中兴通讯超节点白皮书

（2026 年 1 月）

ZTE中兴

目录

图目录

1 AI 算力架构演进：从芯片堆砌迈向系统级协同

随着 AI 模型参数规模突破万亿量级，算力需求已从单纯的 GPU 堆叠，转向全维度的系统架构重构。受限于单芯片物理功耗密度、互连带宽与内存容量瓶颈，其算力增长边际效益递减。当前研究与工程实践表明，系统级协同架构（如高带宽域互联）成为突破单芯片性能上限的主要技术路径。

这一转型的根本动因，在于单颗芯片的物理极限已成为制约算力发展的核心瓶颈。当模型规模远超单芯片的算力与显存容量时，传统分布式训练方法面临通信开销剧增、算力利用率骤降等严峻挑战。在此背景下，通过高速无损互联技术，将数十甚至上百个 GPU 芯片从逻辑层面整合为统一计算单元，对外可视为一台功能极强的“超级计算机”，已成为全球主流 AI 基础设施厂商与研究机构公认的下一代算力架构核心突破方向。这一架构革新不仅实现算力密度的跃升，更是达成系统级高效协同、降低大模型训练与推理综合成本的关键技术路径。

2 超节点系统架构设计

超节点是通过高速互联协议与专用交换芯片构建的高带宽域（High-Bandwidth Domain），将数十至数百颗 GPU 芯片在逻辑上整合为统一编址、低延迟、高带宽的协同计算系统。该架构保留 GPU 的物理独立性，通过统一虚拟内存地址空间与无损互联，实现类单机的编程与调度体验。超节点并非 GPU 的简单物理堆砌，而是融合多芯片、整机硬件、高速互联与配套软件的集成系统，依托算法仿真、工程设计、软硬联合优化等综合手段，构建的极致协同计算系统。超节点对芯片的算传存基础能力，硬

件设计的集成能力，高带宽高可靠可扩展的互联能力，以及面向底层算法要求的软硬协同能力都提出了极高的要求，需实现端到端全链路的平衡与优化，方能构建真正意义上的最优“单一”算力产品形态——超节点。

为实现这一系统级协同，构建超节点，需要遵循以下四大核心前提：

第一，芯片能力的均衡性。构建超节点芯片需要满足算力、显存与互联带宽的均衡，并非所有的 GPU 芯片都具备构建超节点的潜力。比如，算力被裁剪的芯片，其计算能力难以匹配高规格的互联带宽，易造成带宽资源浪费；反之，芯片算力充足，但互联总带宽不足、互联链路数量过少，也无法支撑 GPU 互联规模的扩大，导致算力无法充分发挥。

第二，互联架构的有效性。超节点互联架构需兼顾通信效率、扩展性与场景适配性三大核心要求。原则上超节点内任意 GPU 间的互联带宽是机间互联的 8 倍左右，有助于降低通信开销、提高 GPU 的 MFU（模型 FLOPs 利用率）。而传统总线（例如 PCIe）或低容量交换芯片的方案，无法实现真正意义上的全互联（Full Mesh）。业界虽有厂商在互联技术上进行创新尝试，如定制拓扑或优化交换路径，但在架构的通用性与灵活性之间仍需权衡。面对不同并行策略带来的差异化通信需求，理想的超节点互联架构需具备自适应能力，以更好支持多样化大模型训练的需求。

第三，内存访问的便捷性。超节点内所有 GPU 需支持统一内存编址，以支持各种原语级的内存访问，确保超节点的内存访问与单 GPU、单服务器保持一致的灵活便捷性。同时，由于 GPU 品类的特性差别，以及消息大小对并行访问效率的影响，超节点还需同时支持消息语义和内存语义，在编程易用性与数据访问效率之间达到最佳平衡。

第四，超节点架构扩展的原生性。单体的机柜级超节点需具备灵活扩展能力，可平滑扩展为更大的集群超节点（如从 128 单体超节点可扩展到 8192 的集群超节点）。与 Scale-Out 的互联模式不同，集群超节点的互联依旧属于 Scale-Up 域，且满足任何 GPU 的带宽是机间互联的 8 倍。该设计确保面对未来更大参数量模型训练需求或技术演进时，可以实现算力灵活选择，按需配置，最终达到性能和成本的最佳平衡。

下文将从芯片能力，系统及整机设计等维度，阐述超节点构建的基础要求，并深度分析业界构建超节点的技术方向和技术路线。

2.1 芯片：从计算到互联的协同演进

2.1.1 算力芯片的演进

单纯堆砌低性能计算单元无法实现算力密度的线性增长。系统性能的增益取决于互联带宽、显存容量与算力的协同匹配，而非单元数量的简单叠加。因此，算力密度并非由芯片数量决定，而是指单位体积内可释放的有效算力。

在机柜功耗和物理尺寸受限的前提下，提升单芯片算力密度是实现超节点极致算力密度的首选路径。英伟达历代架构的演进，正是该理念的典型工程化实践：每一代 NVLink 互联带宽的倍增，均与算力、显存容量及显存带宽实现同步提升，确保单位互联带宽所支撑的有效算力持续处于饱和状态，避免资源浪费。在此基础上，英伟达通过 NVLink-C2C（Chip-to-Chip）互联技术，将 CPU 与 GPU 封装于同一基板（Interposer），实现统一内存寻址与高带宽低时延通信，构建逻辑层面的“超级芯片”，完成从“物理多芯片”到“逻辑单芯片”的整合，持续提升芯片级算力密度。

1. **对 GPU 的核心需求：互联先行，算力、显存同步放大**

- **互联可扩展性**：NVLink、UALink、SUE、ETH-X 等主流互联协议，均需支持千卡级 HBD 高带宽域的扩展能力。
- **算力与显存同步升级**：互联带宽每实现一倍提升，FP4 算力、显存容量、显存带宽完成近乎同比例放大，实现三者与互联带宽的精确匹配。

2. **对 CPU 的核心需求：单核性能和 IO 扩展能力**

- **单核性能**：通过更高的主频、微架构设计优化（核心是提升 IPC），将无法并行的控制、预处理、通信框架线程的处理延迟压到微秒级，保障系统调度效率。
- **IO 扩展能力**：原生支持更多的 PCIe 通道数及更加丰富的 IO 接口类型；通过合理的 I/O 设计，可在节点内省去 PCIe Switch，降低系统成本。

超节点的极致算力密度，首先要取决于“单芯片有效算力密度”能否随互联带宽线性甚至超线性增长；其次依赖于 CPU 单核性能与 I/O 扩展能力的同步提升。唯有 GPU/CPU 在算力芯片层级完成“带宽-算力-显存”三角协同匹配，整机柜才能用更少芯片、更低功耗、更简拓扑，释放出更高且可持续的有效算力。

2.1.2 高速互联技术的突破

超节点的实现核心在于构建高带宽、低延迟的 Scale-Up（纵向扩展）通信域。英伟达率先通过 NVLink 互联协议与 NVSwitch 交换芯片的组合，确立了早期超节点的技术范式。以英伟达 Blackwell 架构为例，其 NVSwitch 技术支持集成 18 或 36 个 GB200 超级芯片（对应 36 或 72 颗 GPU），分别构建 NVL36 或 NVL72 超节点，并进一步

借助 NVLink 光互联扩展至 576 卡的集群超节点。该架构下，单卡间 NVLink 双向带宽达 1.8 TB/s，NVL72 超节点内 GPU 间互联总带宽高达 130 TB/s。这种基于专用交换芯片实现的 GPU 直连通信域，打破了传统 PCIe 总线的性能瓶颈，为业界提供了重要的技术参考。然而，随着技术的不断演进，超节点互联正逐步突破单一封闭生态，迈向多元开放的发展路径。

