

# 从Blackwell到Rubin：计算、网络、存储持续升级 ——AI算力“卖水人”专题系列（7）

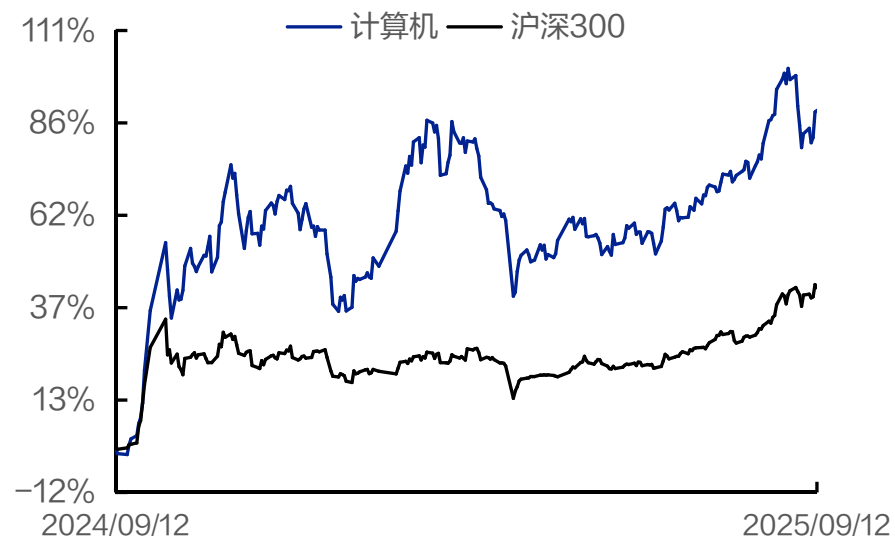
评级：推荐(维持)

刘熹(证券分析师)  
S0350523040001  
liux10@ghzq.com.cn

# 并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

## 最近一年走势



## 相关报告

《从Tokens角度跟踪AI应用落地进展——计算机行业大模型及AI应用专题（推荐）\*计算机\*刘熹》——2025-09-15

《计算机事件点评：甲骨文RPO增至4550亿美元，AI算力强力增长（推荐）\*计算机\*刘熹》——2025-09-12

《液冷：AI算力新一极——AI算力“卖水人”专题系列（6）（推荐）\*计算机\*刘熹》——2025-08-17

## 相对沪深300表现

| 表现    | 1M   | 3M    | 12M   |
|-------|------|-------|-------|
| 计算机   | 7.4% | 23.5% | 97.4% |
| 沪深300 | 8.3% | 17.6% | 44.1% |

◆ 本篇报告解决了以下核心问题：1、英伟达Blackwell和Rubin系列芯片技术和参数更新；2、英伟达算力在计算/网络/存储/液冷等各方面进展；3、AI算力行业投资建议。

◆ 一、GPU核心：GB300计算能力大幅提升，Rubin预期乐观

GB300基于Blackwell Ultra架构，采用TSMC 4NP工艺与CoWoS-L封装，FP4浮点算力可达15PFLOPS，是B200的1.5倍。GB300搭载288GB的HBM3E显存。Vera Rubin NVL144性能是GB300 NVL72的3.3倍。Rubin Ultra NVL576性能较GB300 NVL72提升14倍，内存是GB300的8倍。FY2026Q2，英伟达营收467亿美元，同比+56%。

◆ 二、服务器细节拆分：主板从HGX到MGX，Rubin Ultra NVL576架构升级

GB300 NVL72系统由18个计算托盘和9个交换机托盘组成，搭载72颗Blackwell Ultra GPU与36颗Grace CPU。与Hopper相比，GB300 NVL72的AI工厂总体产出性能最多提升50倍的潜力。GB300 NVL72集成ConnectX-8网卡与BlueField-3 DPU。Rubin Ultra NVL576将于2027年推出，采用Kyber架构，大幅提升机架密度。

◆ 三、网络：CPO取代可插拔光模块，Rubin实现超高速互联

CPO将硅光子器件与ASIC封装，取代传统的可插拔光模块，与传统网络相比，提升能效3.5倍、部署速度1.3倍。Quantum-X和Spectrum-X交换机减少了对传统光收发器的依赖，为超大规模人工智能工厂提供高达400Tbps的吞吐量。Rubin将采用NVLink 6.0技术，速度翻倍至3.6TB/s。NVLink Fusion向第三方开放互联生态，支持异构芯片协同。新一代NVSwitch 7.0扩展至576颗GPU互联，实现非阻塞通信，实现更大规模的GPU互联。

◆ 四、HBM：HBM4预计2026年实现量产，SK海力士主导HBM市场

2025年3月，SK海力士率先向主要客户交付了全球首款12层堆叠HBM4样品。三星已准备在2025年7月底前向AMD和英伟达等客户提供HBM4样品。美光在其AI内存产品组合中展示了下一代HBM4的计划，预计将在2026年实现量产。SK海力士已与NVIDIA、微软与博通展开合作，设计客户专属的定制化HBM芯片。

◆ 五、冷板式液冷较为成熟，GB300 NVL72采用全液冷方案

GB300采用了独立液冷板设计，每个芯片配备单独的一进一出液冷板，而非大面积冷板覆盖方式。液冷技术具备更高散热效率，包括冷板式与浸没式两类，其中冷板式为间接冷却，初始投资中等，运维成本较低，相对成熟，英伟达GB300 NVL72采用全液冷设计。

- ◆ **投资建议：**大模型训推带动AI算力需求增长，GB300、Vera Rubin等新一代算力架构将推出，算力产业链中的AI芯片、服务器整机、铜连接、HBM、液冷、光模块、IDC等环节有望持续受益。维持对计算机行业“推荐”评级。
  
- ◆ **相关公司**
  - 1) **AI芯片：**寒武纪、海光信息、寒武纪、龙芯中科、景嘉微、英伟达、AMD、Intel
  - 2) **服务器整机：**工业富联、浪潮信息、中科曙光、华勤技术、中国长城、高新发展、神州数码、烽火通信、拓维信息、纬创、广达、英业达、纬颖、超微电脑。
  - 3) **服务器组件：**①散热：曙光数创、英维克、飞荣达、申菱环境、高澜股份、川环科技、奇鋆科技、双鸿、VERTIV；②主板：沪电股份、胜宏科技、深南电路、技嘉、华擎；③HBM：SK海力士、三星、美光、赛腾股份、联瑞新材；④铜连接：安费诺、沃尔核材、华丰科技。
  - 4) **光模块：**天孚通信、中际旭创、新易盛、光迅科技、华工科技。
  - 5) **数据中心：**奥飞数据、光环新网、宝信软件、数据港、电科数字、云赛智联、润泽科技。
  - 6) **算力租赁：**协创数据、有方科技、宏景科技、安诺其。
  
- ◆ **风险提示：**宏观经济影响下游需求、大模型产业发展不及预期、AI芯片产业发展不及预期、市场竞争加剧、中美博弈加剧、相关公司业绩不及预期、各公司并不具备完全可比性，对标的相关资料和数据仅供参考。

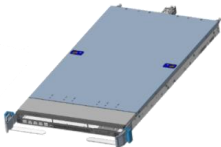
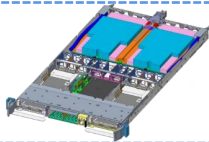


# 投资建议与相关公司

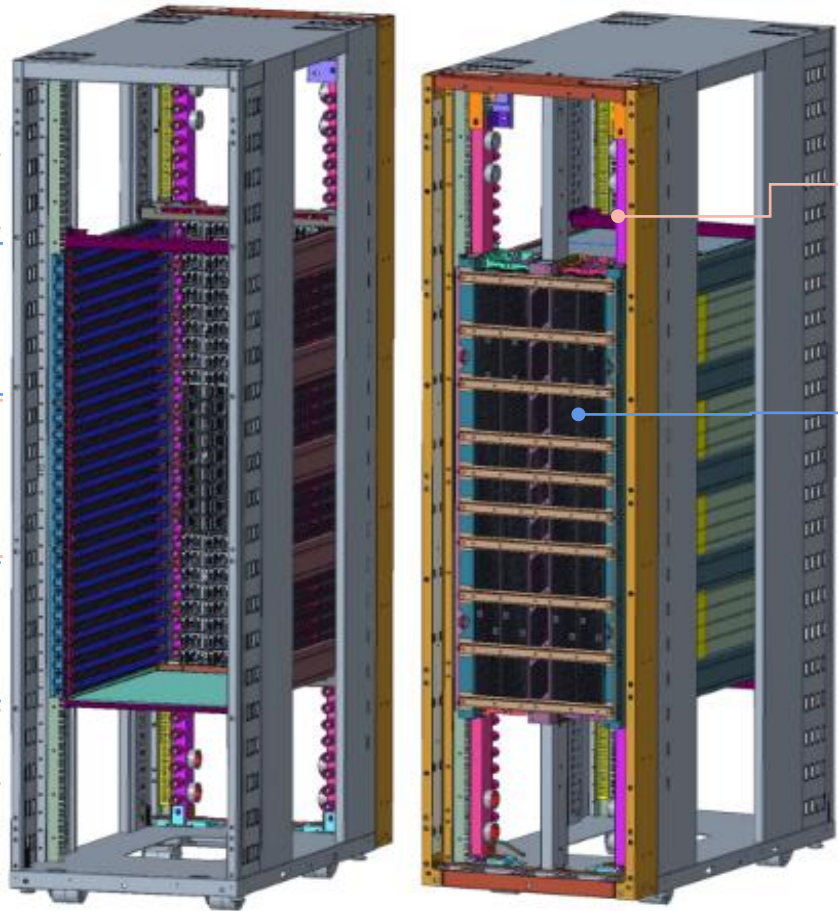


英伟达GB200 NVL72实物图

- **电源**
  - 内地：中国长城、欧陆通、杰华特、泰嘉股份
  - 中国台湾：台达
- **☆GPU**
  - 内地：华为海思、海光信息、寒武纪、龙芯中科。
  - 海外：英伟达、AMD、Intel
- **☆CPU**
  - 内地：华为海思、海光信息、飞腾、龙芯中科
  - 海外：英伟达、AMD、Intel
- **☆主板**
  - 内地：沪电股份、胜宏科技、深南电路、生益科技
  - 中国台湾：泰安电脑（神达）、技嘉、华擎、华硕、英业达、纬创、研华
  - 美国：Intel、Supermicro
- **Compute Tray \* 10**
  - 内地：工业富联
  - 中国台湾：纬创
- **Switch Tray \* 9**
  - 内地：工业富联
  - 中国台湾：纬创
- **Compute Tray \* 8**
- **NV Switch Chip\网卡等**
  - 海外：英伟达、Mellanox
- **电源**



- **存储 NAND**
  - 公司：长江存储、兆易创新、佰维存储、朗科科技
  - 海外：三星、西部数据、铠侠、SK海力士、美光



正面

背面

- **光模块**
  - 内地：中际旭创、新易盛、光迅科技、天孚通信

- **☆分歧管（液冷：支持130KW的制冷能力）**
  - 内地：曙光数创、飞荣达、中航光电、英维克、同飞股份、申菱环境、高澜股份、依米康
  - 中国台湾：奇鋐、双鸿、健策、建准、高力
  - 海外：VERTIV

- **☆线缆（铜连接：5000+根NVLink线缆）**
  - 内地：沃尔核材、华丰科技、兆龙互连、鼎通科技、立讯精密、神宇股份
  - 海外：安费诺

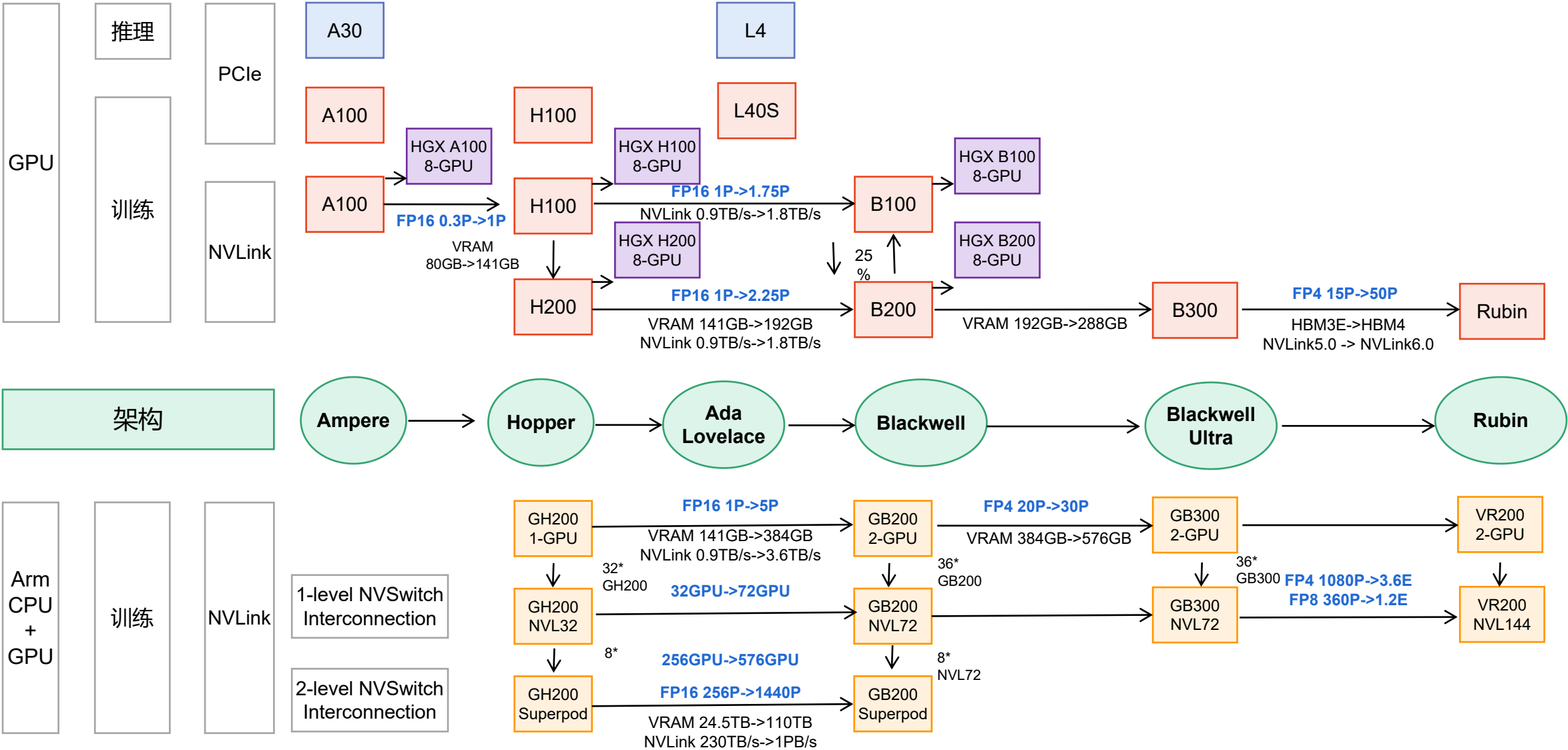
- **☆组装测试**
  - 整机厂商ODM：1) 内地：工业富联、中科曙光、浪潮信息、紫光股份、中国长城、华勤技术、神州数码、拓维信息、烽火通信、软通动力；2) 中国台湾：鸿海、广达、纬创、英业达、纬颖；3) 美股：戴尔、超微电脑
  - BIOS/BMC：1) 内地：卓易信息；2) 中国台湾：新唐、系微、信骅

注：蓝字为零部件。上图展示以英伟达GB200 NVL72整机柜为例，仅列示了部分参与公司

# 第一章 GPU核心

## GB300计算能力大幅提升，Rubin预期乐观

# 1.1 NVIDIA每一代GPU的计算能力、NVLink、内存持续扩大





# 1.1 NVIDIA每一代GPU的计算能力、NVLink、内存持续扩大

- 英伟达 Blackwell Ultra凭借节能的双光罩（dual-reticle）设计、高带宽大容量 HBM3E 内存子系统、第五代 Tensor 核心以及突破性的 NVFP4 精度格式，正在提升加速计算的新标准。

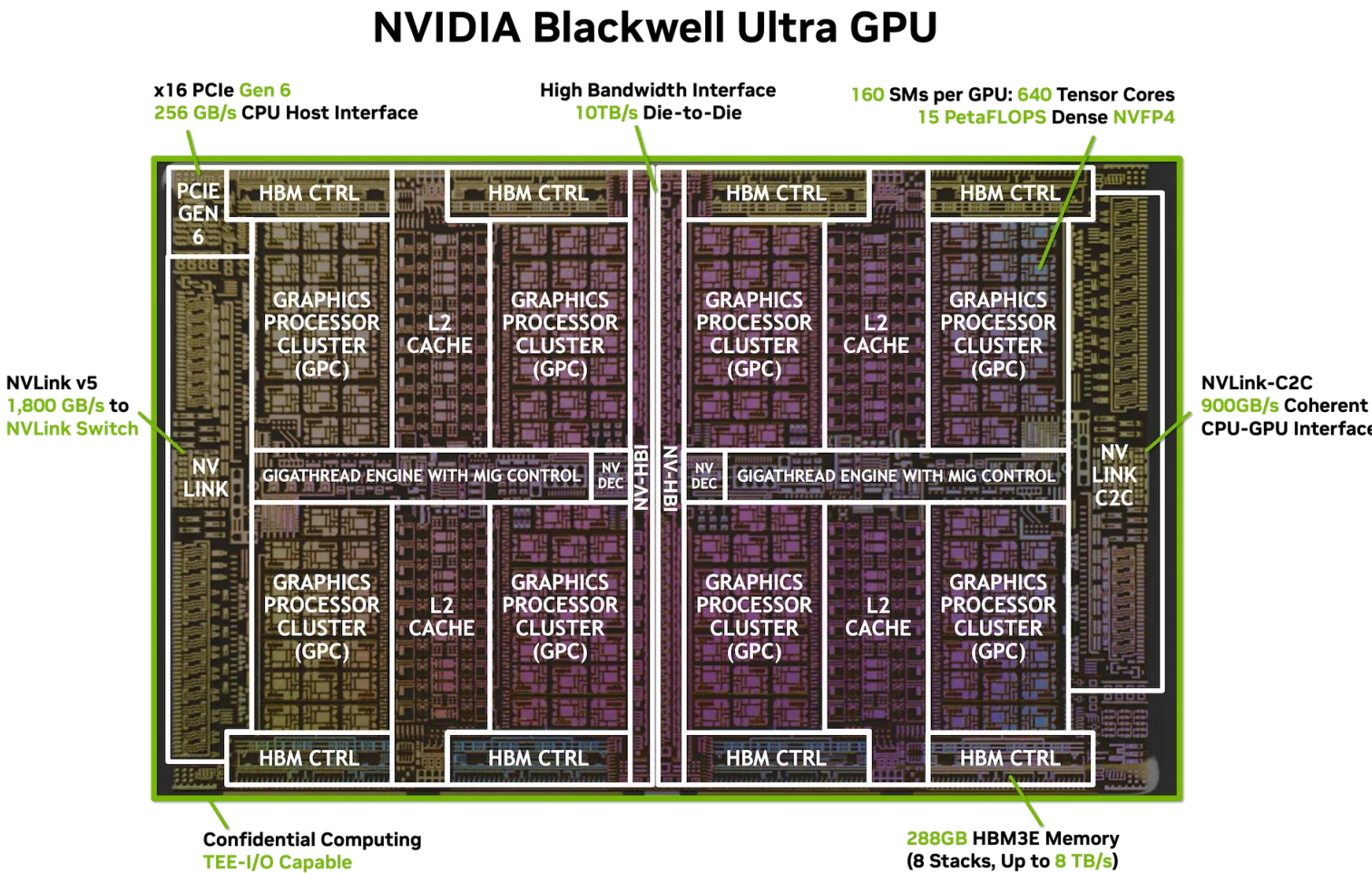
表：英伟达GPU迭代情况

| 2022                       |                       | 2023        | 2024                  |           | 2025         | 2026                                  | 2026                  | 2027                                  |
|----------------------------|-----------------------|-------------|-----------------------|-----------|--------------|---------------------------------------|-----------------------|---------------------------------------|
|                            |                       | Hopper      |                       | Blackwell |              | Rubin                                 |                       |                                       |
| Accelerator                | H100(SXM)             | H200        | B200                  | GB200     | GB300(Ultra) | VR200                                 | CPX                   | VR300(Ultra)                          |
| GPU TDP (W)                | 700                   | 700         | 700                   | 1200      | 1400         | 2300                                  | 800                   | 4000+                                 |
| Foundry Node               | 4NP                   |             | 4NP                   |           |              | N3P(3NP)                              | N3P(3NP)              | N3P(3NP)                              |
| Logic Die Configuration    | 1 x Reticle Sized GPU |             | 2 x Reticle Sized GPU |           |              | 2 x Reticle Sized GPU, 2x I/O chiplet | 1 x Reticle Sized GPU | 4 x Reticle Sized GPU, 4x I/O chiplet |
| Memory                     | 80GB HBM3             | 141GB HBM3E | 192GB HBM3E           |           | 288GB HBM3E  | 288GB HBM4                            | 128GB GDDR7           | 1024GB HBM4E                          |
| HBM Stacks                 | 5                     | 6           | 8                     |           |              | 8                                     | N/A                   | 16                                    |
| Memory Bandwidth           | 3.35TB/s              | 4.8TB/s     | 8TB/s                 |           |              | 20.5TB/s                              | 2TB/s                 | 53TB/s                                |
| NVLink                     | 900GB/s               | 900GB/s     | 1.8TB/s               | 2*1.8TB/s | 2*1.8TB/s    | 3.6TB/s                               |                       | 10.8TB/s                              |
| FP16(FLOPS)                | 1P                    | 1P          | 2.2P                  | 5P        |              |                                       |                       |                                       |
| INT8(FLOPS)                | 2P                    | 2P          | 4.5P                  | 10P       |              |                                       |                       |                                       |
| FP8(FLOPS)                 | 2P                    | 2P          | 4.5P                  | 10P       |              |                                       |                       |                                       |
| FP6(FLOPS)                 | X                     | X           | 4.5P                  | 10P       |              |                                       |                       |                                       |
| FP4(FLOPS)                 | X                     | X           | 9P                    | 20P       |              |                                       | 20P                   |                                       |
| Packaging                  | CoWoS-S               |             | CoWoS-L               |           |              | CoWoS-L                               | FC-BGA                | CoWoS-L                               |
| SerDes speed (Gb/s uni-di) | 112G                  |             | 224G                  |           |              | 224G                                  | 64G (PCIe Gen6)       | 224G                                  |
| Nvidia CPU                 | Grace                 |             |                       |           |              | Vera                                  | Vera                  | Vera                                  |

