东兴证券 DONGXING SECURITIES

超节点与 Scale up 网络行业：谷歌、AMD、国产超节点持续发力，打破英伟达独大格局

2026 年 3 月 2 日

看好/维持

通信 | 行业报告

分析师 | 石伟晶 电话：021-25102907 邮箱：shi_wj@dxzq.net.cn | 执业证书编号：S1480518080001

投资摘要：

超节点与 Scale-up 网络是突破算力与通信瓶颈、支撑万亿级大模型与高实时性应用的关键基础设施。本篇超节点与 Scale up 网络行业深度报告，详细研究英伟达、谷歌、AMD 以及华为四家头部 AI 算力芯片厂商在此领域的布局进展以及各自优势。我们认为，超节点与 Scale-up 网络正处于快速发展期，并将成为算力芯片、网络部件（PCB 板、交换芯片、光器件、高速铜缆）、存储部件、供电和散热设施部件等新兴技术的重要应用市场。

（1）英伟达：超节点领先优势建立在 NVLink 和 NVLink Switch。

在超节点技术方案上，英伟达处于领先优势。2024-2025 年，英伟达陆续推出 GH200 NVL72、GB200/ GB300 NVL72 等成熟超节点解决方案。根据大摩预测，2025 年英伟达 GB200/300 NVL72 出货量约 2800 台。展望 2026-2027 年，英伟达计划推出 Vera Rubin NVL144 和 Rubin Ultra NVL576。互联 GPU 数将从 72 颗进一步向 576 颗发展。届时，英伟达将发布新一代 Kyber 机架，架构引入 NVLink Switch Blade（NVLink 交换机刀片），通过 PCB 中板替代传统 5000+根有源铜缆。可以看到，Rubin Ultra NVL576 仍保持较强的工程创新能力。

英伟达超节点的优势建立在 NVLink 和 NVLink Switch。为实现 AI 训练集群高带宽与低延迟数据传输，NVLink 重新设计通信架构，并引入一系列先进技术，包括网状拓扑、差分信号传输、流量调度信用机制、多 Lane 绑定技术、统一内存空间等。截止 2025 年，NVLink 5 Switch 实现支持单 GPU 到 GPU 带宽 1800GB/s，可构建 72 GPU 的 NVLink 域，总带宽达 130 TB/s（双向），支持 72 GPU 全互联通信。在后续计划中，NVSwitch Gen6 和 Gen7 的 GPU-to-GPU 通信带宽继续升级为 3.6TB/s。

但另一方面，Scale up 网络兴起源于满足大模型分布式训练和推理中的张量并行(TP)与专家并行(EP)。目前 AI 产业也在探索降低 TP 与 EP 规模的技术方案，从而降低 Scale up 网络规模的上限。我们认为，Scale up 网络的发展空间或限制英伟达在超节点领域的领先优势。为保持领先优势，实现 Scale up 网络和 Scale out 网络融合或将成为英伟达超节点新的发展趋势。

（2）华为：对外开放灵衢互联协议，超节点性能追赶英伟达。

国内 Scale Up 协议尚未统一，华为灵衢协议尚未被国内业界广泛接受。在 Scale Up 协议方面，华为推出灵衢协议，并从 2.0 版本起转向开放标准。除此之外，国内其他厂商正探索多种互联协议，包括中移 OISA、腾讯 ETH-X、高通量以太网 ETH+以及中兴通讯 OLink 等。为打破生态壁垒，国内正积极推动标准统一，比如工信部正牵头推动 CLink 协议，旨在形成统一的国内标准。

华为超节点依靠集群化方式实现性能追赶。Atlas 950 超节点预计 2026 年第四季度发布，相比英伟达同样将在 2026 年下半年上市的 NVL144 总算力 2.52 EFLOPS（FP8），其算力达到 8 EFLOPS（FP8）。此外，Atlas 950 超节点在内存容量 1152TB 与互联带宽 16.3PB/s，也实现大幅领先。我们认为，短期内，华为超

节点依靠集群化实现性能追赶，但在超节点复杂性、可靠性、功耗等维度需要平衡。从整体解决方案看，英伟达在超节点的芯片工艺、软件生态与系统集成上的优势仍难以撼动。

Atlas 950 超节点互联方案或将调整，显示华为超节点技术在标准化阶段仍需夯实。相比上一代超节点，华为 Atlas 950 超节点不再使用全光互联架构，其通过“柜内正交铜互联+柜间光互联”的混合设计，在机柜内部利用铜互联实现高可靠、低成本和低功耗的连接，跨机柜则通过光互联保障系统的可扩展性，从而在维持系统可扩展性的同时，有效控制总体拥有成本（TCO）。

（3）谷歌：建立光互联超节点，与英伟达形成不对称竞争。

谷歌 TPU 超节点建立成熟的光互联 Scale up 网络。从技术成熟度看，2023-2025 年谷歌陆续推出 TPU v4、TPU v5p、TPU v7 三代超节点，完成了技术路线探索和方案标准化。此外 TPU v7 也获得外部企业认可。2026 年，Anthropic 将直接从博通采购近 100 万颗 TPU v7 Ironwood AI 芯片，本地部署在其控制的数据中心。2027 年，谷歌将推出第 8 代 TPU，对标 Nvidia Vera Rubin。可以看到，届时谷歌 TPU 超节点的性能指标进一步优化提升。

谷歌 TPU 超节点竞争优势建立在 OCS 交换机，技术路线独树一帜。相比英伟达、华为、AMD 等超节点厂商，谷歌是全球首个将光电路交换机（OCS）大规模商用部署于 Scale up 网络的企业，技术路线独树一帜。谷歌 OCS 交换机，涉及精密光学、机械工程与半导体工艺的深度交叉应用，在光互联领域构筑一道高壁垒的技术护城河。

相较于电分组交换机，光电路交换技术具备诸多优势：光电路交换机可跨多代光收发模块技术复用、光电路交换机的每比特能耗较电分组交换机低数个数量级、光电路交换机引入的时延极小。

OCS 交换机商用落地存在多重困难：光电路交换机需扩展至数百个端口以支撑足够数量 NPU 互连；受限于光电路交换机的控制软件和反射镜配置时延，商用光电路交换机的交换时延通常为 10~20 毫秒；为降低链路功率，光电路交换机插入损需要控制在理想水平。

为搭建高性价比、大规模的光交换层，谷歌创新研发三大核心硬件组件：光电路交换机、波分复用光收发模块和光环形器。其中谷歌 Palomar 光电路交换机的光学核心模块是实现光转向功能的 MEMS 微反射镜；波分复用光收发模块是提升布线效率、支撑大规模且持续扩张数据中心的关键；光环形器是实现光电路交换机链路双向通信的核心器件，将所需的光电路交换机端口和光纤数量减半。

（4）AMD：UALink 成为重要开放标准，超节点有望成为英伟达有力竞品。

作为 Scale up 网络开放技术路线方，UALink 成为重要标准。2025 年 1.0 版本规范正式发布；2026 年，UALink 2.0 版本有望发布。我们认为，目前 UALink 正处于从标准制定阶段走向产品落地阶段，预计 UALink 生态将在 2027 年迎来突破发展，被众多数据中心接纳。目前 UALink 联盟受到业内广泛支持，截止 2026 年 1 月底，成员单位超过 100 家，将成为英伟达 NVLink 有利挑战方。

AMD 超节点 Helios 机架有望成为行业主流选择。Helios 机架采用双宽机架设计，宽度从 1 个机架提升到 2 个机架，在复杂性、可靠性和性能间实现良好平衡。从算力、内存、互联带宽等指标，MI455x 系列 Helios 机柜是目前业界最能挑战英伟达的 NVL72 机柜的竞品；而在功耗领域，对比 GB200 NVL72 机柜，Helios 机架优势显著。此外，双宽结构为未来升级预留物理空间，例如可扩展至 144GPU 配置，而无需重新设计机架基础设施。

投资策略：

自 2025 年开始，超节点成为 AI 算力网络重要的技术创新方向。从 AI 基建竞争维度，AI 芯片厂商从芯片算力性能竞争延续至芯片+Scale up 网络的双战场。因此，除了原先英伟达、华为、AMD 以及谷歌等芯片公司，全球更多厂商加入超节点赛道的竞争，包括微软、Meta、Amazon、中国移动、阿里巴巴、字节跳动、腾讯、百度、中科曙光、中兴通讯、浪潮信息、紫光股份（新华三）、海光信息、沐曦股份、恒为科技等。

全球超节点竞争格局尚未确立。英伟达目前处于领先地位，但谷歌、AMD、华为等巨头在超节点领域的持续发力已经对英伟达一家独大的格局构成挑战。从股价表现看，2023-2024 年，英伟达股价大幅跑赢谷歌、AMD 以及 A 股中证算力指数。但在 2025 年，英伟达股价累计涨幅 38%，显著落后于谷歌、AMD 以及 A 股中证算力指数。我们认为，在超节点技术发展中，市场将继续对谷歌、AMD 以及国产超节点板块价值重估。

投资建议：（1）看好谷歌、AMD 以及国内超节点厂商；（2）看好英伟达、谷歌与 AMD 超节点供应链，包括 PCB 背板、高速铜缆、光模块、供电与液冷系统等；（3）基于交换机及芯片是 Scale up 网络互联的关键设备，看好谷歌光路交换机（OCS）核心零部件供应商以及 UALink 标准下的交换机芯片研发商。

风险提示：（1）LLM 训练与推理技术路径变化；（2）超节点性能与功耗有待平衡；（3）受供应链影响，各厂商超节点出货量低于预期；（4）AI 应用端增长不及预期。

目 录

插图目录

表格目录

超节点与 Scale-up 网络是突破算力与通信瓶颈、支撑万亿级大模型与高实时性应用的关键基础设施。本篇超节点与 Scale up 网络行业深度报告，详细研究英伟达、谷歌、AMD 以及华为四家头部 AI 算力芯片厂商在此领域的布局进展以及各自优势。我们认为，超节点与 Scale-up 网络正处于快速发展期，并将成为算力芯片、网络部件（PCB 板、交换芯片、光器件、高速铜缆）、存储部件、供电和散热设施部件等新兴技术的重要应用市场。

1. LLM 训练要求高带宽与延迟，驱动超节点成为 AI 算力网络创新方向

大语言模型（LLM）参数规模从千亿级向万亿级乃至十万亿级演进，跨服务器张量并行（TP）成为必然选择；此外混合专家（MoE）模型在 Transformer 架构 LLM 中的规模化应用，更使跨服务器专家并行（EP）成为分布式训练和推理的关键技术需求。为应对 TP 和 EP 对网络带宽与延迟的极为严苛的要求，构建超高带宽、超低延迟的 Scale up 网络（纵向扩张网络）成为业界主流技术路径。

表1：AI 大语言模型训练中多种并行计算方式对比

并行方式	带宽要求	延迟要求	说明
张量并行(TP)	数百至数千 GB/s 级	延迟要求极高	将单个运算（如矩阵乘法）拆分到不同 GPU 上运行，通常在机内完成
专家并行(EP)	数百至数千 GB/s 级	延迟要求极高	基于不同的任务选择不同专家进行训练，引入 All to All 流量，适合机内完成
流水线并行（PP）	MB/s 至 GB/s 级	延迟要求较高	将模型的不同层划分为若干个阶段，每个阶段可以在不同的 GPU 上执行，通常在机间完成
数据并行（DP）	GB/s 级	延迟要求较高	将同一批数据分割成多个子集，并将每个子集分配给不同 GPU 上（模型实例相同）运行，通常在机间完成

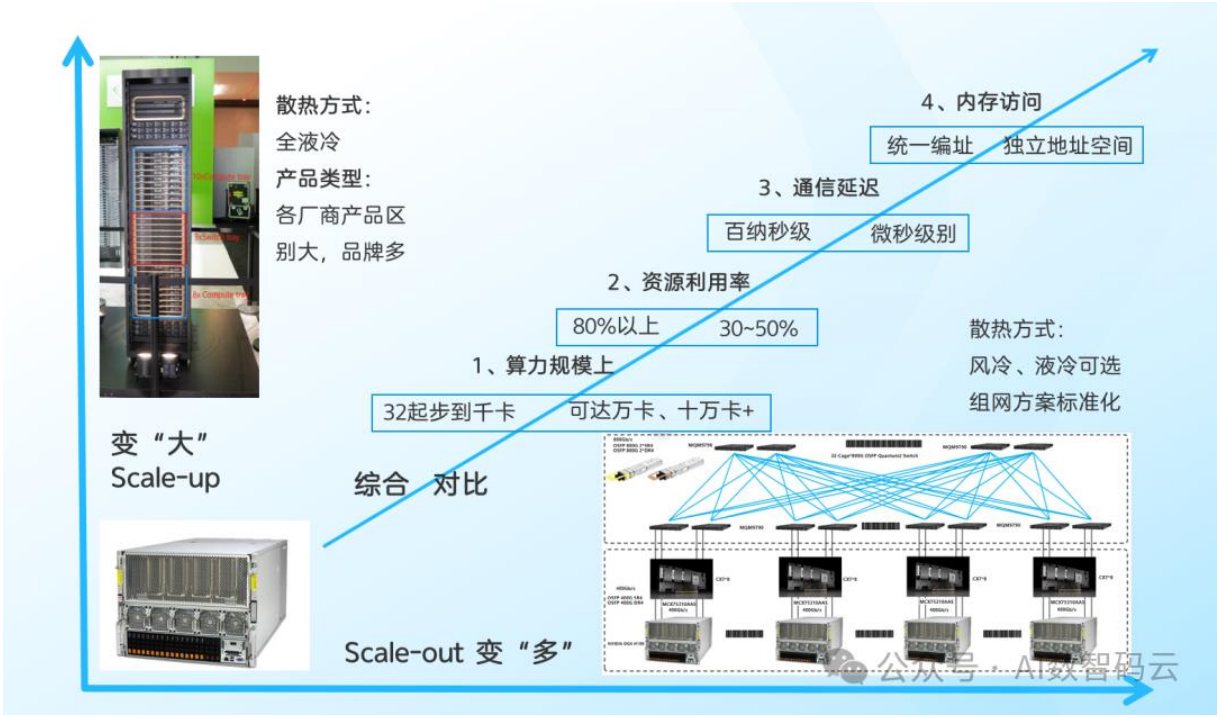
资料来源：网络技术趋势洞察公众号，东兴证券研究所

根据阿里云给出的定义为：Scale up 是在一定范围内，于成本和互联技术约束下实现的超高带宽互联。其范围固定且带宽是 Scale out 的数倍以上，可在协议层面优化以支持内存语义。我们对 Scale up 网络与 Scale out 网络特点对比如下：

Scale up（左）vs Scale out（右）

- 算力规模：数十卡至千卡级 vs 万卡至十万卡级；
- 资源利用率：80%以上 vs 30%-50%；
- 通信延迟：百纳秒级 vs 微秒级；
- 内存访问：统一内存或全局地址空间 vs 独立内存空间；
- 标准化：定制化程度高 vs 基于开放网络标准，相对统一。

图1：Scale up 网络（左）与 Scale out 网络（右）特点对比



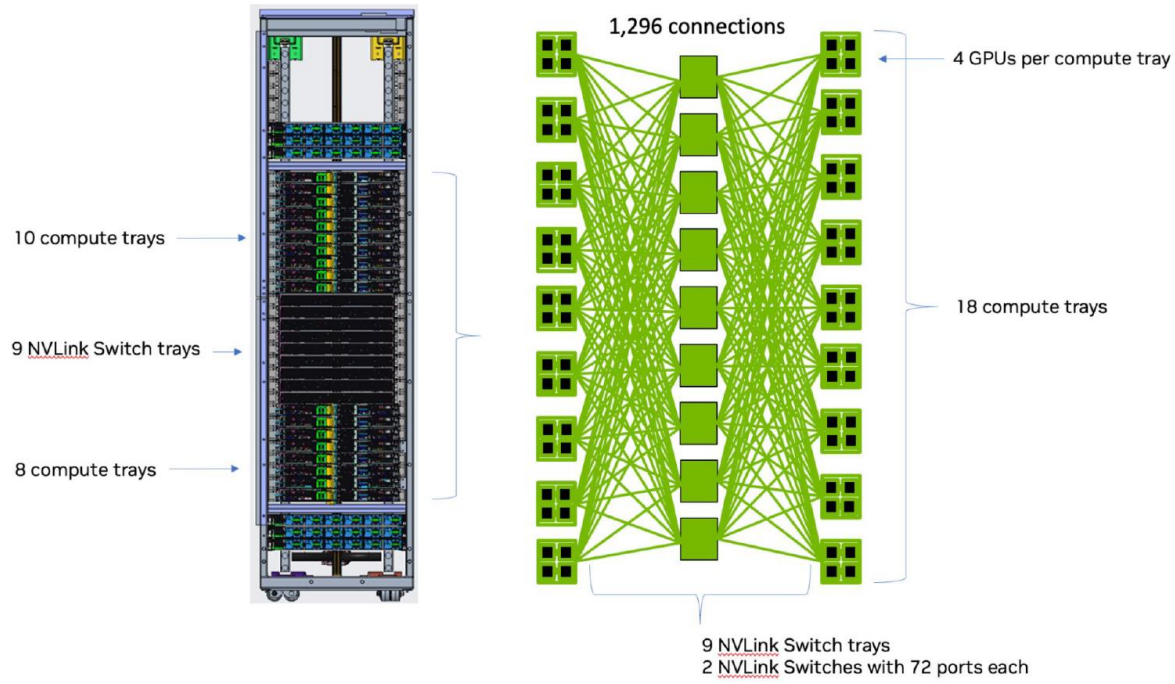
资料来源：AI 数智码云公众号，东兴证券研究所

超节点主要由计算节点、交换节点和 Scale-up 网络互联构成。通过 Scale up 网络，可将几十、上百甚至上千张 XPU 高速互联构建为超节点（SuperPoD），像一台超级 XPU 服务器一样实现高效的计算和通信协同能力。

其中 Scale up 网络互联是超节点的核心要素。Scale up 网络互联方案直接影响超节点系统的功耗、散热、成本、规模、可靠性和可维护性等关键指标。目前主流的互联方案有铜缆互联和光纤互联两大类：

- 铜缆互联方案（如英伟达的 NVL72 超节点及 NVSwitch Scale-Up 网络采用的 DAC 即无源铜缆技术）具有功耗低、成本低、可靠性高的明显优势。不过，受限于铜缆的信号传输距离，单个超节点的规模较小，目前商用的英伟达 NVL72 超节点最大支持 72 张 XPU 卡。
- 光纤互联方案(如华为的 CloudMatrix384 超节点及 Unified Bus (UB)Scale-Up 网络采用的 AOC 技术）则突破铜缆距离限制，超节点规模可以做的更大，目前商用的华为 CloudMatrix384 超节点可支持多达 384 张 XPU 卡，但这种互联技术方案也存在明显短板，如光模块功耗大，成本高，故障率高。

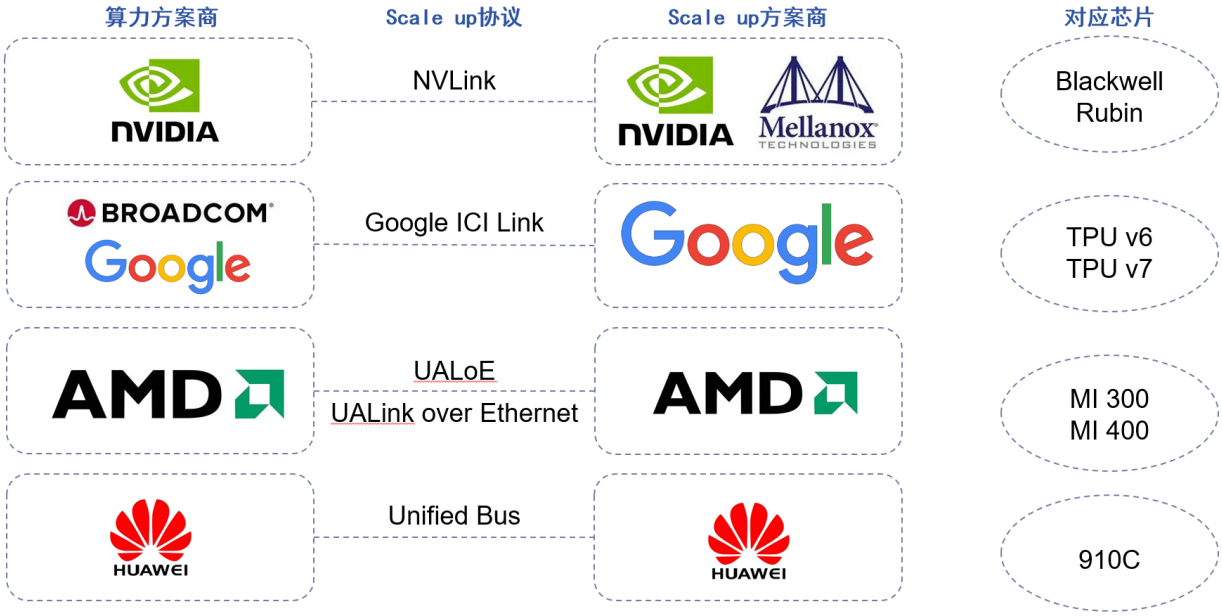
图2：英伟达 NVL72 超节点示意



资料来源：中国移动《超节点 Scale-Up 网络互联技术白皮书》，东兴证券研究所

目前英伟达、谷歌、AMD 以及华为四家头部 AI 算力芯片厂商均推出各自的 Scale up 协议。英伟达在 AI 数据中心的 Scale up 网络中采用自研的 NVLink 高速互连技术；AMD 与 AWS、思科、谷歌等公司组成超以太网联盟（UALink）；Google 采用私有 ICI 协议，机柜之间运用 OCS 光交换技术；华为推出自研的灵衢协议技术（UB）。