2.1.2.1 物理层技术选型

在超节点（Scale-Up）场景中，GPU 间互联需要数百 GB/s 至 TB/s 带宽能力承载 TP（Tensor Parallelism）、EP（Expert Parallelism）并行计算流量，GPU 卡间互联物理层主要有 PCIe 和以太网两种技术路线。

- PCIe：作为通用总线，PCIe 专注于短距、低延迟的设备互联。其设计受功耗和延迟的严格约束，导致 SerDes 速率提升相对保守，PCIe 物理层单通道 SerDes 速率和以太网差距较大。当前主流的 PCIe 5.0 x16 配置，其双向带宽约为 128GB/s，难以满足超节点对 TB 级带宽的需求。
- 以太网物理层：以太网 SerDes 技术迭代迅速，主流速率已达 112Gbps，224Gbps 产品已进入商用阶段。其支持多通道灵活绑定，能够轻松实现 TB/s 级端口带宽。

在超节点 Scale-Up 场景中，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，更契合 AI 训练对 TB 级互联带宽的严苛需求。

2.1.2.2 Scale-Up 互联协议生态格局

当前，Scale-Up 互联协议生态正演变为“垂直整合封闭”与“开放架构”双轨并行的复杂竞争格局，技术路线的分化与重组日益激烈。

在垂直整合路径上，以 NVIDIA 和 Google 为标杆，企业通过全栈自研的专用互联协议（如 NVIDIA NVLink、Google ICI）、硬件接口及软件栈构建高壁垒。然而，这种全栈闭环体系在提供稳定高效服务的同时，也伴随着生态封闭、技术锁定及高昂迁移成本的风险。

在开放架构路径上，产业界致力于打破私有协议垄断，构建多元化生态，目前呈现“国际双轨引领、物理层收敛于以太网”的整体特征：

- 国际路线：UALink 定义了全栈开放协议，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，其交换芯片预计 2027 年商用。ESUN（Ethernet Scale-Up Network）聚焦数据链路层，在现有 L2/L3 以太网基础上进行增强，实现无损、无收敛的交换拓扑，生态构建基于博通 Tomahawk Ultra 等商用交换芯片。
- 国内路线：国内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。

总体而言，Scale-Up 互联协议在物理层上已基本收敛至以太网。未来，国内亟需在兼顾国际路线参考、继承与发展的基础上，加快统一标准的制定，打破生态割裂，形成具备竞争力的开放互联标准落地典范。

2.1.2.3 统一内存编址与访问

超节点支持统一内存地址编址，是解决“多 GPU 协同效率” 与 “数据一致性” 的关键，更是其区别于普通分布式集群、实现高性能算力聚合的前提。这一设计打破硬件孤岛，让所有 GPU 共享同一地址空间，跨 GPU 数据无需物理拷贝，通过地址即可直接读写。无论是 GPGPU 采用 Load/Store 内存语义，还是 DSA（领域专用架构）GPU 采用 DMA 消息语义，均可确保不同 GPU 对同一数据状态的精确感知，实现类本地内存的高效协同，大幅降低软件复杂度。开发者可以根据需要，灵活选择内存语义的便捷编程或消息语义的高效访问，无需手动处理数据同步、地址映射和冲突控制，统一地址简化访问逻辑，标准化事务处理明确交互准则，减少适配成本，让开发者聚焦算法优化。

2.1.2.4 在网计算

Scale-Up 交换芯片除了通过高带宽、低时延的互联能力提升模型训练效率外，其在支持传统稠密模型和动态 MoE 模型的在网计算（In-Network Computing）方面展现出关键优势和实际收益：

- **在传统稠密模型训练中**，交换芯片通过集成在网计算技术，将原本由计算节点承担的 All-Reduce 操作卸载至交换芯片内部完成，将通信交互复杂度从传统的 O（logN）降低至 O（C）（C 为网络层级），大幅减少节点间消息传递次数，降低通信延迟；
- **在动态 MoE 模型训练中，**Dispatch Multicast（专家分发）和 Combine Reduce

（结果聚合）操作带来巨大通信开销，严重制约系统扩展性与效率。通过引入在网计算技术，将数据复制（Multicast）、加权归约（Reduce）等高负载操作从GPU 端卸载至交换芯片：Dispatch 阶段源端 GPU 发送带宽占用与内存复制次数显著下降，GPU 更专注前/反向计算，干线流量减少（常见为>30%），尾时延得到显著改善，分发阶段时延可下降 20%–50%；Reduce 阶段端到端时延下降（常见 40%–60%+），回传链路流量显著降低，尾时延与抖动降低，训练步长更稳定。动态 MoE 模型将 Dispatch Multicast 与 Combine Reduce 卸载到在网计算，带来显著的带宽节省、尾时延下降、GPU 利用率提升与规模扩展能力增强，是支撑大规模 MoE 训练与推理的关键基础能力。

2.1.2.5 Scale-Up 可扩展性

Scale-Up 可扩展性需要从互联协议、互联拓扑、物理形态、和互联介质四个关键方面考虑。

- **互联协议**：为支持未来大规模 GPU 集群的通信需求，协议设计需具备良好的扩展性和前瞻性，建议 GPU ID 的关键标识 bit 位预留足够的空间，以满足未来十万级 GPU 集群规模演进的寻址需求；
- **拓扑层面**：为避免通信瓶颈，需要线性扩展与无收敛扩展架构，支持一级交换或者二级交换无收敛扩展；
- **物理形态**：单机柜高密度扩展与多机柜横向扩展在一定时期内共存，机柜作为基本扩展单元，需模块化设计，支持“即插即用”式扩容；

- **互联介质**：遵循“能铜尽铜，距远用光”原则。单柜内或相邻机柜间优先采用电互联；多柜的远距离连接需要采用光互联。

中兴通讯依托在高速互联接口 SerDes、以太网、在网计算以及网络交换几大关键技术方面长期积累，自主研发凌云大容量交换芯片，为 GPU 提供开放、超带宽的互联能力；互联拓扑层面，从点对点互联升级为大规模全对等互联拓扑，适配数十到数百颗芯片协同；带宽与时延方面，从百 GB/s、微秒级跃升至 TB 级带宽与百纳秒级时延，满足海量数据传输；互联协议方面，凌云芯片除支持开放高速互联协议外，兼容 RDMA、Clink、OISA、Ethlink、SUE、UEC 等主流协议；支持 Reduce、MoE 在网计算，优化通信语义等，提升整体系统效率。

2.2 单体超节点与 Matrix 超节点

超节点硬件形态正加速迭代演进。回顾其演进历程，在超节点探索期，行业普遍采用“8 卡机型互联”的技术路径，试图通过光互连方式构建大规模的 Matrix 超节点。例如，NVIDIA 使用 H100/H200 的 8 卡机型通过两层 NVLink 互联构建 256 卡超节点，但由于光互连的成本及可靠性问题，实际上该产品未能实现大规模商业化落地；国内部分厂商则借鉴了该设计，同样使用 8 卡机型及两层互联构建了百卡规模的超节点形态。

随后，NVIDIA 调整技术路线，转向在单机柜内构建更多卡的互联架构，确立了“去光用铜”策略，以降低成本并提升整体可靠性，成功推出系列化的 NVL36/72 单体超节点机型。单体超节点承袭刀片服务器的设计理念，将计算托盘、交换托盘、液冷分

配与供电背板一体化集成于单机柜，形成一个独立的 HBD 高带宽域。该架构的核心优势体现在高集成度上，可在单位空间实现更高算力密度，提高了数据中心基础设施的利用效率。在硬件架构层面，早期主流采用 Cable Tray（线缆托盘）方案，而随着 SerDes 技术的持续演进和芯片迭代升级的需求驱动，正交架构方案逐步成为行业新的技术方向。

