超节点：从“堆卡”走向系统级协同 20260902

国产超节点 2026 年迎爆发拐点，Scale-up 驱动交换与液冷需求激增

摘要

- 超节点通过 NVLink/UALLink 协议实现 TB 级带宽与内存统一编址，将网络跳转降至 1 次，消除万亿参数模型在 MoE 架构下的通信瓶颈。
- 超节点是推理场景最优解，通过系统工程能力弥补单卡算力差距；2026 年为国产化从 0 到 1 突破元年，华为、摩尔线程、沐曦等厂商将集中发力。
- Scale-up 环节成为核心增量，其交换设备需求密度远超 Scale-out：GPU 与交换芯片配比从 12:1（Scale-out）提升至 4:1（Scale-up）。
- 光互联向 NPO/CPO 演进，硅光技术带动 PIC 上游保偏光纤、MPO 连接器、FAU 等无源器件用量显著弹性增长。
- 短距互联由无源铜缆向 AEC（有源铜缆）升级，或采用铜背板与正交连接方案：单机柜功耗突破 100kW，驱动 2026 年液冷技术大规模落地。

Q&A

什么是超节点架构，其核心特征与传统服务器集群相比有何不同？

超节点是一种通过高带宽、低时延技术将大量 GPU 互联在一起的系统架构。与传统通常基于 8 卡服务器的集群相比，超节点并非简单地增加 GPU 数量，而是通过增强多卡间的内部互联，降低传统架构中的通信墙与通信时延。其核心在于，在不改变单卡算力的前提下，通过系统化的协同能力提升整体算力的利用效率。超节点具备三个基础特征：大带宽、低时延和内存统一编址。 首先，在带宽方面，传统架构下卡间互联主要依赖 PCIe 或以太网，跨服务器互联带宽通常在 200-400 GB/s，时延达到几十微秒，面对参数日益增大的模型和频繁的并行计算交换，容易导致通信阻塞。超节点采用如 NVLink 或第三方 UAL Link 等更高效的互联协议，可支持 TB 级别的互联带宽，打破了传统架构的限制。 其次，在时延方面，传统 8 卡服务器间的连接需要 1 到 3 次网络跳转，信息需经由 GPU、网卡、交换网络再到目标网卡和 GPU，过程中的多次跳转会产生额外的通信时延和损耗。 架构实现了 GPU 之间的直接互联，网络跳转次数减少至一次，显著降低了 延。 最后，内存统一编址解决了大模型分割存储于不同 GPU 时的数据同 问题。传统架构下，一个 GPU 若要调用另一个 GPU 内存中的信息，需要经



网络进行远程访问、复制和粘贴操作。超节点通过构建统一的内存池，使得任何一个 GPU 都可以直接访问共享内存池中的信息，无需进行跨网络的复制粘贴，从而提升了通信效率。

在当前行业背景下，推动超节点成为核心基础设施形态的主要驱动力是什么？

超节点之所以成为核心基础设施，主要源于三点驱动力。 第一，模型规模的持续扩张与架构演进。当前模型参数已进入万亿级别，未来预计将发展至几十万亿甚至百万亿参数。模型规模的增大导致模型切分更为普遍，GPU 间需要同步的信息和次数随之增多，对通信时延的要求愈发严苛。同时，主流模型架构正向 MoE（Mixture of Experts）稀疏模式演进，专家数量的增加也带来了专家并行和动态调度等新的通信需求，这些都凸显了超节点在降低通信时延方面的优势。 第二，超节点架构高度适配推理场景。推理应用天然对时延和成本敏感。超节点方案通过优化网络效率来提升算力利用率，而无需投入更多资金购买额外的计算卡，因此被认为是推理场景下的最优解。 第三，中国的特定产业环境。在先进制程受限的背景下，中国在网络设备与系统架构方面具备较强能力。通过发展超节点，可以利用系统化的工程能力来弥补单卡性能与海外的差距，这使得超节点在国内具有显著的部署必要性。

