NV GB200 互联集群架构

NVIDIA ZOMI

NVIDIA.

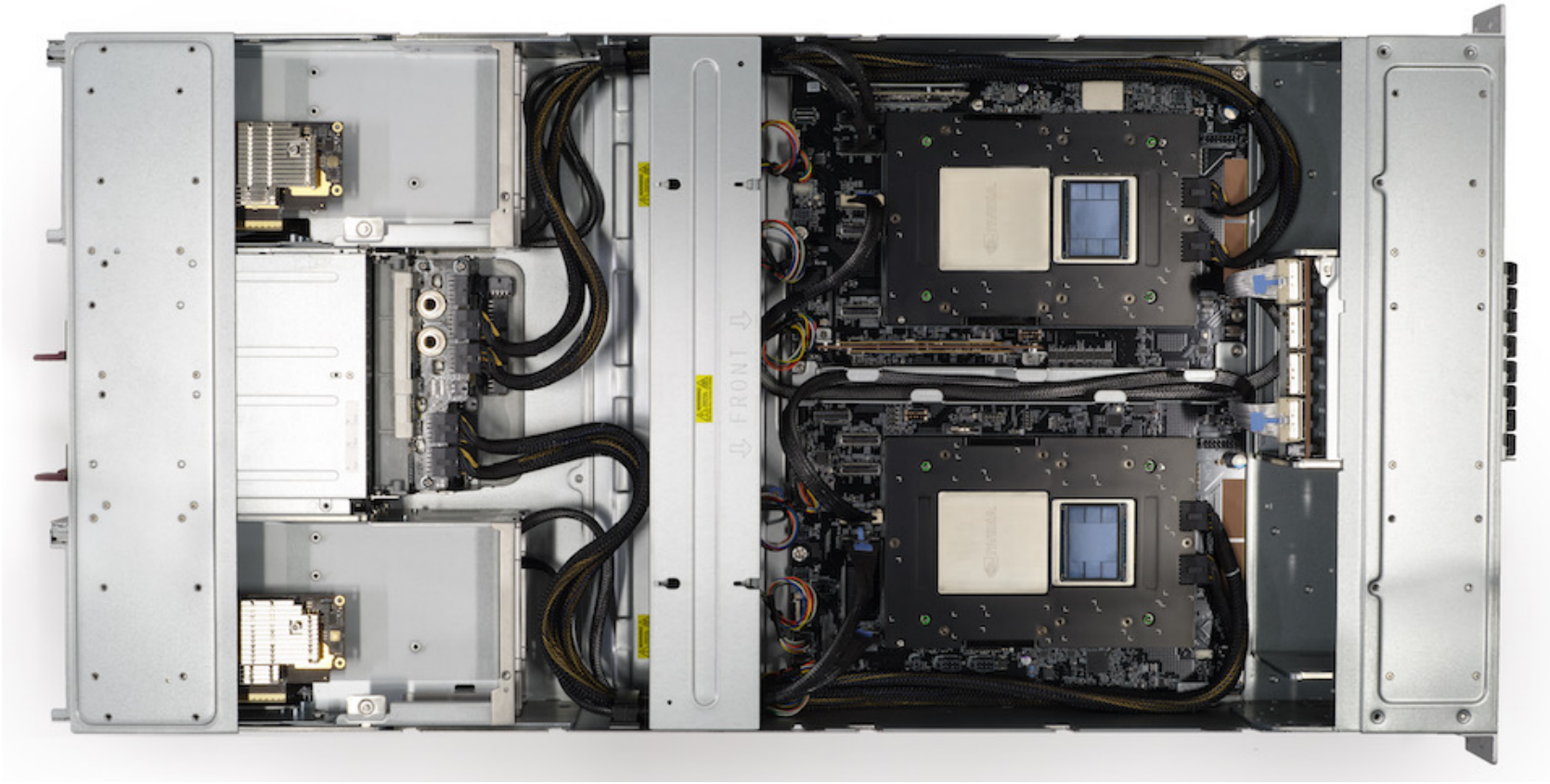
本节内容

1. GH200 to NVL32 & SuperPod
2. GB200 to NVL32 & SuperPod
3. NV 产品演进的小结与思考

01 GH200 NVL32&SuperPod

GH200 Compute Tray

- GH200 Compute Tray 基于 NVIDIA MGX 设计（1U），每个 Compute Tray 上有 2 个 GH200，也就是 2 个 Grace CPU 和 2 个 H200 GPU



NVSwitch Tray

- 第一代 NVSwitch Tray 包含 2 个第 3 代 NVSwitch 芯片，共 128 个 NVLink Port，通信带宽上限为 6.4TB/s

GH200 NVL32

- 一个机柜有 16 x GH200 Compute Tray 和 9 x NVSwitch Tray，共 32 x GH200 GPU + 18 x NVSwitch

