CAICT 中国信通院

集智 专题报告

# AI 计算节点发展研究报告

## （2026 年）

中国信息通信研究院云计算与数字化研究所

2026年3月

# 版权声明

# 前　言

当前，全球人工智能（AI）加速发展，伴随着大模型参数规模与训练数据大幅增长，AI 产业发展推动全球 AI 算力市场规模持续扩大，互联网、金融、制造等重点行业智能化转型进一步放大算力需求。同时，传统算力架构面临单机性能受限、集群扩展瓶颈、资源利用率偏低等多重挑战，新型架构探索成为突围算力瓶颈的关键路径。AI 计算节点作为构建超大规模智能算力集群的核心，依托高速互联技术融合多算力芯片形成规模化计算单元，有效破解 AI 大模型训练中的算力协同与效率难题。

在此背景下，智能算力作为支撑人工智能高质量发展的重要基础，已成为国家战略支点，多国加大政策支持与投资力度，我国也通过多项政策部署，推动 AI 计算节点技术突破与工程落地。同时，我国智能算力正处于从规模化扩张向高效化提升的关键期，AI 计算节点凭借高密集约、高速超宽、高效灵活、高稳可靠的核心特征，通过节点架构重构、超低时延网络、CXL 内存、智能算力调度、绿色低碳供能等核心技术创新，在大模型训练、高并发推理及金融、工业、能源等行业场景应用中发挥着关键支撑作用。

立足新发展阶段，本报告系统分析 AI 计算节点发展概况、核心技术、应用场景、产业生态及未来趋势，为政策制定、技术研发与产业应用提供参考，助力构建先进易用、绿色高效的算力基础设施，推动 AI 与实体经济深度融合，夯实数字经济发展基础。

时间仓促，报告仍有诸多不足，恳请各界批评指正。后续我们

将不断更新完善，如有意见建议请联系中国信通院研究团队：dceco@caict.ac.cn。

# 目 录

一、 AI 计算节点发展概况 ..... 1
（一） 定义与核心特征 ..... 1
（二） 发展背景 ..... 2
（三） 发展阶段与演进脉络 ..... 6
二、 AI 计算节点核心技术分析 ..... 7
（一） 节点架构重构，驱动算力高效聚合 ..... 7
（二） 异构计算技术，实现算力密度突破 ..... 9
（三） 超低时延网络，破解数据传输瓶颈 ..... 10
（四） HBM 与 CXL，突破存储带宽瓶颈 ..... 11
（五） 智能算力调度，提升资源利用效率 ..... 12
（六） 绿色低碳供能，保障系统高效运行 ..... 12
三、 AI 计算节点典型应用场景 ..... 13
（一） 大模型训练场景：支撑万亿参数模型高效训练 ..... 13
（二） 高并发推理场景：保障生成式 AI 服务实时响应 ..... 14
（三） 行业智算场景：适配重点领域定制化需求 ..... 15
四、 AI 计算节点产业生态建设分析 ..... 18
（一） 国际视角：技术引领与生态开放并行 ..... 19
（二） 国内发展：多主体协同与自主生态构建 ..... 20
五、 AI 计算节点未来趋势 ..... 22
（一） 政策聚焦自主创新与多维支撑 ..... 22
（二） 技术关注高效互联与高密集成 ..... 23
（三） 产业格局头部引领与多方协同 ..... 24
（四） 行业应用试点向全域渗透迈进 ..... 25

# 图 目 录

图 1 AI 计算节点组网 ..... 1
图 2 AI 计算节点特征 ..... 2
图 3 全球人工智能服务器市场规模预测 ..... 3
图 4 AI 计算节点组网架构 ..... 9

# 表 目 录

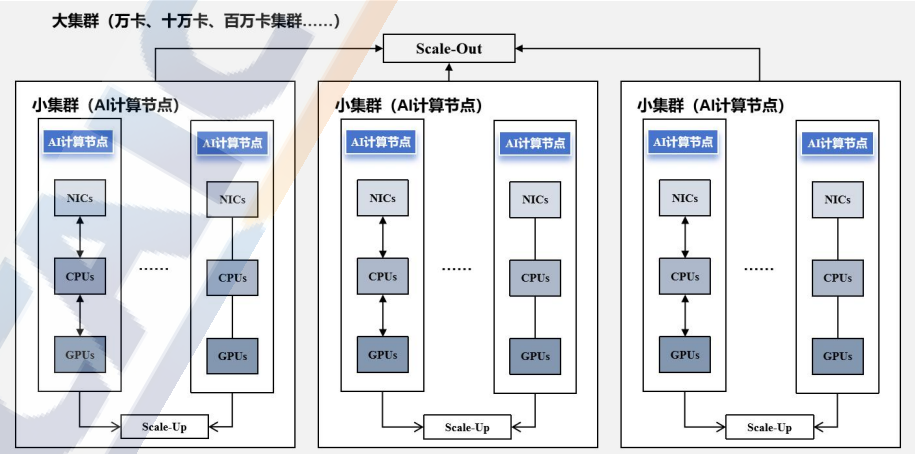
表 1 国内外 AI 计算节点产业生态对比 ..... 18

# 一、AI 计算节点发展概况

## （一）定义与核心特征

在 AI 技术加速迭代演进的背景下，我国智能算力需求正从规模化扩张向高效化提升转型，AI 计算节点已成为支撑智能算力发展的核心算力单元。AI 计算节点是构建大规模算力集群的技术架构，最早由英伟达提出，指将多张 GPU 集成在一个逻辑单元内，形成类似“超级计算节点”的系统。与传统架构相比，该节点通过高速互联技术整合多算力芯片形成计算单元，有效破解 AI 大模型训练中的算力协同与效率问题，实现效率的显著优化。

具体来说，作为以超大带宽实现多卡 GPU-GPU、GPU-CPU 及 GPU-Memory 高效互联的 Scale-Up（纵向扩展）系统，AI 计算节点以高带宽域（High-Bandwidth Domain，HBD）为基本单元，通过传统 Scale-Out（横向扩展）扩展方式构建更大规模的算力集群，有效解决 AI 大模型训练过程中算力协同不足、资源调度效率偏低等突出问题，为 AI 产业高质量发展提供坚实的算力支撑。