图3：全球主流算力方案对应 Scale Up 协议



资料来源：傅里叶的猫公众号，东兴证券研究所

Scale up 网络主要有两个技术方向。一是封闭的私有技术方向，以英伟达、Google 为典型代表，二者均采用专有协议：NVLink 仅向第三方半开放 CPU/Chiplet 接入权限；Google ICI Link 则服务于自研 TPU 集群；二是基于 Ethernet 的开放技术方向，以各大互联网和云计算公司以及一些 GPU 芯片公司为代表。开放标准以 UALink 和华为灵衢为代表，UALink 基于标准以太网组件打造开放互联协议，华为灵衢协议从 2.0 版本起转向开放标准。目前两者均处于生态建设初期。

图4：全球主流算力芯片厂商旗下 Scale up 协议特点

	NVIDIA NVLink™	Google ICI Link	ULTRA ACCELERATOR LINK™	灵衢 UnifiedBus
主导方	英伟达	谷歌	超以太网联盟（Meta、微软、英特尔、AMD 等）	华为
协议性质	专有协议	专有协议	开放标准	自灵衢 2.0 起开放标准
技术标准	基于 SerDes 的私有协议	ICI 协议（支持 3D 环面拓扑）	使用标准以太网组件	基于高速 SerDes 的物理层
连接形式	机柜内通过铜缆实现 GPU 互联	结合铜缆与光缆连接，立方体内部使用铜缆连接，立方体外部使用光缆连接	同时支持铜缆和光缆	机柜内使用电缆互联，不采用光模块
生态支持	NVLink Fusion，通过半定制 CPU 或 Chiplet 向第三方 GPU 开放生态系统	需通过 XLA 编译器将计算图转为 TPU 机器码，深度绑定 Google 生态	兼容多厂商 GPU（如 AMD Instinct MI 系列、英特尔 Gaudi 等）	全栈开放协议，处于建设初期，第三方商用 GPU 产品尚未大规模上市

资料来源：Semi Analysis，CSDN，东兴证券研究所

2. 英伟达：超节点领先优势建立在 NVLink 和 NVLink Switch

2.1 Scale up 网络核心技术：NVLink 与 NVLink 交换机

NVLink 与 NVLink 交换机是英伟达构建单机柜 Scale up 网络的核心技术组合。二者协同演进，从早期点对点互联发展到如今全互联通信，并支持多代 GPU 架构算力芯片。2026 年 1 月，英伟达发布第六代 NVLink 以及 NVLink 交换机，两者支持最新的 Rubin 架构。从性能指标看，第六代 NVLink 交换机支持的 GPU-to-GPU 通信带宽为 3.6TB/s；在 VR NVL72 系统中提供 260TB/s 聚合带宽。其中每 GPU 的 NVLink 带宽保持不变，与 NVLink5.0 一致，仍为 100GB/s。

图5：NVLink 技术规格参数对比

NVLink　NVLink 交换机

	第四代	第五代	第六代
每 GPU 的 NVLink 带宽	900GB/s	1,800GB/s	3,600 GB/s
每 GPU 的最大链路数	18	18	36
支持的 NVIDIA 架构	NVIDIA Hopper™ 架构	NVIDIA Blackwell 架构	NVIDIA Rubin 平台

资料来源：英伟达官网，东兴证券研究所

图6：NVLink 交换机规格参数对比

NVLink　NVLink 交换机

	NVLink 4 交换机	NVLink 5 交换机	NVLink 6 交换机
NVLink GPU 域	8	8 | 72	8 | 72
NVLink 交换机 GPU 到 GPU 带宽	900 GB/s	1,800 GB/s	3,600 GB/s
总聚合带宽	7.2 TB/s	130 TB/s (NVL72)	260 TB/s (NVL72)
支持的 NVIDIA 架构	NVIDIA Hopper™ 架构	NVIDIA Blackwell 架构	NVIDIA Rubin 平台

资料来源：英伟达官网，东兴证券研究所

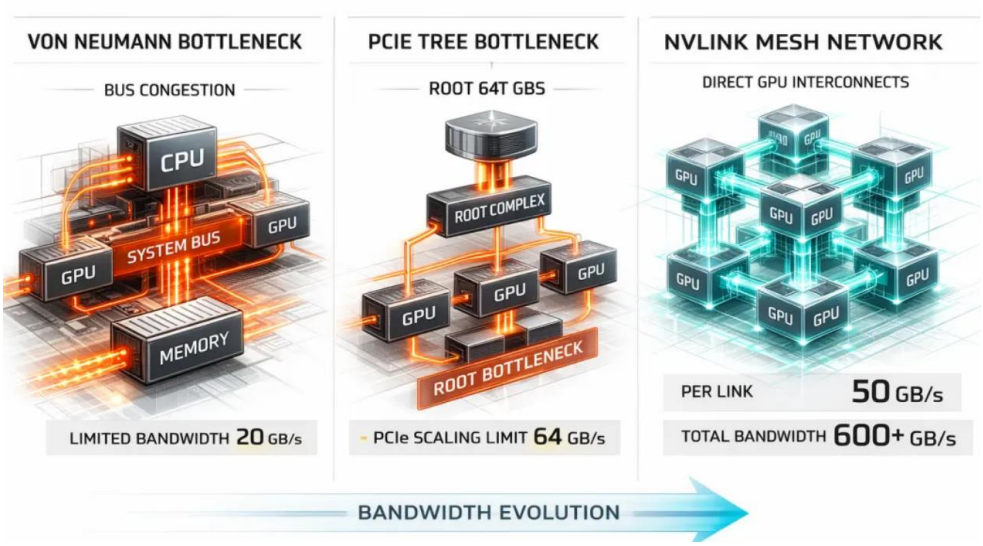
NVLink 重新设计通信架构，推出网状拓扑理念。为实现 AI 训练集群高带宽与低延迟数据传输，NVLink 允许 GPU 之间形成多对多的直接通信网络，每个 GPU 都可以同时与多个其他 GPU 建立高速通信链路。NVLink 协议创新如下：

在物理层面，NVLink 采用差分信号传输技术，具有高带宽和高抗干扰性能。每个链路由多对差分信号线组成，每对信号线负责传输一个方向的数据。SerDes 模块是 NVLink 物理层的核心组件，负责将并行数据转换为高速串行流，并在接收端进行反向转换。NVLink 的 SerDes 设计采用时钟数据恢复技术，以及集成复杂的自适应均衡电路。

在链路层，NVLink 定义多种类型的符号，包括数据符号、控制符号和填充符号，实现复杂的通信协议功能；设计精细的信用机制，实现不同优先级的流量调度。

除此之外，NVLink 其他创新之处包括多 Lane 绑定技术、统一内存空间等。

图7：NVLink 网状拓扑结构提供高速双向带宽



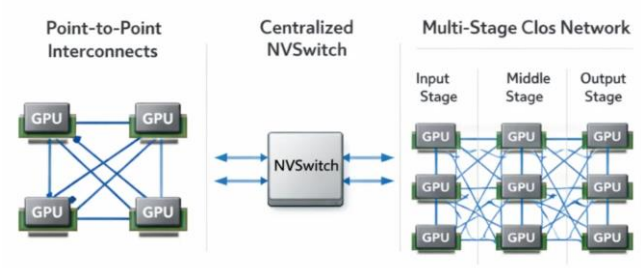
资料来源：仰望7866 公众号，东兴证券研究所

NVSwitch 是实现 Scale up 网络复杂交换的关键设备。

早期的 NVLink 实现主要采用点对点连接模式，GPU 之间通过直接的串行链路进行通信。当系统包含多个 GPU 时，点对点模式的连接复杂度呈平方级增长。

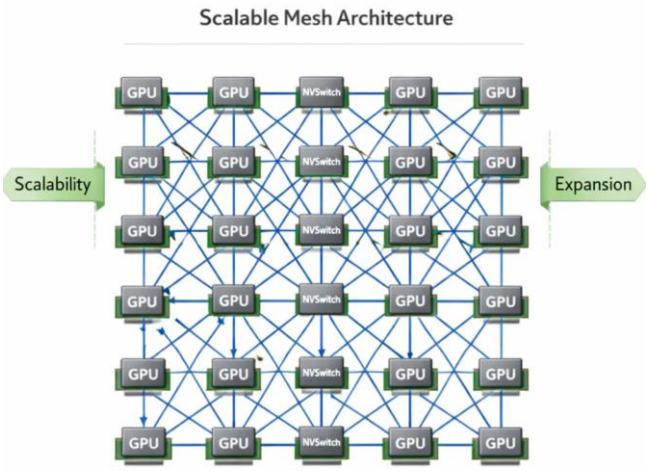
作为专门的交换芯片，NVSwitch 可以提供多端口的高速交换能力。NVLink 的交换网络采用多阶 Clos 网络架构，Clos 网络通过多级交换结构实现输入端口到输出端口的任意连接。

图8：NVLink 交换网络的演进过程（1）



资料来源：仰望 7866 公众号，东兴证券研究所

图9：NVLink 交换网络的演进过程（2）



资料来源：仰望 7866 公众号，东兴证券研究所

2.2 GB200 NVL72 超节点：铜缆互联，总交换容量 129.6TB/s

目前英伟达超节点已经推出成熟方案，在行业中处于领先地位。2024-2026 年，英伟达陆续推出 GH200 NVL72、GB200/ GB300 NVL72、VR200 NVL72 三代超节点。

- **Hopper 架构开启超节点 Scale up 初步探索。**GH200 通过 NVLink 和 NVLink-C2C（Chip-to-Chip）技术，使得每个 GPU 可以访问其他所有 CPU 和 GPU 芯片的内存，实现 GPU 与 CPU 内存统一编址。
- **Blackwell 架构推动 Scale up 标准化。**GB200 NVL72 将 Scale-up 规模稳定在 72 个 GPU/机柜，形成可复制标准化方案。NVL72 由 18 个 Compute Tray（计算托架）和 9 个 Switch Tray（网络交换托架）构成。其中，Compute Tray 是计算核心单元，负责提供强大的计算能力；Switch Tray 是高速通信枢纽，用于实现 GPU 之间的高速数据交换。NVL72 背板通过“NVLink5 私有协议 + 铜线缆”将 18 个 Compute Tray 中的 72 颗 B200 GPU 和 9 个 Switch Tray 中的 18 颗 NVSwitch 芯片进行满带宽全连接。
- **Rubin 架构推动 Scale up 方案带宽倍增。**2026 年 1 月 CES 展会，英伟达发布 Rubin 架构 VR200 NVL72。其中 NVLink 6 Switch 实现单 GPU 的互连带宽提升至 3.6 TB/s，上代为 1.8TB/s。Scale out 方面，Spectrum-6 交换机支持 CPO（共封装光学）技术，将 32 个 1.6Tb/s 硅光光学引擎与交换芯片直接封装集成。

图10：英伟达 GB300 NVL72 超节点外观

资料来源：传热之道公众号，东兴证券研究所

目前全球算力芯片公司进入芯片性能与超节点性能并行竞争的新阶段。GB200 NVL72 作为全球超节点发展的标杆产品，我们将从多个维度拆解其硬件构成以及重点性能指标。

从算力和通信性能看：GB200 NVL72 提供 180 PFLOP 的 TF32 Tensor Core 算力，总内存容量 13.8TB，内存带宽 576TB/s；Scale up 单向带宽 64800 GB/s。

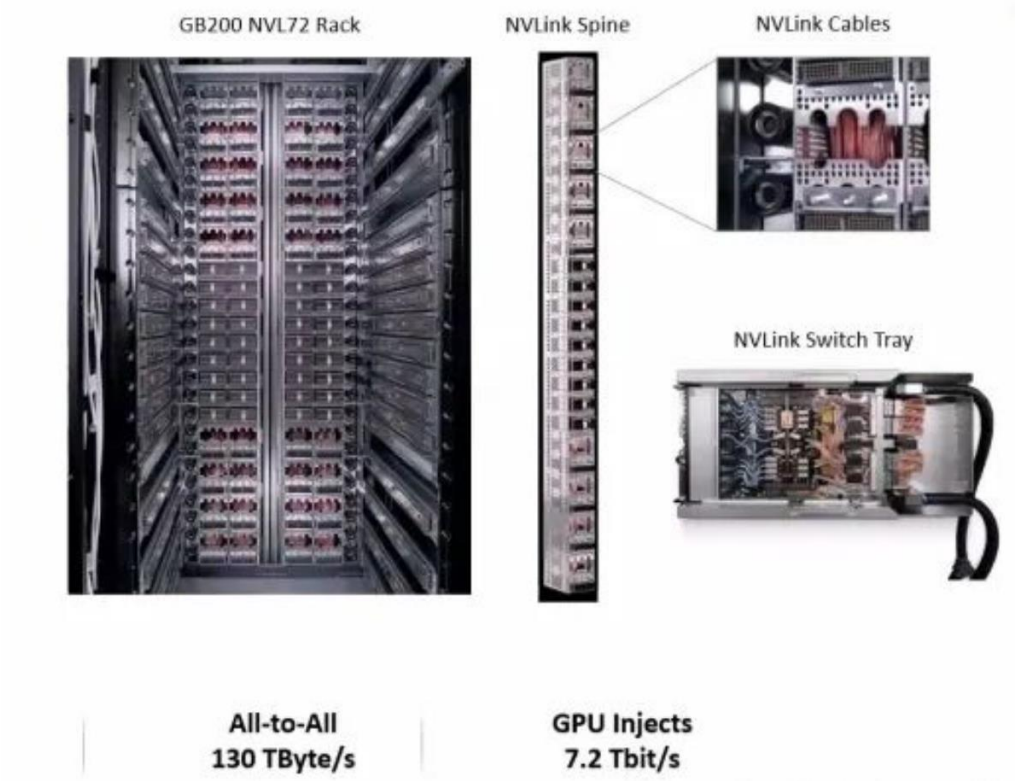
表2：GB200 NVL72 超节点算力与通信性能

	单位	GB200 NVL72
算力（TF32 Tensor 核心）	PFLOPS	180
HBM 内存	TB	13.4
HBM 带宽	TB/s	576
Scale up 带宽	单向 GB/s	64800
Scale up 计算单元	GPUs	72
功耗	KW	145

资料来源：SemiAnalysis，Nvidia，华为，东兴证券研究所

除了算力与通信性能，尺寸、重量、功耗均是超节点 TCO（总体拥有成本）的关键影响因素。GB200 NVL72 机柜尺寸为长 1068 毫米、宽 600 毫米、高 2495 毫米；重约 1.36 吨；功耗 145KW。

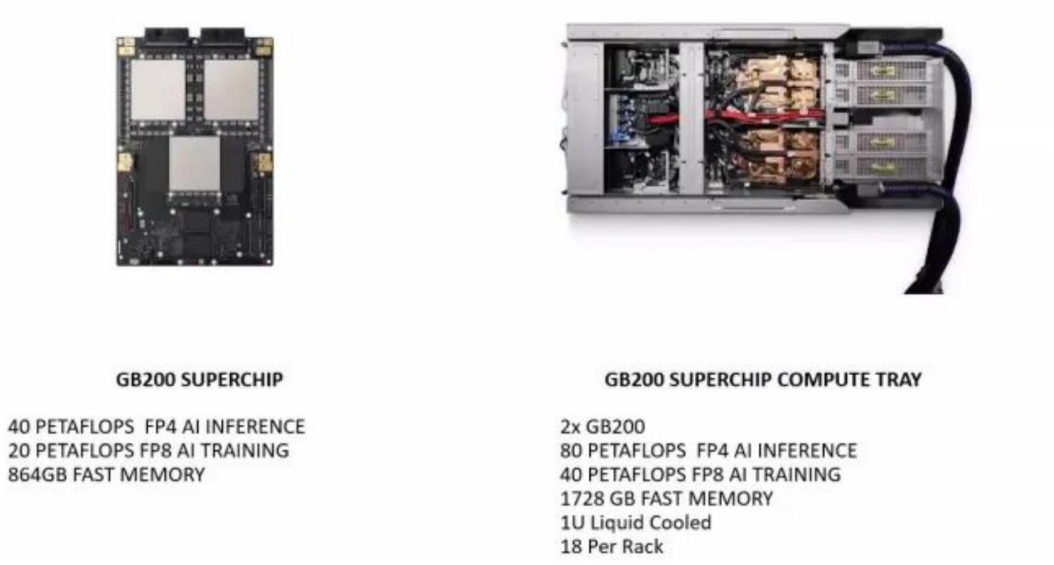
图11：GB 200 NVL72 机柜外观与内部构件细节



资料来源：光芯公众号，东兴证券研究所

单台 GB 200 NVL72 机柜有 18 个计算节点。GB200 NVL72 超节点主要由 18 个 Compute Tray（计算托盘）和 9 个 Switch Tray（网络交换托盘）构成。每个计算托盘容纳 4 颗 B200 GPU 和 2 颗 Grace CPU，构成两个 GB200 超级芯片。

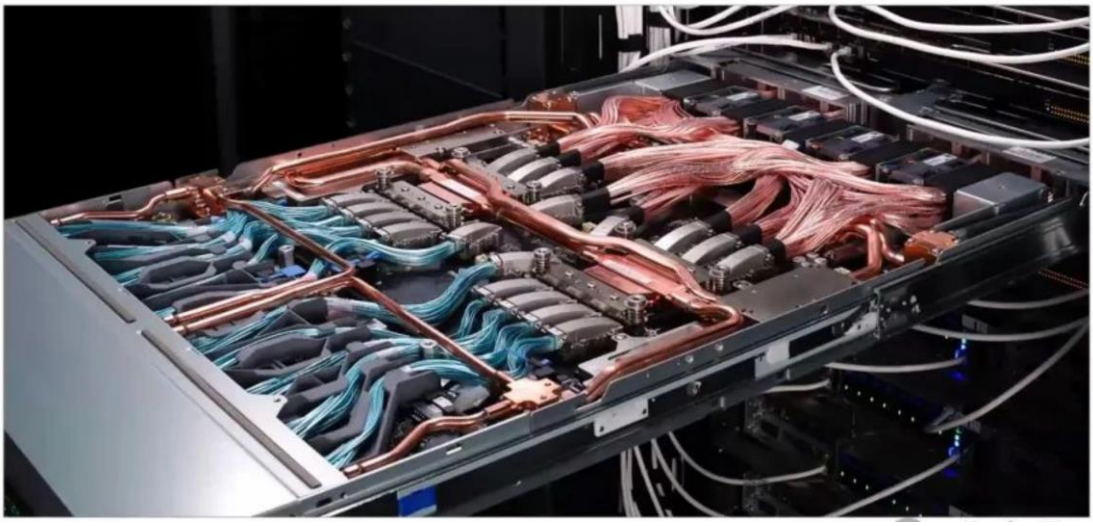
图12：GB200 NVL72 中计算托盘



资料来源：光芯公众号，东兴证券研究所

GB 200 NVL72 机柜有 9 个网络交换托盘。每个网络交换托盘中包含两颗 NVLINK Switch5 芯片，合计 18 颗 NVSwitch5 芯片。单颗 NVSwitch5 芯片交换容量为 7.2TB/s，总交换容量 129.6TB/s。网络交换托架中金色电缆用于 NVLink 连接，与电缆盒相连，机箱前面的蓝色电缆用于 OSFP 接口，实现不同版本的扩展。

