Global Computing Consortium
全球计算联盟

超节点定义与实践白皮书

鹏城实验室

全球计算联盟

本白皮书由鹏城实验室（PCL）与全球计算联盟（GCC）联合发起、组织编写与发布

牵头单位

鹏城实验室

全球计算联盟

参编单位（按拼音顺序）

安谋科技中国有限公司

百度在线网络技术（北京）有限公司

北京大学

北京神州鲲泰信息技术有限公司

北京沃东天骏信息技术有限公司

哔哩哔哩股份有限公司

得瑞领新（重庆）科技有限责任公司

东莞立讯技术股份有限公司

广东远图未来科技有限公司

杭州劲群科技有限公司

杭州天宽科技有限公司

华南理工大学

华为技术有限公司

京东集团股份有限公司

科大讯飞股份有限公司

浪潮计算机科技有限公司

联想集团有限公司

木犀智能（北京）科技有限公司

鹏城实验室

清华大学

奇异摩尔（上海）半导体技术有限公司

软通计算机有限公司

锐捷网络股份有限公司

上海壁仞科技股份有限公司

上海国创驭算人工智能科技有限公司

上海阡视科技有限公司

上海商汤科技开发有限公司

苏州盛科通信股份有限公司

武汉长江计算科技有限公司

行吟信息科技（上海）有限公司（小红书）

新华三技术有限公司

益思芯科技（上海）有限公司

浙江大学

中国电信股份有限公司研究院

中国科学技术大学

中国移动通信有限公司研究院

中科寒武纪科技股份有限公司

中研益企（北京）信息技术研究院有限公司

编写组成员（按参编单位字母排序）

吴枫、田永鸿、陈文光、李诚、余跃、张士勋、贺志学、杨建坤、王孝国、张子尧、陈锴、武正辉、高洪福、张泽华、齐浩、吕宏、陆汉城、杨博程、陈琼南、罗振、王小泉、邢星、马永昊、杨昊、张晓波、龚徐建、张帆、陆璐、张羽、蒋铭、王贵林、席朝飞、鲍中帅、崔希琳、尹宏伟、李鑫、周韬、刘宜龙、朱斌、包贵新、刘军、李亚军、程旭升、孙鹏、辛博、朱鑫、程秋林、宣善明、代文斌、成伟、王俊杰、易超、吴新豫、陈立波、马瑜、汪新新、郭维锋、吴飞、王则可、王总辉 黄志兰、刘圆、李聪聪、陈恣、于涌、张广彬

序

当前，人工智能正以前所未有的广度和深度影响人类社会的生产与生活方式。万亿甚至十万亿参数模型、超长上下文、多模态融合与世界模型等大规模应用，将算力需求推向指数级增长的新高度。然而，单芯片性能的提升远远无法满足这一需求，依靠简单规模堆叠的传统扩展路径也面临着边际收益递减的困局。由多芯片紧密协同组成“超节点”，进而实现算力基础设施的系统级架构重构，已成为突破现有算力瓶颈的必然选择。

鹏城实验室长期深耕前沿算力基础设施的探索与实践。为筑牢国家算力主权根基，实验室正牵头推进“中国算力网（C^2NET）”建设。在以“鹏城云脑”大科学装置为核心枢纽，持续开展大规模智能计算系统的建设与运行中，我们深刻认识到，重构算力基础设施底层架构，已成为支撑智能计算纵深演进的必然选择。基于这一科学研判，实验室不仅积极推动超节点技术的理论研究与概念体系构建，更率先建设了国内首个采用超节点架构的国产智算集群“鹏城云脑 Ⅲ”，并广泛汇聚产业界在算力互联与体系架构创新上的探索与实践。

在全面深化全国一体化算力网总体布局的背景下，鹏城实验室将围绕算力网的基础理论、关键技术、系统架构和工程实践，规划出版算力网系列白皮书。本白皮书作为该系列的开篇之作，由鹏城实验室与全球计算联盟联合发起，汇聚了科研机构与产业界的研究成果和实践经验，系统阐述了超节点的核心定义、架构特征与底层技术体系，探讨了智算、通算及通智融合等核心场景的实践验证，并构建了从物理层到功能层的总线分层能力框架。此举旨在进一步凝聚产业共识、沉淀工程经验，推动自主可控的技术标准加速形成。

期待本白皮书能够成为深化产业协作的新起点，推动芯片、互联、软件与运维全链条的协同创新，加速超节点技术走向开放化、标准化与生态化。我们愿与

各界同仁一道，为构建自主领先的先进算力基础设施、推动全球人工智能产业高质量发展贡献力量。

鹏城实验室主任　高文

二〇二六年七月

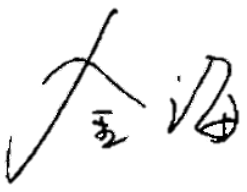
序

当前，人工智能正以前所未有的深度重塑人类社会的生产与生活方式。大模型技术的快速演进，对算力基础设施提出了从“规模堆叠”向“系统级重构”的范式跃迁要求。超节点作为契合 AI 计算特征的系统性架构创新，应运而生。

传统集群架构在应对 AI 大模型“紧耦合、同步并行”的计算特征时，面临通信瓶颈与效率衰减的双重挑战。超节点技术通过构建全局统一内存地址空间、实现内存语义跨节点访问、打造超高带宽与超低时延互联，将分散的物理计算单元融合为逻辑一体的“超级计算机”，从根本上突破了算力扩展的架构层面约束，为 Scaling Law 的持续演进提供了坚实的工程底座。

本白皮书由鹏城实验室、全球计算联盟联合业界科研机构与产业伙伴共同编制，系统阐述了超节点的核心定义、架构特征及技术特征，面向大模型预训练、推理、智能体、科学计算等应用场景进行了深度探索，同时涵盖了智算、通算及通智融合超节点的产业界实践验证，对超节点技术的研究、应用与标准化具有重要参考价值。

超节点正加速推动算力产业从单点硬件竞争迈向系统架构创新。期待本白皮书为“政产学研用”各界提供有益参考，推动超节点技术走向开放与标准化，助力全球人工智能基础设施的高质量可持续发展。



全球计算联盟 金海理事长

二〇二六年七月

序

当前，人工智能正处于从能力涌现迈向规模化应用的关键阶段。以 Scaling Law 为代表的规模化路径仍在持续拓展，多模态、稀疏专家、超长上下文、推理时扩展和智能体等技术加速演进，不断打开大模型能力的边界。与此同时，模型规模、数据规模和 Token 消耗的快速增长，也使算力基础设施面临前所未有的挑战：制程演进趋缓，单芯片性能提升的边际成本不断上升，存储容量、访存带宽、互联效率、能源供给与工程交付日益成为制约智能发展的关键因素。人工智能的持续演进，正在推动计算体系由依赖单点器件突破，转向芯片、互联、存储、软件与工程基础设施协同创新的新阶段。

面对这一趋势，算力竞争的基本单元也在发生深刻变化——从单颗芯片、单台服务器，逐步走向整机柜、超节点乃至数据中心级计算系统。对于我国计算产业而言，正视单芯片能力与国际顶尖水平之间的客观差距固然必要，但更为重要的是认识到：未来算力的竞争，并不只是单颗芯片峰值性能的竞争，更是将大量计算、存储与网络资源高效组织起来的系统能力之争。通过架构创新释放系统潜力，以系统级扩展弥补单点能力的不足，并进一步形成面向具体负载的整体效率优势，是构建先进算力基础设施、提升产业自主创新能力的重要路径。

超节点正是这一计算范式变革的重要载体。它通过高速、低时延、大带宽的互联协议，将数十乃至数百个计算单元紧密连接，在跨物理节点的统一内存编址和内存语义访问基础上，构建逻辑上具备“一台计算机”特征的 Scale-Up 计算系统。超节点是对计算、内存、存储、网络、供电、散热以及系统软件的协同重构。其核心价值，在于打破传统服务器边界，缩短数据移动路径，降低大规模并行中的通信与同步开销，使更多计算单元能够以接近整体化的方式高效协同，为万亿乃至十万亿参数模型、超长上下文、稀疏专家模型、复杂推理与智能体等新型负载提供可持续扩展的算力底座。

架构边界的重构，也必然带来新的复杂性。统一地址空间如何构建，内存语义如何跨节点延伸，互联拓扑与通信协议如何设计，计算、网络和存储如何协同，软件如何感知并利用硬件能力，以及高密度系统如何保障可靠、安全、可运维和可持续演进，都需要产业界形成更加清晰的技术共识与协作框架。超节点要真正从单一产品创新走向广泛产业实践，不仅需要局部技术突破，更需要统一的概念体系、架构语言、能力分级、设计规范和测试评价体系。

本白皮书围绕超节点的核心定义、架构特征、关键技术、设计要求与产业实践展开系统论述，力求厘清超节点“是什么、为什么需要、如何设计、如何应用”，并通过智算、通算及通智融合等场景的实践，呈现其技术价值与演进方向。白皮书的发布，有助于统一产业认知、降低应用与开发复杂度、沉淀可复制的工程经验，促进芯片、系统、软件、网络、存储及基础设施等产业环节协同创新。

期待本白皮书成为产业交流与技术协作的新起点，推动超节点走向开放化、标准化和生态化，加快形成可扩展、可演进、可持续的先进计算体系，为大模型技术持续突破和人工智能产业高质量发展奠定坚实基础。

美团　谢语宸

二〇二六年七月

目录

引言：问题与挑战

AI“奇点”临近的趋势，正对 AI 编程与智能体领域产生颠覆性影响，同时对底层大模型与算力基础设施提出了前所未有的新需求和挑战。AI 智能体的能力表现，直接决定了其完成复杂任务的质量与效率，已成为该领域竞争的核心。而实现能力跃升，又依赖于三大技术基础：更大规模的 MoE 模型、更长的上下文序列、更低的推理时延。以 DeepSeek V4 为标志，技术边界正被快速推高——模型总规模达 1.6T，集成 384 个 MoE 专家，支持 1M 超长序列，并追求 10ms 级别的 token 输出时延，且演进速度持续加快。

面对由此带来的 Token 需求指数级增长，基础设施核心矛盾在于：必须在支持更大模型、更长序列、更低时延的同时，显著降低推理成本、提升算力效率。这已成为支撑 AI 智能体持续高速发展的关键瓶颈。而“超节点”架构，正是解决这一系列矛盾的核心答案。

“并行”与“批处理”是提升大模型计算性能与效率的两大支柱。“更大”的并行规模，是支持更大模型、实现更低时延、降低单卡内存压力的基础，也为“更大”的批处理规模创造了条件。而更大的批处理规模，直接决定了 xPU 等算力单元的利用率。因此，通过超节点实现极大的并行规模，是同时攻克“规模、长度、时延”三角挑战，并保持高算力利用率的关键路径。

然而，控制通信时延是并行计算能否成功扩展（Scale-Up）的决定性因素。若通信开销无法被有效掩盖，并行带来的收益将被抵消甚至适得其反。超节点架构通过内存语义互联、超高带宽、超低时延等关键总线技术，将节点内通信损耗降至最低，从而使计算任务得以在大规模并行中实现近乎线性的性能提升，让真正的极致扩展成为可能。

第一章　超节点发展趋势和影响

过去，AI 基础设施竞争主要围绕 xPU、CPU、HBM、网卡等单点硬件展开；然而目前，行业重心正逐步转向“整机柜即计算机”的系统级架构。将多个 xPU 通过高速互联构成单一高带宽域，互联总带宽可达很高水平；全液冷机柜级架构进一步强化了整机柜交付和整机柜优化能力。

在这一演进过程中，超节点的核心任务不再是简单叠加更多 xPU，而是围绕 xPU 间通信、HBM 容量、机柜功率密度、散热效率和可维护性等系统瓶颈进行协同优化。随着大模型进入 MoE、长上下文、多模态、推理时扩展和 Agentic AI 阶段，单 xPU 性能提升已经难以单独支撑训练和推理效率的持续增长，高带宽、低时延的跨 xPU 互联以及整机柜级资源编排成为关键基础。

因此，超节点更接近一种 AI 数据中心系统架构，而非单一硬件产品。其设计目标是将 xPU、CPU、网络、电源、冷却、机柜和软件调度纳入统一工程体系，使更多加速器在通信、散热、功耗和运维层面形成接近整体化的计算单元。

1.1 超节点的发展阶段

第一阶段：单机多卡，重点是服务器内部互联

该阶段以 8 xPU 服务器为典型形态，关键设计集中在单机内互联拓扑以及风冷/液冷散热能力。此时，市场竞争的主要对象仍是服务器节点本身。随着模型参数、激活数据、KV Cache 和 MoE 路由规模快速增长，8 卡服务器在更大模型训练和高吞吐推理中的资源粒度优势逐渐减弱。跨服务器通信通常基于以太网的 RoCE 等 Scale-Out 网络，网络带宽、尾延迟和拥塞控制会直接影响训练效率与集群利用率。

第二阶段：整机柜超节点，重点是 rack-scale 计算域

整机柜超节点已经从概念验证进入生产应用阶段。其主要价值在于将部分 xPU 集群通信从 scale-out 网络中前移至更高带宽、更低延迟的 rack-scale 互联域，从而提升大规模训练和推理的系统效率。