来源：中国信通院

图 1 AI 计算节点组网

从技术层面看，AI 计算节点的核心特征集中体现在**高密集约、高速超宽、高效灵活、高稳可靠**四大能力。通过四大能力建设，构建起高效处理各类 AI 计算任务的基础架构，为 AI 应用创新发展提供坚实支撑。具体来看，一是高密集约能力，通过硬件架构创新与多芯片集成设计，实现计算资源的高效聚合，提升并行处理效能，为大规模 AI 任务提供核心算力单元支撑。二是高速超宽能力，聚焦构建高带宽、低时延数据传输体系，采用芯片级直连等技术，有效保障计算节点数据的高效流通，破解数据传输瓶颈。三是高效灵活能力，推动异构计算资源池化与软件定义调度，实现根据任务需求动态分配算力资源，提升基础设施利用效率与灵活性。四是高稳可靠能力，通过流量管理、故障冗余等机制，确保长周期、高负载 AI 任务连续稳定执行，强化系统运行的稳定性与容错能力。


来源：中国信通院

图 2 AI 计算节点特征

## （二）发展背景

### 1.人工智能发展催生智算缺口

**当前，全球 AI 产业迭代加速，AI 大模型参数与训练数据量跨越式增长，各行业智能化转型提速，智算资源刚性缺口持续扩大，算力需求激增。**国际数据公司（IDC）数据显示，2025 年全球人工智能服务器市场规模为 1587 亿美元，2028 年有望达到 2227 亿美元[1]。AI 大模型智能水平与性能提升高度依赖算力支撑，依据 Scaling Law（规模法则），扩大模型参数规模、增加训练数据量是提升大模型能力的核心路径。而大模型参数规模已实现从百亿级向万亿级的跨越，主流大模型训练数据量从千亿级 token 跃升至数十万亿级 token，节点间数据传输量几何级增长，进一步加剧了算力资源供给压力。


来源：IDC

图 3 全球人工智能服务器市场规模预测

**AI 技术在互联网、金融、制造业等规模化落地进一步放大了智能算力的需求缺口。**随着 DeepSeek、Llama 等开源大模型的普及，大模型在各行业的落地应用将持续提速，行业模型的智能算力需求也将快速增长。**互联网行业**，头部平台算力需求爆发式增长。如字节跳动全球日活用户达 15 亿，AI 推荐引擎每日处理千亿级数据。**金融行业**，

[1] 国际数据公司（IDC）、浪潮信息，《2025 年中国人工智能计算力发展评估报告》

合规约束加剧算力刚性短缺。如邮储银行在“智慧投行”建设中，将 AI 超算与高性能计算融合，形成以千卡算力集群为核心的算力平台，支撑风控、投研、交易等业务的实时计算。**制造行业**，工业 AI 的深度渗透，催生巨量算力需求。如小鹏汽车研发 720 亿参数模型，已建成万卡智算集群，集群利用率长期稳定在 90%以上，但高峰时仍需外部调配算力才能满足需求。此外，在教育、娱乐等领域的智能问答、个性化推荐等场景，算力需求缺口同样显著。

2.智算中心成为国家战略支点

**多国政府将 AI 基础设施建设上升至国家战略，持续加大投资及政策支持。**美国“网络与信息技术研发计划”(NITRD)人工智能研发投资预算增长至 31 亿美元，占整体年预算的近三分之一，相比于上一年提高 19.2%。2025 年 1 月，美国政府公布“星际之门”的国家级人工智能基础设施计划，预计将投入 5000 亿美元用于美国国内人工智能基础设施建设。2025 年 11 月，美国特朗普政府启动的一项国家级人工智能（AI）科研动员计划—“创世纪计划”旨在整合联邦科学资源，加速 AI 驱动的科学发现，以应对科技竞争，聚焦于先进制造、生物技术、关键材料、核能、量子科学和半导体等六大战略领域。英国在《AI 机会行动计划》中提出“AI 增长区”（AI Growth Zones），通过提供电力、规划审批等专项支持，鼓励在本土建设高密度 AI 数据中心，并计划在 2030 年前将 AI 研究资源容量扩大至少 20 倍。欧盟正在推进设立“人工智能工厂”，鼓励成员国建设人工智能基础设施建设，将向数字欧洲计划拨款 8 亿欧元，用于购买新的 AI 专用计

算资源或升级现有基础设施。加拿大启动《加拿大主权 AI 计算战略》，投入 10 亿美元建设国家级超级计算系统，形成面向科研、产业和政府的公共算力平台。日本发布的《2030 年数字基础设施发展规划》中明确指出数据中心、海底光缆、AI 等“AI 时代新型数字基础设施”的发展规划。

**为抢抓全球 AI 产业竞争主动权，我国持续强化算力网络顶层设计与建设推进，加大政策支持力度。国家层面，明确 AI 计算节点发展方向。**政策出台呈现“梯度推进、重点突出”特征，2023 年印发《算力基础设施高质量发展行动计划》和《关于深入实施“东数西算”工程加快构建全国一体化算力网的实施意见》，旨在不断完善算力基础设施建设，增强算力赋能成效。2024 年《推动工业领域设备更新实施方案》提出推动“云边端”算力协同发展，加大高性能智算供给，在算力枢纽节点建设智算中心。2025 年《关于深入实施“人工智能+”行动的意见》明确提出，“支持人工智能芯片攻坚创新与使能软件生态培育，加快超大规模智算集群技术突破和工程落地。”

### 3.新型架构探索突围算力瓶颈

**随着 AI 模型参数与训练数据不断扩张，传统算力架构面临单机性能受限、集群扩展瓶颈等双重挑战，架构升级迫在眉睫。单芯片算力**提升遭遇技术瓶颈，受“内存墙”等制约，算力效能难以充分释放，先进架构下实际有效算力利用率普遍偏低，大量算力资源处于闲置状态。**集群扩展模式**同样受限，传统方案下集群扩大到一定规模后，有效算力受到限制，而万亿级参数模型需更大规模并行计算。

