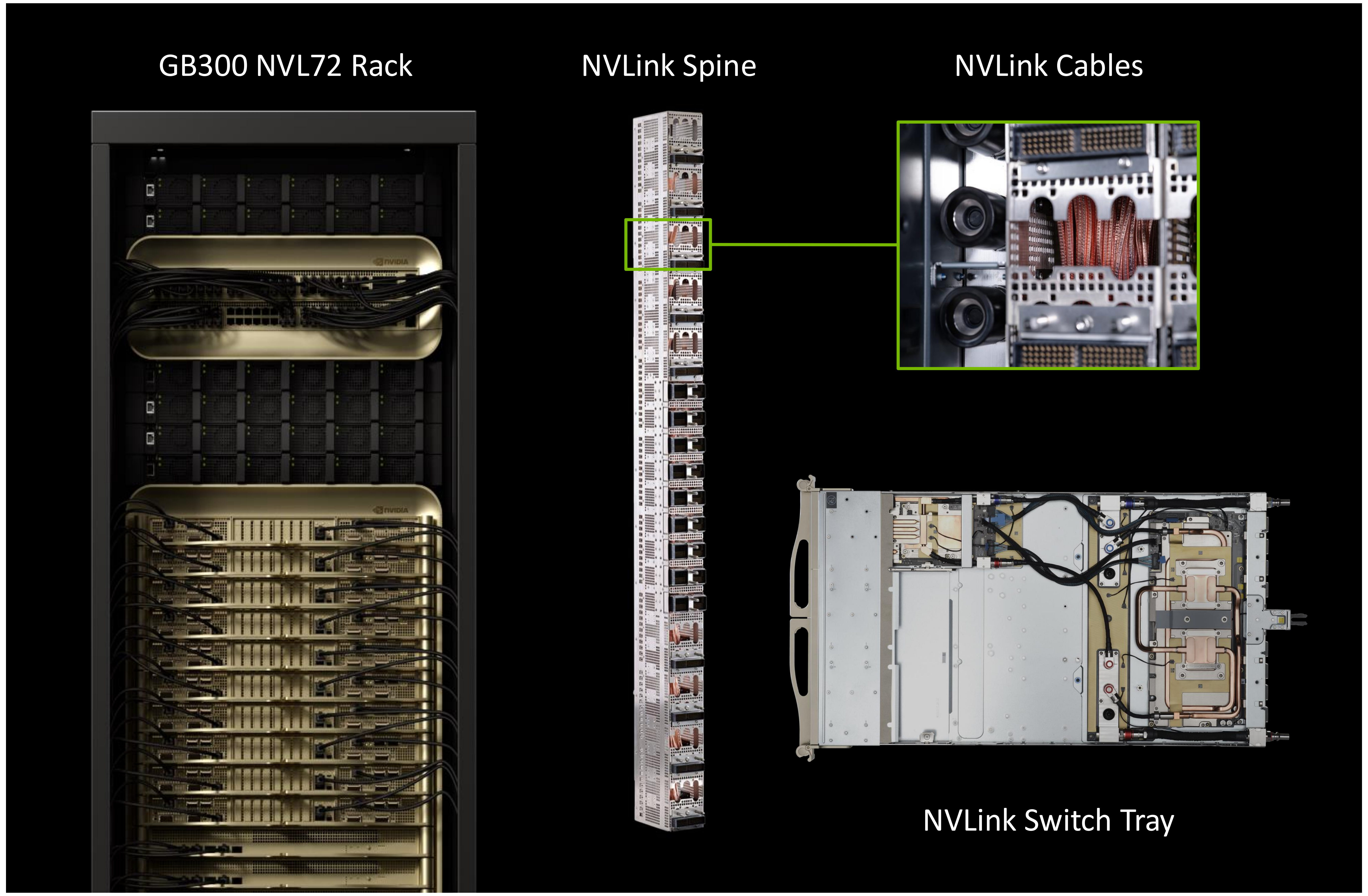
Flexible workload shapes

NVIDIA GB300 NVL72

Scaling GPU domains with NVIDIA NVLink



18x
NVLink Switch
Chips


GB300 NVL72 Rack

NVLink Spine

NVLink Cables

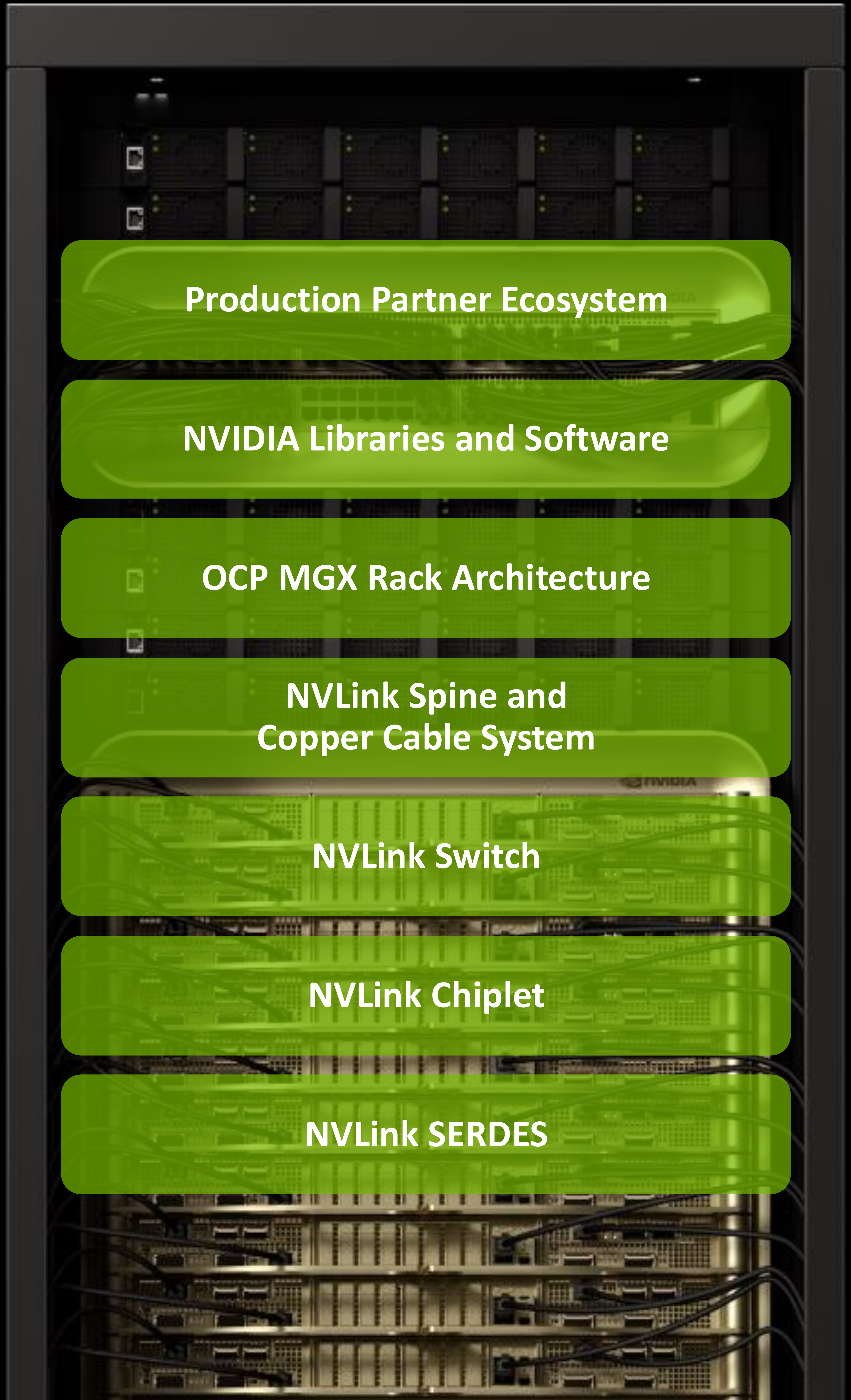
NVLink Switch Tray

GPU Bandwidth	Domain Size	All-to-All	All-Reduce
1.8 TB/s	72 GPUs	130 TB/s	260 TB/s