- 实际 32 x GH200 有 32x18=576 个 NVLink，只需 576/64=9 个 NVSwitch 可以实现全互联

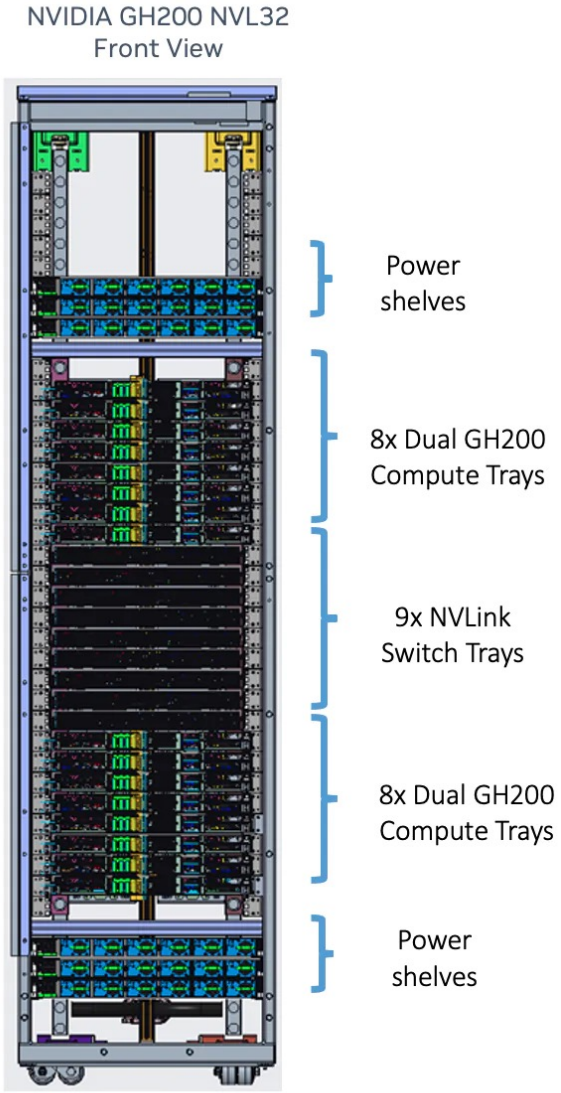
NVLink Switch x9

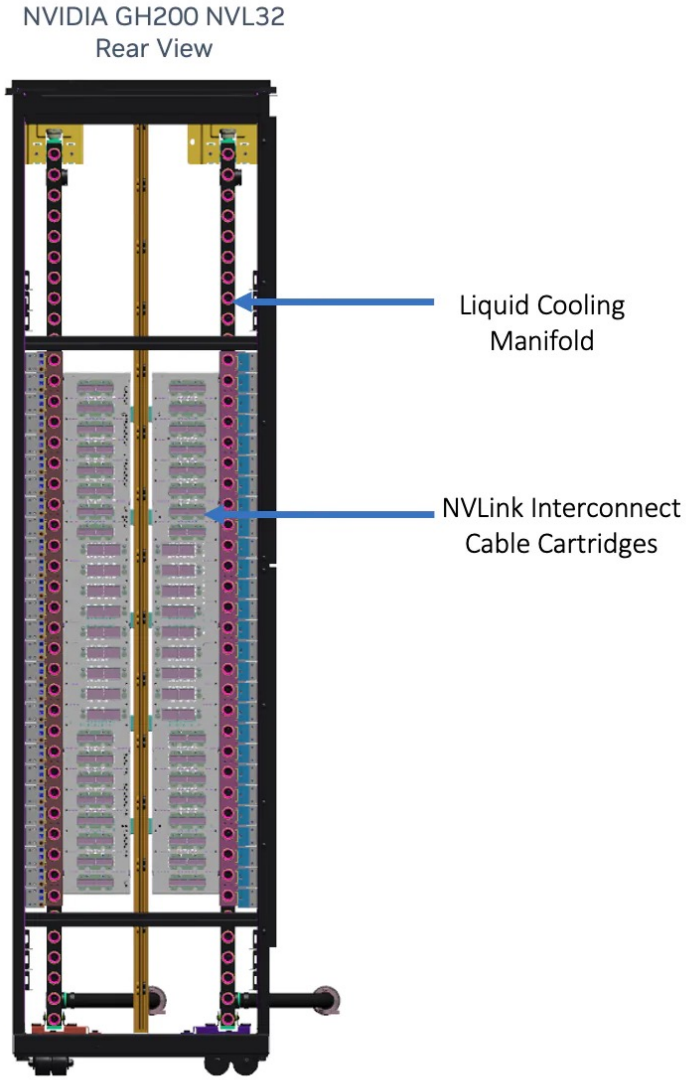
NVIDIA GH200 x32

GH200 NVL32

- 一个机柜有 16 x GH200 Compute Tray 和 9 x NVSwitch Tray，共 32 x GH200 GPU + 18 x NVSwitch

- 实际 32 x GH200 有 32x18=576 个 NVLink，只需 576/64=9 个 NVSwitch 可以实现全互联





GH200 SuperPod

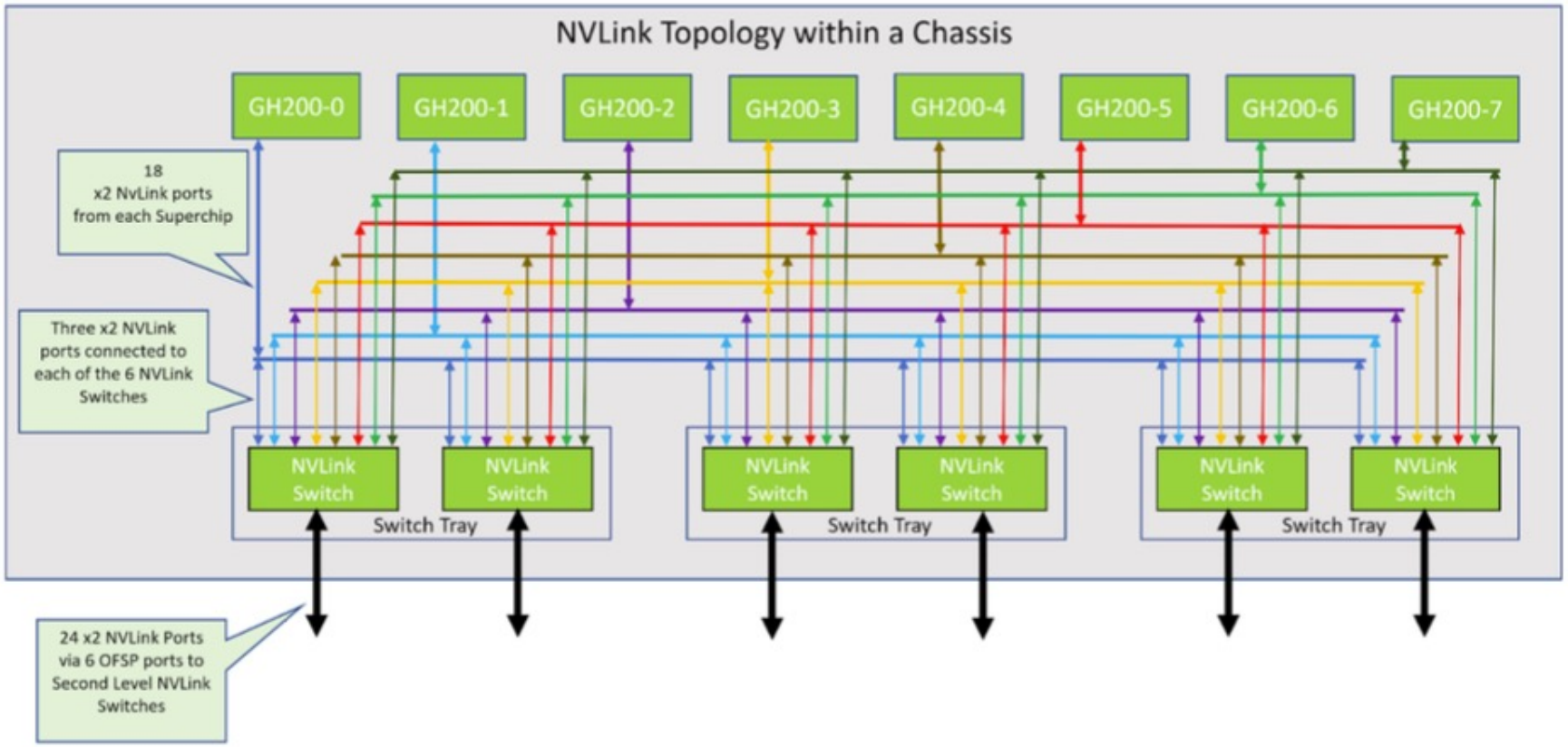
- GH200 SuperPod 由 256 个 GH200 GPU 实现全互联，但并不是由 8 个 NVL32 组成的，而是由 32 个 8-Grace Hopper Superchip 组成。

GH200 SuperPod：32 x 8-Grace Hopper Superchip

- 8 个 Hopper Compute Tray（8U），每个包含：
  - 1 个 GH200 GPU
  - 1 个 ConnectX-7 IB 网卡，400Gb/s
  - 1 个 200 Gb/s 以太网卡
- 3 个 NVSwitch Tray（3U），共 6 个 NVSwitch

