https://crad.ict.ac.cn
63(8): 1905−1913, 2026

观 点

国外超大规模智算集群体系结构发展的 10 个趋势

苏金树[1] 夏雨生[1] 宋丛溪[1] 黄见欣[1] 董德尊[2] 徐 扬[3] 陈 果[4]
[1]（军事科学院 北京 100091）
[2]（国防科技大学计算机学院 长沙 410073）
[3]（复旦大学计算与智能创新学院 上海 200438）
[4]（湖南大学计算机学院 长沙 410082）
（birchsu@139.com）

Ten Development Trends of Overseas Ultra-Large-Scale Intelligent Computing Cluster Architecture

Su Jinshu[1], Xia Yusheng[1], Song Congxi[1], Huang Jianxin[1], Dong Dezun[2], Xu Yang[3], and Chen Guo[4]
[1]（*Academy of Military Sciences*, *Beijing* 100091）
[2]（*College of Computer Science and Technology*, *National University of Defense Technology*, *Changsha* 410073）
[3]（*College of Computer Science and Artificial Intelligence*, *Fudan University*, *Shanghai* 200438）
[4]（*College of Computer Science and Electronic Engineering*, *Hunan University*, *Changsha* 410082）

Abstract Since 2012, the growth rate of intelligent computing has outpaced Moore's Law, achieving an order-of-magnitude increase per year, which poses unprecedented challenges to the architecture of computer. Ultra-large-scale intelligent computing clusters have been proposed as national new infrastructure in "Report on the Work of the State Council" in 2026. Considering the characteristics of human intelligence, i.e "small computing power, large network", the enabling technologies of optical transmission and switching with lower power consumption, and new architectures and mechanisms driven by technological innovation, we present ten preliminary judgments on overseas intelligent computing clusters development trends from four aspects: cluster architecture, computing nodes, network technology, and technical ecosystem. This paper aims to offer insights and contribute to the national implementation of ultra-large-scale intelligent computing clusters as new infrastructure.

Key words ultra-large-scale intelligent computing cluster; system architecture evolution; computing node; high-speed networking

摘 要 2012 年以来，智能计算增长速度已超越摩尔定律，达到年增 1 个数量级，对计算机体系结构提出全新挑战。2026 年国务院《政府工作报告》提出实施超大规模智算集群等新基建工程。综合考量人类智能“小算力、大网络”特征，光传输交换具备更低功耗等使能技术，技术创新带来新体系新机制等，从集群的体系结构、计算结点、网络技术和技术生态 4 个方面，对国外超大规模智算集群发展趋势作出 10 点初步判断，供读者参考。

关键词 超大规模智算集群；系统结构演化；计算结点；高速互联

中图法分类号 TP393

DOI: 10.7544/issn1000-1239.202660242 **CSTR:** 32373.14.issn1000-1239.202660242

收稿日期：2026-04-17；**修回日期**：2026-06-10

基金项目：国家自然科学基金项目（62372462）

This work was supported by the National Natural Science Foundation of China (62372462).

超大规模智算系统在向更大规模、更高能效、更通用化发展的过程中,仍面临着内存墙、通信墙、能耗与散热、生态壁垒等挑战,为了应对这些挑战,本文斗胆对超大规模智算集群体系结构发展趋势作出初步判断。

1 超大规模智算集群的体系结构发展趋势

自 2012 年 AlexNet 在 ImageNet 竞赛中取得突破性成果[1] 以来,深度学习技术成为人工智能(artificial intelligence, AI)发展的核心驱动力,而深度学习依靠高维度的矩阵运算,计算量呈现出指数级增长态势,远超摩尔定律。模型参数从百万级、千万级跃升至百亿级、千亿级甚至万亿级[2],对计算的算力密度、并行能力、能效比提出了极高要求。

以大语言模型(large language model, LLM)为代表的智能计算任务,不仅需要海量的浮点运算能力,而且对内存带宽、数据交互、延迟控制等也有严格要求。为了满足上述要求,超大规模智能计算集群应运而生,通过计算结点定制化设计与网络互联高速化升级,实现算力的规模化聚合与高效协同,成为支撑 AI 大发展的核心基础设施。

1.1 趋势 1: 体系结构上可借鉴人类智能“小算力+大网络”结构

人类智能作为自然界经过亿万年演化形成的高效计算系统,其“小算力、大网络”的特征为智算集群的架构设计提供了极具价值的仿生参考。《美国国家科学院院刊》(PNAS)2021 年的研究成果揭示,人类大脑皮层中网络通信消耗的能量是计算的 35 倍[3],这是人类智能高效运行的重要特征之一。

人类大脑“计算”是单个神经元基于接收的突触输入信号,对局部环境变量做出推断和预测,该过程类似于微型传感器对原始数据的局部分析,算力需求低且高度聚焦;而“网络互联”则是将神经元的推断结果,通过动作电位传导、神经递质释放、离子电位恢复等过程传递至其他神经元的过程,是实现复杂认知、逻辑推理、知识关联的核心环节,占据了大脑能量消耗的绝大部分。

除能量消耗特征外,人类大脑还采用了“极慢但极大”的运行策略,拥有 200 万亿个突触,每个突触可存储约 5 b,总存储容量达 1 Pb 级别,但是以亚 kHz 的极低频率运行,人脑通过大规模的网络连接与高效的信号传递,实现了远超人工计算系统的智能水平。该论文[3] 认为人类“计算”的本质,也就是智能的实现不仅依赖计算能力,更依赖高效的存算协同与网络通信。这个研究结论给大规模智算系统结构发展研究带来新启示。