NVLink Fusion

NVIDIA NVLink Fabric and rack architecture for custom compute

- Single, scalable AI factory architecture
- Brings together entire rack architecture
- Proven scale-up and scale-out roadmap and ecosystem



Production Partner Ecosystem

NVIDIA Libraries and Software

OCP MGX Rack Architecture

NVLink Spine and Copper Cable System

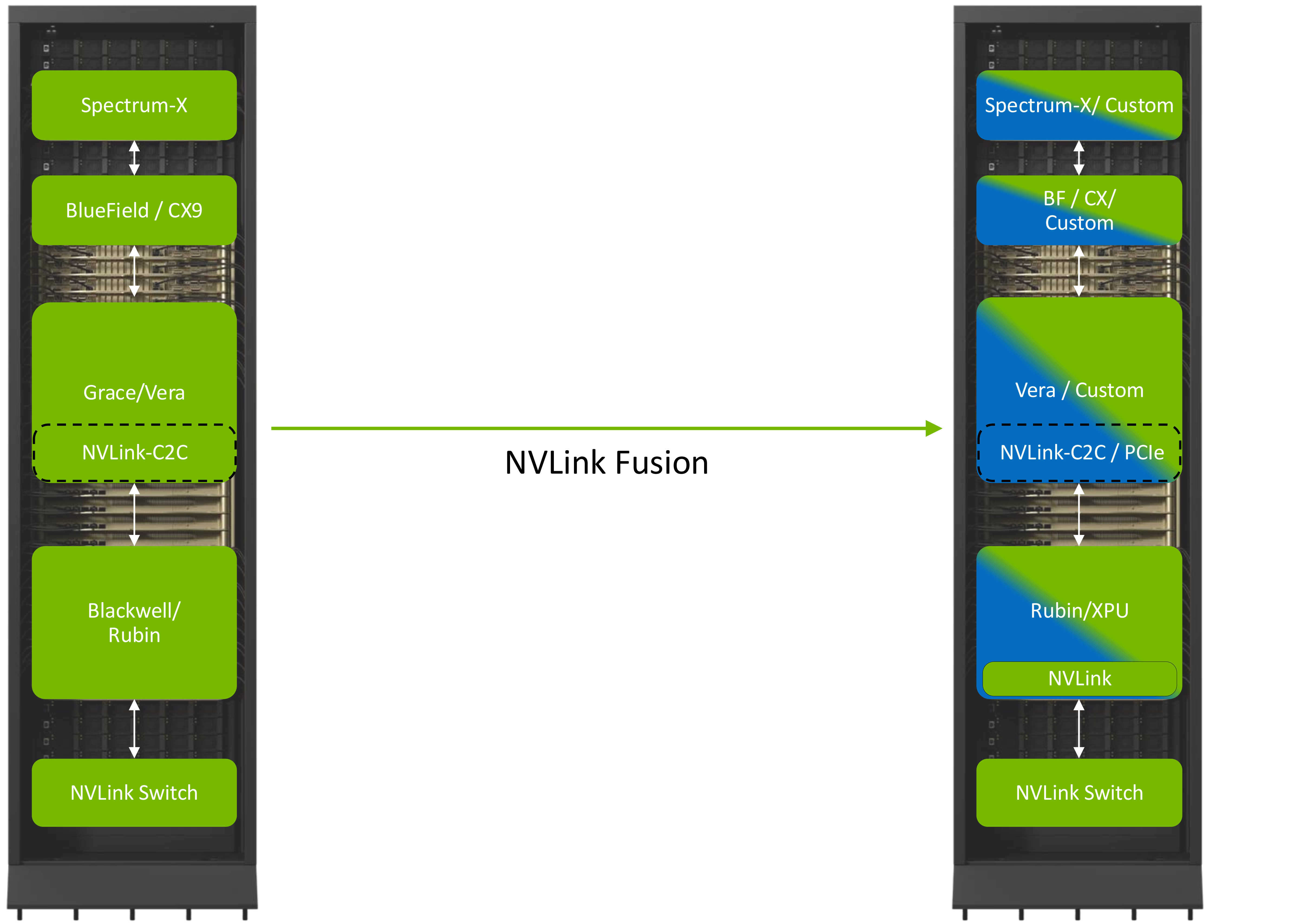
NVLink Switch

NVLink Chiplet

NVLink SERDES

NVLink Fusion

Custom Combinations with Open Standards Integration



