AI大模型与异构算力融合技术白皮书

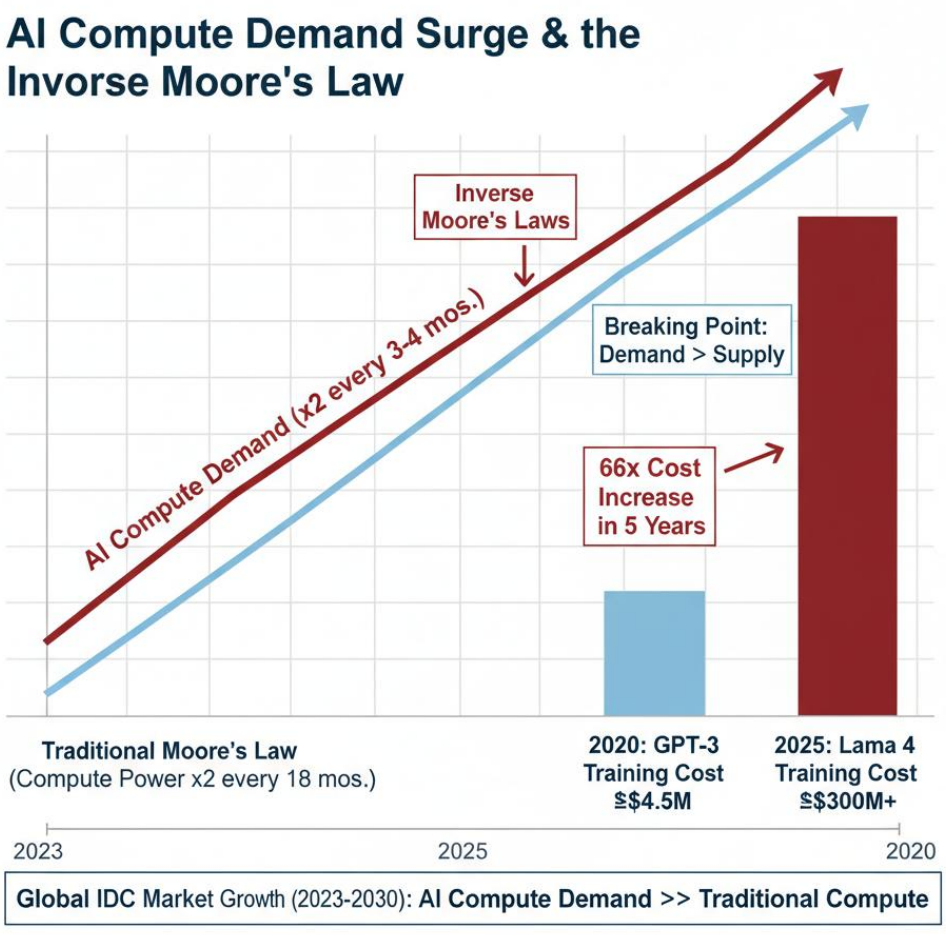
目 录

一、前言

1.1 报告背景与意义

1.1.1 AI 大模型爆发与算力需求激增

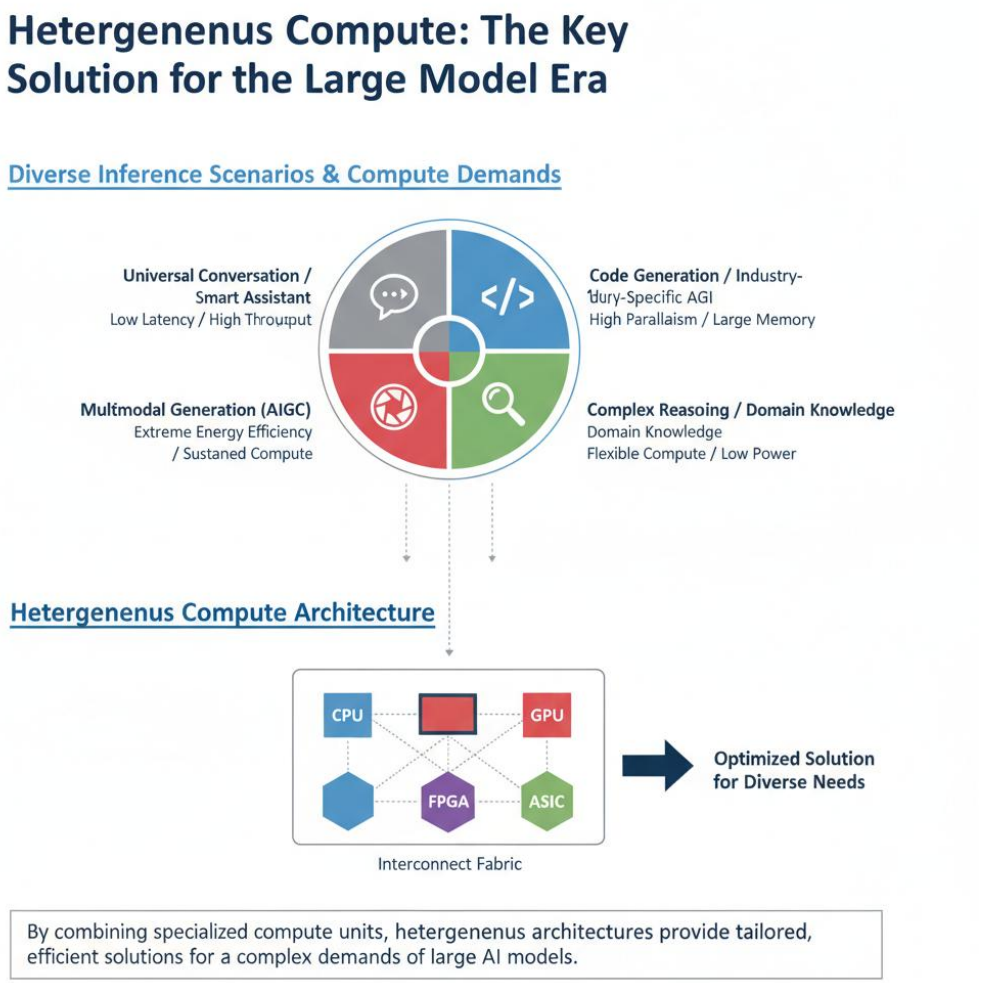
近年来，人工智能大模型技术呈现爆发式增长，模型参数规模从亿级迅速扩展至万亿级。根据最新研究显示，全球 AI 算力需求正以每 3~4 个月翻番的速度突破临界点，远超传统摩尔定律预测的计算能力提升速度（每 18 个月翻倍），形成了所谓的"逆摩尔定律"（Inverse Moore's Law）。IDC 预测，2023-2030 年全球 IDC 市场将保持高速增长，其中 AI 算力需求增速显著高于传统算力。



逆摩尔定律下的 AI 算力需求爆发

随着 GPT-5、Llama 4、Claude Opus 4.1 等大模型的不断演进，模型参数规模持续扩大。2025 年，OpenAI GPT-5 参数规模行业预估从 3 万亿到 52 万亿不等，

业界已开始关注模型效率而非简单扩大参数量，数据质量、数据多样性和领域覆盖度成为更重要的竞争因素。这种"膨胀速度"带来了前所未有的算力需求。据测算，训练 Llama 4 的成本预计花费数亿美元，而 2020 年训练 GPT-3 的成本约为 450 万美元，五年间训练成本增长数十倍。这种算力需求的激增使得单一架构的算力供应难以满足，异构算力成为应对这一挑战的必然选择。