人类大脑的架构特征与当前智算集群的系统结构形成了鲜明对比,也直指智算集群发展的核心瓶颈——数据传输的能耗过高、效率过低。当前计算结构中,存储访问需要的能耗远比计算高,文献 [4] 认为计算的功能操作约 10pJ,而 DRAM 访问是 1~2 nJ,数据访问功耗大约是计算操作的 2 个数量级。文献 [5] 认为执行一次双精度浮点加法或乘法,能耗极低,在几皮焦耳(pJ)(10^{-12} 焦耳)。然而,从主存(DRAM)读取一个 64 位的双精度数字,能耗在几百到几千皮焦耳级别。文献 [6] 研究表明 H100 在 Decode 阶段 GPU 利用率仅 0.13%,但是功耗却仅降 16.2%。

数据传输的问题本质是“计算”与“存储”分离架构导致的。计算单元与存储单元相互独立,数据需要通过总线频繁在二者之间传输,不仅产生了巨大能耗,还带来显著延迟,难以适配大模型对高带宽、低延迟的需求。为破解这一瓶颈,近存计算、存算一体等新型架构应运而生,IBM NorthPole 芯片[7] 是该方向的一种探索。该芯片通过近存计算架构,将计算单元与存储单元深度融合,在 30 亿参数的小规模推理场景中,实现了比最节能 GPU(图形处理器)速度快 47 倍、能效高 73 倍的性能表现。尽管 NorthPole 架构尚未在百亿参数以上的大模型完成验证,但验证了一个重要结论,也就是通过存储部件与计算部件近距离耦合,形成存算一体部件,可以减少大量存算一体部件间的数据传输量,实现计算系统能效比的数量级提升。

该研究为超大规模智算集群的体系结构提供了参考方向,从传统的“以计算为核心”转向“存算协同、网络优化”,通过计算结点定制化设计减少数据传输,通过网络技术的升级提升传输效率,实现算力的高效聚合与整体提升。

1.2 趋势 2: CPU+多样化加速部件是大规模智算逻辑结点的主流结构

早期为了满足科学计算需求,如面向流体力学、空气动力学求解,研制了专用的仿真计算机系统,效率比通用 CPU 高数十倍。银河 II 型仿真计算机是一种专用的计算机,分为 64 算术逻辑 ALU、64 位乘法 MUL、存储部件、控制部件等。我们作为主要完成人利用 ECL100K 电路实现 64 位 ALU 计算部件,直接感受到设计难度和计算速度的优势。其后,面向科学计算和事务处理需求,计算结点以通用 CPU 为

核心，强调逻辑控制能力与通用性。2010年前后，CPU+GPU的异构计算架构取代纯CPU，该变革的核心在于CPU与GPU的架构和性能的互补性。

异构融合计算为解决通用CPU并行算力不足的痛点，发展出多种形态加速部件，如英特尔早期的Phi[8]、英伟达GPU等，以及基于FPGA的可重构计算部件，提升了算力密度与能效比。国防科技大学研制的天河超算系统是重要里程碑，创新CPU+GPU超算异构混合架构，先后6次荣登全球超级计算机TOP 500榜单[9]，证明了异构计算在算力规模化的核心价值。此外，广州国家超算中心天河超算系统在2012年实现了单场景网络达到3万余结点，标志着计算结点的规模化协同技术趋于成熟。

超大规模智算集群既需要满足大模型训练，也需要满足大规模智能体协同等多种场景，CPU擅长复杂逻辑调度、工具调用、长流程编排等，GPU/TPU等加速部件擅长矩阵运算等特定处理，通过CPU+加速部件的高效协同，实现性能倍增。我们认为CPU+多样化加速部件是智算时代逻辑结点的主流结构。英伟达的GPU、谷歌的TPU[10]等是当前的主要加速部件。

2 超大规模智算集群计算结点的发展趋势

计算结点是超大规模智算集群的基本计算单元，计算结点系统结构设计直接决定了集群的算力密度、能效比与应用适配能力。计算结点历经了专用计算机、通用CPU、CPU+GPU异构混合、AI定制芯片4个重要发展阶段，实现从通用计算到智能计算的深度变革，每一次迭代都围绕当时需求、技术条件和工艺条件等，在结点构成、计算精度、存算协同、延迟控制等方面寻求最大突破。

2.1 趋势3：智能加速芯片体系结构将迎来百花齐放的定制化时代

初步判断，智能加速芯片的体系结构迎来了类似20世纪90年代CPU的热潮，进入了百花齐放的AI定制芯片时代。英伟达、谷歌、AMD、亚马逊、微软、Meta、高通、英特尔、华为、百度、阿里等公司纷纷推出自研的AI定制芯片，针对大模型训练、推理、边缘智能等不同场景进行深度优化，形成了多样化的计算结点架构，推动智能计算的算力密度与能效比持续突破。