NVIDIA Rack Architecture

NVLink Fusion Custom Rack Architecture

Production Partner Ecosystem

NVIDIA Libraries and Software

OCP MGX Rack Architecture

NVLink Spine and Copper Cable System

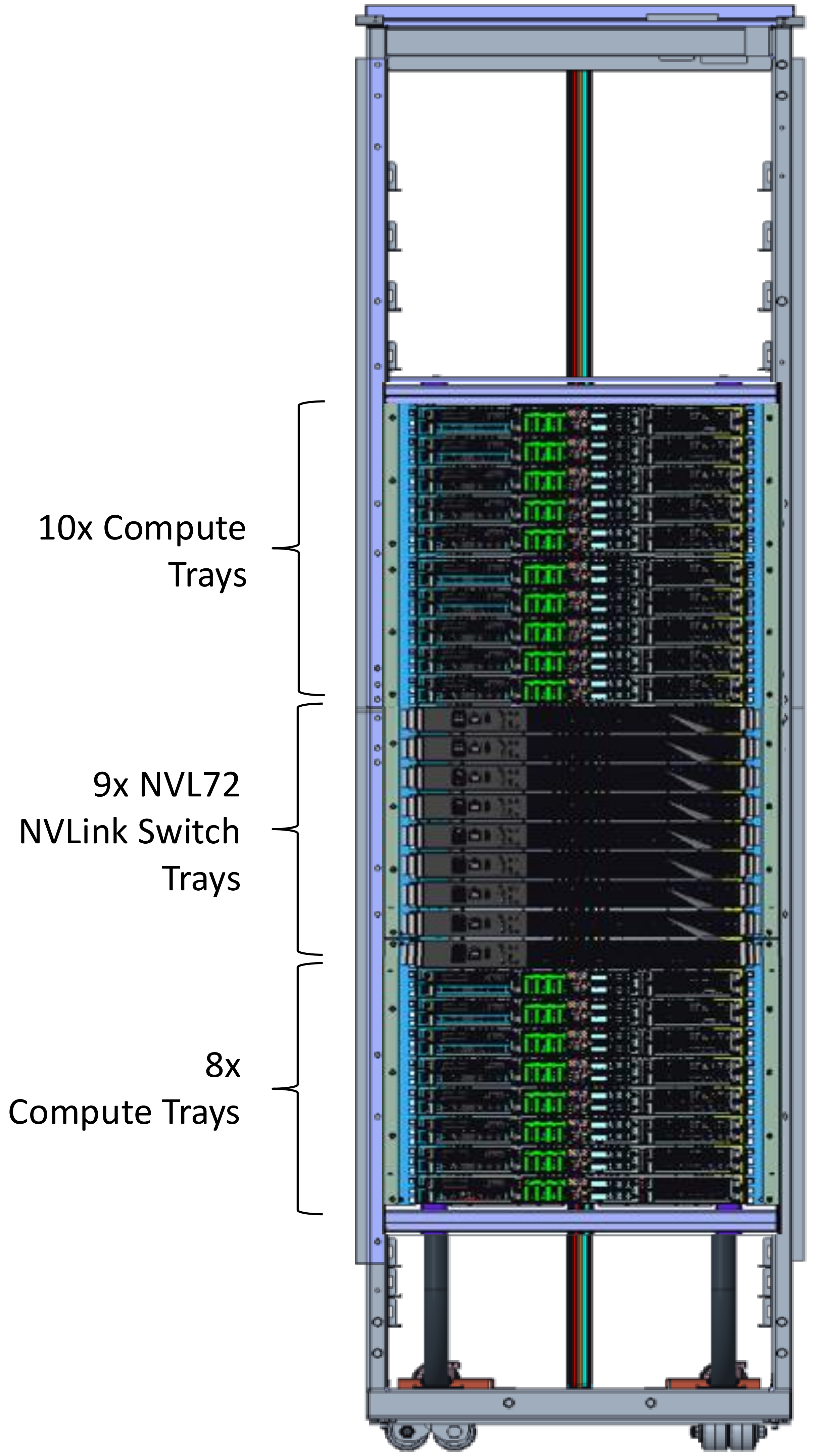
NVLink Switch

NVLink Chiplet

NVLink SERDES

NVIDIA NVL72 Rack Architecture

OCP MGX Rack



Single 72-GPU L1 Domain

Fully Copper Domain

9x Switch Trays

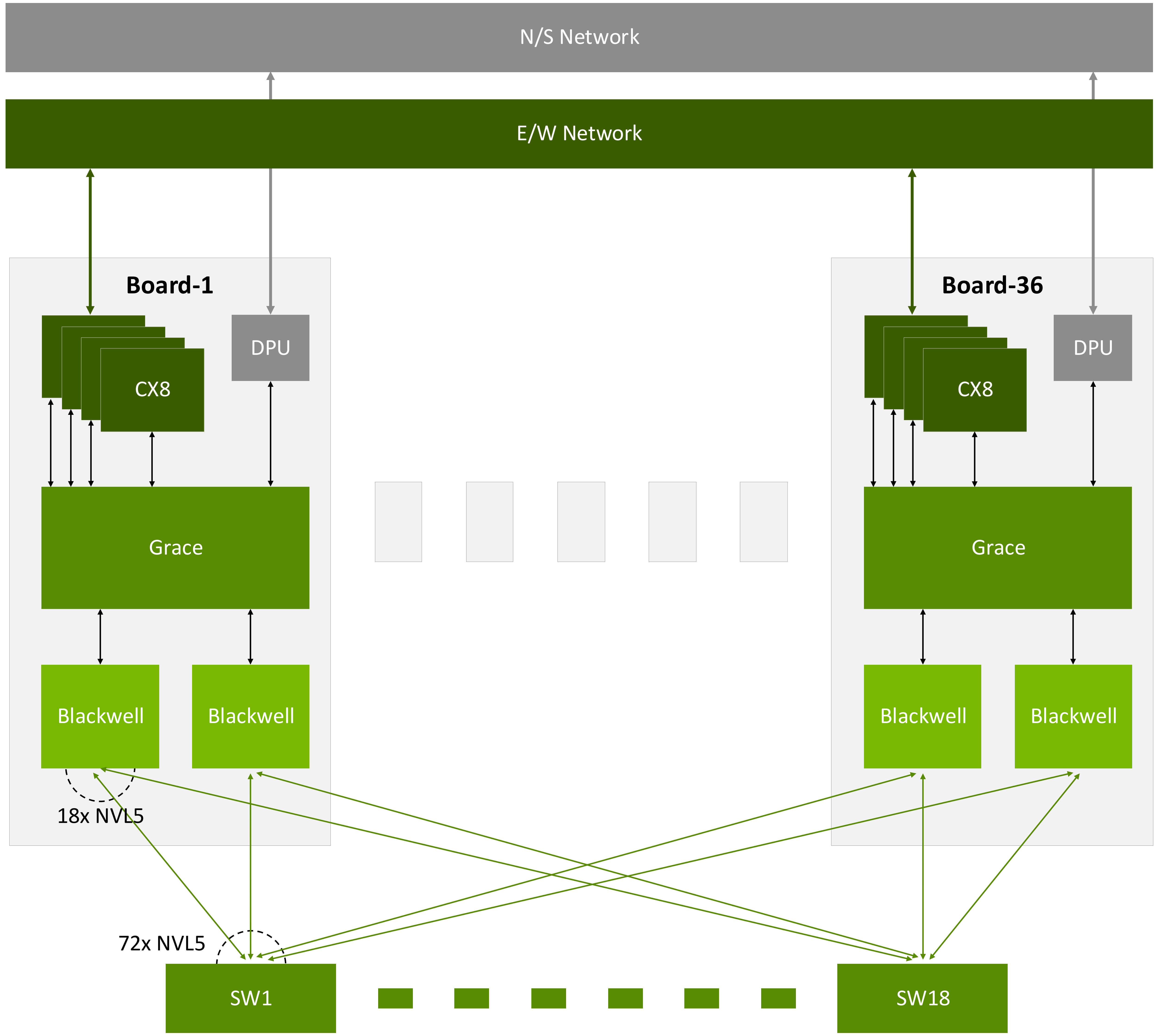
2x Switch ASICs per tray

7.2TB/s per switch

18x Compute Trays