# 1.1 B300: Blackwell Ultra为TSMC 4NP工艺，内存容量扩大

- **B300 GPU:**
- **计算性能:** 基于新一代Blackwell Ultra架构，采用TSMC的4NP定制工艺，搭配NVIDIA的CoWoS-L封装技术。FP4浮点算力可达15PFLOPS，比B200的FLOPS提升50%，部分性能提升将来自200W的额外功耗。
- **内存容量:** 从8-Hi升级到了12-Hi HBM3E，使得每个GPU的内存容量增加到了288GB，不过由于引脚速率保持不变，单GPU的内存带宽仍维持在8TB/s。
- **高效网路互联:** 引入ConnectX-8 NIC，该NIC提供4个200G通道，为InfiniBand实现800G的总吞吐量，与当前的Blackwell CX-7 NIC相比，网络速度提高了一倍。

图：英伟达Blackwell Ultra GPU 芯片说明



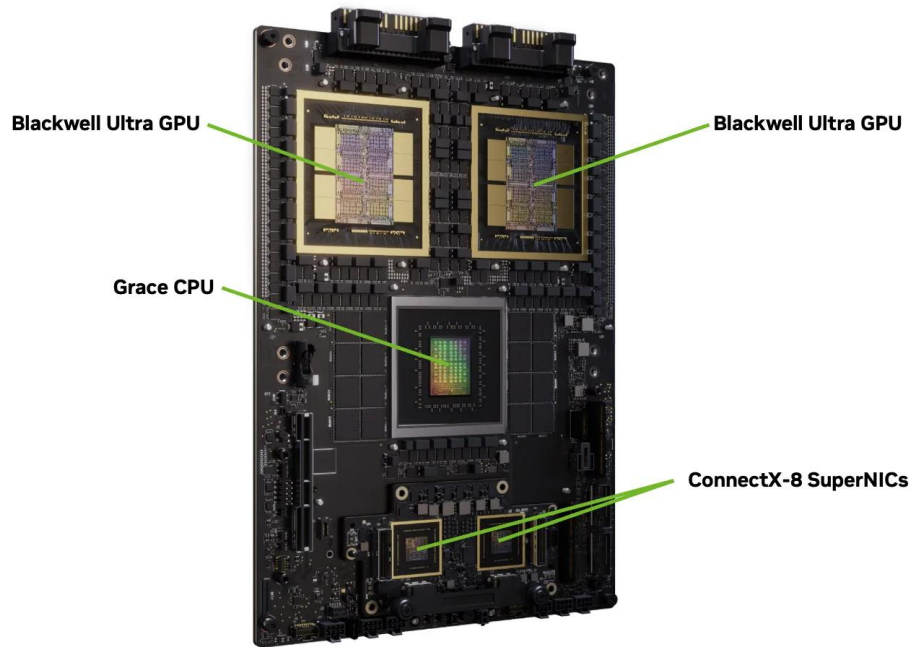
# 1.1 GB300：采用HBM3E技术，FP4的计算能力实现1.5倍的增长

- GB300性能：**集成更强大的B300芯片并采用12层堆叠HBM3E显存，总容量288 GB（较前代192 GB大幅提升），带宽保持8 TB/s；FP4计算能力相比前代产品提升1.5倍，显著增强AI大规模浮点运算性能。
- GB300结构：**通过NVLink-C2C将一个Grace CPU与两个Blackwell Ultra GPU连接，支持NVLink 5.0，每个GPU双向带宽1.8 TB/s；内置ConnectX-8 SuperNIC，提供800 Gb/s高速网络并拥有48条PCIe通道；支持新的架构如空气冷却MGX B300A。对于GB300，Nvidia不再提供完整的Bianca板，而是转而供应B300作为SXM Puck模块，Grace CPU则以BGA封装形式提供。转向SXM Puck为更多OEM和ODM参与计算托盘提供了机会。

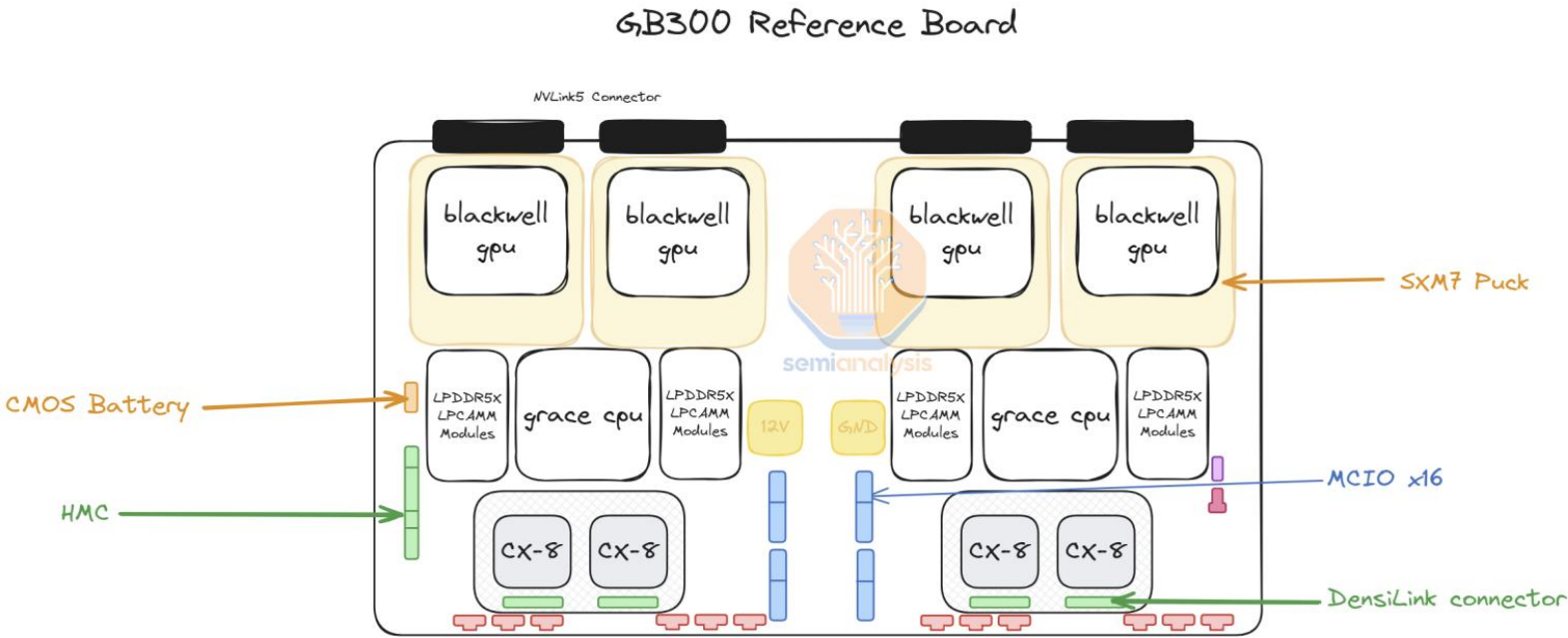
表：Hopper与Blackwell及Blackwell Ultra的互联带宽对比（双向，单位：GB/s）

| Interconnect         | Hopper GPU  | Blackwell GPU | Blackwell Ultra GPU |
|----------------------|-------------|---------------|---------------------|
| NVLink (GPU-GPU)     | 900         | 1,800         | 1,800               |
| NVLink-C2C (CPU-GPU) | 900         | 900           | 900                 |
| PCIe Interface       | 128 (Gen 5) | 256 (Gen 6)   | 256 (Gen 6)         |

图：英伟达GB300实物图



图：英伟达GB300结构图

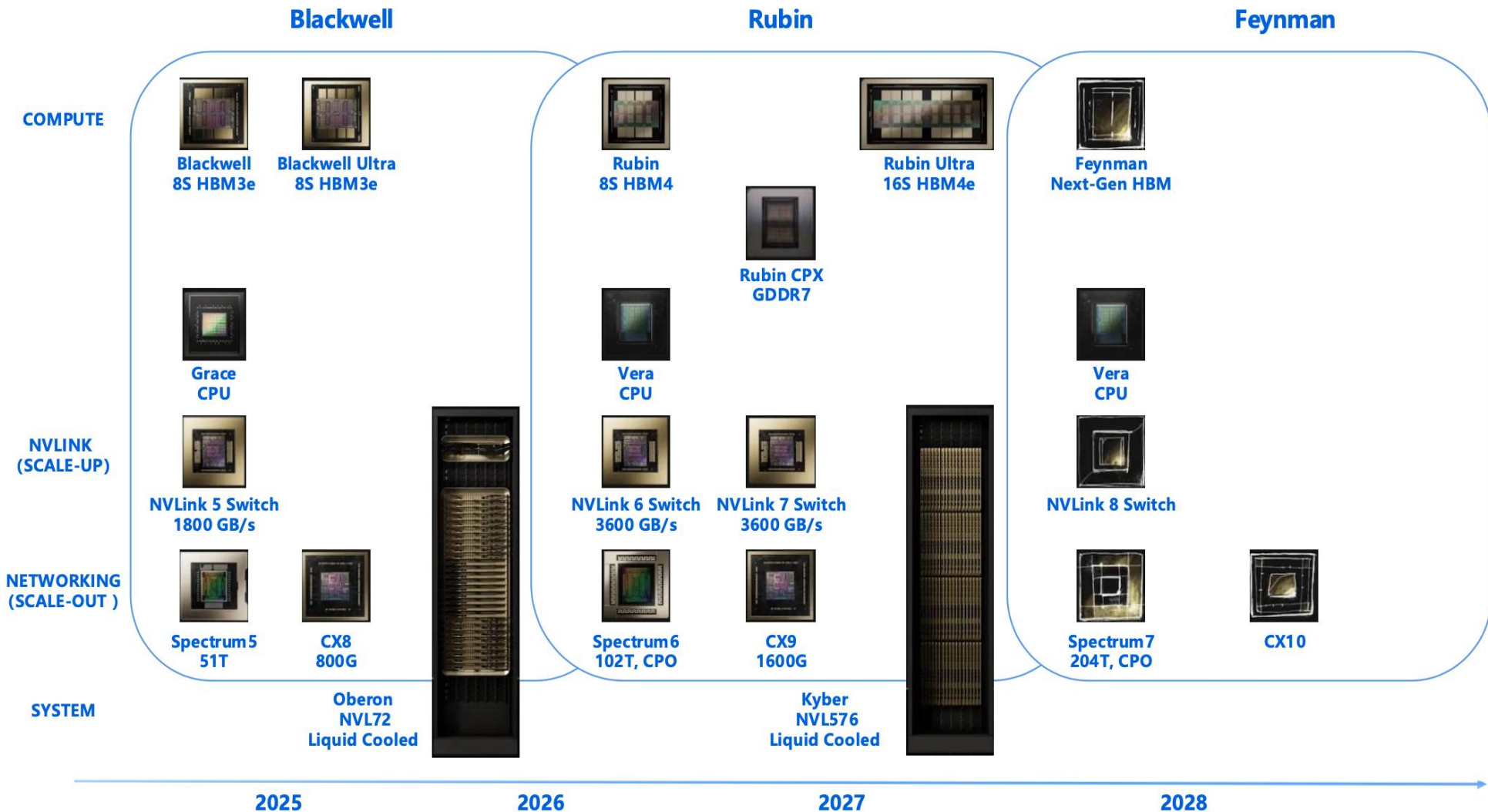




# 1.2 2026年推出Rubin架构，继任者Feynman于2028年登场

- 2026年将推出Rubin架构，基于Rubin的Vera Rubin NVL 144机柜由72颗Vera CPU和144颗Rubin GPU组成。
- 2027年推出Rubin Ultra产品，基于Rubin Ultra的Rubin Ultra NVL 576配备576颗Rubin Ultra GPU组成。
- 2028年将推出Feynman架构。

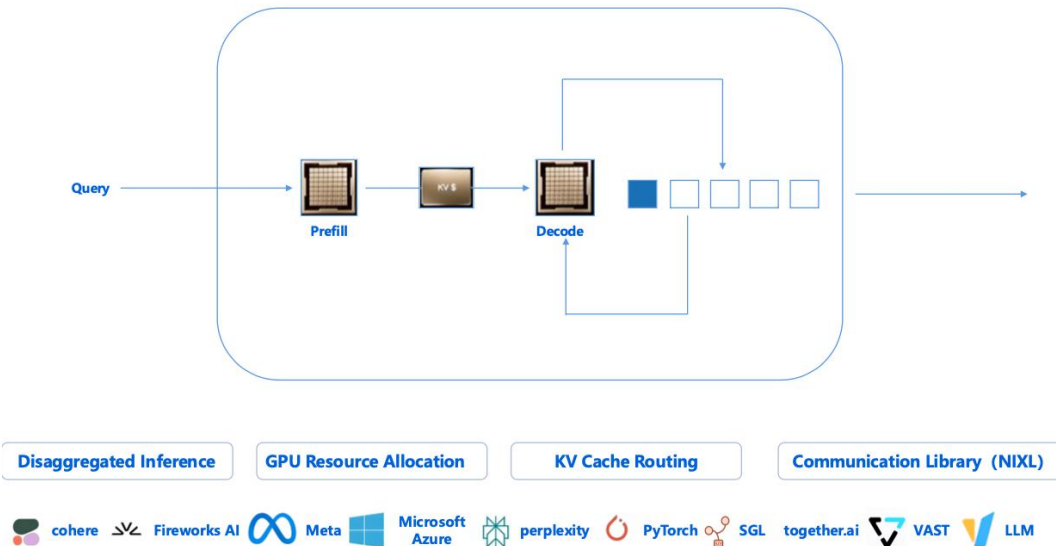
图：英伟达GPU和技术路线图



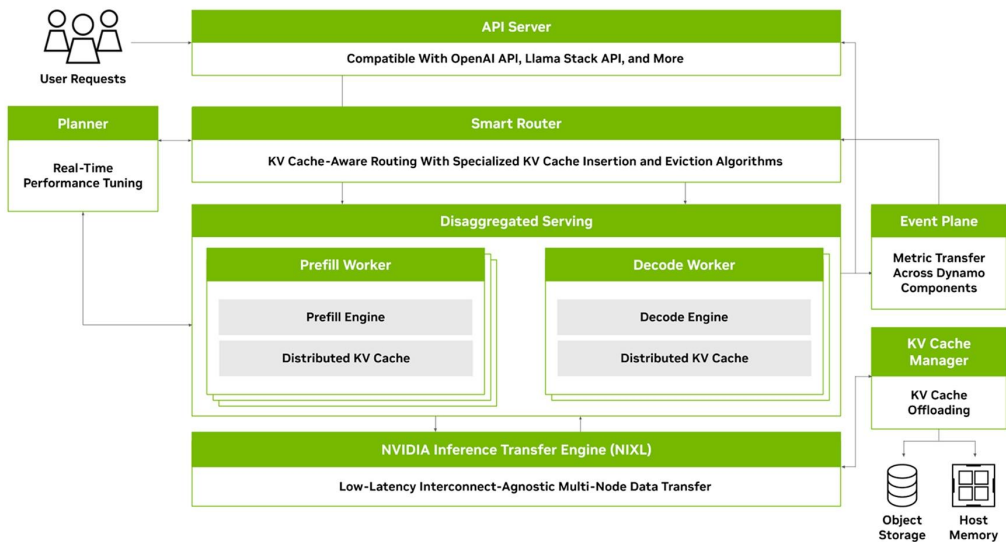
# 1.3 NVIDIA Dynamo：提升多GPU推理效率，AI工厂的操作系统

- **NVIDIA Dynamo**——开源、低延迟的模块化推理框架，在分布式环境中服务生成式 AI 模型。Dynamo可在多个GPU节点之间扩展推理任务，并动态分配GPU工作线程，以缓解流量瓶颈。Dynamo还支持解耦式推理服务，可将大型语言模型推理中的上下文预填（prefill）与生成（decode）阶段在多个GPU间分离执行，以优化性能、提升可扩展性并降低成本。
- NVIDIA Dynamo解决了分布式和分解推理服务的挑战。它包括四个关键组件：
  - ✓ GPU资源规划器：一个规划和调度引擎，用于监控多节点部署中的容量和预填充活动，以调整GPU资源，并在预填充和解码之间分配这些资源。
  - ✓ 智能路由：KV缓存感知路由引擎，可在多节点部署中高效引导大型GPU集群中的传入流量，从而最大限度地减少昂贵的重新计算。
  - ✓ 低延迟通信库：先进的推理数据传输库，可加速GPU之间以及异构内存和存储类型之间的KV缓存传输。
  - ✓ KV缓存管理器：成本感知型KV缓存卸载引擎，旨在跨各种内存层次结构传输KV缓存，在保持用户体验的同时释放宝贵的GPU内存。
- Dynamo的智能推理优化可将每个GPU生成的token数量提高30倍以上。基于Dynamo，相比Hopper，Blackwell性能提升25倍，可以基于均匀可互换的可编程架构。在推理模型中，Blackwell性能是Hopper的40倍。