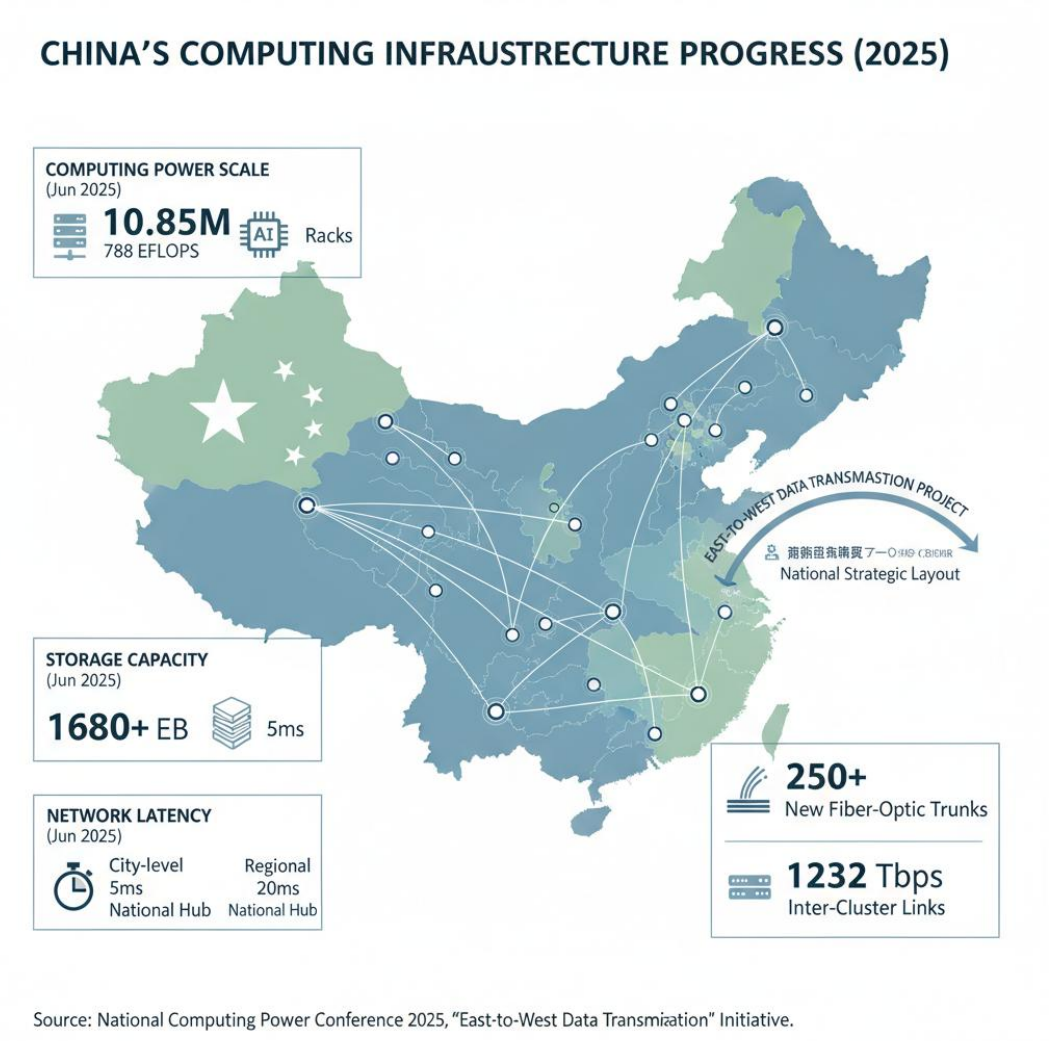


异构算力：大模型时代的关键解决方案

与此同时，推理场景的多样化进一步推动了对异构算力的需求。从通用对话到行业专用应用，从 AIGC 到智能助手、代码生成、多模态生成（视频、音乐、3D、数字人）等场景，对算力的需求各不相同——有的需要高并行计算能力，有的需要低延迟响应，有的则对能效比有极高要求。这种多样化的需求使得单一类型的计算单元难以全面满足，异构算力通过组合不同特性的计算单元(如 CPU、GPU、FPGA、ASIC 等），能够针对不同场景提供最优的算力解决方案，成为大模型时代的刚需。

1.1.2 国内外政策与产业驱动

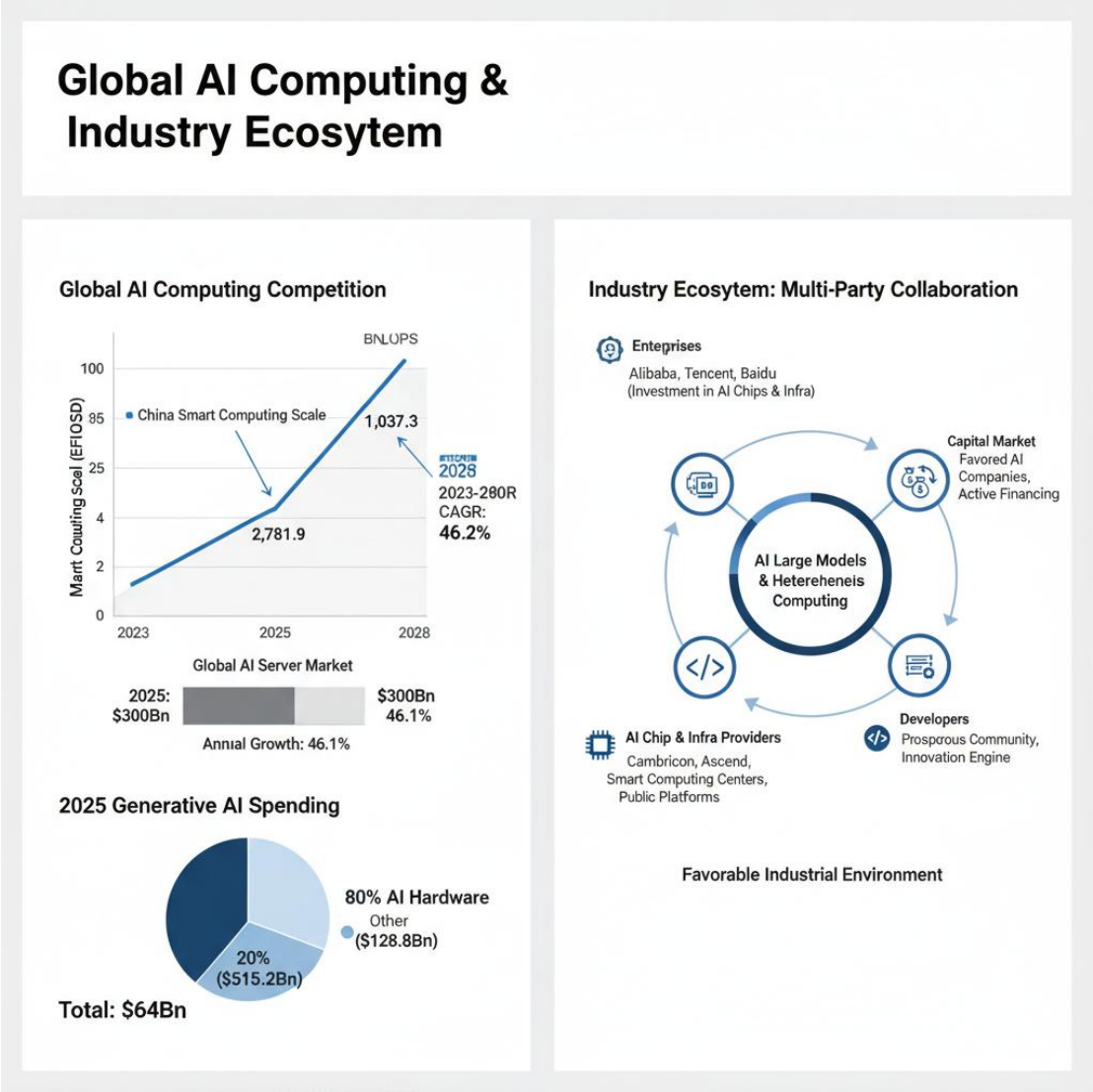
在全球范围内，各国政府纷纷出台政策支持 AI 和算力基础设施发展，形成了强有力的产业驱动力。中国将人工智能和算力基础设施纳入国家战略，明确提出加快数字化发展，建设数字中国。截至 2025 年 6 月，中国在用算力中心标准机架达 1085 万架，智能算力规模达 788EFLOPS（FP16 半精度），算力总规模位居全球第二。中研普华预测,2025-2030 年中国数据中心算力需求将以年均 20%的增速扩张，其中人工智能算力占比将从 30%提升至 50%。



