大模型系列 – AI 集群

NV GB200 网络演进细节

NVIDIA ZOMI

本节内容

1. Blackwell 架构芯片信息
2. GH200 & GB200 产品信息
3. HGX H/B 系列产品信息

01 NV带宽基本分析

带宽计算

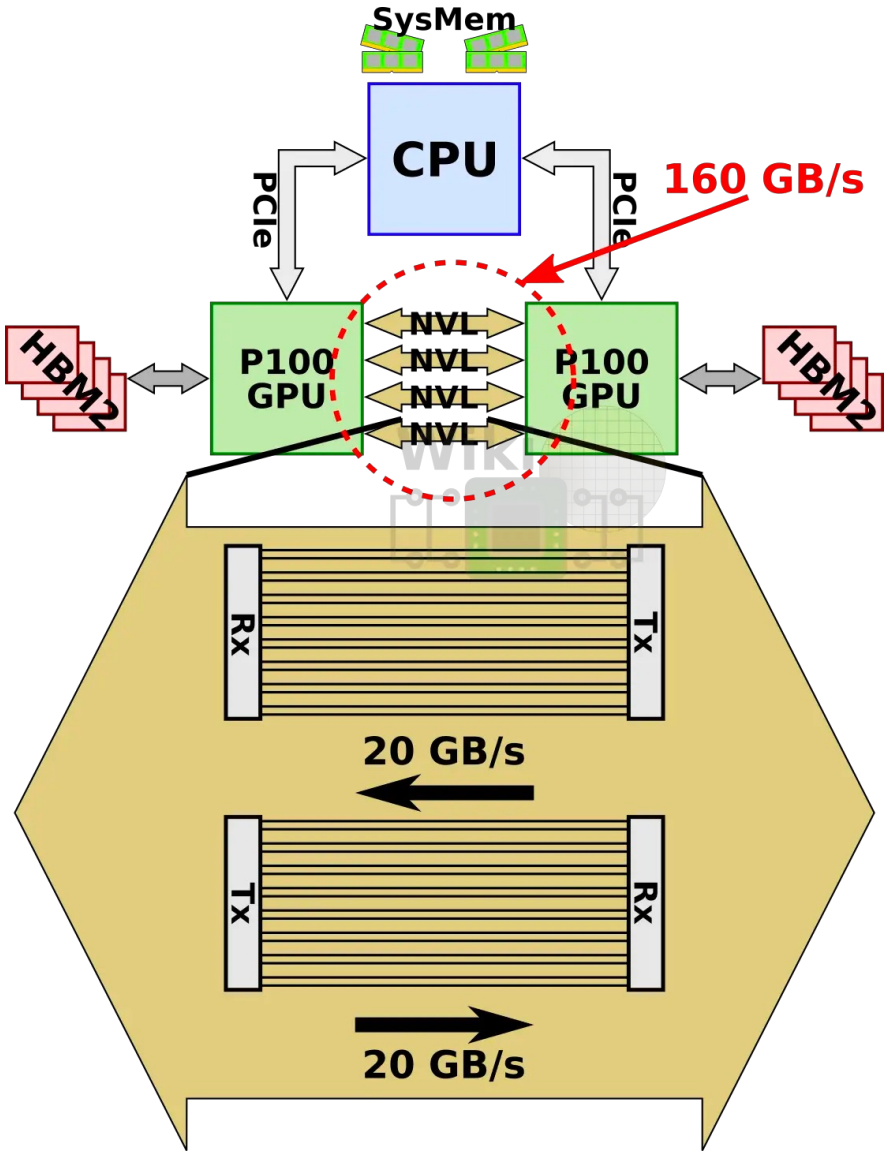
- NVLink 传输带宽计算 B200 NVLink 5th 带宽为 1.8TB/s，这是针对计算带宽，按照内存带宽算法以字节每秒（ Byte/s ）为单位。
- NVLink Switch 或者 IB/Ethernet 交换机和网卡上，以网络带宽来计算，按照传输数据位以比特每秒（bit/s）为单位。

Byte/s	bit/s
GB/s	Gb/s

第一代 NVLink P100 DEMO

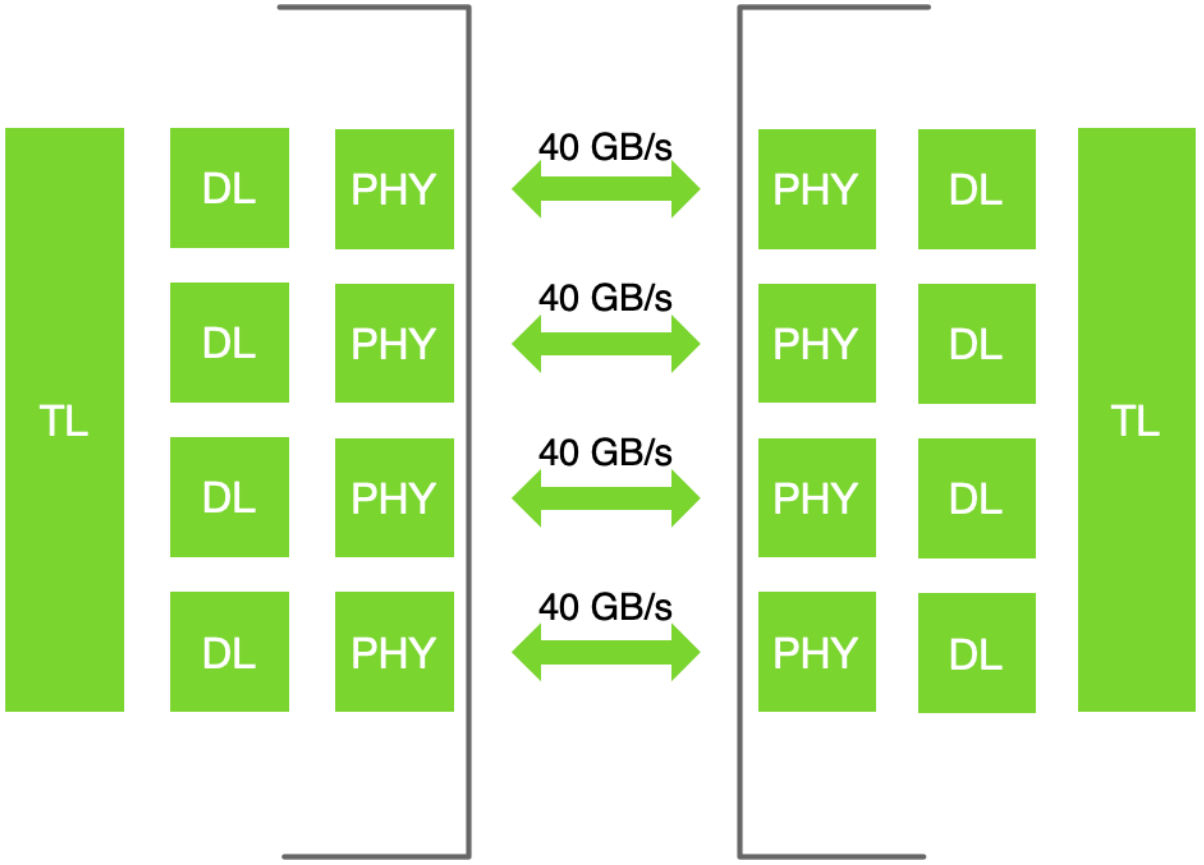
- 单条 NVLink 是一种双工双路信道，其通过组合 32 条配线，从而在每个方向上可以产生8对不同的配对（2bi x 8pair x 2wire = 32wire）