图：Dynamo的系统架构



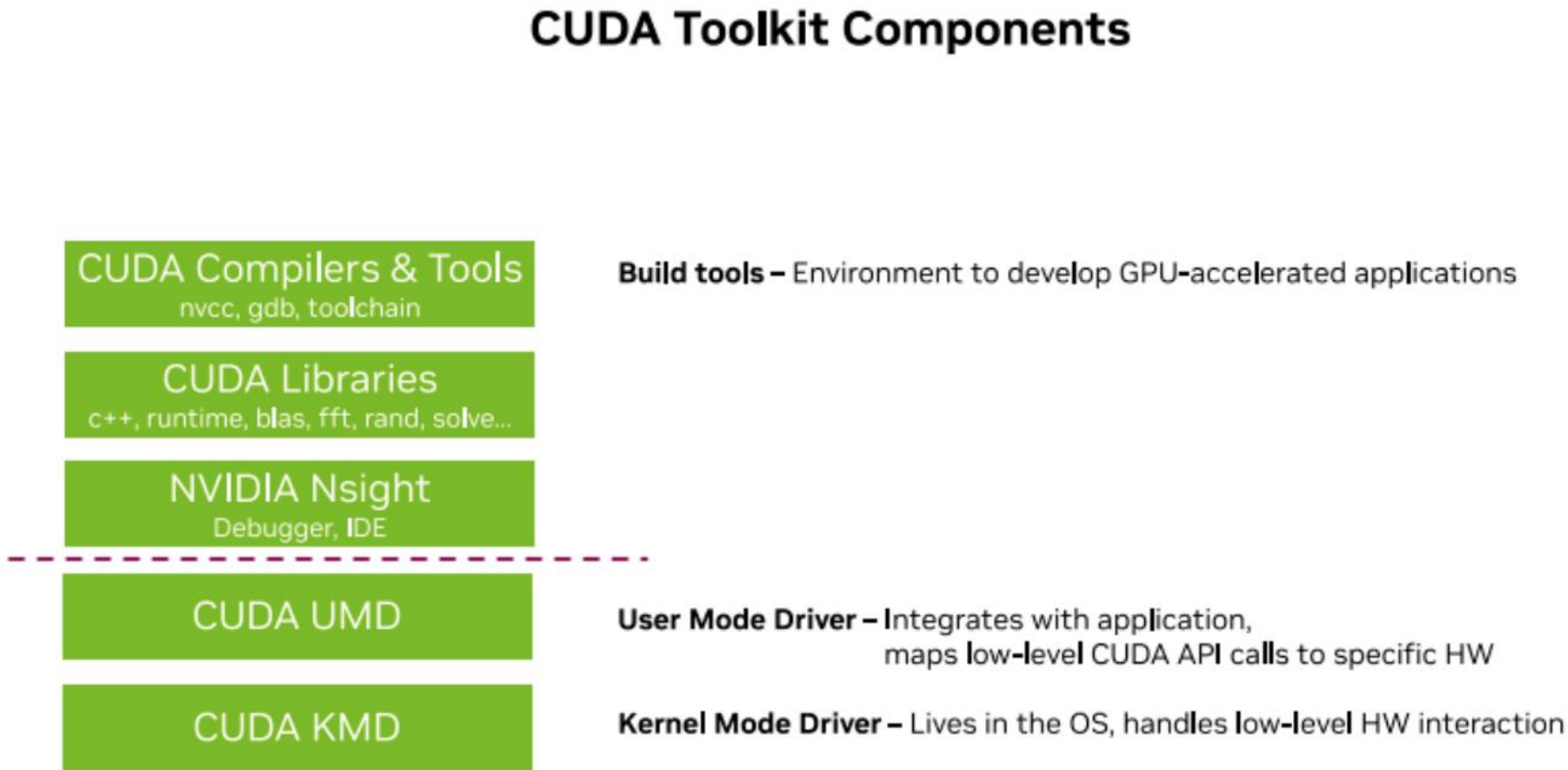
图：Dynamo组成结构图



# 1.4 CUDA：NVIDIA的GPU并行计算生态与架构

- **CUDA**是NVIDIA开发的一个并行计算平台和编程模型，用于在GPU上进行通用计算。借助CUDA，开发者能够利用GPU的强大功能显著加快计算应用程序的运行速度。
- **CUDA**主要包含两个关键组成部分：1 ) **Toolkit**：作为编译器，负责高效的编译功能 2 ) **驱动器**：支撑CUDA在硬件上的运行。

图：CUDA Toolkit 的主要组件

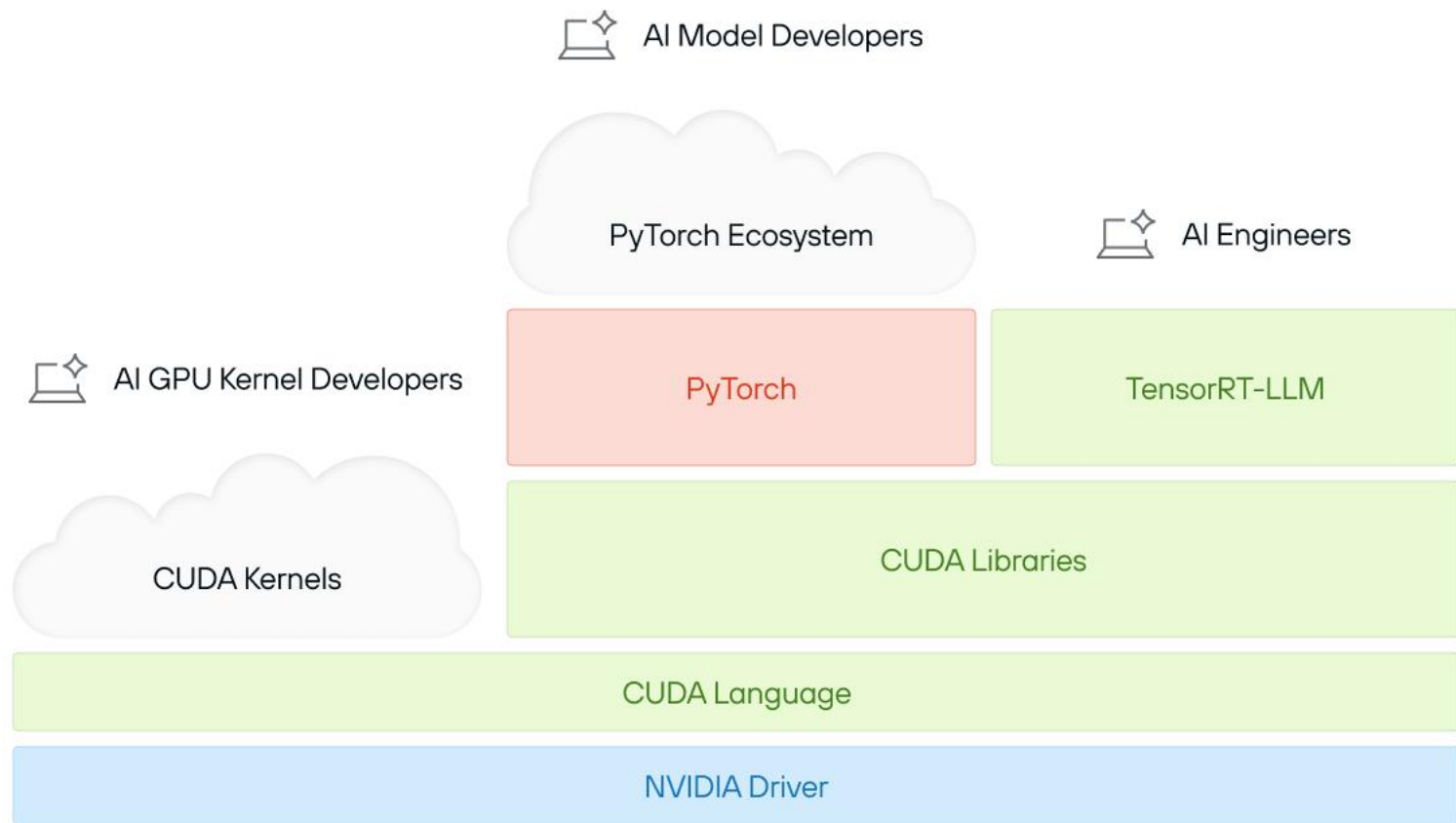




## 1.4 CUDA: NVIDIA的GPU并行计算生态与架构

- CUDA的复杂性不断扩大，涵盖驱动程序、语言、库和框架的多层生态系统。
- CUDA平台的核心包括：
- **庞大的代码库：**数十年优化的GPU软件涵盖从矩阵运算到AI推理的所有领域。
- **庞大的工具和库生态：**从用于深度学习的cuDNN到用于推理的TensorRT，CUDA涵盖了广泛的工作负载。
- **硬件调整性能：**每个CUDA版本都针对NVIDIA最新的GPU架构进行深度优化，确保顶级效率。
- **专有且不透明：**当开发者与CUDA的库API交互时，底层发生的很多操作都是闭源的，并且与英伟达的生态系统紧密相连。

图：CUDA生态系统的分层堆栈



## 1.4 CUDA：Blackwell 与 CUDA-X 驱动半导体制造革新

- 台积电、Cadence 、KLA、西门子和Synopsys正在采用 NVIDIA CUDA-X和NVIDIA Blackwell平台推动半导体制造的发展。
- NVIDIA Blackwell GPU、NVIDIA Grace CPU、高速NVIDIA NVLink结构和交换机以及特定领域的NVIDIA CUDA-X库（如NVIDIA cuDSS和NVIDIA cuLitho）正在改进先进芯片制造的计算光刻和设备模拟。

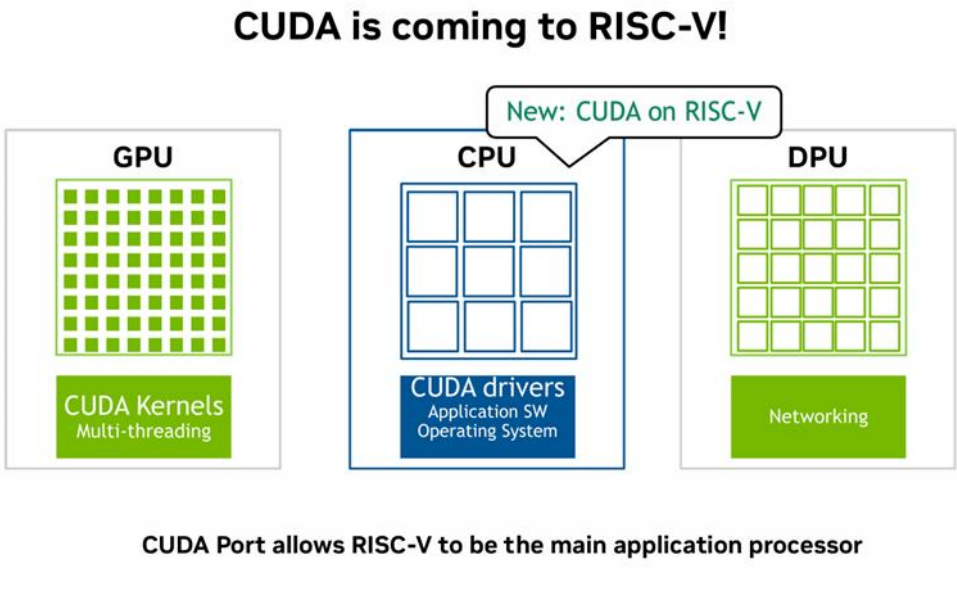
表：NVIDIA Blackwell 与 CUDA-X 推动半导体行业合作进展一览

| 项目       | 详细内容                                                                                                                                                                                                                                                                         |
|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 台积电      | NVIDIA cuLitho和Blackwell将光刻速度提升高达25倍。GPU加速使领先的光刻供应商和台积电等半导体制造商能够以前所未有的速度在生产前预测和纠正光刻问题。                                                                                                                                                                                       |
| Cadence  | 2025年5月初，Cadence宣布推出Millennium M2000 平台，该平台专为EDA市场打造，基于NVIDIA Blackwell架构。M2000是一款可扩展的交钥匙解决方案，可部署NVIDIA Grace Blackwell和CUDA-X库，并配备全套Cadence加速设计工具。                                                                                                                          |
| KLA      | KLA 期待针对特定市场评估带有CUDA-X库的NVIDIA RTX PRO 6000 Blackwell服务器版，以进一步加速为半导体芯片制造过程提供动力的推理工作负载。                                                                                                                                                                                       |
| 西门子      | 西门子正在利用NVIDIA CUDA-X库的并行处理能力和Grace Blackwell平台的突破性性能来显著加速其Calibre平台。                                                                                                                                                                                                         |
| Synopsys | Synopsys正在将NVIDIA CUDA-X库和Blackwell用于其EDA工具。通过集成CUDA-X库，Synopsys在NVIDIA B200上对Sentaurus Device、QuantumATK和S-Litho取得了新的基准测试结果，与同类CPU基础架构相比，性能分别提升了12 倍、15倍和20倍。此外，Synopsys最近在NVIDIA GTC 上宣布，他们预计Synopsys PrimeSim在NVIDIA Blackwell平台上的运行速度将提高30倍，Synopsys Proteus的运行速度将提高20倍。 |

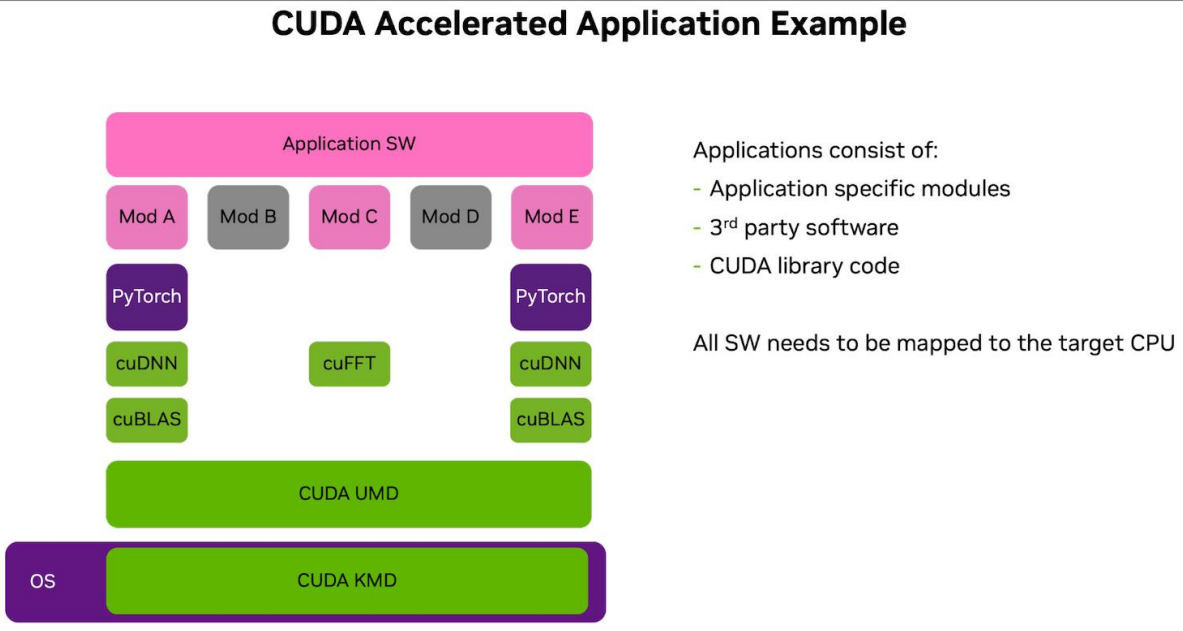
# 1.4 CUDA：支持RISC-V，生态系统拓展至开源领域

- 在2025年RISC-V峰会上，英伟达宣布其CUDA软件平台将在CPU方面兼容RISC-V指令集架构(ISA)。
- 会议上展示的图表展示了一种典型的配置：
  - ✓ GPU处理并行工作负载，而RISC-V CPU执行CUDA系统驱动程序、应用程序逻辑和操作系统。这种设置使CPU能够完全在CUDA环境中协调GPU计算。
  - ✓ 图中还展示了一个处理网络任务的DPU，完善了由GPU计算、CPU编排和数据移动组成的系统。这种配置清晰地展现了NVIDIA构建异构计算平台的愿景，其中RISC-V CPU可以作为管理工作负载的核心，而Nvidia的GPU、DPU和网络芯片则负责处理其余部分。
- 以一个完整的CUDA加速应用示例，包括特定应用模块、第三方软件、CUDA库代码，所有软件都需要映射到目标CPU。目前CUDA的重点移植的是下图中的绿色部分（如PyTorch示例中的CUDA KMD、CUDA UMD）此外，第三方软件和应用软件也需同步移植到RISC-V。

图：CUDA 允许 RISC-V 成为主要的应用处理器



图：CUDA加速应用示例



# 1.5 英伟达：FY2026Q3营收指引540亿美元左右

- **业绩：** FY2026Q2（截止至2025年7月27日），英伟达营收467亿美元，同比+56%。Blackwell数据中心收入环比增长17%，Blackwell实现了非凡的跨越式发展——Blackwell Ultra的产量正在全速提升，市场需求也异常旺盛。指引：FY2026Q3，预期营收为540亿美元左右，上下浮动2%。
- **毛利率：** FY2026Q2，GAAP和非GAAP毛利率分别为72.4%和72.7%；指引：FY2026Q3，预期GAAP和非GAAP毛利率预计分别为73.3% 和73.5%，分别上下浮动50个基准点。

表：英伟达分业务表现

| 亿美元     | Q1FY23 | Q2FY23 | Q3FY23 | Q4FY23 | Q1FY24 | Q2FY24 | Q3FY24 | Q4FY24 | Q1FY25 | Q2FY25 | Q3FY25 | Q4FY25 | Q1FY26 | Q2FY26 | Q3FY26E  |
|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|----------|
| 数据中心    | 37.5   | 38.1   | 38.3   | 36.2   | 42.8   | 103.2  | 145.1  | 184    | 225.6  | 262.7  | 307.7  | 356    | 391.1  | 411    |          |
| yoy (%) | 83%    | 61%    | 31%    | 11%    | 14%    | 171%   | 279%   | 409%   | 427%   | 154%   | 112%   | 93%    | 73%    | 56%    |          |
| 游戏      | 36.2   | 20.4   | 15.7   | 18.3   | 22.4   | 24.9   | 28.6   | 28.7   | 26.5   | 28.8   | 32.8   | 25     | 37.6   | 42.9   |          |
| yoy (%) | 31%    | -33%   | -51%   | -46%   | -38%   | 22%    | 81%    | 56%    | 18%    | 16%    | 15%    | -11%   | 42%    | 49%    |          |
| 专业可视化   | 6.2    | 5      | 2      | 2.3    | 3      | 3.8    | 4.2    | 4.6    | 4.3    | 4.5    | 4.9    | 5.11   | 5.1    | 6.01   |          |
| yoy (%) | 67%    | -4%    | -65%   | -65%   | -53%   | -24%   | 108%   | 105%   | 45%    | 20%    | 17%    | 10%    | 19%    | 32%    |          |
| 汽车      | 1.4    | 2.2    | 2.5    | 2.9    | 3      | 2.5    | 2.6    | 2.8    | 3.3    | 3.5    | 4.5    | 5.7    | 5.7    | 5.9    |          |
| yoy (%) | -10%   | 45%    | 86%    | 135%   | 114%   | 15%    | 4%     | 8%     | 11%    | 37%    | 72%    | 103%   | 72%    | 69%    |          |
| 总收入     | 82.9   | 67     | 59.3   | 60.5   | 71.9   | 135.1  | 181.2  | 221    | 260    | 300.4  | 350.8  | 393    | 441    | 467    | 540(±2%) |
| yoy (%) | 46%    | 3%     | -17%   | -21%   | -13%   | 101%   | 206%   | 265%   | 262%   | 122%   | 94%    | 78%    | 69%    | 56%    | 54%      |



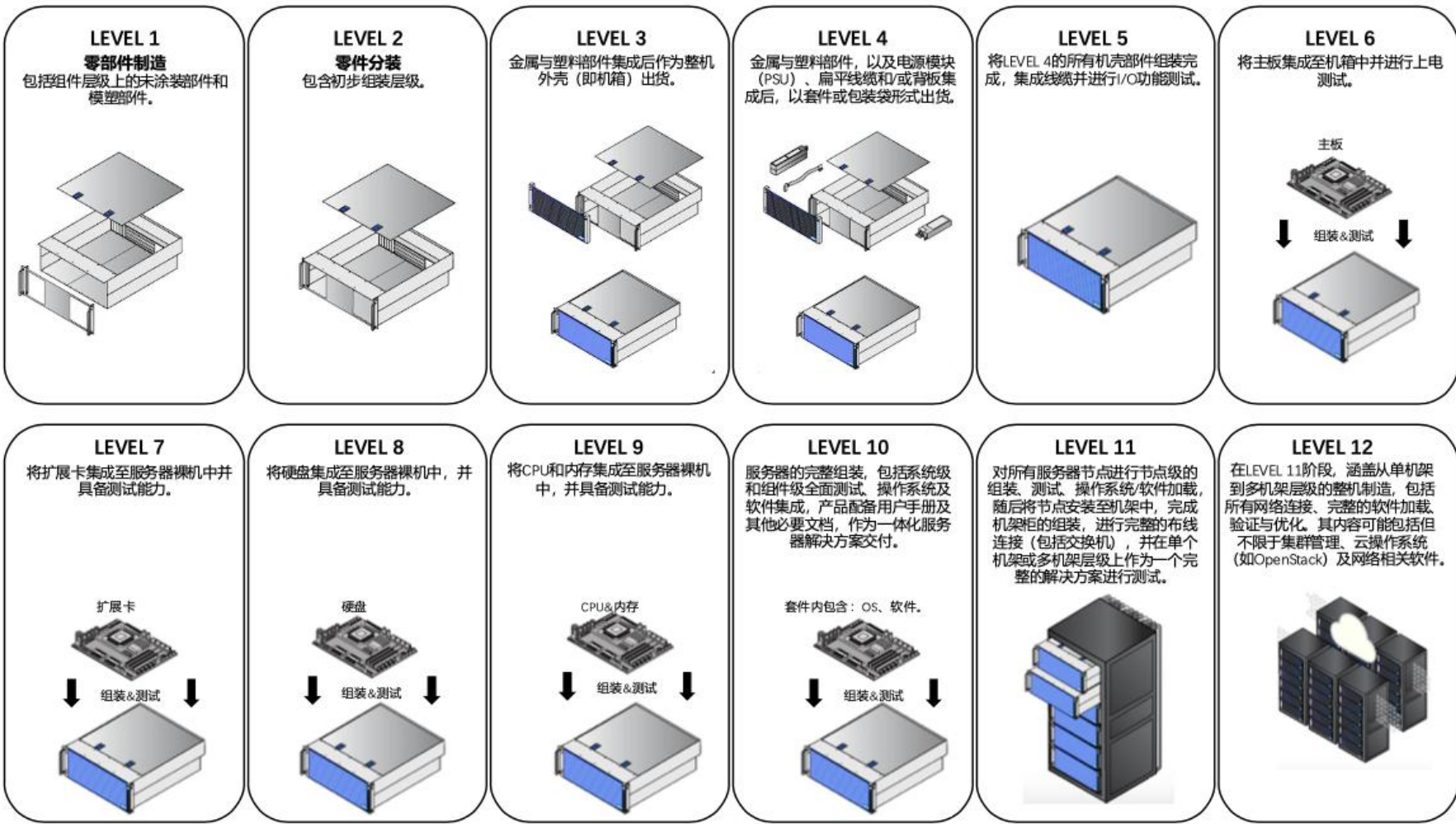
## 第二章 计算

# 主板从HGX到MGX，Rubin Ultra NVL576架构升级

# 2.1 英伟达整机：HGX到MGX计算平台，服务器整机到系统形态

服务器制造级别可分为 Level 1-Level 12。  
 Level 1为零部件制造，Level 6为主板集成，通常为 ODM 发货“服务器准系统”时提供的。  
 Level 10为完整服务器组装和测试，能够达到 10 级制造水平的制造商将提供有效的服务器解决方案。Level 11-12级可将多台服务器联网作为机架级甚至多机架级解决方案。