**大模型“参数-数据-性能”正向循环催生通信密集型场景，传统架构系统性瓶颈凸显。**主流大模型训练数据量从数十 TB 跃升至 PB 级，节点间数据传输量几何级增长，动态交互对低时延、高带宽传输需求严苛。同时，硬件、软件以及集群扩展层面资源利用率偏低问题突出。硬件层面，“训推分离”导致小规模集群 GPU 利用率不足 50%，大规模集群“算力黑洞”效应使利用率低于 30%。软件层面，现有调度系统难以适配大模型训练过程中的动态变化，导致计算资源未能最大化利用，集群扩展成本高且难以实现弹性伸缩。

### （三）发展阶段与演进脉络

**AI 计算节点发展脉络可以分为三个阶段，**从早期分散式设备简单互联，逐步向机间协同组网，再到规模化卡间直连迭代，节点互联效率、算力聚合密度、资源协同能力显著提升。

**在互联网应用发展时期，**业务应用以网页服务、电子商务、在线办公等简单数据交互型业务为主，对算力协同需求较低。算力供给模式以多服务器分布式互联为核心，通过负载均衡机制实现业务流量调度，无需构建复杂的节点协同体系。算力密度维持在单机柜数千瓦阶段，节点间互联以百千兆以太网为主，算力协同局限于单一机柜内少量设备，整体架构灵活性与扩展性较弱，尚未形成规模化的算力聚合与协同调度体系。

**在人工智能发展初期，**业务应用以中小规模模型训练、计算机视觉、语音处理等 AI 任务为主，算力需求从简单数据处理和交互向密集型计算演变，单服务器算力已无法满足需求，多服务器集群协作成

为主流形态。多服务器间通过 InfiniBand、万兆以太网等机间互联技术进行组网，构建中等规模集群，实现高效数据交互，保障模型训练过程中的多节点参数同步与数据传输需求。同时，基础设施指标实现跃升，算力密度随 GPU 等 AI 算力芯片的集中部署提升至单机柜十几至几十千瓦，节点间互联带宽也实现跃升，算力协同范围从“单机柜”扩展到“多机柜”，仍依赖基础集群管理工具，实现计算资源的统一管理与任务分配，算力聚合仍以“堆叠独立算力单元”为主，受限于机间网络延迟与带宽，多机协同效率存在瓶颈，且卡间数据交互需经服务器中转，存在明显时延损耗，未形成规模化的算力聚合能力，算力资源利用率有待提升。

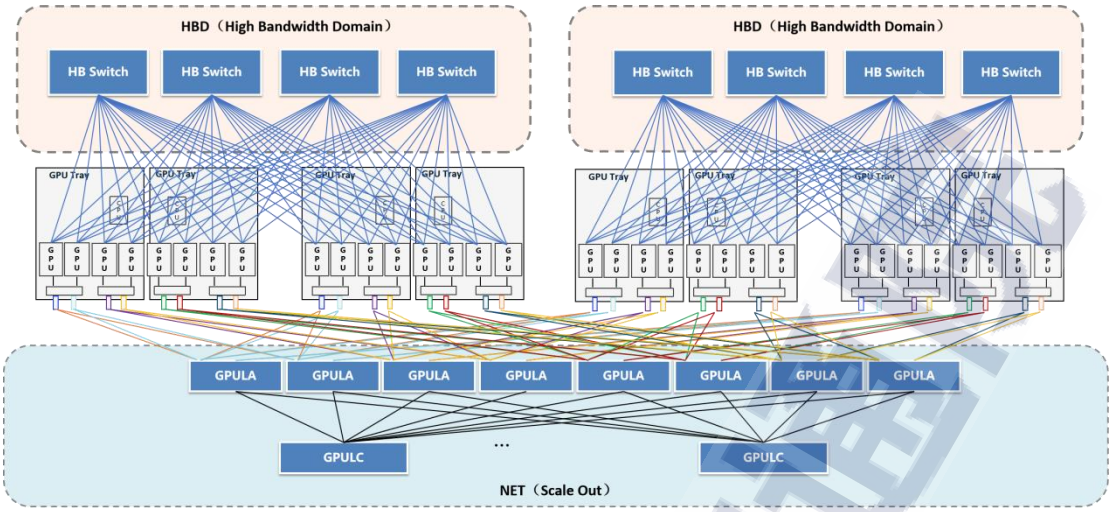
**在大模型快速发展时期，**大模型参数规模向千亿、万亿跨越，单任务算力需求呈指数级增长，传统机间互联架构难以满足低延迟、高带宽的通信要求，AI 计算节点架构迎来变革。业界通过 NVLink 等超高速互联技术实现 GPU 间的直接通信，构建起内存共享、算力一体的 AI 计算节点单元，算力密度进一步跃升，节点内互联带宽较机间网络大幅提升，通信延迟显著降低。同时，借助 InfiniBand、RoCE 等高速网络将多个 AI 计算节点组成集群，配合调度系统实现算力的全局协同，形成了新型算力架构，突破传统“堆叠算力单元”的模式，实现了算力资源的池化与协同，为大模型训练提供算力扩展能力。

## 二、AI 计算节点核心技术分析

### （一）节点架构重构，驱动算力高效聚合

为应对千亿、万亿参数大模型训练对通信性能的极致要求，传统以服务器为单元、依赖外部网络堆叠的算力架构已成为制约算力效能释放主要瓶颈。当前 AI 计算节点的演进主线是从“以连接 CPU 为中心”转向“以 GPU 互联为中心”，通过架构层面的重构，实现算力资源的高效聚合与全局协同。

