大模型系列 – AI 集群

深度解读 Blackwell

NVIDIA ZOMI

NVIDIA.

本节内容

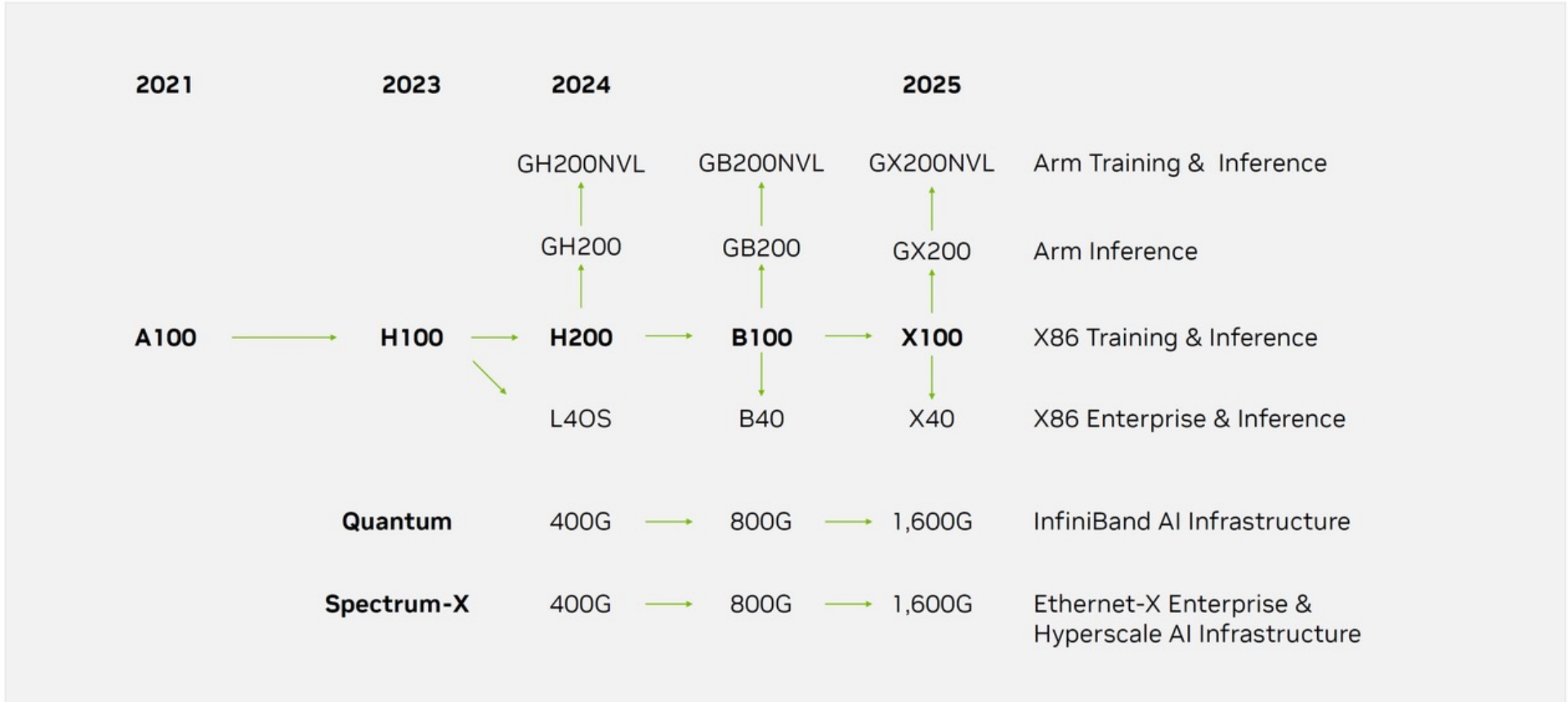
1. Blackwell 架构芯片信息
2. GH200 & GB200 产品信息
3. HGX H/B 系列产品信息

01

GPU产品介绍

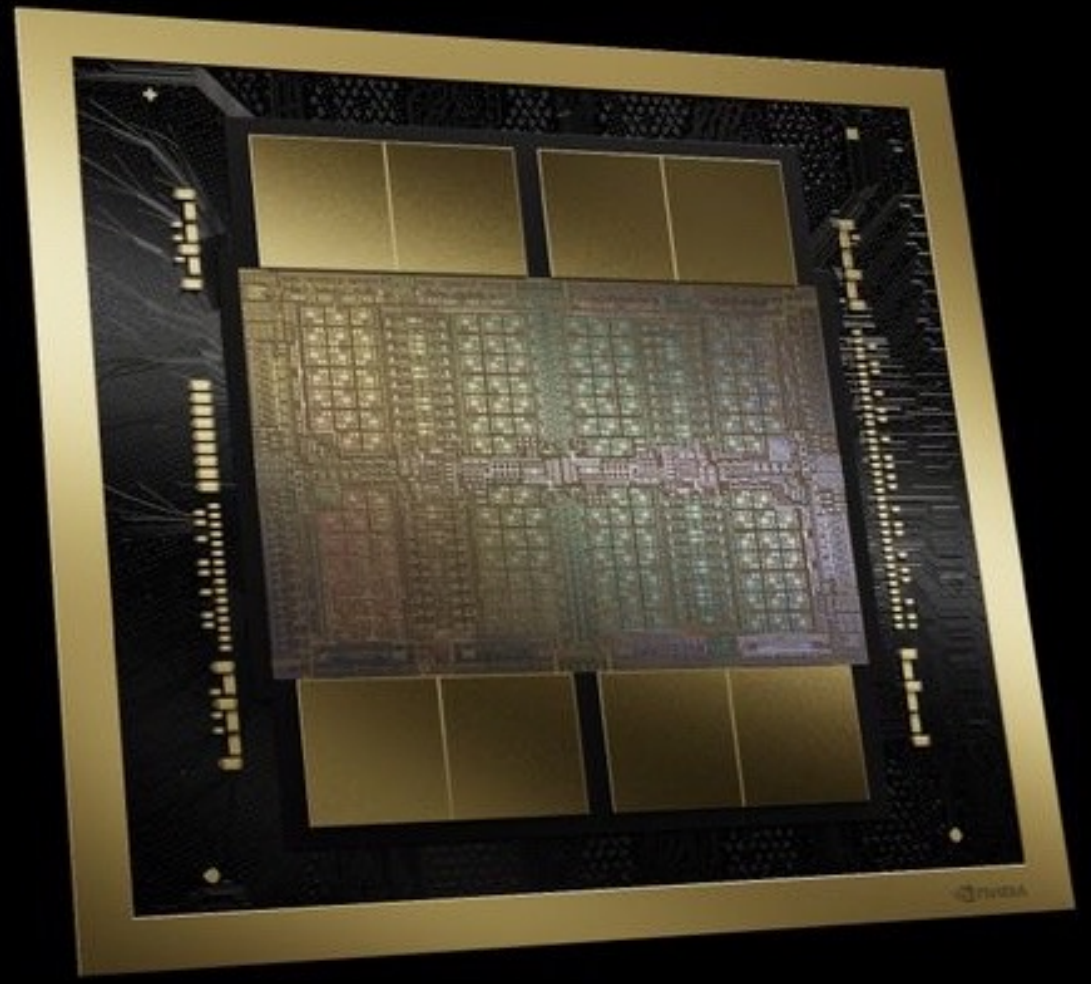
NVIDIA AI – One Architecture | Train and Deploy Everywhere

From Two-Year Rhythm
to
One-Year Rhythm | Training & Inference | x86 & Arm | Hyperscale & Enterprise



Die1

Die2



HV-HBI (10TB/s)