4x GPUs/Tray

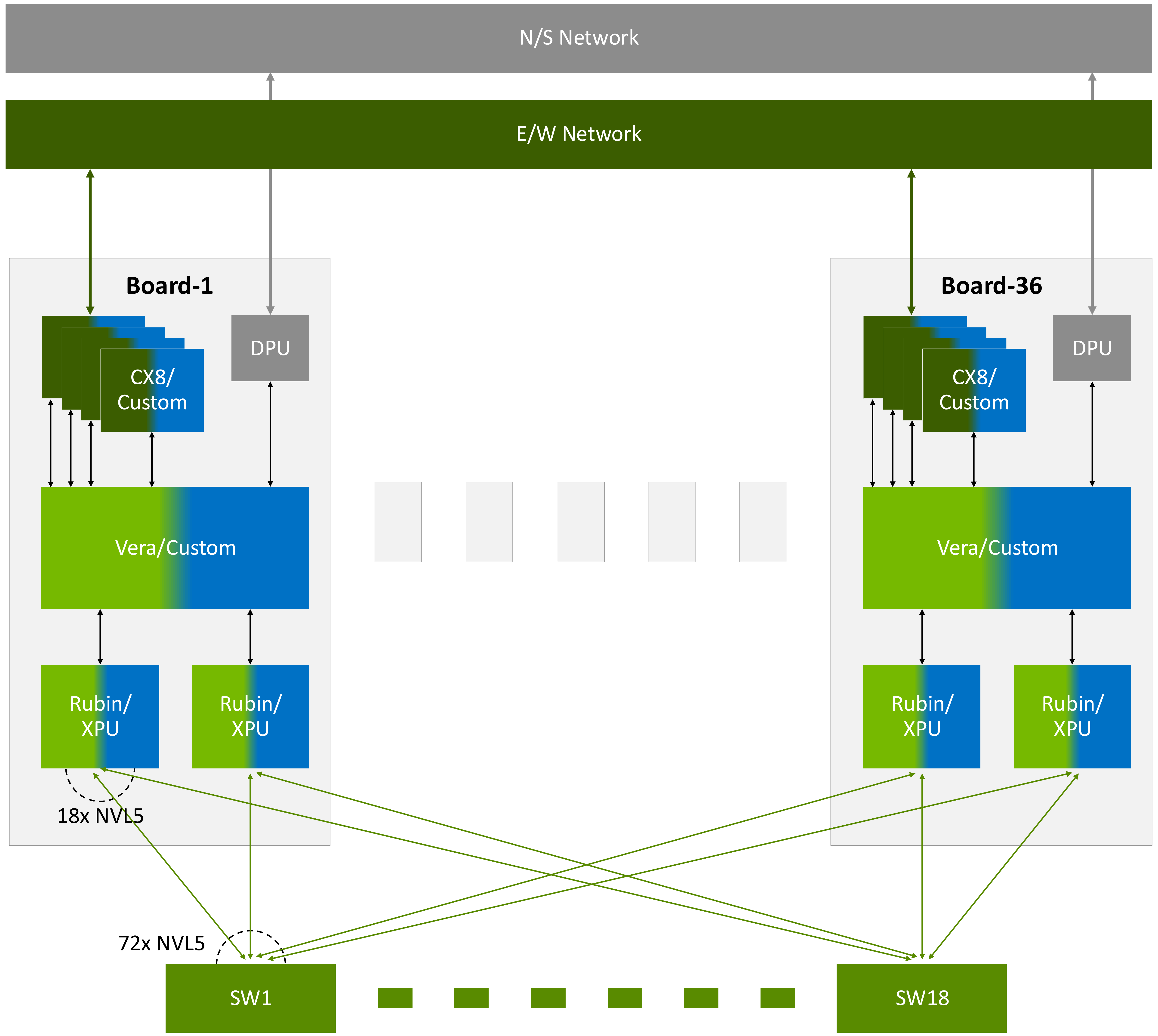
1.8 TB/s per GPU



NVLink Fusion System Integration

Complete Leverage from NVL72 Rack Architecture

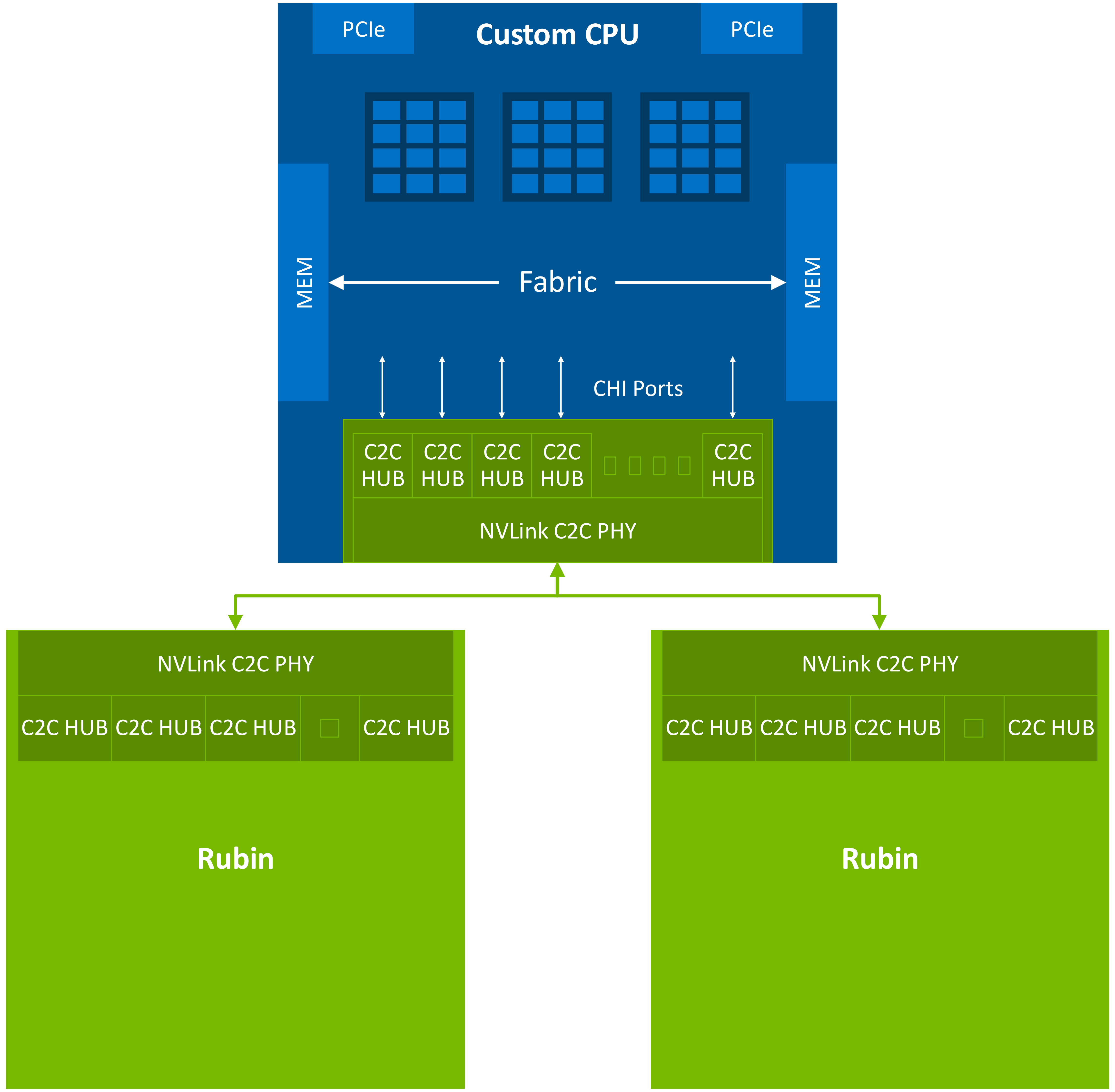
- NVLink Fusion CPU and XPU integrated into NVIDIA rack architecture, like NVIDIA Grace/Vera and NVIDIA GPUs
- NVLink C2C Integration Offered for optimal CPU connectivity
- Complete leverage from NVL72 Architecture
  - System, rack architecture
  - Software tools and libraries
  - Supply chain