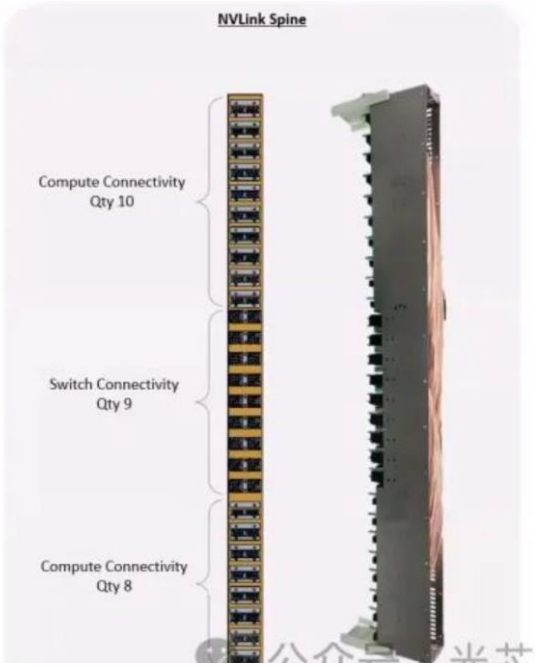
图13：GB 200 NVL72 中 NVLink 交换机托盘



资料来源：光芯公众号，东兴证券研究所

电缆盒负责垂直方向信号重组。电缆盒有 8 个底部连接器和 10 个顶部连接器，每个连接器可处理一个 GPU 的全部带宽。

图14：GB200 NVL72 中 NVLink 电缆盒



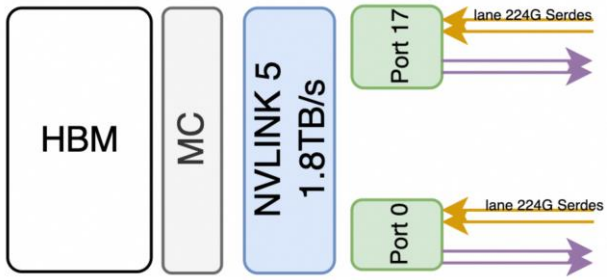
资料来源：光芯公众号，东兴证券研究所

GB200 NVL72 实现 72 颗 B200 完全互联，总交换带宽 129.6TB/s。

计算节点访存带宽为 7.2TB/s：B200 设置 18 个端口（Port）。每个端口采用 224G Serdes，由四对差分线构成。每个端口的传输速率为 200Gbps * 4（4 对差分线）/8 =100GB/s（双向）。每个计算托盘容纳 4 颗 B200 GPU，则每个计算节点 72 个 NVLink5 Port，总访存带宽为 7.2TB/s。

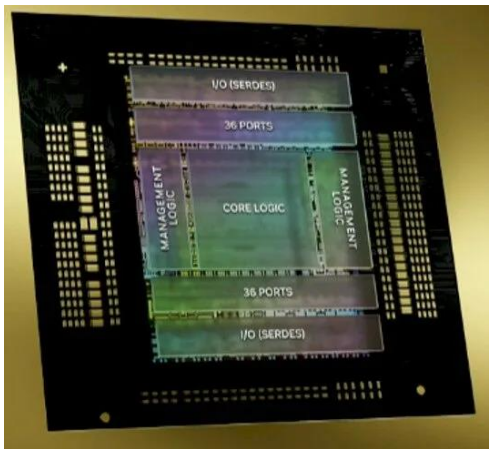
交换节点访存带宽为 14.4TB/s：NVSwitch5 芯片由 72 个 NVLINK Port（上下各 36 个 Port）。同样，每个 Port 采用双路 200Gbps 速率的 SerDes 高速串行接口，则每个 Port 带宽为 100GB/s。每个交换托盘两颗 NVLINK Switch5 芯片。每个交换节点 144 个 NVLINK Port，总访存带宽为 14.4TB/s。

图15：B200 端口 Port 示意图



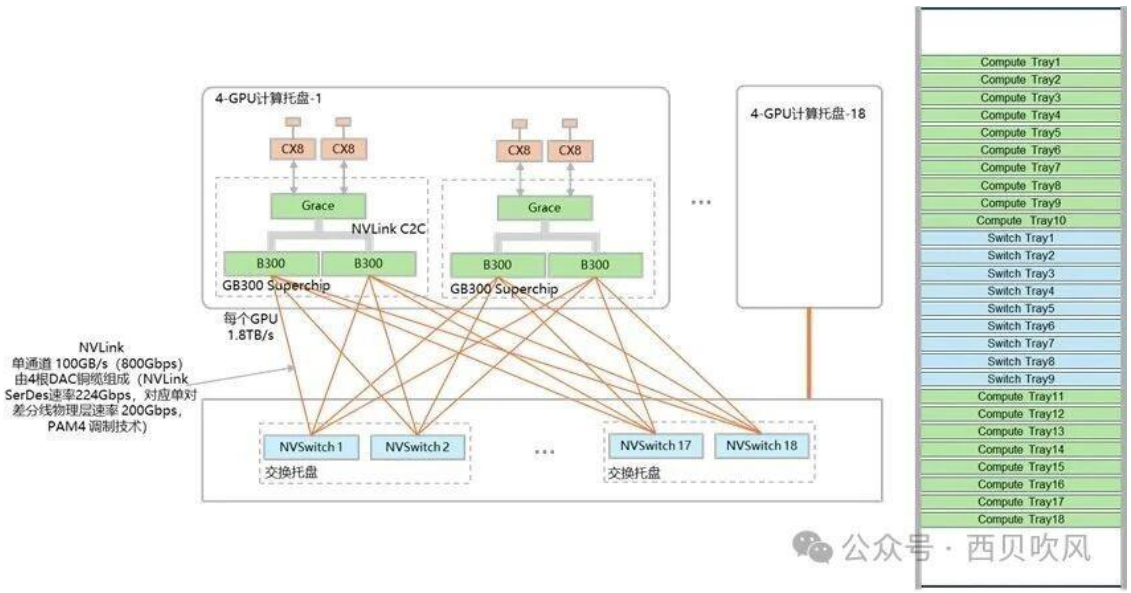
资料来源：zartbot 公众号，东兴证券研究所

图16：NVLINK Switch5 芯片 Port 示意图



资料来源：zartbot 公众号，东兴证券研究所

图17：GB200/300 NVL72 单层计算托架的互联拓扑



资料来源：西贝吹风公众号，东兴证券研究所

GB200 NVL 72 Scale up 方案中以铜缆互联为主。GB200 NVL72 在互联方案中主要采用直连铜缆（DAC），在某些特殊场景（如跨托盘连接或需要稍长传输距离的场景）中，会采用 ACC 铜缆。ACC（主动铜缆，在 DAC 基础上增加有源信号处理芯片）的信号增强能力可以弥补 DAC 在较长距离传输时的信号衰减问题，确保数据传输的稳定性和可靠性。

在 GB200 NVL 72 中所需铜缆数量：18（托盘数量）×4（GPU 数量）×4（GPU 到 NVSwtich 单端口铜缆数量）×18（NVSwtich 数量）=5184 根。（100GB/s 单端口由 4 根 DAC 铜缆组成）

图18：英伟达 GB200 NVL72 机柜后置铜线背板

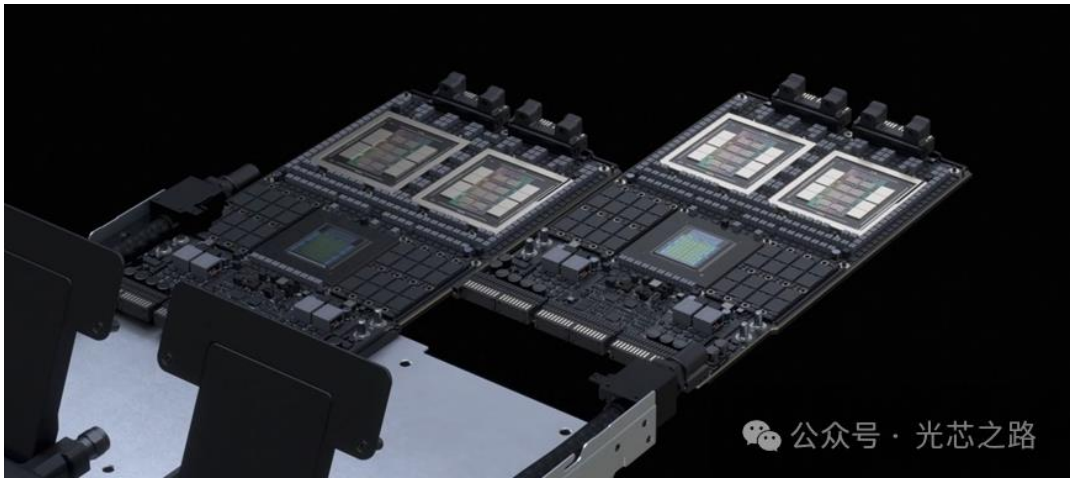
资料来源：英伟达 GTC，东兴证券研究所

2.3 VR200 NVL72 超节点：延续 GB200 NVL72 工程技艺，总交换容量翻倍

2026 年 1 月 6 日在 CES 2026 展会上，英伟达发布新一代超节点 VR200 NVL72。我们认为，**相比 GB200 NVL72，新一代 VR200 NVL72 属于连续性创新，而非破坏性创新**。具体对比如下：

计算节点：Rubin NVL72 机架通过无缆互联架构整合 18 个计算托盘，每个托盘 2 颗超级芯片，每颗超级芯片集成 1 个 Vera CPU 与 2 块 Rubin GPU，共 72 GPU 与 36 CPU。

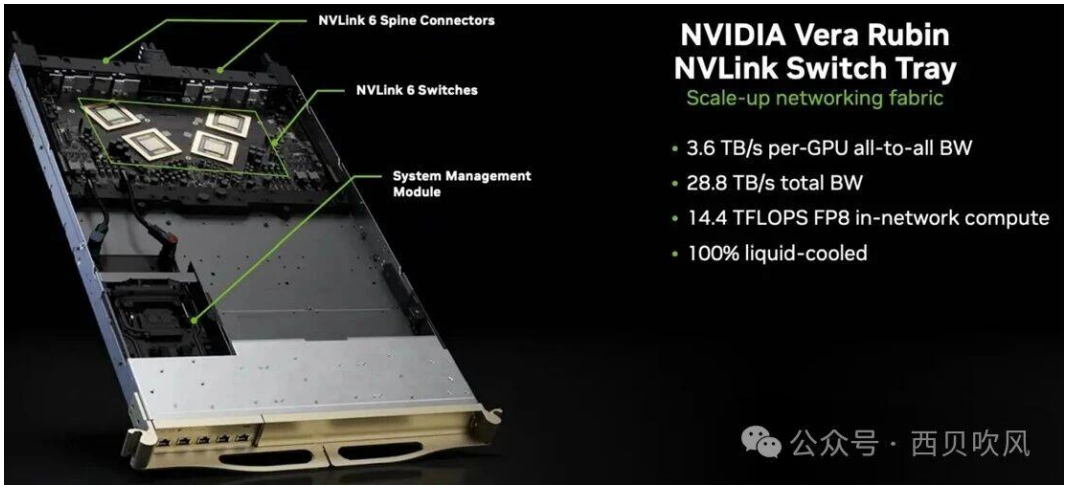
图19：VR200 NVL72 机柜计算托盘



资料来源：光芯之路公众号，东兴证券研究所

交换节点：VR200 NVL72 配置 9 个交换托盘，每个托盘集成 4 颗第六代 NVSwitch 芯片，全机柜部署 36 颗 NVSwitch。相比 GB200 NVL72, NVSwitch 芯片数量实现翻倍。单颗第六代 NVSwitch 交换容量为 7.2TB/s，相比 NVSwitch5 芯片，保持不变。

图20：Vera Rubin NVL72 机柜交换机托盘



资料来源：西北吹雪公众号，东兴证券研究所

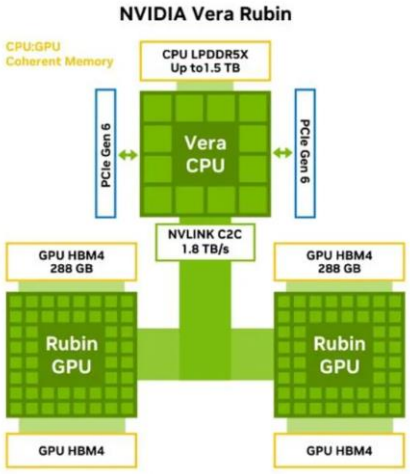
VR200 NVL72 Scale up 方案实现总交换容量 259.2TB/s，对比 GB200 NVL72，提升一倍。

计算节点：VR200 设置 72 个端口。每个端口带宽 100GB/s。每个计算托盘 2 颗 VR200 GPU，则每个计算节点 144 个 NVLink 6.0 端口，总访存带宽为 14.4TB//s。

交换节点：NVSwitch6 芯片 72 个 NVLink 6.0 端口。每个 NVLinkPort 交换容量 100GB/s。每个交换托盘 4 个 NVLink 6 Switch 芯片，每个交换节点 288 个 NVLink Port，总访存带宽为 28.8TB/s。

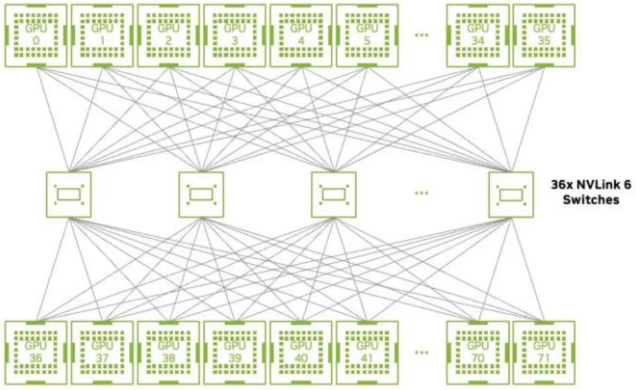
此外，VR200 NVL72 依托 NVLink-C2C 实现 1.8TB/s CPU-GPU 互联，相比 GB200 NVL72 的 NVLink-C2C 的速率为 900GB/s，提升一倍。

图21：英伟达 Vera Rubin GPU 芯片互联方式



资料来源：英伟达官网，东兴证券研究所

图22：VR200 NVL72 机柜中 GPU 互联拓扑结构



资料来源：英伟达官网，东兴证券研究所

VR200 NVL72 Scale up 方案延续铜缆互联方案。稍有不同之处，VR200 用中板取代计算托盘内部的线缆，中板采用覆铜板技术。此外，基于 Rubin 平台 NVLink6.0 升级至 448G SerDes 通道速率。因此，GPU 到每个 NVSwtich 铜缆连接由 4 根变为 2 根。

主干铜缆数量= 18（托盘数量）* 4（GPU 数量）* 2（GPU 到 NVSwtich 铜缆数量）* 36（NVSwtich 数量）= 5184 根。

图23：Vera Rubin NVL72 机柜交换机托盘无缆线设计

资料来源：西北吹雪公众号，东兴证券研究所

2.4 总结：处于领先优势，互联 GPU 数将从 72 颗进一步向 576 颗发展

在超节点技术方案上，英伟达处于领先优势。2024-2025 年，英伟达陆续推出 GH200 NVL72、GB200/ GB300 NVL72 等成熟超节点解决方案。根据大摩预测，2025 年英伟达 GB200/300 NVL72 出货量约 2800 台。展望 2026-2027 年，英伟达计划推出 Vera Rubin NVL144 和 Rubin Ultra NVL576。互联 GPU 数将从 72 颗进一步向 576 颗发展。届时，英伟达将发布新一代 Kyber 机架，架构引入 NVLink Switch Blade（NVLink 交换机刀片），通过 PCB 中板替代传统 5000+根有源铜缆。可以看到，Rubin Ultra NVL576 仍保持较强的工程创新能力。

英伟达超节点的优势建立在 NVLink 和 NVLink Switch。为实现 AI 训练集群高带宽与低延迟数据传输，NVLink 重新设计通信架构，并引入一系列先进技术，包括网状拓扑、差分信号传输、流量调度信用机制、多 Lane 绑定技术、统一内存空间等。截止 2025 年，NVLink 5 Switch 实现支持单 GPU 到 GPU 带宽 1800GB/s，可构建 72 GPU 的 NVLink 域，总带宽达 130 TB/s（双向），支持 72 GPU 全互联通信。在后续计划中，NVSwitch Gen6 和 Gen7 的 GPU-to-GPU 通信带宽继续升级为 3.6TB/s。

但另一方面，Scale up 网络兴起源于满足大模型分布式训练和推理中的张量并行(TP)与专家并行(EP)。目前 AI 产业也在探索降低 TP 与 EP 规模的技术方案，从而降低 Scale up 网络规模的上限。我们认为，Scale up 网络的发展空间或限制英伟达在超节点领域的领先优势。为保持领先优势，实现 Scale up 网络和 Scale out 网络融合或将成为英伟达超节点新的发展趋势。

表3：英伟达超节点 Scale up 迭代路线

架构	Blackwell Ultra	Vera Rubin NVL72	Vera Rubin NVL144	Rubin Ultra NVL576	Feunman
首发时间	2025-03	2026-01	预计 2026 下半年	预计 2027 年	预计 2028 年
核心平台	GB300 NVL72	VR200 NVL72	VR200 NVL144	Rubin Ultra NVL576	Feynman NVL1152
计算托盘	18 个（单盘 4GPU+2CPU）	18 个（单盘 4GPU+2CPU）	36 个（单盘 4GPU+2CPU）	72 个（单盘 8GPU+4CPU）	144 个（单盘 8GPU+4CPU）
CPU	36 Grace CPUs（72 核）	36 Vera CPUs（72 核）	72 Vera CPUs（88 核）	288 Vera Ultra CPUs（176 核）	576 Feynman CPUs（256 核）
	单颗内存带宽 3.6Tpbs	单颗内存带宽 4.8Tpbs	单颗内存带宽 4.8Tpbs	单颗内存带宽 9.6Tpbs	单颗内存带宽 19.2Tpbs
	NVLink C2C 0.9Tpbs	NVLink C2C 1.8Tpbs	NVLink C2C 1.8Tpbs	NVLink C2C 3.6Tpbs	NVLink C2C 7.2Tpbs
GPU	72 GB300 GPUs	72 VR200 GPUs	144 VR200 GPUs	576 VR300 GPUs	1152 Feynman GPUs
	单颗 288G HBM3E	单颗 512G HBM3E	单颗 512G HBM3E	单颗 1TB HBM4E	单颗 2TB HBM5E
	单颗 MVFP4 15PFLOPS	单颗 MVFP4 50PFLOPS	单颗 MVFP4 50PFLOPS	单颗 MVFP4 100PFLOPS	单颗 MVFP4 200PFLOPS
Scale up	铜缆背板	铜缆背板	铜缆背板+板载无源光引擎（非 CPO）	3.2T CPO 硅光（规划）	6.4T CPO 硅光（规划）
	18 个 NVLink5	36 个 NVLink6	72 个 NVLink6	144 个 NVLink7	72 个 NVLink8 硅光交换机
	单个 144*200G 28.8T	单个 72*400G 28.8T	单个 72*400G 28.8T	单个 144*800G 115.2T	单个 288*1.6T 460.8T
	GPU 侧 NVLink 带宽 18*1.8TBps	GPU 侧 NVLink 带宽 18*3.6TBps	GPU 侧 NVLink 带宽 18*3.6TBps	GPU 侧 NVLink 带宽 36*7.2TBps	GPU 侧 NVLink 带宽 72*14.4TBps

架构	Blackwell Ultra	Vera Rubin NVL72	Vera Rubin NVL144	Rubin Ultra NVL576	Feunman
Scale out	Spectrum-5 800G OSFP 可插拔	Spectrum-6 CPO 硅光	Spectrum-6 CPO 硅光	Spectrum-7 CPO 硅光	Spectrum-8 CPO 硅光
	64*800G 51.2T	128*800G 102.4T	128*800G 102.4T	256*1.6T 409.6T	512*3.2T 1638.4T

注：28.8Tbps=3.6TBps

信息来源：光芯之路公众号，东兴证券研究所

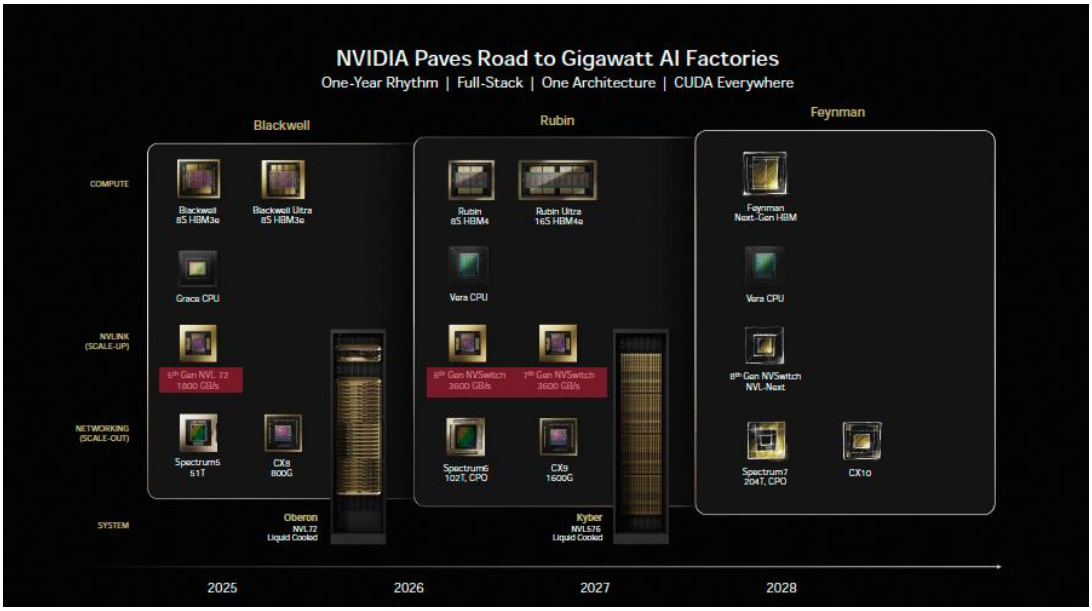
Rubin NVL576 由单个计算柜配置一个 Kyber SideCar 机柜构成。计算柜内由 4 个 NVL144 的计算机框构成，每个计算框包含 18 个 ComputeTray；Switch Blade 垂直插入机架后部，与计算刀片通过中板直接连接。