以国外AI定制芯片为例，主要有10个系列，包括：1）英伟达Vera Rubin GPU，Rubin架构目前是行业标杆；2）Cerebras WSE-3，采用整块晶圆芯片设计，拥有4万亿晶体管和44 GB SRAM，带宽达21 PB/s，颠覆了传统的芯片设计理念；3）AMD MI455X，搭配Helios平台，拥有432 GB HBM4，海量内存为大模型计算提供了有力支撑；4）英伟达LP30，联手Groq采用纯SRAM架构，基于三星4 nm工艺，推理性能实现35倍提升；5）谷歌TPUv7（Ironwood），作为第7代TPU（张量处理单元），搭载192 GB HBM单芯，带宽达7.37 TB/s，成为Google Gemini大模型的计算底座；6）亚马逊Trainium3，基于台积电N3P工艺，为AWS云平台的智能训练提供专用算力；7）微软Maia 200，采用18A先进工艺，搭载288 GB HBM4，成为Azure云平台的核心智算芯片；8）Meta MTIA v3，针对推荐算法优化，大幅降低运营成本；9）高通AI250，以手机等市场为重点，主打低功耗，适配边缘智能计算场景；10）英特尔Jaguar，丰富了智算芯片的生态。多样化AI定制芯片，针对不同的智能计算场景实现了优化，形成了“各有所长、互补发展”的格局。

这一轮竞争的特点不同于上世纪90年代CPU的竞争，过去CPU研制和使用是高度分离的，而AI时代出现研产高度重合或者少量巨头捆绑现象。谷歌的TPU、亚马逊的Trainium3、微软的Maia 200、阿里平头哥的AI芯片和系统都是从自研自用开始。因为一个巨头AI企业每年可以消耗数十万片，足以支撑研发投入。再叠加各自领域的差异，因此需求不同导致设计分化。初步判断，智算加速芯片向头部聚拢的竞争会更激烈，持续时间会更长。

2.2 趋势4：通过混合精度与Prefill和Decode两阶段分离，突破能力和延迟的双重约束

2.2.1 扩大计算部件数量和采用混合精度，大幅提升计算性能

AI计算的核心特征是海量并行运算，因此强化计算单元的核心思路是优化芯片的资源分配，将更多的芯片面积用于计算部件，减少非计算单元的资源占用，实现并行算力的最大提升。

1）优化芯片资源分配。一是增加计算核心数量，GPU的计算核心数量可达数千甚至上万个，远超CPU的数十个核心，能够同时执行大量的并行计算任务，大幅提升数据密集型任务的处理效率；二是优化存储层次，减少非计算资源占用。AI计算的数据局部性弱，可大幅减少Cache等资源占用，将释放的芯片面积用于计算单元，同时弱化Cache一致性，显著提升芯片计算占比。典型案例如英伟达提出的SIMT[11]（单指令多线程）架构。同一行计算核心仅

有1个控制单元，执行相同的指令，简化了指令调度模块；通过大幅减少Cache资源，提升计算核心的数量，成为智能计算的主流加速架构。

2)设计专门针对矩阵运算的硬件单元。AI计算的核心是矩阵运算，包括矩阵乘法、卷积运算等，这些任务占据了智能训练与推理的绝大部分计算量。因此，创新计算部件结构的核心思路是设计专门针对矩阵运算的硬件单元，实现核心任务的硬件加速，大幅提升计算效率。谷歌TPU、英伟达Tensor Core等专用张量计算单元应运而生，相比传统的通用计算单元，矩阵运算效率提升了数十倍甚至上百倍。

3)探索混合精度。传统的科学计算以64位浮点(FP64)为主，追求计算准确性。但智能计算任务对精度要求相对灵活。许多AI任务即使在较低精度下仍能接受。而降低计算精度能显著减少内存带宽需求和芯片功耗，提升计算效率。因此，探索混合计算精度成为提升结点混合计算能力的核心路径。芯片厂商纷纷推出对低精度计算的硬件支持[12-14]，从16位浮点(FP16)、8位整数(INT8)，到8位浮点(FP8)、4位浮点(FP4)等，低精度计算成为智能计算的主流选择。

在模型的正向传播和反向传播过程中，采用FP16、FP8等低精度进行计算，大幅提升计算速度；而在权重更新过程中，采用FP32高精度进行计算，避免精度损失导致的模型收敛困难。“低精度计算+高精度更新”模式，实现了性能与精度的平衡。混合计算精度的探索，实现了算力效率与能效比的大幅提升，成为破解大模型算力瓶颈的关键手段。

2.2.2 推理优化出现以存储和计算为核心的两大技术思路

传统的推理方案采用单一GPU处理Prefill以及Decode两个阶段，导致资源浪费。GPU海量并行算力在Decode阶段无法得到充分利用，硬件资源利用率仅维持在30%~50%[15]；同时，GPU依赖高带宽内存(HBM)，其容量有限、访问延迟较高，长上下文处理时易导致键值缓存(KV Cache)溢出，加剧了延迟问题。

体系结构专家Patterson提出了以HB Flash(高带宽闪存)为代表的存储优化思路[16-17]。该思路通过在传统的存储层级中新增HB Flash层，构建“片上SRAM—HBM(高带宽内存)—HB Flash—DDR5(标准内存)—SSD/NVMe”的分层存储体系，明确了各存储层级的功能定位。片上SRAM用于存储热键值缓存，实现高速访问；HBM用于存储模型权重和部分热键值缓存；HB Flash作为新增层级，用于存储海量的冷键值缓存，提供TB级别的存储容量和远高于DDR5的带宽；DDR5和SSD/NVMe则用于存储模型数据和超大规模的缓存数据。HB Flash提升了存储体系的带宽与容量，减少了数据在不同存储层级之间的迁移，缓解了内存墙问题，适配了长上下文推理的缓存需求。