- P100上，集成了4条nvlink。每条link具备双路共40GB/s的带宽，整个芯片具备整整160GB/s的带宽。



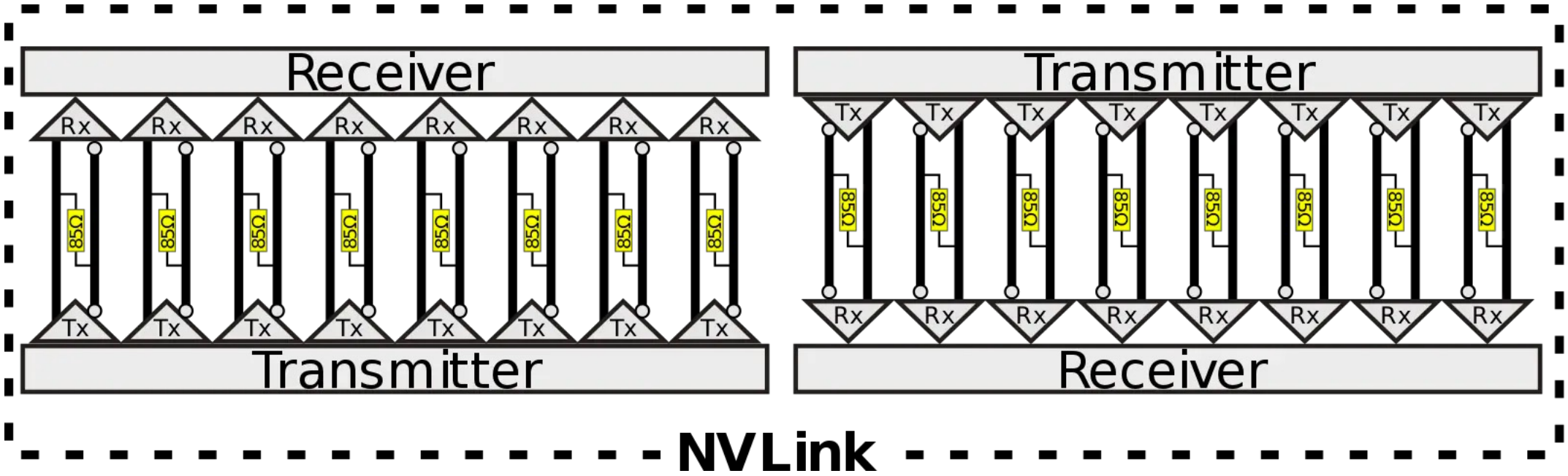
NVLink 连接

- P100 supports 4 NVLinks
- Up to 94% bandwidth efficiency
- Supports read/writes/atomics to peer GPU
- Supports read/write to NVLink-enabled CPU
- Links can be ganged for higher bandwidth



NVLink 连接

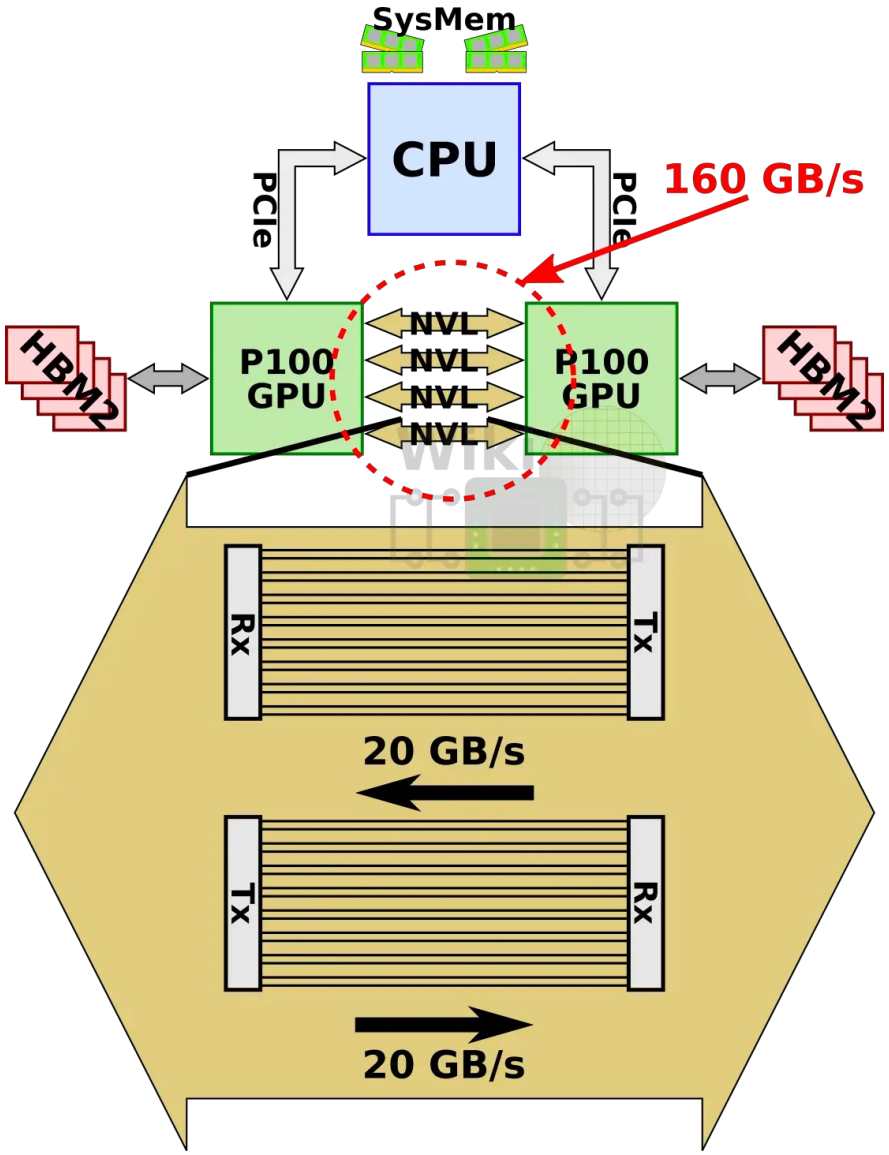
- NVLink 通道称为Brick。单个 NVLink 是一个双向接口（Port），每个方向包含 8 个差分对，总共 32 条线。电气连接线使用直流耦合 DC coupled，带 85 欧姆差分终端。