中国算力基础设施建设进展

"东数西算"工程作为国家战略，已发展成为重大生产力布局战略工程。截至 2025 年 8 月，八大算力枢纽节点协同发展成效显著，规划建设超过 250 条"东数西算"干线光缆，集群间光层直达链路已拓宽至 1232 Tbps。2025 数博会期间，《关于进一步强化"东数西算"工程算力枢纽协同发展的联合倡议》发布，提出要

共建算力监测与调度体系，打破区域壁垒，统一技术标准与安全规范。根据规划，到 2025 年底，我国将初步建成综合算力基础设施体系，国家枢纽节点地区各类新增算力占全国新增算力的 60%以上。



全球 AI 算力竞争与产业生态构建

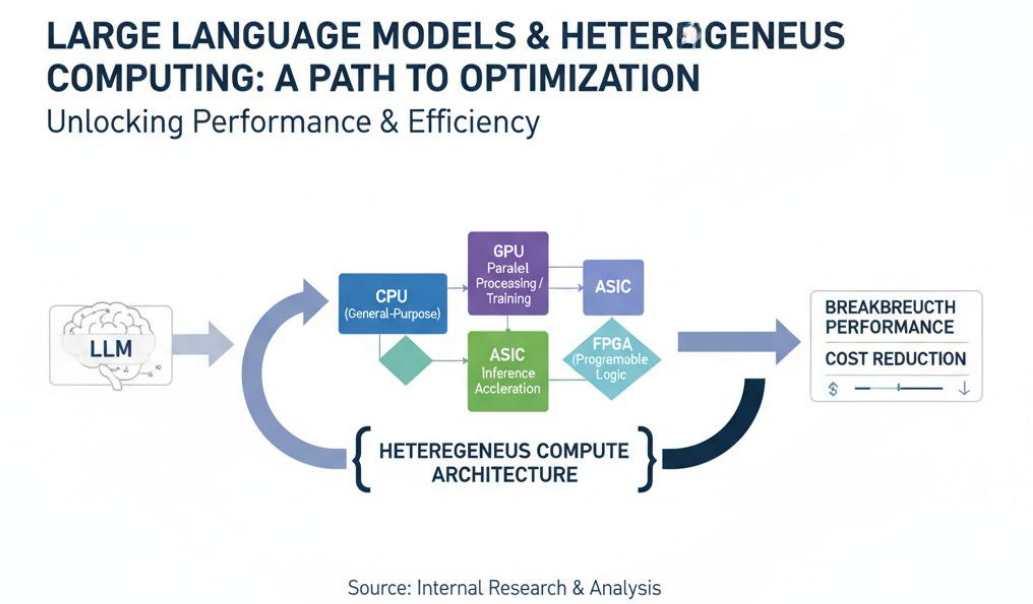
在国际层面，全球 AI 算力竞争日趋激烈。据 IDC 最新预测结果显示，2025 年中国智能算力规模将达到 1,037.3 EFLOPS，并在 2028 年达到 2,781.9 EFLOPS，2023-2028 年中国智能算力规模五年年复合增长率达 46.2%。全球 AI 服务器市场预计到 2025 年将达到 3,000 亿美元，年增长 46.1%。Gartner 预测 2025 年生成式 AI 支出将达 6440 亿美元，其中约 80%用于 AI 硬件。

产业层面，企业、资本、开发者多方参与生态构建。国内互联网巨头如阿里巴巴、腾讯、百度等纷纷加大在 AI 芯片和算力基础设施领域的投入；寒武纪、昇腾等国产 AI 芯片企业快速崛起；各类智算中心、AI 公共算力平台如雨后春笋

般涌现。资本市场上，AI 算力相关企业备受青睐，融资活动活跃。开发者社区日益繁荣，为技术创新提供了源源不断的动力。这种多方参与的生态构建，为 AI 大模型与异构算力的融合发展创造了良好的产业环境。

1.1.3 技术融合与开发者需求

面对大模型带来的算力挑战，单一架构的计算单元已难以满足需求，大模型与异构算力的深度融合成为突破性能瓶颈、降低成本的关键路径。异构计算通过集成不同类型的计算单元（如 CPU、GPU、FPGA、ASIC 等），发挥各自的优势，实现更高的性能和能效。例如，GPU 在大规模并行计算方面表现优异，适合大模型训练；ASIC 在特定任务上能效比极高，适合推理加速；FPGA 则具有灵活可编程的特性，能够适应不断变化的算法需求。通过异构计算架构，可以将不同类型的计算任务分配给最适合的处理单元，从而实现整体性能的最优化。



大模型与异构算力的深度融合

然而，异构算力的应用也带来了新的挑战。不同架构的硬件需要统一的编程模型、编译优化和任务调度机制；异构系统中的数据移动和通信开销可能成为新的瓶颈；开发者需要掌握多种硬件架构的编程技巧，学习曲线陡峭。这些问题使得大模型与异构算力的融合不仅仅是硬件层面的组合，更需要软件栈、编程模型、系统架构等多方面的协同创新。

异构算力融合面临的挑战与解决方案

在这一背景下，开发者亟需系统化的技术指南与实践参考。当前，关于大模型开发的资料虽然丰富，但大多聚焦于算法层面，对于如何在异构算力环境下高效部署和优化大模型的系统性指导相对缺乏。开发者需要了解不同硬件架构的特性、适用场景和性能表现；需要掌握异构环境下的编程模型和优化技巧；需要学习如何设计能够充分发挥异构算力优势的系统架构。本报告旨在填补这一空白，为开发者提供全面、实用的技术参考，推动大模型与异构算力的深度融合。