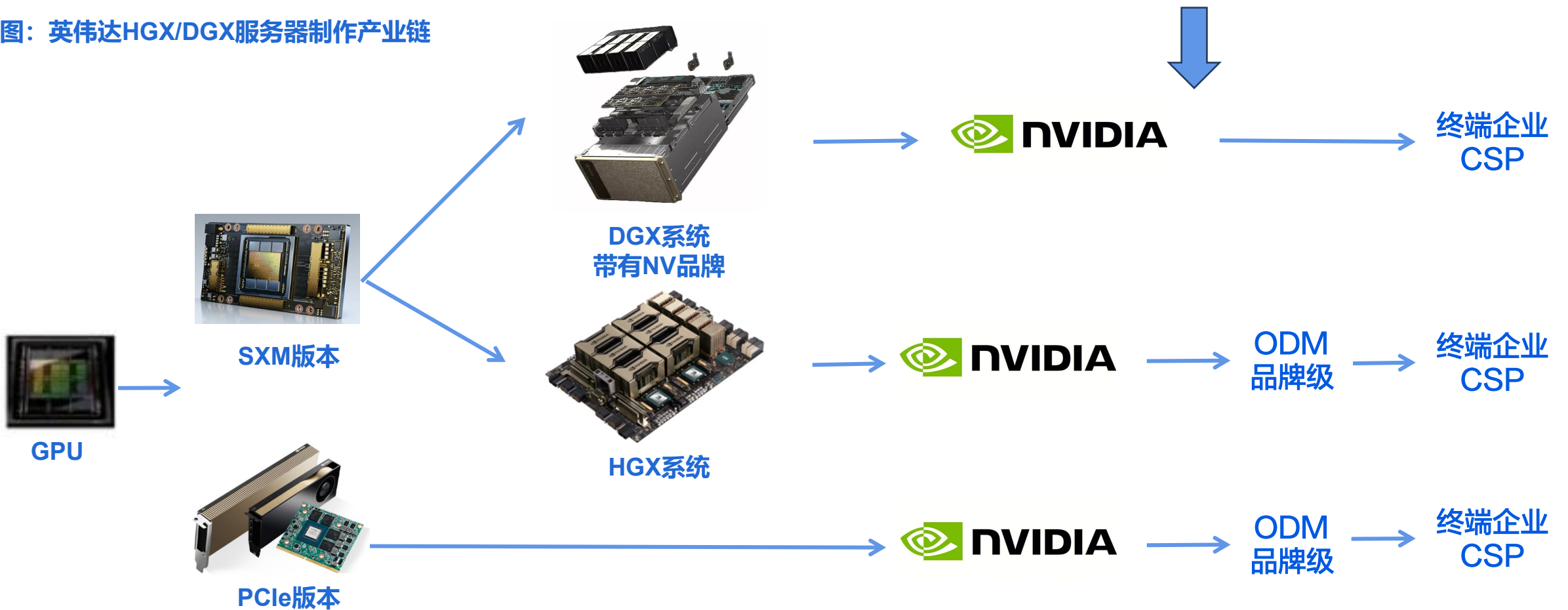
图：服务器制造等级分类



## 2.1 英伟达整机：HGX版本由ODM出货，DGX版本由英伟达交付

- **HGX**：模组主要为SXM版本，随后8个GPU SXM模组构建成一个HGX UBB基板，基本交付给英伟达后，分配给品牌级服务器厂商，包括鸿海、广达、纬创、英业达等厂商。
- **DGX**：SXM模组生产后，交给下游服务器厂商进行组装，整机交付给英伟达，流向下游CSP厂商与企业客户等。

图：英伟达HGX/DGX服务器制作产业链

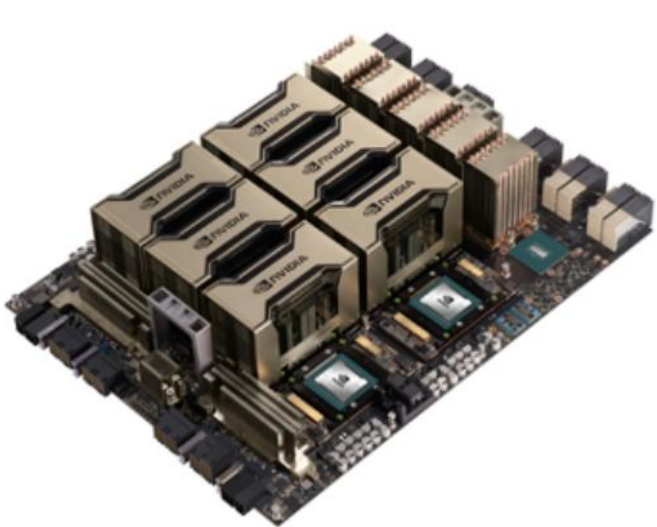




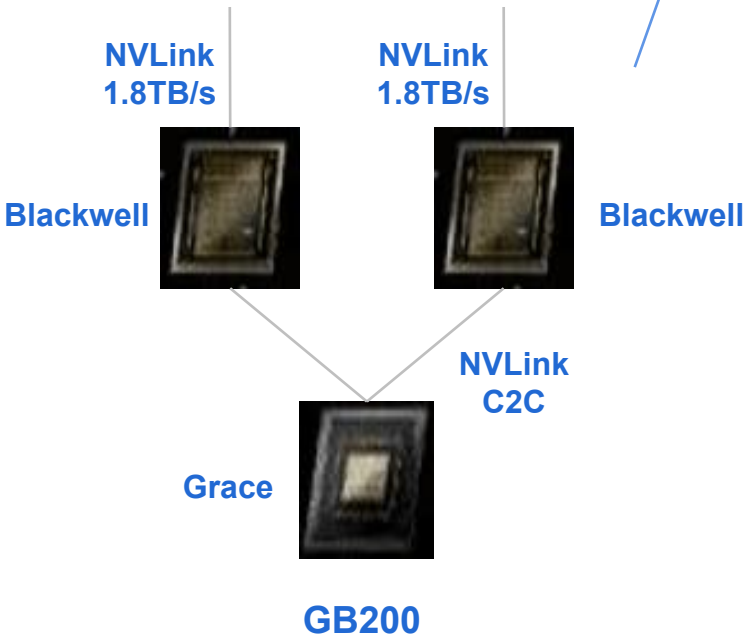
## 2.1 英伟达整机：MDX开放模块化设计，NVL72或采用此结构

- **MGX (Module GPU Accelerator)**：是一个开放模块化服务器设计规范和加速计算的设计。MGX让系统制造商快速且经济高效地为 AI、HPC 和 NVIDIA Omniverse 应用程序构建 100 多种服务器变体。  
**GB300 NVL72**整个系统由18个计算托盘和9个交换机托盘组成。

图：英伟达服务器制作产业链从看HGX-8卡到MGX



HGX系统



**GB200 SUPERCHIP COMPUTE TRAY**  
2x GB200  
80 PETAFLIPS FP4 AI INFERENCE  
40 PETAFLIPS FP8 AI TRAINING  
1728 GB FAST MEMORY  
1U Liquid cooled  
18 Per Rack



**NVLINK SWITCH TRAY**  
2x NVLINK SWITCH  
14.4 TB/s Total Bandwidth  
SHARPV4 FP64/32/16/8  
1U Liquid cooled  
9 Per Rack

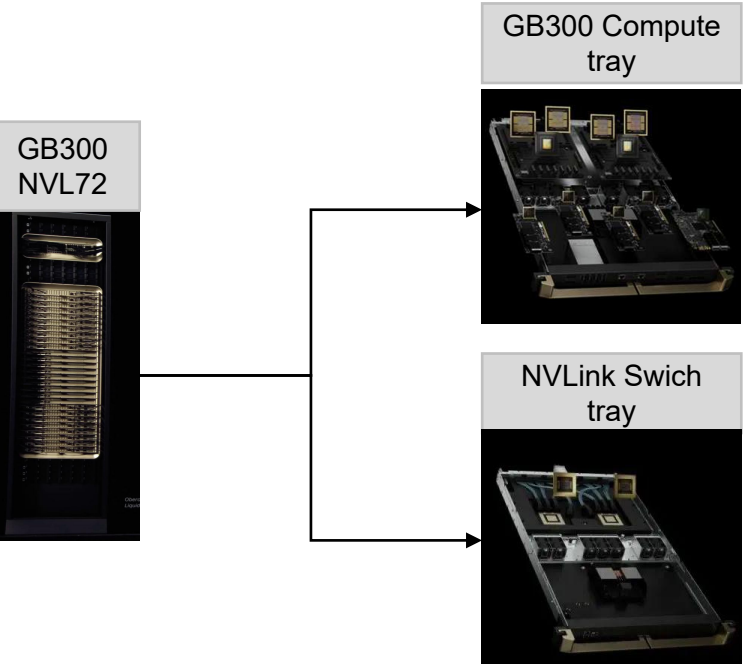




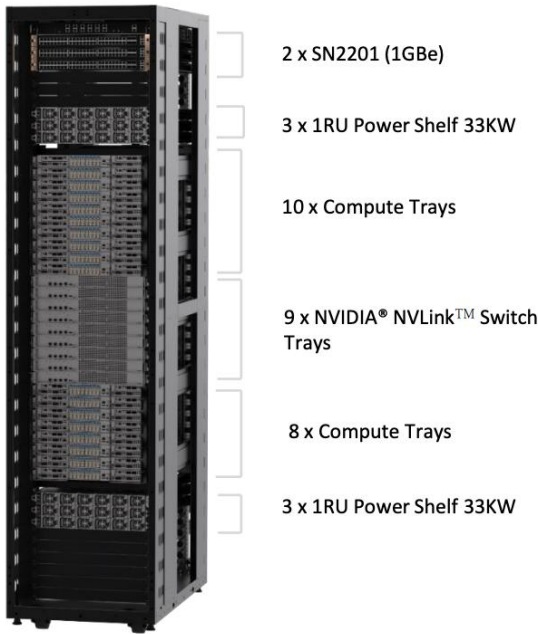
## 2.1 英伟达整机：GB300 NVL72提升性能与带宽，显著增强AI工厂效率

- GB300 NVL72机架：**18个计算托盘，每个计算托盘包含4颗Blackwell Ultra GPU+2颗Grace CPU，总计72颗Blackwell Ultra GPU+36颗Grace CPU；HBM容量达到了20TB，总带宽576TB/s；配备9个NVLink交换机托盘；节点间NVLink带宽130TB/s；内置72张CX-8网卡，提供14.4TB/s带宽。GB300 NVL72 可无缝集成NVIDIA Quantum-X800与NVIDIA Spectrum-X网络平台，使 AI 工厂和云数据中心能够从容应对三大扩展定律的需求。此外，机架还整合了18张用于增强多租户网络、安全性和数据加速BlueField-3 DPU。
- NVIDIA Blackwell Ultra 与 Hopper 相比AI工厂输出性能的整体潜力提升50倍。**与 Hopper 相比，采用NVIDIA GB300 NVL72的AI工厂将在每用户TPS上实现10倍提升，并在单位功耗（每MW）吞吐率上相比Hopper架构实现5倍提升。这一叠加效应将带来AI工厂总体产出性能最多提升50倍的潜力。

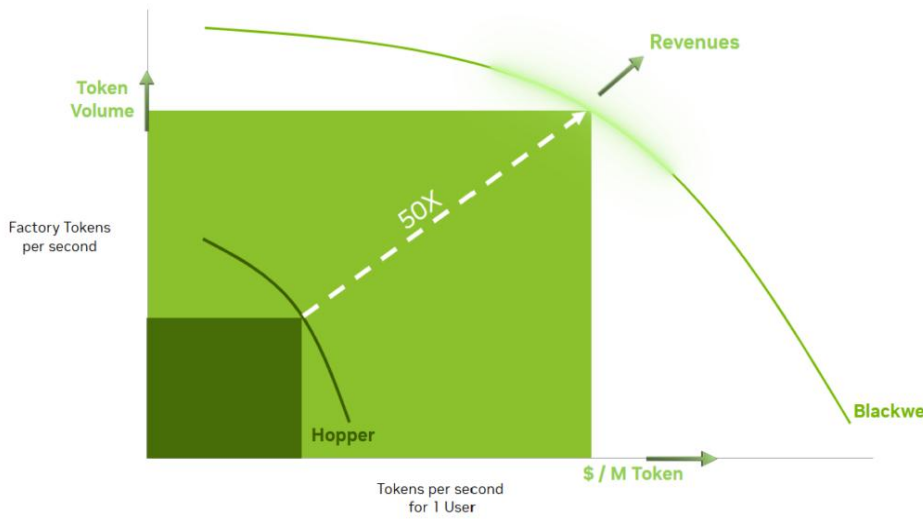
图：英伟达GB300 NVL72 的计算与网络节点



图：英伟达GB300 NVL72整机



图：Blackwell Ultra 与 Hopper 的AI工厂输出性能



## 2.2 NVL72组装模式：计算能力、内存容量实现提升

### ● GB300 NVL72对比GB200 NVL72

- ✓ **计算性能：**密集FP4计算提升至1.5倍，INT8 计算几乎被淘汰：GB200 提供 360P（INT8 稀疏计算），而GB300则降至仅23P（INT8密集计算）。
- ✓ **内存配置：**HBM内存（容量增加1.5倍，带宽保持为每个GPU 8 TB/s）。
  - GB200: 72 个 GPU x 192GB = 总计13.8TB。
  - GB300: 72个GPU x 288GB= 总计20.7TB。
- ✓ **互联网技术：**内置全新的ConnectX-8网卡，替代了之前的ConnectX-7，同时光模块带宽从800G升级到了1.6T，这一改进显著提升了数据传输速度和网络带宽。

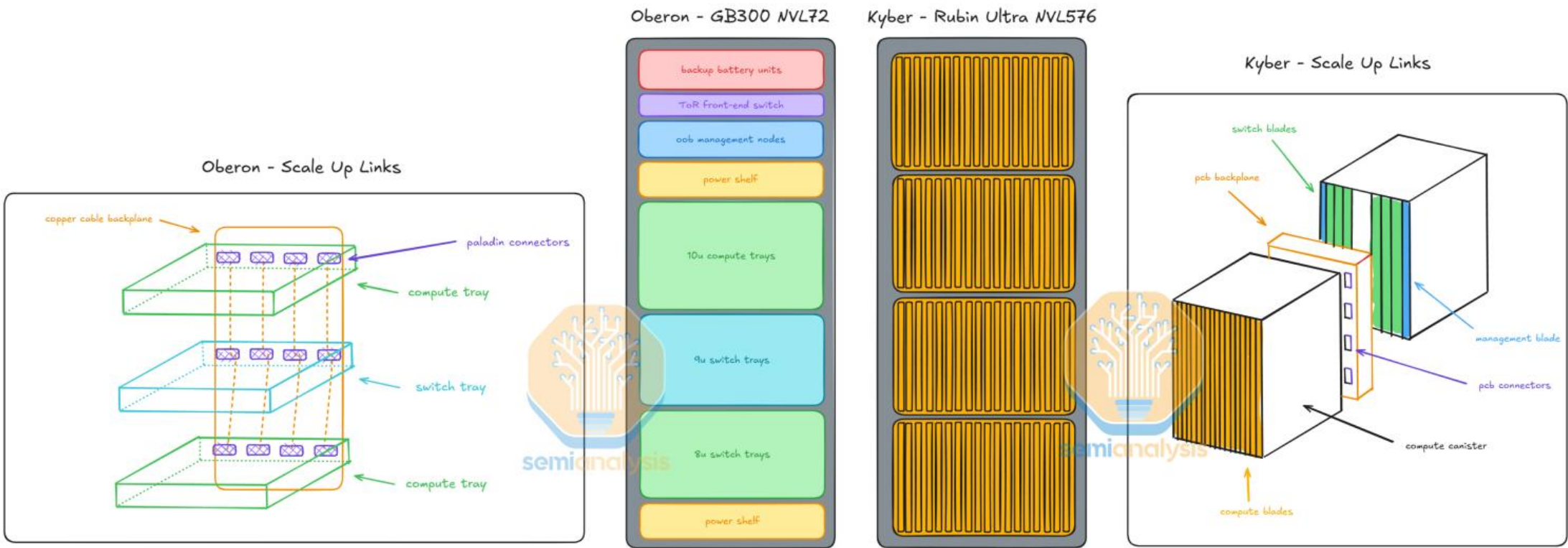
表：Hopper架构和Blackwell架构对应的NVL机柜和SuperPod

| Architecture            | NVL32                 | GH200 SuperPod                                 | GB200 NVL72           | GB200 SuperPod                                  | GB300 NVL72           |
|-------------------------|-----------------------|------------------------------------------------|-----------------------|-------------------------------------------------|-----------------------|
|                         | 32 xGH200             | 256 x GH200                                    | 36 x GB200            | 288 x GB200                                     | 36 x GB300            |
|                         | 32 x GPU              | 256 x GPU                                      | 72 x GPU              | 576 x GPU                                       | 72 x GPU              |
|                         | Hopper                |                                                | Blackwell             |                                                 |                       |
| HBM VRAM Size           | 32 x 144GB=4.6T       | 256 x 96GB=24.5TB                              | 36 x 384GB=13.8TB     | 288 x 384GB=110TB                               | 36 x 2 x 288GB=20.7TB |
| LPDDR5X VRAM Size       | 32 x 480GB=15.4T      | 256 x 480GB=123TB                              | 36 x 480GB=17.3TB     | 288 x 480GB=138TB                               | 36 x 480GB=17.3TB     |
| HBM VRAM Bandwidth      | 32 x 4.9TB/s          | 256 x 4TB/s                                    | 72 x 8TB/s            | 576 x 8TB/s                                     | 72 x 8TB/s            |
| FP16(FLOPS)             | 32P                   | 256P                                           | 180P                  | 1440P                                           | 180P                  |
| INT8(FLOPS)             | 64P                   | 64P                                            | 360P                  | 2880P                                           | 23P                   |
| FP8(FLOPS)              | 64P                   | 64P                                            | 360P                  | 2880P                                           | 360P                  |
| FP6(FLOPS)              | X                     | X                                              | 360P                  | 2880P                                           | 360P                  |
| FP4(FLOPS)              | X                     | X                                              | 720P                  | 5760P                                           | 1100P                 |
| NVSwitch System         | Gen3-64 Port/NVSwitch |                                                | Gen4-72 Port/NVSwitch |                                                 |                       |
|                         | 9 x L1 NVSwitch Tray  | 96 x L1 NVSwitch Tray<br>36 x L2 NVSwitch Tray | 9 x L1 NVSwitch Tray  | 144 x L1 NVSwitch Tray<br>72 x L2 NVSwitch Tray | 9 x L1 NVSwitch Tray  |
| GPU-to-GPU Bandwidth    | 0.9TB/s               | 0.9TB/s                                        | 1.8TB/s               | 1.8TB/s                                         | 1.8TB/s               |
| NVLink Bandwidth        | 36 x 0.9T/s=32TB/s    | 256 x 0.9T/s=230TB/s                           | 72 x 1.8T/s=130TB/s   | 576 x 1.8TB/s=1PB/s                             | 72 x 1.8T/s=130TB/s   |
| Ethernet                | 16 x 200Gb/s          | 256 x 200Gb/s                                  | 18 x 400Gb/s          |                                                 | 18 x 200Gb/s          |
| InfiniBand              | 32 x 400Gb/s          | 256 x 400Gb/s                                  | 72 x 800Gb/s          | 576 x 800Gb/s                                   | 72 x 800Gb/s          |
| Power Consumption       | 32 x 1kw=32kw         | 256 x 1kw=256kw                                | 36 x 2.7kw=97.2kw     | 288 x 2.7kw=777.6kw                             |                       |
| Total Power Consumption |                       |                                                | 120kw                 |                                                 |                       |
| Notes                   | ConnectX-7 NIC        |                                                | ConnectX-8 NIC        |                                                 |                       |