NVLink带宽计算

- 4 对差分信号线同时包含接收和发送方向信号线，在计算网络带宽时，400Gbps Ports 指同时能够收发 400Gbps 数据

- 4 对差分信号线构成 RX/TX 各两对，从网络视角来看是一个单向 400Gbps 链路，而从内存带宽视角是支持100GB/s 访存带宽



01 NVSwitch & NVLink

第 5 代 NVLink

- 第 5 代 NVLink 每个 GPU 上有 18 个 NVLink，单 Link 双向带宽从 H100 NVLink 4th 的 50GB/s 升级到 100GB/s。所B100 & B200 GPU-to-GPU 带宽上限为 1.8TB/s（18 x 100GB/s）。

	Second Generation	Third Generation	Fourth Generation	Fifth Generation
NVLink bandwidth per GPU	300GB/s	600GB/s	900GB/s	1,800GB/s
Maximum Number of Links per GPU	6	12	18	18
Supported NVIDIA Architectures	NVIDIA Volta™ architecture	NVIDIA Ampere architecture	NVIDIA Hopper™ architecture	NVIDIA Blackwell architectur

NVLink 5th 带宽计算

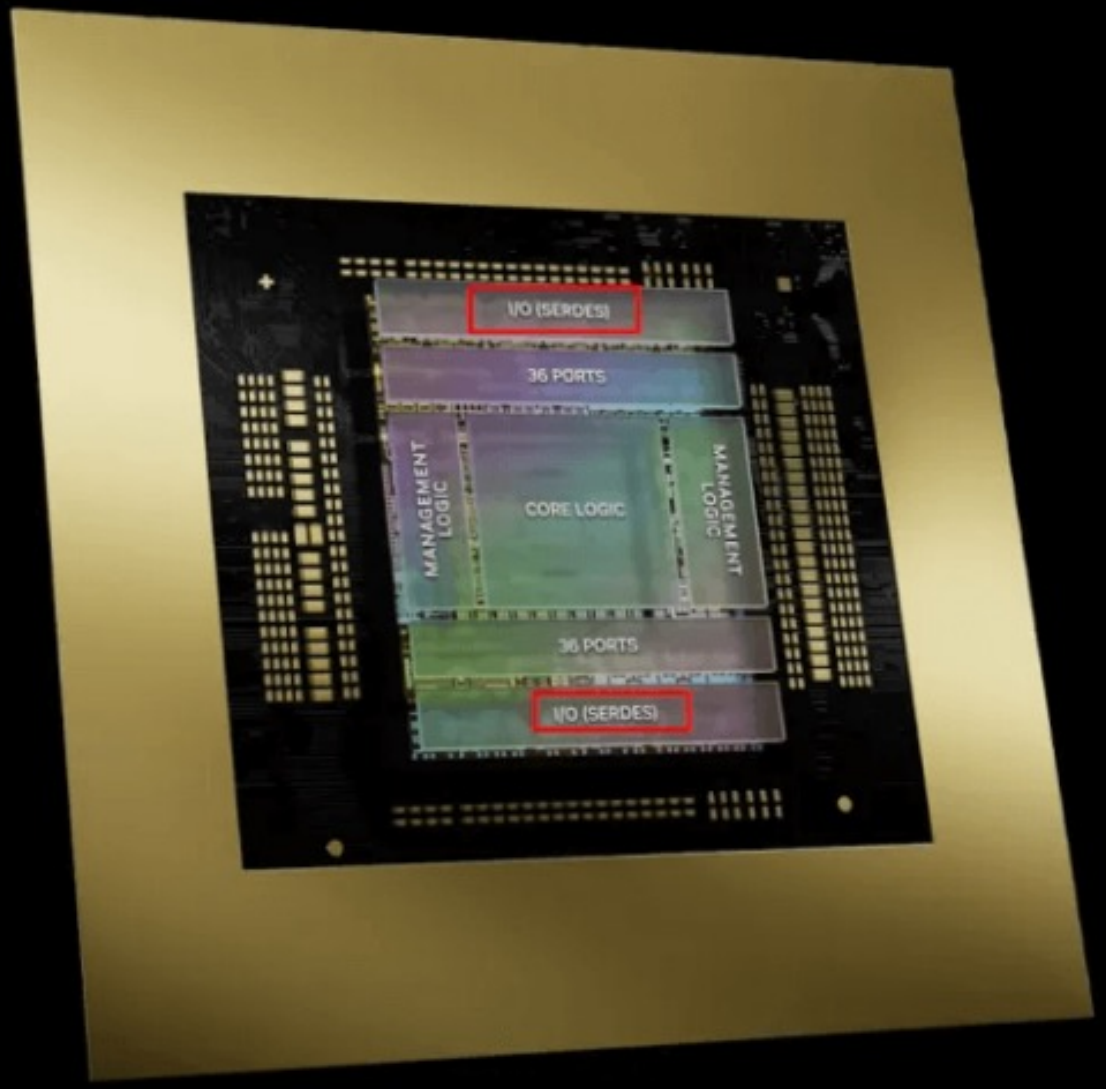
- B200 NVLink 带宽为1.8TB/s，由 18 个 Port 构成，即每个 Port 100GB/s，由四对差分线构成，每个 Port 包含两组 224Gbps 的 SerDes（2x224G-PAM4 按照网络接口算为每端口单向 400Gbps 带宽）