另一个思路是计算重构。该思路通过对计算架构的重构，设计专门针对大模型推理的计算系统，英伟达提出GPU+LPU混合架构。GPU专注于处理计算密集的Prefill阶段，充分发挥其海量并行算力的优势，快速完成上下文信息的计算；LPU则专为大模型推理的Decode阶段设计，针对键值缓存的频繁读取、低延迟串行计算进行深度优化，成为Decode阶段的专用加速单元。二者的协同工作，将推理阶段的硬件资源利用率提升至80%以上[18]。LPU架构典型案例是Groq 3 LPX，它有3方面创新：一是大规模应用片上SRAM，通过SRAM消除内存访问延迟，实现键值缓存高速读取，突破了内存墙限制；二是支持超百万Token(词元)的超长上下文推理，解决了传统GPU长上下文处理时的缓存溢出问题；三是延迟优化，通过架构简化、指令定制等方式，将首个Token延迟降低至8~10 ms，仅为同级别GPU的十分之一，实现了大模型推理的快速响应[18]。

这2个技术思路推动计算结点从“训练与推理通用”向“训练专用+推理专用”的方向演进，实现了智能计算的精细化、定制化设计。

2.3 趋势5：探索和创新面向AI的新型计算结构

伴随智能计算需求的持续升级，计算结点的架构创新不再局限于传统的芯片升级，而是探索和创新新型计算结构。通过计算核心的重构、存算的深度融合、网络的原生集成，打造适配超大规模智算集群的新型计算结点，Tenstorrent公司的Tensix可以视为一种探索[19]。

Tenstorrent公司的计算结点以Tensix核心为基本计算单元(Core)，采用高度集成的设计，融合了计算、存储、通信等多种功能，成为一个紧凑的智能计算单元。每个Tensix核心包含5个RISC-V CPU、2个片上网络(NoC)接口、1个向量运算单元、1个矩阵/张量运算单元、1个数据打包解包单元，以及片上1.5 MB SRAM。其中，RISC-V CPU负责核心的逻辑控制与任务调度，向量运算单元和矩阵/张量运算单元负责并行计算，片上SRAM用于存储本地数据，促进本地组件间的数据交换，片上网络接口实现Tensix核心之间的高速互联。

Tensix 的创新主要有 2 方面：一是 Core 原生支持分块矩阵运算，支持 32×32 分块的矩阵运算，能够高效执行智能计算的核心任务，提升计算效率；二是计算与通信的深度融合，Core 内部集成片上网络接口，实现 Core 间的无阻塞高速互联，减少数据交互延迟，符合计算结点内多核心协同的需求。

Tensix 的探索，打破“计算、存储、通信”相互独立的传统架构，实现三者的深度融合，推动计算结点向高度集成、存算协同、网算一体的方向发展。

3 超大规模智算集群网络技术的发展趋势

在超大规模智算集群中，网络技术与计算结点同等重要，也是实现算力规模化聚合与高效协同的关键基石。随着计算结点的算力密度持续提升，集群的规模从数千卡向数十万卡、百万卡级别扩张，对网络技术的带宽、延迟、可扩展性、可靠性提出了极致要求。

超大规模智算集群的发展，本质是算力的规模化聚合过程，而算力的高效聚合必须依赖高效的网络技术。传统的超算集群以科学计算为主，网络技术以满足结点间基本的通信为目标；而智算集群以大模型训练与推理为核心任务，存在两大显著的通信特征：一是海量数据交互，大模型训练推理过程中，计算结点需要频繁同步模型参数、梯度数据，单卡的通信带宽需求持续提升；二是海量聚合-分发任务、全对全通信（ALL-TO-ALL）、大模型的混合专家（MoE）架构催生了海量的聚合-分发任务，要求网络支持全对全的高速通信，通信量呈指数级增长。

规模扩张使得“网络税”成为不可忽视的问题，网络的延迟、带宽损耗会随其规模的扩大而增加，若网络无法同步升级，将导致集群实际算力远低于理论算力，出现“算力聚合不增效”现象。因此，网络技术的演化一定与计算结点的演化深度绑定，形成了“算力驱动网络、网络支撑算力”的协同演化关系。

从超算/智算系统的整体发展演化来看，结点互联结构历经了从对称多处理（SMP）到大规模并行处理（MPP），再到集群的演化过程，网络的重要性越来越突出。SMP 系统通过共享总线实现通信，能力有限；MPP 系统通过专用高速互连网络实现多维度并行消息传递，数据传输能力得到提升；而超大规模智算集群网络则需要极高速网络、分层组网、光电融合等技术，聚合万卡级以上算力，满足大规模智算中千亿级或者万亿级参数更新的需要。

3.1 趋势 6：大规模智算集群催生多样化的网络定制方案

大规模智能计算对网络的高带宽、大规模、高可靠需求催生了多样化的网络解决方案，总体上是按照领域定制的思路展开[20]，不同的企业基于自身的技术积累、应用场景，定制设计了差异化网络架构，形成了“私有协议与通用协议并存、分层组网与全互联结合”的格局。这些方案在互联协议、网络层级、故障恢复、软件耦合、核心理念等方面存在显著差异，适配了不同规模、不同场景的智算集群需求。

表 1 是部分大规模智算网络定制方案，包括谷歌 TPU G-ICI[21]、AWS Neuron Link[22]、英伟达 NVLink[23]、华为 Cloud Matrix 384[24]、xAI 的 PCIe 互联与 Spectrum-X 以太网结合[25] 等。