第三阶段：Pod 级和多机柜超节点，重点是几百 xPU 的统一域

某些新一代整机柜超节点系统进一步将整机柜超节点通过高速互联连接为机柜级系统。从发展方向看，未来 AI 算力的最小采购、交付与调优单位将不再是单台服务器，而是液冷机柜、Pod、超节点。未来设计会将计算、网络、存储、电源、冷却和设施控制纳入超节点整体系统框架。

第四阶段：超节点集群化，重点是数据中心级拓扑

超节点集群化，即通过横向扩展（Scale-Out）将多个超节点聚合为统一算力资源池，以突破单超节点物理上限。从发展方向看，数据中心级拓扑成为关键——它从全局视角设计网络层级、路由策略与流量调度，旨在构建无阻塞、低跳数、高对称性的互联架构，确保集群在万卡规模下仍保持近似本地访问的通信效率，真正实现“池化”而非“堆叠”的可扩展性。与此同时，物理规模、功耗、调度复杂度、通信效率、软件生态和故障恢复能力也将成为更突出的工程挑战。

1.2 产业影响

对服务器/OEM/ODM：从“卖服务器”转向“交付整机柜系统”

未来高端 AI 服务器厂商的能力边界将从主板设计和整机装配延伸至整机柜级系统交付，重点包括：整机柜级热设计；xPU/网卡/CPU 拓扑设计；液冷歧管、快接头和冷板维护；机柜级 burn-in 和压力测试；整柜运输、吊装、接液、接电和接网；故障托盘快速更换。这将推动传统服务器制造向机架级验证和整柜交付能力升级。具备整机柜验证、交付和维护能力的厂商，更有可能进入 AI 超节点核心供应链。

对液冷企业：从“散热方案”转向“基础设施平台”

液冷厂商的机会主要集中在以下方面：冷板设计与均温能力；快接头与盲插可靠性；歧管压降与流量均衡；CDU 冗余与控制策略；冷却液兼容性、绝缘性、腐蚀性和维护周期；液冷监控、泄漏检测和预测维护；液冷系统相关设计规范的标准化适配。当前，行业内正在推动冷板、通用快接头、盲插快接头、冷却液及 CDU 等关键部件的接口与性能规范统一，液冷产业有望从项目制方案逐步走向标准化供应链。

对电源企业：高压直流和机柜级供电成为新赛道

AI 超节点将带动以下方向发展：800Vdc/HVDC；高功率电源架；侧边电源架构；液冷母排；高密连接器；机柜级储能缓冲；供电遥测与动态功率管理。随着相关能力从数据中心设施层下沉至机柜层，电源厂商、机柜厂商和整机系统厂商之间的边界将进一步交叉。

对网络和光互联企业：网络墙成为下一阶段瓶颈

当单个超节点内部 xPU 数量持续提升后，多个超节点之间的 scale-out 网络将成为影响集群效率的关键环节。业内对 AI 网络瓶颈的关注日益增强，未来性能约束不仅在于链路速率，还包括交换芯片 radix、网络拓扑、拥塞控制以及 MoE 通信模式等方面。这一趋势将继续推动 800G/1.6T、硅光、CPO、LPO、高 radix switch、智能网卡、低延迟 Ethernet 等方向发展。

对运维厂商：从传统单节点维护升级为整机柜级、集群级的统一运维

运维体系应围绕“高可靠、易维护、可观测、可扩展、安全运行”建立，智能化运维平台与标准化诊断工具链是关键。在基础架构上，应支持模块化供电、AB 路供电切换，以及从风冷、局部液冷向全液冷架构演进，满足高功耗 AI 集群的扩展需求。在可靠性方面，应重点保障液冷管路承压密封、冷板防堵防腐、流量均衡、漏液检测与自动告警，降低高密度运行风险。在维护性方面，通过电、液、网盲插接口、快换接头、助力把手、标准化理线和充足维护空间，提升现场维护效率，缩短故障恢复时间。同时，应建设整机柜级统一监控平台，实时采集各类运行状态，结合日志留存、告警编码、远程升级和预测性维护，实现故障快速定位和风险提前预警。持续演进的智能化运维体系围绕供电、散热、液冷、网络、固件和安全建立统一闭环。

1.3 总结

超节点技术的本质是推动 AI 数据中心从 “服务器堆叠” 模式向机柜级、Pod 级、超节点系统工程模式转型。它将 xPU、CPU、HBM、交换、供电、液冷、机柜、运维及软件调度系统统一为一个整体设计对象，实现了 AI 基础设施架构的根本性重构。

行业正在致力于将超节点这一复杂系统进行开放化、模块化和标准化。未来 2-3 年，全球 AI 基础设施竞争的主线将从 “单点 xPU 性能比拼” 升级为 “系统级工程能力竞

争”。谁能够以更低的功耗、更高的可靠性、更低的维护成本，将大量 xPU 高效组织为可交付、可维护、可扩展的超节点系统，谁就将在新一轮 AI 基础设施竞争中占据主导地位。

第二章　超节点的核心定义与特征

超节点是一种在物理上由多个计算节点通过高效的互联协议紧密连接组成，具备跨物理

节点的统一内存编址能力，逻辑上具备“一台计算机”特征的计算系统。

超节点的架构特征为：具备跨物理节点的统一内存编址。

基于此架构特征，超节点有三大核心技术特征，包括：“内存语义”和“统一内存编址”、超低时延、超大带宽。这些特征为 AI 大模型计算等“单体紧耦合”软件构建了一个具备统一地址规划与数据布局的计算环境，使所有处理单元能够实现高效协同执行，从而为大规模、高性能计算场景提供近乎线性的算力扩展能力。

2.1 “Scale-Up”计算系统架构

“统一内存地址空间” 与“超节点总线”共同定义了超节点计算系统的“统一执行环境”，这是超节点“Scale-Up”计算系统架构的核心特征。“Scale-Up”计算系统是一种垂直扩展的紧耦合架构，该架构的本质，在于为“紧耦合型单体并行软件”提供一种可扩展的算力承载机制——即在维持高 MFU（芯片算力利用率）的前提下，通过系统级的地址统一、Load/Store 内存语义操作与访存直通，实现有效算力规模的增长。

从系统结构视角看，有效算力规模的扩展直接支撑了模型容量与计算深度的增长，具体体现为模型参数规模、思维链长度及上下文序列长度的系统性提升。同时，超节点架构通过集群内节点的同步并行执行机制，在指令级和线程级上消除了传统跨节点通信的同步开销，从而在系统层面保证了低时延的推理响应能力。

在模型结构与算法演进方面，稀疏模型与模型记忆系统对全局共享数据提出了更高频、更大范围的访问需求。超节点提供的统一内存地址空间定义与直接内存访问（DMA）能力，使得模型参数与记忆条目可以在无软件地址转换的情况下被计算单元直接引用，从而在系统结构上为稀疏检索和记忆重用提供了最高效的数据通路，进而同时提升智能表现、算力利用效率和推理吞吐。

如果采用“Scale-Out”计算系统架构来承载上述计算负载，其系统结构上的根本约束在于：各计算节点运行于相互隔离的独立地址空间之中，节点间的数据同步须依赖软件协议栈和网络报文交换，而非硬件直通的访存操作。即便为该类集群配置较高的链路带宽与较低

的传输时延，其数据移动路径中大都不可避免地址映射、消息组装、中断处理等软件介入层次。与基于“统一执行环境”的直接访存模式相比，该方式在有效带宽、端到端时延及同步效率上存在数倍乃至一个数量级的结构性劣势。该劣势将导致计算资源在通信等待中闲置，有效算力规模被 I/O 瓶颈所钳制，系统整体的可扩展性在架构层面即受到根本限制。

2.2 “内存语义”跨节点内存访问

现代处理器（CPU、NPU、GPU）均遵循冯·诺依曼计算架构，其指令集主要分为：数据存取指令（以地址为操作对象）、数据计算指令、程序控制指令及系统管理指令等。其中，数据存取指令承载“内存语义”操作，核心指令为 Load 和 Store，由处理器 Core 的硬件单元直接发起对指定地址的读写，是处理器指令集与微架构的固有组成部分。在编译器配合下，软件程序的数据访问模式被映射为这些内存语义指令。

超节点通过系统总线对内存语义操作的可达域进行扩展，使得任一处理器 Core 能够直接对超节点内内存地址执行 Load/Store 操作。该扩展伴随统一内存地址空间的建立，从而在编程模型和执行模型上获得与单系统一致的语义。

在此结构下，跨处理器 Core 的数据同步在机制层面与芯片内部一致，均表现为内存语义访存。这种统一性允许系统充分利用处理器微架构中的预取、乱序执行、多硬件线程等机制，由硬件、编译器和程序算法设计协同完成访存调度与同步隐藏，而无需在软件层引入额外的格式编码或协议栈。

反观不支持内存语义的节点间互联，其数据交换必须依赖软件对信息进行 TLV、KV 等格式封装，并通过通信协议栈收发消息。由于此类路径无法融入处理器的微架构优化流程，其有效带宽和延迟均受限于软件开销，且无法利用硬件自动预取和乱序执行等机制来掩盖通信延迟。

统一内存编址是实现跨节点内存语义访问的前提——只有当不同节点的内存被纳入统一编址，Load/Store 等内存语义操作才能够访问远端内存地址。因此，统一内存编址与内存语义访问构成超节点的核心架构特征。缺失这些特征，即便互联链路提供高带宽和低时延，系统在并行计算中的性能与延迟收益仍将显著受限，且该制约随单芯片性能提升而加剧——因为处理器 Core 执行速度越快，软件通信路径的相对开销越突出。

2.3 超低时延

在提供大模型推理服务时，E2E 推理响应时延是用户体验的关键，让人可以流畅的和 AI 大模型交流，自动执行任务可以更快的执行完成，实时智能系统才能成为可能。超低时延互联，是超节点支持大模型低时延推理的关键之一。

同时，由于 AI 大模型同步并行的计算模式，节点间同步时延如果不能被计算时延掩盖，计算执行将被阻塞等待，导致算力效率下降。而且，由于大模型一次完整的前向计算会通过几十层神经网络、每层多次的节点间数据同步，每次同步包含多次的数据发送、接收确认和同步控制。导致微小的总线 RTT 时延增加，都会被放大成百上千倍，影响大模型 E2E 推理时延和计算效率。

因此，在具备“统一内存编址”和“内存语义访问”核心架构后，超低时延总线也是超节点的关键特征。相较于传统服务器架构，超节点 RTT 通信时延降至微秒级，降低了 50% 以上。

2.4 超大带宽

AI 大模型的激活值通常为具有成千上万个隐藏维的高维向量，相比原始字符输入，其数据量被放大成千上万倍；加之模型由数十至上百层神经网络堆叠而成，一次完整的前向计算相比初始输入，整体数据负载可增加数十万乃至上百万倍。这在多节点并行计算中引发了巨大的跨节点同步数据开销。

因此，超节点总线必须具备 TB/s 级别的高带宽，并能随算力增长持续提升，才能有效承载这些海量的同步流量，从而确保超节点在运行 AI 大模型时的性能优势。否则，跨节点通信开销将严重抵消并行加速收益，使得超节点的实际性能提升大幅受限。

2.5 多算力协同

超节点旨在为计算任务（尤其是并行计算任务）提供系统级架构支撑，其总线互联的核心对象是参与计算的处理器单元（xPU），而非限定于 NPU。CPU 及其附属内存、存储资源会纳入超节点互联域并参与并行计算过程。

超节点的关键架构特征在于：通过超节点总线，将 xPU、CPU、HBM、DDR、存储等

多种算力相关资源互联，并在此互联之上构建统一的地址空间与内存语义访问机制，使得任一互联资源均可被超节点内的计算单元直接、共享地访问，无需软件地址转换或消息封装。

基于上述架构，超节点实现了全局内存池化：任一计算节点可通过标准 Load/Store 指令直接访问远端内存，访问延迟与带宽接近本地内存，且编程模型与单节点一致。这种对等内存访问为系统提供了“Scale-Up”式的架构能力增强，包括跨节点的内存弹性借用、任务状态的快速迁移，以及多节点间低开销的同步与数据交互——这些能力均由硬件和系统底层原生支持，无需依赖上层软件协议。

该架构同样适用于通用计算领域，打破传统服务器节点边界，为以数据为中心的通算应用提供统一内存访问基础，也为通智融合场景提供统一的算力资源池抽象。超节点架构本质上构建了一个跨多种处理器和存储层次的单一执行环境，其核心特性在于地址统一、访存直通和资源池化，而非针对特定应用场景的定制优化。

2.6 超节点规模

规模无疑是超节点的关键特征之一，更大的超节点规模显然让 AI 大模型部署和优化有了更强的灵活性。超节点规模在考虑 AI 大模型并行计算的基本需求后，还要考虑可靠性冗余、灵活部署与调度、模型升级、应用场景变化以及多网合一等诸多因素。

更大规模的超节点，可以让碎片化资源占比更小；按照不同的工作负载模型灵活调整算力配比；在 AI 大模型的参数规模和专家规模变大时，可以通过更大规模的并行，满足运行要求；在应用场景对时延有更高要求时，可以通过扩大并行组规模来满足。