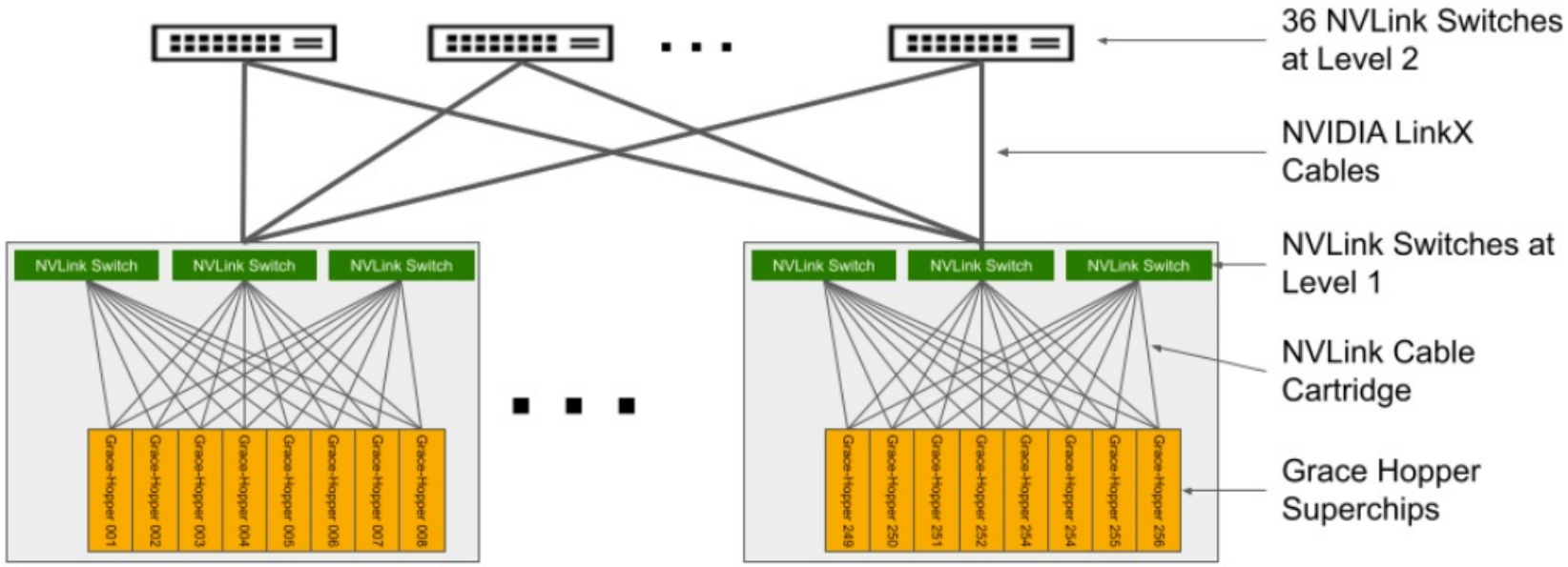
GH200 SuperPod：NVLink

- 每个 GH200 和每个 NVSwitch 上有 3 个 NVLink 连接，每个 NVSwitch 上使用了 24 个 Port
- 每个 NVSwitch 有 24 个 Port 与 L2 NVSwitch 相连，相当于每个 NVSwitch 使用 48 个 Port



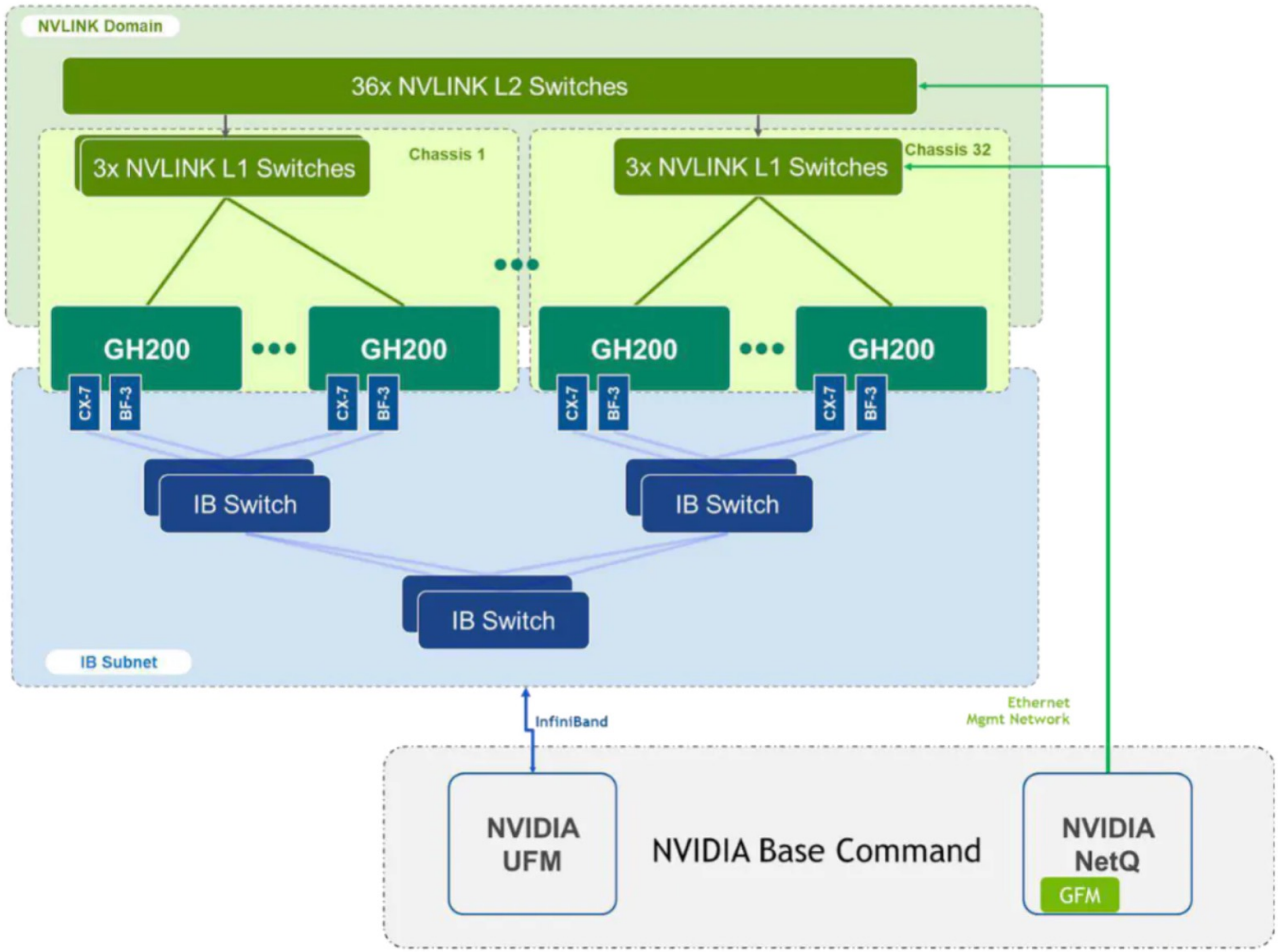
GH200 SuperPod：NVLink

- GH200 SuperPod 由 32 x 8-Grace Hopper Superchip 组成，L1 包含 32 x 3=96 NVSwitch Tray（192 个 NVSwitch），L2 包含 36 个 NVSwitch Tray（64 个 NVSwitch）
- 每个 L1 NVSwitch Tray 有 24 x 2=48 个 Port 与 L2 NVSwitch Tray 相连，因此需要 96x48/(64*2)=36 个 L2 NVSwitch Tray



GH200 SuperPod：NVLink

256 个 GH200 通过两级 IB 交换机互联



02 GB200 NVL72&SuperPod

GB200 NVL72

- 一个 GB200 NVL72 Rack 包含 18 个 GB200 Compute Tray，即 36 个 Grace CPU + 72 个 GPU，9 个 NVSwitch Tray

- 显存为 72*192GB=13.8TB，CPU 对应 Fast Memory LPDDR5X 为 480GB x 36 = 17TB，共 Fast Memory 为 30TB

GB200 Compute Tray

- 基于 NVIDIA MGX 设计（大小为 1U），一个 Compute Tray 包含 2 个 GB200：2 个 Grace CPU、4 个 Blackwell GPU

- 支持 1.7TB Fast Memory，每个 Blackwell 显存为 192GBx4=768GB，1.7TB 包含每个 GB200 额外 480GB LPDDR5X，一共 768GB + 480GBx2=1728GB

GB200 NVSwitch Tray

- 一个 Switch Tray 包含两颗 NVLINK Switch 芯片，累计提供 72x2 = 144 个 NVLINK Port
- 单颗 NVSwitch 芯片，上下各 36 个 Port，带宽为 7.2TB/s
- 28.8 Tbps 交换容量，相对于最领先 51.2Tbps 交换芯片小一些