图24：英伟达 Rubin NVL576 新一代 Kyber 机架

资料来源：GTC2025，东兴证券研究所

后续 NVSwitch Gen6 和 Gen7 的 GPU-to-GPU 通信带宽为 3.6TB/s。

图25：英伟达算力芯片发布时间表



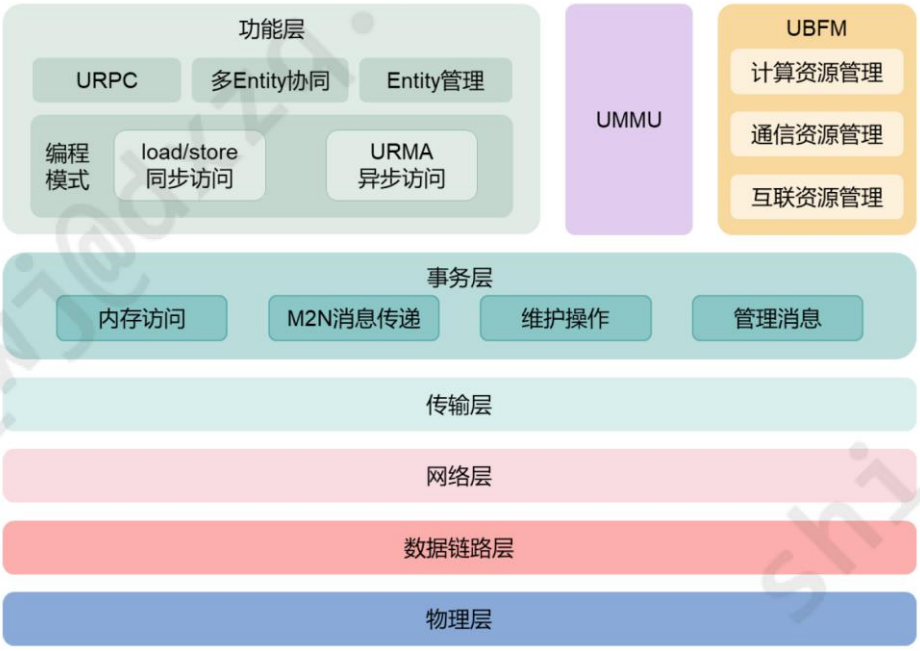
资料来源：GTC2025，东兴证券研究所

3. 华为：对外开放灵衢互联协议，超节点性能追赶英伟达

3.1 华为自研灵衢互联协议，并对外开放

对标英伟达UVLink，华为自研Scale Up网络协议灵衢。2025年，华为在全联接大会上发布的灵衢”（Unified Bus，简称UB）。根据华为官方白皮书定义，灵衢（UnifiedBus，UB）是一种面向超节点的互联协议，将IO、内存访问和各类处理单元间的通信统一在同一互联技术体系，实现高性能数据搬移、资源统一管理、资源灵活组合、处理单元间高效协同和高效编程。我们认为，华为UB协议栈定义全面完整，涉及物理层、数据链路层、网络层、传输层、事务层、功能层、UB内存管理单元、UBFM（负责系统内的计算资源管理、互连资源管理和通信管理）。

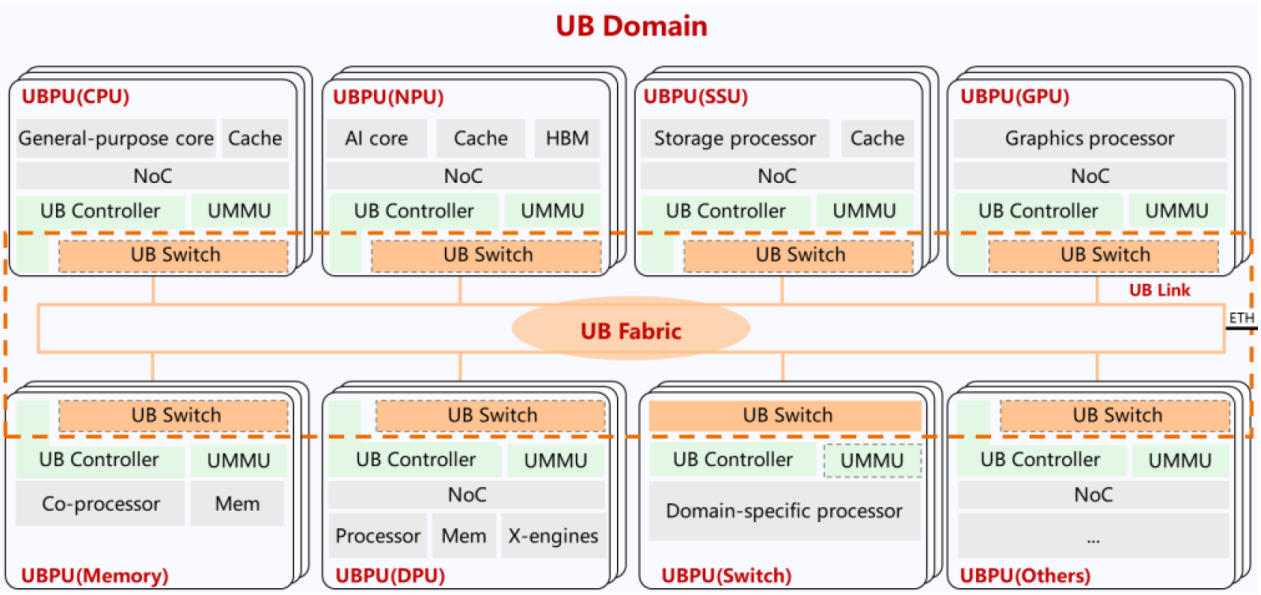
图26：UB 协议栈



资料来源：华为《灵衢基础规范》官方文档，东兴证券研究所

基于灵衢建立的计算系统部署范围可以从单台服务器到全数据中心。基于统一互联，灵衢系统中的所有处理单元地位平等、所有资源均可池化。此外，UB支持链路层的虚通道机制、网络层逐包/逐流多路径路由机制，传输层传输通道组共享机制，从而支持nd-mesh、Clos、torus等在内的任意拓扑或拓扑组合，以提升系统性能、增进容错性、支持大规模连接、降低部署成本。

图27：基于灵衢协议部署的超节点架构



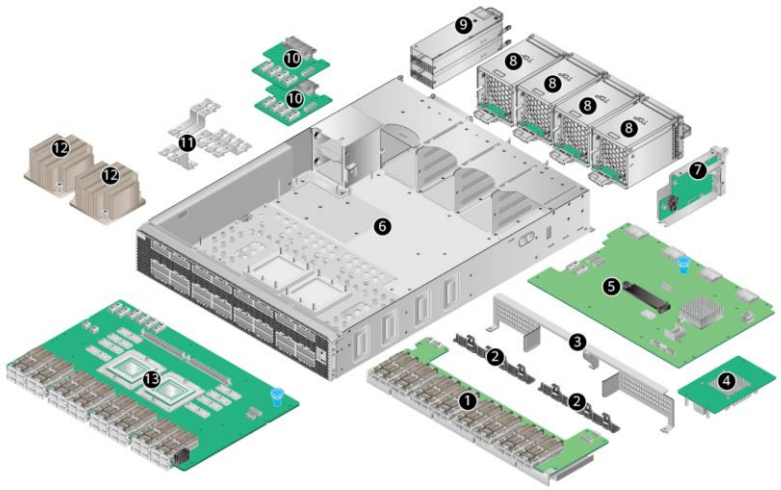
资料来源：华为《灵衢基础规范》官方文档，东兴证券研究所

根据灵衢互联协议，华为超节点自研灵衢总线设备。灵衢总线交换设备（灵衢交换机）内置的高性能交换芯片，从而为超节点的智算服务器提供高速网络连接，该设备具有高性能、高带宽、低延迟等特点。

图28：灵衢总线交换设备外观

资料来源：华为《Atlas 800T A3 超节点技术白皮书》，东兴证券研究所

图29：灵衢总线交换设备物理结构图



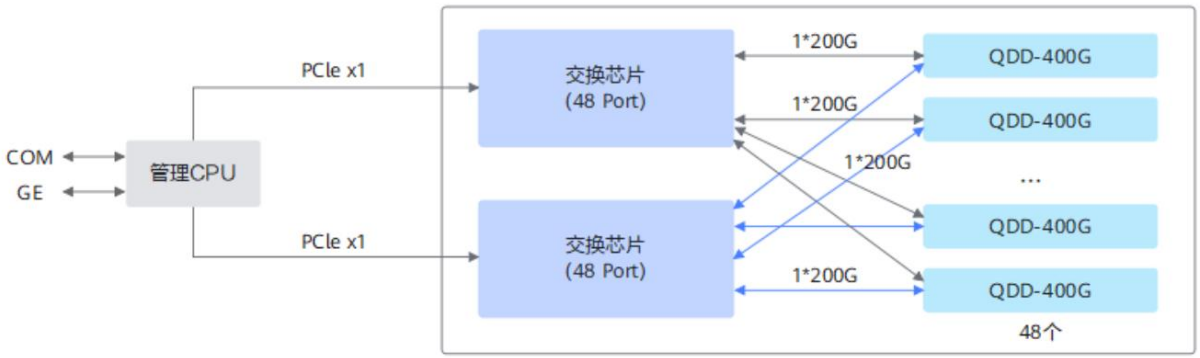
1	QSFP-DD接口板	2	QSFP-DD理线架
3	导风板	4	CPU扣卡
5	CPU底板	6	机箱
7	管理板	8	风扇模块
9	电源模块	10	电源转接板
11	铜排	12	散热器
13	业务交换板	-	-

资料来源：华为《Atlas 800T A3 超节点技术白皮书》，东兴证券研究所

一台灵衢总线交换设备交换容量为 19.2TB/s，约为单颗 NVSwitch5 芯片的 3 倍。灵衢总线交换设备包含两个交换芯片，共有 48 个端口。2 个交换芯片分别出一个 200G 到每个端口，组成一个 400G QSFP-DD 端口。（支持 192*112G SerDes）

（注：单颗 NVSwitch5 芯片交换容量为 7.2TB/s。由 72 个 NVLINK Port（上下各 36 个 Port），每个 NVLlink Port 交换容量 100GB/s 构成。）

图30：灵衢总线交换设备逻辑框图



资料来源：华为《Atlas 800T A3 超节点技术白皮书》，东兴证券研究所

UB 支持电链路，也支持光链路。在 UB 物理层和链路层定义方面，UB 物理层利用创新的串行通信技术，同时支持短距离电缆和长距离光纤。采用 PAM4 调制，支持**线性光组件**场景 53.125 Gbps 和 106.25 Gbps 速率。

图31：UB 物理层支持两种模式

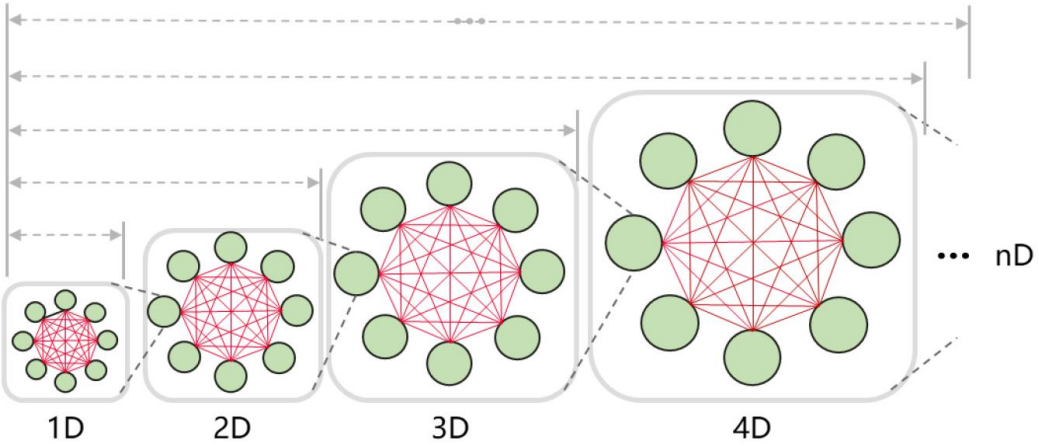
模式	PHY Mode-1	PHY Mode-2
数据速率（Gbps）	4.0，自定义速率	2.578125，25.78125，53.125，106.25
调制模式	4.0 Gbps：NRZ 自定义速率：NRZ 或 PAM4	2.578125/25.78125 Gbps：NRZ 53.125/106.25 Gbps：PAM4

资料来源：华为《灵衢基础规范》官方文档，东兴证券研究所

华为 UB 交换网络支持 UB-MESH。UBPU 是指支持 UB 协议栈的处理单元，其为超节点提供 UB-Mesh 以及基于光交换的组网技术。（UB-Mesh: a Hierarchically Localized nD-FullMesh Datacenter Network Architecture）

UB-Mesh 中的 nD-FullMesh 拓扑技术特征是充分利用业务数据局部性，优先考虑短程直接互连路径，以最大限度减少数据移动距离并减少交换机使用为目标，是一种兼具高性能和低成本的拓扑组网。从互联距离角度出发，可简单理解为单板上的若干块 NPU 之间是 1D 全连接，同一机架内叫 2D 全连接，跨机柜同屋子里叫 3D 全连接，楼层机柜组的 4D 全连接，乃至整栋建筑的 5D 全连接等。

图32：UB-Mesh 中的 nD-FullMesh 拓扑示意

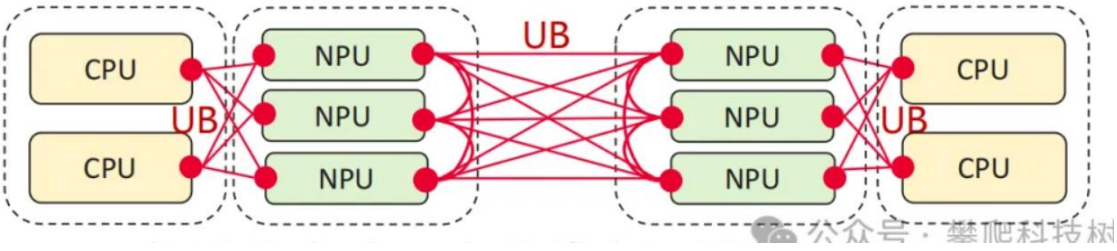


资料来源：华为《基于灵衢的超节点架构参考白皮书》，东兴证券研究所

在华为超节点 1D/2D-FullMesh 拓扑均采用电缆互连方式。

1D-FullMesh 即指 NPU 单板内的若干个 NPU 芯片之间实现 FullMesh 互联。

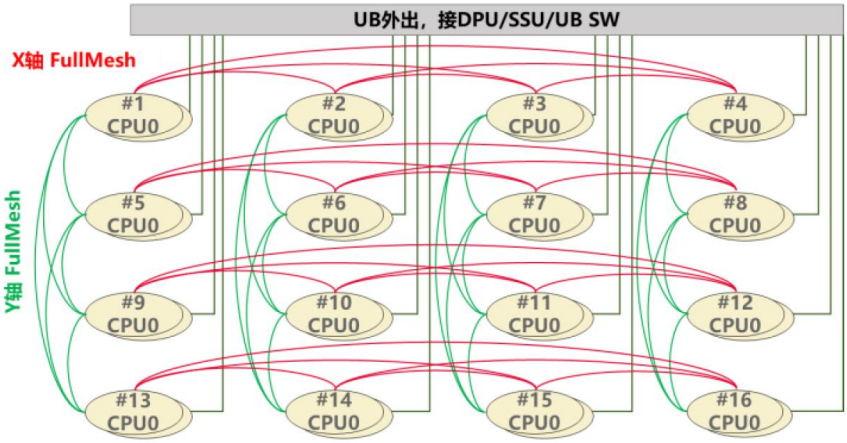
图33：1D-FullMesh 拓扑示意



资料来源：攀爬科技树公众号，东兴证券研究所

2D-FullMesh 是 nD-FullMesh 在极致低时延场景的应用，减少交换引入的时延开销，可用于内存共享和内存借用等场景。

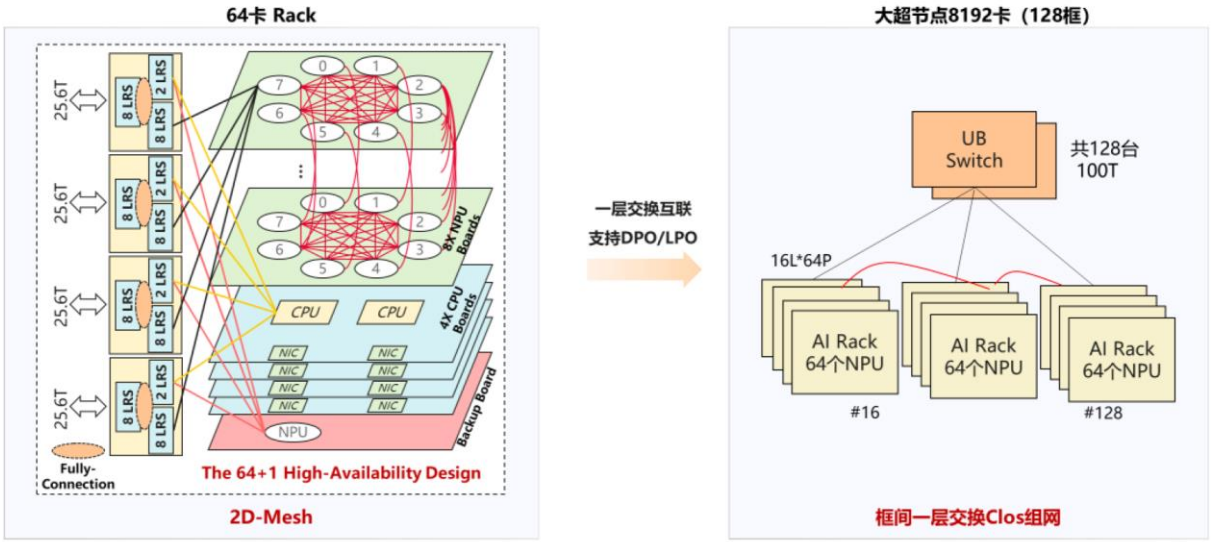
图34：2D-FullMesh 拓扑示意



资料来源：华为《基于灵衢的超节点架构参考白皮书》，东兴证券研究所

混合拓扑一层交换互联支持 DPO/LPO 光模块。UB-Mesh 支持混合拓扑，Rack 内采用 2D-FullMesh 组网，Rack 间采用一层 UB Switch 互连，支持从 64 卡线性扩展到 8192 卡。

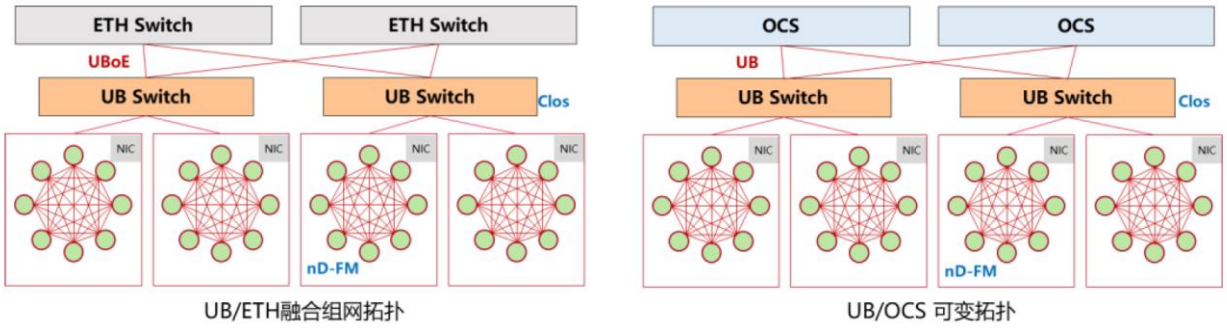
图35：2D-FullMesh + Clos 混合拓扑示意



资料来源：华为《基于灵衢的超节点架构参考白皮书》，东兴证券研究所

UB 支持融合组网、光交换组网。为了进一步扩大组网规模，UB 除了支持采用多级 UB Switch 扩展组网之外，还支持通过 UBoE 与以太 Switch 对接，实现融合组网；以及通过 OCS（光交换）组网，实现可变拓扑，匹配业务动态流量。（OCS 是一种直接在光域完成信号路由与交换的技术，无需像传统设备那样进行“光-电-光”转换。）

图36：UB 融合组网和光交换组网示意



资料来源：华为《基于灵衢的超节点架构参考白皮书》，东兴证券研究所

3.2 华为 CloudMatrix 384 超节点：两层拓扑架构，全光互联

加入超节点发展潮流，华为推出第一代超节点 CloudMatrix 384。2025 年 4 月，华为推出 CloudMatrix 384（Atlas 900 SuperPod）。对标英伟达 GB200 NVL72：从时间维度，CloudMatrix 384 推出时间落后英伟达 GB200 NVL72（2024 年 3 月发布）约 1 年时间。在后续超节点发布节奏方面，华为与英伟达接近，均为一年一代新产品。华为计划 2026 年第四季度发布 Atlas 950 SuperPod，以及 2027 年第四季度发布 Atlas 960 SuperPod。