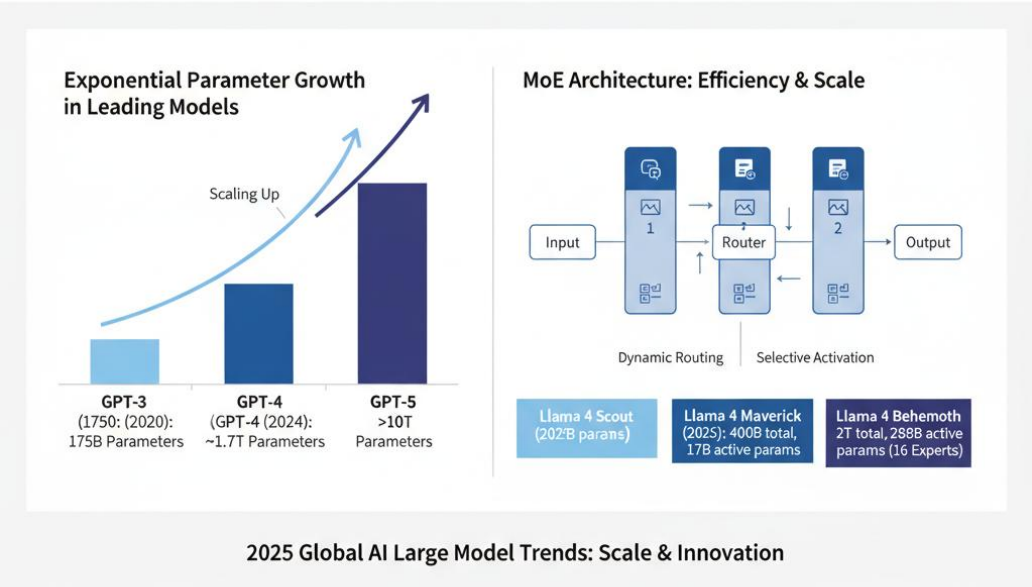
二、AI 大模型与算力行业现状

2.1 全球 AI 大模型发展概况

2.1.1 国际大模型技术演进

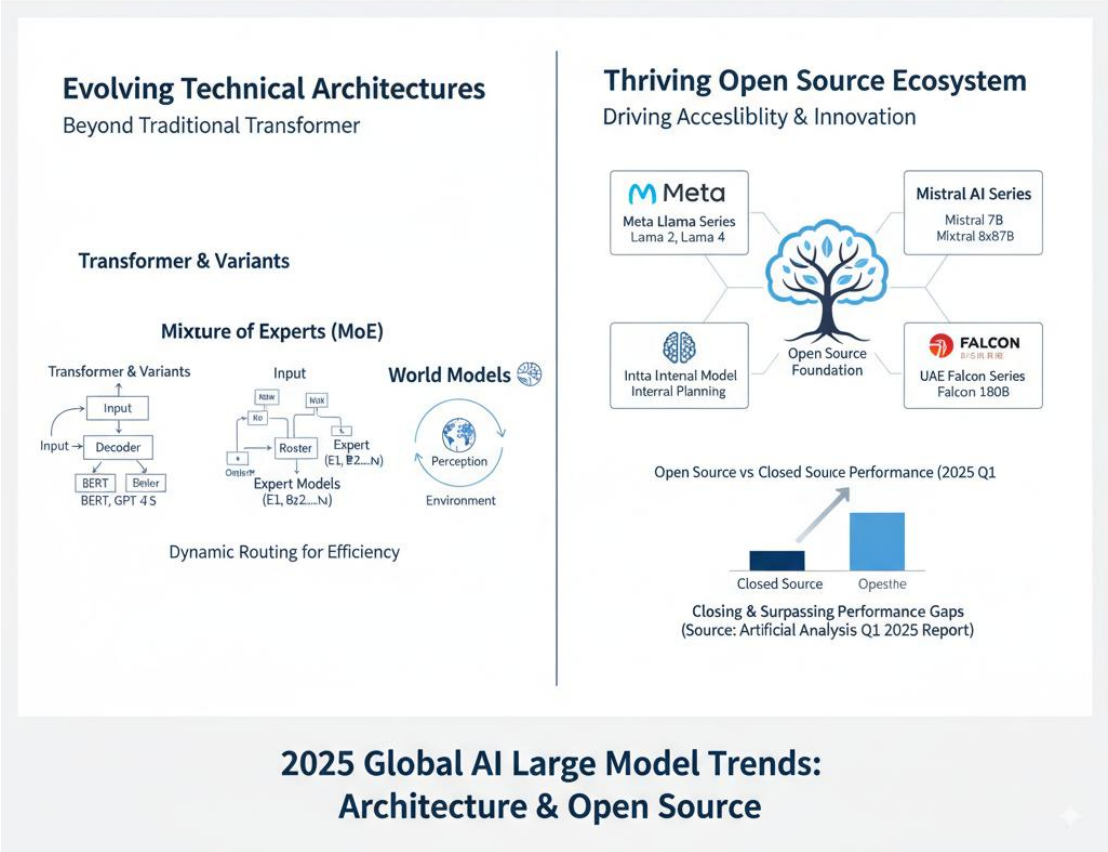
2025 年，全球 AI 大模型技术呈现出快速迭代、规模持续扩大、效率显著提

升的发展趋势。以 OpenAI 的 GPT 系列为代表，从 GPT-3 的 1750 亿参数发展到 GPT-4 的预估 1.7 万亿参数规模，再到 GPT-5 可能达到 3 至 50 万亿参数，模型参数量呈指数级增长。Meta 的 Llama 系列作为开源大模型的标杆，2025 年 4 月发布的 4.0 版本首次采用 MoE（Mixture of Experts）架构，提供了三个不同规模的版本：Llama 4 Scout（1090 亿参数）、Llama 4 Maverick（4000 亿总参数，170 亿激活参数）和 Llama 4 Behemoth（2 万亿总参数，2880 亿激活参数，16 个专家），展现了大模型架构的创新方向。



全球 AI 大模型参数规模的指数级增长与 MoE 架构的创新应用。

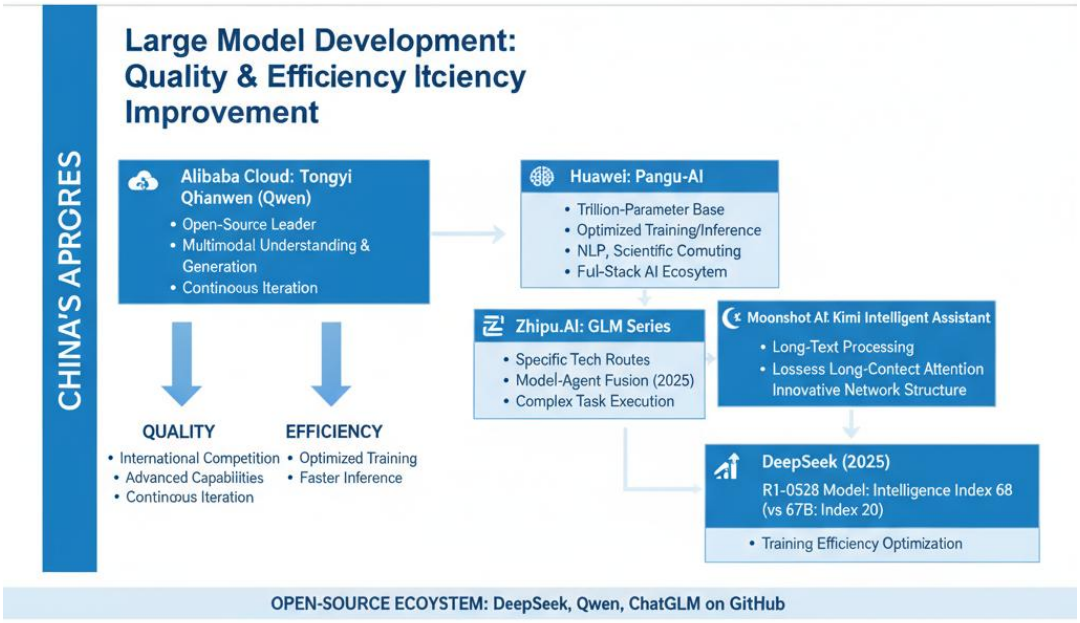
在技术架构方面，Transformer 已成为大模型的主流架构基础，同时各种创新变体不断涌现。MoE（混合专家模型）架构通过动态路由机制，在保持模型容量的同时显著降低了计算成本；世界模型（World Models）探索构建对环境的内部表征，为实现更通用的人工智能提供了新思路；多模态能力成为大模型的标配，从单一的文本处理扩展到图像、音频、视频等多种模态的理解和生成。2025 年 8 月，Anthropic 发布 Claude Opus 4.1，将编码性能提升至 SWE-bench Verified 基准测试的 74.5%，显著增强了深度研究和数据分析能力。