虽然单体超节点是行业主流，但早期的 Matrix 互联设计思路并未被摒弃，反而演进为构建超大规模集群的关键技术支撑。在持续提升单体超节点集成度的同时，行业依然需要通过柜间互联技术构建更大规模的集群超节点，统一满足高带宽互联、全局地址分配、内存语义及消息语义兼容等核心需求。

尽管技术路线仍然存在迭代变数，但行业已形成明确共识：既要通过硬件架构创新，持续提升单体超节点的集成密度与运行稳定性，也要依托灵活的集群扩展模式，实现整体成本优化。基于对整机柜超节点方案的深度工程实践，中兴通讯创新提出 Orthogonal Electrical eXchange （OEX）正交无背板互联交换架构。该架构在保持原有整机柜超节点设计优势的基础上，实现了计算托盘和交换托盘的正交无背板互联，不仅提高了算力密度，保证了高速信号完整性，还进一步增强了系统的可靠性和可维护性；同时通过开放 OEX 机械与电气规范，支持第三方计算/交换托盘标准化接入，向后续多厂家协作共同构建开放、融合、创新的国产化整机柜超节点生态，迈出了关键性一步。

2.2.1 Nebula 单体超节点

2.2.1.1 OEX 架构创新

OEX 是一种正交无背板互联交换架构，其核心在于实现计算托盘与交换托盘之间的垂直交叉物理连接，消除传统线缆托盘（Cable Tray）带来的信号损耗与可靠性风险。该架构通过简化互联路径、提升信号完整性，为构建高密度、高可靠性的单体超节点提供物理基础。在超节点设计中引入 OEX 架构，通过正交连接器与单级交换拓扑，实现计算节点与交换节点之间的垂直交叉互连，从而彻底摆脱了传统线缆的束缚。在高速信号完整性、可靠性和可维护性方面相比传统的线缆（Cable Tray）方案更具优势，也为后续架构扩展和演进预留了足够的空间。

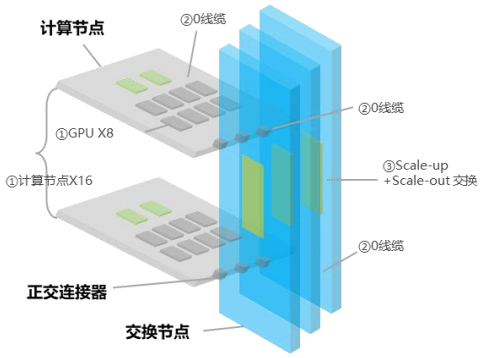


图 2-1 OEX 互联示意图

OEX 架构特性包括：

- **信号完整性**：通过计算和交换节点的正交无背板互联，显著降低了通信损耗，保障了信号完整性。在典型 112G 高速信号场景下，整体 SerDes 链路长度缩短了 30%以上，可以消除 Cable Tray 线缆引入的 6.5dB 插损，降低了误码率，保证

整体端到端链路插损余量大于 3dB，确保大规模集群通信的高速与稳定。

- **集成密度**： 采用无线缆互联设计，通过消除成千上万根高速线缆，极大地释放了机柜内部宝贵空间，为在标准机柜内集成更多的算力芯片提供了物理基础，实现了单位空间算力密度的显著提升。
- **可靠性与可维护性**： 无线缆设计从根本上减少了因线缆松动、老化或连接器故障导致的宕机风险。极短的板间互联路径也显著降低了信号衰减，提升了系统长期运行的稳定性，并简化了运维流程，系统故障修复时间 MTTR 从小时级缩短为分钟级。
- **组网成本**：机柜内部交换板内集成参数面 leaf 交换，消除了传统参数面组网 leaf 层级交换机、光模块和光纤使用，降低了系统组网的成本和复杂度。

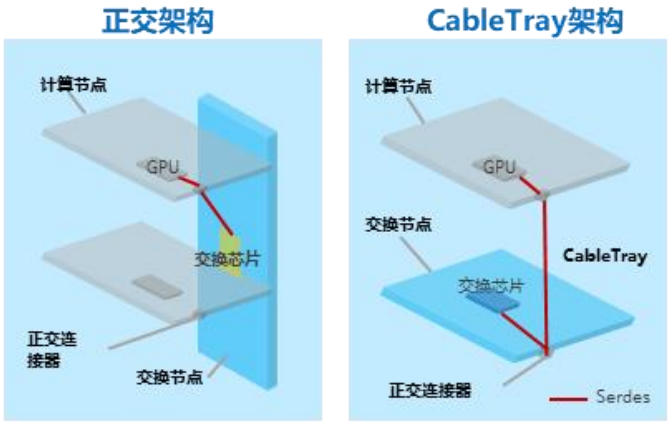


图 2-2 OEX 与 Cable Tray 方案对比

2.2.1.2 工程化验证与标准化进展

中兴通讯率先倡导并实践的 OEX 正交无背板互联交换架构，通过创新的物理布局优

化，实现了信号传输效率与散热性能的双重跃升。该架构凭借卓越的技术先进性与工程价值，已于 2025 年成功入选 ODCC“年度重大技术突破”案例。

秉持“开放解耦”的生态理念，中兴通讯积极推动国产 AI 算力底座的标准化进程，并全面开放 OEX 机械与电气接口规范，支持第三方计算及交换托盘的即插即用，有效降低了系统集成门槛，促进了产业链的协同创新。

此外，中兴通讯已于 2025 年 6 月在 ODCC 网络工作组成功立项《基于正交架构的超节点硬件系统》，旨在通过标准化建设加速国产 AI 基础设施生态的成熟与应用落地。

2.2.1.3 超节点液冷技术

超节点的高密度算力必然带来高能耗，因此供电与散热系统必须与算力架构协同设计。

随着半导体先进封装工艺的持续进步，CPU、GPU 及网络交换芯片的集成度显著提升，单芯片功耗和热流密度持续攀升。过去几年，英伟达主流 GPU 单芯片功耗已从 700 W 跃升至接近 1400 W。根据当前主要厂商的产品路线图，未来 2 至 3 年内单芯片功耗突破 2000 W 已成为高度可预期的趋势。

超节点作为未来智算中心基础设施的核心形态，通过更高密度的芯片集成和高效互联显著提升了计算性能与网络效率。然而这一演进也导致单机柜功耗的快速增长。以 NVIDIA 为例：2022 年 H100 超节点机柜功耗约 50 kW；2025 年 GB300 NVL72 机柜功耗已达 120 – 150 kW；预计 2027 年 Rubin Ultra NVL576 机柜功耗将达到约 600 kW，未来进一步向兆瓦级机柜演进。在此背景下，液冷散热技术已从可选方案转变为大规模 AI 基础设施的必选方案。

从技术成熟度和市场应用角度，当前液冷方案主要分为以下几类：

- **单相冷板式液冷**：这是目前应用最广泛、工程化最成熟的液冷技术，市场占有率超过 70% - 80%。其原理是通过导热冷板与芯片直接接触，利用单相冷却液对流带走热量。该方案结构相对简单、可靠性高、维护成本较低，能够有效支撑当前百千瓦级机柜的散热需求。

- **硅基微通道冷板技术**：硅基微通道冷板被视为承接未来极高热流密度芯片散热需求的重要方向。其核心创新在于在硅基材料上直接蚀刻微米级流道，使冷却液在极小尺度内紧贴热源表面流动，从而显著提升单位面积换热系数。相比传统铜/铝冷板，硅微通道冷板具有更低的界面热阻、更高的热流密度承受能力，特别适用于 HBM 堆叠、Chiplet 多芯片模块等热源高度集中的先进封装形态。目前该技术仍处于加速验证与小规模商用阶段，但已被英伟达（如：Vera Rubin）等多家厂商视为下一代高功率 GPU/ASIC 的关键散热路径。