## 2.2 NVL72组装模式：优化PCB与机架设计，提升互联性能

- PCB：全称Printed Circuit Board，即印刷电路板，是指在通用基材上按预定设计形成点间连接及印制元件的印刷板。
- Rubin Ultra NVL576采用Kyber架构，PCB背板替代铜缆互联。英伟达将于2027年下半年推出Rubin Ultra NVL576产品，使用代号为“Kyber”的新型液冷机架。在此架构下，计算托盘被旋转90度，显著提升机架密度。每个机架包含4个计算单元，每层配置18个计算刀片。为了克服在有限空间中布设铜缆的难度，PCB板背板取代了铜缆背板，作为机架内GPU与NVSwitch之间的扩展互联链路。

图：Oberon机架架构和Kyber机架架构示意图

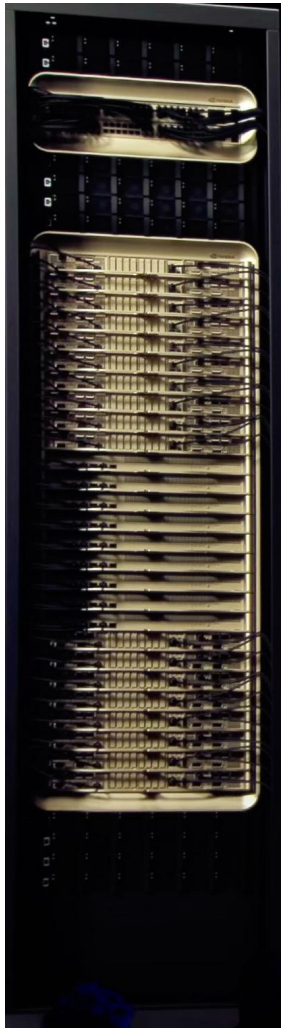




## 2.2 Vera Rubin NVL144：配备HBM4内存，算力与互连全面升级

- Vera Rubin NVL144将于2026年下半年推出。
- 规格：Rubin GPU将采用两颗Reticle大小的芯片，FP4性能高达50PFLOPS，并配备288GB的HBM4内存。这些芯片还将搭载一颗88核Vera CPU，该CPU采用定制的Arm架构，拥有176个线程，并支持高达1.8 TB/s的NVLINK-C2C互连。
- 性能扩展：Vera Rubin NVL144平台将具有3.6Exaflops的FP4推理能力和1.2Exaflops的FP8训练能力，比GB300 NVL72提升3.3倍，13TB/s的HBM4内存和75TB的快速内存，比GB300提升60%，并且NVLINK和CX9功能是前代的2倍，额定速度分别高达260TB/s和28.8TB/s。

图：Vera Rubin NVL144



### Vera Rubin NVL144 Second Half 2026

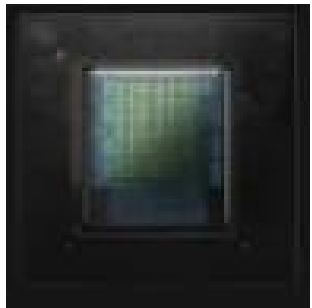
3.6 EF FP4 Inference  
1.2 EF FP8 Training  
3.3X GB300 NVL72

13 TB/s HBM4  
75 TB Fast Memory  
1.6X

260 TB/s NVLink6  
2X

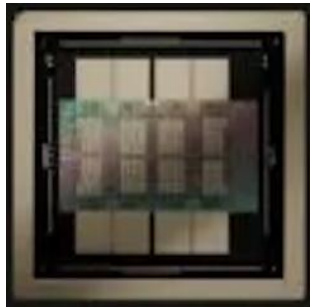
28.8 TB/s CX9  
2X

Vera



88 Custom Arm Cores  
176 Threads  
1.8 TB/s NVLink-C2C

Rubin



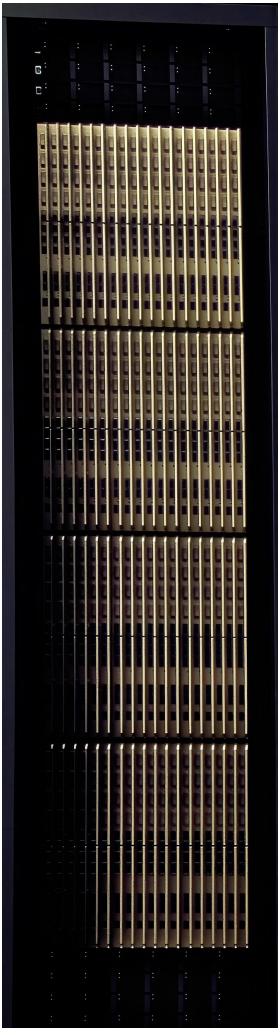
2 Reticle-Sized GPUs  
50PF FP4 | 288GB HBM4



## 2.2 Vera Rubin Ultra NVL576：配备HBM4E内存，性能显著跃升

- Vera Rubin Ultra NVL576计划2027年下半年推出。
- 规格：Rubin Ultra GPU将配备四个Reticle-Sized GPU，提供高达100 PFLOPS的FP4性能，并配备1TB的HBM4E内存。
- 性能扩展：Rubin Ultra NVL576平台将具有15Exaflops的FP4推理能力和5Exaflops的FP8训练能力，比GB300 NVL72提升14倍；4.6PB/s的HBM4E内存和365TB的快速内存，是GB300的8倍；NVLINK能力是前代的12倍，CX9能力是前代的8倍，额定速度分别高达1.5PB/s和115.2TB/s。

图：Vera Rubin Ultra NVL576



### Rubin Ultra NVL576 Second Half 2027

15 EF FP4 Inference  
5 EF FP8 Training  
14X GB300 NVL72

4.6 PB/s HBM4e  
365 TB Fast Memory  
8X

1.5 PB/s NVLink7  
12X

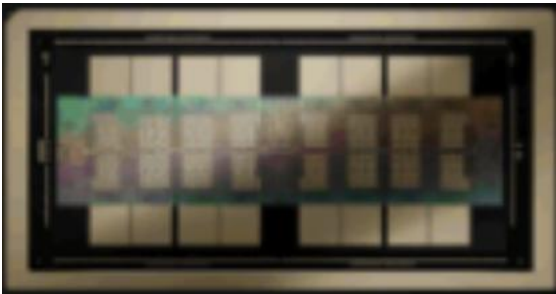
115.2 TB/s CX9  
8X

Vera



88 Custom Arm Cores  
176 Threads  
1.8 TB/s NVLink-C2C

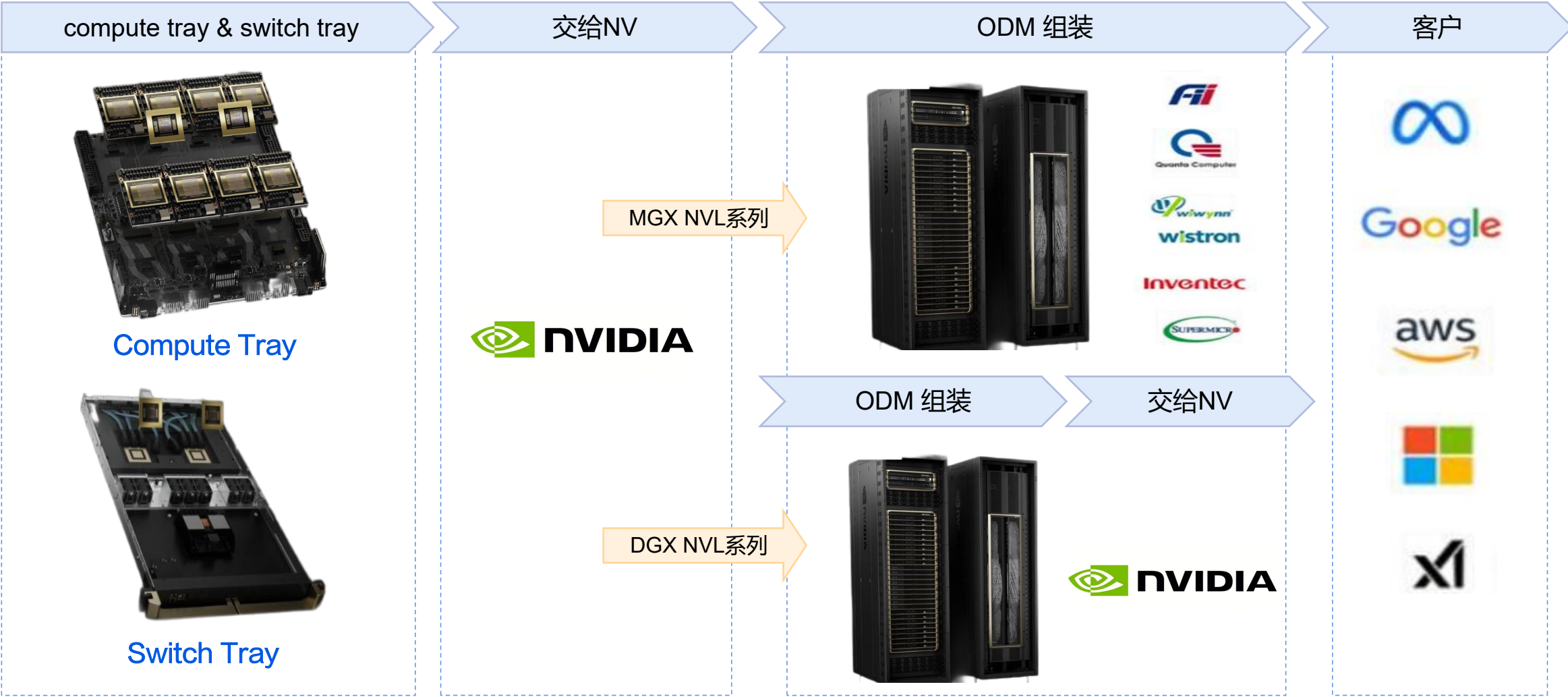
Rubin Ultra



4 Reticle-Sized GPUs  
100PF FP4 | 1TB HBM4

## 2.2 NVL72组装模式：MGX-->整机柜-->集群

图：NVL72的组装流程



## 第三章 网络

# CPO取代可插拔光模块，Rubin实现超高速互联

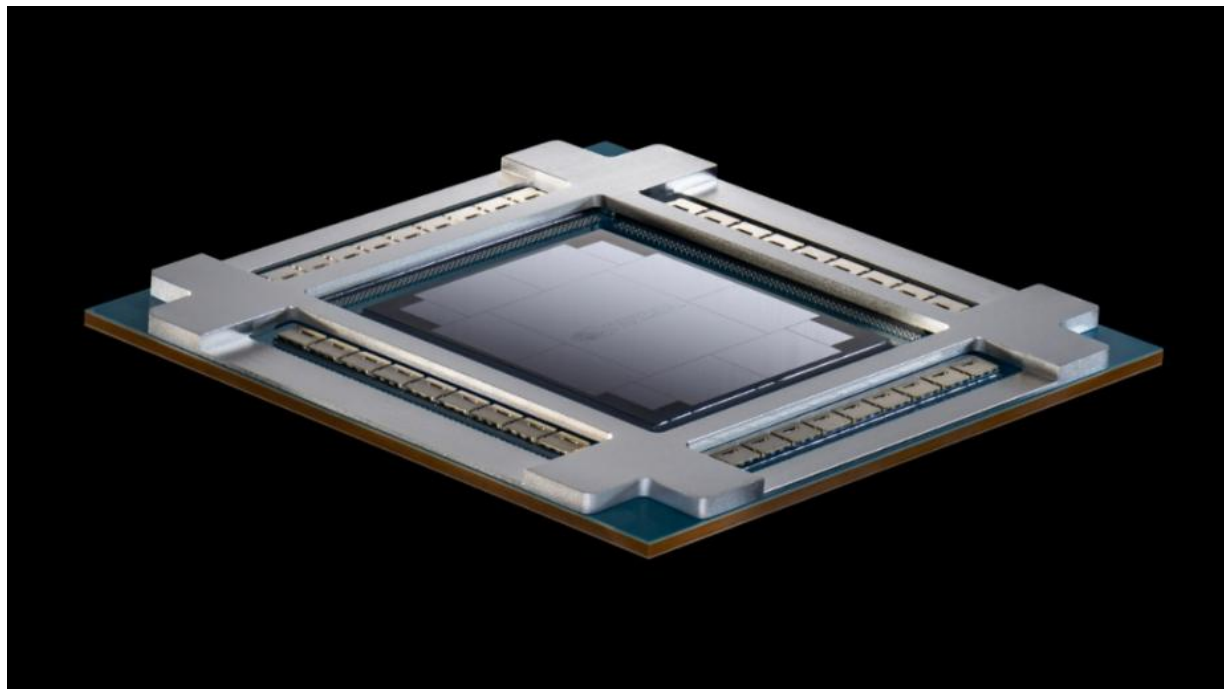
### 3.1 NVIDIA CPO：集成硅光子技术，打破超大规模与企业网络旧限制

- NVIDIA集成硅光子的共封装光纤 (CPO) 交换机是面向代理AI时代的全球最先进的网络解决方案。NVIDIA CPO创新技术将可插拔收发器替换为与ASIC封装在同一封装中的硅光子器件，与传统网络相比，其能效提高了3.5倍，网络弹性提高了10倍，部署速度提高了1.3倍。
- 根据SemiAnalysis分析，在部署400,000个GB200 NVL72设备的场景中，从传统的基于DSP收发器的三层网络迁移到基于CPO的两层网络，可以实现高达12%的集群总功耗节省，将收发器功耗占比从计算资源的10%降低到仅1%。

图：用于NVIDIA Quantum-X800 InfiniBand平台的CPO



图：用于NVIDIA Spectrum-X以太网平台的CPO

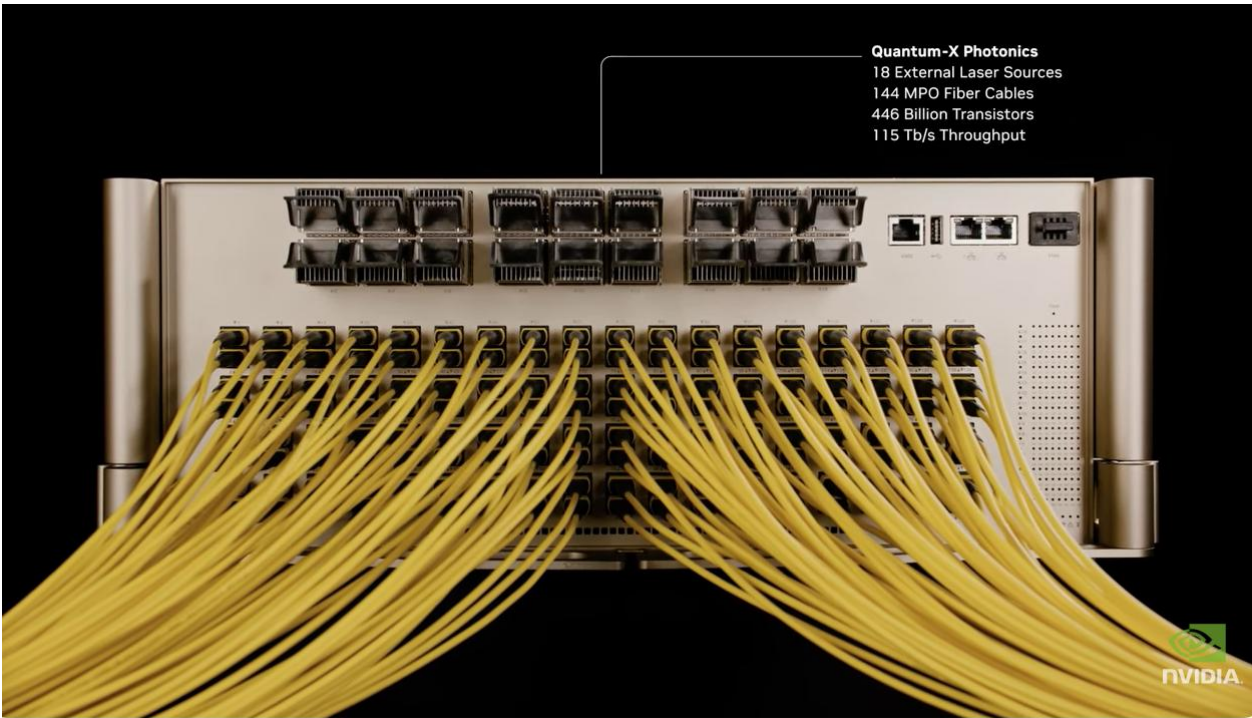




### 3.1 NVIDIA CPO：为构建百万级GPU的AI工厂提供保障

- NVIDIA CPO带来了多项关键优势，为构建百万级GPU的AI工厂提供了必要的网络可扩展性保障。
  - ✓ **功耗方面：**CPO技术显著降低网络能耗，相较于传统可插拔光模块，实现3.5倍的能效提升。
  - ✓ **网络弹性：**在支持自主智能所需的超大规模部署下，相较于可插拔光模块提供10倍的网络可靠性与弹性。
  - ✓ **部署速度：**CPO简化数据中心网络的安装与维护流程，使系统从部署到产生洞察的时间加快1.3倍。
  - ✓ **延迟速度：**CPO不依赖数字信号处理器（DSP）重定时器，从而显著降低网络延迟。
  - ✓ **TCO方面：**通过消除可插拔光模块的使用，NVIDIA CPO简化了物料清单（BOM），从而降低总体成本。
  - ✓ **维修方面：**CPO所需组件更少，安装与更换操作远比传统可插拔光模块更为简便。
- Nvidia宣布其首款CPO解决方案，该解决方案将部署在其横向扩展交换机中。借助CPO，收发器现在被外部激光源取代，这些激光源与直接放置在芯片硅片旁边的光学引擎（OE）一起促进数据通信。光纤电缆现在不再插入收发器端口，而是插入交换机上的端口，将信号直接路由到光学引擎。

图：Quantum-X Photonics交换机细节图



## 3.2 NVIDIA 交换机：力图将AI数据中心功耗降低50%以上

- Nvidia正在通过将CPO集成到其Quantum InfiniBand和Spectrum以太网交换机中来推进其网络技术，此举有望降低AI数据中心的功耗和成本。Quantum-X和Spectrum-X交换机减少了对传统光收发器的依赖，为超大规模人工智能工厂提供高达400Tbps的吞吐量。
- **Quantum-X Photonics交换机**：上市时间预计在2026年初上市。NVIDIA Quantum-X800平台正在扩展，新增基于CPO技术的交换机，首款产品为Q3450-LD，支持144个800Gb/s InfiniBand端口。这一突破性交换机采用液冷设计，高效冷却板载硅光子器件。NVIDIA Quantum-X光子InfiniBand交换机支持全新网络架构创新，实现前所未有的扩展能力，可在非阻塞的两级胖树拓扑结构下，以800Gb/s的速率连接超过 10,000 个 GPU。
- 相比Quantum家族的上代产品，速度快2倍，扩展性提升5倍。就Quantum-X系统来看，每颗CPO芯片都配有18个硅光chiplet（3D堆叠的硅光引擎），每个硅光引擎采用TSMC N6工艺，2.2亿晶体管、1000个集成的光器件；每个硅光引擎连接2个激光器以及16条光纤（一颗CPO芯片也就要连36个激光器、288条数据连接）。
- **Spectrum-X Photonics交换机**：上市时间预计为2026年下半年。NVIDIA也在其Spectrum-X平台中引入了面向以太网交换机的CPO技术。该系列包括两个型号：一个型号支持512个800Gb/s的以太网端口，另一个型号支持128个800Gb/s的以太网端口。