AI 大模型技术架构的演进与开源生态的繁荣。

开源生态的繁荣是国际大模型发展的另一重要特征。智谱的 GLM 系列、Meta 的 Llama 系列、阿里的 Qwen 系列、腾讯混元系列、Mistral AI 的 Mistral 系列、阿联酋的 Falcon 系列等开源模型的发布，极大地推动了大模型技术的普及和创新。这些开源模型不仅提供了强大的基础能力，还通过开放的权重和代码，为研究者和开发者提供了宝贵的实验平台，催生了大量基于开源模型的改进和应用。据 Artificial Analysis 公司 2025 年 Q1 报告显示，开源模型在性能上与闭源模型的差距正在缩小，在某些特定任务上甚至实现了超越。

2.1.2 国内大模型技术进展



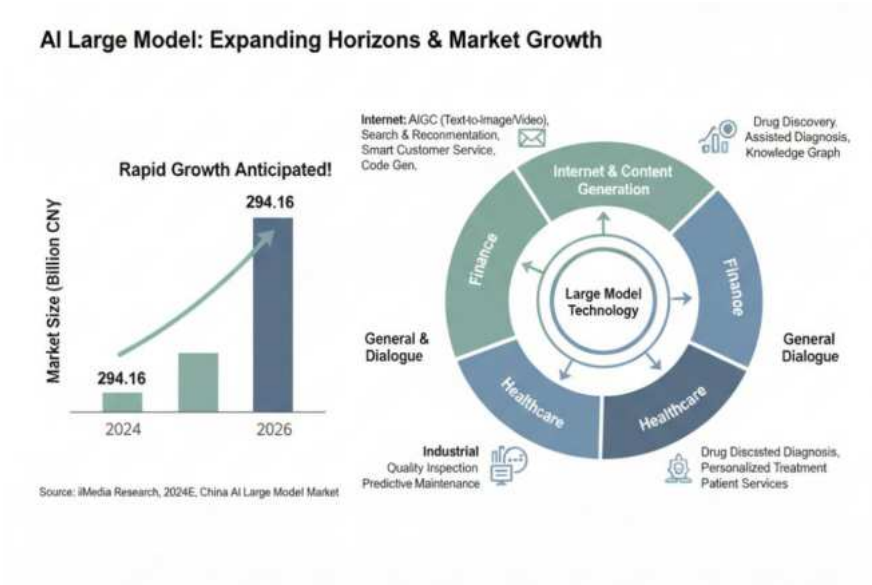
中国大模型“提质增效”及主要参与者

中国在大模型领域的发展呈现出"提质增效"的态势，涌现出一批具有国际竞争力的模型和产品。阿里巴巴的通义千问（Qwen）系列在开源社区备受关注，通过持续迭代优化，在多模态理解和生成方面取得显著进展。华为的盘古大模型在千亿级参数基础上，进一步优化了训练效率和推理性能，覆盖 NLP、科学计算等多个领域，并在华为的全栈 AI 生态中得到广泛应用。

智谱AI的GLM系列和月之暗面的KIMI智能助手代表了国内大模型在特定技术路线上的突破。KIMI通过创新的网络结构和工程优化，在长文本处理方面形成了差异化优势，实现了无损的长程注意力机制。GLM系列则在2025年进一步融合了原生Agent能力，实现模型与Agent的深度融合，提升了复杂任务的执行能力。

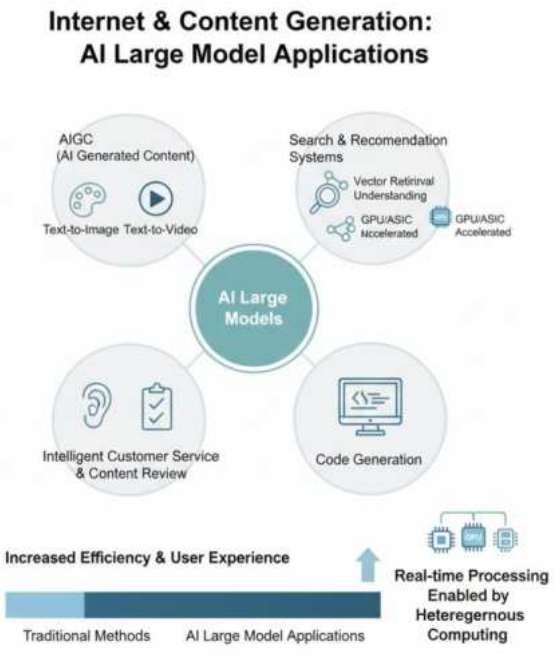
2025年，DeepSeek系列模型在国内外引起广泛关注，其R1-0528模型智能指数已达到68，相较于最初的67B模型有了显著提升，展现了中国在大模型训练效率优化方面的实力。国内大模型在开源生态方面也取得了显著进展，Deepseek、Qwen、ChatGLM等开源模型在GitHub等平台获得了大量关注和应用，形成了活跃的开发者社区。