- **两相冷板液冷**：在传统冷板基础上引入相变机制：冷却工质在冷板内部沸腾吸热，汽化蒸汽在冷凝段回流，实现高效热量迁移。该方案在较低流量条件下即可实现极高的散热效率，理论上更适合未来极端高功耗（>2000 W/芯片）场景。目前工程化程度正在快速提升，但面临工质选择、相变稳定性及系统可靠性等挑战。

- **浸没式液冷**：工程化程度较高的全液冷方案，将服务器整体浸没于绝缘冷却液（单相或两相）中，通过自然/强制对流直接带走热量。其优势在于散热能力强、结构简化、对高密度器件兼容性好，能够支撑百千瓦乃至兆瓦级机柜部署。近年来，随着氟化液等工质的成熟，单相/两相浸没式液冷在超大规模数据中心中的部署比

例持续上升。

当前阶段，单相冷板液冷（或单相冷板结合局部风冷）仍为主流方案，能够可靠满足百千瓦级机柜的散热需求。然而，随着芯片功耗和热流密度的持续提升，单相冷板正逐步接近其物理极限。未来 2 – 5 年内，硅基微通道冷板、两相冷板液冷及浸没式液冷等高性能方案将逐步成为主流，共同支撑兆瓦级 AI 工厂的热管理需求。

液冷技术的全面普及，不仅是散热能力的升级，更是智算基础设施向高能效、低碳化、可持续方向演进的必然趋势，标志着数据中心从“算力导向”向“能效导向”的结构性转型。数据中心运营商需提前规划电源、冷却基础设施及运维体系，以适应这一技术代际跃迁。

2.2.1.4 极端功率密度的供电方案

当前，数据中心机柜内主流配电方案已由早期的 12V 演进至 48V/54V。该方案通过提升电压等级显著降低了承载电流，在一定阶段内有效缓解了配电路径上的欧姆损耗，优化了材料成本与整体转换效率。然而，随着超节点功率密度的爆发式增长，48V 体系的物理局限性正日益凸显。

以额定功率 120 kW 的超节点机柜为例，若采用 54V 总线（Busbar）配电，其承载电流将高达约 2222 A。如此极端的电流强度不仅会产生巨大发热，还要求大幅增加铜质母排的截面积以控制压降。这不仅导致母线系统自重与成本激增，甚至在极端场景下需为供电路径配置专用液冷回路，严重侵占了 IT 设备的可用算力空间。

另外从经济效率考虑，电网侧到芯片核心的完整供电链路也涉及复杂的能量转换级联

（AC-DC、DC-DC 等）。

- **级联损耗：** 每一级转换通常伴随 2%–5% 的能量损失。累加效应导致市电到处理器核心的端到端转换效率显著下降，典型总效率损失维持在 5%–12% 区间。
- **OPEX 敏感度：** 电力成本通常占据数据中心运营支出（OPEX）的 30%–50%。在超节点时代，能源成本的权重被进一步放大。供电效率每提升 1%，对于大规模智算集群而言都意味着巨大的经济回报与可持续性价值。为突破功率壁垒，行业正加速转向 HVDC（High-Voltage Direct Current 高压直流） 配电架构。其核心逻辑是将机柜级或排级配电电压从传统的 48V/54V 提升至更高量级，目前行业主流演进方向包括 ±400V DC（等效 800V） 与 800V DC 直流系统。

HVDC（高压直流） 架构的核心优势：

- **电流强度指数级下降：** 在同等功率下，电流可降低 8–16 倍。这允许采用更精细、更轻便的配电组件，铜材用量可减少 40%–50%，为计算与冷却组件释放关键空间。
- **配电损耗大幅缩减：** 有效抑制传输热损，预计可提升整体端到端效率 3%–5%。
- **支撑兆瓦级部署：** 轻松承载从当前 100–150 kW 向 250 kW 至 1 MW+ 级机柜的演进需求。
- **架构扁平化：** 减少中间能量变换层级，从根本上缓解功率因数校正（PFC）与无功功率管理压力。

目前，全球头部科技巨头正通过开源社区与技术白皮书加速推进 HVDC 的商业化进程：

- **OCP Diablo 400 项目：** 由 Microsoft、Meta 与 Google 联合推动。该标准定义了 ±400V DC 解耦式电源机架（Sidecar），支持机柜功率从 100 kW 向 1 MW 的平滑扩展，并标准化了机械接口与安全电气要求，构建了跨厂商的兼容生态。
- **NVIDIA 800VDC 生态：** 英伟达正协同 CoreWeave、Oracle 等合作伙伴，布局 800V 直流电源架构，以支撑单机柜 1 MW 的超高密度计算环境。其发布的《面向下一代 AI 基础设施的 800VDC 架构》明确了当前至 2030 年的三阶段演进路径，为产业链提供了清晰的投资与技术路线指引。

尽管 HVDC（高压直流） 潜力巨大，但在大规模落地前仍需解决以下工程化命题：

- **安全规范与绝缘防护：** 需针对高压直流下的电弧抑制、绝缘击穿风险制定严苛的工业级标准，保障运维安全。
- **供应链成熟度：** 高功率密度 DC-DC 转换器、高压连接器及直流断路器等关键组件需进一步验证其长期可靠性并降低成本。
- **生态互操作性：** 跨厂商的物理接口、冗余协议及监控 telemetry 的统一仍需行业协作。
- **HVDC（高压直流）** 是数据中心供电体系迈向更高能效、更高密度的必然路径。智算中心运营商应前瞻性地布局高压电源基础设施与冷却体系，以应对下一代 AI

算力集群的能源挑战。

2.2.1.5 方案先进性

- **正交架构满足未来演进趋势：** 中兴通讯 OEX 超节点架构符合当前数据中心向高密度、高能效比和块化演进的技术方向，为下一代 AI 基础设施在算力密度、互联带宽与能效比方面提供了可落地的工程实现方案。该架构采用正交无背板互联设计，彻底摒弃传统 Cable Tray 线缆架构的物理限制，采用正交连接器与单级交换拓扑，实现计算节点与交换节点的垂直交叉互连，显著提升系统性能和扩展性。
- **大容量交换芯片+Link 协议，兼容多厂家 GPU：**大容量交换芯片与多样化 Link 协议的支持，使 OEX 架构能够兼容多厂家 GPU，满足不同应用场景需求。
- **组件化设计，灵活适配不同 GPU：**关键模块采用组件化设计，通过更换 UBB 模组可实现不同厂家 GPU 兼容。这种设计对各种协议交换芯片的快速适配开发，支持不同 GPU 超节点的快速部署和升级，最小化改动即可满足多样化需求。

综上所述，中兴通讯 OEX 超节点架构通过正交设计、大容量交换芯片和组件化设计，实现了高性能、高兼容性和高灵活性的统一。该架构面向未来数据中心建设，支持构建可扩展、高能效的 AI 推理与训练平台，能够满足业务智能化的演进需求。

2.2.2 Nebula Matrix 集群超节点

2.2.2.1 Matrix 集群超节点的演进路径

在单体超节点技术趋于成熟，并且实现 64 卡、128 卡等高密度集成后，为满足超大规模模型训练的极致需求，业界开始探索基于单体超节点构建更大规模的 Matrix 集群超节点。这一发展阶段主要形成了两条核心技术路线：

- **“电交换+光互联”技术路线**

该路线通过高性能电交换机实现跨机柜 GPU 间的互联。受铜缆传输距离限制，跨机柜场景需采用光纤介质完成互联。传统电交换机采用包级交换机制，在业务适配性上具备显著的灵活性，可满足多样化的互联需求；因涉及光电转换环节，相较于全光方案，在功耗控制与时延表现上可能面临一定挑战/损耗。