第 4 代 NVSwitch

- 第 4 代 NVSwitch GPU-to-GPU 带宽扩大一倍，变为 1.8TB/s，最多可以支持 576 个 GPU，总带宽上限为 576 x 1.8TB/s=1PB/s

	First Generation	Second Generation	Third Generation	NVLink Switch
Number of GPUs with direct connection within a NVLink domain	Up to 8	Up to 8	Up to 8	Up to 576
NVSwitch GPU-to-GPU bandwidth	300GB/s	600GB/s	900GB/s	1,800GB/s
Total aggregate bandwidth	2.4TB/s	4.8TB/s	7.2TB/s	1PB/s
Supported NVIDIA architectures	NVIDIA Volta™ architecture	NVIDIA Ampere architecture	NVIDIA Hopper™ architecture	NVIDIA Blackwell architecture

NVLink Switch Chip



NVLink Switch Chip

50B Transistors in TSMC 4NP

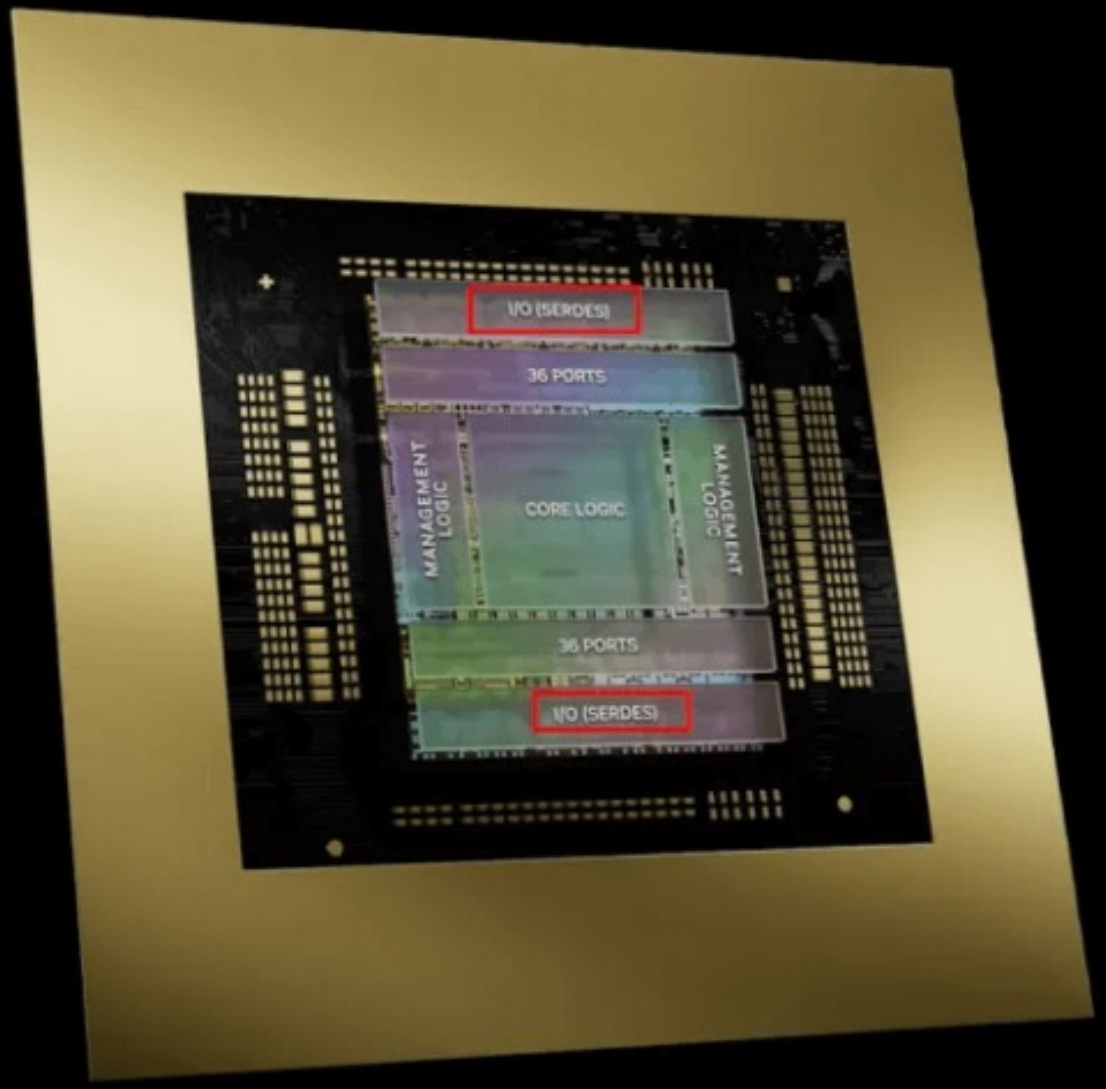
72-Ports Dual 200 Gb/sec SerDes

4 NVLinks at 1.8TB/sec

7.2TB/sec Full-Duplex Bandwidth

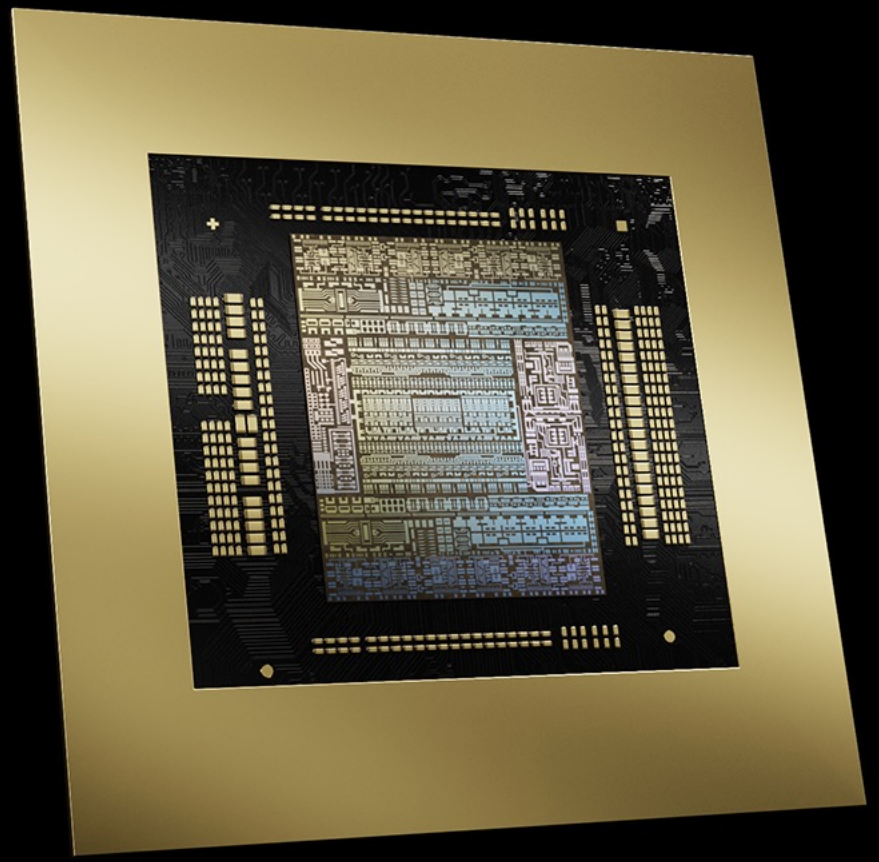
SHARP In-Network Compute – 3.6 TFLOPS FP8

NVLink Switch Chip



- NVSwitch Chip 上有 72 个 NVLink Port，每个 Port 2 个 lane，双向带宽为 2 x 2 x 200 Gb/s = 100GB/s
- 72 个 Port 对应 7.2TB/s，1.8TB/s NVLink 对应 18 个 Port。
- 传输 20 GB 仅需 22 毫秒，对于大模型一层 Transformers 的反向聚合完全够用