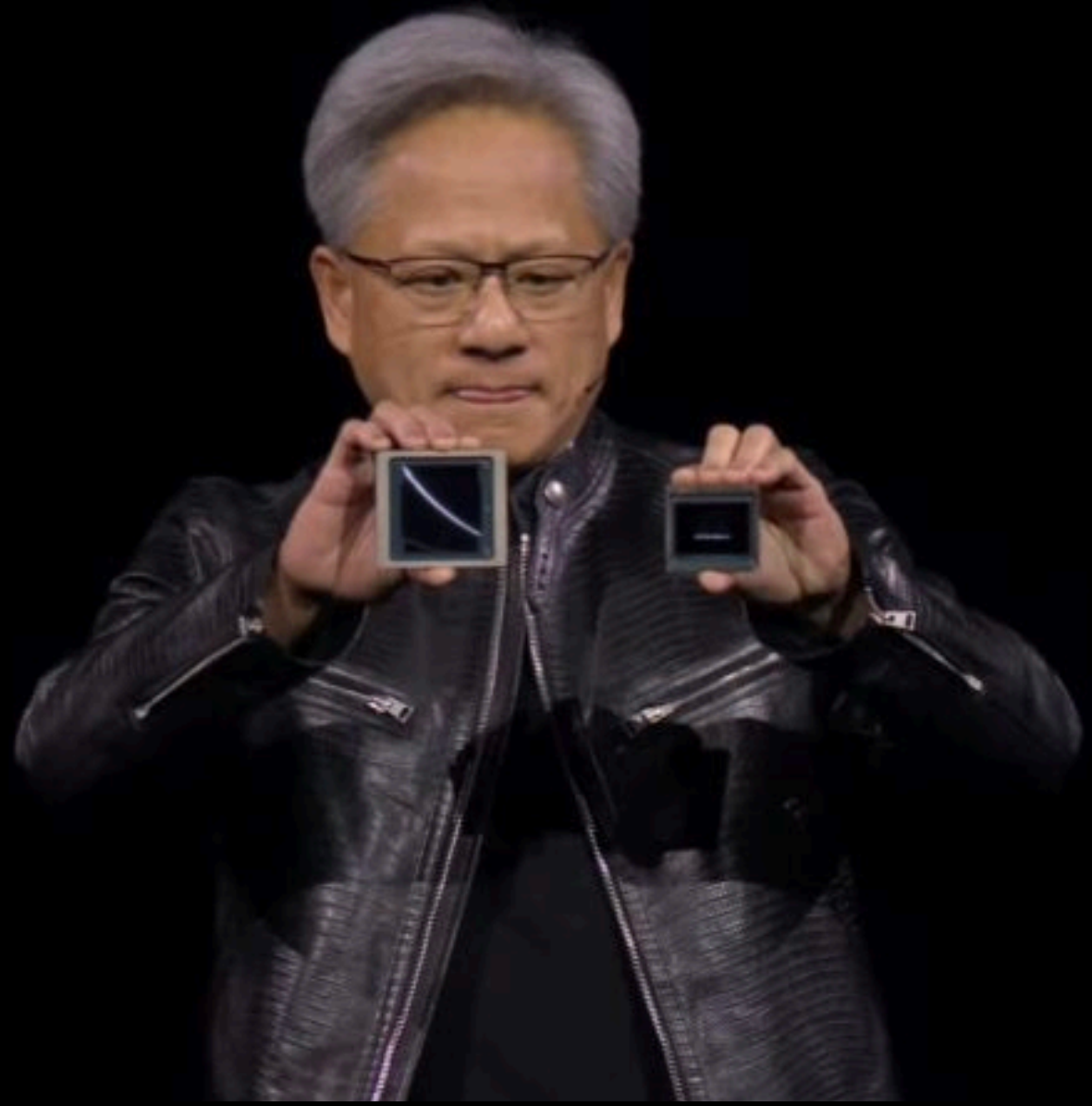
HBM (8TB/s)

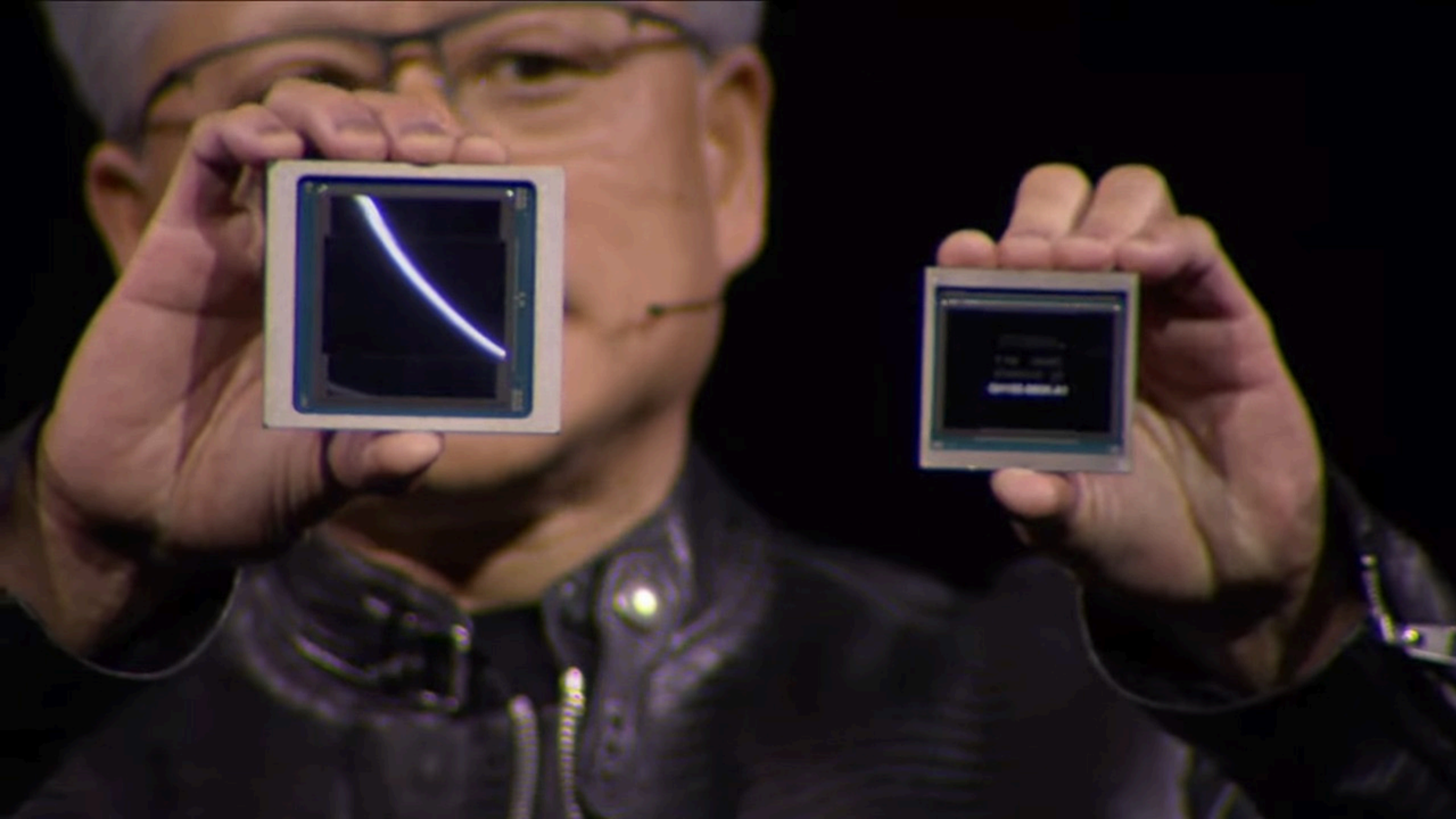
GPU 芯片

1. **制造工艺：**
   - Blackwell GPU 和 H100 GPU 都采用台积电 4N 工艺
2. **晶体管数：**
   - H100 包含 800 亿晶体管，Blackwell GPU 包含 2080 亿晶体管，B200 GPU 晶体管数为 H100 2X
   - 这意味着 B200 芯片封装密度比 H100 进一步提高，对管理散热和功耗也提出更高要求
3. **封装方式：**
   - H100 单 Die（单个完整半导体单元） 封装，Blackwell GPU 封装 2 个 Die
   - B200 每个 Die 算力大概为 H100 1.25X，两个 Die 合封为 H100 2.5X
4. **NV 高带宽接口（Nvidia High Bandwidth Interface）：**
   - B200 2Die，高速连接通道NV-HBI 达 10TB/s，并占用部分芯片面积

GPU 芯片

1. **FP4 精度：**
   - 引入一种新计算精度格式，位宽比 FP8 进一步降低，B200 峰值算力达 18P
2. **FP6 精度：**
   - 位宽介于 FP4 和 FP8 之间
3. **内存与带宽：**
   - 采用 HBM3e，每片大小 ~24GB，理论带宽上限为 1.2TB/s，实际为 1TB/s，Blackwell 芯片上共 8 个
   - 每个 Die 4 个 HBM3e stack，一个 GPU 有 192GB 内存 (8 * 24GB)，内存带宽达 8TB/s (8 * 24GB)
   - 相比 H200 6 个 HBM2，可以减少内存接口芯片面积，使计算面积更大