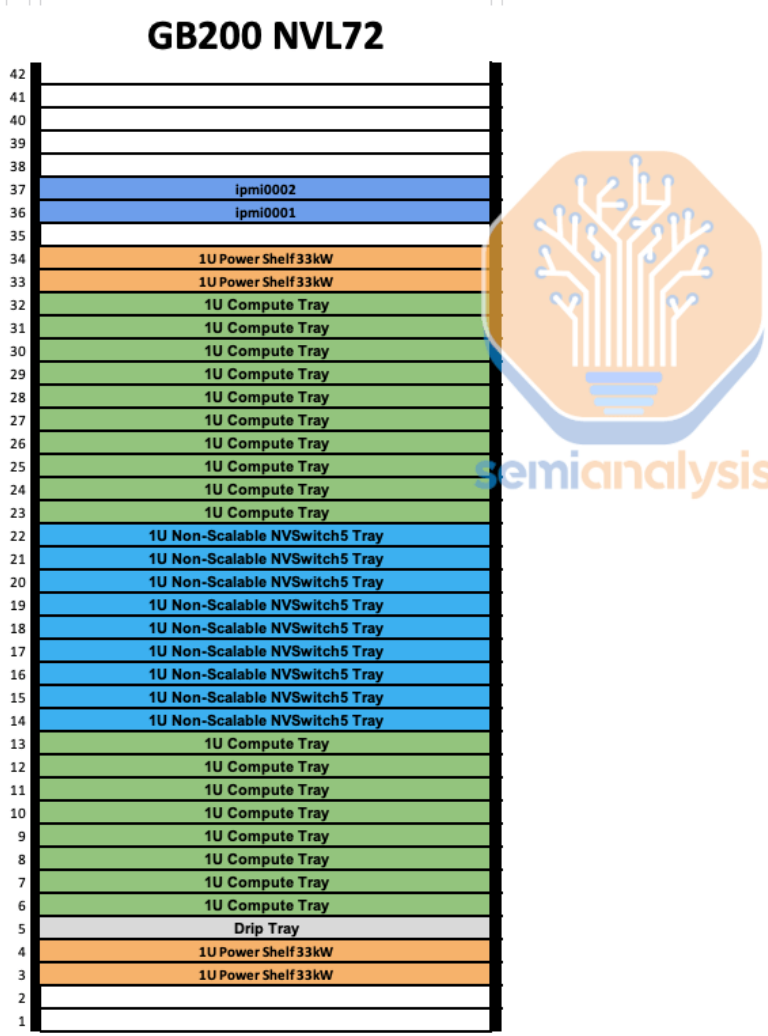
GB200 NVL72

- 单个 GB200 子系统则包含 2 x 18 = 36 个 NVLink 5th Port
- 整个系统对外互联上并没有采用 OSFP 光模块接口，而是直接通过后置的铜线背板连接

GB200 NVL72

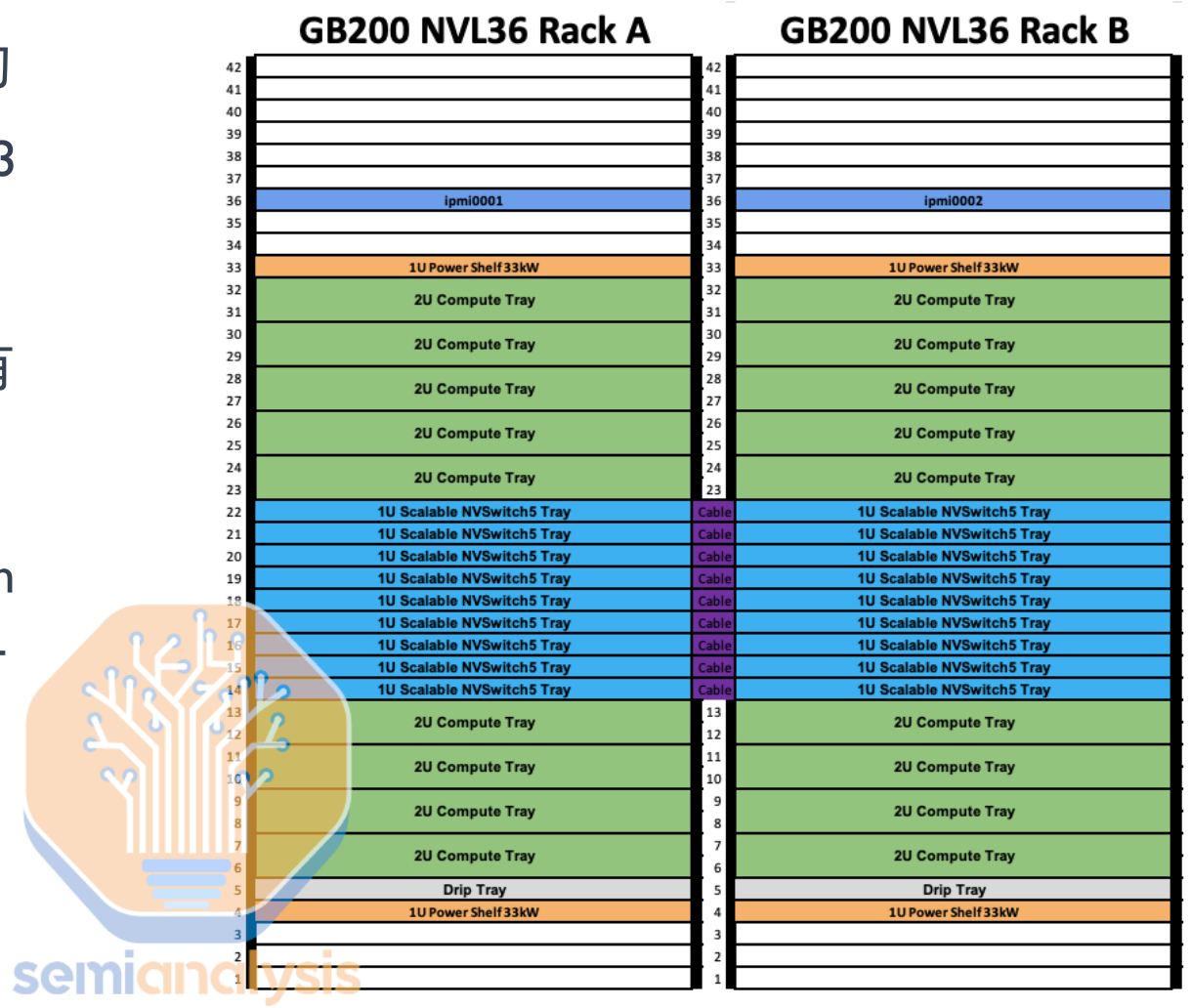
- GB200 NVL72 外形尺寸，需要大约 120kW/机架。每机架超过 40kW，因此 GB200 需要液冷。
- 通用 CPU 机架支持高达 12kW/机架，而更高密度的 H100 风冷机架通常仅支持大约 40kW/机架。

- 18 个 1U Computer Tray 和 9 个 NVSwitch Tray 组成。每个 Computer Tray 高 1U，包含 2 个 Bianca 板。每个 Bianca 板包含 1 个 Grace CPU 和 2 个 Blackwell GPU。 NVSwitch Tray 有两个 28.8Tb/s NVSwitch5。