多样化的网络解决方案背后是不同的设计思路：一是静态定制优化，以 Google TPU 为代表，通过构建私有协议、物理隔离的光交换网络，实现静态场景下的高带宽与低延迟，适配大规模、高稳定的大模型训练需求；二是英伟达采用了极高带宽的思路，通过 NVLink 实现结点内的高速互联，通过 InfiniBand/RoCE 实现结点间的规模化组网，依托先进芯片工艺，提供无阻塞带宽，成为当前智算集群网络的主流选择；三是基于通用以太网协议，灵活优化和扩展。以 AWS Trainium 为代表，通过多路径、重传机制解决网络拥

Table 1 Partial Specific Schemes for Intelligent Computing Network

表 1 部分智算网络定制方案

主体	互联协议	网络层级	故障恢复	软件耦合	核心理念
谷歌 TPU (v4/v7)	G-ICI（私有轻量级、基于信用）	物理隔离：ICI 和 DCN 存储分离	物理重构：OCS 旋转镜面隔离	强耦合：XLA 编译器需感知物理拓扑	静态高速：构建 OCS 光交换网络，确定拓扑
AWS Trainium (Trn2/3)	Neuron Link（环）+ EFA/SRD（以太网）	相对融合：内部 2D Torus，外部走 EFA	协议重传：SRD 协议处理乱序与多路径重传	解耦：软件层网络扁平化，依赖底层硬件重传	动态灵活：依靠以太网多路径解决拥塞
英伟达 NVlink	NVLink + InfiniBand/RoCE	分层架构：Scale-up 与 Scale-out 分层	协议重传：依赖 IB/RoCE 重传机制	解耦：CUDA 生态屏蔽底层拓扑差异	带宽堆叠：顶级芯片提供无阻塞带宽
华为 Cloud Matrix 384	以太网 RoCE	两级架构	协议重传：依赖 IB/RoCE 重传机制	-	数量突破：依靠一定规模破除芯片工艺技术受限
xAI	结点内：PCIe；结点间：以太网	两级架构	-	耦合：基于 CUDA 生态	折中优化：兼顾价格与性能

塞问题,适配多样化、动态化的智能计算需求。华为则通过规模突破的思路,基于以太网 RoCE 构建两级架构,依靠计算结点与网络设备的规模,破除芯片工艺受限的现实,实现超大规模智算集群的组网。这些方案为超大规模智算集群的设计提供了丰富的选择,推动网络技术向“场景化、定制化”方向发展。

3.2 趋势 7: 结点互联成为核心竞争点,网络带宽实现跨越式提升

伴随智能计算的发展,结点互联成为超大规模智算集群的核心竞争点,其核心变革在于通信连接点的深度下沉与网络带宽的跨越式提升,实现了“计算部件间带宽>总线带宽”的突破,解决了传统总线的通信瓶颈。

在传统的服务器架构中,服务器间的通信连接点是 PCIe 总线,计算部件之间的互联必须通过 PCIe 总线,总线成为通信效率的天花板;而在高性能的 AI 服务器中,网络连接点必须深入到计算部件内部,计算芯片之间可以直接通过高速互联协议实现通信,无需经过 PCIe 总线,大幅提升了通信效率。这一变革使得计算结点间的带宽超过总线带宽,开启了结点互联新篇章。2016 年,英伟达在 Pascal 架构中首次引入 NVLink 技术,其带宽达到 160 GB/s,大约是当时 PCIe3.0 x16(15.8 GB/s)的 5 倍,实现了 GPU 之间的高速直连。此后,NVLink 技术持续迭代,NVLink4.0 采用四电平脉冲幅度调制技术(PAM4),双向带宽提升至 900 GB/s,成为当前结点内高速互联的代表性技术。NVLink5 已经用于 DGX 300 NVL72,NVLink6 已发布技术规格。

除了英伟达的 NVLink,其他企业也纷纷推出自研的专用高速互联技术,例如谷歌的 G-ICI、亚马逊的 Neuron Link、华为的灵衢网络等,这些技术针对智能计算的通信需求进行定制化优化,实现了计算部件间的高速、低延迟通信,成为超大规模智算集群的核心技术支撑。

3.3 趋势 8: 充分发挥光传输交换优势,提高全系统稳定性和能效比

智算集群的算力密度持续提升,导致系统稳定性差、能耗过高等问题。超大规模智算集群引入光技术,改善系统稳定性和能效比。

1)不断缩短电传输距离,提升系统稳定性

电传输技术受限于传输损耗、延迟等问题,难以满足万卡级以上集群的稳定性、高带宽需求,这使得光传输技术成为重要选项。光传输技术聚焦于近封装光模块(NPO)[26] 和共封装光模块(CPO)[27] 的竞争,二者通过将光引擎与计算芯片/交换芯片进行近距离封装,减少光信号的传输损耗,提升稳定性、传输效率。

NPO 是指光引擎位于紧邻 ASIC(专用集成电路)的独立载板上,与 ASIC 间的物理距离小于 5 cm,光引擎围绕 ASIC 排列,通过短距离的电信号实现与 ASIC 的通信。NPO 的优势在于设计灵活、兼容性强,光引擎与 ASIC 相互独立,可灵活更换光引擎或 ASIC,无需对整体架构进行大幅调整;NPO 通过缩短光引擎与 ASIC 的距离,减少了电信号的传输损耗,提升了通信效率和信号完整性,相比传统的外置光模块,延迟与功耗显著降低。NPO 是光传输技术从“外置”向“封装”演进的重要阶段,为 CPO 技术的发展奠定了基础。