**其核心突破体现在三个层面：一是卡间高速互联构建紧耦合计算单元。**通过在单节点或高密度机柜内大规模部署 NVLink 等卡间直连技术，将数十至上百张加速卡整合为一个内存统一寻址、算力无缝调用的“超级计算单元”。卡间通信带宽显著提升，延迟明显降低，解决了张量并行等紧耦合任务中的通信瓶颈。**二是节点内互联拓扑优化通信效率。**AI 计算节点普遍采用非阻塞的 Mesh、全连接或胖树拓扑，替代传统的多层收敛架构。**三是集群级网络实现大规模弹性扩展。**AI 计算节点作为基本算力模块，通过 InfiniBand 或 RoCE 高速网络进行互联，并借助全局调度系统实现跨节点的任务协同与资源池化。计算集群既能通过 Scale-Up 在节点内获得极致性能，又能通过 Scale-Out 实现近乎线性的算力扩展，支撑起万卡级别的训练任务。




来源：中国信通院，ODCC

图 4 AI 计算节点组网架构

## （二）异构计算技术，实现算力密度突破

**“CPU+GPU+专用加速器”的异构架构是支撑多样化 AI 算力的基础。**当前，AI 工作负载已从单一训练任务演变为涵盖训练、微调、推理及科学计算的多元混合体。CPU+GPU+XPU（专用 AI 加速器如 NPU、TPU、DPU）的异构架构成为选择。架构体系中，各计算单元分工明确、协同高效。CPU 作为控制平面，负责复杂逻辑调度与任务编排；GPU 凭借其数千个计算核心与高并行度计算能力，仍是支撑大规模模型训练与高吞吐批量推理的主力；而 NPU 等专用芯片则针对 Transformer 等特定模型的核心算子，进行从指令集到计算单元的全流程定制，实现能效比的提升。

**芯片级定制与先进封装技术是突破算力瓶颈的核心路径。**随着摩尔定律放缓，传统单一通用大芯片的发展模式，正遭遇功耗持续攀升与制造成本高的双重挑战。Chiplet（芯粒）技术作为模块化、异构集成的芯片设计和制造方法，通过将大型 SoC 解耦为多个功能化、模

块化的小芯片，显著提升了产品良率并降低了制造成本。同时，技术突破了传统工艺限制，支持不同工艺节点、不同材质芯片的灵活组合，有效缩短了计算单元与 HBM 内存的物理距离，提升数据交换带宽，有力缓解了制约算力有效输出的“内存墙”瓶颈，为 AI 计算节点突破算力密度瓶颈，提升计算效能及实现高可扩展性提供重要支撑。

**高密度计算硬件是推动 AI 计算节点规模化、集约化部署，实现密度和能源利用效率同步提升的重要载体。**为应对万卡级集群对机房空间、功耗及散热面临的严峻挑战，亟需在芯片与系统架构两端协同突破。在芯片层面，聚焦 GPU、AI 加速卡等关键器件的创新，提升单芯片算力峰值与数据吞吐能力。在服务器层面，依托 NVLink、高速以太网等先进 Scale-Up 互联技术，构建超低时延通信架构，实现单机柜内 64 卡及以上的高带宽互联，优化张量并行、专家并行等大模型训练场景下的通信效率。

### （三）超低时延网络，破解数据传输瓶颈

**针对大模型训练“通信密集型”特征，构建“节点内-节点间-集群间”高速互联网络体系，通过协议优化与架构升级，大幅降低数据传输延迟、提升网络带宽。节点内**，聚焦全互联拓扑与专用协议，打造 GPU 间、GPU 与 CPU 及内存的全互联网络，通过专用协议或开放标准实现“高带宽、低时延、内存共享”，进一步突破卡间通信瓶颈，提升大模型训练效率。开放数据中心委员会（ODCC）在 ETH-X 项目下推出了支持以太 Scale Up 网络传输层协议的 IP 解决方案。该方案支持内存语义，具备低延迟、高带宽和高可靠性的特点，满足多

达 512 卡的全互联。同时在 ODCC AI 网络实验室的协议承载效率、单向时延、基础吞吐、All to All 通信场景等多项测试中表现出优异性能。

**节点间**，主要依托高性能互连协议与专用交换设备，构建低延迟、高吞吐、无损的集群网络平面，将多个超节点紧密耦合。业界主要采用 RoCE 或 IB 网络协议，结合拥塞控制、流量控制、扁平化拓扑等技术，构建无损网络环境。如英伟达通过 InfiniBand 将数十至数百台服务器连接为统一集群。

**集群间**，技术焦点在于打破地理空间限制，实现跨域异构智算集群高速高可靠互联，解决大模型跨域分布式训练中数据远距离传输、多中心算力协同的核心痛点。目前，业界主要利用高速光传输网络、IPv6/SRv6、广域 RDMA 和软件定义网络（SDN）等技术，构建跨数据中心的广域互联网络，支撑跨域智算协同开展大模型训练。

### （四）HBM 与 CXL，突破存储带宽瓶颈

**在芯片层面，高带宽内存（HBM）通过三维堆叠与超宽位宽接口等解决 AI 数据密集型任务的瓶颈。**HBM 技术是突破“内存墙”的核心方式之一，如同在计算核心旁构建了超大型、超高速的“数据仓库”，能将海量参数和数据流持续、高速地服务计算单元，避免因等待数据而造成的计算核心闲置。以 Blackwell GPU 为例，其通过 8 层 HBM 实现每秒数 TB 的带宽。同时，HBM 与计算芯粒的 3D 封装设计，将数据路径缩短至毫米级，显著降低功耗与延迟。

**在内存池层面，CXL 协议与内存池化技术实现跨节点内存共享，**

**为千亿参数大模型提供弹性资源。**CXL 协议通过一致性内存访问，打破传统内存隔离限制。英伟达 GH200 将 Grace CPU 与 Blackwell GPU 互联，支持跨设备内存直接访问，为千亿参数大模型训练提供弹性扩展能力，降低硬件冗余成本。

### （五）智能算力调度，提升资源利用效率

**以软件层技术创新为核心，构建算力调度体系，着力解决传统架构下资源闲置、适配低效等问题。在调度系统方面**，产业界研发支持“训练-推理一体化”的协同调度平台，通过实时监测算力负载与模型计算需求，破解传统架构下的资源闲置、适配低效等问题，搭建分布式智能调度引擎，基于实时负载感知动态匹配算力需求与资源供给。**在生态适配方面**，产业界研发平台提供兼容英伟达、昇腾、寒武纪等多芯片架构的调度系统，大幅提升平台适配能力。