图：NVIDIA Quantum-X Photonics Switch



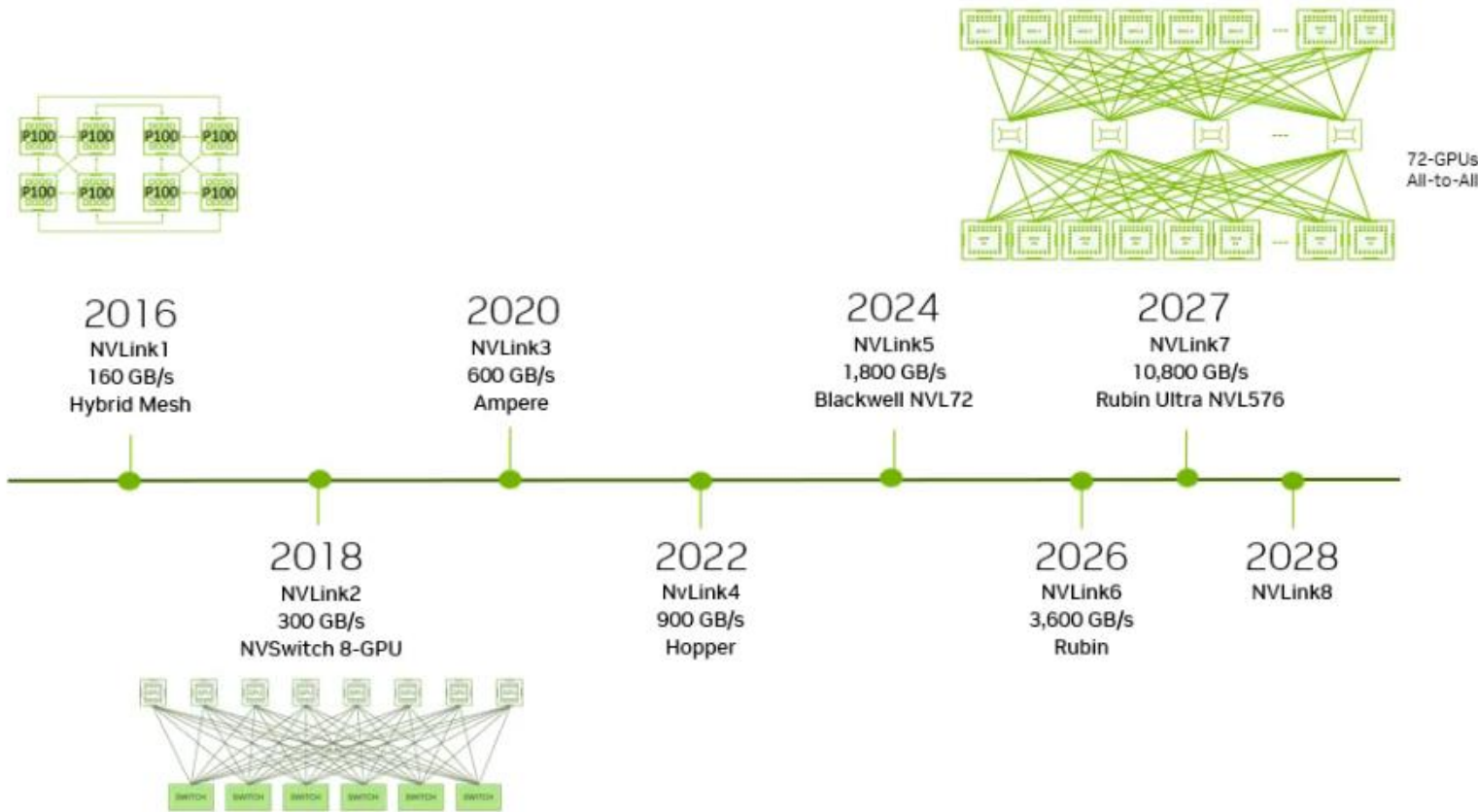
图：NVIDIA Spectrum-X Photonics Switch



# 3.3 NVIDIA NVLink：持续迭代升级，支持纵向扩展的需求

- **NVLink——GPU到GPU双向互连技术**，可在服务器内扩展多GPU输入和输出 (IO)。NVIDIA保持NVLink年度迭代，不断突破技术极限，对未来三代的NVLink 产品，会保持每年推出一代的节奏。这一迭代策略推动了持续的技术进步，有效满足了 AI 模型在复杂性和计算需求方面的指数级增长。
- **2016年，NVIDIA首次推出NVLink**，旨在克服PCIe在高性能计算和人工智能工作负载中的局限性。该技术实现了更快的GPU间通信，并构建了统一的内存空间。
- **2018年，NVIDIA推出了NVLink Switch技术**，实现了在8个GPU的网络拓扑中每对GPU之间高达300 GB/s的all-to-all带宽，为多GPU计算时代的scale-up网络奠定了基础。随后，在第三代NVLink Switch中引入了NVIDIA可扩展分层聚合与归约协议 (SHARP) 技术，进一步提升了性能，有效优化了带宽性能并降低了集合操作的延迟。
- **2024年，第五代NVLink发布**，进一步增强的NVLink Switch支持72个GPU实现全互联通信。通信速率达1800 GB/s，聚合总带宽高达130 TB/s，较第一代产品提升了 800 倍。
- **2026年，Rubin将采用NVLink 6.0技术**，速度翻倍至3.6TB/s（双向）。

表：英伟达NVLink 路线图

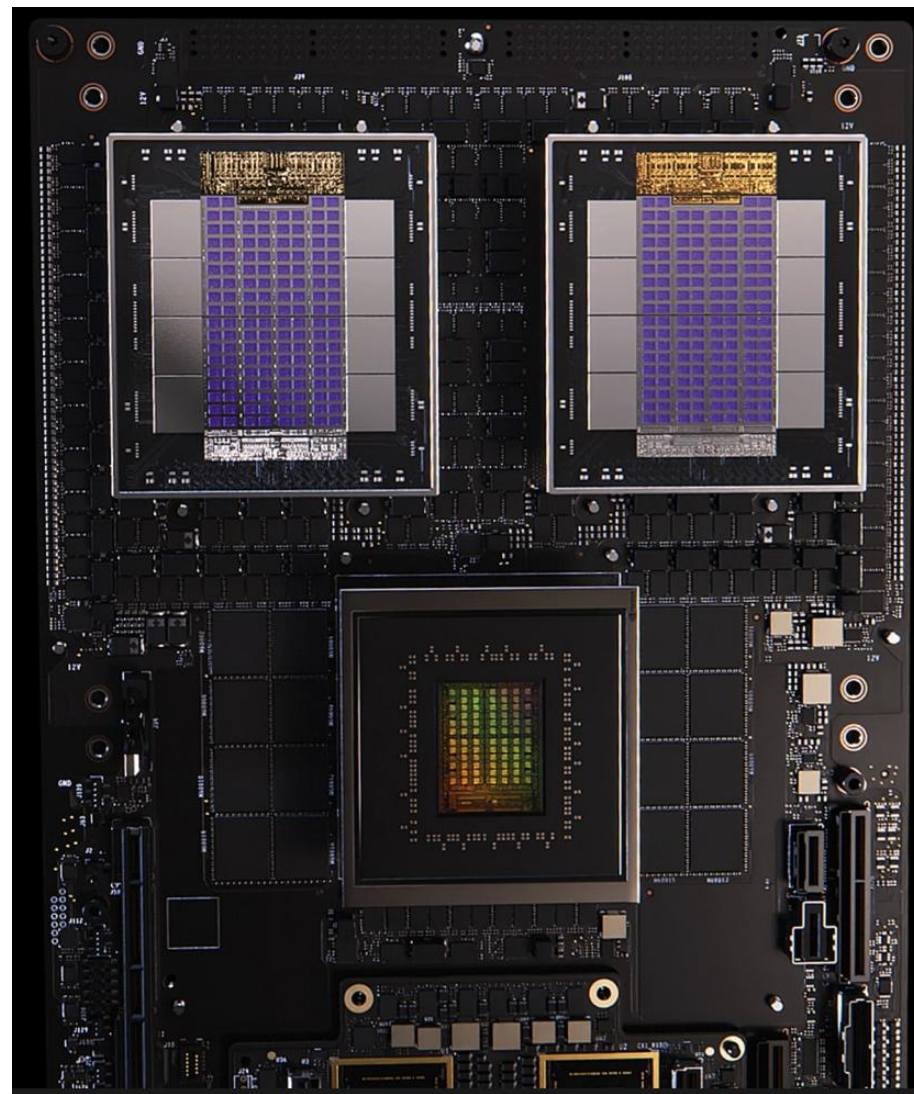




### 3.4 NVIDIA NVLink Fusion：开放半定制设计，满足客户的灵活需求

- **NVIDIA NVLink Fusion——全新互联芯片，构建半定制AI基础设施。**2025年5月19日，NVIDIA发布NVLink Fusion，旨在向第三方CPU和加速器开放NVLink生态系统，通过发布IP和硬件推动第三方设计与自家芯片互操作，虽系统仍需包含部分NVIDIA芯片，但目标是让合作伙伴构建融合英伟达芯片与定制芯片的半定制机架系统。
- **NVLink Fusion的优势：**
  - ✓ **出色的纵向扩展性能：**要充分发挥AI工厂的潜力，关键在于每个加速器之间需要快速、无缝的通信。NVIDIA NVLink 1.8 TB/s 互连速度可扩展至72个加速器，助力释放AI性能。
  - ✓ **强健的生态系统：**NVIDIA生态系统合作伙伴(包括ASIC设计师、CPU和IP提供商以及OEM/ODM)助力超大规模企业借助NVLink Fusion部署定制芯片，并缩短上市时间。
  - ✓ **易于部署和管理：**超大规模数据中心已经部署完整NVIDIA机架解决方案，NVLink Fusion支持异构芯片产品，同时围绕通用机架设计实现标准化，加速AI工厂部署并简化管理。
  - ✓ **持续迭代性：**NVIDIA保持年度路线图节奏，确保为NVLink Fusion采用者提供高性能的NVLink和整机柜架构性能。

图：NVIDIA NVLink Fusion

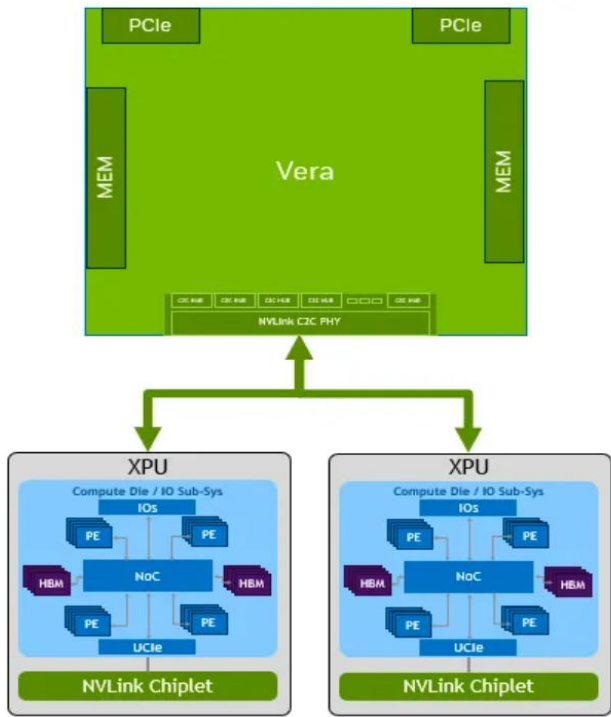




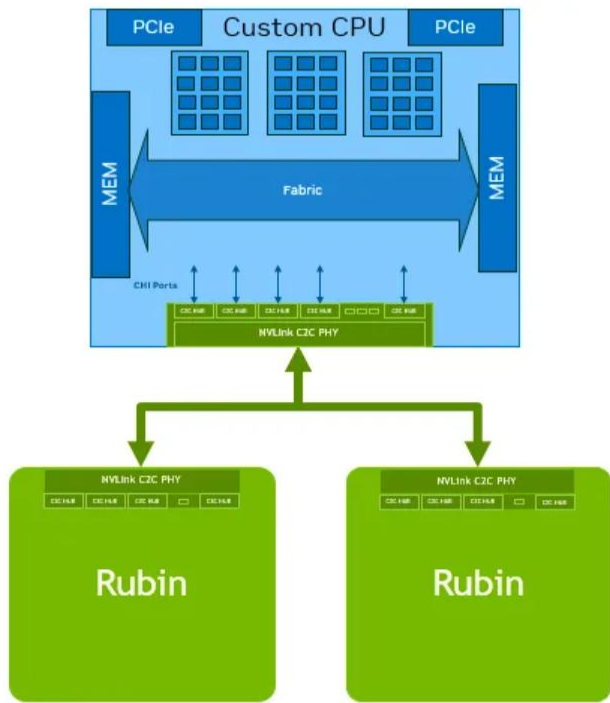
### 3.4 NVIDIA NVLink Fusion：开放半定制设计，满足客户的灵活需求

- **NVLink Fusion**为定制CPU、定制XPU或两者的组合配置提供了灵活的解决方案。作为模块化开放计算项目（OCP）MGX机架架构的一部分，NVLink Fusion可与任何网卡（NIC）、数据处理器（DPU）或横向扩展交换机集成，使客户能够根据需求灵活构建理想的系统。
- 对于定制XPU配置，NVLink通过通用芯粒互连（UCIe）IP与接口实现集成。NVIDIA提供支持UCIe的NVLink桥接芯片，既能实现极高性能，又便于集成，使客户能够像 NVIDIA 一样充分利用NVLink的功能。UCIe作为一项开放标准，采用该接口进行NVLink集成可让客户为其XPU灵活选择当前或未来平台的多种方案。
- 对于定制CPU配置，建议集成NVIDIA NVLink-C2C IP，以连接NVIDIA GPU，从而实现最佳性能。采用定制CPU与NVIDIA GPU的系统可平滑访问CUDA平台的数百个NVIDIA CUDA-X 库，充分发挥加速计算的高性能优势。

图：XPU通过 NVLink Chiplet访问 NVLink



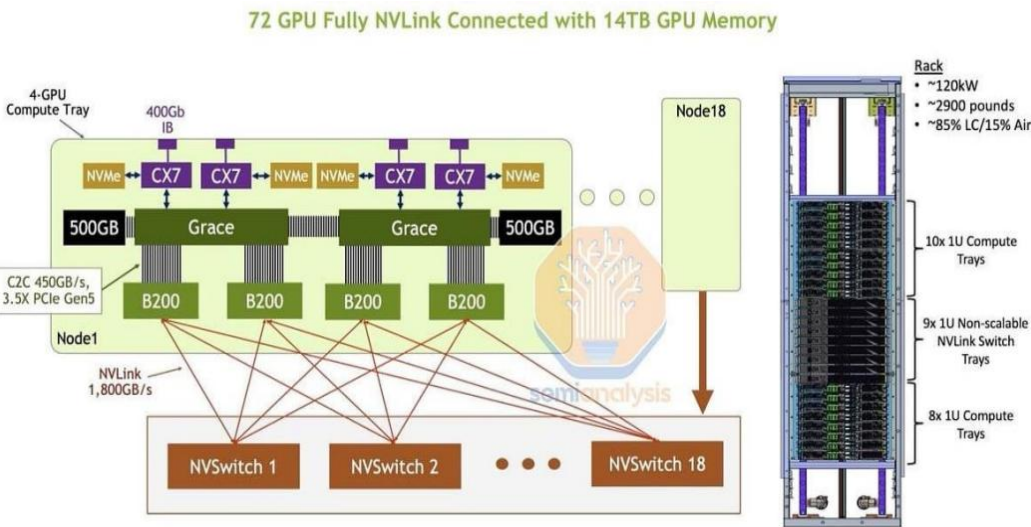
图：定制CPU通过NVLink-C2C访问NVLink



# 3.5 NVIDIA NVL72：优化铜缆布局，预计线缆长度将增加50%

- 英伟达GB200NVLink Switch和Spine由72个Blackwell GPU采用NVLink全互连，有5000根NVLink铜缆（合计长度超2英里）。GB300进一步优化了铜缆布局，预计线缆长度将增加50%，以满足更高性能和更大数据传输需求。
- 作为新一代AI服务器，GB300标配支持1.6T光模块（800G×2）的CX8网卡，对内部数据传输带宽要求较高。机柜内短距离互联场景（<5米）需要低时延、低功耗、高密度布线，1.6T铜缆（DAC或AEC）是最佳技术方案。这些铜缆单通道速率高达224Gbps，能够满足多通道聚合带来的高带宽需求。

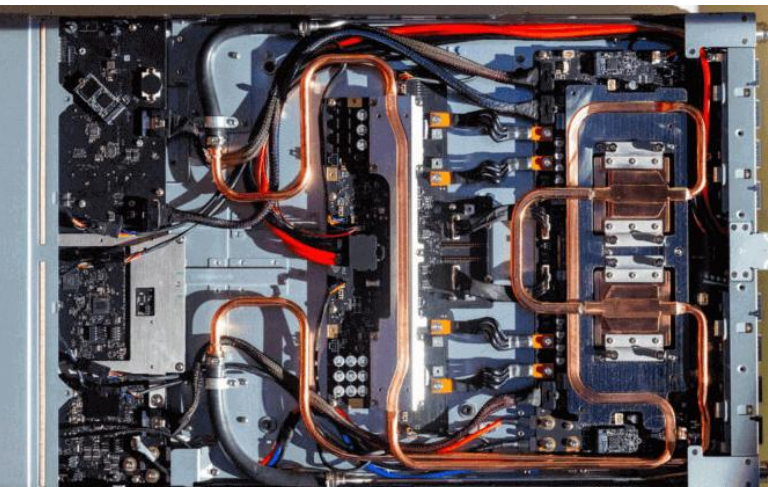
图：英伟达GB200 NVL72 互联结构



图：英伟达GB200 NVL72铜缆组成

图：英伟达GB300 NVL72 Switch Tray

图：英伟达GB300 NVL72 Compute Tray

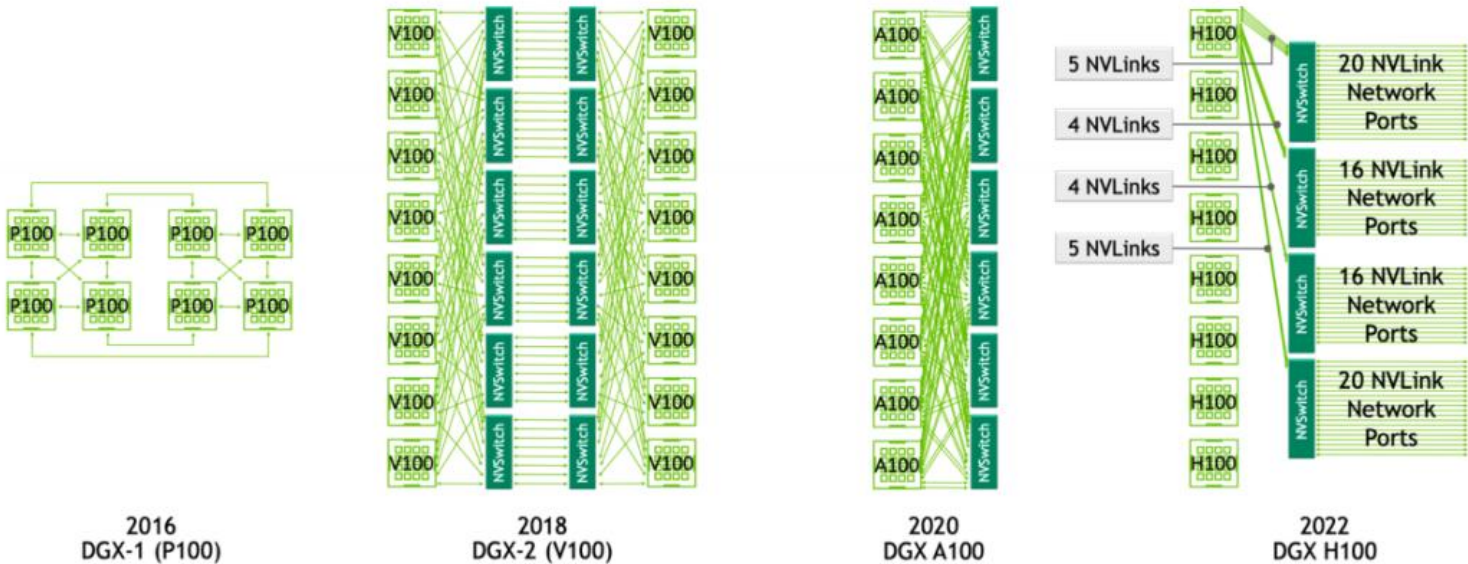




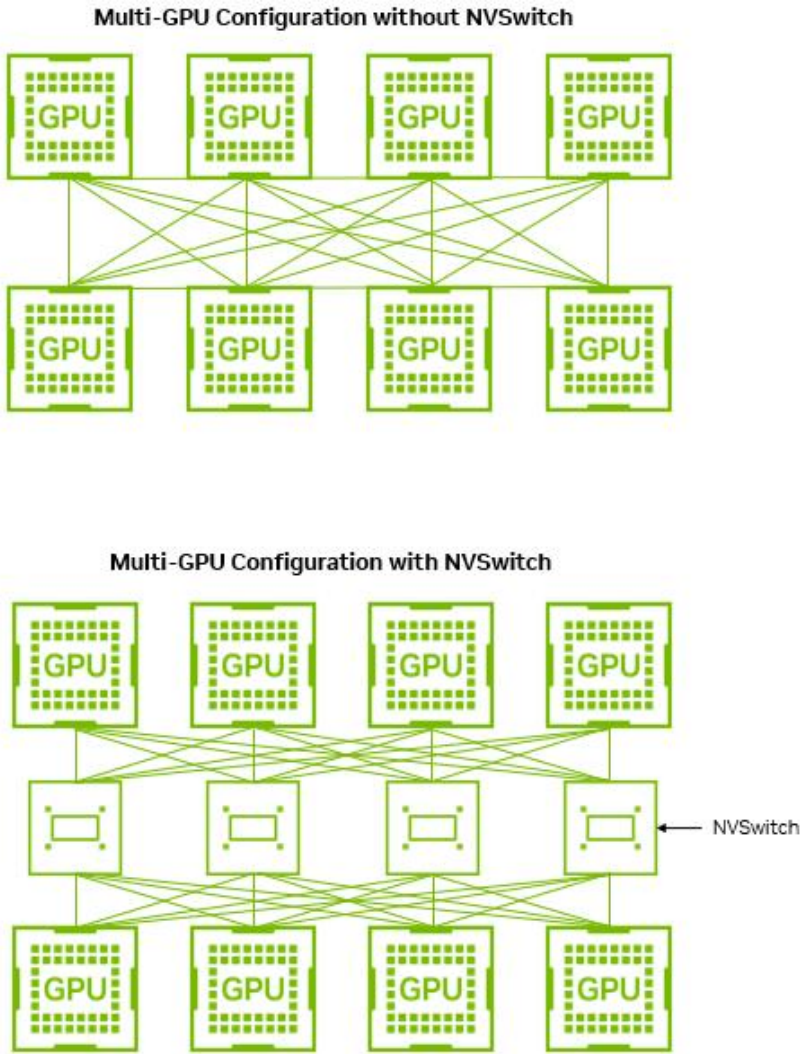
### 3.6 NVIDIA NVSwitch：实现非阻塞通信，实现更大规模的GPU互联