CPO 是光传输技术的发展方向,光引擎与 ASIC 共享同一个封装基板,二者的物理距离一般小于 1 cm,实现了彻底的融合,不可物理拆分。CPO 的优势在于传输损耗小、能效比高,光引擎与 ASIC 的“零距离”融合,消除了电信号传输损耗和信号完整性差等,提升了通信带宽与速度,显著降低了功耗。CPO 通过将光引擎与 ASIC 集成在同一个封装基板上,减少了设备的体积,适合智算集群高密度组网的需求。尽管 CPO 拥有显著的性能优势,但目前仍面临工艺复杂、成本较高、兼容性差等问题。封装工艺要求极高,光引擎与 ASIC 的协同设计难度大,且一旦封装完成,无法单独更换光引擎或 ASIC,灵活性较低。

2)探索光交换技术在大规模智算集群应用

传统的电交换技术存在延迟高、功耗大、带宽受限等问题。超大规模智算集群的网络技术需要传输层的光电融合,还需要考虑在交换层从电交换到光交换的迭代升级,打造“电交换负责短距离、光交换负责长距离”的高效交换体系,从而满足万卡需求。

光交换技术以光信号为交换对象,通过光器件实现信号的交换与路由,具有带宽高、延迟低、功耗小、抗干扰能力强等核心优势,能够满足超大规模智算集群长距离、高带宽的通信需求。

谷歌的 Pods/Cube[28] 架构是典型代表。谷歌基于光交换技术,借鉴超算互联的立方体结构,设计了 Pods/Cube 架构,实现了 9 216 卡的超大规模 POD 组网,采用 TPUv7 Twisted 3D Torus 拓扑,满足了 Google Gemini 大模型训练的超大规模通信需求,证明了光交换技术与分层组网在超大规模智算集群中的核心价

值。OCS(光核心交换)成为超大规模智算集群网络架构的一种新方向。OCS 采用 2D MEMS(微机电系统)和二向色分光镜构建核心交换组件，通过两级 2D MEMS 实现三维空间内的精准光束操纵，二向色分光镜则实现了业务光与监控光的分离(允许 1 310 nm 业务光透射，反射 850 nm 监控光)，实现了光信号的高效、精准交换。

4 超大规模智算集群的技术生态发展趋势

超大规模智算集群的发展，不仅是硬件技术的迭代，更是生态体系的构建，核心在于人才、产业、市场的协同发展，是一个长期的、系统性的工程，无法一蹴而就。智能计算结点进入了百花齐放的 AI 定制芯片时代，也带来了生态碎片化与生态壁垒问题，成为制约芯片技术规模化应用的核心挑战。AI 定制芯片的性能发挥高度依赖软硬件的深度协同，导致 AI 定制芯片均构建了自身的封闭生态。例如，谷歌 TPU 依托 XLA 编译器与 TensorFlow 框架，英伟达 GPU 依托 CUDA 软件栈。生态之间相互独立，难以兼容。我们要发展智能技术，赋能千行百业，必须在抓好智能技术研发的同时，高度重视技术生态建设。

4.1 趋势 9: 人才培养是智算生态竞争的核心要素

生态构建的核心是人才，尤其是掌握硬件原理、编译优化、底层驱动的底层技术人才。AI 定制芯片的生态壁垒，本质上是人才壁垒，即只有拥有足够的底层技术人才，才能实现芯片硬件的自主设计、编译器的深度优化、软硬件的协同适配，最终构建起自主可控的生态体系。这类人才存在严重的短缺，尤其是掌握硬件原理与软件优化的复合型人才，成为制约智算集群技术创新与生态构建的关键因素。例如，苹果、Anthropic 等企业为适配谷歌 TPU 架构，需要大量挖掘谷歌的“校友”，就是因为缺乏掌握 JAX/XLA 深度调优能力的专业人才。

从行业发展的经验来看，2010 年前后 IBM 的 NPSG 系列网络处理器[29] 也面临与目前 AI 定制芯片类似问题，性能发挥高度依赖底层技术的优化。要解决问题，核心途径就是加强人才培养。因此，超大规模智算集群的生态构建，首先需要加强计算机硬件、编译原理、芯片设计、高速互联等领域的人才培养，通过高校专业设置、企业人才培养、产学研协同等多种方式，培育一支庞大的底层技术人才队伍，为生态构建奠定人才基础。清华、北大等双一流高校等成为 AI 发展重镇和源头，也证明了人才培养是竞争的核心要素。

4.2 趋势 10: 时间积累与市场验证是生态构建的关键路径

生态构建是一个长期的时间积累过程，需要经过技术的持续迭代、产品的不断优化、开发者的逐步认可，才能形成成熟的生态体系。英伟达的 CUDA 生态从早期黄仁勋亲自背着 GPU 到各高校推广，历经 20 余年的发展，形成了今天的规模，成为 GPU 生态的标杆[30]；谷歌的 TPU 生态从内部自用走向对外开放，也经历了多年的技术积累与优化。