更大的超节点规模还可以更好的在超节点内部署 CPU 资源、DDR 内存资源和存储资源，在 AI 大模型不断向存算协同、通智协同发展的趋势下，持续满足未来的发展要求。具体的超节点部署规模，需要在实际应用中，根据具体的业务述求合理选择。

2.7 软硬件协同

和云计算架构“Scale-Out”倡导和严格遵循的软硬件解耦不同，超节点和 AI 大模型计算是存在软硬件耦合的，模型的开发和部署调度需要根据超节点的互连拓扑、总线协议和性能规格，进行针对性的优化才能充分发挥超节点的性能优势。

例如：部署 DeepSeek R1/V3 Decoding 推理集群。DeepSeek R1/V3 具有 256 个路由专家、并每次激活其中 8 个。当部署在 256 个节点构成的超节点上，就具备了最大并行规模，实现了极低时延的支持。而考虑到共享专家和故障冗余，还需要额外增加 32 个计算节点到超节点中。而当追求低时延和复杂度的平衡时，128/64 节点的超节点部署会是更好的选择。

所有这些部署规划需要在超节点集群硬件和互联拓扑设计、AI 大模型并行部署和调度算法设计、以及相应的集合通信算法设计等软硬件各环节协同考虑，并通过 PyTorch、vLLM 等框架加以支持，才能针对不同的场景和需求充分发挥超节点的性能优势。

第三章　超节点架构

超节点以总线技术为核心，通过一系列芯片、硬件、软件、算法技术组合，最终实现超大规模、高性能的“Scale-Up”计算系统。

3.1 总线与协议概述

超节点总线定义了超节点设备模型、操作模型、互联规模等关键技术要素，是构建超节点的基础。超节点设备模型中“超节点统一内存地址空间”定义、“地址空间规格”决定了超节点的构建方式和能力；功能层提供异步编程模型和 Load/Store 同步编程模型确定了超节点中总线的操作方式和实现机制；“Entity”规格确定了超节点支持的最大互联规模。

超节点协议定义了超节点总线的实现算法和互操作方式，是超节点各硬件单元按统一标准，实现超节点总线机制的基础。

3.2 硬件集群

超节点超大规模、高性能的基础是硬件集群技术，需要尽可能通过液冷、高功率供电、高密布局布线等硬件技术实现高密度的算力部署，并要通过电、光互联的方式把各计算单元高速互联起来，使超节点高算力性能的目标真正可以落地实现。

3.3 互联拓扑与交换

超节点的互联拓扑和硬件集群形态、互联规模密切相关，可以采用 Full Mesh、CLos、或 Full Mesh&Clos 混合方式，恰当的互联拓扑是发挥超节点性能的关键。交换技术决定了超节点的规模，在支持大容量超节点时，还必须通过超大容量的交换单元实现超节点总线的扩展。

3.4 芯片（含算力、存储、HBM 芯片）

超节点要构建“Scale-Up”算力集群，需要算力芯片在指令层面实现超节点总线的跨节点操作，并和算力芯片的调度、cache、预取、乱序执行、同步等微架构机制深度融合和优化，才能真正实现超节点内所有计算单元在“一个单一系统”中、“同步并行”高效计算的目标。否则，超节点中的跨节点操作将充满“断点”，宝贵算力资源将会白白浪费，E2E 业务时延将大大提升。

在超节点中，NPU/GPU 等加速器主要承担张量、向量与专用算子的高吞吐计算，CPU 与系统软件则构成控制面与管理面的支点。超节点的系统价值不仅来自高带宽、低时延互联，

也来自系统层面对计算、内存、I/O、虚拟化、安全与管理资源的统一编排能力。随着计算重心从处理器峰值性能转向数据流动效率、资源利用率、服务质量与平台总拥有成本，系统级性能越来越取决于内存层次、I/O 路径、虚拟化开销与数据搬移效率。综合来看，CPU 承担通用计算、控制面执行、系统软件承载、资源编排、虚拟化隔离与平台管理。

3.5 算子与算法

超节点域内数据交互通过集合通信库、共享内存库、点对点通信库等为上层应用封装好发送、接收、复制、节点间进程同步、内存访问等基础操作，对外呈现为一组或一类算子。同时，在 AI 大模型框架层，也需要相应的 EP、TP、CP 等并行算法。使 AI 大模型应用能更便捷的运行在超节点上、利用超节点大规模的算力优势。

3.6 超节点软件

超节点系统软件对硬件能力进行抽象与封装，通过拓扑感知、故障切换、负载均衡等机制实现算力资源池化，并为应用程序提供内存对等访问、统一编址与亲和性调度等核心能力。这些能力被封装为高阶算子库（如集合通信、共享内存库）和底层操作接口，供不同层次开发者使用。

普通开发者通过集成高阶算子库，即可便捷利用超节点的性能优势。资深开发者则可基于提供的内存语义访问、RDMA 等底层接口，实现通信与计算的深度融合，定制高性能算子，完成性能优化。

此外，超节点通过运行时框架、管控系统及配套工具链，支持应用在超节点集群上的快速部署与高效执行。

3.7 超节点设备管控

超节点是作为整体运行的，良好的设备管控可以让超节点中所有部件获得一致、有序的运行环境，并灵活应对故障、变更、扩容、升级的各自日常维护和操作。

3.8 可靠性与可服务性

作为一个紧耦合的大规模计算系统，整个超节点范围内的部件规模大，单个器件或节点

的故障如何做到不扩散到整个超节点，如何保证 RAS 是一个不容忽视的挑战。这需要创新超节点硬件架构设计，例如无线缆正交架构，将计算节点与交换节点通过背板直接对插，可将因线缆连接导致的系统故障率大幅降低。同时需要在超节点中实现全面、完善的故障检测、冗余恢复机制，降低单个部件故障对系统的影响，例如，通过数据重传、器件冗余、链路冗余、资源负荷分担等机制避免单个部件的故障被传播到整个超节点。

针对超节点，需要通过预防、检测、隔离、恢复的分层防护机制，把故障影响控制在可度量、可恢复的范围内。这里的故障包含了业务中断、性能裂化等各类问题，相关的度量指标包括 MTBF、MTTR、可用度、MFU、吞吐量、时延等，超节点的可靠性指标范围及要求与智算集群保持一致。

预防包括超节点的硬件、拓扑、连通性、带宽、时延等的上线前检查和例行巡检；检测包括硬件、软件类异常的检测和 SLA 指标裂化的检测；隔离主要是把故障影响隔离在超节点内，并将故障单元从作业中隔离出来；恢复包括数据重传、器件冗余、链路冗余、资源负荷分担、业务重调度等机制，恢复机制需要考虑超节点的亲和性调度。

3.9 系统协同

超节点是 AI Infra 的全新构建单元，然而仍然需要与 AI Infra 的其他存储、网络协同一致，因此，超节点的架构设计需要与存储系统、网络扩展系统协同设计。具有“算存网”协同设计的整体产品解决方案才能构建高质量的算力基础设施。

第四章　超节点价值应用场景

在面向人工智能计算与通用计算领域的 10 大核心业务场景以及行业场景，如大模型预训练、推理、后训练与强化学习、多模态内容理解与生成、Agentic AI、虚拟化、大数据、数据库、分布式存储和高性能计算等，超节点均可提供领先的系统能力，带来计算业务性能和资源利用率提升。

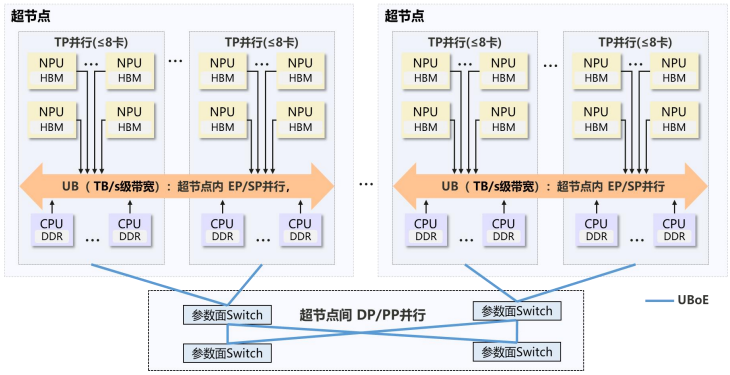
4.1 大模型预训练

为了提升大模型性能，业界主流基础大模型训练的参数量和数据量快速增长，从早期的数十亿参数发展到如今的万亿级别，大模型预训练的业务负载呈现高并行特征，单一训练任务横跨整个算力中心的所有节点，随着预训练所需要的算力规模持续增长，对计算基础设施高效并行提出挑战：

模型参数增长需要多卡并行部署，带来大量集合通信开销。以 GPT4 1.8T 为例（基于行业数据评估），显存超 10TB，采用 TP8/PP16/EP8/SP4/DP64@seq8k/bs8k 时，单次迭代每卡 TP 通信 336GB、EP 268GB、DP 86GB，对网络时延和带宽要求极高。

针对上述瓶颈，超节点架构通过 NPU 间大带宽低时延互联，有效降低并行通信开销，并在 MoE 模型及长序列输入等复杂场景中验证其性能增益，如图 4-1 所示。在 DeepSeek V3 训练中（PP8/DP64/EP64 配置），超节点集群相较传统服务器集群，未掩盖通信比例降低 80%；在单 NPU 裸算力提升 2 倍的条件下，系统训练性能实现 3 倍以上提升。

图 4-1 大模型预训练超节点参考架构



4.2 推理

大语言模型正向万亿参数稀疏化演进，总参数量与专家数量持续增长，推理模式从单卡单机走向多机大专家并行，推理序列长度不断扩展（DeepSeekV4 以支持 1M 级别），未来多模态应用将推动序列长度进入 1M 甚至 10M 级别。因此，算力中心推理面临如下挑战：

多专家低时延通信挑战。多机大专家并行推理中，小数据高频通信占比高。以 DeepSeek

V3 为例，每 Token 推理需经 58 层 MoE、58 轮跨节点 Dispatch 与 Combine 动态通信，最小粒度仅 7KB。推理服务需在 TPOT≤10ms 约束下尽量增大 Batch Size，通信比例须控制在 20%以内，单次通信耗时需<10us。

长序列 KV Cache 缓存挑战。推理过程需保存用户全量 KV Cache 并卸载至内存或 NAND Flash，以提升 Batch Size 及吞吐。基于 DeepSeek V3 测算，64~128K 序列下 KV Cache 卸载可提升 4~6 倍 Batch Size 及 30%~40%推理吞吐。

超节点架构通过 NPU-to-NPU 大带宽低时延互联与高效统一语义能力，实现高性能 Dispatch/Combine 通信，显著降低多专家通信时延。同时，基于全局资源池化架构提供多层次 KV Cache 卸载与加载能力，以查代算提升计算吞吐，有效应对大专家并行与长序列带来的推理性能瓶颈，如图 4-2 所示。

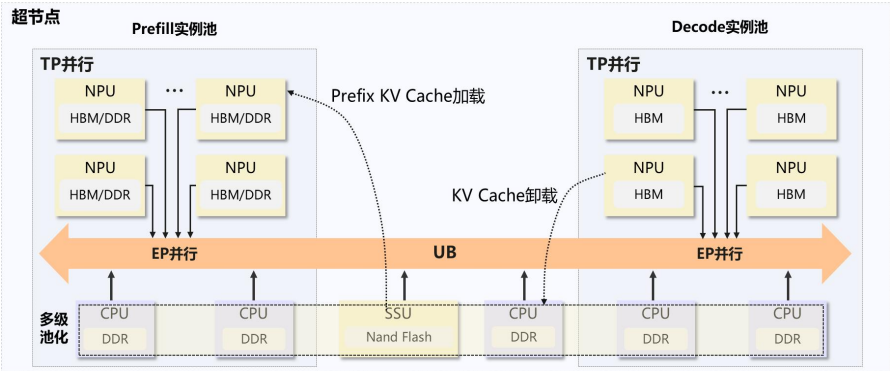


图 4-2 推理超节点参考架构

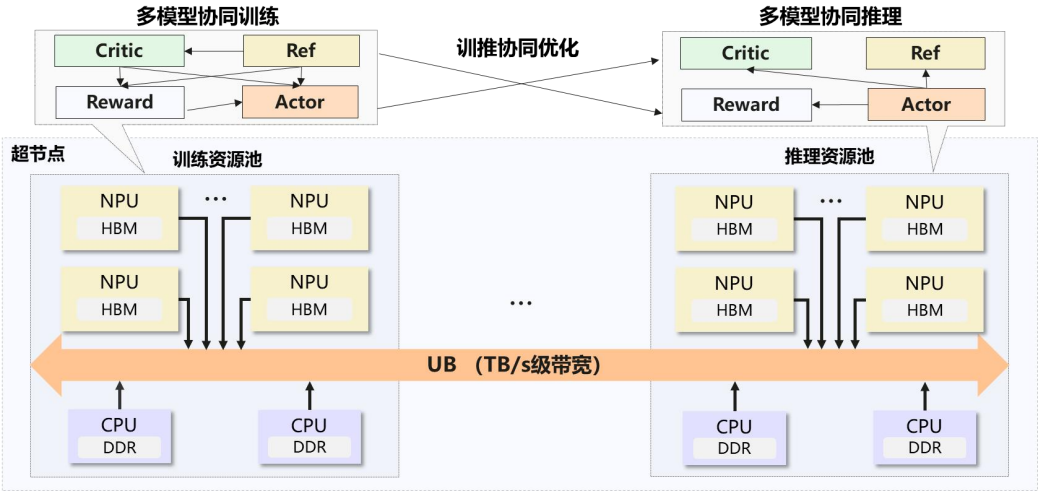
4.3 后训练与强化学习

大模型强化学习训练（经典流程）涵盖 SFT、RM 和 PPO 三个阶段，需同时维护策略、价值、奖励及参考模型，总参数量达数万亿级，计算与通信挑战随模型规模扩大日趋严峻。