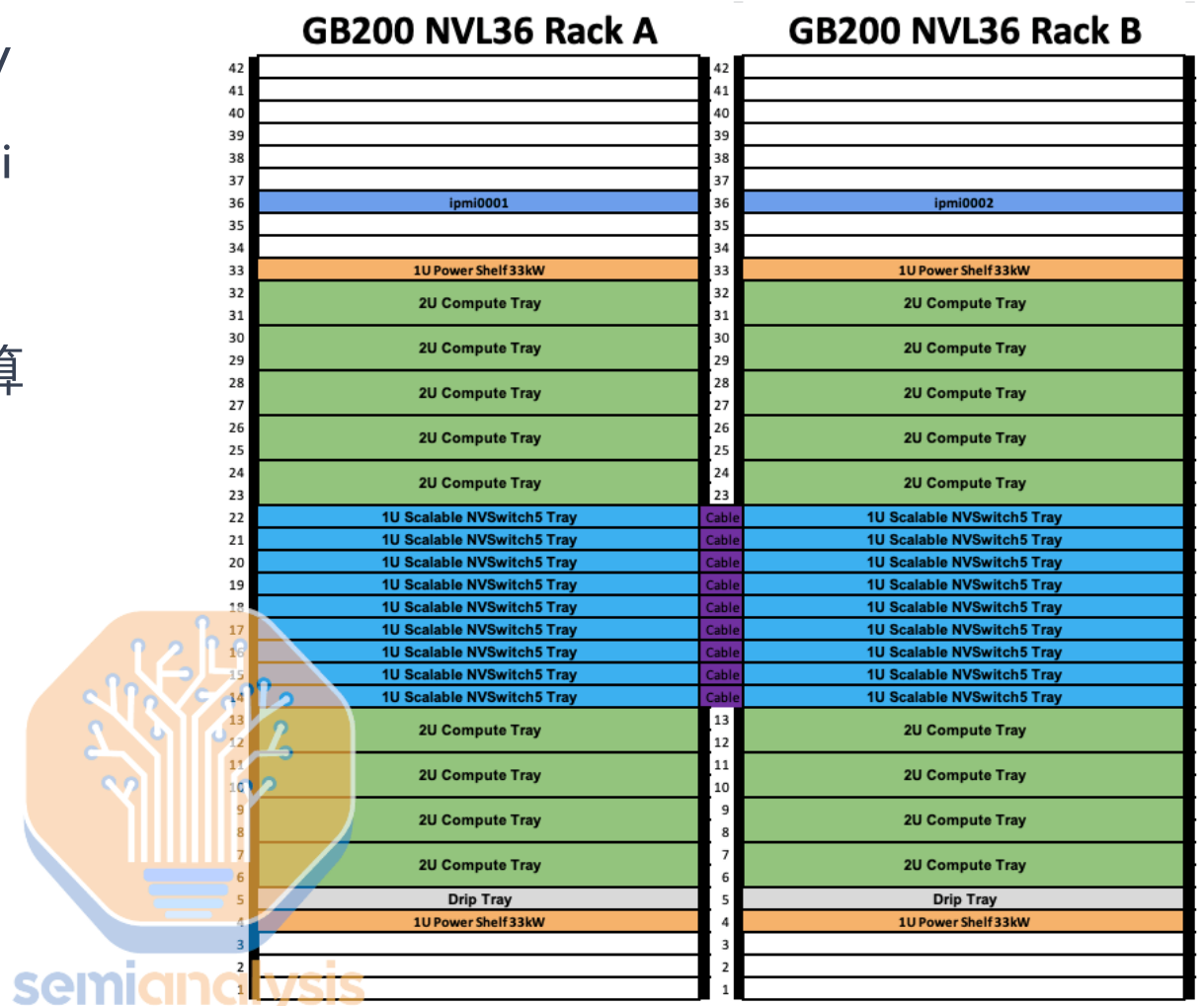
GB200 NVL36

- GB200 NVL36 * 2，两个并排互连在一起的机架。每个机架包含 18 个 Grace CPU 和 36 个 Blackwell GPU。
- 在 2 个机架之间，仍然保持 NVL72 中所有 72 个 GPU 之间的无阻塞全对全。
- 每个计算托盘的高度为 2U，包含 2 个 Bianca 板。每个 NVSwitch 托盘都有两个 28.8Tb/s NVSwitch5 芯片。



GB200 NVL36

- 2025 年第二季度将推出B200 NVL72 和 NV L36x2 规格，将使用 x86 CPU 而不是 Nvidi a 内部的 grace CPU。
- 这种规格称为 Miranda。semi 认为每个计算托盘的 CPU 到 GPU 的比例将保持不变，即每个计算托盘 2 个 CPU 和 4 个 GPU。



思考

A： H100 时代大家预测光模块的需求会激增，不过很快 B200 时代大家有改变方向了，预测后面光模块需求急剧下降，因为类似于大型机的交付时代后面选用铜互联。ZOMI 老师怎么看？

思考

A： H100 时代大家预测光模块的需求会激增，不过很快 B200 时代大家有改变方向了，预测后面光模块需求急剧下降，因为类似于大型机的交付时代后面选用铜互联。

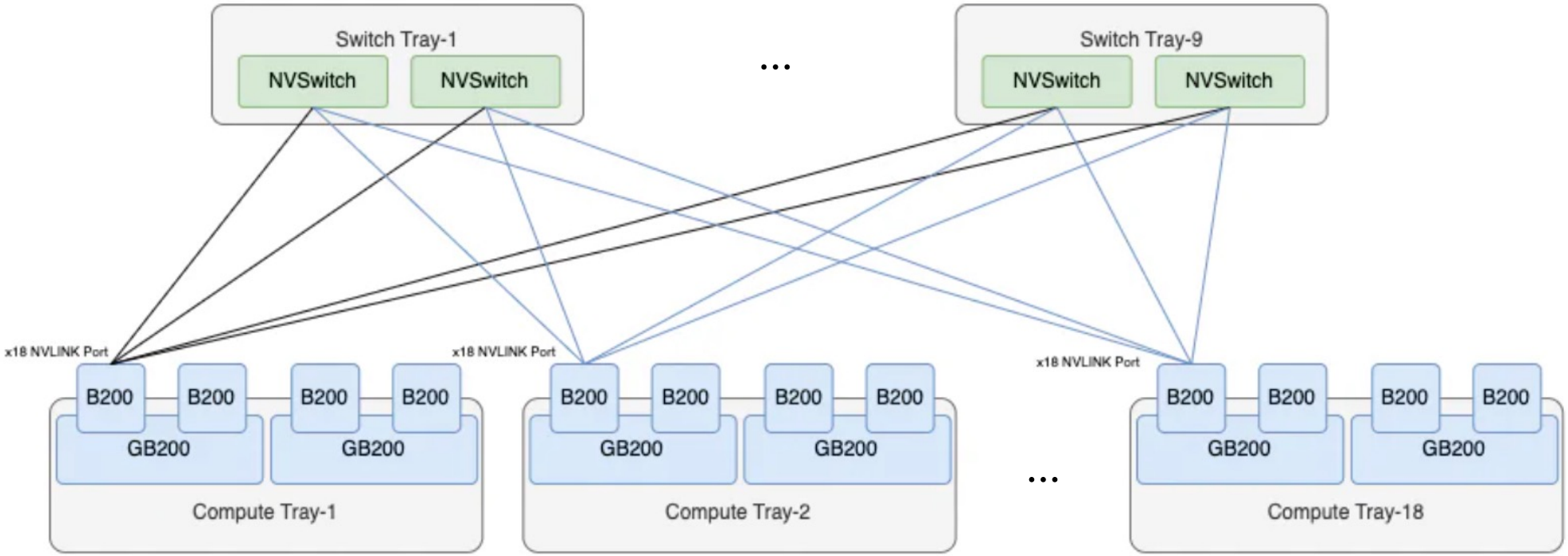
Q： 看每一代架构演进：Hopper 使用松耦合连接方式，机柜的散热和部署会相对灵活，而且同时提供 HGX 方式部署；B200 采用大型机交付方式，铜互联有效降低整体功耗和增加超节点内的稳定性。