综合来看，电交换技术成熟度高、业务普适性强，凭借这些核心优势，它已成为当前业界构建集群超节点的主流技术选择。

- **“光交换+光互联”技术路线**

与传统电交换技术不同，光交换机采用光路交换机制，可支持任意两条光路间的直接映射。由于无需进行光电转换环节，光交换在时延优化与功耗降低方面具备天然优势。但受限于光路交换的技术特性，其难以实现传统电交换机那样的包级路由能力，在逻辑拓扑设计上受到较强的约束。

以 Google 为例，其利用 OCS 光交换机构建了 3D-Torus 拓扑，但此类拓扑对上层业务提出了特殊适配要求，需要针对业务调度、集合通信等核心环节进行定制化优化。

尽管近年来光交换相关技术实现了快速迭代，且有 Google 等互联网巨头完成了商业化落地，但整体生态仍不够完善，同时对系统整体构建的技术门槛要求更高，因此目前业界多数企业仍处于观望阶段。

当前主流集群超节点部署方案多采用电交换+光互联架构，因其技术成熟、生态完善、兼容性强。基于该技术方案，中兴通讯现有 Nebula X32 单体超节点可灵活扩展，构建形成 Nebula Matrix X256/800 集群超节点；面向未来，依托更高密度的 Nebula X128 单体超节点，更可进一步扩展至 Nebula Matrix X8192/16384 超大规模集群，充分满足超大规模模型训练的算力需求。

与此同时，中兴通讯并未止步于此，而是积极探索光交换与电交换的互补协同，旨在融合光传输的高效与电交换的灵活，以支持未来超大规模集群的可扩展性需求。

2.2.2.2 Scale-Up/ Scale-Out 融合设计

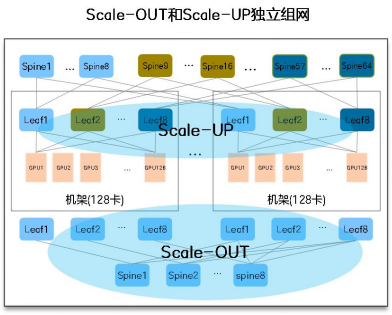
在追求构建大规模 Matrix 集群超节点的同时，必须思考一个核心问题：Matrix 超节点的物理规模边界与收敛比设计，需在性能与成本间寻求平衡。

AI 智算业务场景下 GPU 间的高性能互联网络，根据承载业务的不同，通常分为 Scale-Up（纵向扩展）和 Scale-Out（横向扩展） 网络。

- **Scale-Up 网络**：承载 AI 智算中对网络性能要求极高的张量并行和专家并行等业务通信流量，属于 HBD（高带宽域）。
- **Scale-Out 网络**：承载数据并行和流水并行等对网络性能要求相对低一些的业务通信流量。

随着模型参数规模的增加，张量并行和专家并行规模随之扩大，对超节点内 HBD 域规模的需求也越来越大，这必然超越单个机柜的物理极限。因此，如何构建大规模的集群超节点网络，既需要满足超节点内部 GPU 间对于网络的高性能要求，又要支持灵活的规模扩展，成为了一个需要平衡性能、成本等多方面因素的设计课题。

模型测算显示，扩大 HBD 域对于模型性能会有一定的收益，但 HBD 域到了一定规模以后，收益就会逐渐趋缓。在未来 2-3 年内模型参数普遍达到 10 万亿量级时，超节点内部的 HBD 域可以局限在机架内部，机柜内和机柜间互联带宽采用一定的收敛比，达到应用与工程实现之间最佳的平衡。针对未来模型演进可能会存在更大的互联域需求，机柜间 GPU 也可以带宽无收敛的互联设计以构建更大的 HBD 域。相对于传统 8 卡 GPU 服务器，在跨越多个单体超节点的集群超节点内部，Scale-Up 网络和 Scale-Out 网络的边界日益模糊。为了既可以满足集群超节点对于 HBD 域灵活的规模需求，又可以满足集群超节点间的互联需求，构建一张 Scale-Up 和 Scale-Out 融合的超节点互联网络，成为大势所趋。这种融合架构不仅能保障集群超节点部署和扩容的平滑性，相比独立组网模式，更能显著降低 TCO（总拥有成本）。



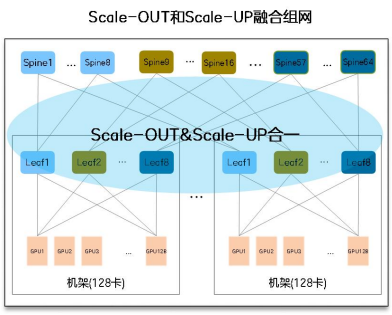


图 2-3 Scale-Up 和 Scale-Out 融合和独立组网对比

基于此，在机架间通过光互联+交换芯片+高性能网络协议构建一张更大规模、更高性能的网络，实现 Scale-Up 和 Scale-Out 网络的融合，统一承载 GPU 间的所有 AI 计算通信业务，构建超级算力的集群超节点已成为业界迫切的诉求。正交架构解决了“算力如何做得更密”的问题，而光互联+交换芯片+高性能网络协议则应对了“算力如何连得更广”的挑战。这种从底层物理架构到上层系统拓扑的全方位思考，使其在激烈的市场竞争中，展现出独特的技术韧性和发展潜力。

3 以超节点为核心：打造 AI 工厂

3.1 核心理念：从项目到工厂的范式转变

AI 工厂，是以超节点为核心，集成全栈软硬件协同能力，实现从数据输入到智能输出（Token）的标准化、规模化、自动化生产系统。

传统以项目为中心的 AI 开发模式，往往受困于基础设施孤岛、资源利用率低效及部署周期漫长等瓶颈。AI 工厂范式旨在彻底颠覆这一现状，其核心在于将 AI 能力建设从传统的“手工作坊”升级为标准化的“现代化流水线”。

AI 工厂通过全栈软硬协同优化，将数据输入高效转化为 Token，正如传统工厂将原材料精炼为高价值制成品。构建 AI 工厂，其战略意义远不止于缓解当下的算力瓶颈，更在于数字时代对技术主权与敏捷性的重新定义。

3.2 构建路径

AI 工厂是一个以“超节点”为核心的生产力平台，集灵活性、可扩展性与高可靠性于一体。客户可以依据自身业务场景，像搭积木般自由定义“工厂”的规模、性能与成本模型。这种从底层芯片到上层软件的全栈协同与深度定制能力，正是算力竞争下半场的决胜焦点。

要实现以超节点为核心的 AI 工厂，关键在于超越传统的硬件堆叠思维，将分散的算力资源系统性地转化为可高效输出的“智能生产力”。具体可通过以下三个层面展开：

首先，在物理层，重塑底层算力单元，构建高性能基础模组。 利用先进的光互联与高性能交换技术，突破传统机柜的物理边界，将成千上万个 GPU 互联为一个统一的高带宽、低延迟网络域，形成如同超级芯片般的“集群超节点”。这彻底解决了大规模并行训练中的通信瓶颈，为万亿参数模型的运行提供了极致性能的物理底座。

其次，在系统层，实现软硬全栈垂直优化，激活系统协同效能。 AI 工厂不仅仅是硬件的集合，更强调软件栈对硬件资源的深度调度与优化。通过定制化的集群操作系统，实现对超节点内异构算力、分布式内存及复杂网络拓扑的统一编排与智能调度。这种软硬一体的设计，能够最大化资源利用率，并通过重叠计算与通信来隐藏延迟，确保每一份算力都转化为实际产出。