同时，市场验证是生态构建的关键路径，只有让产品在实际的市场场景中得到应用，才能发现技术与生态的问题，进而持续优化；也只有通过市场的规模化应用，才能吸引更多的开发者加入生态，形成“技术创新—市场应用—生态完善—技术再创新”的良性循环。因此，超大规模智算集群的生态构建，需要给予足够的时间与市场空间，让技术与产品在实践中不断成熟。

我们应该看到，TPU 阵营正成为英伟达的重要竞争对手。苹果已于 2024 年确认采用 TPU 训练其 AI 模型；Anthropic 在 2025 年开始基于 TPU 构建超大规模集群；Meta 计划 2026 年将 TPU 纳入租赁体系，同时与 AMD 达成 6 GW 算力的 5 年战略合作。驱动这些巨头变化的核心逻辑，不再是单一芯片的峰值算力，而是系统级工程效率与规模化推理的性价比，TPU 在大模型推理中的能效比优势，使其成为大模型时代选择的重要理由。

5 总　　结

超大规模智算集群作为智能计算时代的核心基础设施，其体系结构演化是智能算力需求与底层技术创新共同驱动的结果。根据当前发展态势，我们尝试提出 10 点趋势判断，抛砖引玉以供参考。总体而言，超大规模智算集群不仅是结点技术与网络技术的创新，更是技术生态与智能体系的重构。马斯克于 2026 年 5 月宣布将 xAI 旗下 22 万张算力卡全部转租给 Anthropic 公司，从而盘活资产，调整 AI 布局，可见 AI 底座大模型生态竞争程度之激烈。2026 年 6 月 24 日，OpenAI 宣布首颗定制芯片 Jalapeño 已经部署，Anthropic 宣布计划为 Claude 打造专用芯片。大模型公司开始研制计算结点芯片，将进一步加剧该领域的竞争。

作者贡献声明：苏金树设计论文的总体框架，提出10个趋势的核心观点，撰写论文的主体部分；夏雨生和董德尊完善了体系结构的2个趋势；宋丛溪和徐扬完善了计算结点的3个趋势；黄见欣完善了技术生态的2个趋势，并负责整理参考文献和校对论文；陈果完善了网络互联的3个趋势。

参 考 文 献

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84−90

[2] Sevilla J, Roldán E. Training compute of frontier AI models grows by 4−5x per year[EB/OL]. 2024[2026-05-27]. https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

[3] Levy W B, Calvert V G. Communication consumes 35 times more energy than computation in the human cortex, but both costs are needed to predict synapse number[J]. Proceedings of the National Academy of Sciences, 2021, 118(18): e2008173118

[4] Horowitz M. Computing's energy problem (and what we can do about it)[C]//Proc of 2014 IEEE Int Solid-State Circuits Conf Digest of Technical Papers (ISSCC), Piscataway, NJ: IEEE, 2014: 10−14

[5] Kestor G, Gioiosa R, Kerbyson D J, et al. Quantifying the energy cost of data movement in scientific applications[C]//Proc of 2013 IEEE Int Symp on Workload Characterization (IISWC). Piscataway, NJ: IEEE, 2013: 56−65

[6] Moon S, Cha J, Park H, et al. Hybe: GPU-NPU Hybrid System for efficient LLM inference with million-token context window[C]//Proc of the 52nd Annual Int Symp on Computer Architecture (ISCA'25). New York: ACM, 2025: 808–820

[7] Cassidy A S, Arthur J V, Akopyan F, et al. 11.4 IBM NorthPole: An architecture for neural network inference with a 12 nm chip[C]//Proc of 2024 IEEE Int Solid-State Circuits Conf (ISSCC). Piscataway, NJ: IEEE, 2024: 214−215

[8] Sodani A, Gramunt K, Corbal J, et al. Knights landing: Second-generation Intel Xeon Phi product[J]. IEEE Micro, 2016, 36(2): 34−46

[9] TOP500. The 66th TOP500 list[EB/OL]. 2025[2026-04-14]. https://www.top500.org/TOP500

[10] Patterson D. Domain specific architectures for deep neural networks: Three generations of tensor processing units (TPUs)[R/OL]. Cambridge: MIT CSAIL, 2019[2019-10-16]. https://www.csail.mit.edu/news/domain-specific-architectures-deep-neural-networks-three-generations-tensor-processing-units

[11] NVIDIA. Parallel Thread Execution ISA Version 2.0[M/OL]. Santa Clara: NVIDIA Corporation, 2009[2009-03-09]. https://developer.download.nvidia.com/compute/cuda/2_0/toolkit/docs/PTX_ISA_2.0.pdf

[12] Fang B, Hari S, Tsai T, et al. Towards precision-aware fault tolerance approaches for mixed-precision applications[C]//Proc of IEEE/ACM the 12th Workshop on Fault Tolerance for HPC at eXtreme Scale (FTXS). Piscataway, NJ: IEEE, 2022: 47−52

[13] NVIDIA. NVFP4 trains with precision of 16-bit and speed and efficiency of 4-bit[EB/OL]. 2025[2026-04-14]. https://developer.nvidia.com/blog/nvfp4-trains-with-precision-of-16-bit-and-speed-and-efficiency-of-4-bit/

[14] Jouppi N P, Young C, Patil N, et al. Datacenter focused TPU design: TPU v4 and TPU v5[J]. IEEE Micro, 2022, 42(6): 8−17

[15] Huang J S. NVIDIA GTC 2026 Keynote Official Transcript[EB/OL]. Santa Clara: NVIDIA Corporation, (2026-03-16)[2026-04-14]. https://www.nvidia.com/en-us/on-demand/session/gtc26-s81595/