- NVSwitch——基于NVLink的多GPU互联解决方案。它不仅支持更多的NVLink链路，还允许多个GPU之间实现全互联，较大地优化了数据交换的效率和灵活性。
- NVIDIA将推出一款全新的NVSwitch 7.0。这是新款NVSwitch首次应用于中端平台。这允许更大的交换机聚合带宽和基数，从而在单个域中扩展至576个GPU芯片（144个封装），尽管拓扑结构可能不再是全对全无阻塞、轨道优化的单层多平面拓扑。相反，它可能是一个多平面轨道优化的双层网络拓扑，具有超额认购，甚至是非封闭拓扑。

表：英伟达NVLink 和NVSwitch相关的服务器



表：有无 NVSwitch 全互联交换时 GPU 与 GPU 间带宽比较



## 第四章 存储

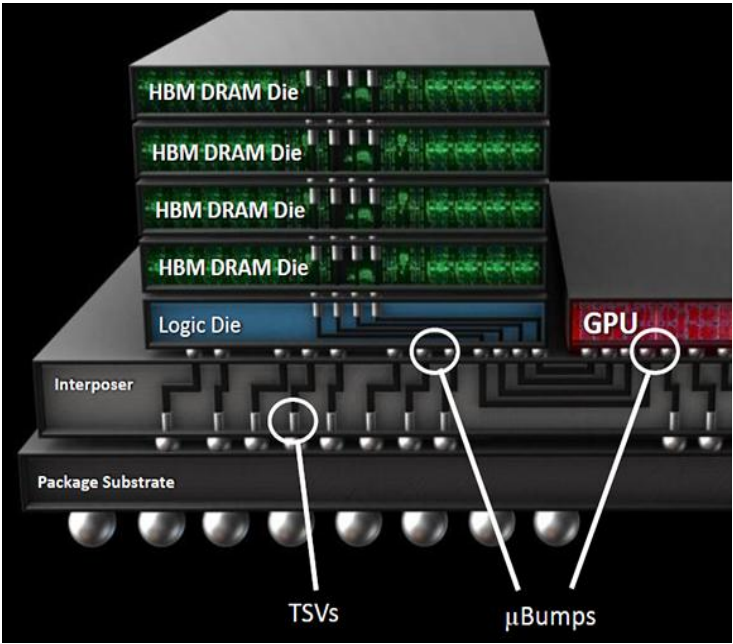
**HBM4预计2026年实现量产，SK海力士主导HBM市场**



## 4.1 HBM市场： 预计制造难度更高的HBM4溢价幅度将突破30 %

- **HBM（High Bandwidth Memory）**：是一款新型的CPU/GPU 内存芯片（即“RAM”），是将很多个DDR芯片堆叠在一起后和GPU封装在一起，实现大容量，高位宽的DDR组合阵列。
- 根据TrendForce最新研究，HBM技术发展受AI Server需求带动，三大原厂积极推进HBM4产品进度。由于HBM4的I/O（输入/输出接口）数增加，复杂的芯片设计使得晶圆面积增加，且部分供应商产品改采逻辑芯片架构以提高性能，皆推升了成本。据Trendforce数据，鉴于HBM3E刚推出时预计的溢价比例约为20%，预计制造难度更高的HBM4溢价幅度将突破30%。

图：HBM DRAM 3D图形



中间的die是GPU/CPU，左右2边4个小die是DDR颗粒的堆叠，Die之间用TVS方式连接，现在一般只有2/4/8三种层数。

图：HBM产品规格对比

A Comparison of HBM Product Specs

| Product | Release Time | Core Die Density | Layer     | Speed (Gbps) | I/O  |
|---------|--------------|------------------|-----------|--------------|------|
| HBM4    | 2026F        | 24Gb             | 12, 16    | 8-10         | 2048 |
| HBM3e   | 2024         | 24Gb             | 8, 12, 16 | 8.0-9.8      | 1024 |
| HBM3    | 2022         | 16Gb             | 8, 12     | 5.6-6.4      | 1024 |
| HBM2e   | 2020         | 16Gb             | 4, 8      | 3.2-3.6      | 1024 |
| HBM2    | 2018         | 8Gb              | 4, 8      | 2.0-2.4      | 1024 |

Source: TrendForce, May 2025



## 4.1 HBM市场：HBM4预计于2026年推出，自2026年起开始量产

- HBM4预计规划于2026年推出，预计将在2026年第二季度量产。HBM迭代快速，集邦咨询预计HBM3E将占据2025年出货份额超过90%，2026年HBM4将开始渗透入市场，供货商预计2026年第二季度量产。随着客户对运算效能要求的提升，在堆栈的层数上，HBM4除了现有的12hi外，也将再往16hi发展。HBM4 12hi产品将于2026年推出；而16hi产品则预计于2027年问世。此外，受到规格更往高速发展带动，将首次看到HBM最底层的Logic die采用12nm制程wafer，该部分将由晶圆代工厂提供，使得单颗HBM产品需要结合晶圆代工厂与存储器厂的合作。

表：HBM产品进度情况

| 项目  | 详细内容                                                                                                                                                                                                 |
|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 海力士 | 2025年3月，SK海力士率先向主要客户交付了全球首款12层堆叠HBM4样品，引发行业关注。公司计划在2025年下半年完成12层HBM4产品的量产准备。<br>据财联社7/21讯，SK海力士正在讨论扩大HBM3E 8层产品的生产规模，其出货量预计将超出最初预期。<br>据财联社6/20讯，SK海力士首款定制HBM预计将于2026年下半年推出，目前已将英伟达、微软和博通作为其主要客户。    |
| 三星  | 据财联社7/21讯，三星已准备在2025年7月底前向AMD和英伟达等客户提供HBM4样品。三星计划应用10纳米级第六代（1c）DRAM工艺，开发更精密、良率更高的HBM4。                                                                                                               |
| 美光  | 据 TechNews 报道，美光在其AI内存产品组合中展示了下一代HBM4的计划，预计将在2026年实现量产，随后在2027至2028年推出HBM4E。<br>美光表示，其HBM4的带宽相比HBM3E提高超过60%，在数据吞吐能力上实现大幅跃升。与SK海力士类似，美光也选择与台积电（TSMC）合作开发定制化基底裸晶，并在接受采访时确认，HBM4E产品的研发进展顺利，且已有多家客户参与合作。 |

## 4.2 HBM：预计2025年三星产能领先，SK海力士市占超50%主导HBM市场

- 以HBM产能来看，三星、SK海力士（SK hynix）至2025年底的HBM产能规划较为积极，三星HBM总产能至年底将达约170K（含TSV）；SK海力士约150K，但产能会依据验证进度与客户订单持续而有变化。
- Counterpoint Research表示，SK海力士扩大了其在HBM领域的领先地位，第一季度市场份额达到70%。TrendForce预测，今年SK海力士的HBM市场份额将保持在50%以上，三星的份额将降至30%以下，美光科技的份额将升至近20%。

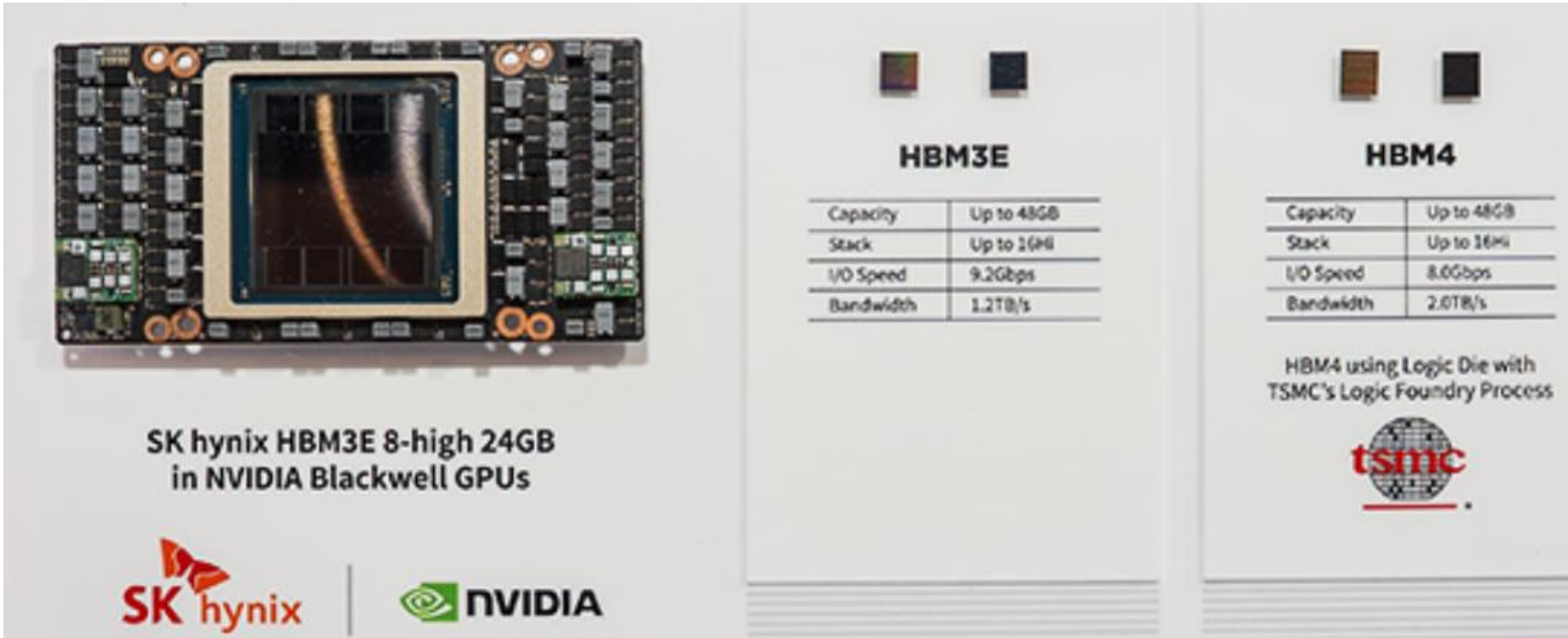
表：2023~2025年各供货商HBM/TSV产能预估

| HBM供货商          | 三星        | SK海力士     | 美光       |
|-----------------|-----------|-----------|----------|
| 2023年底HBM TSV产能 | 45000片/月  | 45000片/月  | 3000片/月  |
| 2024年底HBM TSV产能 | 120000片/月 | 120000片/月 | 25000片/月 |
| 2025年底HBM TSV产能 | 170000片/月 | 150000片/月 | 45000片/月 |

### 4.3 HBM：定制化HBM直击AI时代需求，预计2025年下半年实现

- **SK海力士领跑定制化HBM4E。**SK海力士已与NVIDIA、微软与博通展开合作，设计客户专属的定制化HBM芯片。首批商用定制HBM产品预计将于2025年下半年出货。SK海力士此前曾表示，从HBM4E开始，将全面转向定制化生产模式。当前主流为第五代HBM3E，产业正加速迈向第六代HBM4。为支撑该转型，SK海力士已与台积电合作代工逻辑HBM堆栈中的“核心大脑”。
- **HBM4逻辑基片设计定制化，提升AI性能与功耗效率。**据TrendForce称，目前的HBM3基础芯片采用纯内存架构，仅充当信号通路。相比之下，SK海力士和三星正在与代工厂合作，为HBM4采用基于逻辑的基片设计。这种新方法使HBM与SoC之间能够更紧密地集成，从而降低延迟、提高数据路径效率，并在高速传输环境中提供更高的稳定性。定制化HBM旨在满足AI时代日益复杂的性能与功耗效率需求，尤其是当超大规模客户逐步从通用内存迁移至针对AI工作负载优化的专属模组。

图：SK海力士的12层HBM4和16层HBM3E与NVIDIA的B100 GPU的合作展台





### 4.3 HBM： KAIST发布技术演进路线，展望未来发展方向

- 2025年6月，韩国国家级研究机构——韩国科学技术院（KAIST）发布HBM相关论文，详细介绍了HBM技术到2038年的演进，展示了带宽、容量、I/O 宽度和散热性能的提升。该路线图涵盖了从HBM4 到 HBM8 的技术发展，涵盖了封装、3D 堆叠、以内存为中心的嵌入式 NAND 存储架构，甚至还有基于机器学习的功耗控制方法。
- HBM性能参数全面跃升：容量、功耗、带宽与接口技术持续突破。**在容量方面，预计单堆栈容量将从HBM4的288-348GB跃升至HBM8的5,120-6,144GB；功耗管理上，单堆栈功耗将随性能提升从75W(HBM4)增至180W(HBM8)，并配套机器学习动态调频技术；带宽性能方面，预计2026年至2038年间，内存带宽将从2TB/s增长到64TB/s，数据传输速率将从8GT/s提升到32GT/s。每个HBM封装的I/O宽度也将从目前HBM3E的1,024位接口增加到HBM4的2,048位，之后将一路提升到HBM4的16,384位。

图：KAIST TERALAB制定的下一代HBM技术路线图

| Next-Generation HBM Roadmap by KAIST TERALAB |                                                |                                                      |                                                                |                                              |                                                                    |
|----------------------------------------------|------------------------------------------------|------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------|--------------------------------------------------------------------|
| Ver 1.2 / updated.250521                     |                                                |                                                      |                                                                |                                              |                                                                    |
|                                              | HBM4 (2026)                                    | HBM5 (2029)                                          | HBM6 (2032)                                                    | HBM7 (2035)                                  | HBM8 (2038)                                                        |
| Data Rate                                    | 8 Gbps                                         | 8 Gbps                                               | 16 Gbps                                                        | 24 Gbps                                      | 32 Gbps                                                            |
| # of I/O                                     | 2,048                                          | 4,096                                                | 4,096                                                          | 8,192                                        | 16,384                                                             |
| Bandwidth                                    | 2.0 TB/s                                       | 4 TB/s                                               | 8 TB/s                                                         | 24 TB/s                                      | 64 TB/s                                                            |
| Capacity/die                                 | 24 Gb                                          | 40 Gb                                                | 48 Gb                                                          | 64 Gb                                        | 80 Gb                                                              |
| # of die stack                               | 12/16-Hi                                       | 16-Hi                                                | 16/20-Hi                                                       | 20/24-Hi                                     | 20/24-Hi                                                           |
| Capacity /HBM                                | 36/48 GB                                       | 80 GB                                                | 96/120 GB                                                      | 160/192 GB                                   | 200/240 GB                                                         |
| Power/HBM                                    | 75 W                                           | 100 W                                                | 120 W                                                          | 160 W                                        | 180 W                                                              |
| Die stacking                                 | Microbump (MR-MUF)                             |                                                      | Bump-less Cu-Cu Direct bonding                                 |                                              |                                                                    |
| Cooling Method                               | Direct-to-Chip (D2C) Liquid Cooling            | Immersion Cooling                                    |                                                                | Embedded Cooling                             |                                                                    |
| HBM Architecture                             | Custom HBM Base Die HBM-LPDDR                  | 3D NMC-HBM & stacked cache / decap                   | Multi-tower HBM Active / Hybrid Interposer                     | Hybrid HBM Architecture HBM-HBF HBM-3D LPDDR | Full-3D / HBM Centric Computing Architecture                       |
| Additional Features (Patent)                 | NMC processor + LPDDR Ctrl                     | + Cache + CXL + on-die/stacked decap + HBM shielding | + Network switch + Bridge die + Asymmetric TSV                 | + HBF/LPDDR Ctrl + Storage network           | + HBM Centric Interposer + Double side Cooling + Edge-expand Stack |
| AI Design Agent                              | ubump & TSV-array Decap placement Optimization | I/O Interface Optimization considering PSIJ          | Hybrid Equalizer + Generative AI based SI/PI Metric Estimation | LLM based Human Interactive AI Design Agent  |                                                                    |

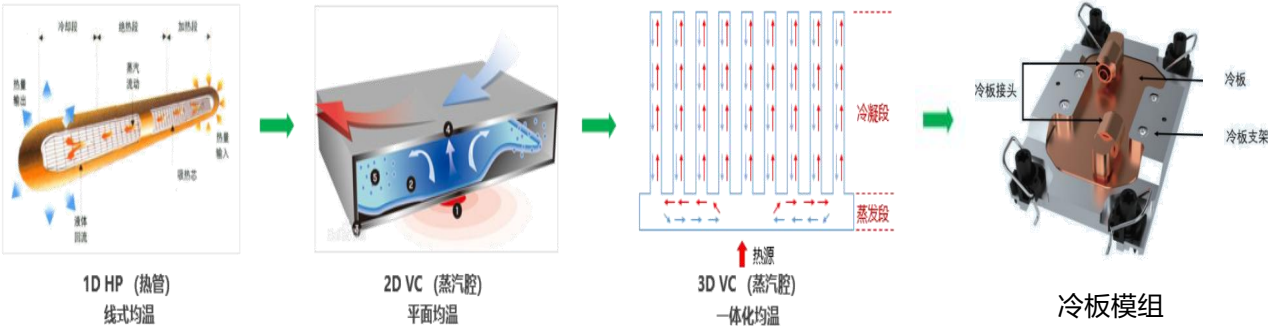
## 第五章 液冷

冷板式液冷较为成熟，GB300采用液冷方案

# 5.1 芯片散热革新：相比热管/VC/3DVC，冷板式散热范围广

- 随着芯片功耗的提升，从一维热管的线式均温，到二维VC的平面均温，发展到三维的一体式均温，即3D VC技术路径，最后发展到液冷技术。
- 热管与VC的散热能力较低，3D VC风冷散热上限扩大至1000W，冷板式具备1000W+的广阔散热范围。