表4：华为超节点迭代路线及性能对比

	CloudMatrix 384（Atlas 900 SuperPod）	Atlas 950 SuperPod	Atlas 960 SuperPod
推出时间	2025 年 4 月	预计 2026 年第四季度发布	预计 2027 年第四季度发布
NPU 数量	384 昇腾 910C NPUs	8192 昇腾 910DT NPUs	15488 昇腾 960 NPUs
计算机柜数	12	128	176
互联机柜数	4	32	44
系统算力	300 PFLOPS（BF16）	8 EFLOPS（FP8） 16 EFLOPS（FP4）	30 EFLOPS（FP8） 60 EFLOPS（FP4）
内存容量	49.2TB	1152TB	4460TB
互联协议	灵衢 1.0	灵衢 2.0	灵衢 2.0
总互联带宽	269TB/s	16.3PB/s	34PB/s
训练总吞吐	280k TPS	4.9mn TPS	15.9mn TPS
推理总吞吐	740k TPS	19.6mn TPS	80.5mn TPS

资料来源：华为全连接大会，SemiAnalysis，科技攀爬树公众号，东兴证券研究所

CloudMatrix 384以芯片互联规模实现Scale up网络性能。GB200 NVL72采用整机柜型超节点方案，Scale up计算单元72个GPU芯片；华为CloudMatrix 384则采用分机柜超节点方案，其计算节点和交换节点分别安装在不同机柜（包含12个计算柜和4个交换柜），Scale up计算单元由384个Ascend 910C 芯片组成。昇腾芯片数量增加五倍，弥补每个图形处理器（GPU）性能仅为英伟达布莱克韦尔（Blackwell）芯片三分之一的不足。从算力性能维度，华为CloudMatrix 384的BF16密集算力约300 PFLOPS，与GB200 NVL72接近；

此外，华为CloudMatrix 384Scale up单向带宽134400 GB/s，约是GB200 NVL72的2.1倍。

表5：GB200 NVL72 超节点与 CloudMatrix 384 算力与通信性能对比

	单位	GB200 NVL72	CloudMatrix 384	倍数
算力	PFLOPS	180（TF32 Tensor 核心）	300（BF16 dense）	接近
HBM 内存	TB	13.8	49.2	3.6X
HBM 带宽	TB/s	576	1229	2.1X
Scale up 带宽	单向 GB/s	64800	134400	2.1X
Scale up 计算单元	GPUs	72	384	5.3X
功耗	kW	145	600	4.1X

资料来源：SemiAnalysis，东兴证券研究所

除了算力与通信性能，尺寸、重量、功耗均是超节点 TCO（总体拥有成本）的关键影响因素。CloudMatrix 384 超节点 16 个机柜，占地面积约 20 平米；总功耗约 600kW。（GB200 NVL72 机柜尺寸为长 1068 毫米、宽 600 毫米、高 2495 毫米；重约 1.36 吨；功耗 145KW。）

图37：CloudMatrix 384 超节点外观

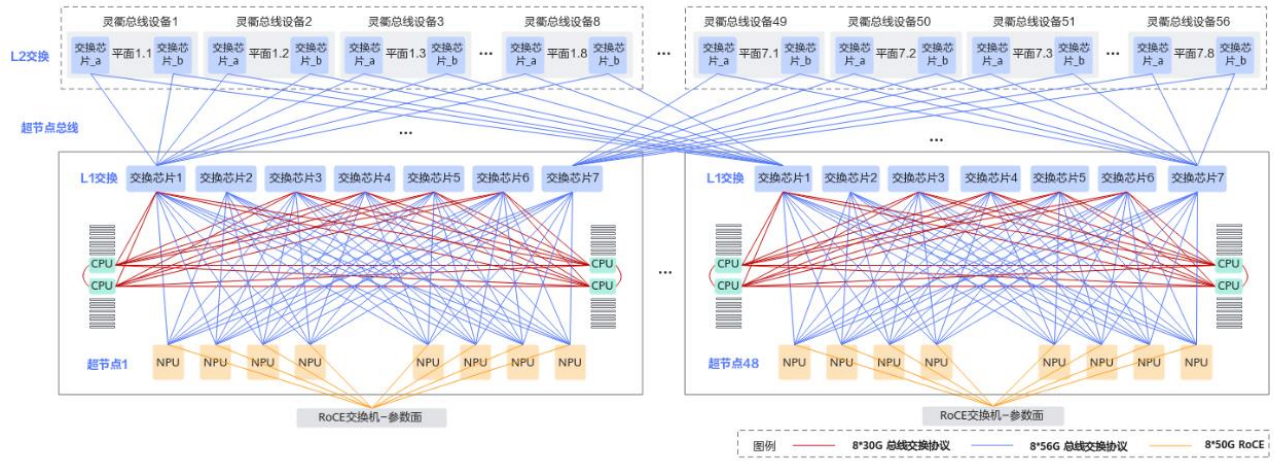
资料来源：华为，东兴证券研究所

CloudMatrix 384 Scale up 网络采取两层扁平拓扑架构。CloudMatrix 384 互联网络通过华为自研的灵衢网络和灵衢总线设备实现互联组网。灵衢网络 L1 层由超节点的交换网板承载，L2 层通过总线设备柜中的灵衢总线设备组成，L1-L2 分别通过光纤组成不同规模的超节点集群。

L1 层：每个计算节点集成了 8 个昇腾 910C NPU、4 个鲲鹏 CPU，每个计算节点内部放置了 7 颗板载 UB 交换芯片。

L2 层（机柜间）：划分为 7 个独立子平面，每个子平面包含 16 个 L2 UB 交换芯片。L1 交换芯片扇出 16 条链路到对应 L2 子平面的每个交换芯片，实现无阻塞全对等拓扑。

图38：CloudMatrix 384 超节点组网方案

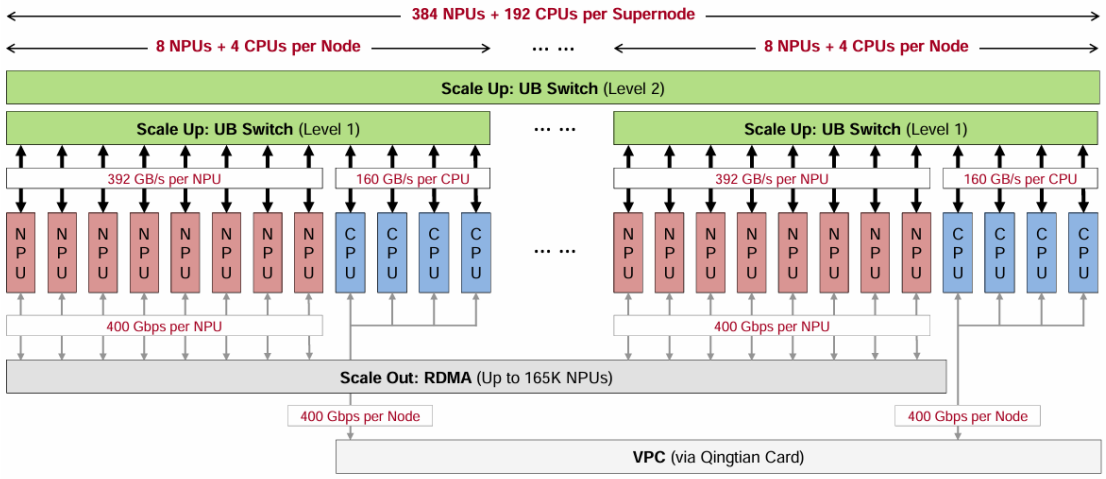


资料来源：华为《Atlas 800T A3 超节点技术白皮书》，东兴证券研究所

CloudMatrix384 两层扁平拓扑架构形成三个网络平面。

- UB 平面构成超级节点内主要的超高带宽 Scale-UP 扩展结构。它以无阻塞的全对全拓扑直接互连所有 384 个 NPU 和 192 个 CPU。
- RDMA 平面支持 CloudMatrix384 超级节点和外部 RDMA 兼容系统之间的向外 Scale-OUT 通信。
- 虚拟私有云（VPC）平面是通过高速网卡（华为的青天河卡）将 CloudMatrix384 超级节点连接到更广泛的数据中心网络。

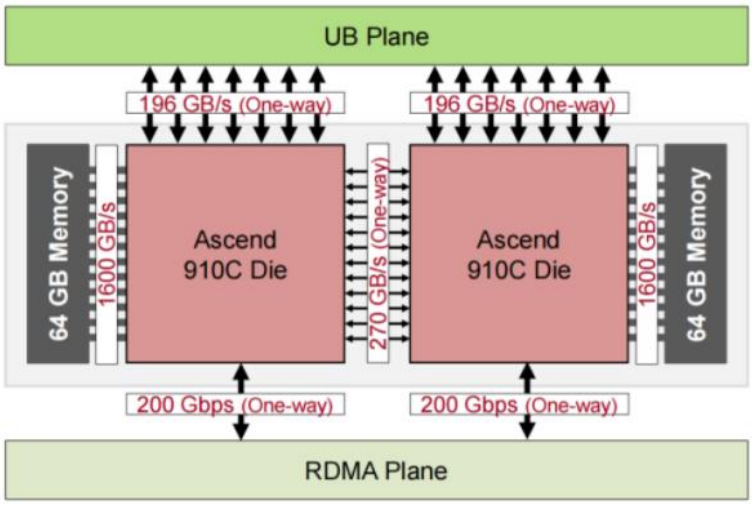
图39：CloudMatrix 384 三层网络平面



资料来源：架构师技术联盟公众号，东兴证券研究所

Ascend 910C NPU 芯片采用双 Die 封装，通过高带宽总线实现芯片互连，单向传输速率为 270 GB/s。对比 GB200 NVL72，其芯片内部通过 NVLink-C2C 的双向传输速率为 900GB/s。

图40：CloudMatrix 384 中 Ascend 910C NPU 芯片架构



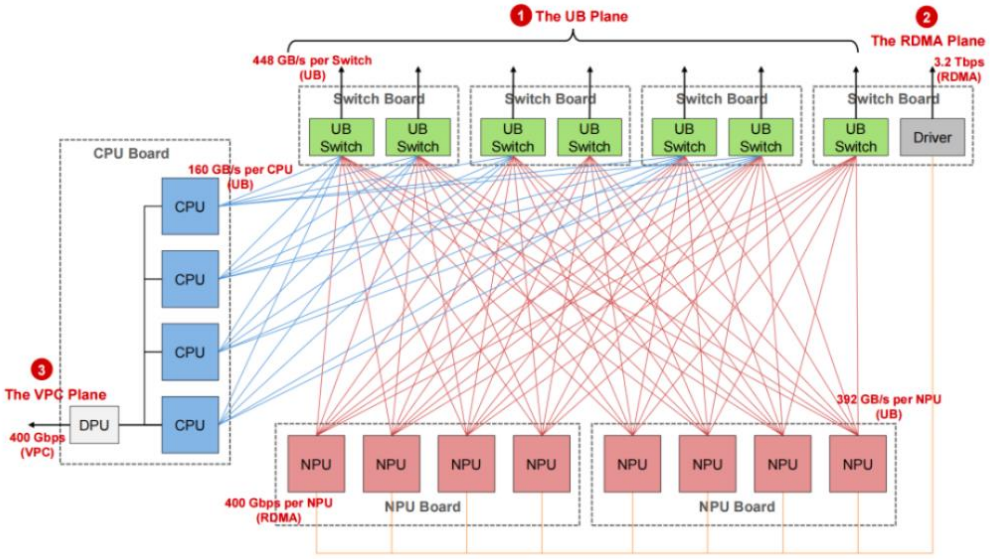
资料来源：曦微有光公众号，东兴证券研究所

CloudMatrix 384 计算节点 48 个；单个计算节点传输带宽 5.6TB/s。

CloudMatrix 384 设置 12 个计算柜，每个计算柜 4 个计算节点。华为 CloudMatrix384 每个计算节点集成了 8 个昇腾 910C NPU、4 个鲲鹏 CPU，每个计算节点内部放置了 7 颗板载 UB 交换芯片。12 个处理器（8 个 NPU + 4 个 CPU）通过 UB 链路连接到这些板载交换机，在节点内创建单级 UB 平面。

每个昇腾 910C 贡献 392GB/s 单向带宽（底层通过 14*400Gbps 以太接口互联），按 400Gbps 接口换算，每个 NPU 分别与 7 颗交换芯片双 400G 接口互联，通过提供 14*400GB/s =5.6TB/s 传输带宽。

图41：CloudMatrix 384 单个计算节点网络拓扑



资料来源：曦微有光公众号，东兴证券研究所

CLoudMatrix 384 通过 3168 根光纤和 6912 个 400G LPO 模块构建高速互连总线。

（1）UB 平面：384（NPU 总数）×7（每个 NPU 需要 7 个 400G 光模块）×2（双向互联）=5376 个，L1 层和 L2 层之间不采用 400G 光模块互联；

（2）RDMA 平面：384（与服务器互联，每个 GPU 配 1 个 400G 网卡）+ 768（RDMA 网络平面采用 2 层胖树架构，叶层交换机端口翻倍）+ 384（脊层交换机需同等数量）= 1536 个；

（3）VPC 平面：将 CloudMatrix384 超节点连接到更广泛的数据中心网络，是超节点外的运用，使用的光模块数量不计入统计。

表6：华为 CloudMatrix 384 超节点网络架构与互联方案

	CloudMatrix 384
NPU 数量	384 昇腾 910C NPUs
CPU 数量	192 鲲鹏 CPUs
系统算力	BF16 密集算力 300 PFLOPS，与 GB200 NVL72 接近
总内存容量	49.2TB，是英伟达 GB200 NVL72 的 3.6 倍
总内存带宽	1229TB/s，是英伟达 GB200 NVL72 的 2.1 倍
网络架构	三平面网络设计：（1）UB 平面（392GB/s 卡间带宽，全对等互联）；（2）RDMA 平面（400Gbps/NPU，支持 165K NPU 扩展）；（3）VPC 平面（管理控制与存储）
互联方案	去铜全光，每个 NPU 通过 7 个 400G LPO Sipho 光模块实现互联，单节点总用量达 6912 个光模块，配合 3168 根光纤构建全光互联网络
UB 平面	完成超节点 Scale-Up，采用两层扁平拓扑架构：（1）L1 层（节点内），每个超节点集成 8 个昇腾 910C NPU+4 个鲲鹏 CPU+7 颗板载 L1 UB 交换芯片+1 擎天卡；（2）L2 层（机柜间），划分为 7 个独立子平面，每个子平面包含 16 个 L2 UB 交换芯片；L1 交换芯片扇出 16 条链路到对应 L2 子平面的每个交换芯片，实现无阻塞全对等拓扑
L1 UB 交换芯片	上行链路带宽为 448GB/s，可支持 48 个 400G 接口
RDMA 平面	支持跨 CloudMatrix 384 超节点和外部 RDMA 兼容系统的 Scale-out 通信，采用 RoCE 协议以兼容标准的 RDMA 生态，该平面主要连接 NPU，每个 NPU 提供 400 GB/s 的单向 RDMA 带宽
VPC 平面	通过高速 NIC（华为擎天卡）将 CloudMatrix 384 超节点接入更广泛的数据中心网络，每个超节点提供 400 GB/s 的单向带宽
光模块用量	（1）UB 平面：384（NPU 总数）×7（每个 NPU 需要 7 个 400G 光模块）×2（双向互联）= 5376 个，L1 层和 L2 层之间不采用 400G 光模块互联 （2）RDMA 平面：384（与服务器互联，每个 GPU 配 1 个 400G 网卡）+768（RDMA 网络平面采用 2 层胖树架构，叶层交换机端口翻倍）+384（脊层交换机需同等数量）=1536 个 （3）VPC 平面：将 CloudMatrix384 超节点连接到更广泛的数据中心网络，是超节点外的运用，使用的光模块数量不计入统计

资料来源：SemiAnalysis，架构师技术联盟公众号，曦微有光公众号，东兴证券研究所

3.3 总结：灵衢协议尚未被国内业界广泛接受，集群化方式实现性能追赶

国内 Scale Up 协议尚未统一，华为灵衢协议尚未被国内业界广泛接受。在 Scale Up 协议方面，华为推出灵衢协议，并从 2.0 版本起转向开放标准。除此之外，国内其他厂商正探索多种互联协议，包括中移 OISA、腾讯 ETH-X、高通量以太网 ETH+以及中兴通讯 OLink 等。为打破生态壁垒，国内正积极推动标准统一，比如工信部正牵头推动 CLink 协议，旨在形成统一的国内标准。

华为超节点依靠集群化方式实现性能追赶。Atlas 950 超节点预计 2026 年第四季度发布，相比英伟达同样将在 2026 年下半年上市的 NVL144 总算力 2.52 EFLOPS（FP8），其算力达到 8 EFLOPS（FP8）。此外，Atlas 950 超节点在内存容量 1152TB 与互联带宽 16.3PB/s，也实现大幅领先。我们认为，短期内，华为超节点依靠集群化实现性能追赶，但在超节点复杂性、可靠性、功耗等维度需要平衡。从整体解决方案看，英伟达在超节点的芯片工艺、软件生态与系统集成上的优势仍难以撼动。

Atlas 950 超节点互联方案或将调整，显示华为超节点技术在标准化阶段仍需夯实。相比上一代超节点，华为 Atlas 950 超节点不再使用全光互联架构，其通过“柜内正交铜互联+柜间光互联”的混合设计，在机柜内部利用铜互联实现高可靠、低成本和低功耗的连接，跨机柜则通过光互联保障系统的可扩展性，从而在维持系统可扩展性的同时，有效控制总体拥有成本（TCO）。

表7：华为超节点 Scale up 迭代路线

	Atlas 800T A3	Atlas 900 SuperPod	TaiShan 950 Superpod	Atlas 850	Atlas 950 SuperPod
首发时间	2025 年 4 月		2025 年 9 月华为全连接大会		
上市时间	2025 年 Q2		预计 2026 年 Q1	预计 2026 年 Q4	
产品定位	企业级 AI 服务器	旗舰级 AI 集群	首个通用计算超节点	企业级 AI 服务器	旗舰级 AI 集群
卡数	单机 8 NPUs	384 NPUs	单机 32 CPUs	单机 8 NPUs	8192 NPUs
形态	单机柜，支持多柜灵活部署	12 计算柜+ 4 总线设备柜	单机柜，支持多柜灵活部署	单机柜，支持多柜灵活部署	128 计算柜+32 总线设备柜
系统算力	6 PFLOFPS（FP16） 12 POPS（INT8）	300 PFLOPS（FP16）	未公布	8 PFLOPS（FP8） 16 PFLOPS（FP4）	8 EFLOPS（FP8） 16 EFLOPS（FP4）
内存容量	1024GB	48TB	48TB	1152GB	1152TB
D2D 互联带宽	784GB/s	784GB/s	/	2TB/s	2TB/s
总互联带宽	/	269TB/s	/	/	16.3PB/s

资料来源：华为官网，华为全连接大会，东兴证券研究所

4. 谷歌：建立光互联超节点，与英伟达形成不对称竞争

4.1 Scale up 网络核心技术：创新应用光电路交换机（OCS）

相比英伟达 NVLink，谷歌超节点 Scale Up 协议具有显著差异。谷歌超节点 Scale Up 协议主要基于其自研的 ICI（Chip-to-Chip Interconnect）协议，结合光电路交换（OCS）技术，用于实现 TPU 集群内的高速互联。对英伟达而言，NVLink 与 NVLink 交换机是英伟达构建单机柜 Scale up 网络的核心技术组合，顶级交换机芯片提供无阻塞高性能带宽；而谷歌 ICI 协议精简不必要的功能模块，专注于芯片间高效数据传输，降低协议开销和延迟，协议复杂度低，追求实用主义，与英伟达形成不对称竞争。