GB200 NVL72

	GB200 NVL72	GB200 Grace Blackwell Superchip
Configuration	36 Grace CPU : 72 Blackwell GPUs	1 Grace CPU : 2 Blackwell GPU
FP4 Tensor Core[2]	1,440 PFLOPS	40 PFLOPS
FP8/FP6 Tensor Core[2]	720 PFLOPS	20 PFLOPS
INT8 Tensor Core[2]	720 POPS	20 POPS
FP16/BF16 Tensor Core[2]	360 PFLOPS	10 PFLOPS
TF32 Tensor Core	180 PFLOPS	5 PFLOPS
FP32	6,480 TFLOPS	180 TFLOPS
FP64	3,240 TFLOPS	90 TFLOPS
FP64 Tensor Core	3,240 TFLOPS	90 TFLOPS
GPU Memory | Bandwidth	Up to 13.5 TB HBM3e | 576 TB/s	Up to 384 GB HBM3e | 16 TB/s
NVLink Bandwidth	130TB/s	3.6TB/s
CPU Core Count	2,592 Arm® Neoverse V2 cores	72 Arm Neoverse V2 cores
CPU Memory | Bandwidth	Up to 17 TB LPDDR5X | Up to 18.4 TB/s	Up to 480GB LPDDR5X | Up to 512 GB/s

GB200 NVL72 互联拓扑

- 每个 B200 有 18 个 NVLINK Port，9 x Switch Tray 共计 18 颗 NVLINK Switch 芯片，每个 B200 的 Port 连接一个 NVSwitch 芯片，Switch Tray 上单 NVSwitch 有 72 个 Port，刚好构成 NVL72，把 72 x B200 芯片全部连接起来。



思考

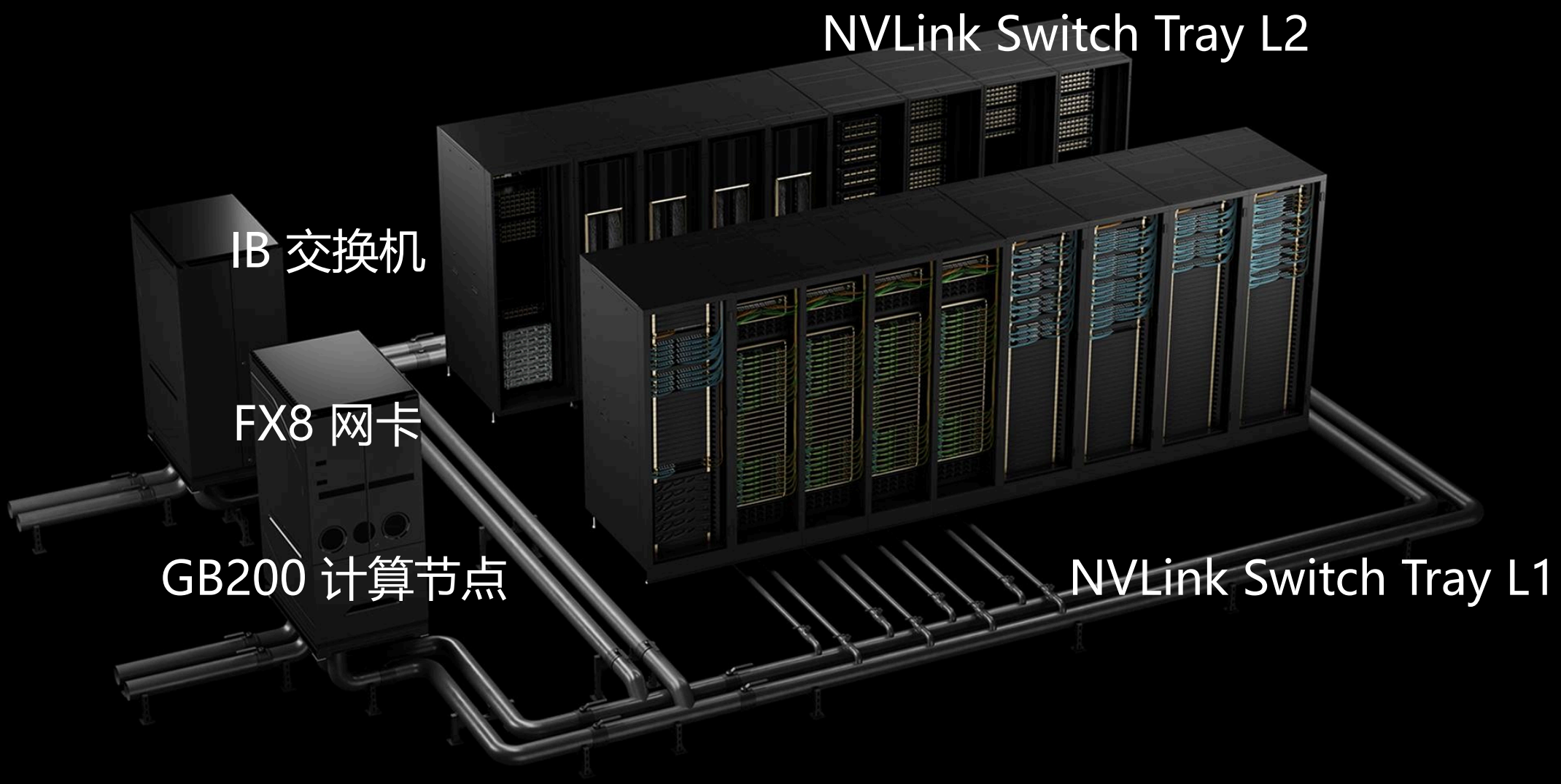
1. 在 NVL72 机柜中，所有 NVSwitch 交换机已经没有额外接口互联构成更大规模的两层交换集群。那么 NVL576（72 x 8）是怎么把 8 个 NVL72 链接起来的吗？