NVLink Fusion

Custom CPU integration

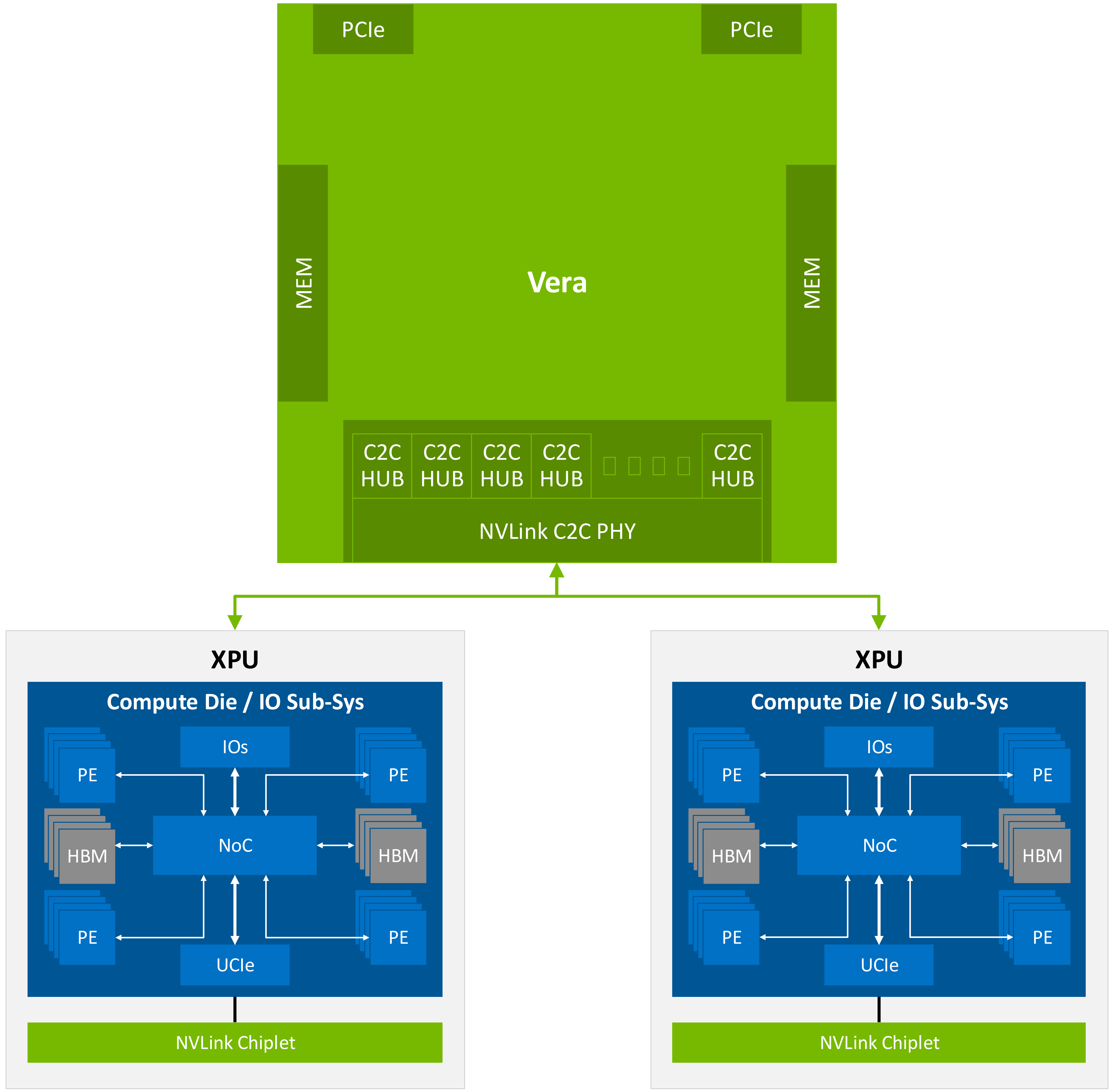
- CPU-GPU High Bandwidth Interface
  - > 100 lanes of PCIe GEN6
- Unified Memory Architecture
  - GPU and CPU can access all memory
- Heterogenous Memory Model Support
  - CPU follows CPU memory model
  - GPU follows GPU memory model
  - Full support for CPU or GPU defined atomics
- GPU Memory expansion
  - Native use of Host Memory from GPU
  - Supports all native operations to host memory
- Protocol support
  - Symmetric functionality for Host and Accelerator
  - Compatible with industry standard IP interfaces
- NVLink IP Integration
  - Soft IP + Optimized IO Implementation