表8：谷歌 ICI Link 协议 VS 英伟达 NVLink 协议

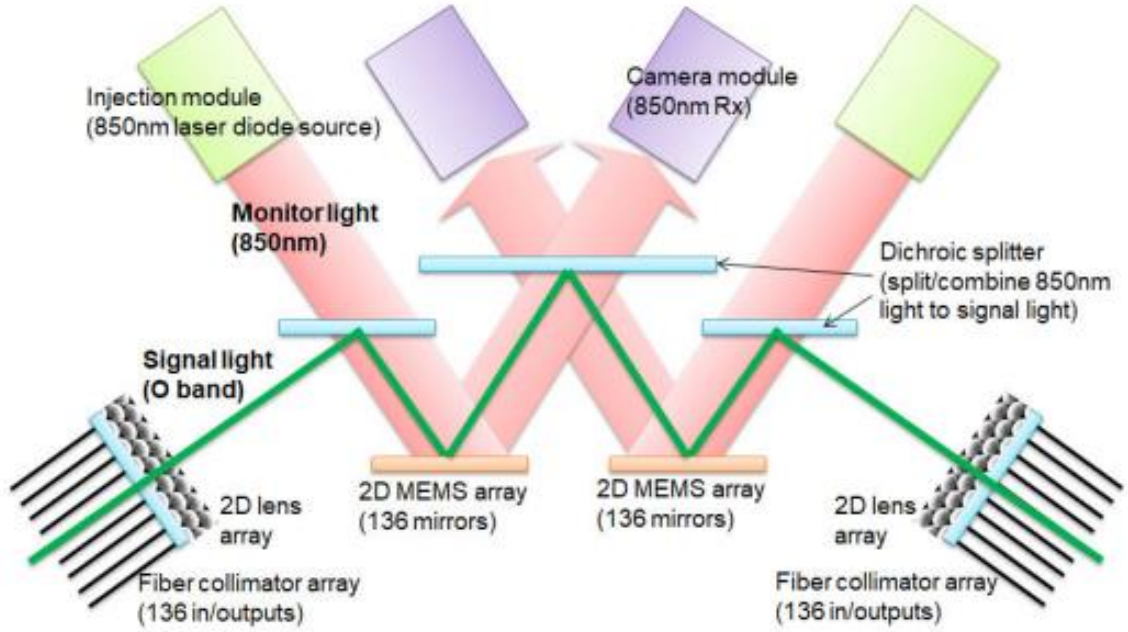
维度	Google TPU（v4/v7）	NVIDIA（H100/GB200）
互联协议	G-ICI(私有轻量级，Credit-based)	NVLink+InfiniBand/RoCE
网络层级	物理隔离：ICI 和 DCN 存储分离	分层架构：Scale-up 与 Scale-out 分层
故障恢复	物理重构：OCS 旋转镜面隔离坏点	协议重传：依赖 IB/RoCE 重传机制
软件耦合	强耦合：XLA 编译器需感知物理拓扑	解耦：CUDA 生态屏蔽底层拓扑差异
核心哲学	静态极致：通过 OCS 光交换网络构建确定拓扑	带宽堆叠：顶级芯片提供无阻塞带宽

信息来源：AGI 小咖公众号，东兴证券研究所

对比英伟达 Nvlink 电交换机，谷歌自研光交换机。“阿波罗计划”（Project Apollo）致力于革新传统的“Clos”网络架构，用光路交换机（OCS）取代原有的电子分组交换机（EPS）。谷歌第一代光交换机代号“帕洛玛”（Palomar），与传统架构需要在核心层多次进行电-光-电信号转换不同，OCS 采用全光互联技术，不读取数据包头、不进行光电转化。而是通过镜面反射引导携带数据的光束，实现源端口到目标端口的直接传输。在 Palomar OCS 机箱内部光信号的传输路径呈现出一个经典的“W”形状，最大限度减少插入损耗和实现任意端口间的互联。

W 形光路设计：光纤准直器> 二向色分光镜 > 2D MEMS 阵列 I > 二向色分光镜 > 2D MEMS 阵列 II> 二向色分光镜 >光纤准直器。

图42：谷歌 Palomar OCS 光信号传输路径



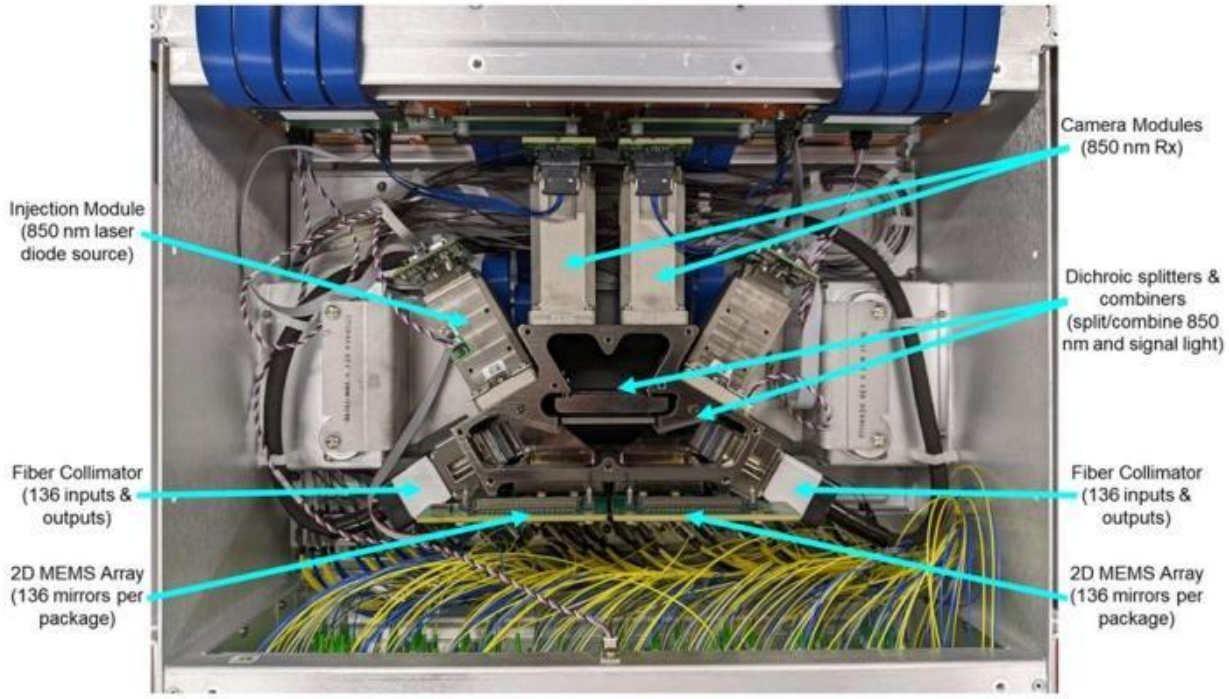
资料来源：Google 论文《Mission Apollo_Landing Optical Circuit Switching》（作者：Ryohei Urata, Hong Liu 等），东兴证券研究所

谷歌 Palomar OCS 设备主要构成：a) 光纤准直器（fiber collimators）；b) 相机模块（camera modules）；c) MEMS 微镜模块（packaged MEMS）；d) 监视模块（injection modules）；e) 二向色分光镜&合光镜（dichroic splitter and combiners）。

工作原理：输入光信号通过二维光纤准直器阵列进入光学核心模块，每个准直器阵列由 N×N 光纤阵列和二维透镜阵列组成；光学核心模块包含两组 2D MEMS 微镜模块，光信号将依次通过两个准直器阵列的端口和两个 MEMS 微镜模块。通过驱动反射镜倾斜，将光信号导向对应的输出准直器光纤。

在信号光路的基础上，增设监测信道辅助反射镜的调谐：两级 2D MEMS 设计实现了三维空间内的精准光束操纵，二向色分光镜作为允许 1310nm 业务光透射，同时反射 850nm 监控光的核心滤光组件，与 Injection Module + Camera Module 联动实现带内运维监控和驱动 2D MEMS 的微秒级微调。

图43：谷歌 Palomar OCS 实物图

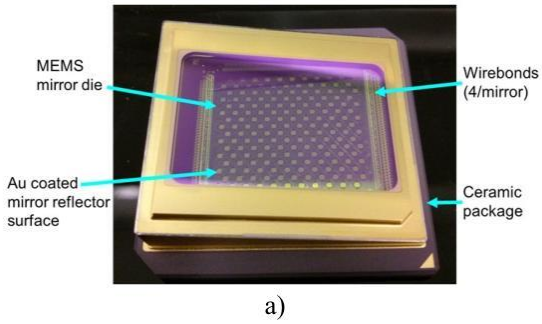


资料来源：Google 论文《Mission Apollo_Landing Optical Circuit Switching》（作者：Ryohei Urata, Hong Liu 等），东兴证券研究所

谷歌 Palomar OCS 设备端口数为 136×136，支持输入输出端口的任意双向互连。Palomar OCS 为实现任意输入到任意输出的光束转向，系统采用两个微机电系统反射镜封装组件，提供四个自由度，确保光信号从输入端口到输出端口的最佳耦合。MEMS 陶瓷封装内部包含一块大型微机电系统芯片，集成 176 个独立可控的反射镜，经筛选后，在全校准系统中保留 136 个可用反射镜。

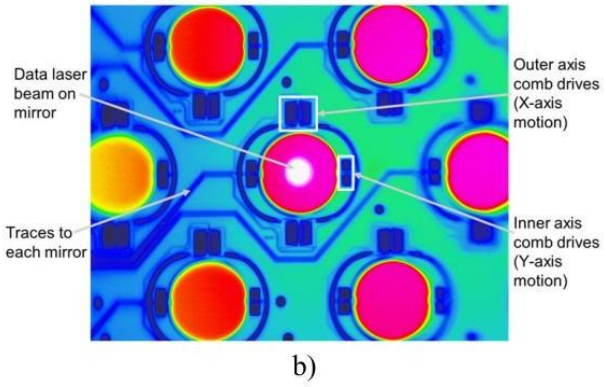
MEMS 是微机电系统（Micro-Electro-Mechanical Systems）的缩写。这类器件将微型机械结构与电子元件集成，通过微细加工技术制造，实现了机械与电学功能的微型化融合。

图44：MEMS 微镜模块实物图



资料来源：Google 论文《Mission Apollo_Landing Optical Circuit Switching》（作者：Ryohei Urata, Hong Liu 等），东兴证券研究所

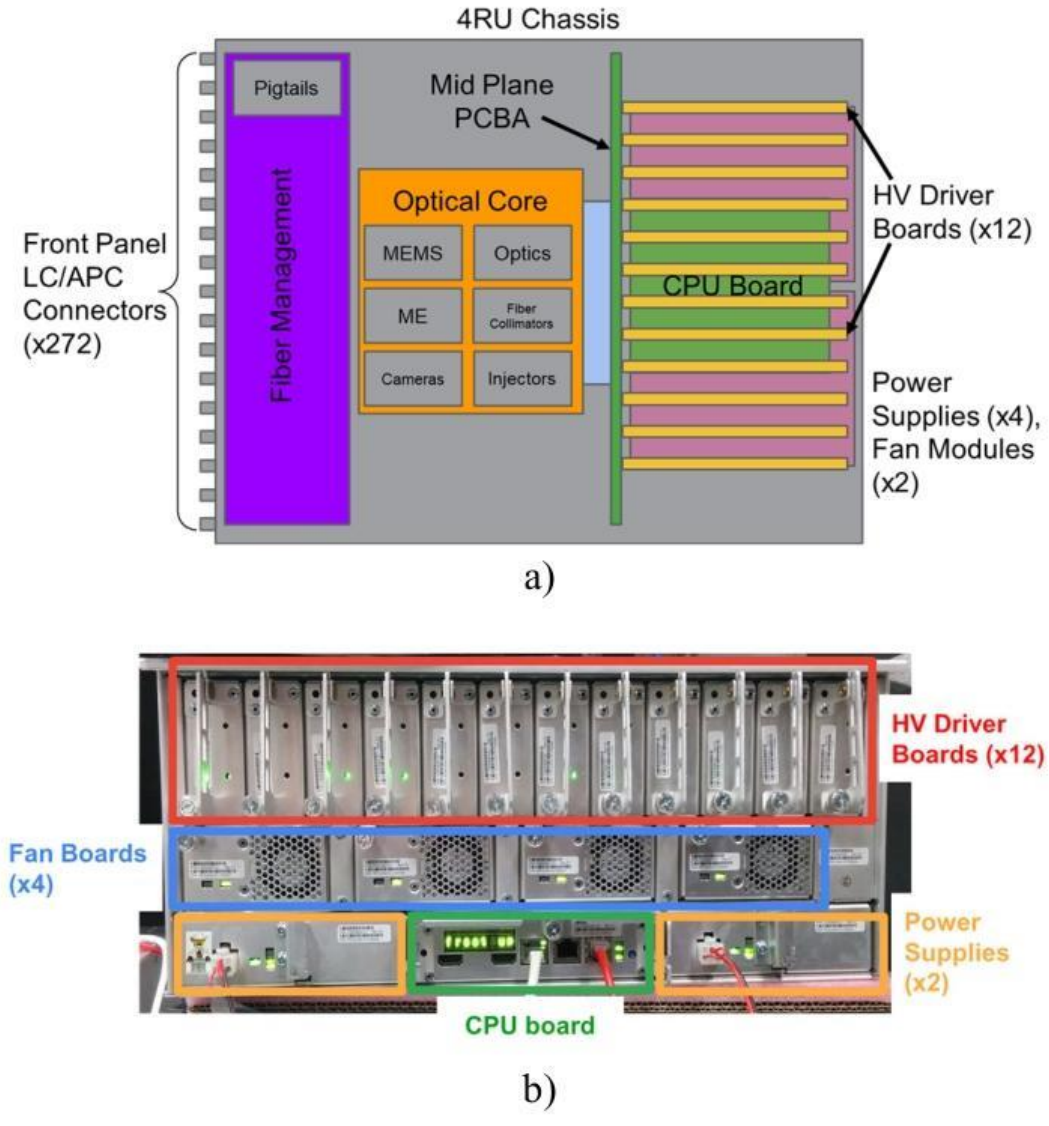
图45：MEMS 微镜模块热成像图



资料来源：Google 论文《Mission Apollo_Landing Optical Circuit Switching》（作者：Ryohei Urata, Hong Liu 等），东兴证券研究所

谷歌 Palomar OCS 设备最大功耗为 108W，仅为同交换容量电分组交换机功耗的一小部分。谷歌在过去十年间制造并部署了数万台 136×136 全双工端口的光电路交换机（含 8 个备用端口）。为量产 Palomar 光电路交换机，谷歌研发了定制化的测试、对准和组装设备，覆盖 MEMS 微镜模块、光纤准直器、光学核心模块等组件，乃至整台交换机。每个 MEMS 微镜模块反射镜在晶圆代工厂完成全量测试，采用探针台方案，单次接触即可对芯片上的 176 个反射镜进行检测；研发了定制化的自动对准设备，将二维透镜阵列以亚微米级精度贴合在光纤准直器表面；最终，每台交换机均需完成微机电系统反射镜的全量校准，以确定端口选型。

图46：谷歌 Palomar OCS 机箱机构图以及实物机箱后视图



资料来源：Google 论文《Mission Apollo_Landing Optical Circuit Switching》（作者：Ryohei Urata, Hong Liu 等），东兴证券研究所

基于 MEMS 路径的谷歌 Palomar OCS 设备具备长期迭代发展的潜力。从实际商业落地角度，压电驱动（Piezo）、Robotic 和 MEMS 三类技术路径下的光电路交换系统已经实现有限的商用落地。从发展潜力看，基于成本可控的前提下支撑数据中心应用所需的大端口数，MEMS 具备长远发展潜力，已实现超 1000×1000 端口的互连规模。

表9：各类光电路交换技术的成本、规模、性能及可靠性/可用性对比

技术类型	相对成本	端口数	交换时延	插入损耗	驱动电压（伏特）	锁存功能
MEMS	中等	320×320	毫秒级	＜3dB	数百伏	无
Robotic	中等	1008×1008	分钟级	＜1dB	—	有
Piezo	高	384×384	毫秒级	＜2.5dB	十余伏	无
Guided Wave	低	16×16	毫秒级	＜6dB	1 伏	无
Wavelength Switching	待定	100×100	纳秒级	＜6dB	0	有

信息来源：Google 论文《Mission Apollo_Landing Optical Circuit Switching》（作者：Ryohei Urata, Hong Liu 等），东兴证券研究所

4.2 谷歌 TPU v7 超节点：Cube+3D Torus+OCS 光交换实现扩展

基于光互连技术，谷歌建立独特且成熟的超节点技术方案。自 2017 年 TPU v2 至 2025 年 TPU v7，谷歌逐步将光互连技术融入 TPU 系统，超节点互联芯片数量从 256 颗增长至 9216 颗。正如英伟达超节点发展历程，谷歌在超节点领域也同样经历技术路线探索（TPU v4）与方案标准化（TPU v5p），目前已经处于性能指标优化提升阶段。2025 年谷歌发布 TPU v7（代号 Ironwood），超级集群芯片数最多可支持 9216 颗，保持 3D 环面拓扑，OCS 技术继续沿用。

表10：谷歌超节点迭代路线及性能对比

	TPU v2	TPU v3	TPU v4	TPU v5p	TPU v7
首发时间	2017	2018	2022	2023	2025
单芯片峰值算力（FP16）	45.9T	123.2T	275T	459T	2307T
单芯片 HBM 内存	16GB	32GB	32GB	95GB	192GB
机柜数	4	16	64	140	144
互连拓扑	2D 环面	2D 环面	3D 环面	3D 环面	3D 环面
分布	16×16	32×32	4×4×4	4×4×4	4×4×4
OCS 数量	-	-	48	48	48
芯片数	**256**	**1024**	**4096**	**8960**	**9216**

信息来源：AI 闲谈公众号，东兴证券研究所

谷歌 TPU v7 单芯片算力不及 GB200，但在大规模训练扩展性上占优。在单个算力芯片性能指标上，TPU v7 提供 2307 TFLOPs 的 BF16 算力，内存是 192G HBM3e，内存带宽是 7.3TB/s。与 GB200 相比，TPU v7 提供的 FLOPs 和内存带宽差距较小，产品上市时间落后约一年时间。在 Scale-Up 网络性能上，TPU v7 可通过 4×4×4 Cube+3D Torus+OCS 光交换的层级架构实现从单芯片到全 Pod 的无缝扩展，Scale-Up 最多可支持 9216 块芯片集群，适用于千亿/万亿参数 LLM 训练、大规模 MoE 模型。

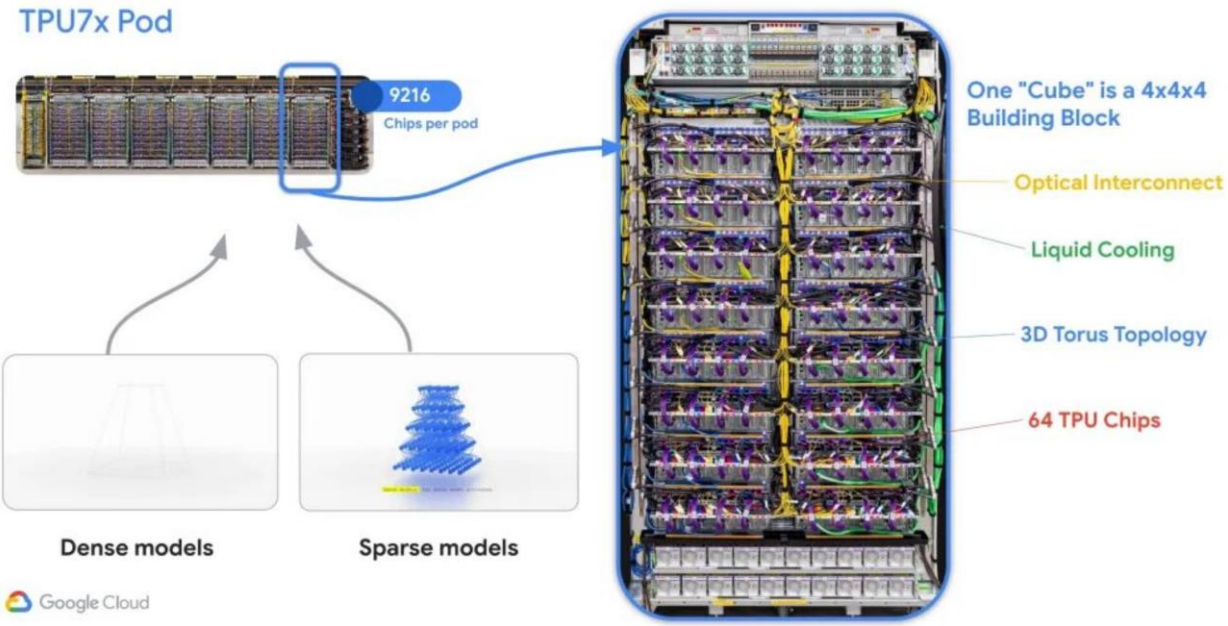
表11：英伟达 GB200 芯片与与谷歌 TPUv7 性能对比

	Unit	GB200	TPU v7 (internal)	TPU v7 (external)
FP8 dense TFLOPS	TFLOPS	5000	4614	4614
BF16 dense TFLOPS	TFLOPS	2500	2307	2307
HBM capacity	GB	192	192	192
HBM bandwidth	TB/s	8.0TB	7.3	7.3
单向互连带宽	Gb/s	7200	4800	4800
TCO per Marketed FP8 Dense PFLOP	$/hr per PFLOP	0.46	0.28	0.4
TCO per Memory Bandwidth	$/hr per TB/s	0.28	0.18	0.25
TCO per Memory Capacity	$/hr per TB	11.87	6.67	9.65

资料来源：SemiAnalysis，Nvidia，Google，东兴证券研究所

谷歌单个机架采用“4×4×4 立方体构建块”，包含 64 颗 TPU。在过去的几代产品中，谷歌 TPU 机架设计基本保持一致。每个机架包含：16 个 TPU Tray、16 个或 8 个 Host CPU Tray（取决于散热配置）、一台 ToR 交换机、电源模块，以及 BBU。每个 TPU tray 包含一块带有 4 个 TPU 封装的 TPU 板。每个 Ironwood TPU 配备 4 个用于 ICI 连接的 OSFP cage，以及 1 个用于连接 Host CPU 的 CDFP PCIe cage。