思考

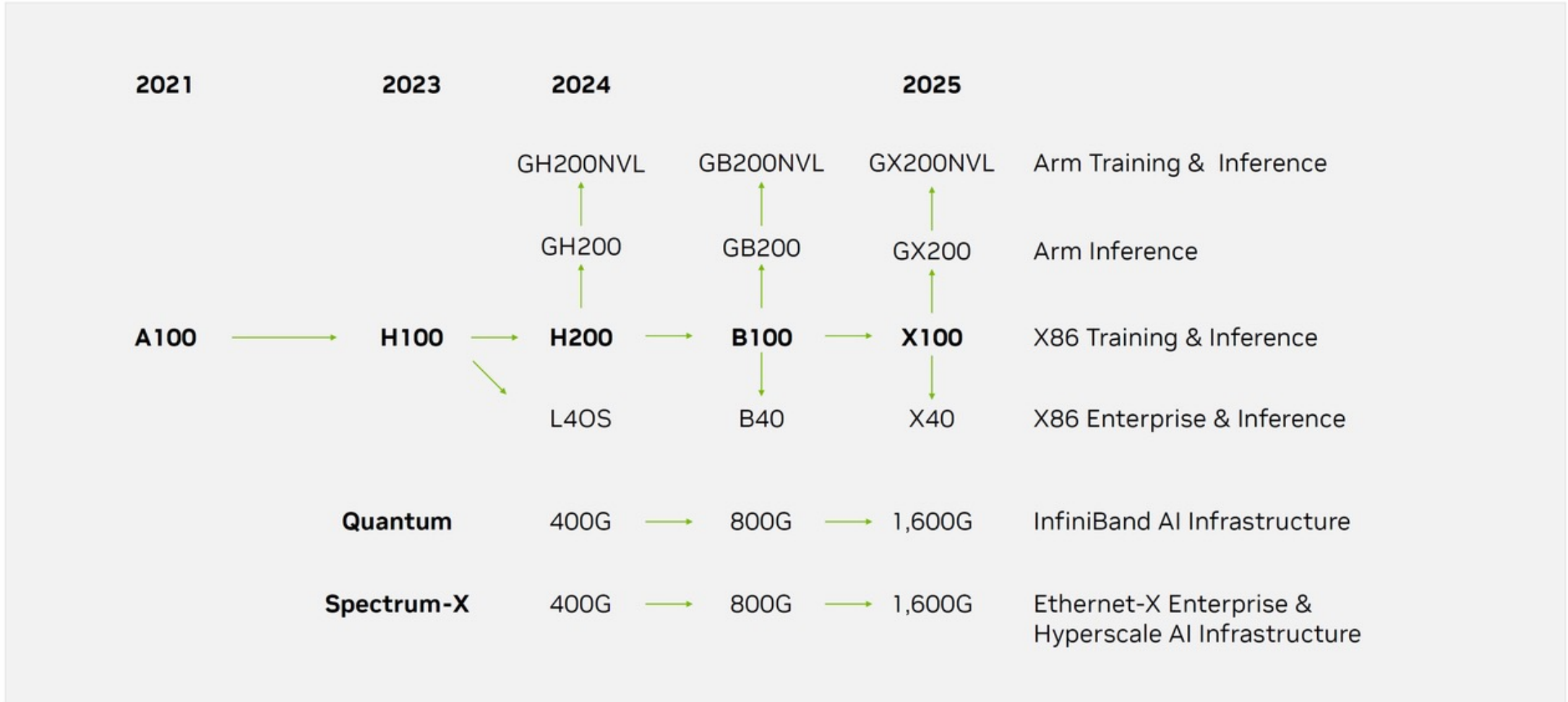
- 制造工艺的性能提升变缓（B200 每个 Die 算力大概为 H100 1.25X，两个 Die 合封为 H100 2.5X）
- 封装能力提升解决工艺制造，拼封装技术从传统单 die 主处理芯片，演进到双 die 合封

02

GH200 & GB200

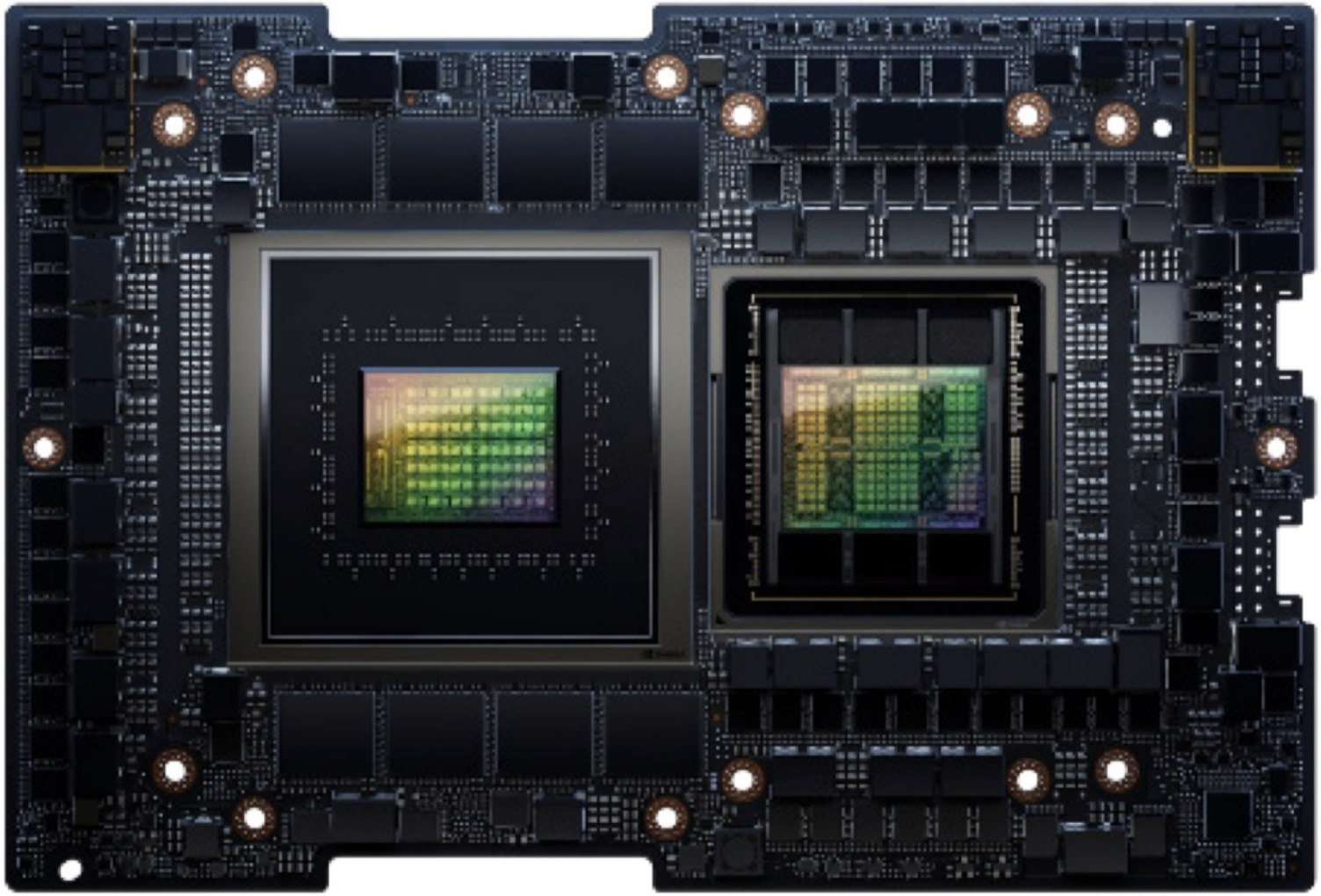
NVIDIA AI – One Architecture | Train and Deploy Everywhere

From Two-Year Rhythm
to
One-Year Rhythm | Training & Inference | x86 & Arm | Hyperscale & Enterprise