2.1.3 大模型应用场景拓展



大模型市场规模及应用生态概览

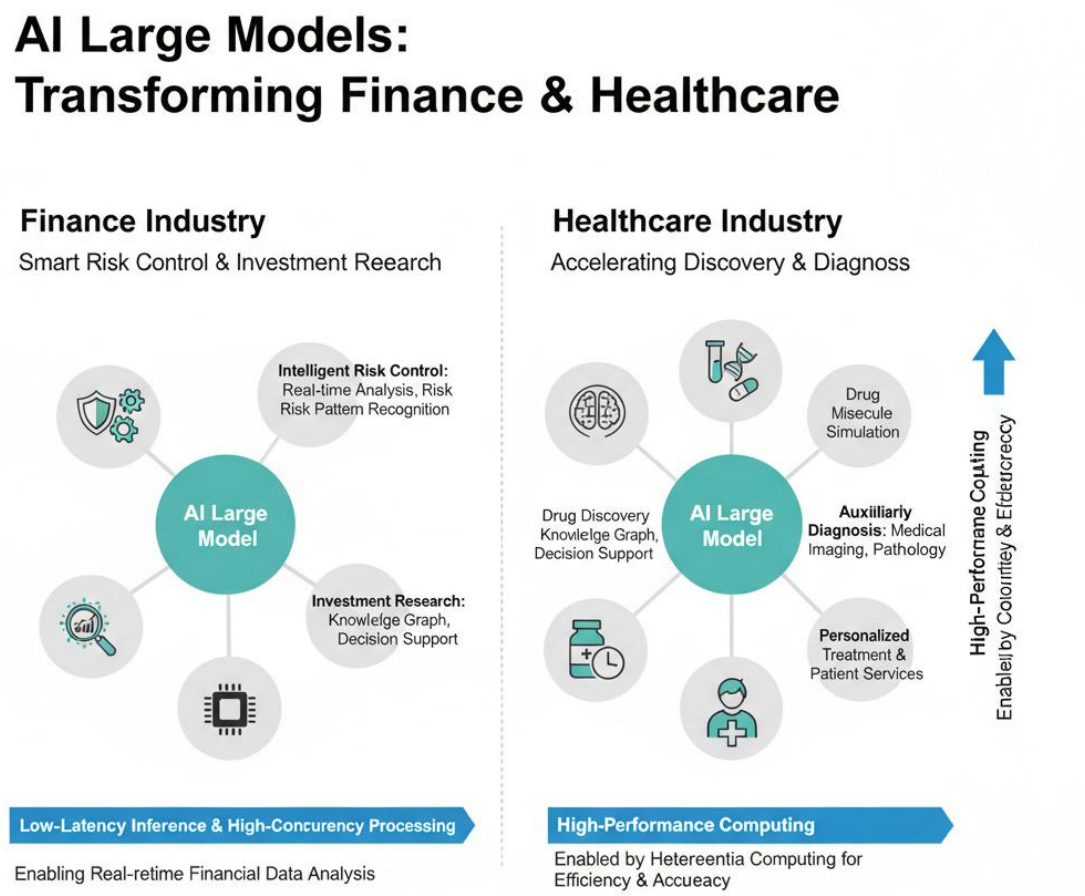
随着大模型技术的不断成熟，其应用场景也在不断拓展和深化。从最初的通用对话场景，逐步扩展到金融、医疗、工业等垂直行业，形成了丰富的应用生态。据艾媒咨询数据显示，2024 年中国 AI 大模型市场规模约为 294.16 亿元，预计 2026 年将持续快速增长。



互联网与内容生成领域的大模型应用

在互联网与内容生成领域，AIGC（AI 生成内容）应用蓬勃发展，包括文生图、文生视频等应用，异构算力的支持使得实时生成成为可能。大模型搜索与推荐系统通过向量检索、语义理解等技术，GPU/ASIC 加速推荐系统推理，提升了

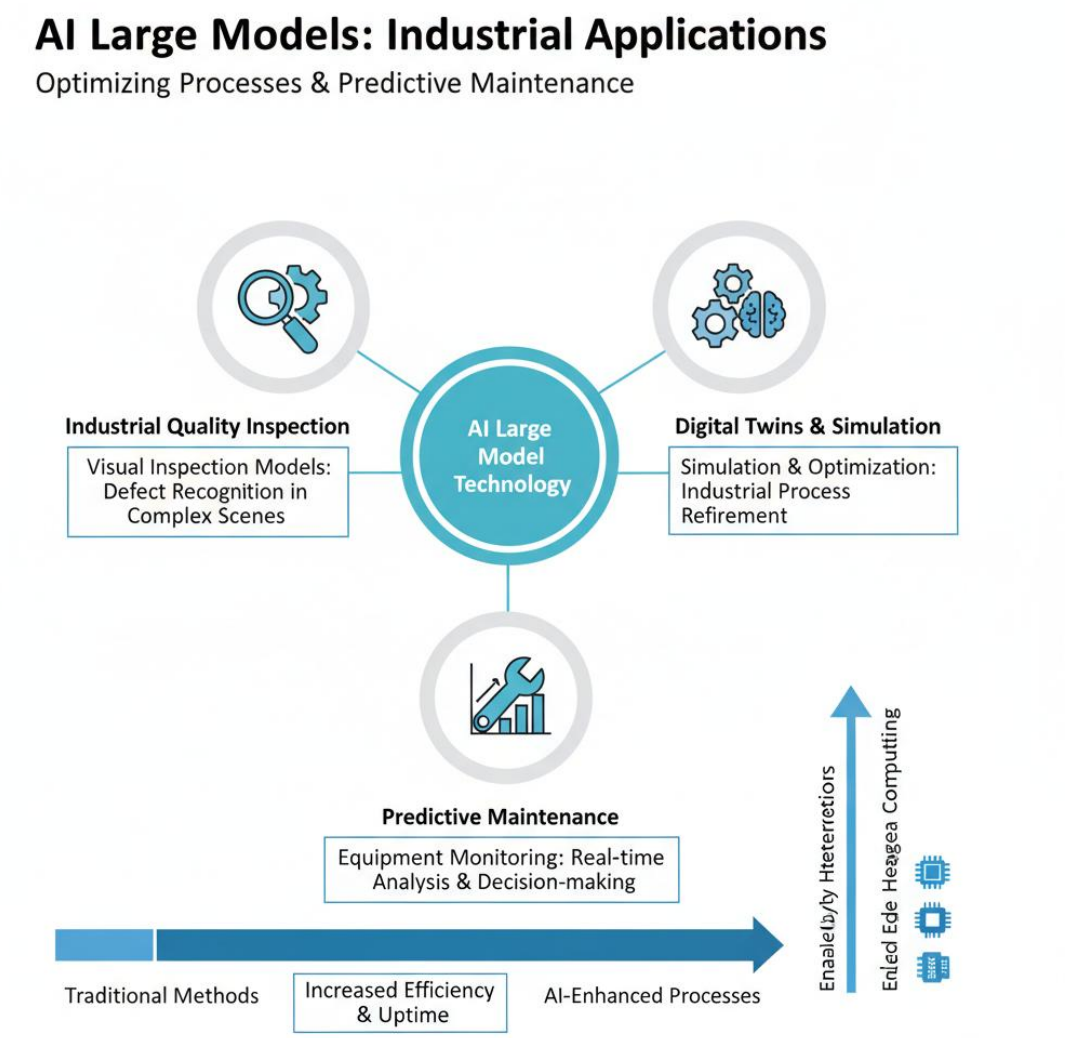
用户体验和系统效率。智能客服、内容审核、代码生成等应用也在互联网企业中得到广泛应用，大幅提升了业务效率和用户体验。



金融与医疗领域的大模型应用

在金融领域，大模型与知识图谱结合，在智能风控与投研方面发挥重要作用。低延迟推理、高并发处理能力使得大模型能够实时分析海量金融数据，识别风险模式，辅助投资决策。国产 AI 芯片在金融客户案例中表现出色，为金融行业的智能化转型提供了有力支撑。

在医疗领域，大模型应用场景迅速拓展，涵盖药物发现、辅助诊断、个性化治疗、医患服务等各个方面，展现出加快药物开发、早期发现疾病、提升诊疗效率的巨大潜力。医学影像分析、病理诊断、药物分子模拟等应用对算力要求极高，异构算力的引入显著提升了处理效率和准确性。