NVLink Fusion

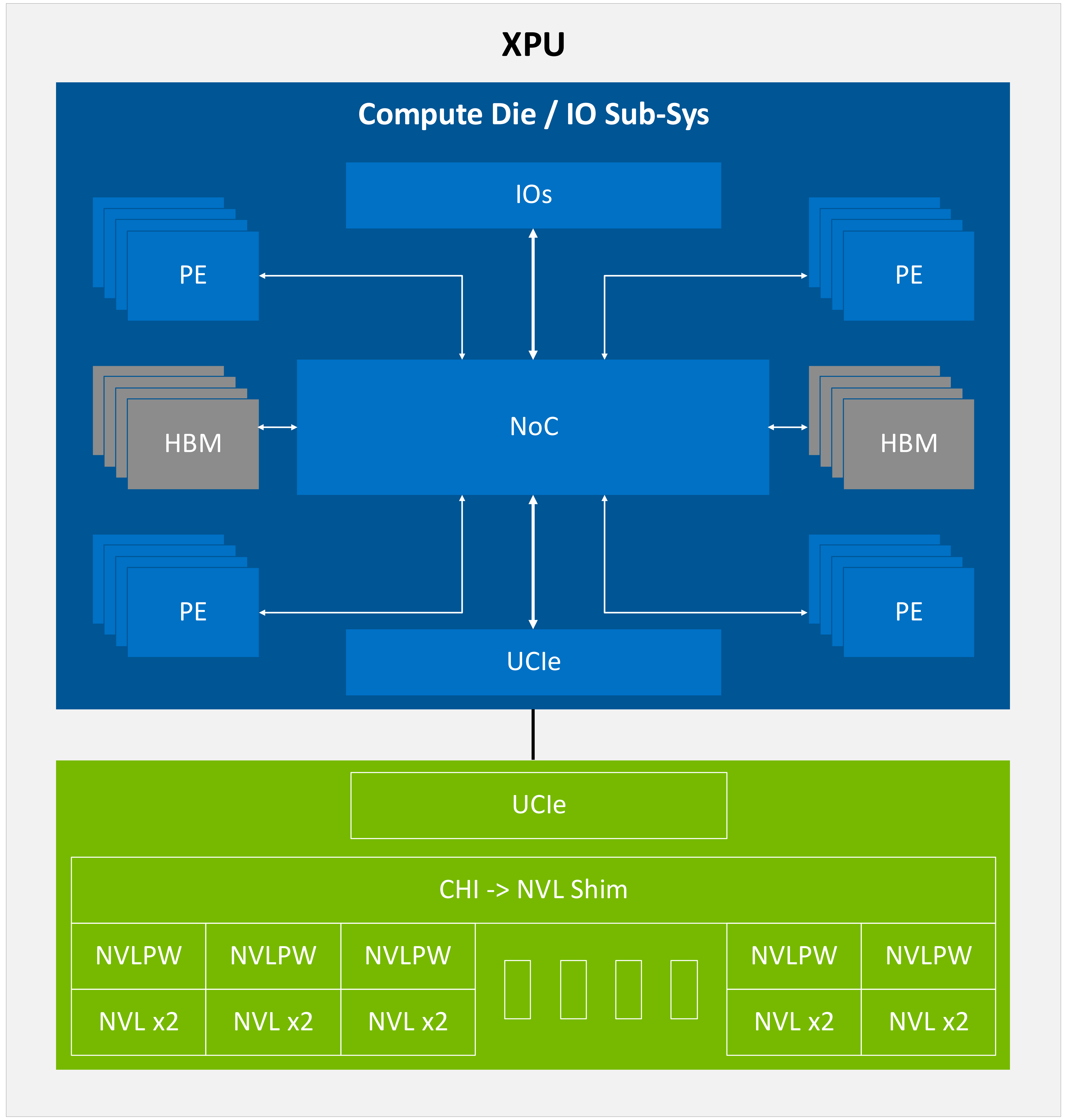
Custom XPU integration

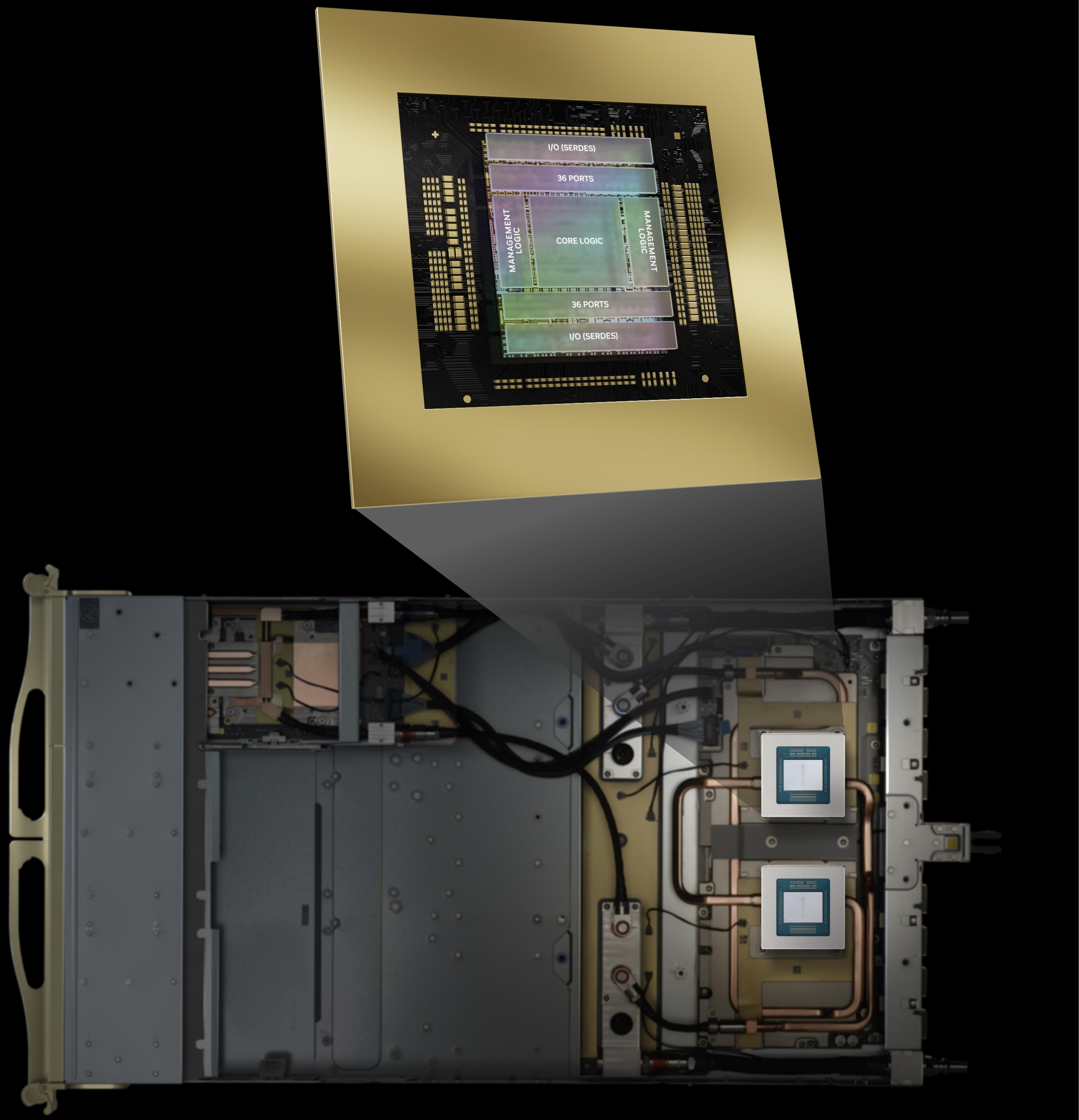
- NVLink Integration to XPU
  - Chiplet-based Integration
  - CHI-like protocol, UCIe Phy Layer
- Zero friction access to peer-XPU memory
  - Load/Store architecture
  - Access through DMA engines
- Ultra-High Bandwidth Architecture
  - Slim Network Layer
  - Low latency, area and power overhead
- In-Network Computing
  - Multicast, programmable reductions
- Fabric Management
  - NVLink configuration, decoding, mapping and routing
- NVLink C2C IP Integration
  - Optimal CPU:XPU interface option
- Custom CPU/NIC Support



NVLink Integration with XPU Compute

- NVLink chiplet encapsulates NVLink functionality
  - Exposes a standard interface to the XPU Compute
  - NVLink specifics contained within the chiplet
  - Link and Physical Layer for D2D compatible with standard UCIe specifications
- Protocol based is CHI-like, optimized for NVLink
  - Packetization leverages CHI C2C on UCIe
  - Support for scale-up operations
    - Peer-XPU Memory reads/writes
    - Collective operations for reductions
    - Atomic operations
- XPU has flexible choices on integration into XPU fabric
  - Peer XPU memory can be exposed only to DMA engines, or more deeply integrated into the PEs