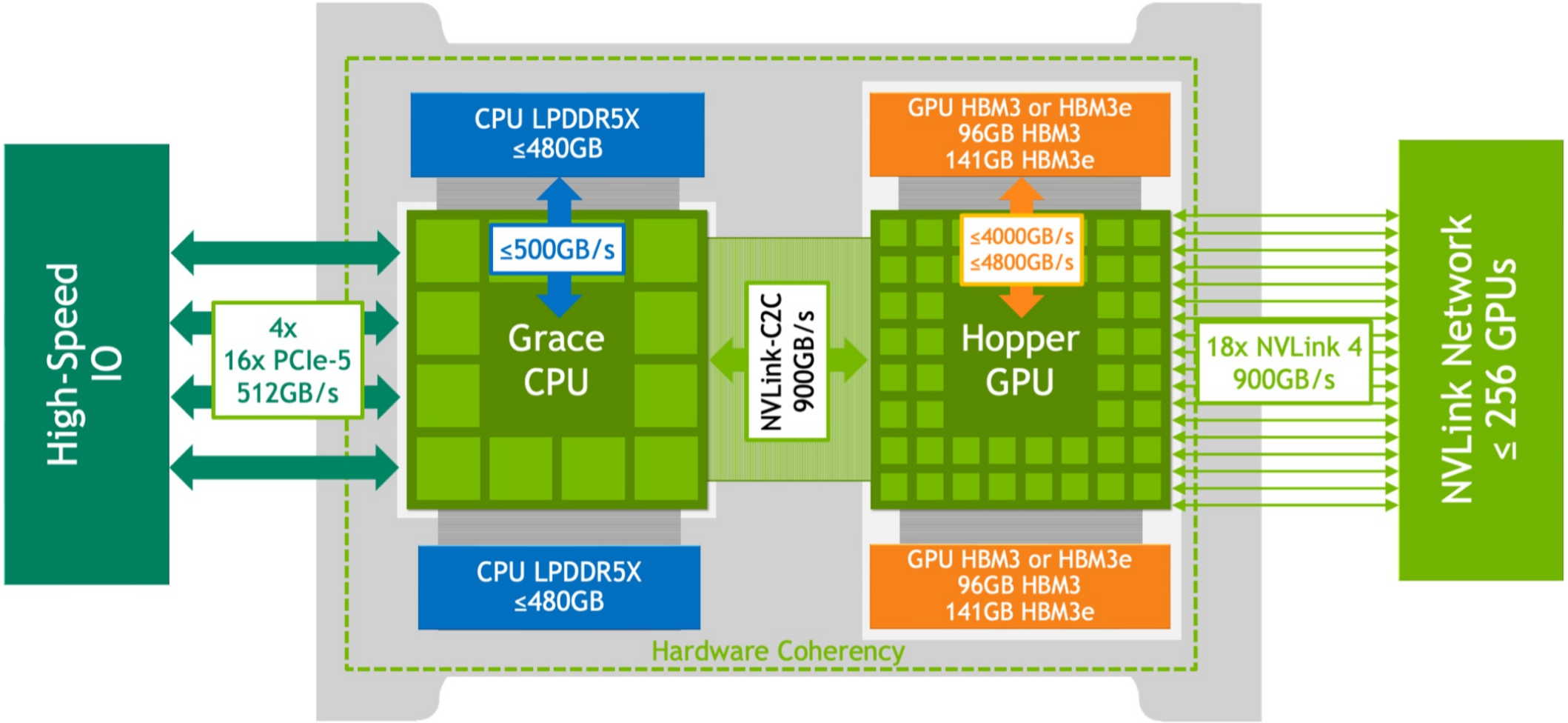


NVIDIA's First GPU-CPU Superchip： GH200

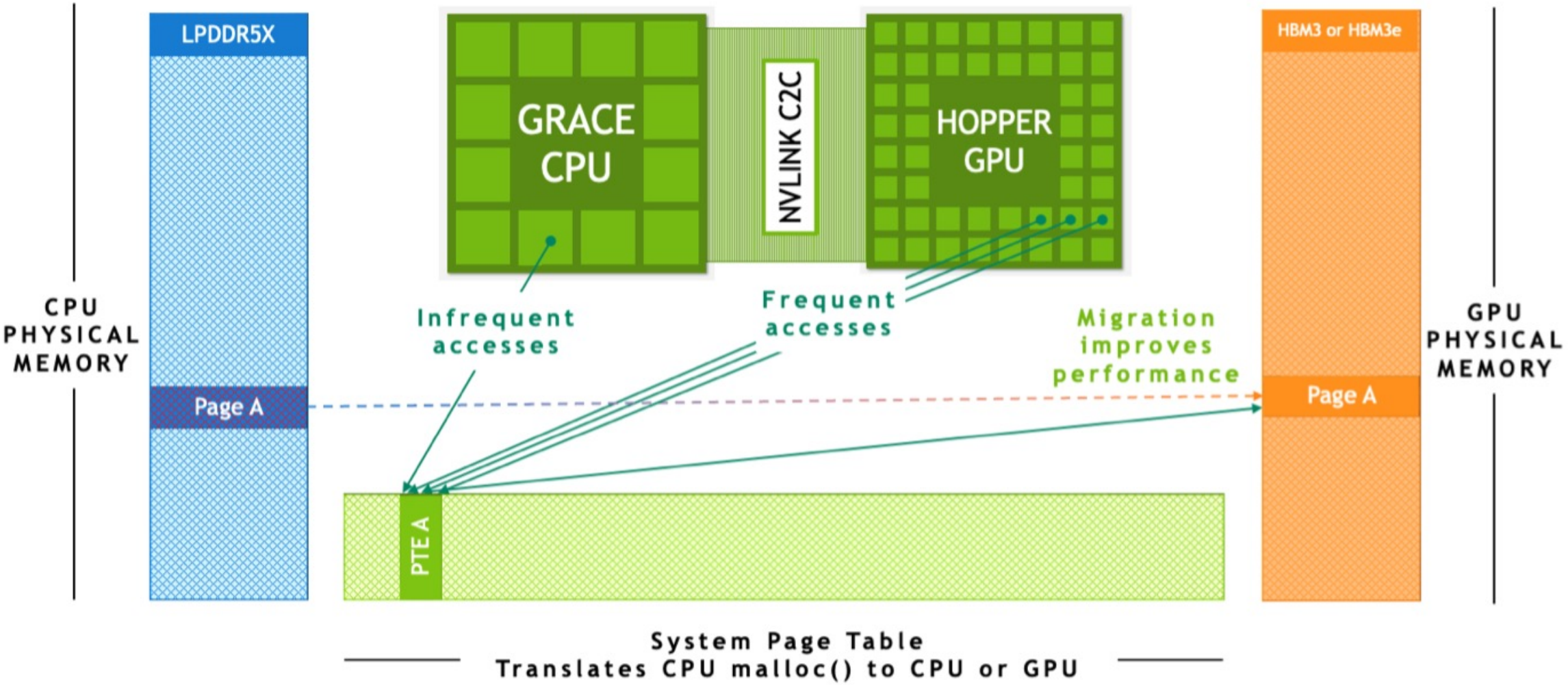
- **H200 GPU + Grace CPU**
  - H200 GPU 显存为 96GB or 144GB。
  - Grace CPU 和 Hopper GPU 间通过 NVLink-C2C 互联，带宽为 900GB/s
  - HBM3e 外 Grace CPU 外接 480GB LPDDR5X，带宽 ~500GB/s。