### （六）绿色低碳供能，保障系统高效运行

**AI 计算节点单机柜功耗普遍将突破 100 kW，液冷技术已从“可选项”转为“刚需”，成为降低 PUE、实现绿色低碳的关键抓手。**冷板式液冷通过直接贴合芯片散热片，将冷却液直接送到 CPU、GPU 等高功耗计算单元，散热效率远超风冷。浸没式液冷将整机柜浸入绝缘冷却液，实现热量均匀传导，提升能效。同时，配合智能温控与 AI 预测模型，实现冷却功耗的动态调节和故障预警，进一步压缩能耗波动，提升系统可靠性。

# 三、AI 计算节点典型应用场景

## （一）大模型训练场景：支撑万亿参数模型高效训练

AI 计算节点通过高效互联协议、统一内存编址、故障主动预防与恢复，有效突破了传统架构在大规模分布式计算中的性能瓶颈，成为支撑万亿参数模型高效训练的重要方案。**在互联协议方面，**AI 计算节点实现了硬件互联的革新，采用高速总线技术构建大规模节点对等直连架构，显著提升节点间通信带宽，大幅降低数据传输时延，解决大规模集群中数据交互效率不足的问题。**在内存管理方面，**AI 计算节点通过全局内存统一编址技术，将分散的节点内存虚拟为统一地址空间，消除参数同步过程中的“序列化-网络传输-反序列化”传输损耗。**在安全保障方面，**AI 计算节点采用多平面链路设计、链路故障秒级切换与算子级故障恢复等技术，延长系统平均无故障运行时长，缩短故障恢复时间，避免因单点故障或局部异常导致的训练中断，保障大规模模型训练的连续性与稳定性。

**业界厂商积极推动 AI 计算节点技术创新和产品研发，全面提升模型训练效率。**2025 年 7 月，新华三发布 H3C Uni-PoD 系列计算节点，其核心 S80000 单柜可部署 64 张 AI 加速卡，柜内可实现卡间全互联通信。2025 年 8 月，浪潮信息发布了其面向万亿参数大模型的 AI 服务器元脑 SD200，在单机内集成 64 路加速计算芯片，单机支持 DeepSeek、Kimi 等国产开源模型运行。2025 年 9 月，华为发布 Atlas 950 SuperPoD，基于 Atlas 950 实现零线缆电互联，通过直连拓扑网络架构支持 NPU 全互联，并扩展至 8192 卡。2025 年 11 月，中科曙

光发布单机柜级 640 卡 AI 计算节点 scaleX640，产品以高密架构为核心设计理念，成功实现单机柜 640 卡的超高速总线互连。

### （二）高并发推理场景：保障生成式 AI 服务实时响应

面向公众的生成式 AI 服务，如对话交互、文生图、代码生成等，需同时应对百万级用户的实时请求与超长上下文的处理需求，对显存带宽和计算资源消耗巨大。

**AI 计算节点通过高效互联协议、异构资源利用、动态资源调度，实现高并发、低时延、低成本推理，显著提升 AI 推理效率。在互联协议层面，**AI 超节点支持高速互联协议及智能数据路由，将跨节点传输中的推理数据按优先级分类缓存至节点间共享的高速缓存区，仅对增量推理数据进行实时传输，大幅降低节点间数据交互时延。同时通过协议层数据压缩与校验优化，在保证数据传输可靠性的前提下，进一步提升传输带宽利用率，减少网络资源占用。**在异构资源利用层面，**AI 计算节点融合 CPU、GPU 及专用 AI 加速卡，构建异构协同计算架构。同时依托算子级细粒度任务拆分与动态协同调度机制，实现 CPU 高效承担数据预处理与后处理，加速卡专注核心推理计算，避免单一资源瓶颈，提升集群算力利用率与能效比。**在资源调度层面，**AI 计算节点采用基于实时负载预测的弹性算力分配机制，结合共享显存池技术，将模型参数从“单实例独占”转为“多实例共享调用”，减少重复加载带来的内存开销。同时引入流量削峰算法，动态调整计算单元数量，避免峰值时段算力不足、谷值时段资源浪费。

**业界厂商围绕高并发推理需求，提升 AI 计算节点推理性能产品，加速生成式 AI 服务落地。**昆仑芯超节点采用超高密度集成设计，单个机柜支持 32 至 64 张加速卡的灵活部署，具备优异的扩展性和资源利用效率，在 DeepSeek V3/R1 PD 分离推理架构的优化下，实现了单卡性能提升，单实例推理性能大幅提升。

### （三）行业智算场景：适配重点领域定制化需求

随着 AI 技术的不断发展，行业智能化转型高速发展，金融、工业、能源等行业作为国民经济的核心领域，其智能化转型对算力需求呈现“定制化、高精度、高稳定”的鲜明特征。传统算力架构因实时服务能力不足、大算力支撑薄弱、灵活适配性差，难以突破行业智算瓶颈。AI 计算节点通过技术创新，对接金融风控实时性、工业质检精准性、能源调度长期性等差异化场景需求，构建专业化、高效能的智算解决方案，成为推动行业向智能化跨越的核心基础设施。

**在金融风控领域，**金融行业的核心风险点贯穿“交易、结算、融资”全过程，随着高频交易、实时信贷、线上支付等业务形态快速发展，其对风险甄别的实时性、精准性与并发处理能力要求也在不断提升，促使金融行业的风险控制进入毫秒级竞争时代。传统算力架构因数据读取延迟高，模型更新周期长，无法应对新型风险。AI 计算节点通过构建集约化、高性能的智算底座，为现代金融风控提供了高效的解决方案。**在实时推理方面，**AI 计算节点凭借其极致的高速互联与并行计算能力，将模型推理延迟稳定压缩至毫秒级别，能够并行处理海量实时交易流，在用户无感知的情况下，同步完成多维度的特征