最后，在架构层，采用模块化灵活组装，实现业务敏捷适配。 基于超节点的标准化与解耦设计，企业可以根据业务规模和模型需求，灵活调整工厂的产能。同时，引入算力仿真平台构建“数字孪生”，在虚拟环境中预先推演不同配置下的性能与成本，精准定位最优方案。这种“仿真指导组装”的模式，使 AI 工厂能灵活应对多样化需求：一

方面通过仿真规避试错风险，精准规划；另一方面通过弹性扩展快速响应业务变化。最终，它得以演进为一个能持续自我优化、赋能业务的现代化 AI 生产中心。

3.2.1 大规模集群网络：突破集群扩展的规模限制

为了突破单点瓶颈、整合分散资源，同时匹配指数级增长的算力需求，大规模智算集群的建设通过“两步走”的方式来突破物理边界，实现算力的极致扩展：

第一步，通过 Scale-Out 网络实现单数据中心内的集群构建。

以 Nebula 单体超节点为基本单元，利用高性能 Scale-Out 网络进行横向扩展，搭建基础的智算资源池。针对万亿参数级超大模型的极致性能需求，可进一步利用光互联技术将多台单体超节点进行逻辑整合，形成 Nebula Matrix 集群超节点，并在此基础上叠加 Scale-Out 网络，实现算力在数据中心（DC）内部的高密度、高性能聚合，打造单数据中心智算集群。

第二步，通过 Scale-Across 网络实现跨数据中心的广域算力互联。

随着算力需求的持续增长，单数据中心受空间、供电、散热等物理条件制约，算力架构亟需突破建筑边界。在完成单数据中心集群构建的基础上，利用 Scale-Across 网络配合长距光互联技术（如 OTN），将地理位置分散的多个智算数据中心实现全域互联。同时引入独立的算力网关设备，凭借其大缓存特性及快速拥塞控制反馈机制，解决长距传输的时延与拥塞挑战。这一步标志着智算集群从“单点算力极致”走向“广域算力协同”，为 AI 工厂提供了近乎无限的算力扩展空间。

综上所述，AI 工厂以单体超节点和集群超节点为起点，通过 Scale-Out 网络横向扩

展构建单 DC 集群，再通过 Scale-Across 网络跨越数据中心边界，实现多 DC 算力的广域聚合，形成大规模跨域智算集群。这一端到端的全栈互联演进路径，可全方位突破集群扩展的规模限制，助力打造具备真正无限扩展能力的 AI 工厂的算力底座。

3.2.2 软件栈：超节点的“操作系统”

3.2.2.1 集群管理

超节点的强大硬件能力，需要通过一套深度协同、全栈优化的软件系统才能被充分抽象、调度与释放。这套软件栈扮演着超节点“操作系统”的角色，其核心作用在于将离散的高性能芯片、异构内存与高速网络等物理资源，转化为高效、稳定、易用的一体化算力服务。其主要价值体现在以下六个层面：

- **统一虚拟化资源池与智能编排**：软件层首先对超节点内所有硬件资源进行抽象与池化，形成统一的虚拟化算力、内存与存储资源池。通过智能资源调度器，根据 AI 训练、推理等不同工作负载的需求，动态、弹性地分配和隔离资源，实现多任务、多租户环境下的共享与安全隔离。这包括对异构内存（如 HBM、DDR）的统一纳管与池化，使应用程序能够超越单卡物理显存限制，透明地使用聚合后的分布式大内存空间。

- **极致通信优化与拓扑感知**：针对超节点内部高带宽域及跨节点互联的复杂网络拓扑，软件栈提供深度优化的通信库（如集合通信库）和运行时系统。这些组件具备拓扑感知能力，能够自动识别最优的数据传输路径，避免网络拥塞，最大化利用 TB 级互联带宽。同时，通过实现计算与通信的高效重叠、梯度压缩、异步化

等技术，将通信开销隐藏于计算过程之中，从而将系统整体效率推向理论峰值，使大规模分布式训练的线性加速比接近理想值。

- **异构计算统一调度与编译器优化**：支持 CPU/GPU/DSA 等异构单元统一调度；高级编译器自动切分计算图，进行算子融合、内核生成、流水线调度，提升单卡效率与跨芯片协同。
- **全栈可观测性与智能运维**：构建芯片→节点→集群多级监控体系，实时可视化功耗、温度、性能等指标；结合 AI 运维，实现故障预测、根因分析，定位时间从小时级缩短至分钟级；支持检查点续训、服务无缝迁移，保障业务连续性。
- **高可靠冗余机制**：在超节点层面，超大规模的集群故障概率显著增高。芯片、机柜数量的剧增，使得硬件故障（如芯片损坏、链路中断）成为常态。且大模型训练中，故障爆炸半径会随部署规模扩大而变大，任意硬件故障都可能导致整个训练任务不可用，冗余节点可避免因此类问题引发的业务中断。冗余节点搭配对应的故障切换机制，能大幅缩短故障恢复时间，保障训练任务按计划推进，避免算力和时间成本的浪费。软件调度系统可引入冗余算力节点避免单点故障导致的任务中断、降低故障恢复的时间与算力成本、保障大规模并行计算的性能稳定性、支撑集群的弹性扩展与灵活调度。
- **“算力-电力”协同的绿色调度**：引入“算电协同”策略，结合任务优先级、功耗模型与实时电价，动态调整调度与频率，在保障 SLA 前提下平滑功率波动，降低能耗与运营成本。

3.2.2.2 集群调度

集群部署的大规模 LLM 推理服务，需要兼顾资源成本效率与用户体验，在吞吐量和实时响应性之间取得权衡。帕累托最优作为多目标优化领域的核心理论，在评估 LLM 推理系统性能时，其帕累托前沿曲线能清晰展现吞吐与时延的权衡关系，为集群算力调度提供了核心理论依据。中兴通讯的 Turbo 集群算力调度服务，能够实时感知集群负载和业务流量波动，动态调整资源配比，通过智能调度实现帕累托寻优。

在集群环境部署 LLM 推理服务时，主流架构可分为 PD（Prefill/Decode）聚合以及 PD 分离两大类。本质上是围绕推理的两个核心阶段（Prefill 和 Decode），在集群资源分配和任务调度逻辑上的不同实现。

- **PD 聚合架构**：Prefill 和 Decode 共享相同的 GPU 资源和并行策略，适合追求高吞吐的批量处理场景，如离线推理、批量内容生成。
- **PD 分离架构**：将推理过程中 Prefill 与 Decode 两个阶段解耦，为每个阶段独立分配 GPU 资源，分别设置并行策略。通过规避 Prefill 和 Decode 两阶段间的相互干扰，能够满足严苛 SLA 约束下的服务需求，更适合时延敏感的场景。

针对当前主流框架在国产算力卡上的适配挑战，中兴通讯联合主流 GPU 厂家通过框架调优、算子优化、通讯加速、智能调度等技术手段，实现了国产高性能算力卡性能提升。

- **框架层面**：采用 Chunk Prefill、计算与通信双流重叠、MTP（Multi-Token Prediction）等优化方案，大幅提升推理效率。Chunk Prefill 通过分块预填充技术，通过将长输入序列分割成多个较小的子块，逐块进行处理，降低了显存峰值

使用量，提升并发度；MTP 策略在 Decode 阶段将单个 token 的生成，转变成多 token 的生成，从而提升推理的性能；EPLB（Expert Load Balancing with Redundancy）优化策略通过复制热点专家副本来创建冗余专家，实现专家负载均衡优化，解决不同 EP 专家的负载不同导致快慢卡问题，从而提升推理性能。

- **算子层面**：通过 MLP 的 UP 和 Gate 矩阵合并，只进行一次矩阵乘，提升计算密集程度，在推理的 Decode 阶段的性能有显著收益。
- **通信层面**：支持 DeepEP 功能，优化 MoE 的分发（Dispatch）与合并（Combine）操作，支持 IBGDA（Intelligent Bandwidth-Guided Data Aggregation）功能，提升机间通信效率、降低延迟、提升吞吐。
- **服务层面**： 通过推理服务多副本部署、丰富的调度机制以及自动扩缩容，实现业务灵活调度和高可靠性。