NVIDIA GH200 Grace Hopper Superchip Logical Overview

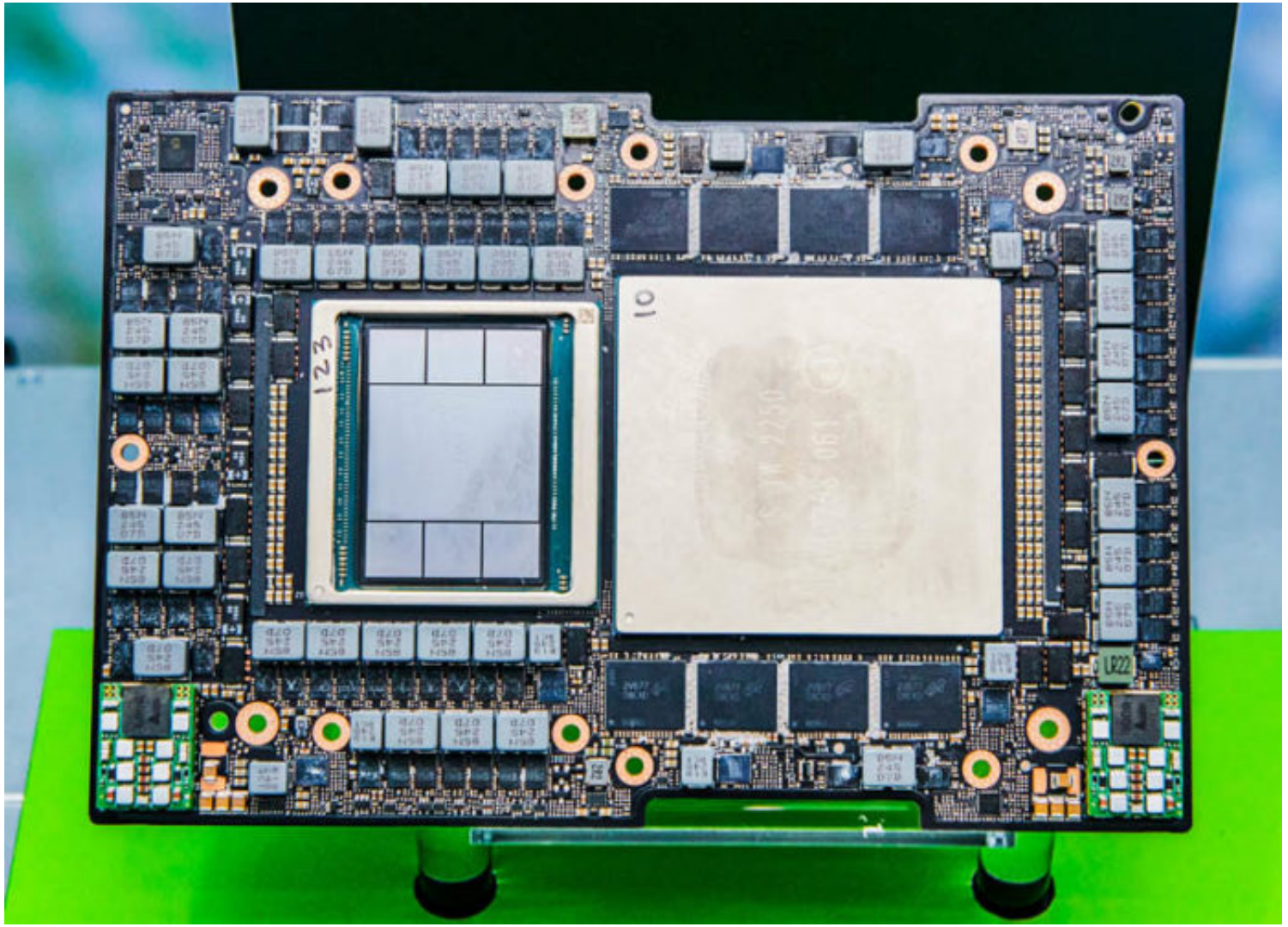


Access-Frequency-Based Automatic Memory Migration



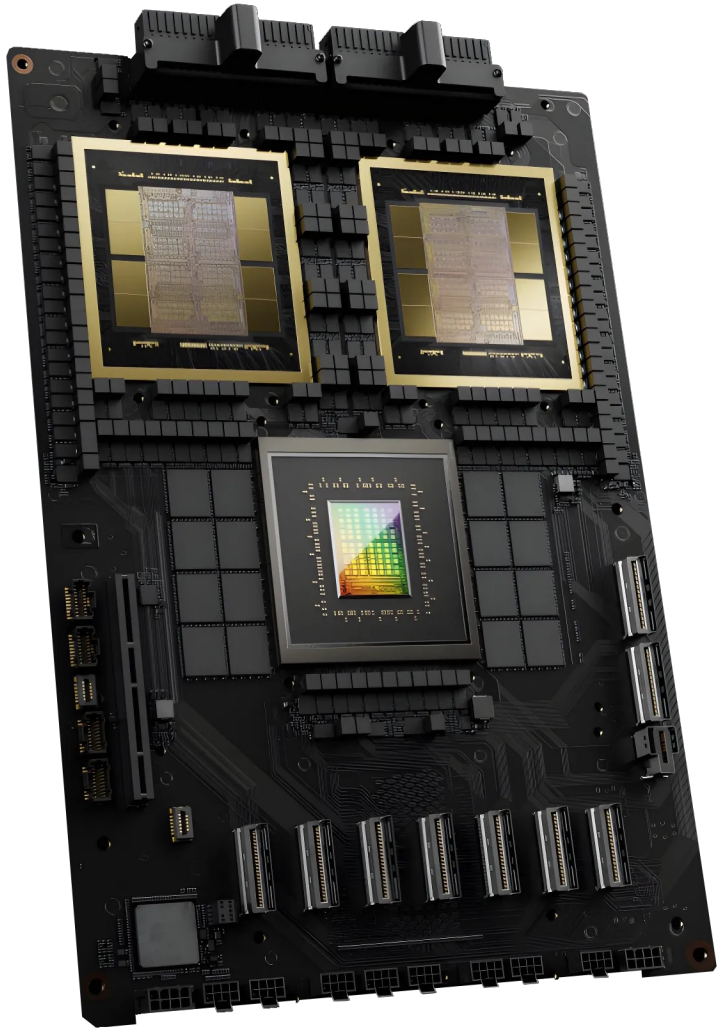
NVIDIA's First GPU-CPU Superchip

- **H200 GPU + Grace CPU**
  - H200 GPU 显存为 96GB or 144GB。
  - Grace CPU 和 Hopper GPU 间通过 NVLink-C2C 互联，带宽为 900GB/s
  - HBM3e 外 Grace CPU 外接 480GB LPDDR5X，带宽 ~500GB/s。

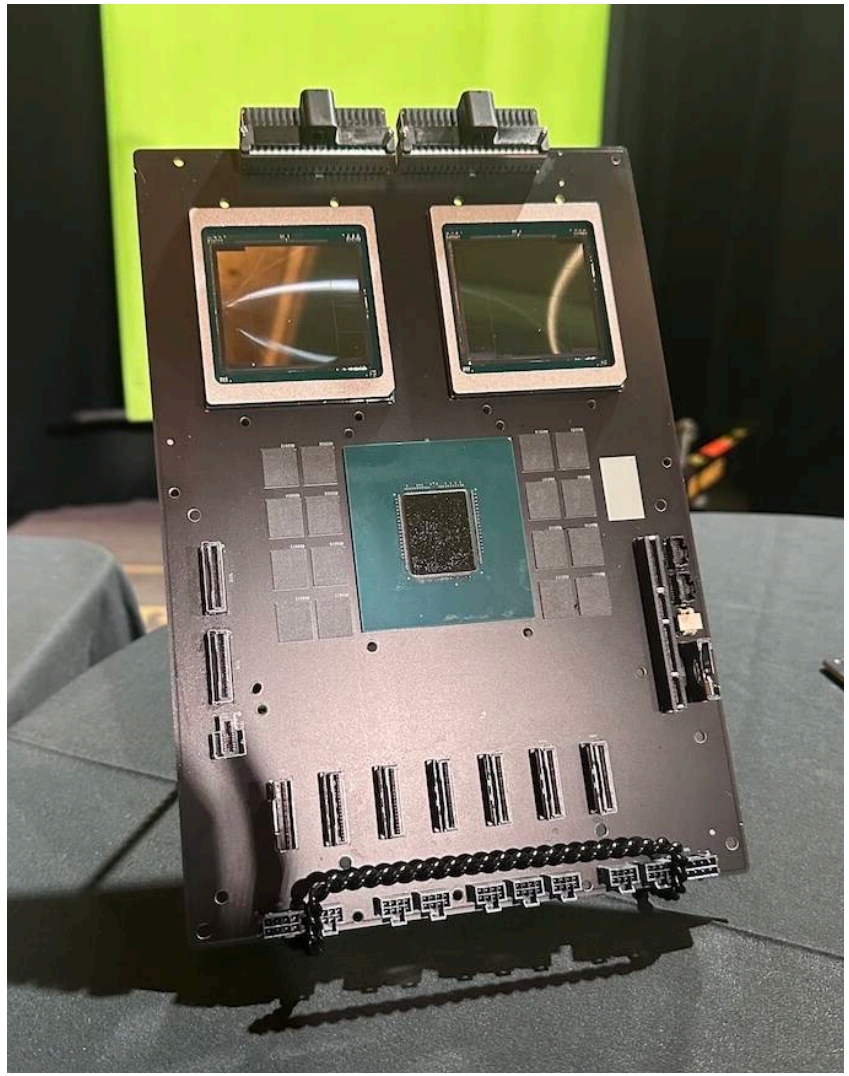


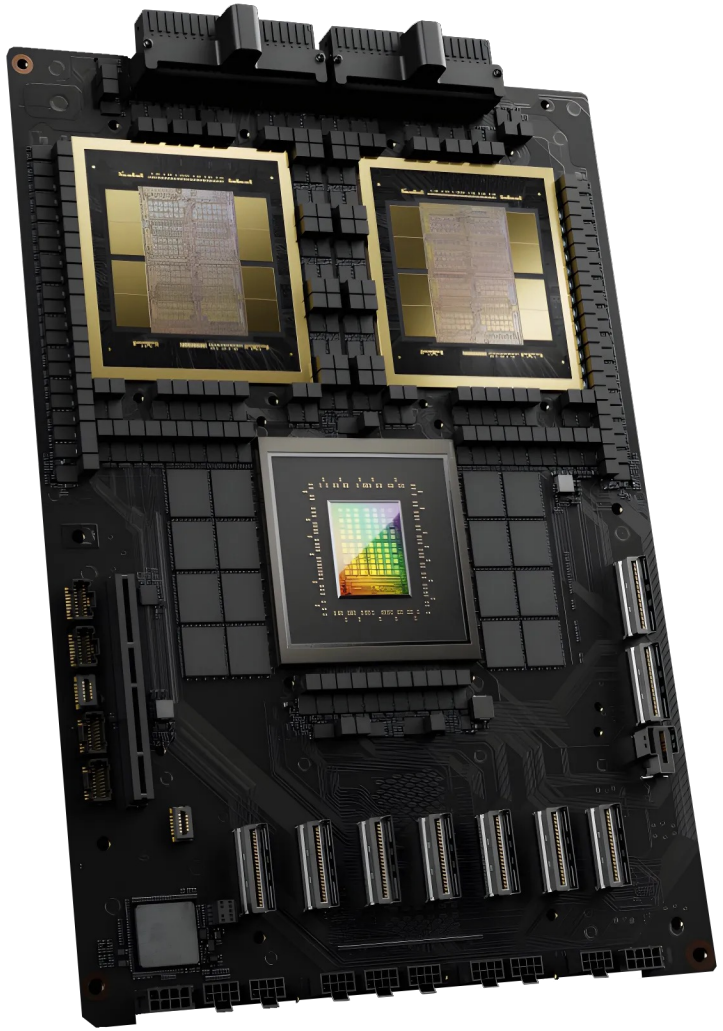
GB200

- 一个 GB200 由 1 个 Grace CPU + 2 个 Blackwell GPU 组成，相应 GPU 算力和显存加倍。
- CPU 和 GPU 间通过 900GB/s NVLink-C2C 实现高速互联，功耗为 1200w。
- 包含 384 GB HBM3e 显存，以及 GH100 同样 480GB LPDDR5X，Memory 为 384GB + 480GB = 864GB。