计算与复杂模型推断，实现对欺诈交易、信用透支等风险的精准拦截与实时决策。**在训练方面，**针对信用风险、市场风险等复杂模型，AI 计算节点提供强大的分布式训练能力，能够将长达数周的传统模型训练周期缩短至数天甚至数小时，快速迭代优化风控策略，应对瞬息万变的市场环境。**在安全保障方面，**AI 计算节点支持构建安全可信的智算环境，通过硬件隔离、加密计算等技术，确保客户敏感数据在计算全过程的安全，满足金融行业的强监管要求。2025 年 2 月，中国太保规划建设的“太平洋保险智算中心”正式落地，是保险行业“算-网-存-云”协同的千亿级全栈 AI 基础设施，“太平洋保险智算中心”基于国产芯片的千卡智算集群，重点支持保险精算、健康管理等场景的大模型训练需求。

**在工业质检领域，**随着高端制造、精密工艺、柔性生产等模式的快速普及，其对缺陷检测的精准度、效率与复杂场景适应性要求也达到了前所未有的高度，推动工业质检向微米级精度与秒级响应演进。AI 计算节点支持构建高吞吐、高精度的视觉智算平台，为智能制造质检提供了有力支撑。**在实时检测方面，**AI 计算节点能够同步处理来自数十至上百个工业相机的高清图像，在秒级甚至毫秒内完成对微小划痕、异色、装配错误等上百种缺陷类型的像素级识别与精准分类，实现对不合格品的实时分拣与告警。**在模型优化与迭代方面，**面对不断出现的新缺陷类型和个性化的产线需求，AI 计算节点提供强大的分布式训练与样本学习能力，能够利用小样本数据在数小时内快速完成新模型的训练或已有模型的优化迭代，极大缩短质检模型从开发到

部署的周期。**在系统稳定性方面，**AI 计算节点具备的容错设计与自动负载均衡机制，能够确保在部分硬件发生故障时，整个质检系统依然稳定运行，满足智能工厂对连续作业与生产零中断的严苛要求。联想为吉利汽车打造 HPC 智算节点集群，通过异构算力平台融合 HPC 与智算能力，支持多种仿真应用，提升研发效率并缩短产品迭代周期。

**在能源领域，**随着新能源占比快速提升、多元负荷复杂接入以及电力市场改革深化，对供需平衡预测的精准性、调度决策的智能化与系统运行的安全性要求不断提升，推动能源调度进入多时空尺度协同与秒级响应的智慧化新阶段。AI 计算节点支持构建融合多源数据、智能预测与优化决策的智慧调度平台，为新型电力系统提供了核心算力支撑。**在功率预测与多时间尺度调度方面，**AI 计算节点依托其强大的时空数据处理与并行计算能力，能够融合数值天气预报、卫星云图、机组运行状态等海量多源数据，实现对未来风电、光伏发电的高精度预测。基于精准的功率预测，调度系统可自动生成实时平衡的调度方案，降低新能源弃电率。**在仿真与 AI 融合决策方面，**面对电网安全约束与经济运行的双重目标，AI 计算节点支持传统物理仿真模型与深度强化学习算法的协同计算，实现安全约束下的经济最优调度，将调度决策效率提升数个数量级。中国电力报数据显示，国家电网公司明确了超节点算力集群将成为超大模型应用的主流支撑。国家电网正依托智算节点设施，推动构建光明电力大模型，围绕智能运维、智能客服、智能调度等领域，推广无人机巡检、变电巡视、营销客服等上百个应用场景。

# 四、AI 计算节点产业生态建设分析

当前，AI 计算节点产业正经历从硬件性能单点竞赛向“硬件-软件-生态”综合体系竞争的关键跃迁。国际竞争围绕生态主导权展开，呈现出技术垄断与开放协作两条路径。国内 AI 计算产业立足自身市场条件与发展需求，探索出一条通过系统级架构创新弥补单点短板，以开放兼容策略融入国际主流，借多主体协同加速构建自主生态的特色发展路径，推动产业生态从“可用”到“好用”的稳健升级。国内外 AI 计算节点产业生态对比如下:

表 1 国内外 AI 计算节点产业生态对比

| 对比维度 | | 国际产业生态 | 国内产业生态 |
|---|---|---|---|
| 核心驱动力 | | 市场机制主导，企业基于商业利益开展技术竞争与生态联盟。 | 政策引导与市场需求结合，在追求商业成功的同时，强调产业链的自主可控与安全可靠。 |
| 生产类主体 | 芯片/硬件 | 格局：“一超多强”，垄断与竞争并存。<br>路径：依托尖端制程与架构创新构建壁垒，同时通过开放联盟不断演进。 | 格局：多路线并行与协同攻关。<br>路径：采取“芯片研发+集群架构创新”系统级策略，弥补单点差距，积极参与国际开放生态。 |
| | 协议/架构 | 依赖尖端制程与架构领先，强调垂直整合与生态锁定。开放路径侧重于接口标准化与互操作性。 | 强调系统级创新与软硬件协同。通过集群架构、互联协议等优化弥补单点短板，采取“自主核心+开放接口”发展模式。 |
| 应用类主体 | | 核心：以云与科技巨头为主导。 | 核心：形成云厂商、运营商多元协同格局。 |

| | | |
|---|---|---|
| 协作模式 | 以市场化联盟为主，合作与竞争并存，动态性强。 | 强调产学研用协同，产业链上下游通过联盟等形式深度合作。 |

来源：IDC，中国信息通信研究院

## （一）国际视角：技术引领与生态开放并行

AI 计算节点作为支撑 AI 训练和推理的核心基础设施，产业生态围绕生产类与应用类两类核心主体协同构建。生产类企业聚焦硬件研发、标准制定与设备供应，筑牢计算节点底层支撑体系。应用类企业侧重计算节点架构下的算力部署、场景落地与服务输出，加速技术商业化转化。

### 1.生产类主体：筑牢技术壁垒、主导标准制定

生产类企业涵盖芯片设计、设备制造及系统集成领域，通过核心技术突破与生态联盟构建，主导计算节点产业底层架构。**芯片领域呈现“一超多强”格局**。英伟达凭借技术先发优势形成垄断地位，其 GB200 NVL72 平台单机柜提供 1440 PFLOPS 算力，NVLink/NVSwitch 互联技术实现极低时延，重点服务万亿参数模型训练需求，构筑起“硬件+软件”的生态护城河。AMD 以 Instinct MI300X 芯片结合 Infinity Fabric 架构构建差异化竞争力，推动开放技术方案。博通则聚焦高带宽交换芯片领域，凭借开源以太网协议产品占据全球市场领先地位，为 RoCE 等主流技术路线提供硬件支撑，保障计算节点规模化部署的网络基础。