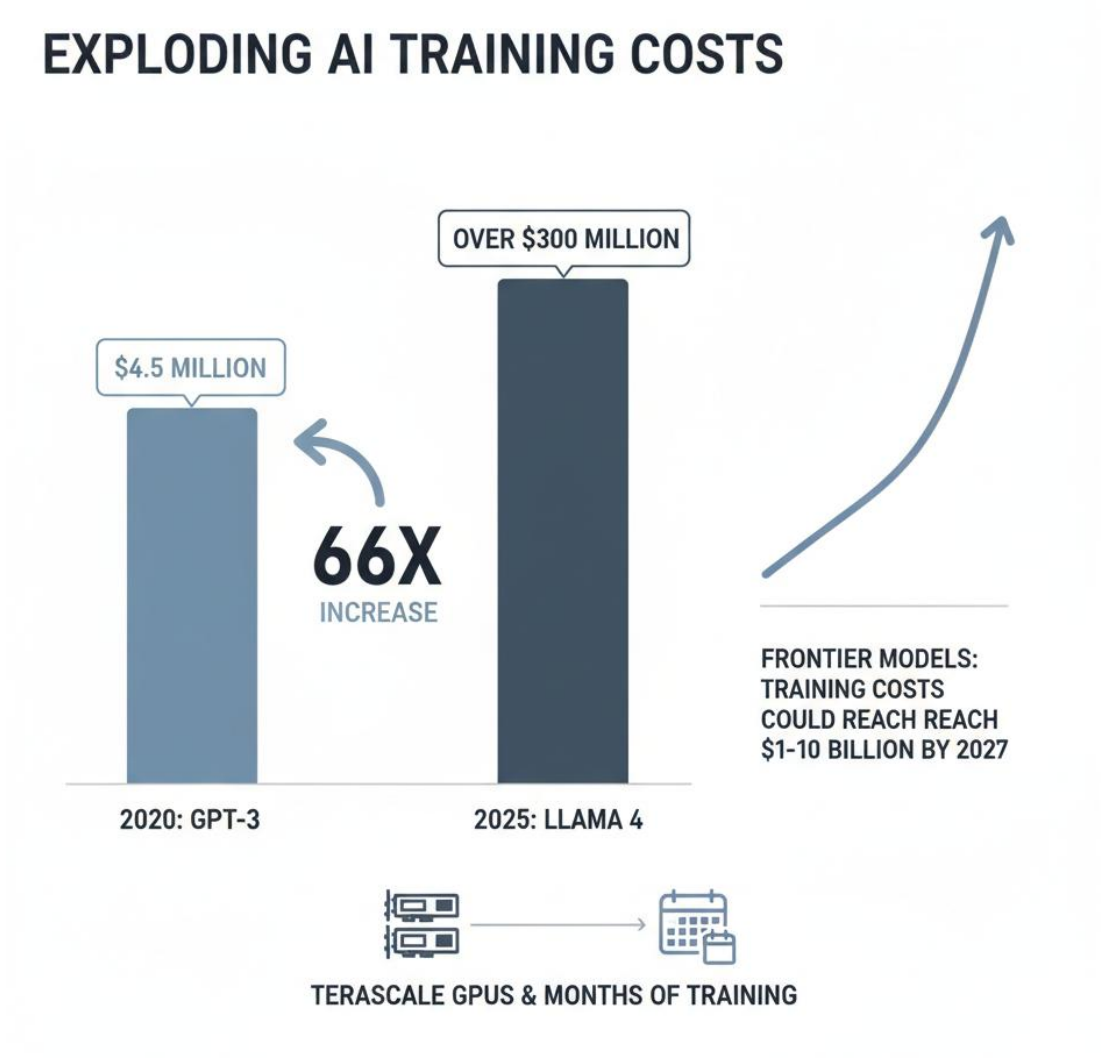
工业领域的大模型应用

在工业领域，大模型在工业质检、数字孪生、设备预测性维护等方面发挥重要作用。视觉质检大模型能够识别复杂工业场景中的缺陷，数字孪生技术通过大模型仿真优化工业流程，边缘异构算力的部署使得实时分析和决策成为可能。

2.2 算力需求爆发与挑战

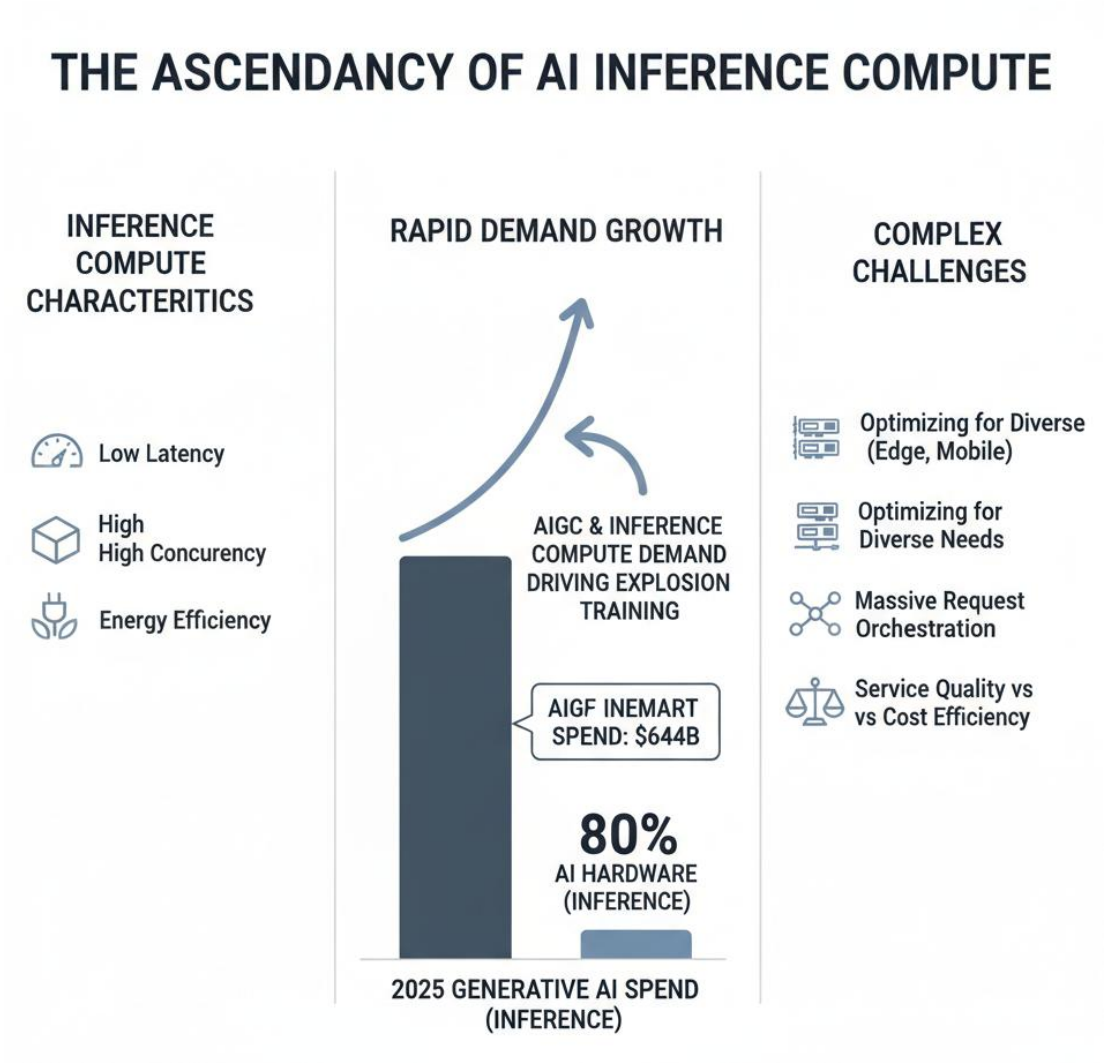
2.2.1 训练与推理算力需求分析

大模型训练对算力的需求呈现出前所未有的增长态势前沿模型的训练成本正以惊人的速度膨胀，Anthropic CEO 预测训练成本可能在 2027 年达到 100 亿至 1000 亿美元级别。千亿参数模型训练一般需要上千张高性能 GPU 卡支撑，训练时间长达数月，对算力基础设施提出了极高要求。