大量模型参数与梯度同步通信挑战。PPO 阶段需在 Actor、Critic、Reward、Reference 间频繁同步参数与梯度，以 GPT3 为例，175B 策略模型与 6.7B 奖励模型，单次迭代通信超 500GB，传统集群带宽瓶颈致训练效率下降 40%~60%。大批量样本生成与并行推理性能挑战。 为每提示生成数十个候选响应（总数达数万级），要求高效并行推理，通信时延直接影响收敛稳定性。

超节点架构基于 NPU-to-NPU 大带宽互联，将模型传输时延从数十秒级降至毫秒级，并支持数百卡并行推理，承载更大 Batch Size。在 70B 模型 RLHF 训练中（Batch Size 32K），推理时延由数十秒降至秒级，收敛速度与样本生成效率均获倍级提升，如图 4-3 所示。

图 4-3 强化学习超节点参考架构



4.4 多模态内容理解与生成

多模态模型向 T 级参数量、稀疏模型架构演进，负载从单一模型扩展至多模型多任务，不同任务对算力与访存需求差异显著，对算力基础设施在通信带宽、异构调度与实时同步等方面提出更高要求。

分布式推理通信挑战：序列并行与专家并行需百 GB/s 级多机带宽，当前网络难以满足。

异构部署与多任务并发挑战：Token 级模态分组与 Re-order 要求动态分配计算量，大量小包通信对时延极为敏感。

高性能 3DGS 重建挑战：千万级高斯球场景下，单轮迭代需压缩至数十毫秒，同步百万级椭球梯度（位置/协方差/透明度）通信需求达百 GB/s。

超节点架构将多个 xPU 组成逻辑资源池，提升分布式 3DGS 重建效率；支持不同类型 NPU 异构组网，增强资源动态调度能力；基于内存语义通信实现小包超低时延，支撑 Token 级分组与 Re-order，显著提升多模态推理与 3D 渲染生成效率，如图 4-4 所示。

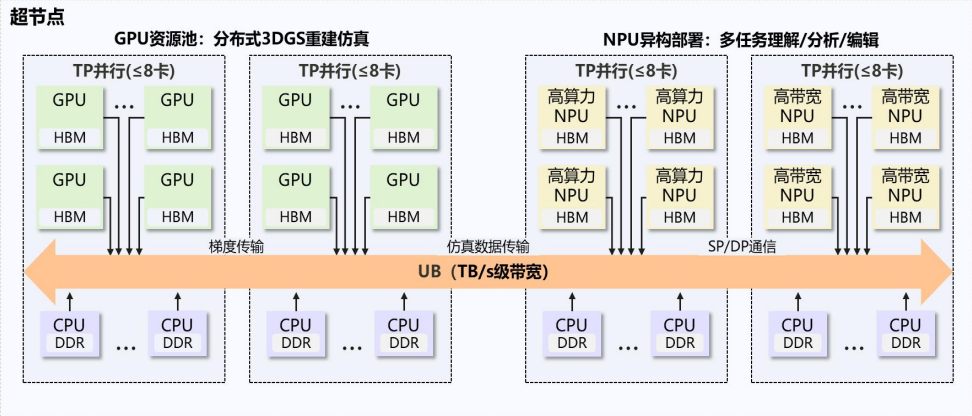


图 4-4 多模态理解与生成超节点参考架构

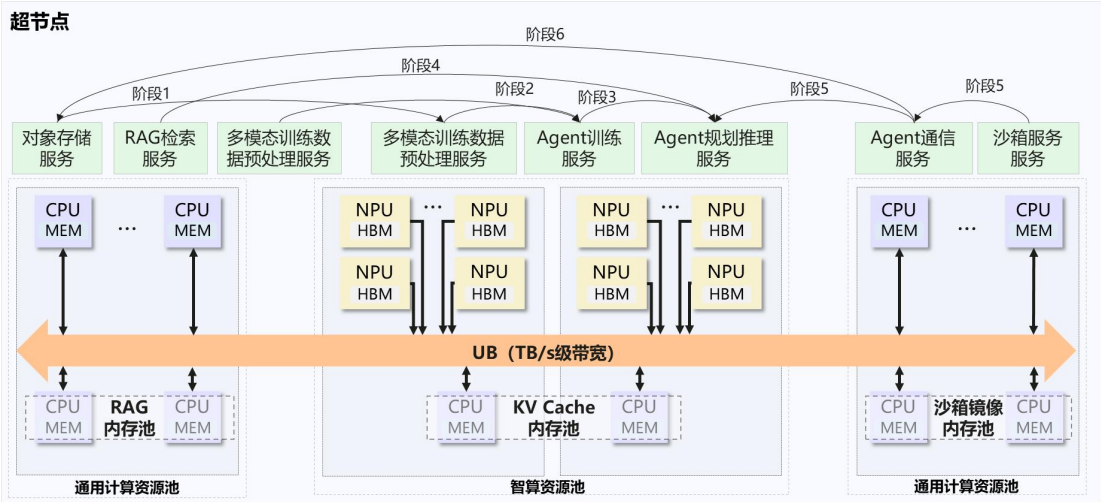
4.5 Agentic AI

Agentic AI 已成为 AI 领域关键趋势，预计到 2028 年，全球 15%的日常工作决策将由自主 AI 代理完成。AI 正从“被动响应”转向“主动行动”，融合多模态感知与物理世界交互能力，并与互联网结合形成分布式智能体网络，支撑大规模协作。上述趋势对算力基础设施带来全新挑战。

业务负载层面挑战： Agentic AI 结合训练、推理、通用计算与数据库管理四个核心环节，多 LLM 协作加剧系统复杂度。除既有训练推理挑战外，长链条推理与长期记忆引发存储及访存瓶颈，长尾任务与异构硬件阻塞降低资源利用率，多任务多副本导致频繁上下文切换与镜像拉取，开销显著增加。

系统架构层面挑战：负载模式的升级换代带来内存需求激增、资源调度复杂化、通信随机性增加及存储持久化管理难题，系统须在多样化动态环境中确保高效内存使用、精确调度与持续数据存储，满足长期运行需求。

超节点架构提支持多样性算力平等协同，实现多模态数据高吞吐预处理、RAG 检索及慢思考推理优化；全局资源池化支持热迁移，应对 KV Cache 与 RAG 数据指数级增长；高速互联架构实现沙箱镜像毫秒级热加载，提升 Agentic AI 负载亲和性，为分布式智能体网络构建高效算力底座。



Agentic AI工作6阶段：阶段1-数据加载；阶段2-预处理；阶段3：同步参数；阶段4-RAG检索；阶段5-Reflection；阶段6-agent探索轨迹收集&存储

图 4-5 Agentic AI 超节点参考架构

互联总线为 Agentic AI 提供了芯片级的高速互联底座，除了具备智算超节点与通算超节点价值外，可以将通算与智算资源紧耦合起来服务于 Agentic AI。 SSU/内存具备物理池化能力，可灵活管理与分配，发挥总线高带宽低时延对等互联能力与 xPU 数据面直通，极大提升 Agentic AI 资源利用率与 Agentic AI 吞吐性能；融合超节点具备灵活配比的 SSU 池与内存池等，实现容量时延等 SLA 按需的 Agent 知识库、记忆上下文、状态及过程输出、检索与更新；NPU Direct SSU 提供比 DDR 成本更低的记忆方案选择，满足不同场景（时延和带宽要求、KVCache 换入换出）的低成本低时延 Agentic AI 方案，如图 4-6 所示。

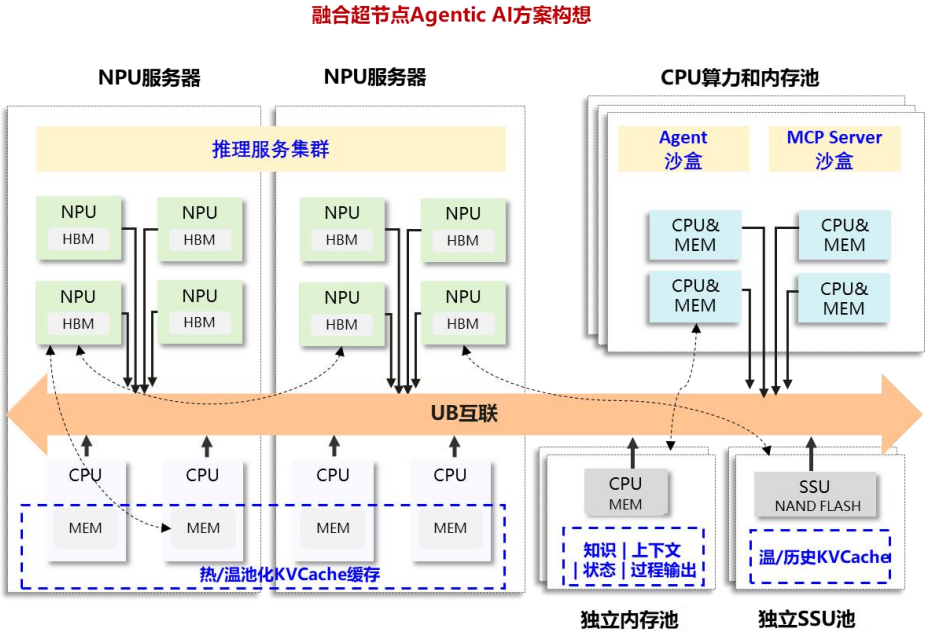


图 4-6 超节点 Agentic AI 方案架构

4.6 虚拟化

虚拟化当前仍面临多重挑战。

资源碎片化挑战：虚拟机的动态创建与销毁导致物理服务器资源产生碎片，传统云服务20%~25%内存碎片无法有效利用，整体利用率偏低。

热迁移效率瓶颈挑战：高负载虚拟机迁移时，过高的内存脏页率延长迁移时间甚至导致失败，目标主机的连续资源需求亦受碎片化阻碍。

内存利用率低下挑战：预留分配机制导致近半数虚拟机 30%~50%内存闲置，造成资源浪费。

固定配比限制挑战：通用服务器预设 CPU 与内存配比（如 1:4）难以适配内存密集型或计算密集型应用的差异化需求，造成资源配置失衡。

超节点架构针对虚拟化场景提供系统性优化方案。基于全局内存资源池实现弹性分配，内存分配率从 80%提升至 95%；依托大带宽低时延互联与统一语义技术，支持低至 50ms 的极速热迁移，有效规避内存不足风险；通过透明冷热内存分级机制，性能损失控制在 5%以内，显著提升内存利用率，如图 4-7 所示。

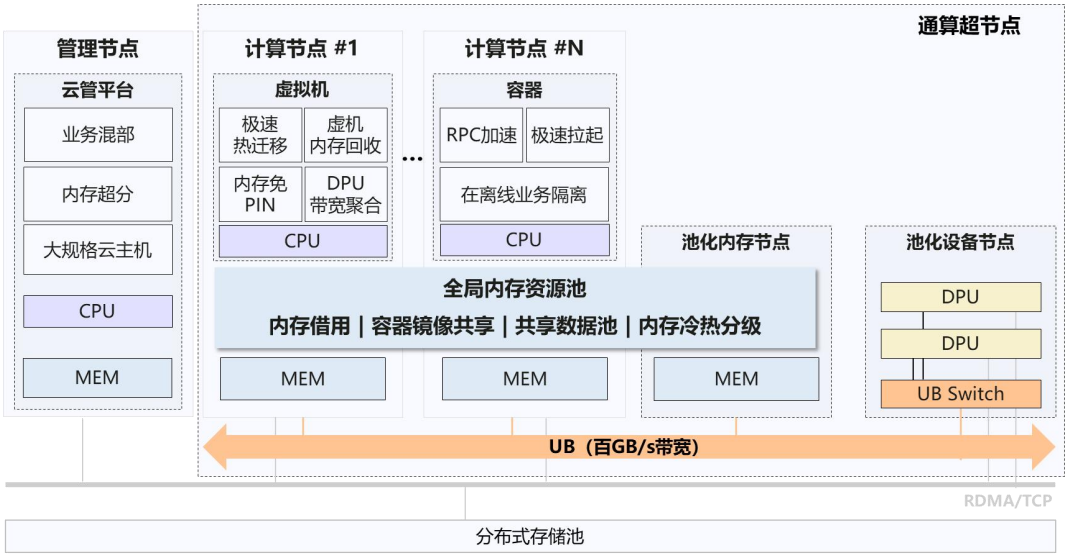


图 4-7 虚拟化超节点参考架构

4.7 大数据与数据库

大数据平台与数据库作为企业数据洞察与业务交易的核心计算基础设施，其数据处理效率与响应能力直接影响业务决策与服务连续性。然而两类系统在传统架构下面临相同的挑战：资源配置失衡导致大数据场景仅 30%~40%内存有效利用，数据库主从架构线性度低于 70%；跨节点数据 Shuffle 与分布式事务同步带来巨大网络开销，CPU 利用率仅达峰值 50%，协议栈开销超 33%，长尾 SQL 响应超 10ms。