NVLink Switch

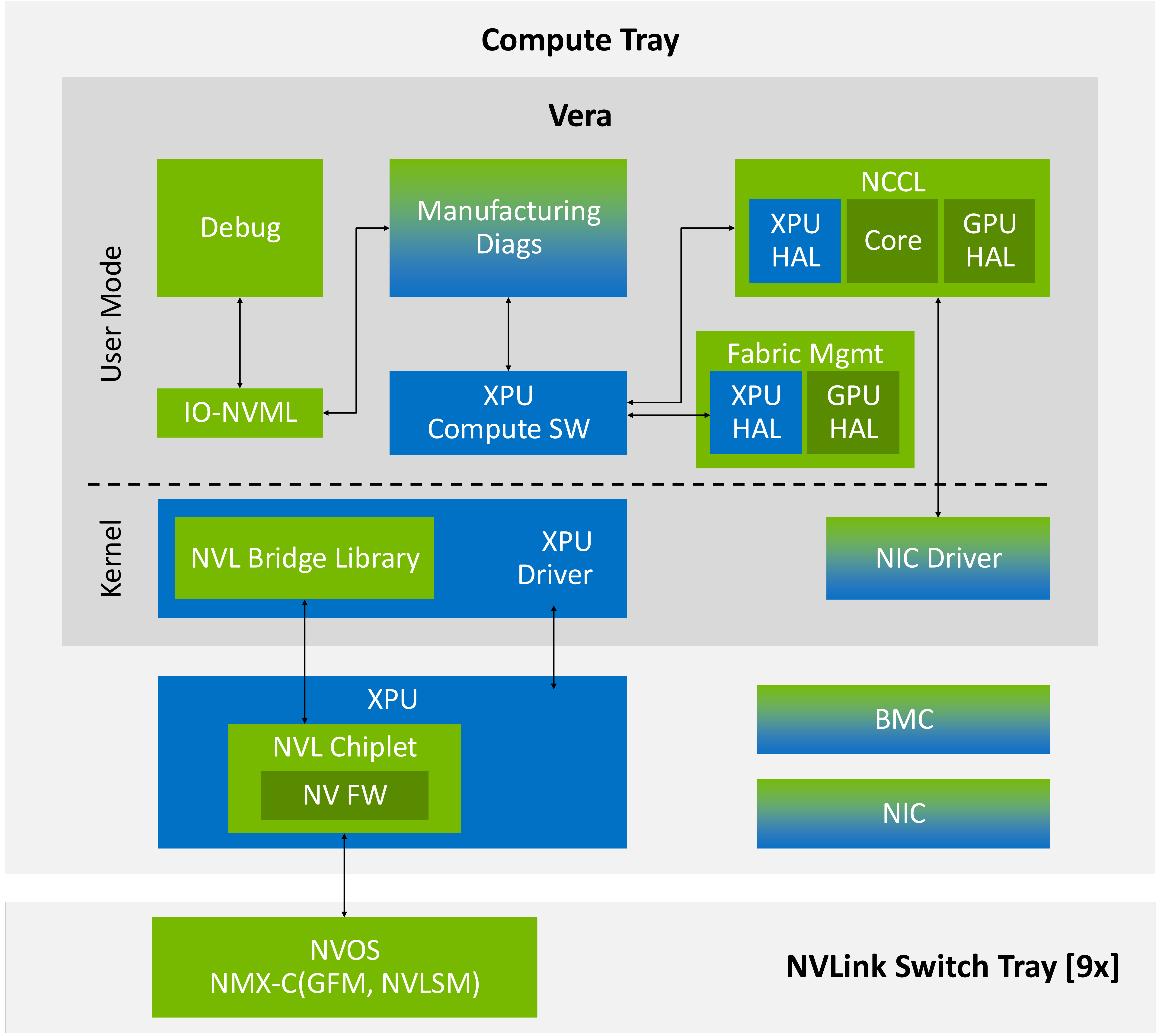
- NVLink5 Switch
  - Single monolithic die for minimal latency
  - 7.2 TB/s full all-to-all bidirectional BW over 72 ports
- SHARP$^{\text{TM}}$ * In-Network Compute
  - 3.6 TFLOPS of compute
  - Unicast, Multicast writes, Multicast reads with data reduction
  - Multiple type of operands – from 32bits to 8bits
  - Multiple reduction groups in parallel
- Fabric Partitioning Support
- NVLink5 Switch Tray
  - 2x NVLink5 Switch chips
  - 14.4 TB/s total bandwidth

** SHARP (Scalable Hierarchical Aggregation and Reduction Protocol)*

NVLink Fusion Software

NVIDIA's best in class communication libraries, telemetry and debug tools

- NCCL
  - Best in class algorithms for low latency and high bandwidth
  - Topology graph search for finding optimal data paths
  - Integration with PyTorch, VLLM, SGLang, and many other frameworks
  - Ten years of performance tuning and production operation
- Fabric and Memory Management
  - Address space management APIs
  - Extendable for XPU-specific memory semantics
  - Routing and forwarding setup
- User Mode Tools
  - NVLink telemetry and debug
- Diagnostics
  - Manufacturing diagnostic capabilities for compute tray and rack
  - Fabric Testing at rack scale



NVLink Fusion Ecosystem

Custom Silicon Partners

alchip AsteraLabs GUC MEDIATEK MARVELL

CPU Partners

FUJITSU Qualcomm

Technology Partners

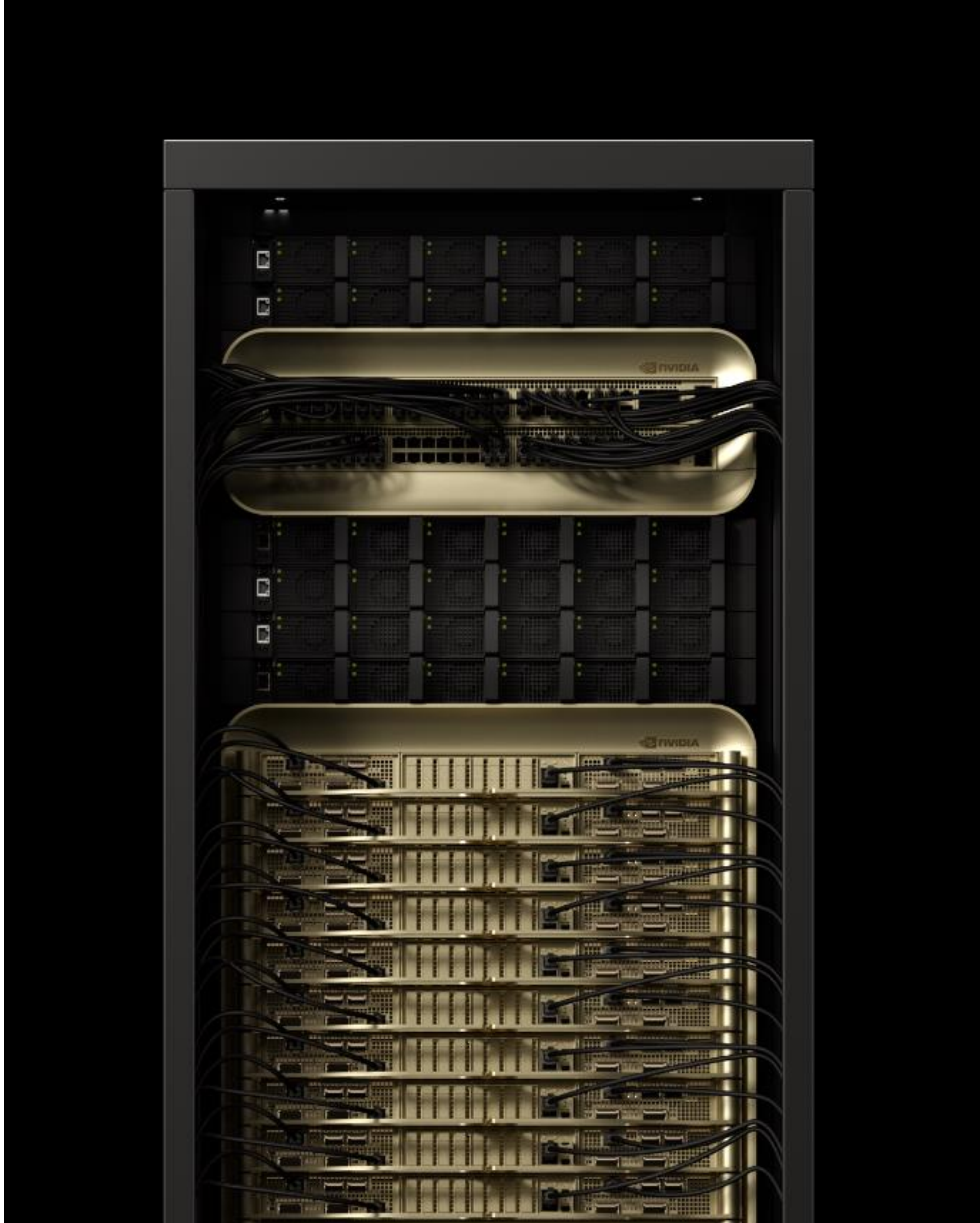
cadence SYNOPSYS

System Partners

AIVRES ASUS DELL Technologies GIGABYTE

Hewlett Packard Enterprise ingrasys Inventec Inventec Data Center Solutions Lenovo PEGATRON