3.2.2.3 算力仿真平台

在 AI 大模型训练与推理需求爆发式增长的背景下，算力资源的高效规划、硬件选型的科学决策以及并行策略的优化配置，已成为企业降低成本、提升研发效率的核心诉求。传统依赖实际硬件部署测试的方式，不仅存在周期长、成本高、资源浪费的问题，还难以全面覆盖不同模型、硬件、超参组合下的性能表现，无法快速锁定最优技术方案。

在此背景下，算力仿真平台应运而生。它如同为 AI 基础设施构建了一个“数字孪生”体，通过数字化建模，在虚拟环境中精准复刻算力系统的运行逻辑，实现对硬件选型、

并行策略和系统配置的性能预测与方案验证。这不仅是降低研发成本的工具，更是提升决策科学性、加速 AI 创新的关键引擎。

算力仿真平台基于硬件参数（包括显存、通信带宽等），超参（如 GBS 等），模型结构等（如 Attention 头数等），并结合算子和通信带宽实测数据，对不同规模、不同硬件配置下的训练/推理场景下的端到端性能进行建模，并结合优化特性，对关键性能指标进行遍历和评估。通过持续分析与迭代，选择端到端性能最优的并行策略和方案，输出端到端性能数据。

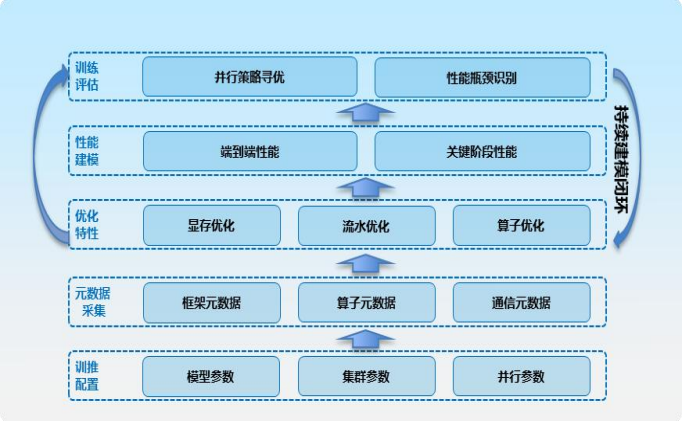


图 3-1 算力仿真平台

以 Qwen3-235B，某国内 GPU 卡为例，分析不同超节点形态的训练性能。

（1）算子建模：分析 QKV_Linear、Flash-Attention、O_Linear、Gating_Linear、MoE MMA 五类算子的算力强度，发现只有 MoE MMA（Matrix Multiply-Accumulate）算子强度随 EP 增加而变化，其他模型算子都不变。MoE MMA 算子强度随 EP 增加而变化，但随着 EP 增大增长变缓而逐步逼近硬件上限。

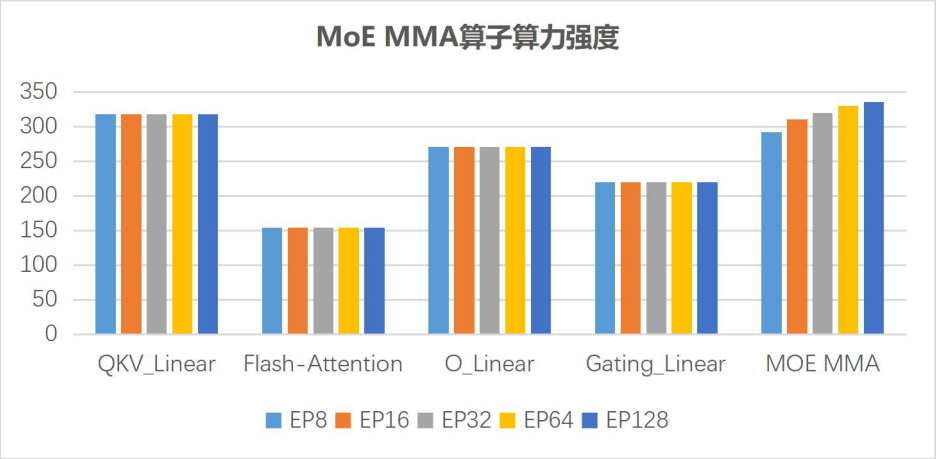


图 3-2 MoE MMA 算子算力强度

（2）通信建模：分析不同超节点形态下计算和通信耗时影响。分析发现，在 2000 卡的集群规模下，随着超节点规模增大，最优切分的性能逐渐增加，收益主要来源于 MoE 算子性能提升；该收益存在边际效应：当超节点规模可达 64 卡及以上时，性能基本趋于一致。

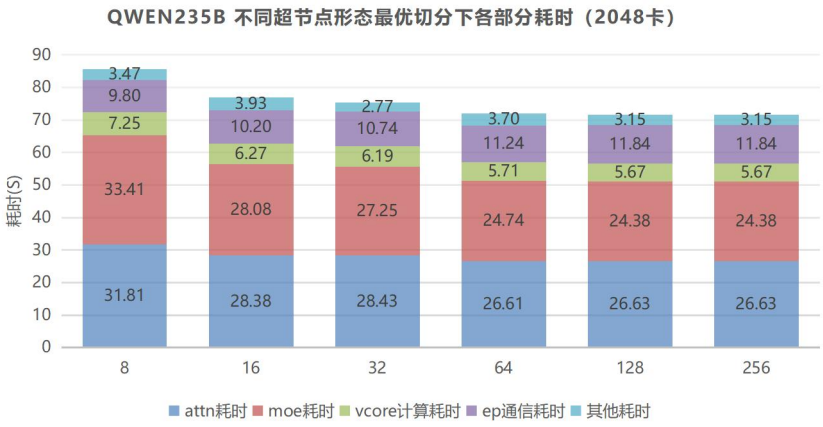


图 3-3 Qwen3-235B 不同超节点形态最优切分下各部分耗时

（3）TFLOPS 性能评估：在不同集群规模下，探究超节点形态对模型性能的影响。分析发现，不同超节点形态最优切分不同，随着规模增大，最优 EP 也逐渐增大；64、128、256 超节点性能基本一致，在 2K 卡规模下，256 卡超节点相比 32、16 卡超节点高 4%，相比 8 卡服务器高 15%；

通过算力仿真平台进行业务模拟和性能推演，分析表明千亿参数规模的大模型（Qwen3-235B）训练场景，在同样规模下，随着超节点形态增大，单卡训练性能逐渐增加，收益主要来源于 MoE 算子性能提升；但收益存在边际效应。

同样分析发现千亿参数规模大模型（DeepSeek-671B）在推理场景，在同样规模下，随着超节点形态的增加，HBD 域扩大，All-to-All 通信时间的减小，单卡推理性能逐渐增加；但收益存在边际效应。

算力仿真平台通过建模与预测，显著降低硬件选型与并行策略设计的试错成本，已成为 AI 系统设计流程中的关键辅助工具。中兴通讯已将其深度集成至研发闭环，支持从架构设计到部署优化的全链路决策。