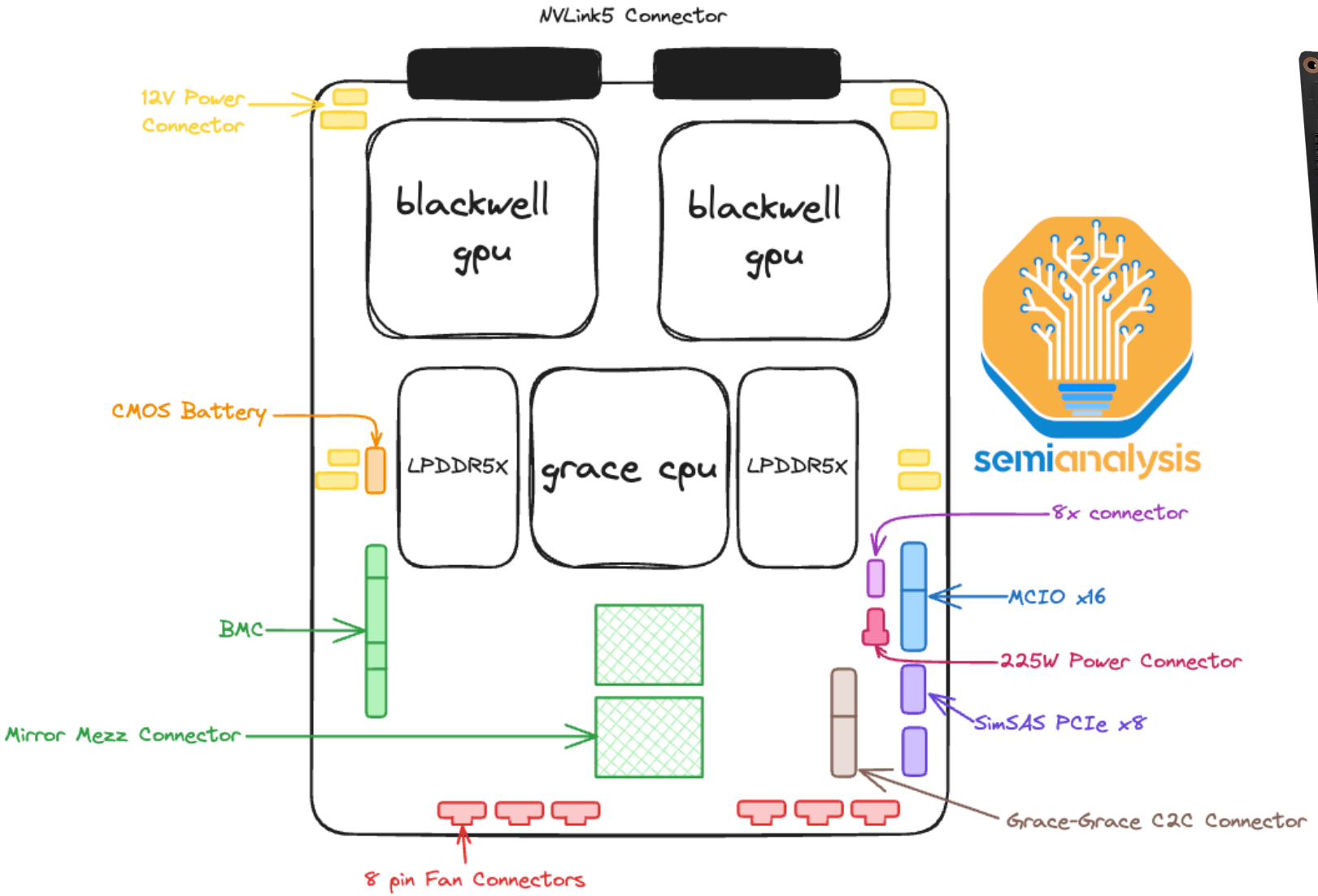


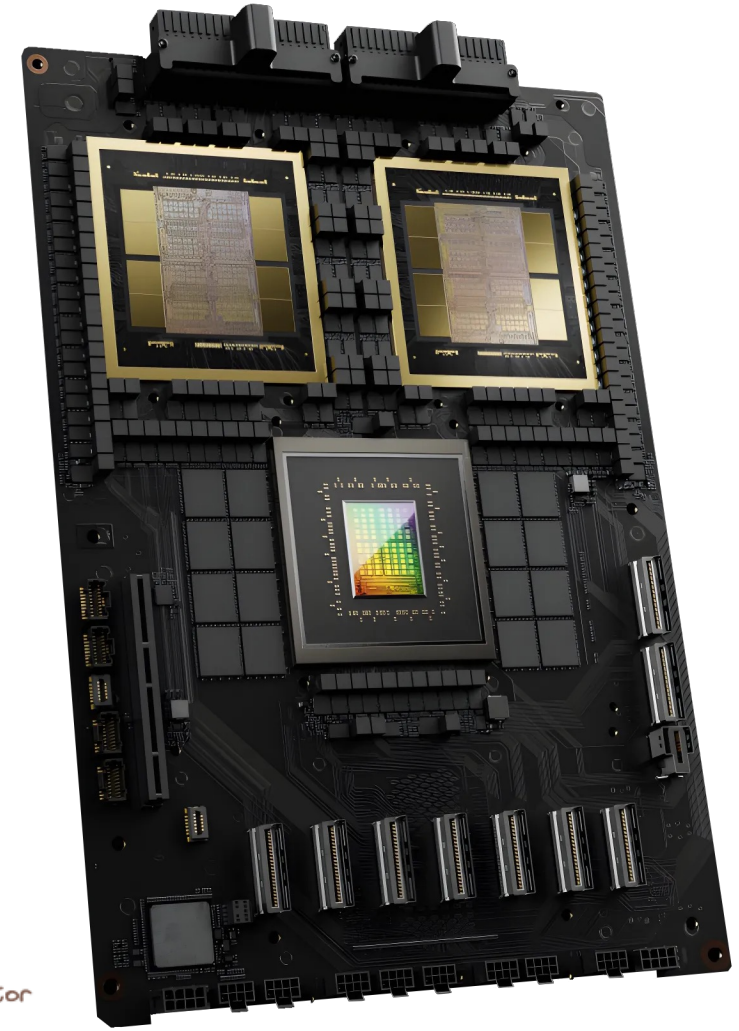
GB200





GB200 Bianca Board





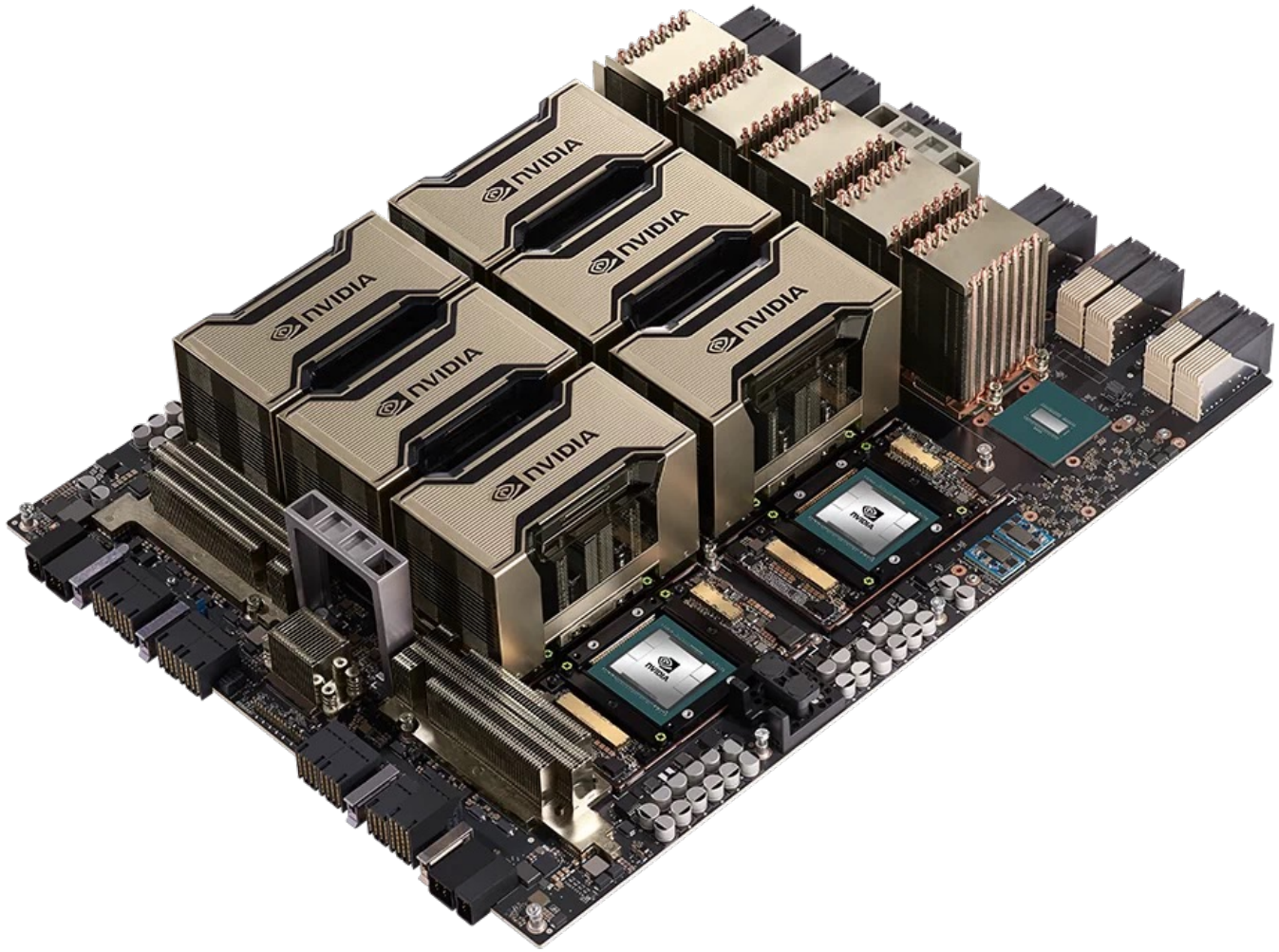