超节点架构通过全局资源池化与大带宽低时延互联实现：大数据场景下，内存弹性分配与智能超分使 Spark 任务效率提升 30%，Shuffle 深度优化提升 10%处理性能；数据库场景下，多主架构消除写瓶颈（线性度>0.75），全局缓冲池与冷热迁移使 TPC-C 提升 20%，行列混存加速使 TPC-H 提升 65%，为数据计算构建高效加速底座，如图 4-8 所示。

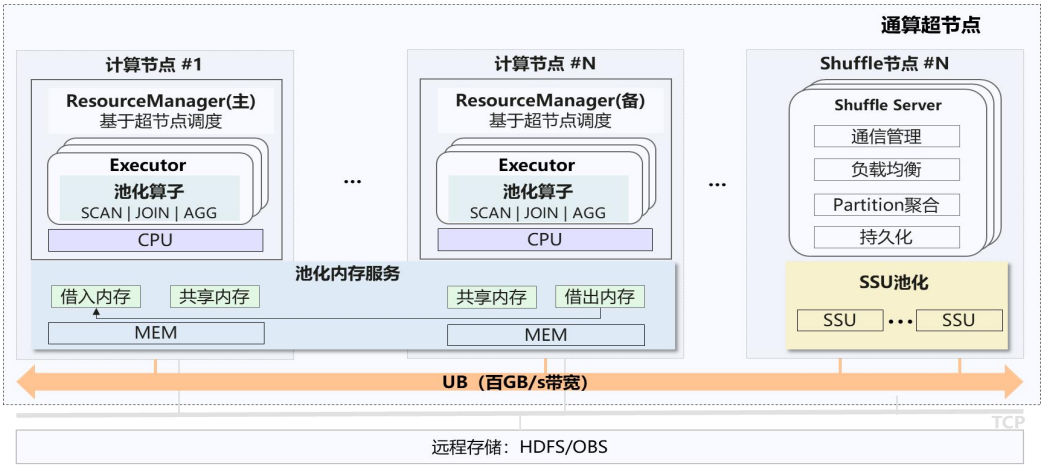


图 4-8 大数据超节点参考架构

4.8 分布式存储

分布式存储作为海量数据持久化与高可用访问的底层基础设施，其性能直接决定上层业务的存储效率。传统架构面临两大瓶颈：SSD 资源受限于单机物理边界，带宽利用率仅约 25%；以 CPU 为中心的架构中，EC 编解码、垃圾回收等任务抢占主机算力，有效 IOPS 下降 25%~35%。

超节点架构通过多样性算力平等协同提供系统性方案：构建跨节点存储资源池，SSD 带宽利用率最高提升 60%；将垃圾回收与 EC 重构等任务卸载至 SSU，IOPS 提升 25%~35%，重构速度提升 6 倍，为高密存储场景提供性能与可靠性保障，如图 4-9 所示。

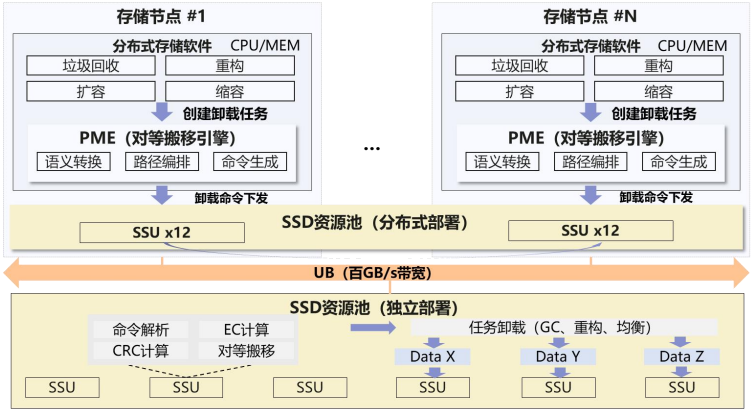


图 4-9 分布式存储超节点参考架构

4.9 高性能计算

随着科研计算需求持续攀升，高性能计算系统算力规模向 EFLOPS 级乃至 10EFLOPS 级演进。同时，AI4S（AI for Science）的兴起推动高性能计算与智算深度融合，对计算架构和性能提出更高要求。

通信性能挑战：应用通信占比可达 30%，大包场景需提升带宽，小包场景需降低时延；AI4S 中分布式矩阵计算引入 KB 级中小包通信，时延劣化问题突出。

超大规模组网挑战：集群向 10 万节点演进，网络拥塞加剧，传统胖树拓扑面临光模块数量与可靠性瓶颈。存储 I/O 性能挑战：传统客户端架构存在多次内存拷贝，增加 I/O 时延。

超节点架构提供系统性优化方案，如图 4-10 所示。基于高带宽低时延互联与拓扑感知调度，分子动力学等通信密集型应用性能提升 10%；内存语义通信使小包时延从微秒级降至百纳秒级，融合算子减少访存次数，AI4S 推理性能提升 8%。采用优化网络拓扑与动态路由技术，减少拥塞，提升带宽利用率与组网可靠性。通过数据直通机制绕过存储节点内存，I/O 性能提升 20%，统一存储架构减少数据搬移开销，为下一代高性能计算与 AI4S 融合构建高效算力底座。

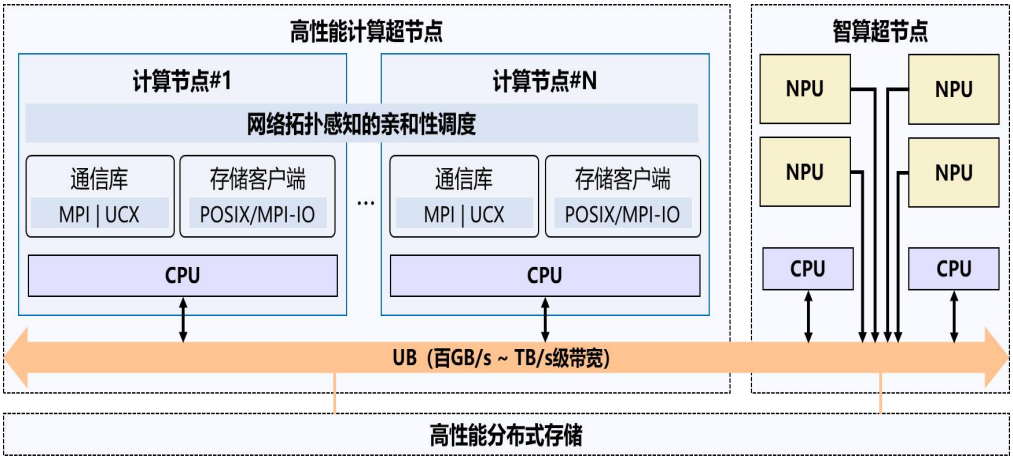


图 4-10 高性能计算超节点参考架构

4.10 行业场景

当前，自动驾驶、智慧金融、城市治理、工业仿真与医疗健康等领域正加速智能化升级，应用场景从单点试验走向规模化落地，从训练为主转向训练推理并重，从单一任务处理走向多任务并发，对算力基础设施的性能密度、数据处理能力和安全合规水平提出更高要求。

然而，行业应用落地仍面临共性瓶颈。自动驾驶 L3+需百亿公里仿真验证，数据量达 PB 级，传统算力难以支撑；金融行业大模型参数量达千亿级，须在训练推理中兼顾数据安全与合规要求；城市治理需十万路级视频实时分析，要求端到端毫秒级响应；工业仿真 CAE 计算周期长达数周，制约研发效率；医疗健康领域分子对接计算量达万亿级，混合精度算力需求迫切。各行业在算力规模、数据处理、安全合规与实时性等方面的差异化诉求，亟需定制化算力基础设施予以应对。

超节点架构通过场景深度定制提供差异化解决方案。自动驾驶领域，贵阳汽车超节点专区提供百 Eflops 级算力支撑虚实结合仿真训练；金融领域，邮储银行超节点与实现数据不出域的安全推理，支持 2000+金融智能体并发运行；城市治理领域，龙岗项目基于超节点实现 25 万路摄像头实时分析，端到端延迟低于 3 秒；工业仿真领域，长虹“云河”平台将数周仿真周期缩短至数小时，提供即开即用的 CAE 服务；医疗健康领域，济南超算中心基于超节点训练 AI 辅助诊断大模型，加速疾病筛查与新药研发进程。超节点正成为千行百业智能化升级的关键算力底座。

第五章　总线分层能力框架

超节点互联总线是一种全新的计算总线，顺应智能时代数据和算力爆炸增长趋势，实现统一内存编址、内存语义访问、高带宽、低时延、大规模的总线互联，提供多种 XPU、CPU、内存、存储设备之间的统一、对等、高性能访问机制。实现高性能并行同步计算能力，以及多样算力的统一执行环境。互联总线的核心设计理念包括：

1、以内存访问为核心，主要包括三种方式：1.1、Load/Store 方式，在本程序空间，使用处理器指令直接访问内存空间；1.2、异步 DMA 方式，在用户进程空间，通过异步方式发起对另一个用户进程内存空间访问；1.3、函数调用方式，使用互联总线 RPC，组合多项基础内存操作，实现分布式节点的复杂事务操作。

2、高性能高效率总线交互，主要包括：2.1、支持乱序和并发，将计算系统的乱序执行、多任务并发等事务高效运行在总线上；2.2、计算事务和互联传输分离，互联总线在设备之间建立可靠传输，承载解耦的计算访存事务，内存访问能够小代价扩展通信规模；2.3、消减不同互连协议引入的不必要的通信过程和开销；

3、多样算力对等访问架构，主要包括：3.1、总线上各资源互不从属，实现资源池化，提升资源利用率；3.2、数据传输不经过集中控制节点，芯片间完成直接通信，多算力高效协同完成任务；3.3、在总线上实现访问权限分区隔离、错误抑制等处理。

互联总线的协议功能设计全景如下图：

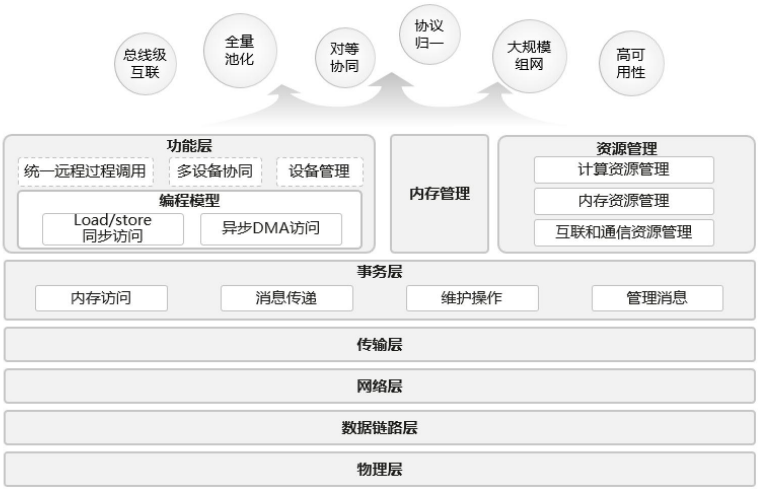


图 5-1 互联总线协议功能全景图

5.1 物理层和数据链路层

图 5-2 物理层和数据链路层

面对超节点互连的超大带宽、超低时延、超大规模、高可靠等关键诉求，在传统的物理层和链路层进行多维度的方案创新，能够最高效的使能最先进的光电互连能力。

物理层协议：

1.通过适配和编码子层，提供灵活的可变速率机制，最大化利用物理信道带宽能力，支持 112G/118G/128G/224G/236G/256G 等多种速率；同时可选兼容以太等其他物理层的 106.25G/212.5G 等系列速率；

2.链路状态管理支持故障自动检测、自动降 Lane，光模块通路冗余保护机制、实现光模块故障无丢包，实现大规模总线互连高可靠；

3.极低时延物理层编码和动态轻量化 FEC 编码选择，大幅优化互连时延，对比传统以太降低百 ns 级别；

4.支持 LPO/NPO/OIO/CPO 等光互连，实现 k 级别大规模总线互连，同时大幅降低时延和功耗，提升可靠性。

链路层协议：

1.采用 Flit 数据结构，大幅优化纠错时延、提升重传效率和小包转发性能，实现高效的链路层数据传输；

2.支持多 Link、多端口聚合方式，提供更大组合带宽；

3.使用基于信用的流控，并通过多虚通道机制，支持丰富的拓扑结构，实现多流量无死锁、流量冲突场景合理分配各业务带宽；

4.支持有 CRC/无 CRC 配置模式，平衡时延和功耗。

5.2 网络层和传输层

图 5-3 传输层

传统的节点内总线协议不支持网络和传输层，面对智算和通算大规模超节点应用，需要增加对应的处理机制，并通过灵活的多模式适配，在高效率和大规模组网间取得平衡，形成适合的超节点协议设计。