**设备与标准层面呈现“技术升级与规则主导”双重竞争态势。在设备侧，**主流厂商提供 AI 计算节点整机柜解决方案，融合高密度计

算模块与液冷技术，破解单机柜较高功耗散热难题，支撑 AI 计算节点规模化落地。**在标准侧，**行业联盟成为标准制定核心力量，由 AMD、英特尔、谷歌、微软等巨头组建的 UALink 推出 1.0 版本，实现跨厂商处理器直接数据传输。UEC 联盟 2024 年发布的超以太网规范 1.0 预览版，在软件 API、拥塞控制等方面实现优化，为 Scale-Out 提供标准化通信栈。2025 年，开放数据中心委员会（ODCC）与 UALink 签署 MOU，双方将围绕 Scale-Up 技术开展技术研讨、测试评估等合作，推动相关标准规范交流与技术应用，助力算力生态的开放协同。

2.应用类主体：强化算力部署，深化场景赋能

应用类企业以云服务商、科技巨头为主，通过大规模部署 AI 计算节点架构，构建算力平台并赋能各类 AI 场景落地。**云服务商方面，**微软 Azure 推出搭载 AMD MI300X 芯片的计算节点实例，为企业用户提供高性能 AI 训练与推理服务。亚马逊 AWS 基于自研 Trainium2 芯片与 EFA 网络架构，打造专属计算节点算力集群，保障电商智能推荐、云原生 AI 应用等场景的算力供给。**科技巨头层面**，Meta、微软等作为 UEC 与 UALink 联盟核心成员，深度参与技术标准制定，同时通过内部 AI 计算节点部署支撑大模型研发、自动驾驶等核心业务，以场景化需求牵引生产类企业技术升级方向。

## （二）国内发展：多主体协同与自主生态构建

国内 AI 计算节点生态形成“生产类企业攻坚核心技术、应用类企业拓展场景市场”的协同格局，运营商、云厂商、硬件企业与科研

机构深度合作，形成以系统级创新对冲单点短板、以开放协同汇聚产业力量的独特路径，助力算力产业从“可用”向“好用”跨越。

### 1.生产类主体：坚持自主创新，推进开放兼容

国内生产类企业以芯片自主化、设备国产化与协议标准化为核心，形成多技术路线并行的产业格局。**芯片领域实现国产化替代突破，集群架构补足性能短板。**面对单芯片算力差距，国内企业采取“芯片研发、架构创新”双轮驱动策略。寒武纪的思元 590、海光信息的深算系列 DCU 等芯片实现性能与能效的双重突破。沐曦通过差异化布局，推出光互连、耀龙 3D Mesh 等多形态 AI 计算节点，形成特色化技术优势。

**互联协议与系统架构的创新是筑牢底层技术支撑、促进产业协同发展的关键环节。当前，国内坚持自主创新与开放合作相结合，已形成“自主研发与开放兼容相互促进、协同演进”的良性发展格局。在自主研发方面**，国内领军科技企业面向大规模算力协同的迫切需求，积极开展核心技术攻关，取得系列重要突破。腾讯依托自身技术积累，在腾讯云平台中通过自研互联优化技术显著提升大规模集群协同效率。**在生态构建方面，**产业界倡导构建开放解耦、分层协同的产业生态。阿里磐久基础设施采用的 ALink 协议支持 UALink 等国际开放标准，实现自主与开放的融合。由腾讯、中国信通院联合快手科技、京东、燧原科技等企业启动的 ETH-X 超节点项目，提出 ETH-X 系统架构理念，并发布《ETH-X Scale Up 互联协议白皮书 V1.0》等多项成果，为 AI 计算节点等关键领域提供标准化指引。同时，由中国信通

院与腾讯牵头，联合合见工软、燧原等企业在 ODCC 成立“AI 网络实验室”，聚焦组网技术创新，突破物理层、数据链路层等瓶颈，构建开放兼容的智算中心网络标准体系与基础测试平台。

### 2.应用类主体：推进规模部署，强化场景落地

**应用类企业以云厂商、运营商为核心，通过计算节点部署构建算力平台，赋能千行百业 AI 转型。云厂商方面，**头部云厂商将 AI 计算节点作为 AI 基础设施建设的核心载体。百度智能云在百舸 AI 计算平台中深度融合昆仑芯，为搜索优化、智能驾驶等核心业务提供高效算力支撑。字节跳动针对万卡集群互联瓶颈，发布自研 EthLink（Ethernet Link）高速互联协议，利用以太网生态优势替代部分私有互联方案。**运营商层面**，依托网络资源与服务优势，构建自主可控的 AI 计算节点算力集群，聚焦重点领域实现精准赋能。中国电信依托天翼云底座，打造 AI 计算节点解决方案，重点聚焦政务服务、智慧医疗、工业互联网等关键领域，通过算力下沉与场景化适配，为地方政府数字化治理、医疗机构智能诊断、制造企业产线升级等提供低时延、高可靠的算力支撑，助力行业 AI 转型落地见效。此外，鹏城实验室等国家级科研机构建设的“鹏城云脑”等超大规模智算设施，广泛服务于科研、政务和产业领域。