QCT SUPERMICRO wistron wiwynn



NVLink Fusion Ecosystem

Rich eco-system of partners

MGX Rack

Auras, DELTA, ingrasys, Interplex, KARRIE, LEAD WEALTH, legrand, LITEON, OuRack, RITTAL, Schneider Electric, YS YUANS

Rack Manifold

Auras, AVC, COOLER MASTER, CoolIT systems, DELTA, ingrasys, LEAD WEALTH, LITEON, nVent, P.D.M, READORE TECHNOLOGY

CDU

AIVRES, AVC, BEEHE, BOYD, COOLER MASTER, CoolIT systems, DELTA, Envicool, ingrasys, LEAD WEALTH, LITEON, motivair, Nidec, nVent, QCT, Schneider Electric, VERTIV

UQD and MQD

Auras, AVC, BEEHE, EJN, CPC, C.TECH, Danfoss, Envicool, ingrasys, LEAD WEALTH, LOTES, Nidec, Parker, READORE TECHNOLOGY, refas s.r.l., STÄUBLI

Cold Plate

Auras, AVC, BOYD, COOLER MASTER, CoolIT systems, C.TECH, DELTA, ingrasys, LEAD WEALTH, LITEON, Nidec, P.D.M, READORE TECHNOLOGY

Slide Rail

AVC, King Slide, REPON, YS YUANS

Power Shelf

DELTA, flex., LEAD WEALTH, LITEON, MEGMEET

12V Busbar

Amphenol, BizLink, Interplex, JPC connectivity, LOTES, molex

1400A Busbar

Amphenol, BizLink, DELTA, ingrasys, Interplex, LEAD WEALTH, TE connectivity

Powerwhip

Amphenol, BizLink, JPC connectivity, TE connectivity

Fan

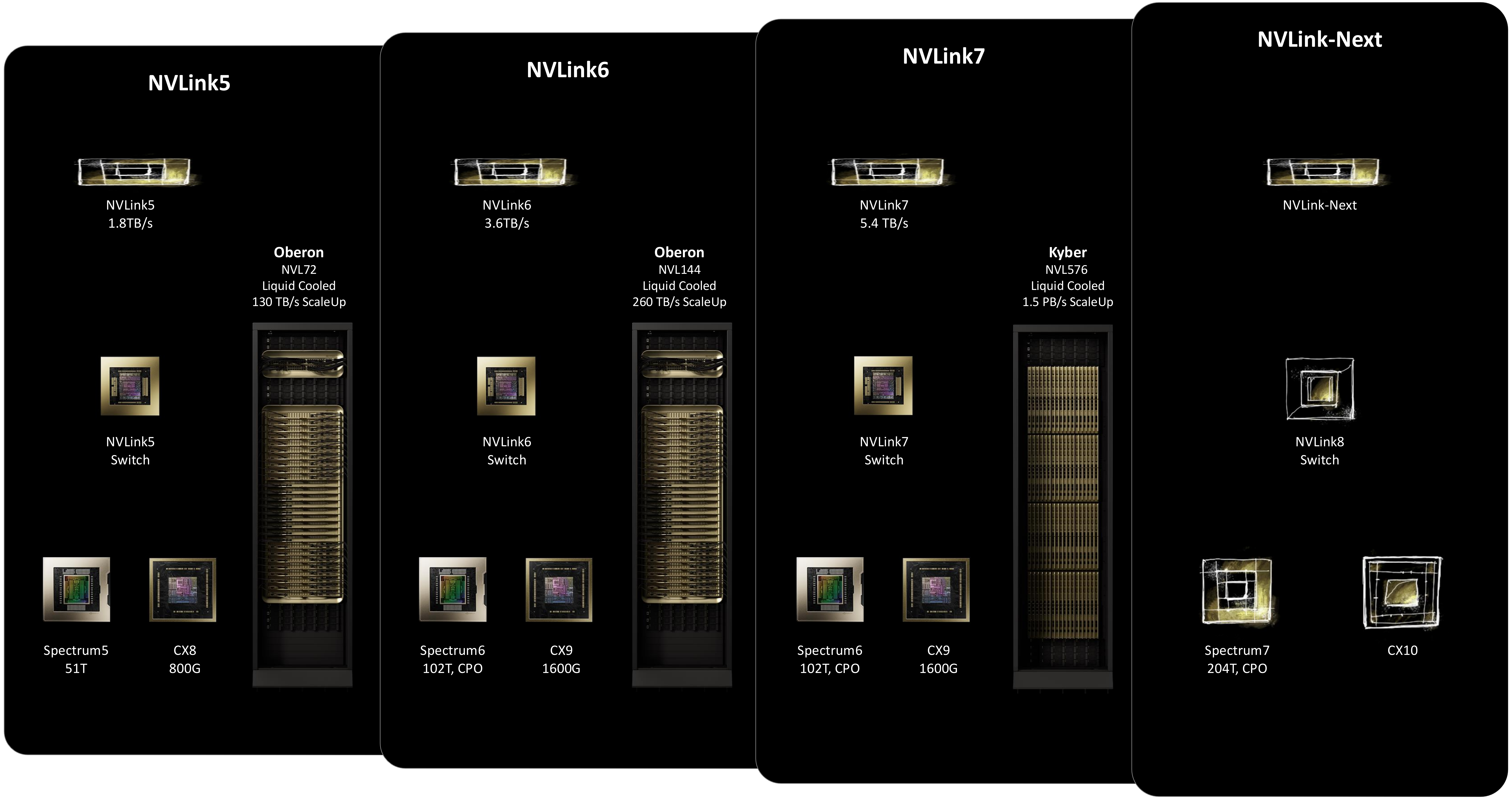
AVC, DELTA, ingrasys, Nidec, SANYO DENKI, SUNON

Chassis

AVC, CHENBRO, ingrasys, Interplex, KARRIE

NVLink Fusion System Roadmap

One-Year Rhythm | Full-System | One Architecture



	NVLink5	NVLink6	NVLink7	NVLink-Next
Semi Custom NVLink Chiplet	NVLink5 1.8TB/s	NVLink6 3.6TB/s	NVLink7 5.4 TB/s	NVLink-Next
	Oberon NVL72 Liquid Cooled 130 TB/s ScaleUp	**Oberon** NVL144 Liquid Cooled 260 TB/s ScaleUp	**Kyber** NVL576 Liquid Cooled 1.5 PB/s ScaleUp	
Networking (Scale-up)	NVLink5 Switch	NVLink6 Switch	NVLink7 Switch	NVLink8 Switch
Networking (Scale-out)	Spectrum5 51T CX8 800G	Spectrum6 102T, CPO CX9 1600G	Spectrum6 102T, CPO CX9 1600G	Spectrum7 204T, CPO CX10