网络层协议：

1.支持网络地址编址和转发功能，可支持不同的短地址模式，可以通过线性表、分段表等方式大幅降低转发时延、降低交换复杂度；

2.支持拥塞标记，协同传输层拥塞控制降低动态时延；

3.支持 IPv4/IPv6 格式，支持超大规模部署，支持超节点总线报文桥接到以太报文，实现超节点和以太网高效互通、支持超节点和参数面网络的融合组网；

4.支持多种路由策略指示，可以精细化的控制不同的流量，使用逐包或者逐流均衡模式，

有效提升不同流量的路由效率；

5.支持自动排除故障路径，包括近端故障、远端故障、快速故障路径通告和切换等能力，提升传输可靠性；

6.支持网络隔离，实现高性能的超节点多租户组网能力。

传输层协议：

1.原生多路径 Group 机制，通过内存地址均衡、或网络地址均衡等不同模式，支持单设备能够使用更大的聚合带宽、优化负载分担的均衡度；

2.原生连接共享机制，支持物理节点间仅建立 1 个传输队列，多个事务层可共享 1 个传输队列，大幅消减多事务下重复连接开销、降低内存占用；

3.原生提供多种传输模式，如标准模式支持 E2E 重传、拥塞控制等全量特性，精简模式可精简部分复杂特性；在不同的组网场景下，提供不同的性能、可靠性、调度能力等不同维度的最优模式；

4.原生提供拥塞控制，保证多流量抢占下的公平调度、降低拥塞、控制动态时延；

5.事务和传输分离，TP 连接独立，在传输层 TP 之上，支持各类事务如内存语义、异步 DMA、RPC 等，实现原生支持各计算语义在大规模超节点总线互连的扩展；

6.支持高效重传机制，支持 E2E 的选择性重传等能力，实现数据中心范围内互联总线高可靠部署。

5.3 事务层和功能层

图 5-4 功能层

超节点总线的各类通信功能，需要封装为统一的事务层和功能层接口，基于内存调用简化编程，有效降低通信的调用开销，同时可以支持 RPC、集合通信等的复杂通信加速能力。

功能层协议：

1.同时提供同步 Load/Store/Atomic 访问和异步 DMA 访问，单报文 MTU 长度可支持 1kB~8kB，数据访问单事务长度最大可支持 MB~GB 级，可在不同业务场景选择适合的功能使用；

2.提供基于用户态内存指针引用的过程调用，大幅简化传统过程调用的复杂协议栈功能接口，提升编程效率；

3.总线设备是系统中可寻址的最小粒度功能实体，能够发起或响应事务操作、可以被管理；在一个 Domain 内提供唯一标识设备 ID；

4.提供原生的多设备协同机制，支持功能层封装和提供多总线设备之间的通信加速，例如 HPC 场景的同步并行计算，AI 训练推理的集合通信，数据库场景的功能。

事务层协议：

1.提供内存操作（Ld/St/Atomic 等）、消息操作（Read/Write/Send 等）、维护操作（状态同步等）以及管理操作等四大类事务，实现计算单元间高效交互；

2.支持多种可靠性和序模式，例如：可靠保序、目标保序、可靠乱序、不可靠乱序等不同模式和事务层接口；结合不同传输/网络模式提供高效易用的编程接口；

3.事务操作保持原子化，含地址注册、生命周期、权限、原子宽度、作用域和内存序等，并实现周边解耦，保证协议栈核心的稳定。

5.4 资源和内存管理

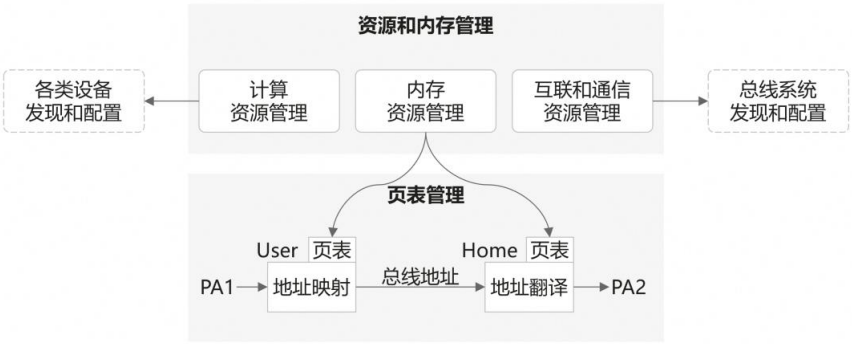


图 5-5 资源和内存管理

通过资源和内存管理，实现超节点的全局统一编址，支持多样性 XPU 平等互连、池化共享、支持大规模和多样性拓扑的高效管理应用；支持带内和带外等不同方式的管理架构，高效的被各种管理系统所集成。

资源管理：

1.定义原生系统管理权限和消息，在超节点中通过总线带内或带外方式完成资源管理；

2.完成各类计算和互联资源管理，实现多算力对等访问；

3.采用分区方式支持多用户，一个设备支持属于且仅属于一个 Partition，属于不同 Partition 的设备之间访问隔离，提供高性能的多租户部署方案；

内存管理：

1.内存使用方（User），通过地址映射单元完成本地 PA 地址到总线虚拟地址的映射；提供方（Home），通过地址翻译单元完成总线虚拟地址到本地 PA 的映射；通过总线虚拟地址实现大规模的内存统一编址映射和部署；

2.基于设备 MMU、页表管理、缺页处理和权限管理等机制实现 Shared Virtual Memory Address、non-pinned 内存、zero-copy 等能力，物理内存按需加载，提高动态内存管理效率；

3.提供 Ld/St/Atomic 等语义的安全隔离机制、支持完整的鉴权特性，实现多设备对相同内存空间的安全访问处理。

第六章　超节点案例

6.1 多尺度超节点

超节点作为突破传统计算集群通信、功耗与复杂度等瓶颈的新技术形态，重点面向 AI 大模型全流程训推等新式计算需求构建了梯度化、多尺度的架构体系，可延展覆盖百、千、万卡级多尺度架构体系，按需匹配差异化的业务场景。

百卡级超节点以 384 卡规格为商用主流形态，通过高速互联总线实现超百颗芯片的统一内存编址与协同计算，通信带宽较传统架构提升一个数量级，是当前千行百业规模化落地的主力方案，可高效支撑千亿参数级模型的训练与高并发推理任务。

千卡级超节点以 1024 卡为典型构建单元，通过新一代光互联与统一总线技术突破单集群规模上限，面向万亿参数级大模型预训练等前沿科研场景，是下一代超节点技术的核心演进方向。

万卡级超节点是超节点技术体系的未来发展趋势，将突破多套千卡单元横向扩展的组网模式，依托全光无阻塞交换、全域内存池化等前沿技术，逐步形成单域紧耦合的万级芯片规模超节点形态，可承载十万亿参数级超大规模基础模型、通用人工智能复杂智能体等前沿研发任务，成为未来国家级智算基础设施与大科学装置的核心算力架构方向。

多尺度超节点体系既通过标准化单元实现工程化、规模化交付，又可通过灵活组网适配从行业本地化到大科学装置的全谱系算力需求，为新型算力基础设施的分层建设与高效利用提供了架构支撑。

6.2 超节点案例

6.2.1　鹏城云脑 III 架构案例

鹏城云脑 III 是国内首个采用超节点架构的国产智算集群。在物理实现上，该系统采用了计算刀片与交换刀片直接正交盲插的全电交换架构，将 64 颗 NPU 通过总线互联为一个高度耦合的计算单元。这一架构摒弃了传统的 PCB 背板走线和光纤传输方式，大幅缩短了高速信号的传输距离并减少了连接器级数，展现出高密度、低时延与高可靠性的显著优势。系统采用的三维浮动全盲插设计，将电源连接器、液冷快接头及信号连接器均集成于机柜内部，

实现了即插即用。这不仅极大地简化了系统复杂度，更确保了大规模量产部署的可制造性与稳定性。

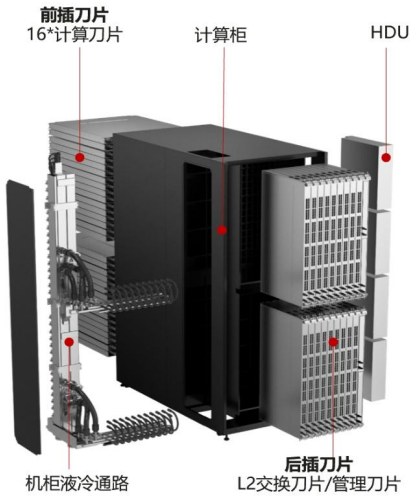


图 6-1 鹏城云脑 III 计算节点

在通信协议层面，超节点采用全局统一编址的内存机制，支持内存语义直接访问和无序访问。交换芯片和 XPU 可直接携带目的及源内存地址进行交互，大幅简化了软件处理流程，在时延和带宽规格上均实现了显著优化。基于此架构，超节点内 NPU 间的双向通信带宽超过 650GB/s，较传统网络提升了一个数量级，点对点最小通信时延则降低至 1.2 微秒以内。

在千亿级参数模型的千卡并行训练中，该集群的 MFU（模型算力利用率）达 43%。在从 128 卡扩展至 1024 卡的过程中，其扩展效率达 98.4%，充分体现了超节点架构在大规模分布式训练中的高效扩展能力。

6.2.2 模型服务厂商智算集群案例

超节点方案解决智算中心两大痛点：一是解决大规模算力集群能源消耗问题，二是解决算力密度问题。

华为 Atlas 900 A3 SuperPoD 超节点采用 384 卡设计，灵衢全互联架构。满足行业智算高性能、高能效液冷训练集群的需求，提供业界领先的 AI 集群方案。

超节点采用风液混合架构，超高风液比带来了超低的 PUE。同时超节点也解决了算力密度问题。超节点方案提供的算力相当于 125 台风冷设备提供的算力，传统机房需要占用

一个 400 ㎡的机房，超节点方案占用机房约 52 ㎡，算力密度提升 7.7 倍。超节点的灵衢传输方案，让卡间单跳，从传统 2-3us 的时延降低至 200ns 左右，通讯时延降低了 10 倍以上。算力密度的提高，同时也解决了通讯距离过远的问题，通讯传输距离的缩短，可以大大降低光模块功率、线缆规格，网络建设辅材成本降低 50%。

6.3 应用场景案例

6.3.1 大吞吐推理案例：云服务商基于超节点方案大 EP 实现 DeepSeekV3 大吞吐推理

云服务厂商采用超节点方案大 EP 方案，实现了 DeepSeekV3 1 卡 1 专家推理方案，借助超节点带宽和时延优势，通过增大 EP 卡数，实现单卡权重降低和单卡并发增大，Decoding 性能超越 H1xx。

在 DeepSeekV3 大 EP 方案中，EP 卡数从 16 卡增加到 288 卡，单卡权重 HBM 占用从 42GB 降低至 2.33GB，可以保持更多的 KVCache，实现 BatchSize 从 8 增长到 30。

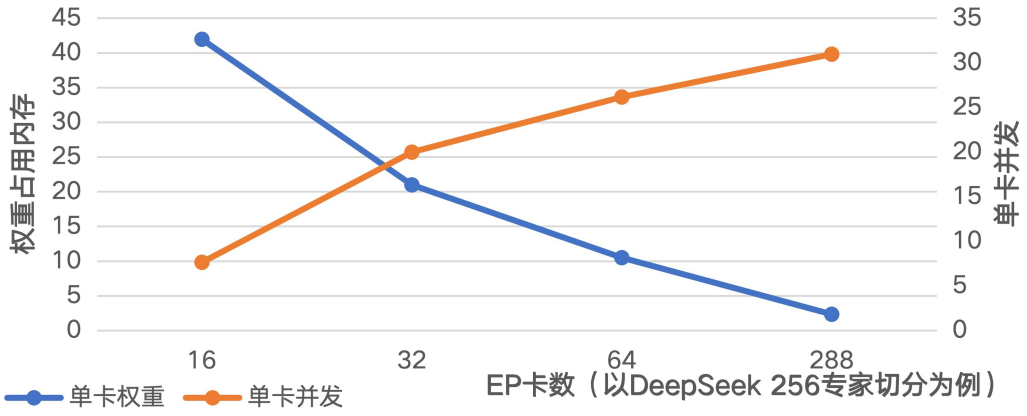


图 6-2 deepseekV3 大 EP

EP 卡数越多， Decode 计算耗时越小，从 37ms 降低至 17.5ms， EP 卡数越多，多卡之间 All2All 通信挑战越大，虽然通信耗时占比从 5%提升至 38%，但是借助于超节点的大带宽，整个系统仍处于最优状态。