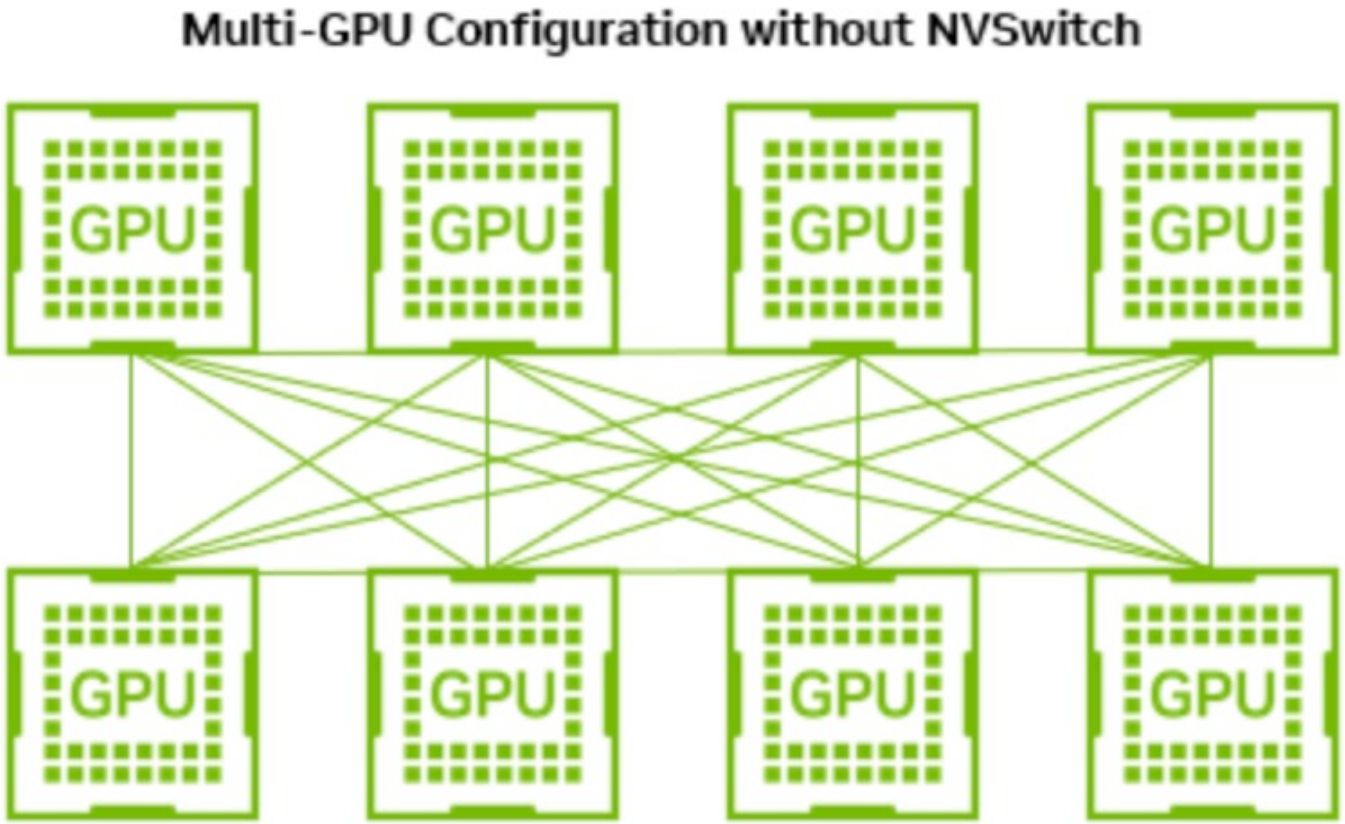
NVSwitch Tray

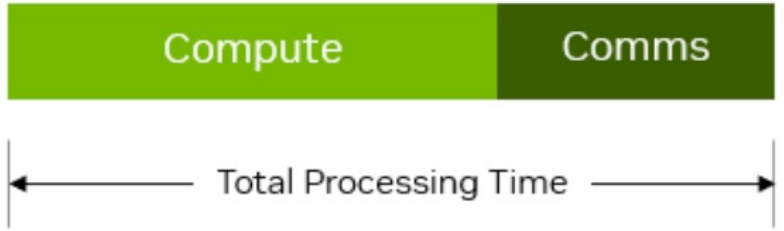



NVIDIA

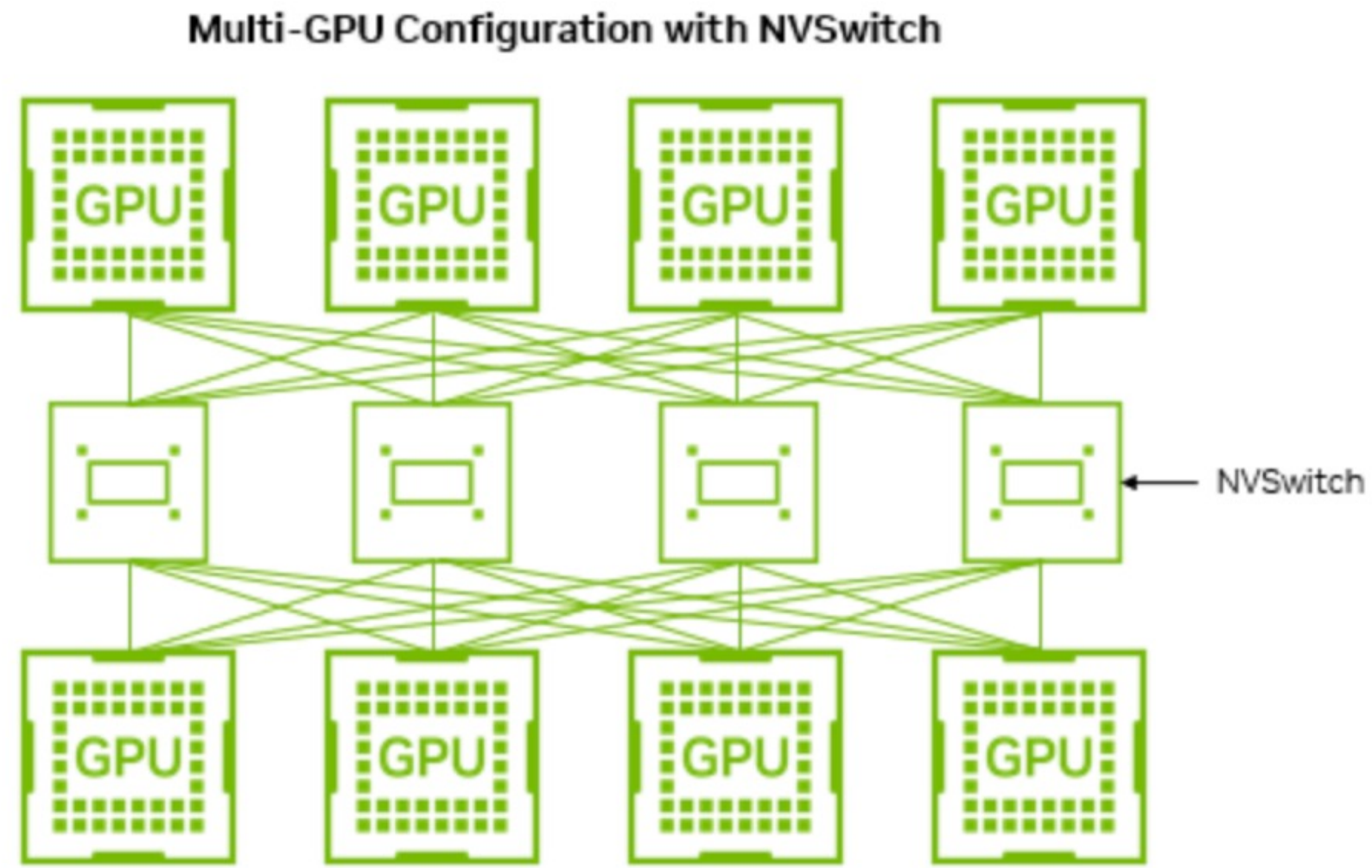

NVIDIA

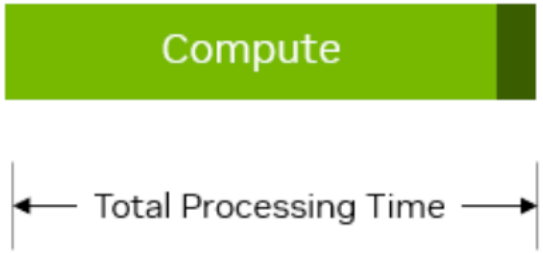
NVSwitch 作用


Multi-GPU Configuration without NVSwitch
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU


Compute
Comms
Total Processing Time

NVSwitch 作用


Multi-GPU Configuration with NVSwitch
GPU
GPU
GPU
GPU
NVSwitch
GPU
GPU
GPU
GPU


Compute
Total Processing Time

NVSwitch 作用

- 使用 Llama3.1 70B 来测试在 30~50 tokens/s/user 的网络吞吐带宽，其中 ISL/OSL = 8k/256

Real-time Response Budget tok/s/user	Throughput tok/s/GPU (batch size)			NVSwitch Benefit
	Single GPU TP=1	Point-to-Point TP=2	NVSwitch TP=2	
30	67 (2)	80 (6)	115 (9)	**1.4x**
35	Does Not Meet	74 (5)	104 (7)	**1.4x**
40	Does Not Meet	67 (4)	87 (5)	**1.3x**
45	Does Not Meet	56 (3)	76 (4)	**1.4x**
50	Does Not Meet	43 (2)	63 (3)	**1.5x**

NVSwitch 作用

- 固定批处理大小下的整体服务器吞吐量，更大的批处理大小意味着可以一次处理越来越多用户的请求，从而提高整体服务器利用率并降低每次推理的成本

Batch Size	Throughput tok/s/GPU		NVSwitch Benefit
	Point-to-Point	NVSwitch	
1	25	26	**1.0x**
2	44	47	**1.1x**
4	66	76	**1.2x**
8	87	110	**1.3x**
16	103	142	**1.4x**
32	112	168	**1.5x**