[16] Ma X, Patterson D A. Challenges and research directions for large language model inference hardware[J]. Computer, 2026, 59(5): 55−64

[17] Sandisk. Sandisk Unveils the future of memory architecture for AI introducing: High nandwidth flash[EB/OL]. 2025[2026-05-27]. https://documents.sandisk.com/content/dam/asset-library/en_us/assets/public/sandisk/collateral/company/Sandisk-HBF-Fact-Sheet.pdf

[18] Aubrey K, Ghodsian F. Inside NVIDIA Groq 3 LPX: The low-latency inference accelerator for the NVIDIA Vera Rubin platform[EB/OL]. 2026[2026-05-27]. https://developer.nvidia.com/blog/inside-nvidia-groq-3-lpx-the-low-latency-inference-accelerator-for-the-nvidia-vera-rubin-platform/

[19] Brown N, Barton R. Accelerating stencils on the Tenstorrent Grayskull RISC-V accelerator[C]//Proc of SC'24: Workshops of the Int Conf for High Performance Computing Network, Storage, and Analysis. Piscataway, NJ: IEEE, 2024: 1690−1700

[20] Su Jinshu. Thinking on networking and computing promoted by domain-specific method[J]. Journal of Computer Research and Development, 2025, 62(12): 2889−2894 (in Chinese)
(苏金树. 领域定制方法促进网络与计算发展的思考[J]. 计算机研究与发展, 2025, 62(12): 2889−2894)

[21] Makhija P. TPU rack overview[EB/OL]. 2025[2026-07-03]. https://hc2025.hotchips.org/assets/program/tutorials/HC2025.T1DCRacks.S7.Pankaj.pdf

[22] AWS NEURON. Amazon EC2 Trn1/Trn1n architecture[EB/OL]. 2025[2026-04-15]. https://awsdocs-neuron.readthedocs-hosted.com/en/latest/general/arch/neuron-hardware/trn1-arch.html

[23] Nvidia. NVIDIA NVLink and NVLink switch[EB/OL]. 2026[2026-05-27]. https://www.nvidia.com/en-us/data-center/nvlink/

[24] Zuo P, Lin H, Deng J, et al. Serving large language models on Huawei CloudMatrix384[EB/OL]. (2025-06-05) [2026-04-15]. https://arxiv.org/pdf/2506.12708.pdf

[25] Crosley B. Xai's Memphis Colossus: Anatomy of a 100, 000 GPU supercomputer[EB/OL]. 2026[2026-05-27]. https://introl.com/blog/xai-memphis-colossus-100000-gpu-supercomputer-infrastructure

[26] Link-PP Community. Beyond pluggables: What is NPO (near-packaged optics) and why does it matters[EB/OL]. 2025[2026-05-27]. https://resources.l-p.com/glossary/what-is-near-packaged-optics-benefits-network-upgrades

[27] Mahajan R, Li X, Fryman J, et al. Co-packaged photonics for high performance computing: Status, challenges and opportunities[J]. Journal of Lightwave Technology, 2022, 40(2): 379−392

[28] Poutievski L, Mashayeki O, Ong J, et al. Jupiter evolving: Transforming Google's datacenter network via optical circuit switches and software-defined networking[C]//Proc of the 2022 ACM SIGCOMM Conf. New York: ACM, 2022: 1−17

[29] Allen J R, Bass B M, Basso C, et al. IBM PowerNP network processor: Hardware, software, and applications[J]. IBM Journal of

Research and Development, 2003, 47(2.3): 177−193

[30] Tae Kim. The Nvidia Way: Jensen Huang and the Making of a Tech Giant[M]. Beijing: CITIC Press Group, 2025 (in Chinese)
（金泰. 英伟达之道：黄仁勋和他的科技帝国. 北京：中信出版集团, 2025）



Su Jinshu, born in 1962. Professor, PhD supervisor. Fellow of CCF. His main research interests include computer networks and cyberspace security.
苏金树, 1962 年生。研究员，博士生导师。CCF 会士。主要研究方向为计算机网络、网络空间安全。



Xia Yusheng, born in 1992. Associate professor. His main research interests include computer networks and space-air-ground integrated networks.
夏雨生, 1992 年生。副研究员。主要研究方向为计算机网络、空天地一体化网络。



Song Congxi, born in 1994. Assistant professor. Her main research interests include multipath transmission technology and video transmission.
宋丛溪, 1994 年生。助理研究员。主要研究方向为多路径传输技术、视频传输。



Huang Jianxin, born in 1985. Assistant professor. His main research interests include computer networks and cyberspace security.
黄见欣, 1985 年生。助理研究员。主要研究方向为计算机网络、网络空间安全。



Dong Dezun, born in 1980. Professor. His main research interests include computer architecture, and high-performance and intelligent computing.
董德尊, 1980 年生。研究员。主要研究方向为计算机体系结构、高性能与智能计算。



Xu Yang, born in 1979. Professor. His main research interests include AI training/inference systems, data center networks, and programmable networks.
徐　扬, 1979 年生。教授。主要研究方向为 AI 训练/推理系统、数据中心网络、可编程网络。



Chen Guo, born in 1989. Professor. Senior member of CCF. His main interest includes high performance networks.
陈　果, 1989 年生。教授。CCF 高级会员。主要研究方向为高性能网络。