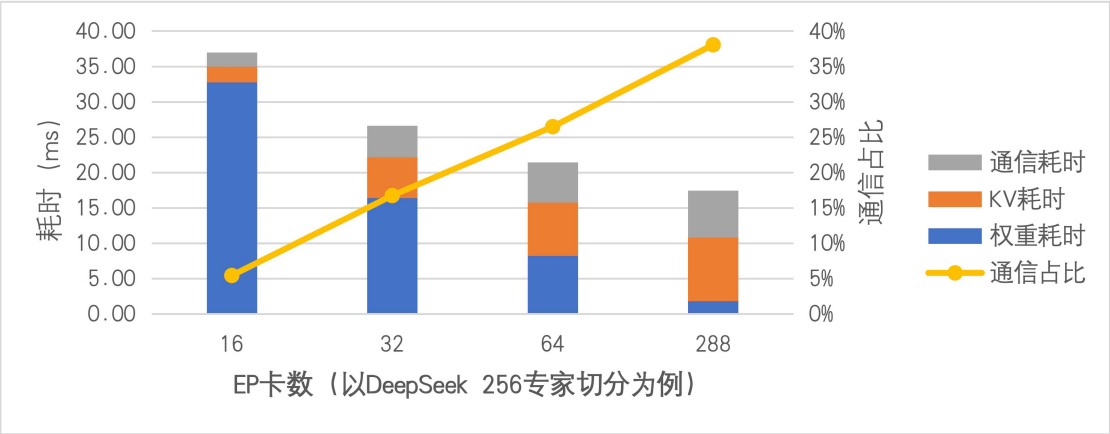


图 6-3 deepseekV3 大 EP 通信耗时和通信占比

6.3.2 低时延推理案例：能源、制造、互联网客户基于超节点方案实现 DeepSeekV3 低时延推理

能源、制造和互联网行业客户采用超节点方案提供 DeepSeekV3 推理服务。在 50ms TPOT 约束下，超节点方案单卡推理性能相比服务器单卡实现了 3x+性能提升，极大降低了推理服务成本。

DeepSeekV3 不同专家之间需要进行 Dispatch & Combine 通信后才能进行后续计算，模型每个 token 需要通信 58*2=116 次。Dispatch & Combine 通信时延直接影响大模型推理的 TPOT 时延。

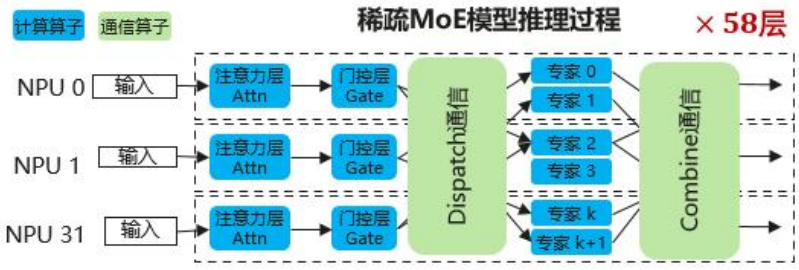


图 6-4 MoE 模型推理过程

Decoding 阶段由于存在 HBM Bound，计算耗时不会随着计算量等比增长，借助超节点低时延优势缩小了 dispatch 和 combine 的通信耗时，提升 batchsize，实现更大的的吞吐。

6.3.3 TTFT 时延降低案例：互联网多轮对话场景基于超节点方案实现 TTFT 时延降低

互联网客户基于 LLM 构建智能客服系统，需支撑大规模多轮对话交互。由于用户基数大，存在大量重复问题和高频任务，需通过 Prefix Cache 技术降低 Prefill 阶段 TTFT。超节点提供内存池化能力，通过全局统一编址和内存语义，实现超节点内 NPU HBM 与 Host DDR 的直接数据交换，KV Cache 跨节点读写时延较 RoCE 降低 10 倍。

在此基础上，客户基于 DeepSeekV3.1-671B 模型构建金融管家服务。开启 Prefix Cache 后，TTFT 从 12119.31ms 降至 9289.28ms；叠加超节点内存语义（NPU D2H 访问 Host DDR）后，TTFT 进一步缩短至 7235.28ms，QPS 提升 37%。

在 RL 后训练场景中，Agent 进行多轮对话，每轮在上一轮基础上新增 128 token 输入与 128 token 输出，KV Cache 命中率超过 90%。超节点内存池架构实现 Prefill 与 Decoding 节点的 KV Cache 共享，借助内存语义 rH2D/rD2H 跨节点低时延访问，TTFT 最大降幅达 75%。

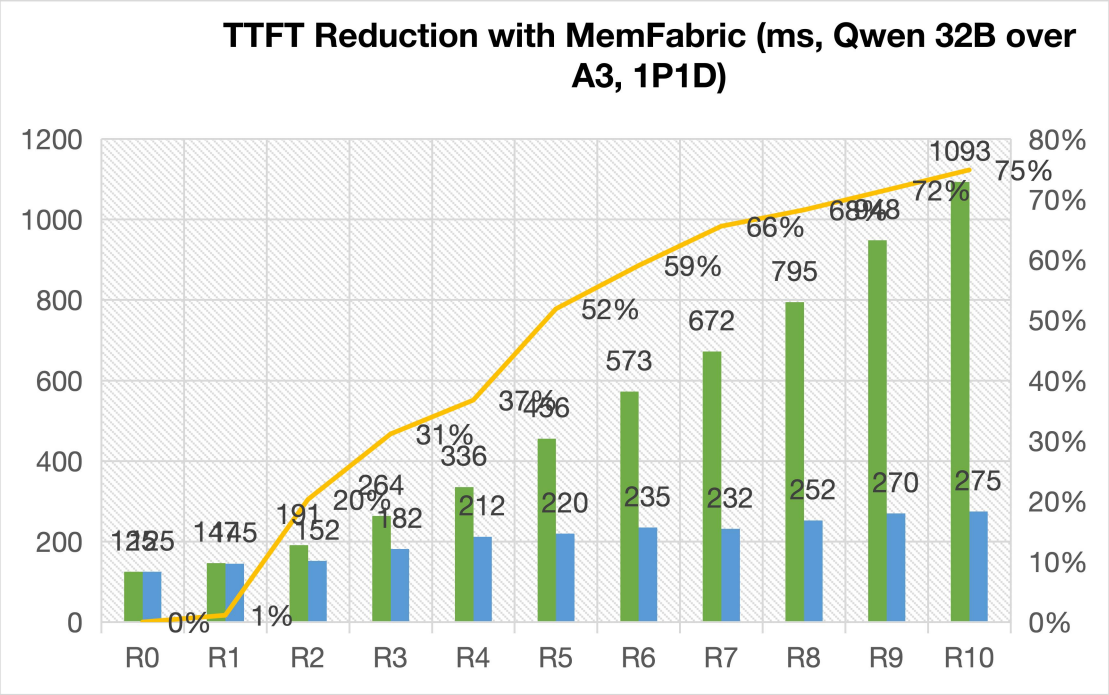


图 6-5 互联网客户 RL 后训练场景

6.3.4 大模型训练加速案例：互联网客户基于超节点方案实现稀疏 MoE 大模型训练加速

DeepSeekV3 是当前国内外被广泛使用的模型，大量客户基于 DeepSeekV3 进行二次训练调优，DeepSeekV3 模型训练的难点之一在于如何实现 all2all 通信的计算通信掩盖。在传统服务器中，由于 RoCE 带宽较小，all2all 通信无法被掩盖，需要使用较为复杂的并行策略。在超节点方案中，由于节点间带宽无收敛，all2all 性能相比 RoCE 提升 5-6x。使用 EP64 可以实现 ALL2ALL 通信全掩盖，单步耗时 49.2s，而相同配置下如果使用 RoCE 网卡，单步耗时需要 103s，超节点大带宽实现了性能翻倍。

图 6-6 计算和通信分布

6.3.5 大模型训练易用性提升案例：互联网客户基于超节点方案实现大模型训练易用性提升

FSDP 由于其简单易用，被头部互联网客户广泛使用。FSDP 采用参数分片策略，将参数均分切分在不同的 NPU 上，在计算之前，提前预取，在百卡规模使用 FSDP 训练时，FSDP 存在计算通信无法掩盖的问题。

以 Qwen3-235B 为例，在 256 卡进行 FSDP 训练时，单次通信量在 5GB，超节点方案预计耗时是 RoCE 的 56%。8K 序列计算无法掩盖通信，超节点性能是同规模 H8xx 集群的 1.6x；16K 训练时，超节点性能是同规模 H8xx 集群的 1.04x。

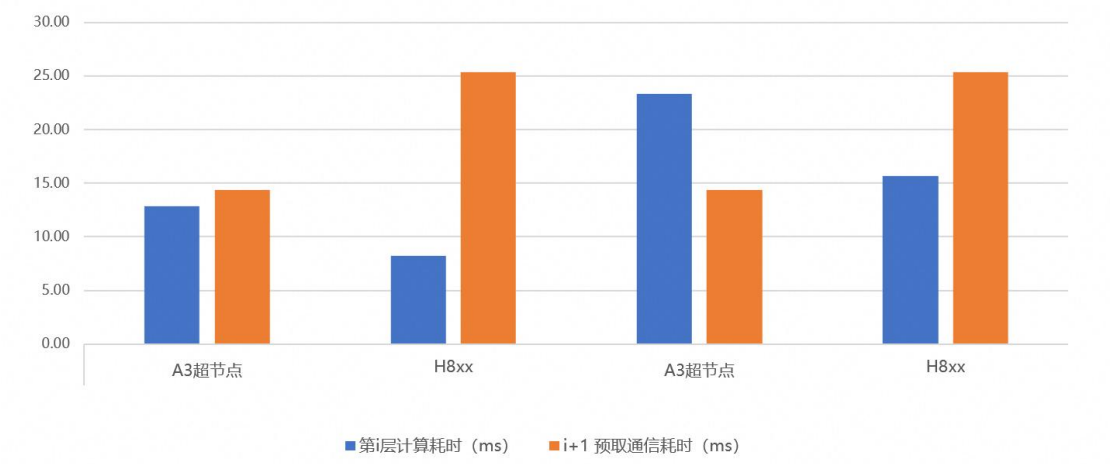


图 6-7 性能对比

6.3.6 搜推广案例：互联网搜推业务基于通算超节点构建极低时延生成式推荐业务

搜推广业务作为互联网电商平台核心营收场景，正全面迈入生成式推荐时代，召排一体化、长序列推理、万亿级向量检索与 PB 级 KVCache 缓存成为必备关键能力。但是，传统生成式推荐业务采用四级缓存，架构复杂，末级缓存访问时延高达 20ms，缓存命中率低。同时 RPC 通讯开销高，跨服务通信损耗占生成式推荐整体时延 30%以上，不同用户行为序列大小各异，多次内存拷贝损耗，极易引发时延毛刺，存在“推理精度提升”与“响应时延达标”的核心矛盾，造成推理精度低，影响广告转化率。

某头部互联网电商基于通算超节点的灵衢大带宽低时延和内存语义通信能力，以及 openEuler、UB Service Core、openFuyao、openYuanrong 等系列化开源组件，构建高性能弹性调度算力底座。通过大容量 KVCache 和 RPC 通信加速，大幅降低生成式推荐端到端时延。

大容量 KVCache：基于灵衢构建百 TB 级全内存 KVCache 共享缓存资源池，全面替代传统四级缓存架构。实现多级缓存直访、共享内存免拷贝、零冗余传输和分布式数据缓存等技术创新，并通过冷热数据智能迁移，有效提升缓存命中率，KVCache 访问 TP99 时延降低 80%。

RPC 通信加速：基于灵衢传输层与事务层分离的创新设计，有效降低建链内存开销，

搭配通算超节点 K 级大规模组网能力，可稳定支撑十万级容器大规模通信；同时通过灵衢 U 异步编程模型技术全面加速 RPC 通信，消除多层数据拷贝与协议封装开销，大幅削减容器间通信损耗，实现 RPC 通信 TP99 时延降低 40%。基于 K8s 标准插件机制提供 ub-network-device-plugin 实现 UB 设备直通容器，兼容业界生态；基于 UB Service Core 提供的高阶服务 Socket 接口兼容，数据面 Bypass TCP/IP 协议栈，应用零修改。

经真实业务场景实战验证，整套方案实现搜推广业务全维度性能跃升，生成式推荐全链路端到端时延下降 40%，释放的时序余量可以用于更长序列用户行为推理、支撑更大参数规模生成式模型落地，兼顾业务响应速度与推送精准度，大幅优化用户体验与商业转化效率，为互联网生成式搜推广业务规模化、高性能迭代提供了成熟可复制的算力底座方案。

6.3.7 虚拟化案例：云服务提供商基于通算超节点构建云&容器基础能力底座

传统虚拟化虽通过资源抽象提升数据中心部署灵活度，但存在明显短板：虚拟机频繁启停造成 CPU、内存、存储碎片化，内存搁浅率达 20%~25%；高负载虚拟机迁移受脏页、资源碎片制约，迁移效率差甚至失败；内存采用预留分配模式，大量虚拟机闲置 30%~50% 内存；固定 CPU 内存配比无法适配计算、内存密集型业务差异化诉求，整体资源利用率偏低。