从发展节奏来看，中国国产超节点的演进历程和未来渗透率趋势如何？

国产超节点的发展呈现出清晰的加速态势。2025 年被视为国内超节点初步探索的元年，该年以来，国内多家芯片厂商陆续推出了相关产品，其中华为在 WAI 2025（世界人工智能大会）上展示的超节点产品获得了广泛关注。进入 2026 年，国产超节点的发展进入加速期，在 2026 年的人工智能大会上，新增了如摩尔线程、沐曦等更多厂商展出其超节点产品。这标志着 2026 年是国产超节点实现从 0 到 1 突破的关键时间点。尽管今年（2026 年）的渗透率可能尚不高，但预计到明年（2027 年）及未来，超节点的市场渗透率将实现快速提升。

超节点架构的引入对数据中心组网中的 Scale-up 和 Scale-out 部分分别产生了哪些影响，其核心增量体现在哪里？

在数据中心组网中，Scale-out 主要负责跨节点连接，通常采用 IB 网络或开放以太网协议；而 Scale-up 则负责节点内部多颗 GPU 的直连，多采用如 NVLink、UAL Link 或华为的 HCCS 等定制化协议。超节点产品的诞生，其核心在于显著增加了芯片间直连的 Scale-up 部分。因此，探究超节点带来的核心增量，主要体现在 Scale-up 环节。但需要明确的是，Scale-up 的增强并非孤立的，它与 Scale-out 紧密相关，共同构成了完整的组网体系。

在AI基础设施中，如何界定“scale-up”与“scale-out”的概念，以及“scale-up”环节为产业链带来了哪些核心增量？

“Scale-up”与“scale-out”并非简单以机柜内外进行划分，而是指节点内与节点外的互联。一个超节点（super node）内部的互联，例如一个包含384张或数千张加速卡的计算集群，即为“scale-up”。这种互联可能跨越多个机柜，但仍属于同一个超节点内部。因此，“scale-up”是超节点方案的核心增量环节。从硬件视角看，“scale-up”环节主要拉动了通信和散热领域的增量需求，具体体现在以下四个方面： 一、光模块：在“scale-up”内部，为满足高密度集成要求，新增了大量光互联需求，主要采用NPO和CPO方案。其中，NPO方案因可复用现有可插拔光模块的供应链，具有良率高、易于维护和生态开放的优势，预计将更快落地。NPO和CPO方案的核心增量来自硅光子集成电路，因为这两种方案基本都采用硅光技术。PIC的上游有源和无源器件用量也随之增加，例如保偏光纤、MPO连接器、FAU等无源环节的用量弹性显著高于传统可插拔方案。 二、铜连接：在短距离互联场景中，铜连接因其在成本、功耗、时延及可靠性方面的优势而得到应用。“scale-up”增加了短距电互联的需求，具体方案与厂商设计紧密相关，主要包括三种形式：首先是DAC（无源铜缆）和AEC（有源铜缆），目前行业趋势正从无源向有源方案升级，AEC通过增加集成的Retimer芯片来补偿高速信号，从而在更高功率和更长链路上拓展铜互联的覆盖范围。其次是铜背板方案，例如NVIDIA GB200 NVL72就在机柜背部采用了铜背板进行互联。最后是正交连接方案，例如升腾910C超节点就采用了此方案。 三、交换机与交换芯片：“scale-up”网络是在传统“scale-out”网络之上新增的互联层次，构成了纯粹的增量市场。与“scale-out”相比，“scale-up”域内需要配置更高密度的交换设备。以两台GB200 NVL72互联为例，在“scale-up”域内，需要18个交换托盘，对应36颗NVSwitch芯片，此时GPU与交换机的配比为8:1，GPU与交换芯片的配比为4:1。而在“scale-out”域内，互联需要4台Leaf交换机和2台Spine交换机，合计12台400G端口交换机（每台对应一颗交换芯片），此时GPU与交换机及交换芯片的配比均为12:1。数据显示，“scale-up”环节的交换设备需求量显著高于“scale-out”环节。 四、液冷：超节点方案的特点是单柜集成度和功率密度显著提高，使得液冷成为其重要的伴生技术。在超节点方案出现前，无论国内还是海外，液冷并非数据中心的标配。海外市场自NVIDIA NVL36产品推出后，液冷开始成为标配；同样，国内液冷的大规模应用也将伴随超节点方案的落地而展开。目前，超节点单机柜的峰值功耗已普遍突破100千瓦。因此，2026年作为国内超节点的落地元年，也将成为国内液冷的落地元年，为相关厂商带来从0到1的市场机遇。