图47：谷歌超节点单个机架实物图



资料来源：Google，东兴证券研究所

谷歌 v4、v5p 及 v7 超节点架构组网均采用“4×4×4 立方体构建块”作为核心架构单元。

单元组成：由 4(X)×4(Y)×4(Z)共 64 块 TPU 芯片构成。

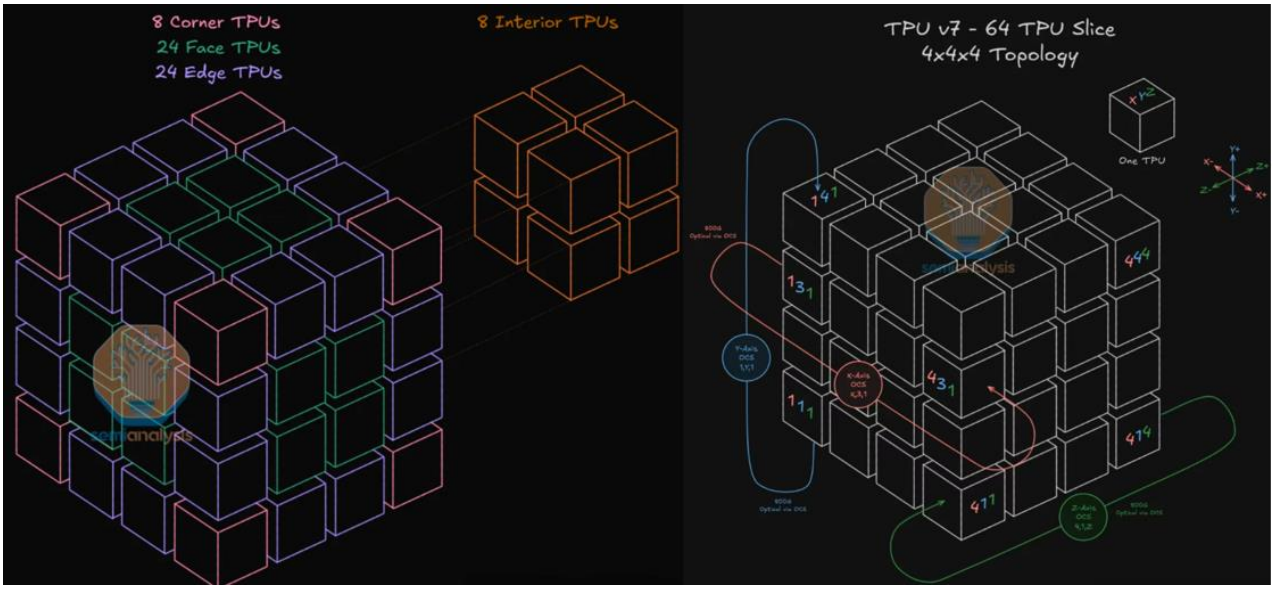
链路总数：每块 TPU 芯片引出 6 条 ICI (Inter-Chip Interconnect)高速互联链路，分别对应三维坐标系的±X,±Y,±Z 六个方向，构建起 3D Torus 的基础网格。

具体来说，每个 TPU 的互联方式如下：

- 位于立方体内部的 TPU（8 个）：2 根 PCB+4 根 DAC copper
- 位于立方体面的 TPU（24 个）：2 根 PCB+3 根 DAC copper+1 个光模块
- 位于立方体边缘的 TPU（24 个）：2 根 PCB+2 根 DAC copper+2 个光模块
- 位于立方体角落的 TPU（8 个）：2 根 PCB+1 根 DAC+3 个光模块

因此，单个 4×4×4 的立方体需要 96 个光模块，平均下来单个 TPU 需要 1 根 PCB、1.25 根 DAC copper 和 1.5 个光模块。光模块能够连接 OCS 交换机，使得立方体（机架）与立方体（机架）之间得以互连。

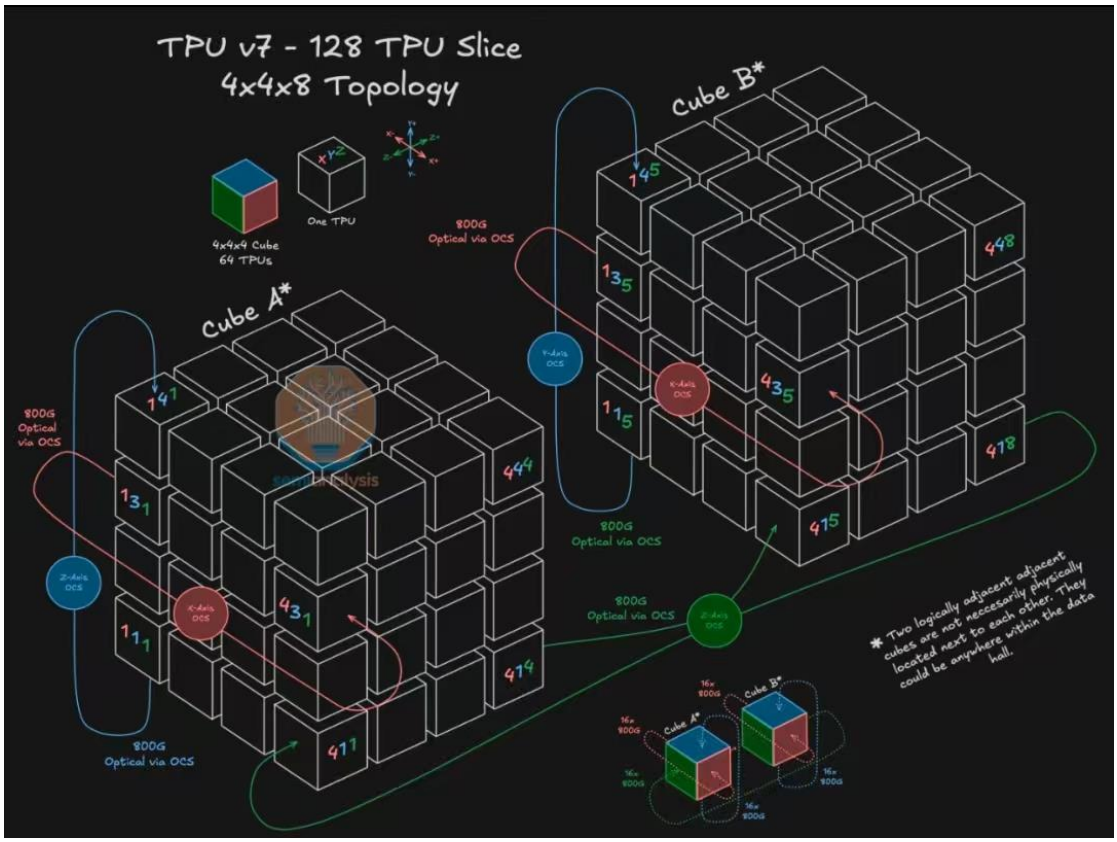
图48：TPU 4×4×4 立方体互联逻辑示意图



资料来源：SemiAnalysis，东兴证券研究所

TPU v7 引入 Twisted 3D Torus，不局限于 4×4×4 立方体的拓扑单元。TPUv7 升级了标准的 3D Torus，引入了步长的概念来构建了 Twisted 3D Torus（扭曲环面）拓扑以降低通信跳数。以 TPUv7 架构中 128 TPU Slice （4x4x8 拓扑）为例，位于 Cube A 边界的节点 TPU 414 并没有像标准 3D Torus 4×4×4 拓扑那样回环至自身的起点 TPU 411 而是通过 Twisted 3D Torus 和 OCS 构建类似于“虫洞”的跳跃式链接至逻辑相邻的 Cube B 起始节点 TPU 415 上。这意味着网络拓扑可以被重构，从而支持大量不同拓扑——理论上可达数千种。目前，谷歌支持将 TPU v7 配置成从 4 颗 TPU 到 2048 颗 TPU 之间的 10 种不同 slice 大小。

图49：TPUv7 128（4×4×8）TPU 拓扑示意图



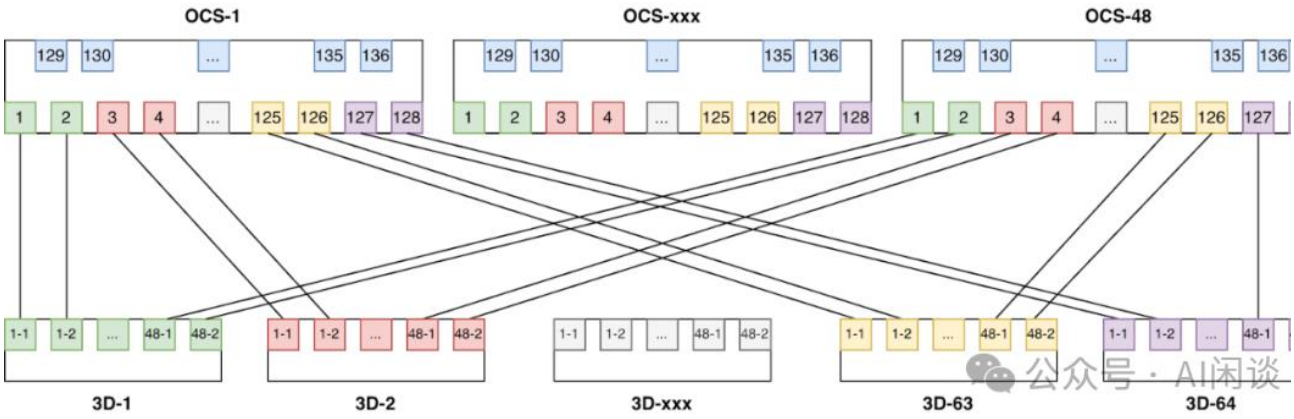
资料来源：SemiAnalysis，东兴证券研究所

谷歌 TPU v4 超节点设置 64 个 3D Torus，单个 3D Torus 对外互联带宽 4.8TB/s。可以看到，谷歌 TPU v4 超节点单个 3D Torus 对外互联带宽，低于英伟达 GB200 NVL72 计算节点访存带宽 7.2TB/s，以及低于华为 CloudMatrix 384 单个计算节点传输带宽 5.6TB/s。

谷歌 TPU v4 超节点最大支持 4096 个 TPUv4 芯片互联，由 64 个 3D Torus 以及 48 个 OCS 交换机构成。具体来讲：每个 3D Torus 连接 6*16/2 = 48 个 OCS；一个 OCS 对应 136 个端口，其中 128 个端口用于连接 3D Torus TPUv4，8 个用于测试或备份。考虑到 3D Torus 中需要两个相对 Link 连接到一个 OCS，因此每个 OCS 连接 128/2=64 个 3D Torus。

3D Torus 之间互联通过光模块和 OCS 光交换机实现。单个 3D Torus 对外引出 96 条光链路，TPUv4 每 Link 50GB/s，则单个 3D Torus 对外互联带宽 4.8TB/s。

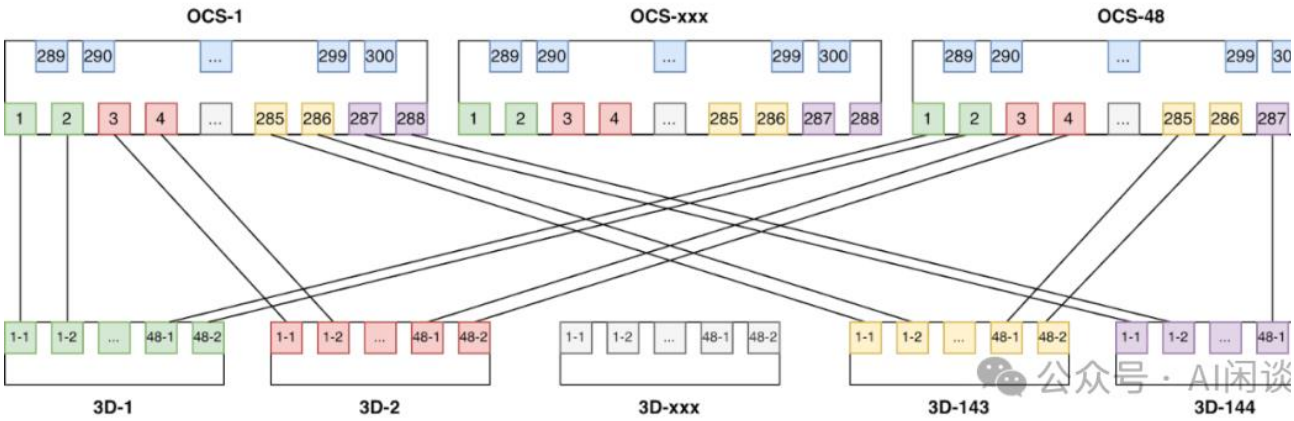
图50：谷歌 TPU v4 超节点网络拓扑



资料来源：AI 闲谈公众号公众号，东兴证券研究所

同理，谷歌 TPUv7 Ironwood 超节点（9216 芯片）设置 144 个 3D Torus，单个 3D Torus 对外互联带宽 19.2TB/s。

图51：谷歌 TPU v7 超节点网络拓扑



资料来源：AI 闲谈公众号公众号，东兴证券研究所

附：谷歌 TPU Scale up 网络演进与 TPU 代际发展紧密同步。

- TPU v2 首次引入 ICI Link，配置 4 条 Link，每条带宽 62 GB/s，ICI 总带宽为 248 GB/s，尚未引入光模块。
- TPU v3 保持了与 v2 相同的 Link 数，每条带宽从 62 GB/s 提升至 82GB/s，ICI 总带宽达到 328GB/s，首次引入光互连技术，采用 400Gbps 有源光缆（AOC），光通道波特率为 50G。
- TPU v4 采用 3D 拓扑，在降低单条 Link 带宽的情况下，Link 数量提升至 6 条，ICI 总带宽达到 300 GB/s，光模块升级为 400G OSFP，同时引入光交叉连接（OCS），光通道波特率仍为 50G。
- TPU v5p，采用 3D 环面设计，Link 数量保持 6 条，单链路带宽从 50GB/s 翻倍提升至 100GB/s，ICI 总带宽达到 600GB/s，光模块升级为 800G OSFP，光通道波特率提升至 100G。
- TPU v7 延续 3D 环面设计，Link 数量保持 6 条，单链路带宽从 100GB/s 翻倍提升至 200GB/s，ICI 总带宽提升至 1200GB/s，沿用 800G OSFP 光模块，光通道波特率提升至 200G。

表12：谷歌 Scale up 网络演进与 TPU 代际发展紧密同步

	TPU v2	TPU v3	TPU v4	TPU v5p	TPU v7
首发时间	2017	2018	2022	2023	2025
互连拓扑	2D 环面	2D 环面	3D 环面	3D 环面	3D 环面
Link 数	4	4	6	6	6
带宽/Link	62GB/s	82GB/s	50GB/s	100GB/s	200GB/s
OCS 数量	\	\	48	48	48
ICI 带宽	248GB/s	328GB/s	300GB/s	600GB/s	1200GB/s
ICI 光模块	\	400Gbps AOC	400G OSFP	800G OSFP	800G OSFP
光通道速率	\	50G	50G	100G	200G

信息来源：AI 闲谈公众号，东兴证券研究所

4.3 总结：光互联 Scale up 网络实现技术标准化，技术路线独树一帜

谷歌 TPU 超节点建立成熟的光互联 Scale up 网络。从技术成熟度看，2023-2025 年谷歌陆续推出 TPU v4、TPU v5p、TPU v7 三代超节点，完成了技术路线探索和方案标准化。此外 TPU v7 也获得外部企业认可。2026 年，Anthropic 将直接从博通采购近 100 万颗 TPU v7 Ironwood AI 芯片，本地部署在其控制的数据中心。2027 年，谷歌将推出第 8 代 TPU，对标 Nvidia Vera Rubin。可以看到，届时谷歌 TPU 超节点的性能指标进一步优化提升。

谷歌 TPU 超节点竞争优势建立在 OCS 交换机，技术路线独树一帜。相比英伟达、华为、AMD 等超节点厂商，谷歌是全球首个将光电路交换机（OCS）大规模商用部署于 Scale up 网络的企业，技术路线独树一帜。谷歌 OCS 交换机，涉及精密光学、机械工程与半导体工艺的深度交叉应用，在光互联领域构筑一道高壁垒的技术护城河。

相较于电分组交换机，光电路交换技术具备诸多优势：光电路交换机可跨多代光收发模块技术复用、光电路交换机的每比特能耗较电分组交换机低数个数量级、光电路交换机引入的时延极小。

OCS 交换机商用落地存在多重困难：光电路交换机需扩展至数百个端口以支撑足够数量 NPU 互连；受限于光电路交换机的控制软件和反射镜配置时延，商用光电路交换机的交换时延通常为 10~20 毫秒；为降低链路功率，光电路交换机插入损需要控制在理想水平。

为搭建高性价比、大规模的光交换层，谷歌创新研发三大核心硬件组件：光电路交换机、波分复用光收发模块和光环形器。其中谷歌 Palomar 光电路交换机的光学核心模块是实现光转向功能的 MEMS 微反射镜；波分复用光收发模块是提升布线效率、支撑大规模且持续扩张数据中心的关键；光环形器是实现光电路交换机链路双向通信的核心器件，将所需的光电路交换机端口和光纤数量减半。

5. AMD：UALink 成为重要开放标准，超节点有望成为英伟达有力竞品

5.1 UALink：代表开放标准路线，受到业内广泛支持

相比英伟达 NVLink 以及谷歌 ICI 协议，UALink 通过联合制定标准来避免被单一厂商锁定。UALink（Ultra Accelerator Link）是用于连接 AI 加速器的开放、高效的 Scale-Up 互联标准，代表开放标准路线，旨在打破单一厂商垄断，使得基于开放标准的异构融合将成为可能。我们认为，UALink 在 Scale-up 中的产业定位类似移动互联网发展历程中的“安卓”开放生态。经历一年多发展，UALink 协议 1.0 持续完善，具体历程如下：

- 2024 年 5 月，AMD、博通、思科、谷歌、惠普（HPE）、英特尔、Meta 和微软携手成立 UALink 小组，旨在推动数据中心 AI 连接。
- 2024 年 10 月，博通退出董事会，新增亚马逊网络服务（AWS）、Astera Labs 两家公司为董事会成员，UALink 联盟正式成立，主推 AI 服务器 Scale UP 互连协议——UALink。
- 2025 年 1 月，阿里巴巴、Apple、Synopsys 当选为 UALink 联盟新增董事会成员，进一步扩大联盟影响力。
- 2025 年 2 月，发布 UALink 200G 候选规范，为正式标准奠定基础。
- 2025 年 4 月，正式发布 UALink 200G 1.0 规范，并充分考虑中国市场落地，定义了 AI 计算舱中加速器和交换机之间通信的低延迟、高带宽互连，支持最多 1024 个加速器实现每通道 200G 的扩展连接。
- 2025 年 12 月，UALink 1.0+协议陆续发布，新增 INC 在网计算、IO Die、12G DL/PL。

图52： UALink 发展时间线

资料来源：半导体行业观察公众号，东兴证券研究所

UALink 联盟受到业内广泛支持，成员单位超过 100 家（截止 2026 年 1 月底）。目前 UALink 联盟董事会成员有阿里巴巴、AMD、苹果、Astera Labs、AWS、CISCO、谷歌、HPE、Intel、Meta、Microsoft、Synopsys 等行业巨头以及产业链关键厂商。在 100 多家 UALink 联盟成员单位中，美国企业约 44 家，中国企业约 37 家，其中字节跳动、盛科通信、新华三、海光信息、中兴通讯等中国企业均加入 UALink 联盟。