GB200 SuperPod

- Q：通过 Scale-Out RDMA 网络互联，而并不是通过 Scale-Up NVLINK 网络互联

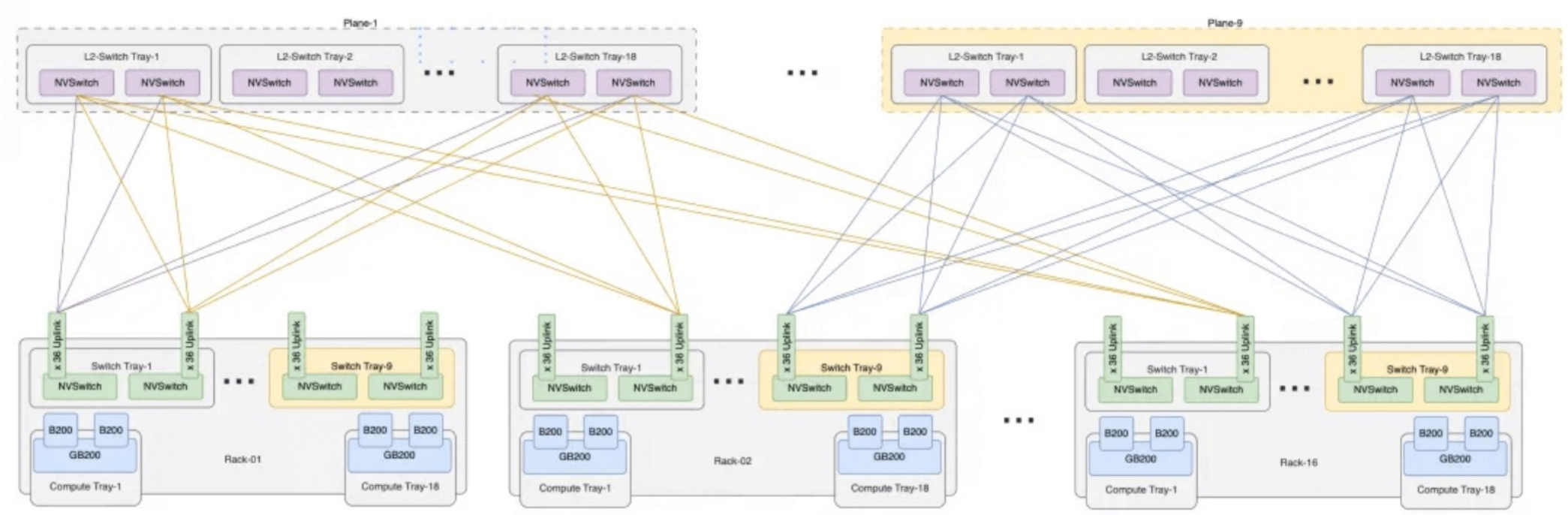
- 由 8 个 NVL72 组成，共 576 个 Blackwell GPU，要实现其全互联，和上一代 256 个 GH200 全互联类似两级 NVSwitch Tray：
  - L1 NVSwitch Tray 一半 Port 连接 576 个 Blackwell GPU，需要 576x18/(144/2)=144 个 NVSwitch Tray（剩余 144 * 72 个 Port）。
  - L2 NVSwitch Tray Port 全部与 L1 NVSwitch 剩余 Port 连接，需要 144x72/144=72 个 NVSwitch Tray。L2 每一个 NVSwitch Tray 都与 L1 所有 NVSwitch Tray 连接。

GB200 SuperPod



GB200 SuperPod 互联拓扑

- 为以后扩展到576卡集群，每个交换机都提供36个对外互联的Uplink，累计单个机柜有36 * 2 * 9 =648个上行端口，构成NVL576需要有16个机柜，则累计上行端口数为 648 * 16 = 10,368个，实际上可以由9个第二层交换平面构成，每个平面内又有36个子平面，由18个Switch Tray构成



思考

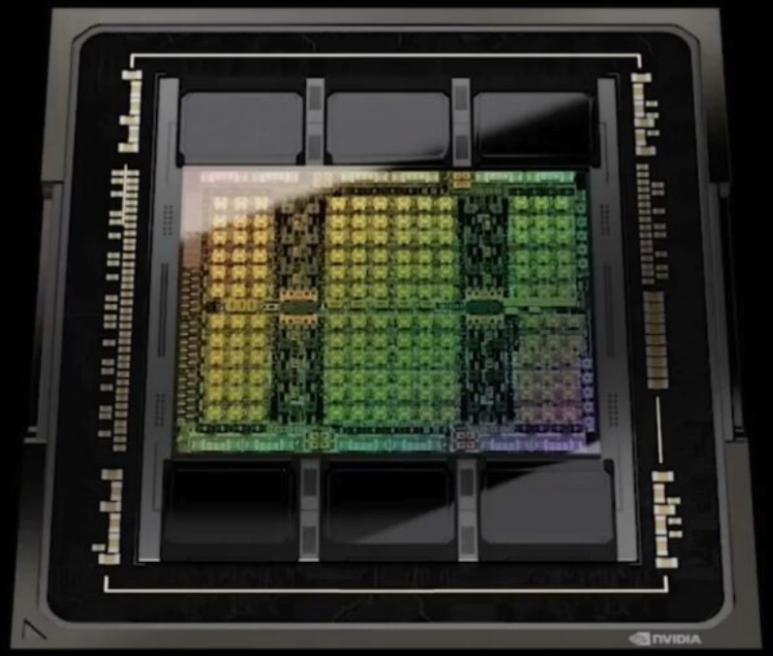
1. 从下一代大模型本身的算力需求来看，超节点已经成为 High Bandwidth Domain，当 Scale-Up 的 NVLink 网络规模增大时，实际 HBD 之间互联 RDMA 带宽带来的性能收益在减小。所以 SuperPod 的需求量到底有多大？到底会不会迎来思科时刻？

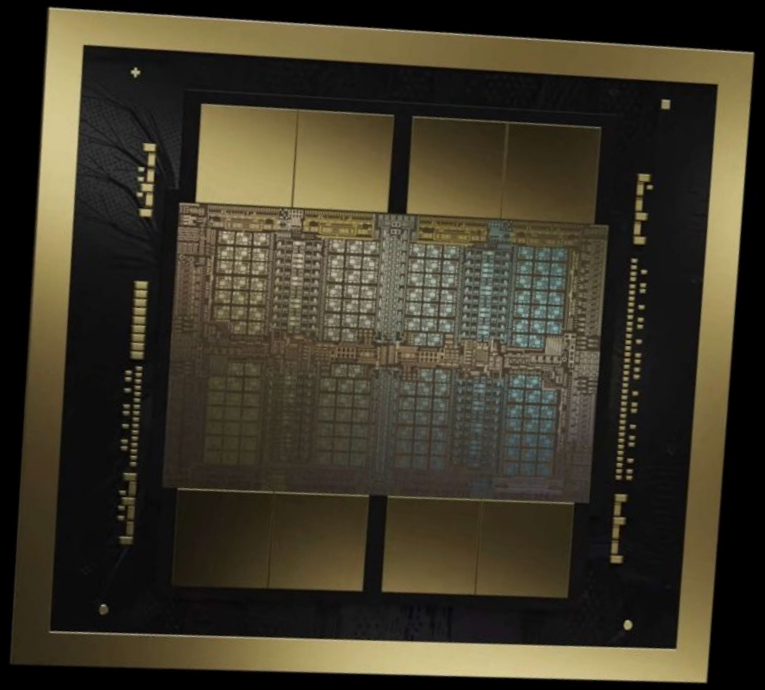
03

小结与思考

分析 1：制造工艺对性能提升有限

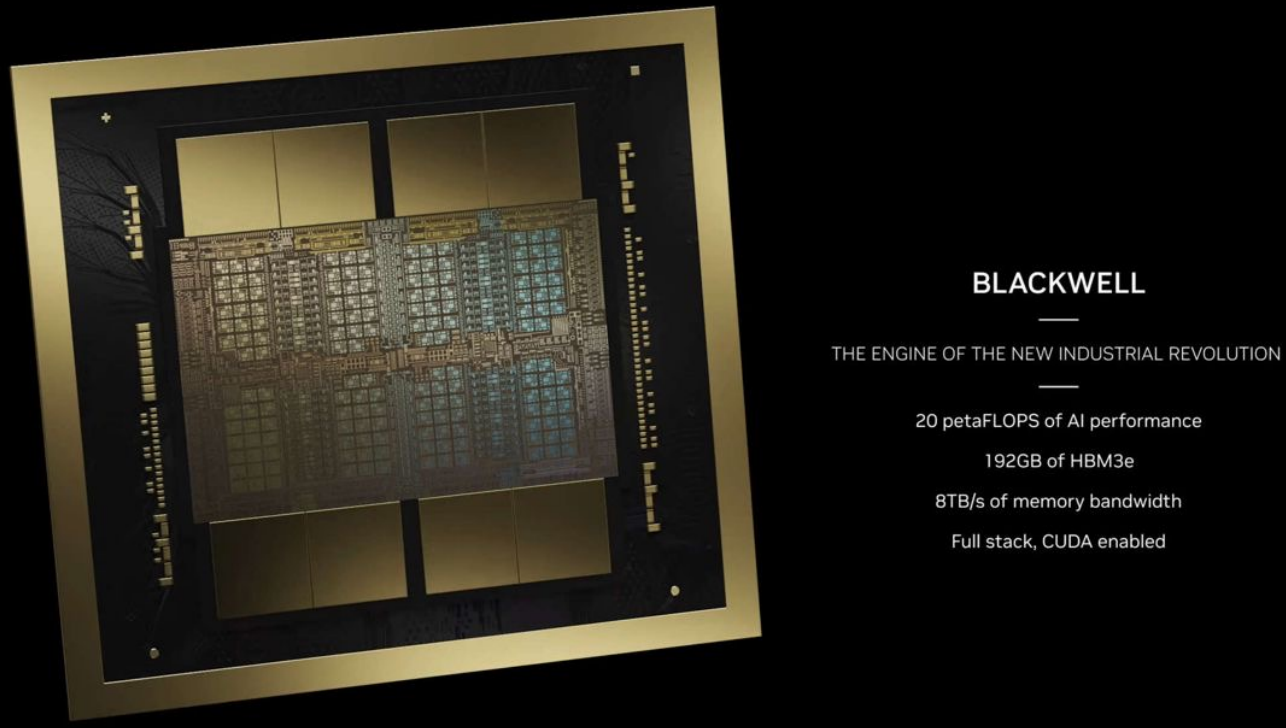
- 单 die GPU 性能并没有大幅提升，NV 代际间受先进工艺限制约束：
  1. 先进工艺代际升级时间长（ 2~3 年）；
  2. 代际间单芯片面积性能提升~15%，功耗提升35%；