思考

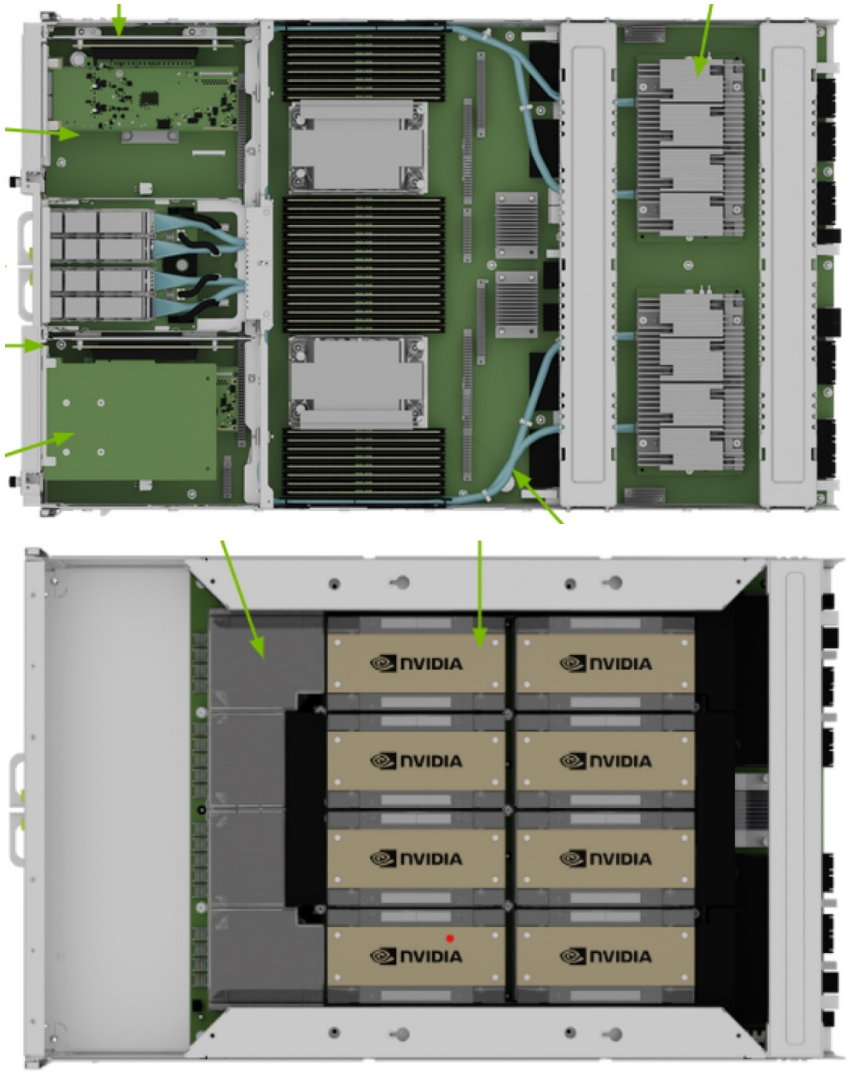
- Nvidia 不仅使用封装技术来解决部件性能瓶颈，现在使用板卡（Rack）新概念
- Rack 为超节点提供更加高密度的集成解决方案
- Host - Devices 统一编址利于内存管理，特别是在大规模的推荐模型上