某云服务商通过通算超节点，通过内存池化能力，极速热迁移能力，通过全局资源弹性调配，实现虚拟化场景编排不同资源主机之间内存借用实现内存整体售卖率从 80%->95%，依托 UB 低时延大带宽通信与统一语义实现 50ms 极速热迁移，规避内存超分风险；借助页面访问统计完成透明冷热内存分级，性能损耗控制在 5% 内；同时可支持创建最大 6T 超大规格内存云主机，支撑 HPC EDA 场景/金融量化，为新一代虚拟化云平台提供完整优化方案。

6.3.8 数据库案例：运营商以及互联网分别基于通算超节点提升数据库一写多读以及多写多读数据库能力

运营商某数据库，存在传统网络与 I/O 性能制约性能，通讯瓶颈，网络协议栈开销最大超过 33%；存储 IO 瓶颈，缓存不命中导致 SQL 长尾时延，响应时间大于 10ms；通过

通算超节点，加速日志同步、全局资源分配、扩展 Buffer Pool，TP 性能提升 20%，优化算子减少数据溢盘、缓存列存加速 shuffling，AP 性能提升 50%，通过内存池缓存脏页降低 RTO、故障主动通知加速故障检测，达成 RTO<6s、故障检测时间<=1s。

互联网某数据库，游戏开服、电商大促等写密集场景存在写瓶颈，需要多写能力；多写存在页面 ping-pong、并发控制、资源分配等业界难题。基于 UB 实现全局 Buffer Pool、分布式锁、全局事务资源、消息交互等功能，达成线性度>=0.8@8 节点、内存节省 10%。

6.3.9 大数据案例：运营商大数据基于通算超节点提升集群资源效率

传统大数据平台存在静态资源分配失衡、跨节点 Shuffle 开销过高两大问题，内存闲置严重。运营商客户大数据 Spark 集群静态分配机制导致资源无法充分利用，数据交换效率制约性能，资源利用率 30%-40%。基于实际资源使用率的超分调度，利用率 60%，超节点资源调度，性能提升 20%+；基于内存借用/共享，减少数据溢盘，算子效率提升 10%。基于 UB 低时延通信，shuffle 通信效率提升 10%。

6.3.10 分布式存储案例：运营商分布式存储基于通算超节点降低 TCO

传统分布式存储存在资源孤岛、CPU 算力瓶颈两大难题：单机 SSD 相互隔离，带宽利用率偏低；EC 编码、垃圾回收等存储任务挤占 CPU，叠加多层 I/O、网络开销，扩容受限。

以某云服务商分布式存储为例，存在硬件资源离散化，SSD 带宽利用率仅 25%，CPU 为中心大量 IO 栈与网络栈开销占比超过 30%。分布式存储超节点基于 UB 全局资源池与异构协同算力优化，打通跨节点存储资源，通过 SSU 直通，无带宽收敛，设备对等互联免 CPU 开销，可以实现盘即存储，大幅降低 TCO。

6.3.11 通智融合应用案例：OpenClaw 智能体高效应用部署

通智融合超节点也是重要的超节点形态之一，可支撑具备自主规划、决策与执行能力的 Agentic AI 的高效应用。其中，通用计算单元主要负责处理任务规划、工具调用、数据存取、环境交互与逻辑调度等复杂流程，智算单元专注于大模型推理、即时决策与多模态信息生成。

该场景下典型应用实例为 OpenClaw 智能体工作流。以 OpenClaw 为代表的自动化任务执行场景，充分展现了通智融合超节点的价值。在其运行过程中，需同时处理：

多任务并行：网页访问、邮件处理、任务规划、记忆管理、模型推理等。

实时操作：餐厅预订、金融交易、报告生成等远程操控。

此类工作流的核心需求在于，任务流需在通算与智算单元间进行多轮、实时的反复交互，并保障数据能够快捷、安全、持续地流动。高效、低时延的通算智算协同与数据访问机制，使得 OpenClaw 能够更迅捷、可靠地执行任务，满足日益复杂的自动化需求。

通智融合超节点能够在一个统一的运行环境中，动态、弹性地支撑从感知交互、思考决策到行动执行的完整智能闭环。这不仅加速了通用人工智能在实体世界中的深度融合与落地，也为下一代自主智能系统的规模化部署奠定了坚实基础。

第七章　超节点的未来发展

超节点是人工智能时代，计算机系统技术发展的最新成果，已经成为 Tokens 时代的核心技术引擎。当前，人工智能技术还在迅猛发展，尤其是当 AI 大模型技术融入生产系统时，

必然面临和传统软件技术的深度融合，形成以 Agentic AI 为代表的新型人工智能软件，支撑下一代自主智能系统的演进。这类软件兼具模型推理、工具执行、复杂记忆、情景思维、群体智能等综合智能能力，将给超节点带来新的需求与挑战。为此，超节点需在多样性算力、存储、高速互联、工程设计（供电、散热，全液冷成为主流、NPO/CPO/OIO）、软件耦合等关键技术领域持续演进，以迎接人工智能计算时代的全面到来。

7.1 多样性算力

单一 xPU 算力无法覆盖模型预处理、大模型训练、轻量化推理、长文本序列运算全业务场景，多样性算力融合部署成为超节点重要形态，CPU、GPU、NPU、LPU 多类型芯片混合集成成为常态。对于物理 AI、多模态、AI for Science 等复杂应用场景，超智融合架构成为创新方向，以满足复杂任务所需的多精度混合计算。

CPU 承担集群调度、业务逻辑处理、基础控制任务；GPU/NPU 负责通用大模型训练、推理；LPU 针对长文本、大语言模型序列运算优化，弥补 GPU/NPU 在语言处理场景的算力短板。多样性算力超节点依托硬件层多芯片高速互通、软件层统一算力调度，实现不同算力单元优势互补，按需弹性分配算力资源，显著提升整机算力利用率与业务适配范围。

7.2 存储

在 AI 大模型场景中，KV cache 长度关乎智能表现，其海量在线访问是性能瓶颈。SSU 通过低时延、高带宽的灵衢总线，实现计算单元对存储的高性能直访，支持“以存代算”，并为未来 Engram 等记忆技术发展奠定基础。

在通用计算中，SSU 同样价值显著。数据库场景下，SSU 提供的超低时延直访能力可提升 TPS 性能，降低访问时延，优化内存与算力效率；其卸载数据搬迁功能能加速数据重平衡，提升系统可靠性。分布式存储场景下，针对 EC 编解码等任务抢占 CPU 导致的瓶颈，SSU 构建的协同计算架构利用直通与大带宽，打造高密存储底座，最大化主 IO 利用率，通过后台流量盘间卸载实现性能提升。

7.3 高速互联

超大规模节点是智算基础设施核心演进趋势。伴随大模型训练、通用人工智能算力需求爆发，算力集群规模持续扩容，演进路径从单机柜，逐步迈向千卡、万卡乃至十万卡级巨型算力池。传统电互联存在带宽、时延与传输距离瓶颈，唯有光电融合技术可突破物理约束，搭建全域互联底座，实现跨机柜、跨集群纳秒级高速通信与全局资源池化。

行业需开展全链路系统性创新，底层深耕硅光、LPO/NPO/OIO/CPO 光模块等核心器件；中层迭代统一总线等互联协议与集合通信调度算法；上层融合光交换、混合 WSS/OCS 光组网技术，打通端到端光电协同通路。基于超节点架构，采用混合拓扑及总线交换设备，通过器件、协议、光电组网、软硬件调度多维度协同突破，构建高带宽、低时延、低能耗的统一算力互联架构，全方位优化资源访问效率。整套体系兼顾故障自愈、负载均衡与全域隔离能力，支撑十万卡级集群 7×24 小时稳定运转，以长稳可靠的算力底座，承载千行百业超大规模 AI 业务落地。

7.4 功耗

传统通用服务器机柜功耗普遍维持在 10kW 至 30kW，早期 AI 算力机柜峰值功耗不超过 120kW，伴随机柜宽度拓宽、大于 128 卡超高密集成、高速光模块、液冷辅助设备叠加，未来超节点峰值功耗突破 MW，正式进入兆瓦级时代。

兆瓦级机柜对机房底层基础设施提出全新要求，传统低压配电、常规线缆、消防、空调系统均无法适配瞬时大功率负载，倒逼机房配电架构、线缆规格、冷却系统、消防体系同步重构，算力机房设计逻辑彻底区别于传统 IDC 机房，形成高功率、高密度专属智算机房建设标准。

7.5 供电

传统 48V 低压直流母线供电方案在兆瓦级机柜、5KW 单 xPU 高负载场景下缺陷突出，同等传输功率下电流数值过大，线缆/Busbar 发热严重、电能传输损耗高，机柜内部供电线缆排布占用大量空间。未来超节点供电体系完成核心升级，±400V/800V 高压直流 HVDC 成为主流母线规格，未来持续向 1500V HVDC 演进。

高压直流供电可在传输同等功率时大幅降低母线电流，减少线缆发热与电能损耗，提升整机供电稳定性；同时精简机柜内部供电线缆数量，释放更多空间用于算力硬件堆叠，匹配兆瓦级整机柜功耗的大功率供电需求，成为高密超节点标准化供电演进方向。

7.6 散热

随着单机柜、单芯片功耗持续攀升，传统风冷散热风量上限、散热效率短板暴露，超节点散热演进路径清晰，当前主流风液混合冷却将逐步退出高端算力场景。风液混合方案仅对高功耗 xPU 配置冷板液冷，其余部件保留风冷，仅适用于中低功耗场景；下一代超高密超节点将全面采用全液冷架构，整机取消风扇等风冷部件。

除上述关键技术领域的持续演进，超节点的下一步发展还需面向开放生态与标准化能力建设。随着超节点从 AI 训练与推理扩展到云计算、大数据、数据库、分布式存储、通智融合与 Agentic AI 等更广泛场景，其产业价值将不仅取决于单点硬件性能，也取决于系统架构、软件生态、开放互联、虚拟化、安全隔离、可靠性与运维体系能否形成可持续协同。为此，超节点宜在统一术语与能力分级、设备与拓扑的标准化描述、跨厂商互操作，以及面向功能、性能、可靠性与安全的测试认证体系等方面持续推进，使硬件能力可被操作系统、Hypervisor、运行时框架、通信库与云平台稳定地发现、调度、计量与治理。

7.7 标准化

当前超节点产业正处于蓬勃发展阶段，仍存在诸多技术与工程难题亟待攻克。需持续推进超节点统一标准与产业生态构建，围绕机柜硬件、机房部署、跨节点互联、软件栈、运维管理接口等全链路开展分层标准化建设。

附录一：符号表/术语表

术语	全称/说明
All2All	集合通信原语，每个参与者向每个其他参与者发送不同的数据块，

	MoE 的 EP 并行场景下为主要通信模式。
Clos	一种多级交换网络拓扑（也称 Fat-Tree），通过中间交换层实现无阻塞或近似无阻塞互联。
CPO	Co-Packaged Optics，光引擎与交换芯片共封装，消除可插拔光模块走线，适用于极高密度场景。
DMA	Direct Memory Access，由专用引擎驱动的批量数据搬运，Kernel 分离，适用于大块集合通信。
DP	Data Parallelism，数据并行，每张卡持有完整模型副本、各自处理不同数据批次。
DPU	Data Processing Unit，基础设施平面自治中枢，卸载网络、存储、安全与遥测等非计算负载。
EP	Expert Parallelism，专家并行，MoE 模型中不同专家分布在不同设备上，触发 AllToAll 通信。
Flit	Flow Control Unit，链路层最小传输单元；总线型协议多采用固定长度 Flit，以太型采用可变帧。
HBM	High Bandwidth Memory，高带宽堆叠存储，通过硅中介层（interposer）与计算芯片互联。
KV Cache	Key-Value Cache，Transformer 推理中缓存已计算的 Key/Value 向量，避免重复计算。
LDST	Load/Store，指令级细粒度远端访存

LPO	Linear-drive Pluggable Optics，线性驱动可插拔光模块，省去 DSP 重定时，降低功耗与延迟。
MoE	Mixture of Experts，混合专家模型架构，通过门控路由激活部分专家子网络，降低计算量。
NPO	Near-Packaged Optics，光引擎紧邻芯片封装但不共基板，缩短电走线至 <25 mm。
OCS	Optical Circuit Switching，光电路交换，通过光学路径重配置实现拓扑动态调整。
RAS	Reliability, Availability, Serviceability，可靠性、可用性、可维护性。
SLA	Service Level Agreement，服务等级协议，定义延迟、吞吐等性能约束的量化承诺。
TCO	Total Cost of Ownership，总拥有成本，涵盖采购、部署、运维、能耗和折旧的全生命周期费用。
TP	Tensor Parallelism，张量并行，将单个算子的权重矩阵切分到多设备，要求高带宽同步通信。
TPOT	Time Per Output Token，推理场景下每生成一个 token 的延迟，衡量 Decode 阶段性能。
TTFT	Time To First Token，推理场景下从请求到达到首个 token 输出的延迟，衡量 Prefill 阶段性能。

Global Computing Consortium

全球计算联盟

GCC 官方微信

GCC 官方网站