对于客户而言，选择具备强大仿真能力的合作伙伴，意味着能在复杂的算力迷宫中找到通往高效、低成本 AI 创新的捷径。

3.3 AI 工厂的核心优势与商业价值

通过部署经过全栈验证的 AI 工厂，企业将在战略高度构建起多维度的竞争优势，并在以下四个层面实现商业价值的深度释放：

- **缩短业务上线周期**：依托经过预验证的软硬件协同配置，减少现场集成与功能调

试的时间。通过标准化的部署流程，提高资源利用效率，加快应用从开发到生产环境的部署速度，从而缩短项目交付周期。

- **支持架构平滑演进**：基于模块化与解耦设计，实现计算、存储与网络资源的独立弹性伸缩。该架构能够适应业务规模的增长，支持算力容量的线性扩展，避免因业务升级而频繁重构基础设施，延长硬件资产的使用生命周期。
- **优化总体拥有成本（TCO）**：通过架构设计优化提升资源密度与利用率，配合自动化的运维体系，降低对人工干预的依赖。在保障计算性能指标的前提下，有效控制资本支出（CAPEX）与运营支出（OPEX），实现性能与成本的平衡。
- **降低系统集成风险**：采用经过大规模实践验证的架构设计及经过兼容性认证的组件列表，减少异构集成带来的不确定性。规避因硬件选型差异或接口不匹配导致的兼容性问题，保障系统运行的稳定性与业务的连续性。

4 中兴通讯：全栈协同的 AI 基础设施构建者

构建以超节点为核心的 AI 工厂，是一场涉及底层芯片、整机、集群与软件的复杂系统工程。中兴通讯将通信领域的系统工程方法、大规模组网技术及高可靠性设计经验应用于 AI 基础设施建设，重点解决智算中心在互联带宽、系统稳定性及工程交付方面的技术挑战。

作为全栈协同的 AI 基础设施构建者，中兴通讯的核心能力不仅体现在技术的深度整合，更体现在对开放生态的坚定承诺，具体包括：

图 4-1　中兴通讯：全栈协同的 AI 基础设施构建者

- **芯片与基础算法**

中兴通讯具备 CPU、DPU 及全系列交换芯片的自主设计能力。依托自研交换芯片与 SerDes 技术的深厚积累，我们将通信领域的高性能互联机制应用于智算场景，解决传统集群中的通信瓶颈问题，提供支持 Scale-Up 与 Scale-Out 融合的开放网络方案。同时，基于底层算法优化能力，通过对国产 GPU 架构的算子调优及垂直领域模型适配，实现算法与硬件的匹配，提升系统有效算力。

- **复杂架构设计能力**

我们将通信设备在长期高可靠、高并发、低时延运行中积累的系统设计经验应用于 AI 基础设施。基于在硬件结构、散热工程及 EDA 等领域的技术储备，设计了正交无背板互联架构，实现无外部线缆的高密度超节点，提升信号完整性与散热效率。依托覆盖芯片、整机、集群、软件及数据中心的跨领域研发体系，实现软硬件的协同设计。此外，引入电信级运维标准，构建具备故障自愈与性能自优功能的智能运维体系，保障大规模 AI 集群的连续运行。

- **全球工程交付能力**

  依托覆盖全球 160 个国家的服务网络与本地化团队，建立了针对超大规模、复杂环境的工程交付体系。通过标准化的模块化设计与自动化运维平台，将 AI 集群建设转化为可复制、流程化的交付作业，确保客户 AI 工厂按计划上线并保持高效运营。

- **标准引领与开源开放**

  中兴通讯致力于构建开放解耦的国产 AI 生态。通过开放 OEX 架构规范，支持第三方算力组件的兼容接入。同时，积极参与国内互联标准制定，开源 Co-Sight 智能体通信协议，推动 AI 工厂软件生态的标准化与共建。

中兴通讯通过底层硬件至顶层软件的技术整合，形成涵盖芯片、整机、网络与软件的全栈解决方案，支持智算基础设施的构建与部署。

展望未来，Token 经济学已成为衡量智算基础设施竞争力的核心理论框架。其内涵不再局限于物理算力的简单堆叠，而是聚焦于智能产出的实际效能与综合成本。中兴通讯 AI 工厂与超节点架构的设计逻辑，正是遵循 Token 经济学原理，通过架构重构与全栈协同，推动价值导向从“每秒浮点运算次数（FLOPS）”向“每瓦 Token 数”的关键转变，从而在激烈的产业竞争中确立成本与效率优势。

面向智能化浪潮，中兴通讯将继续秉持开放解耦的理念，提供涵盖硬件、软件及交付的全栈完整方案。中兴通讯将携手全球产业伙伴，共同构建面向未来的开放智算生态，推动 AI 技术的普及化与标准化，赋能千行百业。

5 缩略语表

缩略语	英文全称	中文全称
AI	Artificial Intelligence	人工智能
ASIC	Application-Specific Integrated Circuit	专用集成电路
CAPEX	Capital Expenditure	资本性支出
CXL	Compute Express Link	计算快速链接
DSA	Domain-Specific Architecture	领域专用架构
EPLB	Expert Load Balancing with Redundancy	基于冗余的专家负载均衡
EP	Expert Parallelism	专家并行
HBD	High-Bandwidth Domain	高带宽域
HBM	High Bandwidth Memory	高带宽内存
IBGDA	Intelligent Bandwidth-Guided Data Aggregation	智能带宽引导数据聚合
IPC	Instructions Per Cycle	每周期指令数
LLM	Large Language Model	大语言模型
MoE	Mixture of Experts	混合专家模型
MTP	Multi-Token Prediction	多 Token 预测
MMA	Matrix Multiply-Accumulate	矩阵乘加内核
MFU	Model FLOPs Utilization	模型算力利用率
NIC	Network Interface Card	网络接口卡
NVMe	NVM Express	非易失性内存主机控制器接口规范
OCS	Optical Circuit Switch	光交换机
OEX	Orthogonal Electrical eXchange	正交无背板互联交换
OISA	Omni-directional Intelligent Sensing Express Architecture	全向智感互联
OPEX	Operating Expense	运营支出
OTN	Optical Transport Network	光传送网
PCIe	Peripheral Component Interconnect Express	高速串行计算机扩展总线标准
PD	Prefill & Decode	预填充与解码
PHY	Physical Layer	物理层

RoCE	RDMA over Converged Ethernet	以太网上的 RDMA
Scale-Across	Scale-Across Data Center	跨数据中心扩展
Scale-Out	Scale-Out	横向扩展
Scale-Up	Scale-Up	纵向扩展
SerDes	Serializer/Deserializer	串行器/解串器
TCO	Total Cost of Ownership	总体拥有成本
TP	Tensor Parallelism	张量并行
TTFT	Time To First Token	首字延迟
UALink	Ultra Accelerator Link	超级加速器链路
UEC	Ultra Ethernet Consortium	超级以太网联盟

6 参考文献

[1] 中兴通讯. 面向高带宽域的 Scale-Up 算力高速互联技术[J]. 《中兴通讯技术》, 2024, 30(3): 1 - 8.

[2] “面向 AI 大模型训练的高性能网络”，《中兴通讯技术（简讯）》2024 年第 3 期

[3] NVIDIA. 800 VDC architecture for next-generation AI infrastructure[R/OL]. (2025) [2025-11-15]. https://developer.nvidia.com/blog/building-the-800-vdc-ecosystem-for-efficient-scalable-ai-factories/.

[4] Open Compute Project. Diablo 400 project: rack and power 0.5.2 [S/OL]. 2024 [2025-11-15]. https://www.opencompute.org/.

[5] 开放计算技术委员会. 全液冷冷板系统参考设计及验证白皮书[R/OL]. (2024) [2025-11-15].

[6] 中兴通讯. 中兴通讯液冷技术白皮书[R/OL]. 2025 [2025-10-15] .

[7] 中国信息通信研究院. 算力中心冷板式液冷发展研究报告[R/OL]. (2025) [2025-11-15] .

[8] 中兴通讯. 基于正交架构的超节点系统[C/OL] // 2025 阿里云栖大会. 杭州, 2025 [2025-11-15].