02

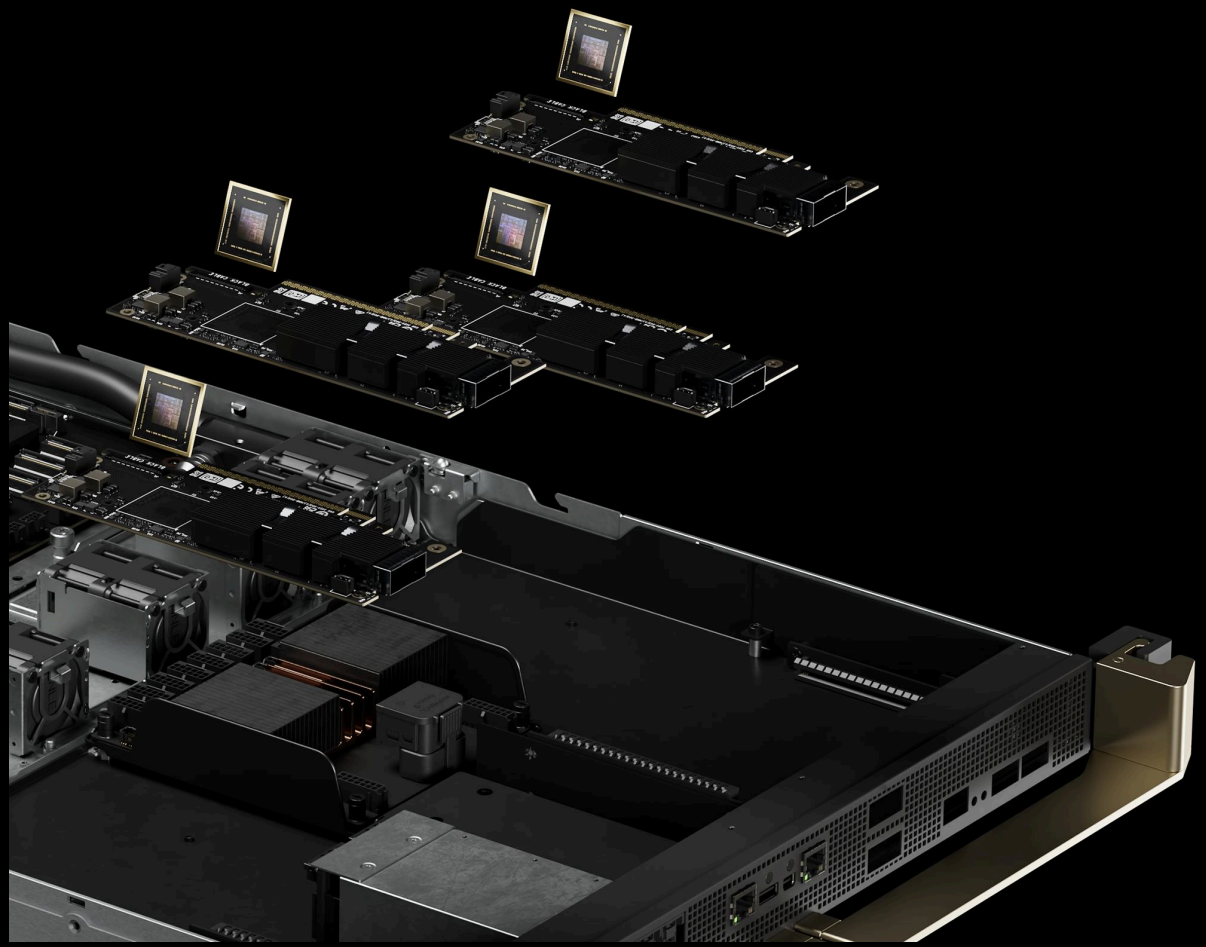
网卡&交换机

ConnectX-8 网卡

- DGX B100/B200 发布新一代 IB 网卡 ConnectX-8，相应通信带宽为 800Gb/s
- H100/H200 采用的为 ConnectX-7 网卡，对应的通信带宽为 400Gb/s
- A100 采用的 ConnectX-6 网卡的通信带宽为 200Gb/s