图：热管、VC、3DVC 散热技术情况



图：散热方案趋势



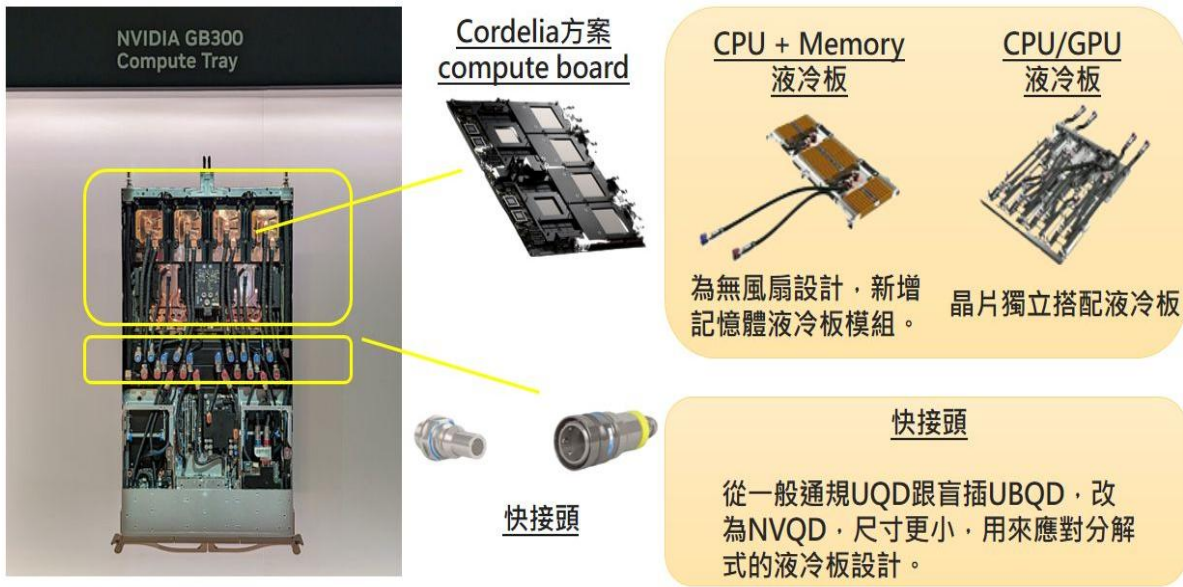
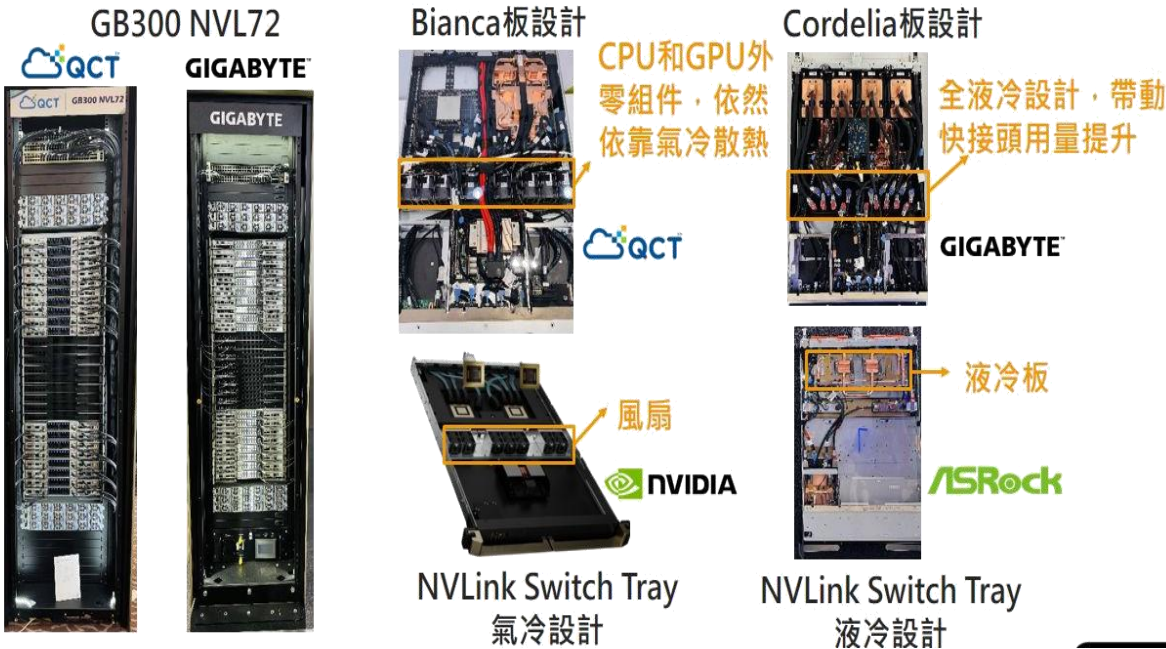


## 5.2 首批GB300出货仍沿用Bianca架构

- 首批GB300 NVL72出货仍沿用Bianca架构。根据EEPW公众号，为了加快部署速度，英伟达舍弃原先规划导入的Cordelia（多层PCB设计）的新芯片板设计，通知合作伙伴暂时回归目前GB200所使用的Bianca（HDI PCB）设计。

图：首批GB300 NVL72出货仍沿用Bianca

图：Cordelia方案Fanless架构提升液冷板及快接头用量





### 5.3 Cordelia架构将采用独立冷板设计，UQD尺寸缩小至1/3

- GB300的GPU单卡功耗达到1.4kW，假设比GB200增加25%，则机架潜在功耗约为150kW。
- 变化1：独立液冷板设计，精准冷却每个芯片。GB300的液冷方案与GB200不同，每个芯片配备单独的液冷板，一进一出通道，这一创新大大提升了冷却效率，避免了大面积冷板所带来的散热不均和系统复杂性。据汉深流体官网和CDCC公众号，Compute tray冷板用量将由GB200的54（2\*18+2\*9）片提升至GB300的126（6\*18+2\*9）片。
- 变化2：UQD尺寸缩小，至原来的1/3。GB300采用NVUQD03快接头，尺寸缩小至原来的1/3，虽单价下降但用量翻倍。

表：GB200和GB300架构计算托盘液冷需求增量对比

|                  | GB200      | GB300      | 增量幅度    |
|------------------|------------|------------|---------|
| 单机柜功耗            | 120kW      | 150kW      | +25%    |
| Compute tray冷板数量 | 54片/系统     | 126片/系统    | +133%   |
| 快接头数量            | 126对（至少）   | 270对（至少）   | +114%   |
| 快接头单价            | 70-80美元/颗  | 40-50美元/颗  | 约-40%   |
| 整柜液冷价值量          | 84830美元/机柜 | 99050美元/机柜 | +16.76% |

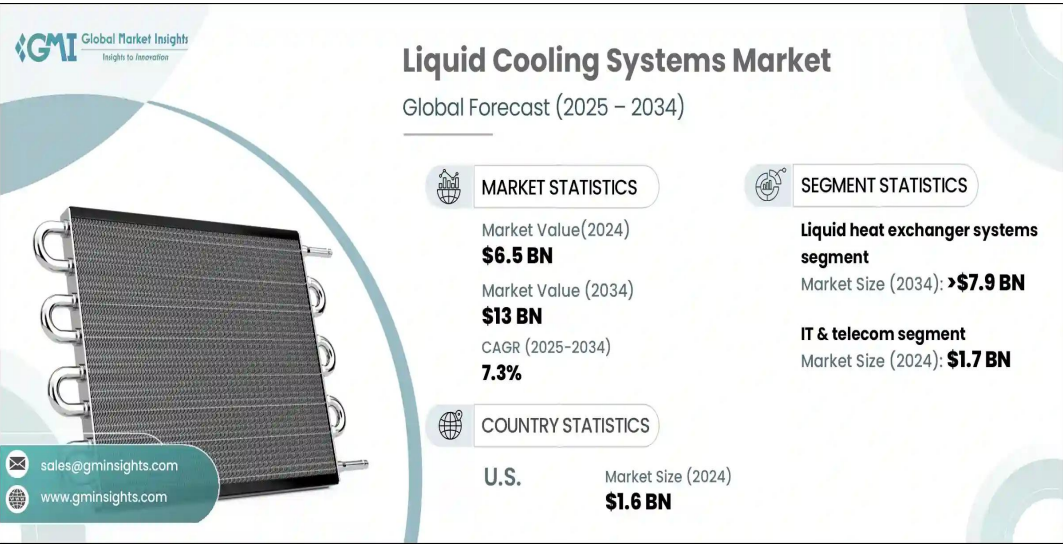
表：GB300 Cordelia架构液冷BoM拆解

| Rack-level | ASP（美元） | 数量                        | 金额（美元） | 占比    |
|------------|---------|---------------------------|--------|-------|
| 计算托盘冷板     | 300     | 6*18=108                  | 32400  | 32.7% |
| 计算托盘风扇     | 25      | 8*18=144                  | 3600   | 3.6%  |
| 交换托盘冷板     | 300     | 2*9=18                    | 5400   | 5.5%  |
| 交换托盘风扇     | 25      | 6*9=54                    | 1350   | 1.4%  |
| CDU        | 30000   | 1                         | 30000  | 30.3% |
| Manifold   | 12000   | 1                         | 12000  | 12.1% |
| UQD        | 50      | 14*18+2*9=270<br>（至少270对） | 13500  | 13.6% |
| 其它         |         |                           | 800    | 0.8%  |
| 合计         |         |                           | 99050  | 100%  |

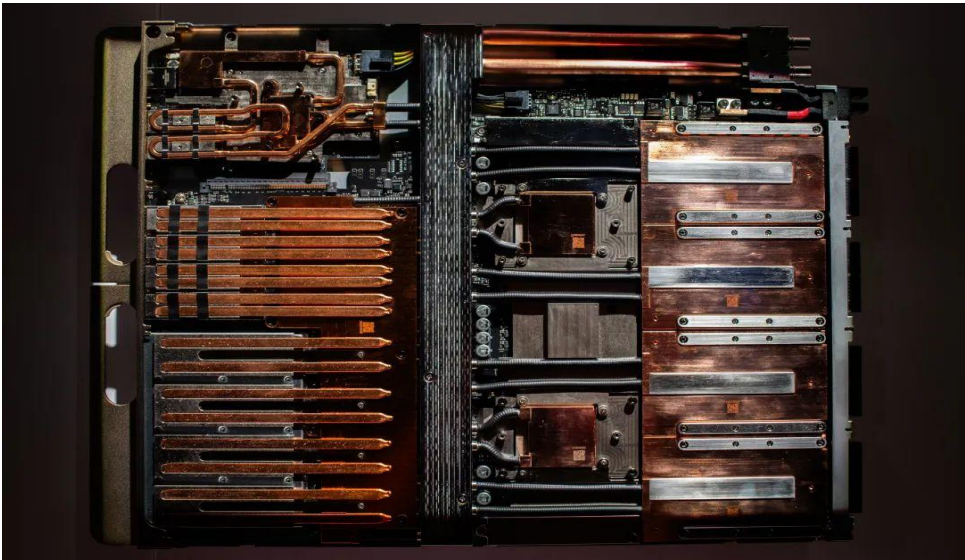
# 5.4 2026年GPU液冷市场有望达800亿元

- GB200/GB300出货量的扩大使液冷渗透率不断提升，液冷市场空间大开。TrendForce预估今年Blackwell GPU将占NVIDIA高阶GPU出货比例80%以上，随着GB200/GB300出货的扩大，液冷散热方案在高阶AI芯片的采用率正持续升高。**按照单机柜液冷价值量70万元计算，假设明年出货10-12万个机柜，我们预计仅GPU液冷市场就有望达800亿。**以Global Market Insights 2024年规模数据65亿美元为基准，2025-2026年CAGR约为30%。
- 非GPU/CPU部分也存在散热需求，整体制造成本有望显著提高。根据热管理行家公众号，英伟达计划2027年推出的Rubin Ultra NVL576——600kW等级的Kyber机架，将彻底摒弃风冷，实现100%液冷，并采用计算刀片（compute blade）。

图：Global Market Insights预测全球液冷市场规模

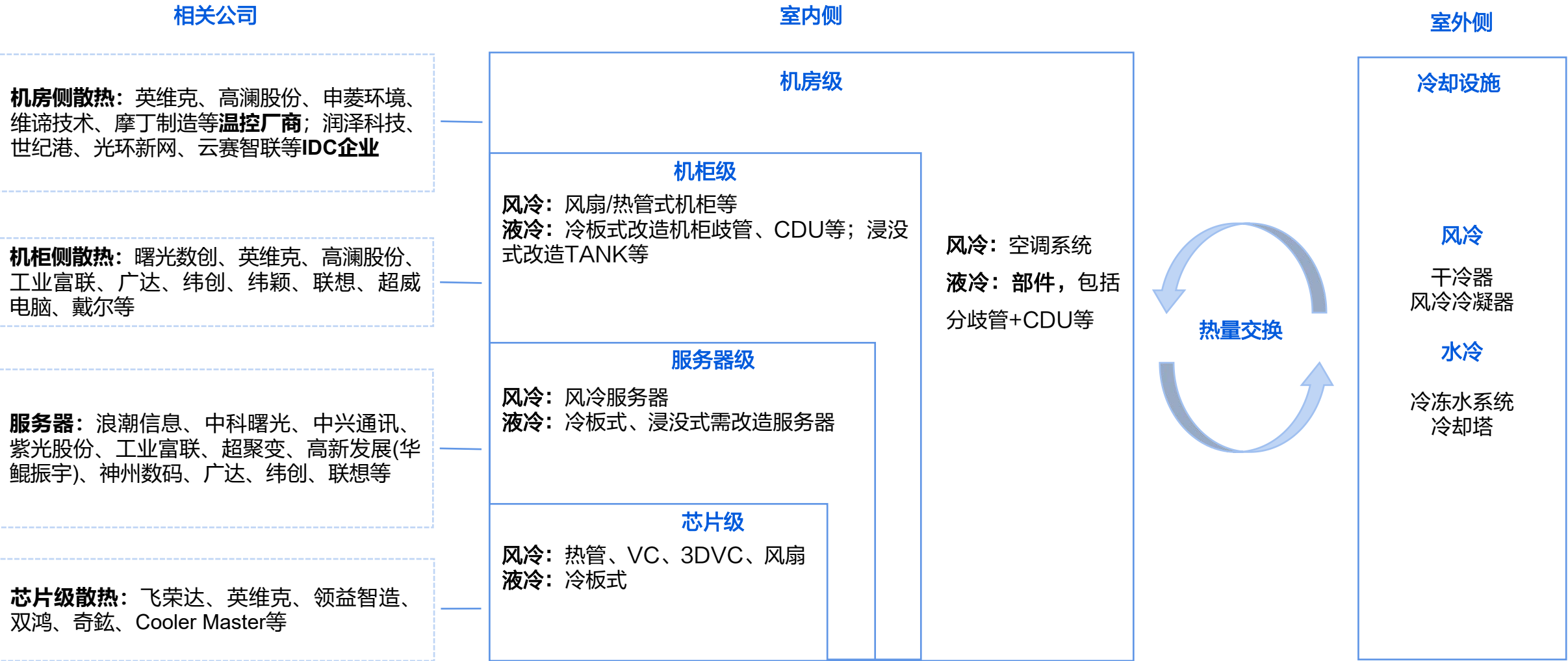


图：英伟达GTC 2025释出的刀锋服务器内部结构



# 5.5 服务器/机架/机柜级散热

图：数据中心温控系统框架





# 第六章

## 投资建议与风险提示

- 投资建议：大模型训推带动 AI算力需求增长，GB300、Vera Rubin等新一代算力架构将推出，算力产业链中的AI芯片、服务器整机、铜连接、HBM、液冷、光模块、IDC等环节有望持续受益。维持对计算机行业“推荐”评级。
- 相关公司
  - 1) AI芯片：寒武纪、海光信息、寒武纪、龙芯中科、景嘉微、英伟达、AMD、Intel
  - 2) 服务器整机：工业富联、浪潮信息、中科曙光、华勤技术、中国长城、高新发展、神州数码、烽火通信、拓维信息、纬创、广达、英业达、纬颖、超微电脑。
  - 3) 服务器组件：①散热：曙光数创、英维克、飞荣达、申菱环境、高澜股份、川环科技、奇鋆科技、双鸿、VERTIV；②主板：沪电股份、胜宏科技、深南电路、技嘉、华擎；③HBM：SK海力士、三星、美光、赛腾股份、联瑞新材；④铜连接：安费诺、沃尔核材、华丰科技。
  - 4) 光模块：天孚通信、中际旭创、新易盛、光迅科技、华工科技。
  - 5) 数据中心：奥飞数据、光环新网、宝信软件、数据港、电科数字、云赛智联、润泽科技。
  - 6) 算力租赁：协创数据、有方科技、宏景科技、安诺其。

- 1) 宏观经济影响下游需求：宏观经济环境下行，将影响客户对信息化基础设施的采购需求；
- 2) 大模型产业发展不及预期：行业的核心驱动力是人工智能大语言模型的训练和推理对算力的需求，如果大模型行业发展不及预期将会影响AI算力的相关需求；
- 3) AI芯片产业发展不及预期：人工智能芯片的推广与其技术和生态息息相关，尤其是针对国内的相关公司，如果AI芯片迭代放缓，或将影响AI算力行业的需求；
- 4) 市场竞争加剧：IT 产品和服务行业是成熟且完全竞争的行业，新进入者可能加剧整个行业的竞争态势；
- 5) 中美博弈加剧：国际形势持续不明朗，美国不断通过“实体清单”等方式对中国企业实施打压，若中美紧张形势进一步升级，将可能导致中国半导体供应链供应受到影响；
- 6) 相关公司业绩不及预期：市场环境变化、公司治理情况变化、其他非主营业务经营不及预期等原因或将导致相关公司的整体业绩不及预期。
- 7) 各公司并不具备完全可比性，对标的相关资料和数据仅供参考。



## 计算机小组介绍

刘熹，计算机行业首席分析师，上海交通大学硕士，多年计算机行业研究经验，善于把握由技术、政策驱动的科技产业新趋势，致力于进行前瞻重磅推荐。2024年金牌分析师，2024年同花顺最受欢迎分析师。

唐锦珂，计算机行业研究助理，中山大学岭南学院硕士，主要覆盖AI芯片、服务器、算力租赁、AIDC。

## 分析师承诺

刘熹，本报告中的分析师均具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立，客观的出具本报告。本报告清晰准确的反映了分析师本人的研究观点。分析师本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收取到任何形式的补偿。

## 国海证券投资评级标准

### 行业投资评级

- 推荐：行业基本面向好，行业指数领先沪深300指数；
- 中性：行业基本面稳定，行业指数跟随沪深300指数；
- 回避：行业基本面向淡，行业指数落后沪深300指数。

### 股票投资评级

- 买入：相对沪深300 指数涨幅20%以上；
- 增持：相对沪深300 指数涨幅介于10%～20%之间；
- 中性：相对沪深300 指数涨幅介于-10%～10%之间；
- 卖出：相对沪深300 指数跌幅10%以上。

## 免责声明

本报告的风险等级定级为R3，仅供符合国海证券股份有限公司（简称“本公司”）投资者适当性管理要求的客户（简称“客户”）使用。本公司不会因接收人收到本报告而视其为客户。客户及/或投资者应当认识到有关本报告的短信提示、电话推荐等只是研究观点的简要沟通，需以本公司的完整报告为准，本公司接受客户的后续问询。

本公司具有中国证监会许可的证券投资咨询业务资格。本报告中的信息均来源于公开资料及合法获得的相关内部外部报告资料，本公司对这些信息的准确性及完整性不作任何保证，也不保证其中的信息已做最新变更，也不保证相关的建议不会发生任何变更。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。报告中的内容和意见仅供参考，在任何情况下，本报告中所表达的意见并不构成对所述证券买卖的出价和征价。本公司及其本公司员工对使用本报告及其内容所引发的任何直接或间接损失概不负责。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或者金融产品等服务。本公司在知晓范围内依法合规地履行披露义务。

## 风险提示

市场有风险，投资需谨慎。投资者不应将本报告为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向本公司或其他专业人士咨询并谨慎决策。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息。本报告不构成本公司向该机构之客户提供的投资建议。

任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

## 郑重声明

本报告版权归国海证券所有。未经本公司的明确书面特别授权或协议约定，除法律规定的情况外，任何人不得对本报告的任何内容进行发布、复制、编辑、改编、转载、播放、展示或以其他方式非法使用本报告的部分或者全部内容，否则均构成对本公司版权的侵害，本公司有权依法追究其法律责任。

国海证券 · 研究所 · 计算机研究团队

# 心怀家国，洞悉四海



## 国海研究上海

上海市黄浦区绿地外滩中心C1栋  
国海证券大厦

邮编：200023

电话：021-61981300

## 国海研究深圳

深圳市福田区竹子林四路光大银行大厦28F

邮编：518041

电话：0755-83706353

## 国海研究北京

北京市海淀区西直门外大街168号腾达大厦25F

邮编：100044

电话：010-88576597