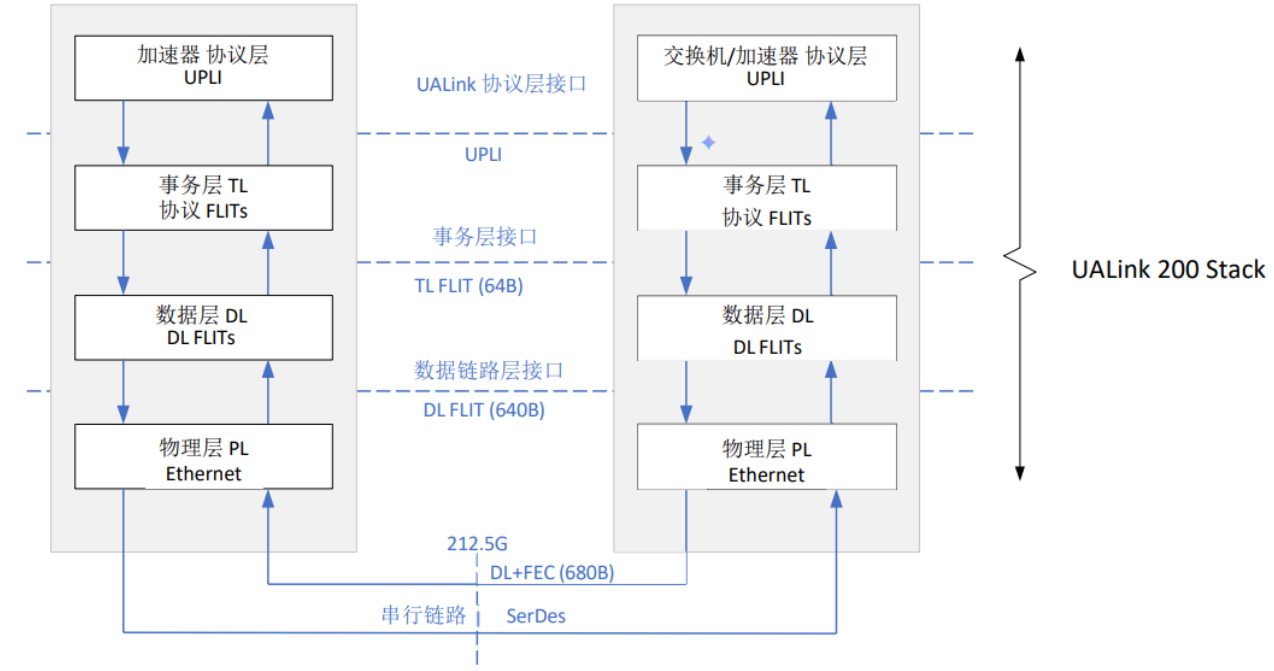
图53： UALink 联盟成员名单

资料来源：UALink 官网，东兴证券研究所

Scale-Up 互联协议在物理层上已基本收敛至以太网。在超节点 Scale-Up 场景中，GPU 卡间互联物理层主要有 PCIe 和以太网两种技术路线。以太网物理层 SerDes 技术凭借更高的单通道速率（当前主流达 112Gbps，224Gbps 已商用），相较 PCIe 5.0 x16（双向约 128GB/s）具备显著的带宽扩展潜力。

UALink 定义全栈开放协议，物理层采用标准以太网 PHY，但链路层和传输层完全重新定义，旨在实现总线级的性能。UALink 协议栈自底向上分为物理层（PL）、数据链路层（DL）、事务层（TL）和协议层（UPLI）。

图54： UALink 协议栈架构



资料来源：UALink《Scale-Up 互联技术白皮书》，东兴证券研究所

对比博通 SUE 协议，UALink 可实现单节点 1024 个加速器互联。博通作为以太网交换芯片的领导者，选择中途退出 UALink 董事会转向推动 SUE，采用“网络总线化”思路，在传统以太网协议和交换技术基础上，针对 Scale-Up 需求进行协议优化和交换芯片架构创新，意在将其市场主导地位从 Scale-Out 自然延伸至 Scale-Up 领域。而 UALink 采用“总线网络化”思路，以传统总线技术为基础，结合网络技术元素，满足 Scale-Up 高带宽和扩展性需求。

（1）延迟优化方面，UALink 通过更短线缆（≤4 米）和优化的协议栈实现更低的端到端延迟。SUE 则通过报头压缩和交换芯片优化，在较长线缆（≤10 米）条件下仍保持较低延迟。

（2）互连规模方面，UALink 通过专用交换机实现更大规模的单 Pod 互联（1024 个加速器）。SUE 则通过与标准以太网设备的兼容性，支持更灵活的扩展架构，如两层 Clos 架构可支持 10 万+XPU。

（3）应用场景方面，UALink 更专注于超节点内部的 GPU 协同计算，特别是需要显存共享的大模型训练场景。SUE 则兼顾单机架 Scale-Up 和跨机架 Scale-Out，支持更广泛的混合架构部署。

表13：UALink 与 SUE 技术对比

性能指标	UALINK	SUE
物理层	基于以太网 SerDes	完全兼容标准以太网 SerDes
数据链路层	封装 64B 事务为 640B 数据帧，添加 CRC 校验	保留以太网 MAC 层，新增 10BA 转发包头
传输层	事务层（压缩寻址，直接内存操作）和协议层（内存语义逻辑）	简化协议栈，直接处理内存语义通信，避免 IP/UDP 层开销
单通道带宽	200Gb/s	200Gb/s
链路数量	x1，x2，x4	x1，x2，x4
延迟	交换机延迟 100-150ns，端到端 RTT<1μs	交换机延迟 250ns，端到端 RTT<400ns
互联规模	单 Pod 支持最多 1024 个加速器	单跳支持 512 个加速器，扩展至 1024 个加速器
线缆长度限制	不超过 4 米	不超过 10 米
显存共享支持	完整支持，实现 GPU 间显存直接访问	不直接支持，通过集体操作卸载优化通信效率
生态兼容性	需专用交换机，兼容 PCIe/CXL 等服务器协议	完全兼容标准以太网设备和工具链

资料来源：中国信息通信研究院信息化与工业化融合研究所《超节点关键技术与产业发展态势研究》，架构师技术联盟公众号，东兴证券研究所

UALink 交换芯片预计 2027 年商用。UALink 交换机 ASIC 芯片主要供应商有 Cisco、Arista、Marvell、HPE 等。产品落地进展方面，Siemens EDA 已推出 UALink VIP（验证 IP），Broadcom、Marvell 等正在开发 UALink 兼容交换机芯片，Astera Labs 等厂商已发布支持 UALink 的连接芯片和模块。

5.2 AMD 超节点：英伟达 NVL72 系列有力竞品，有望实现市场突破

AMD 超节点方案进入标准化阶段。2024 年，MI300 系列超节点方案单节点最高 8 卡，被超微、HPE 等集成到 4U/8U 高密度服务器，用于中小模型训练与推理。2025 年-2026 年，AMD 陆续正式发布 MI400 系列 Helios 机柜，单柜最大 72 卡，面向超大规模模型训练。其中，在 2026 年 CES 大会，AMD 发布 Helios 机架级平台，搭载 72 颗 MI455 GPU 的 UALoE72 纵向扩展域。

从时间维度，MI455 超节点推出时间落后英伟达 GB200 NVL72（2024 年 3 月发布）约 2 年时间，但由于 AMD 在算力芯片先进制程优势，双方差距有望快速缩小。在后续超节点发布节奏方面，AMD 将于 2027 年推出下一代超节点 MI500 UAL256。该超节点支持 256 卡规模，采用开放解构架构，整体由三个互联的机柜构成。

图55： AMD Helios AI Rack MI455X 72x GPU 超节点外观

资料来源：企业存储技术公众号，东兴证券研究所

MI455x 系列 Helios 机柜具有较强竞争力，有望实现市场突破。MI455x 系列 Helios 机柜是 AMD 在机柜设计上的一次跨越级产品，AMD 称之为 UALoE72（UALink over Ethernet），是目前业界最能挑战英伟达的 NVL72 机柜的竞品，计划 2026 年下半年出货，有望实现市场突破。算力方面，UALoE72 的 FP4 算力达 2.9EFLOPS，与 Vera Rubin 的 3.6 EFLOPS 差距不大；内存方面，UALoE72 配备 31TB HBM4 内存容量，是 Vera Rubin NVL72 的 1.5 倍；此外，Helios 采用开放标准的 UALink 协议，互联总带宽与 Vera Rubin 持平。

表14：AMD MI455x 系列 Helios 与英伟达 Vera Rubin NVL72 参数对比

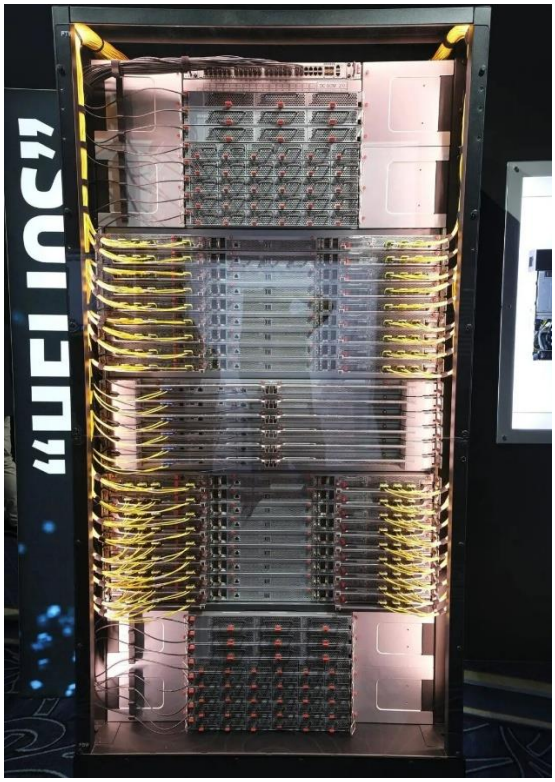
	AMD Helios	Vera Rubin NVL72
GPU 数量	72 Instinct MI455X GPUs	72 Rubin GPUs
CPU 数量	18 EPYC Venice CPUs	36 Vera CPUs
CPU 核心数量	4608 Zen6 Cores	3168 NVIDIA Olympus Cores
制造工艺	Advanced 2nm/3nm	TSMC Custom 2nm/3nm
FP4 算力	2.9EFLOPS	3.6EFLOPS
互联技术	UALink	NVLink6
互联总带宽	260TB/s	260TB/s
内存容量	31TB HBM4	20.7TB HBM4

资料来源：公司官网，智能计算芯世界公众号，新智元公众号，东兴证券研究所

对比 GB200 NVL72 机柜，Helios 机架在功耗领域优势显著。Helios 机架采用双宽机架设计,宽度从 1 个机架提升到 2 个机架。双宽机架设计主要源于大型数据中心中的关键限制因素是功耗而不是占地面积，双机架宽度在复杂性、可靠性和性能间实现平衡。Helios 超宽机柜总功耗 64.8kW，重量约 3.2 吨。

（GB200 NVL72 机柜尺寸为长 1068 毫米、宽 600 毫米、高 2495 毫米；重约 1.36 吨；功耗 145KW。）

图56： AMD MI455x 系列 Helios 机架外观



资料来源：AMD，东兴证券研究所

单台 Helios 机架布置 18 个计算节点与 6 个交换节点。MI450 Helios 机架由 18 个计算托盘（顶部 9 个，底部 9 个）与 6 个交换托盘交错排列，每个计算托盘容纳 4 颗 MI450 GPU，整机柜形成 72 GPU 的统一计算域；每个交换托盘包含两个 Tomahawk 6 102.4T 以太网交换机。双宽结构为未来升级预留物理空间，例如可扩展至 144GPU 配置，而无需重新设计机架基础设施。

图57： AMD MI450X UALoE72 Helio 机架示意图

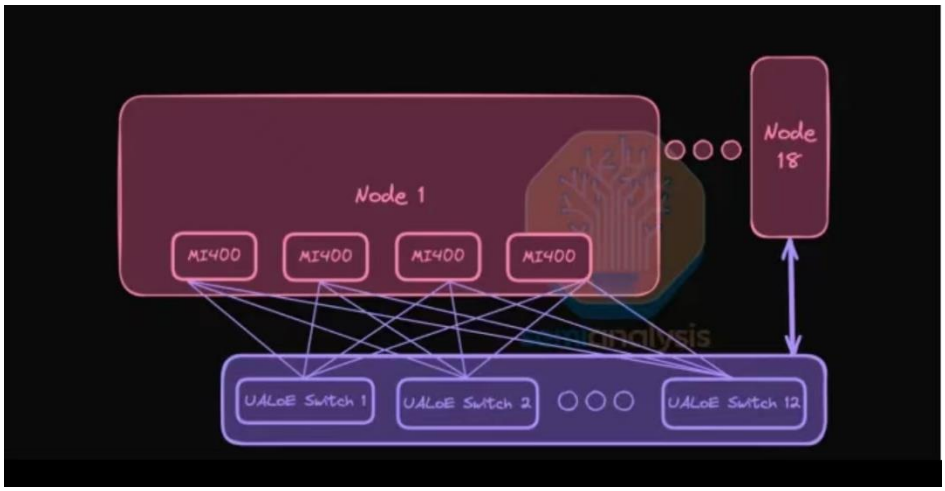
资料来源：semianalysis，东兴证券研究所

MI400 系列 Helios 机柜 Scale up 方案实现总交换容量 260 TB/s，与 VR200 NVL72 持平。

计算节点：MI400 使用每条 200Gbit/s 的 72 条通道实现的单 GPU 1.8TB/s（单向）纵向扩展带宽；

交换节点：交换托盘包含两个 Tomahawk 6 102.4T 以太网交换机和四个连接器。每个连接器有 432 个差分对，相当于 216 条 UALink 通道，每条通道 200G。每个 102.4T 交换机提供 512 个端口，交换托盘总共有 1024 个端口。由于每个连接器仅输入 216 条通道，总共 864 条通道，导致 TH6 以太网交换机有 160 个未使用的端口。

图58： AMD MI400s UALoE72 Scale Up 拓扑示意图



资料来源：semianalysis，东兴证券研究所

5.3 总结：UALink 成为重要标准，Helios 机架有望成为行业主流选择

作为 Scale up 网络开放技术路线方，UALink 成为重要标准。2025 年 1.0 版本规范正式发布；2026 年，UALink 2.0 版本有望发布。我们认为，目前 UALink 正处于从标准制定阶段走向产品落地阶段，预计 UALink 生态将在 2027 年迎来突破发展，被众多数据中心接纳。目前 UALink 联盟受到业内广泛支持，截止 2026 年 1 月底，成员单位超过 100 家，将成为英伟达 NVLink 有利挑战方。

AMD 超节点 Helios 机架有望成为行业主流选择。Helios 机架采用双宽机架设计,宽度从 1 个机架提升到 2 个机架，在复杂性、可靠性和性能间实现良好平衡。从算力、内存、互联带宽等指标，MI455x 系列 Helios 机柜是目前业界最能挑战英伟达 NVL72 机柜的竞品；而在功耗领域，对比 GB200 NVL72 机柜，Helios 机架优势显著。此外，双宽结构为未来升级预留物理空间，例如可扩展至 144GPU 配置，而无需重新设计机架基础设施。

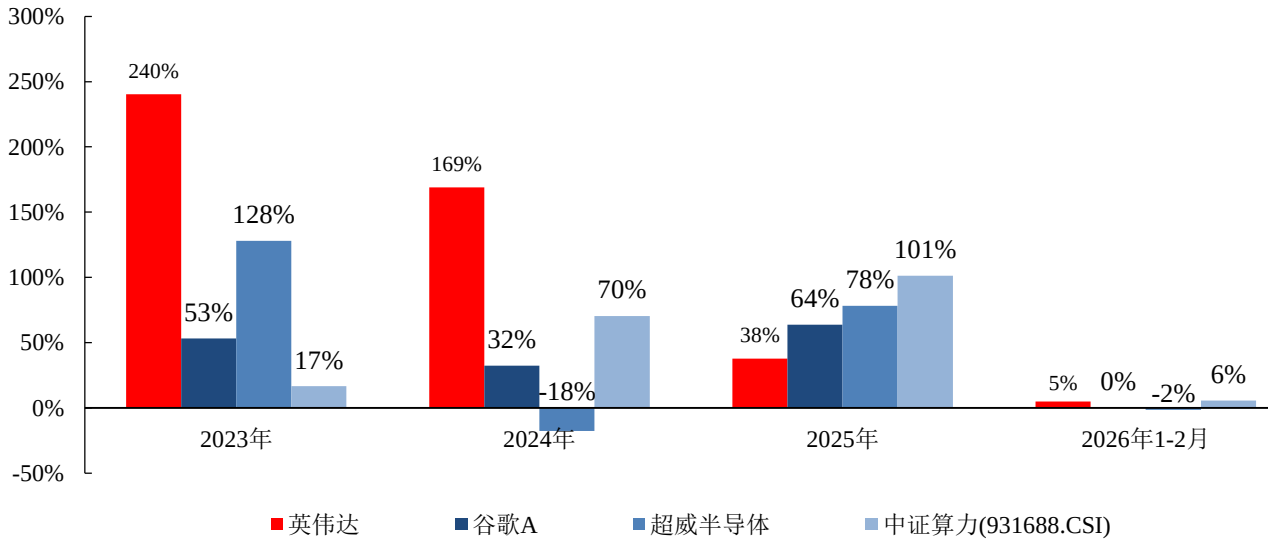
6. 投资建议

自 2025 年开始，超节点成为 AI 算力网络重要的技术创新方向。从 AI 基建竞争维度，AI 芯片厂商从芯片算力性能竞争延续至芯片+Scale up 网络的双战场。因此，除了原先英伟达、华为、AMD 以及谷歌等芯片公司，全球更多厂商加入超节点赛道的竞争，包括微软、Meta、Amazon、中国移动、阿里巴巴、字节跳动、腾讯、百度、中科曙光、中兴通讯、浪潮信息、紫光股份（新华三）、海光信息、沐曦股份、恒为科技等。

全球超节点竞争格局尚未确立。英伟达目前处于领先地位，但谷歌、AMD、华为等巨头在超节点领域的持续发力已经对英伟达一家独大的格局构成挑战。从股价表现看，2023-2024 年，英伟达股价大幅跑赢谷歌、AMD 以及 A 股中证算力指数。但在 2025 年，英伟达股价累计涨幅 38%，显著落后于谷歌、AMD 以及 A 股中证算力指数。我们认为，在超节点技术发展中，市场将继续对谷歌、AMD 以及国产超节点板块价值重估。

投资建议：（1）看好谷歌、AMD 以及国内超节点厂商；（2）看好英伟达、谷歌与 AMD 超节点供应链，包括 PCB 背板、高速铜缆、光模块、供电与液冷系统等；（3）基于交换机及芯片是 Scale up 网络互联的关键设备，看好谷歌光路交换机（OCS）核心零部件供应商以及 UALink 标准下的交换机芯片研发商。

图59：2023-2026 年 2 月英伟达/谷歌/超威半导体/中证算力当年累计涨跌幅对比



资料来源：iFinD，东兴证券研究所

7. 风险提示

（1）LLM 训练与推理技术路径变化；（2）超节点性能与功耗有待平衡；（3）受供应链影响，各厂商超节点出货量低于预期；（4）AI 应用端增长不及预期。

分析师简介

石伟晶

首席分析师，覆盖传媒、互联网、云计算、通信等行业。上海交通大学工学硕士。10 年证券从业经验，曾供职于华创证券、安信证券，2018 年加入东兴证券研究所。

分析师承诺

负责本研究报告全部或部分内容的每一位证券分析师，在此申明，本报告的观点、逻辑和论据均为分析师本人研究成果，引用的相关信息和文字均已注明出处。本报告依据公开的信息来源，力求清晰、准确地反映分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与，未来也将不会与本报告中的具体推荐或观点直接或间接相关。

风险提示

本证券研究报告所载的信息、观点、结论等内容仅供投资者决策参考。在任何情况下，本公司证券研究报告均不构成对任何机构和个人的投资建议，市场有风险，投资者在决定投资前，务必要审慎。投资者应自主作出投资决策，自行承担投资风险。

免责声明

本研究报告由东兴证券股份有限公司研究所撰写，东兴证券股份有限公司是具有合法证券投资咨询业务资格的机构。本研究报告中所引用信息均来源于公开资料，我公司对这些信息的准确性和完整性不作任何保证，也不保证所包含的信息和建议不会发生任何变更。我们已力求报告内容的客观、公正，但文中的观点、结论和建议仅供参考，报告中的信息或意见并不构成所述证券的买卖出价或征价，投资者据此做出的任何投资决策与本公司和作者无关。

我公司及报告作者在自身所知情的范围内，与本报告所评价或推荐的证券或投资标的不存在法律禁止的利害关系。在法律许可的情况下，我公司及其所属关联机构可能会持有报告中提到的公司所发行的证券头寸并进行交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。本报告版权仅为我公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。如引用、刊发，需注明出处为东兴证券研究所，且不得对本报告进行有悖原意的引用、删节和修改。

本研究报告仅供东兴证券股份有限公司客户和经本公司授权刊载机构的客户使用，未经授权私自刊载研究报告的机构以及其阅读和使用者应慎重使用报告、防止被误导，本公司不承担由于非授权机构私自刊发和非授权客户使用该报告所产生的相关风险和责任。

行业评级体系

公司投资评级（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数）：

以报告日后的 6 个月内，公司股价相对于同期市场基准指数的表现为标准定义：

强烈推荐：相对强于市场基准指数收益率 15%以上；

推荐：相对强于市场基准指数收益率 5%～15%之间；

中性：相对于市场基准指数收益率介于-5%～+5%之间；

回避：相对弱于市场基准指数收益率 5%以上。

行业投资评级（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数）：

以报告日后的 6 个月内，行业指数相对于同期市场基准指数的表现为标准定义：

看好：相对强于市场基准指数收益率 5%以上；

中性：相对于市场基准指数收益率介于-5%～+5%之间；

看淡：相对弱于市场基准指数收益率 5%以上。

东兴证券研究所

北京
西城区金融大街 5 号新盛大厦 B 座 16 层
邮编：100033
电话：010-66554070
传真：010-66554008

上海
虹口区杨树浦路 248 号瑞丰国际大厦 23 层
邮编：200082
电话：021-25102800
传真：021-25102881

深圳
福田区益田路 6009 号新世界中心 46F
邮编：518038
电话：0755-83239601
传真：0755-23824526