## 五、AI 计算节点未来趋势

### （一）政策聚焦自主创新与多维支撑

**当前全球科技竞争日益加剧，算力规模与效率已成为衡量科技竞争力的核心指标之一，**AI 计算节点作为突破智算性能瓶颈、保障算

**力安全的关键技术，战略地位逐步凸显。**为进一步抢占全球智算技术制高点、夯实产业智能化转型基础，我国将进一步聚焦算力以及产业上下游多方协同能力建设，全面推动算力发展质量和应用效率的提升，**在技术引领层面，**国产建设持续加速，针对 AI 计算节点芯片互联、架构设计等关键技术环节，我国将进一步加速发展，依托揭榜行动等方式，推动国际 AI 算力竞争从“单卡性能比拼”向“系统级效率竞争”转变，减少对外部算力技术的依赖。**在产业支持层面，**将重点强化产业链协同与规模化落地能力，将进一步发挥算力产业发展方阵、ODCC 等政产研合作平台，推动 AI 计算节点上下游企业加强合作，完善行业标准，优化产业生态，降低企业应用和部署门槛。同时，鼓励行业龙头企业率先应用 AI 计算节点解决方案，形成可复制的落地经验，为 AI 计算节点的创新发展和产业应用提供全面支撑。

### （二）技术关注高效互联与高密集成

**为支撑大模型的研发与应用，AI 计算节点的技术演进将聚焦于系统性能突破，架构设计、互联协议、资源调度等关键环节的创新迭代将驱动智算性能实现跨越式提升。在架构设计层面，**全柜级 AI 计算节点将实现计算芯片、存储单元与网络组件的深度集成，支持多卡全带宽互联，通过新型高速总线架构适配混合专家模型显著提升大模型训练与推理效率。同时，训推一体架构将逐步普及，通过芯片级技术创新实现训练与推理资源的动态切换，避免算力浪费。**在互联协议层面，**卡间新型高速互联协议将逐步替代传统协议，同时网络传输协议体系将加速成熟，实现多类型数据并行传输，避免协议转换损耗。

光互连技术将进行探索，推动从“铜缆主导”向“全光互联”的升级，通过分布式光交换矩阵实现机柜间的动态连接，减少转发时延，满足 AI 计算节点跨机柜协同的高带宽需求。**在资源调度层面**，AI 驱动的动态调度算法将广泛应用，通过实时感知模型类型、数据规模等负载特征，实现计算、存储、带宽资源的最优分配，针对不同大模型的训练需求自动优化算力配比，缩短训练周期。跨节点协同调度能力将显著增强，通过大规模分布式并行计算技术实现多节点的高效协作，支撑万亿参数级大模型的稳定运行。**在绿色低碳层面，**全液冷技术将成为高密度 AI 计算节点的标配，通过冷板、浸没式等多元方案实现精准散热，结合数字孪生与 AI 运营技术，进一步降低 PUE。

### （三）产业格局头部引领与多方协同

**当前，国内外科技巨头正在积极推动 AI 计算节点技术研发、标准制定和应用推广，产业竞争逐步从技术攻关向生态体系建设与规模化应用迁移，头部引领、生态协同成为 AI 计算节点未来产业发展的基本格局。在市场主体层面，**具备全栈技术能力的云厂商、基础电信企业依托其资本、技术与生态优势，主导大规模智算基础设施的投建与运营，成为 AI 计算节点技术方案的重要用户，牵引先进、成熟技术方案落地。同时，头部云厂商、基础电信企业、芯片厂商、网络设备运营商将进一步加强合作，结合智算训练及推理需求开发并构建 AI 计算节点技术方案。

**在商业模式层面，**为降低用户使用门槛，AI 计算节点的交付形态将从单一的硬件设备，向模型即服务、算力即服务等一体化解决方

案转变。头部企业将通过构建开放平台与标准接口，聚合算力、算法、工具链等关键要素，为用户提供开箱即用的全链路服务。

**在生态构建层面，**为避免多种 AI 计算节点技术方案间，以及各 AI 计算节点方案与各类异构芯片间兼容适配不足等问题，产业发展将重点关注共性标准制定与开放生态建设。在产业联盟、开源社区的引领下，将逐步构建一个接口标准化、组件模块化、软硬件解耦的开放产业生态，硬件互联、软件框架、应用接口等关键领域共性标准建设加快推进，有效降低不同技术路线与芯片架构的集成复杂度与适配成本。同时，通过共建共享基础软件栈、驱动与编译器，有效降低底层硬件差异，打破技术壁垒，使产业竞争从封闭系统间的竞争转向开放平台上的协同创新，最大化汇聚产业力量，加速 AI 计算节点技术的规模化落地与应用创新。

### （四）行业应用试点向全域渗透迈进

**随着大模型技术迭代与多行业智能化需求升级，AI 计算节点应用正逐步从行业试点向全域渗透加速迈进，全面提升算力对产业发展的支撑能力，成为推动数字经济与实体经济深度融合的关键力量。在应用广度上，**继赋能大规模模型训练、高性能计算等场景后，AI 计算节点依托其强大的并行计算与数据处理能力，将加速向金融风控、智能制造、生物医药研发、智慧城市管理等传统行业核心业务系统渗透。通过提供集约化、高效率的算力供给，降低传统企业部署与应用前沿 AI 技术的门槛，成为支撑产业数字化、智能化转型的通用底座，加速算力红利释放。

**在应用深度上，**专业化与场景化定制正成为释放算力价值的关键路径。面对不同行业的独特工作负载与业务需求，通用的算力堆砌模式已难以满足要求。未来，针对垂直行业特性深度优化的“行业 AI 计算节点”解决方案将不断涌现。例如，在生物医药领域，将出现专注于分子动力学模拟与药物虚拟筛选的 AI 计算节点架构。“AI 计算节点算力+行业场景”的应用模式，将推动 AI 算力从资源消耗型投入，转变为直接创造业务价值的核心生产工具。

**在赋能模式层面，**AI 计算节点正从算力供给中心升级为一体化能力输出平台。角色不再局限于提供原始计算能力，而是通过深度融合模型、工具链与行业知识，形成“算力、算法、数据、服务”的一体化赋能范式。通过模型即服务、AI 开发平台等形态，AI 计算节点将封装并输出从模型微调、推理部署到持续学习的全链路能力，使各行各业能够以更低的技术门槛和更短的周期，将前沿 AI 技术转化为实际生产力，实现算力赋能水平的跃升，为经济社会高质量发展筑牢智能底座。

中国信息通信研究院 云计算与数字化研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62300095

传真：010-62304980

网址：www.caict.ac.cn