- PS： HGX B100/B200 依旧使用的上一代 ConnectX-7

ConnectX-8 网卡



ANNOUNCING

NVIDIA BLUEFIELD-3

400 Gbps Data Center Infra Processor

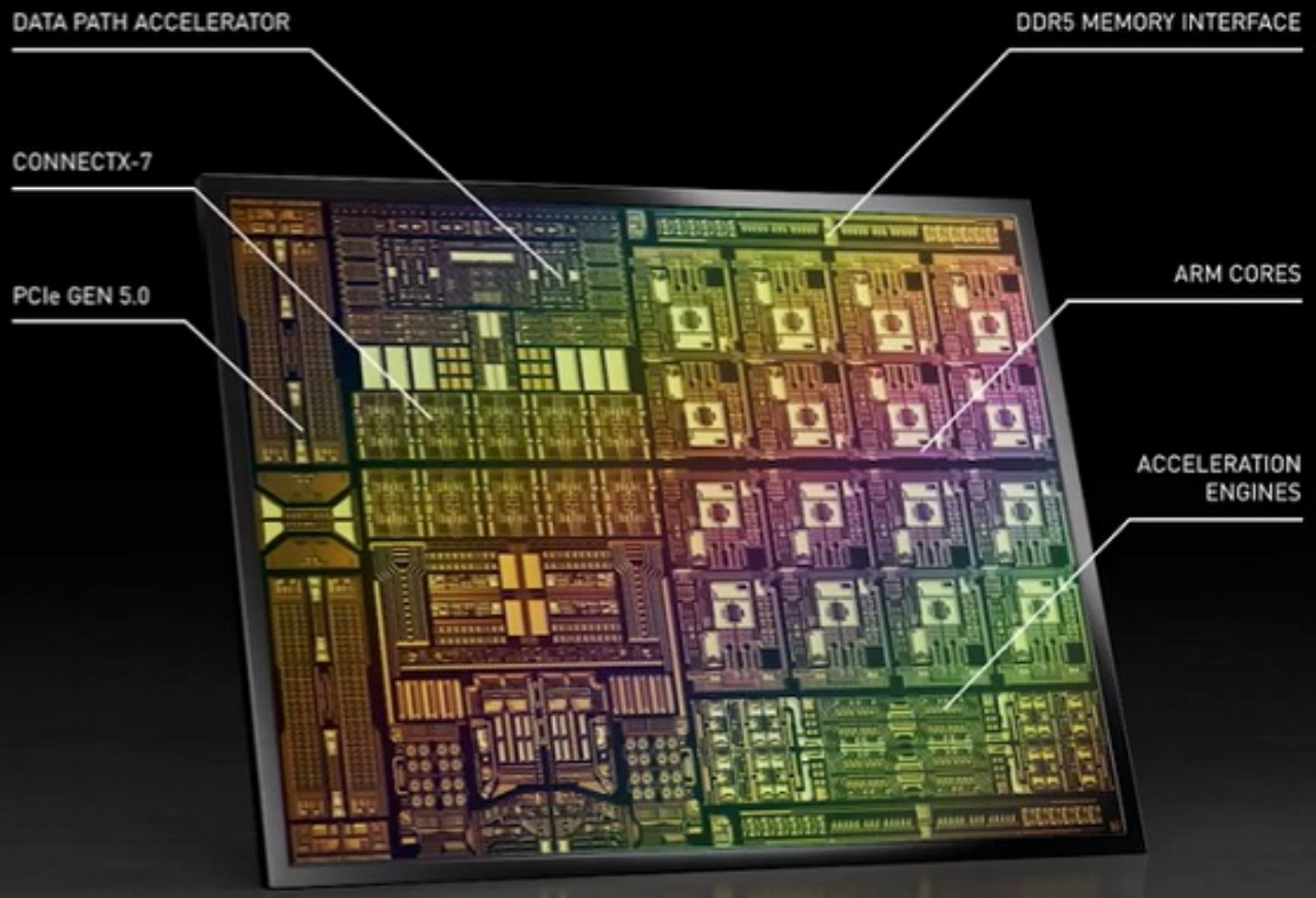
Offloads and Accelerates Data Center Infrastructure

Isolates Application from Control and Management Plane

Powerful CPU – 16x Arm A78 Cores

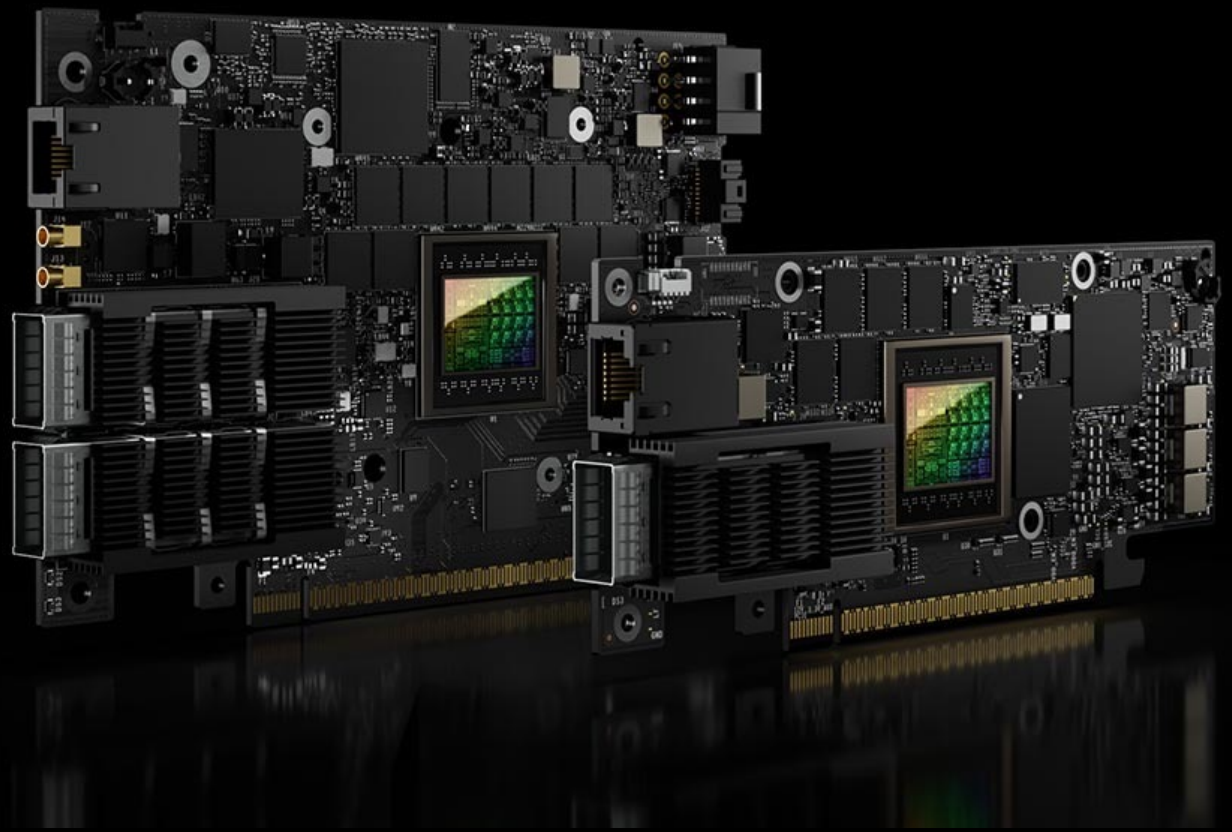
Process Networking, Storage, and Security at 400 Gbps

22 Billion Transistors



BlueField-3 DPU/SuperNIC

- BlueField-3 支持以太网和 IB，速度高达 400Gb/s，与网络、存储等结合
- BlueField-3 SuperNIC 可以在 GPU 节点间通过以太网直连
- 支持单 Port 400Gb/s，或双 Port 各 200 Gb/s



Quantum-X800 IB 交换机

Spectrum-X800 以太网交换机

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem

Reference 参考&引用

1. https://www.fibermall.com/blog/nvidia-b100-b200-gh200-nvl72-superpod.htm