03

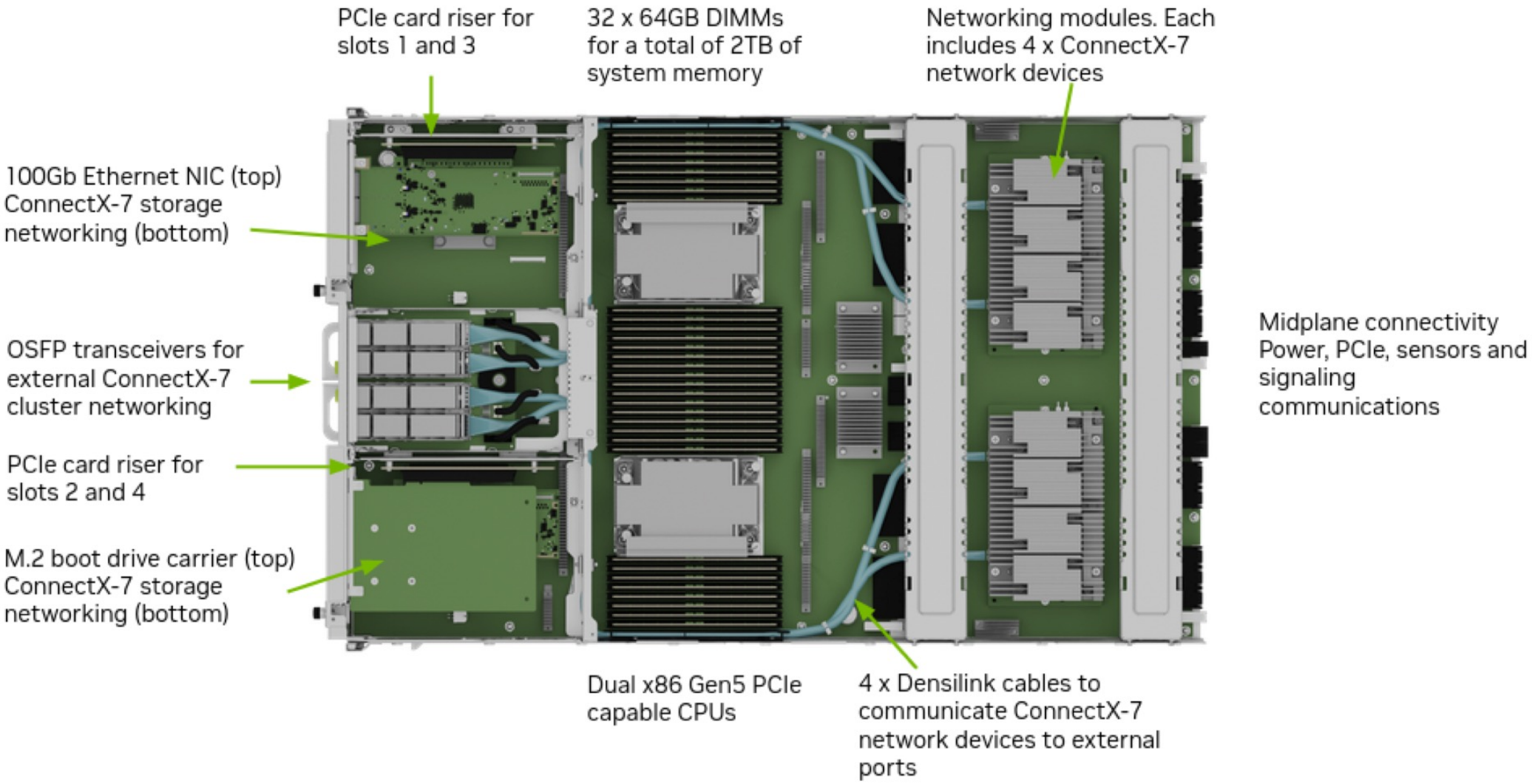
HGX H/B 100/200

HGX H100 & H200



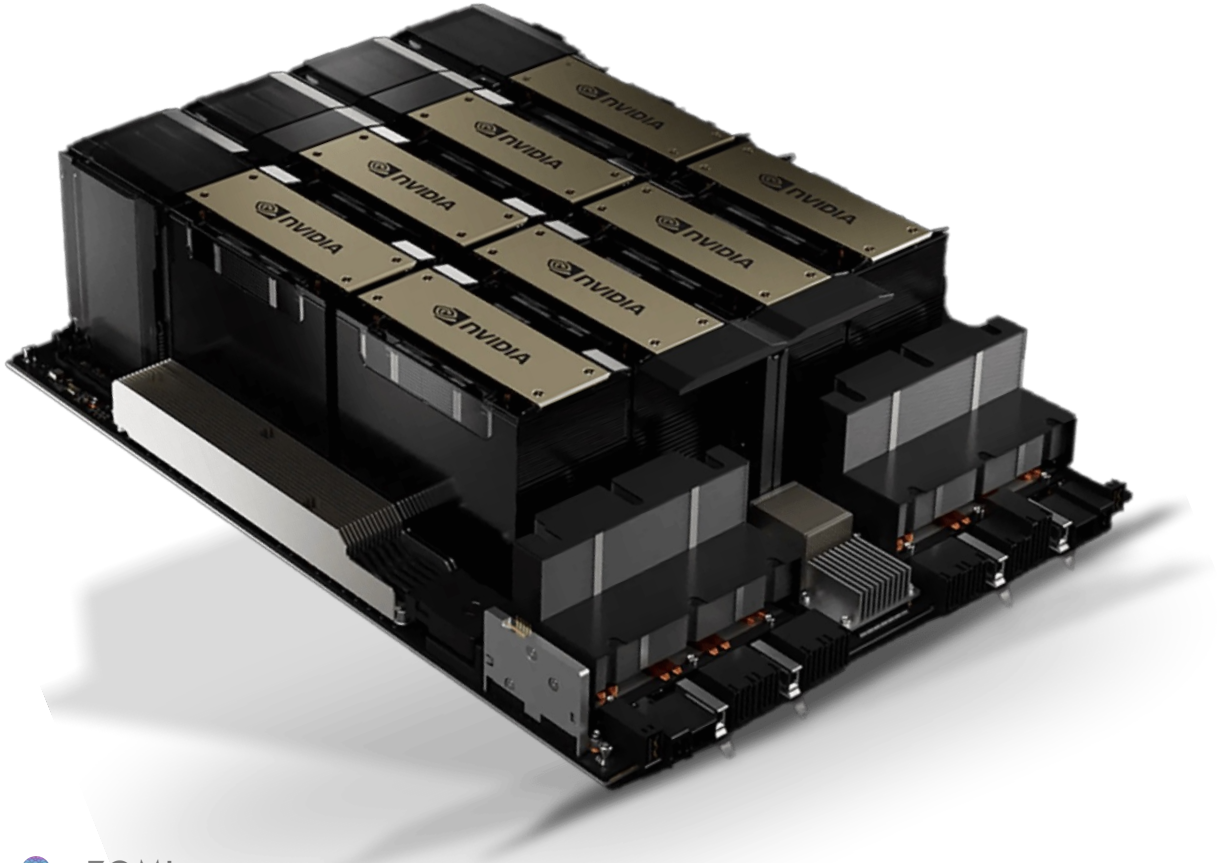


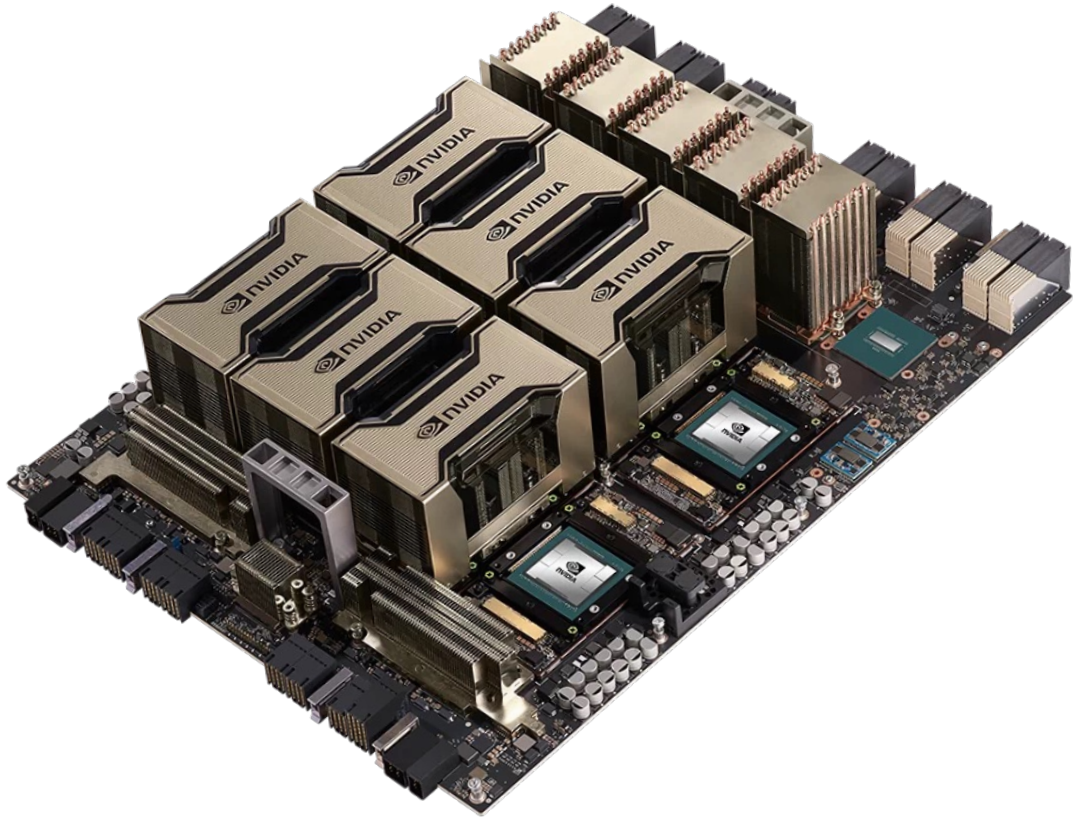
https://www.nvidia.com/en-us/data-center/hgx/



HGX 系列

- B100 和 B200 分别与之前的 H100 和 B200 对应，都是不带 Grace CPU 的版本，可以配合 Intel CPU 或 AMD CPU 使用。





HGX H100 & H200

	HGX H100	HGX H200
	8-GPU	**8-GPU**
Form Factor	8x NVIDIA H100 SXM	8x NVIDIA H200 SXM
FP8 Tensor Core*	32 PFLOPS	32 PFLOPS
INT8 Tensor Core*	32 POPS	32 POPS
FP16/BF16 Tensor Core*	16 PFLOPS	16 PFLOPS
TF32 Tensor Core*	8 PFLOPS	8 PFLOPS
FP32	540 TFLOPS	540 TFLOPS
FP64	270 TFLOPS	270 TFLOPS
FP64 Tensor Core	540 TFLOPS	540 TFLOPS
Memory	640GB HBM3	1.1TB HBM3
GPU Aggregate Bandwidth	27GB/s	38GB/s
NVLink	Fourth generation	Fourth generation
NVSwitch	Third generation	Third generation
NVSwitch GPU-to-GPU Bandwidth	900GB/s	900GB/s
Total Aggregate Bandwidth	7.2TB/s	7.2TB/s

HGX B100 & B200

	HGX B200	HGX B100
Form Factor	8x NVIDIA B200 SXM	8x NVIDIA B100 SXM
FP4 Tensor Core*	144 PFLOPS	112 PFLOPS
FP8/FP6 Tensor Core*	72 PFLOPS	56 PFLOPS
INT8 Tensor Core*	72 POPS	56 POPS
FP16/BF16 Tensor Core*	36 PFLOPS	28 PFLOPS
TF32 Tensor Core*	18 PFLOPS	14 PFLOPS
FP32	640 TFLOPS	480 TFLOPS
FP64	320 TFLOPS	240 TFLOPS
FP64 Tensor Core	320 TFLOPS	240 TFLOPS
Memory	Up to 1.5TB	Up to 1.5TB
NVLink	Fifth generation	Fifth generation
NVIDIA NVSwitch™	Fourth generation	Fourth generation
NVSwitch GPU-to-GPU Bandwidth	1.8TB/s	1.8TB/s
Total Aggregate Bandwidth	14.4TB/s	14.4TB/s

其他信息

- B100 算力 B200 ~3/4，8 x B100 稀疏 FP16 算力为 28PF，8xB200 的稀疏 FP16 算力为 36PF，可见 8*B200 的稀疏 FP16 算力为 8*H100/H200 的 2.25 倍。相当于单个 B200 的稀疏 FP16 算力为 4.5PF。
- Blackwell 的 Tensor Core 相比 Hopper 添加了对 FP6 和 FP4 的支持，从上图也可以看出其 FP4 算力为 FP8 算力 2 倍，FP16 算力 4 倍。而 Blackwell 的 CUDA Core 不再支持 INT8。此外，从 Hopper 开始都不再支持 INT4。

04

总结与思考

NVIDIA GPU架构发展

- B200、B100、GB200、NVL72、NVL32、SuperPod、GH200、H200、H100、L20、SuperPod-576
- ConnectX-800G 网卡、网络交换机

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