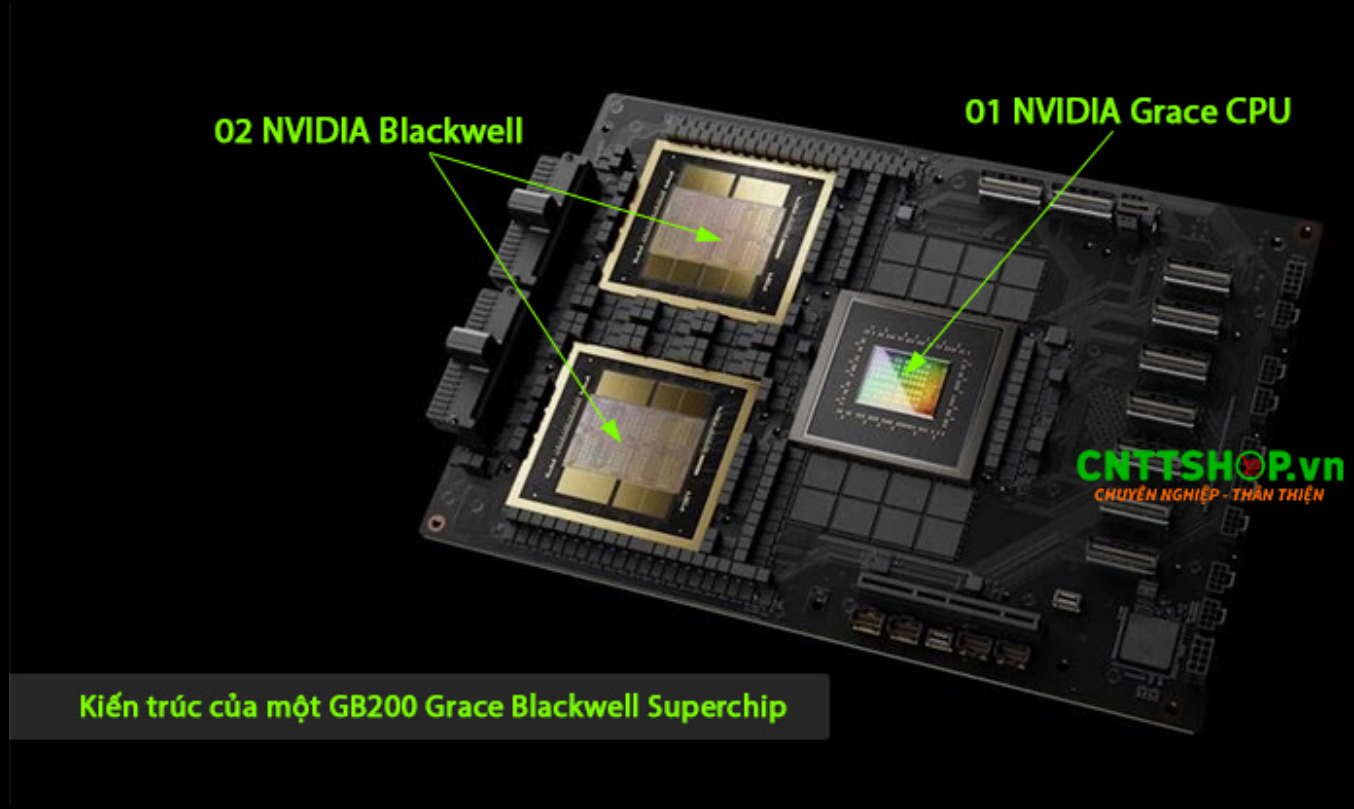
分析 2：封装能力提升解决工艺提升慢

- 拼封装技术的时代，从传统单 die 主处理芯片，演进到双 die 合封。
  - HBM 实现多 die 和多层合封技术，解决内存墙问题；
  - 多 die 合封实质上是在解决工艺发展受限；



分析 3：引入 Tray 板卡概念

- Nvidia 不仅使用封装来解决部件性能瓶颈，也运用板卡（Tray）概念。
  - Hopper 板卡支持 1CPU + 1 GPU；
  - Blackwell 板卡支持 1 CPU + 2 GPU，同制造工艺下算力增加 4 倍+， 单 GPU 包括两 die 主芯粒。



分析 4：网络带宽提升

- NVLink 交换机端口数和单端口带宽都增加了1倍，为更大规模 NVLink 组网提供支撑。
  - 单柜支持 72 GPU 组网（NVL72）， SuperPod 支持 576 GPU 组网（NVL576）；
  - 中等规模微调大模型，NVL72 + IB 可实现千卡互联更有性价比优势；
  - 大规模训练大模型，NVL576 + IB + Internet 可以组成万卡超级集群；

GB200 NVL72
- 36 GRACE CPUs
- 72 BLACKWELL GPUs
- Fully Connected NVLink Switch Rack

Training	720 PFLOPs
Inference	1,440 PFLOPs
NVL Model Size	27T params
Multi-Node Bandwidth	130 TB/s
Multi-Node All-Reduce	260 TB/s

分析 5：网络带宽提升弥补计算性能

- Serdes当前 224Gbps 继续提升到 336或 448
- 单芯片 NVLink 数量再提升 1 倍，如从现在 18 提升到 36
- 单柜集群实现 72 to 144 GPU 超级互联，计算密度继续提升1倍，带宽提升 4 倍。

GB200 NVL72	36 GRACE CPUs 72 BLACKWELL GPUs Fully Connected NVLink Switch Rack
Training	720 PFLOPs
Inference	1,440 PFLOPs
NVL Model Size	27T params
Multi-Node Bandwidth	130 TB/s
Multi-Node All-Reduce	260 TB/s

分析 6：

- 交换机端口数更多，以支持更大规模集群。交换机端口数量越多越能减少层级数，这不仅可以减少网络跳数，还可以提升单柜的计算密度。
- NV 未来还会享受先进工艺的红利，可以持续发展多年。并且，光通信这-块的能力，还有待进步释放，包括芯片内光互联，芯片直接出光等。

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
2. https://www.facebook.com/photo.php?fbid=1048191653980302&id=100063684318606&set=a.747927134006757