大模型训练成本的爆炸式增长

推理场景对算力的需求同样快速增长，但特点与训练有所不同。推理更注重低延迟、高并发和能效比。在实际应用中，大模型推理需要同时服务大量用户，对并发处理能力提出高要求；在实时交互场景，如智能客服、实时翻译等，对响应延迟极为敏感；在边缘设备和移动终端，对能耗和计算效率有严格限制。这些多样化的需求使得推理算力的优化和调度面临复杂挑战。

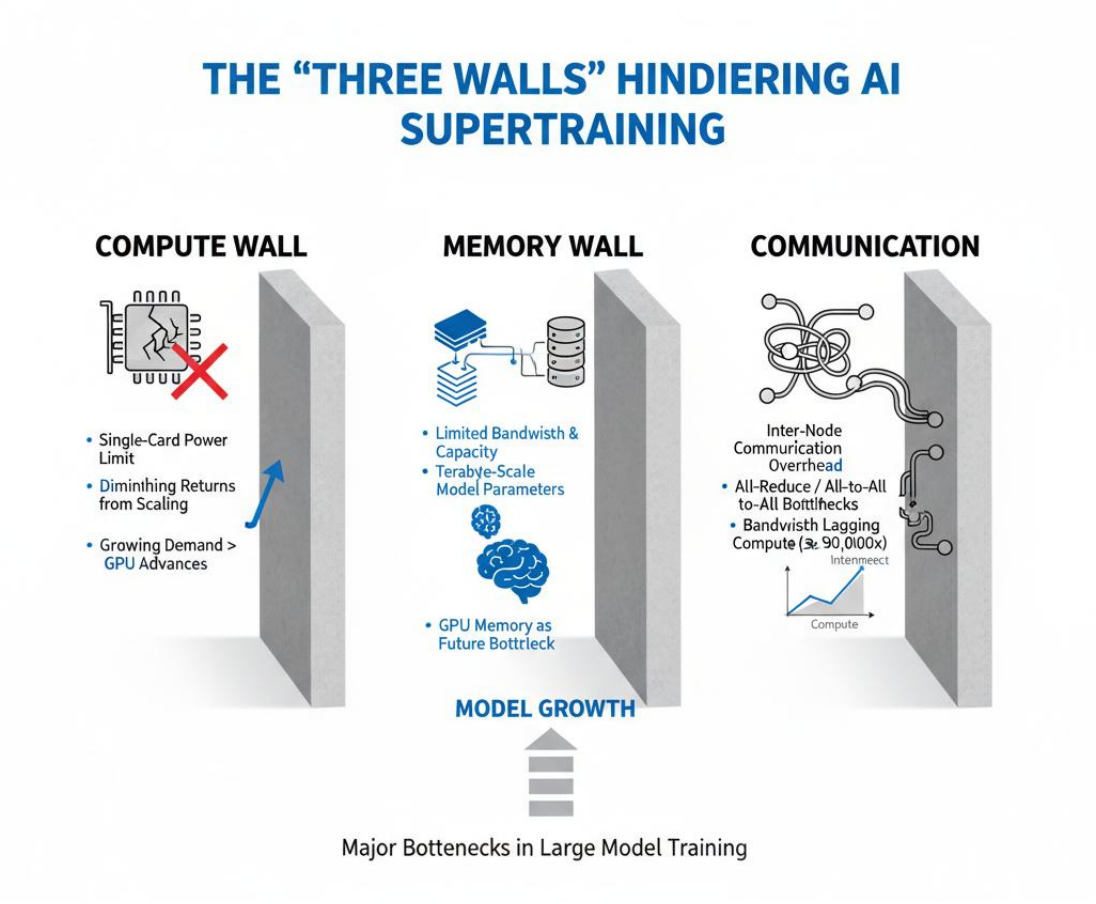


推理算力需求的特征、增长及其挑战

随着大模型应用的普及，推理算力的总需求已超过训练算力，成为算力消耗的主要部分。特别是在 AIGC、智能助手等大规模应用场景，推理算力需求呈现爆发式增长。Gartner 预测 2025 年生成式 AI 支出将达 6440 亿美元，其中约 80%用于 AI 硬件，主要用于推理场景。如何高效满足海量推理请求，同时保证服务质量和成本效益，成为算力基础设施面临的重要课题。

2.2.2 算力墙、存储墙、通信墙

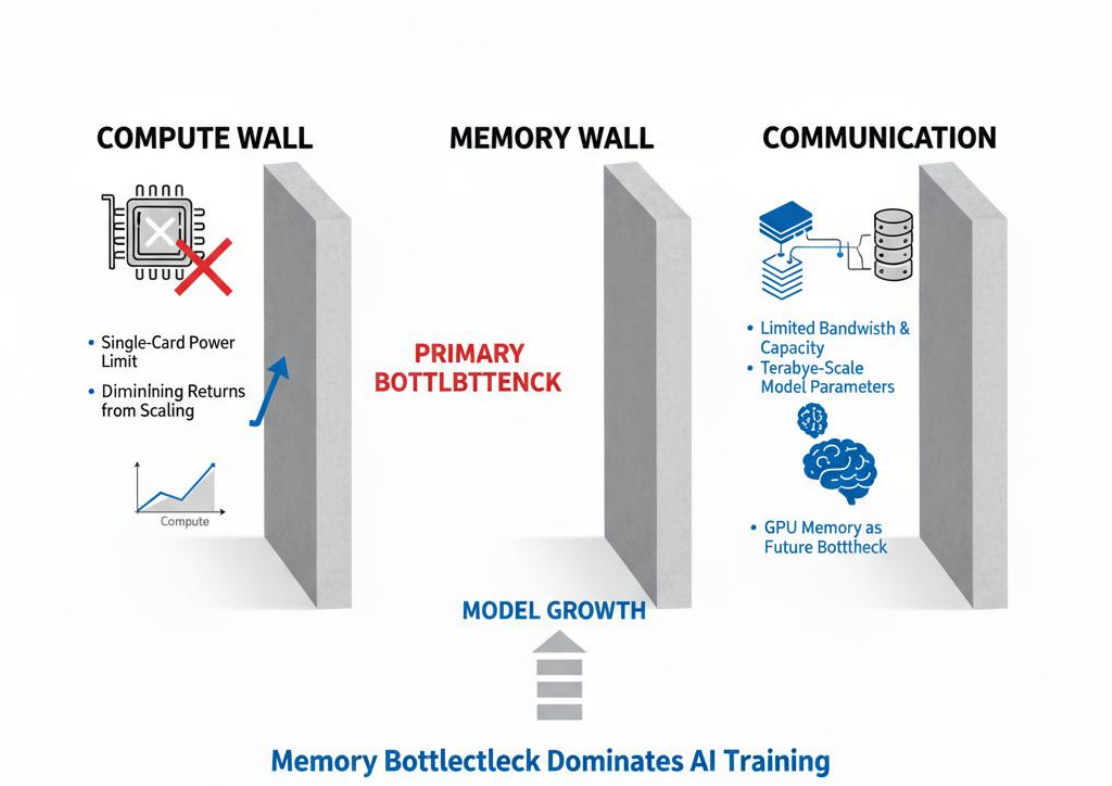
在大模型训练过程中，"三堵墙"——算力墙、存储墙和通信墙成为制约性能的主要瓶颈。



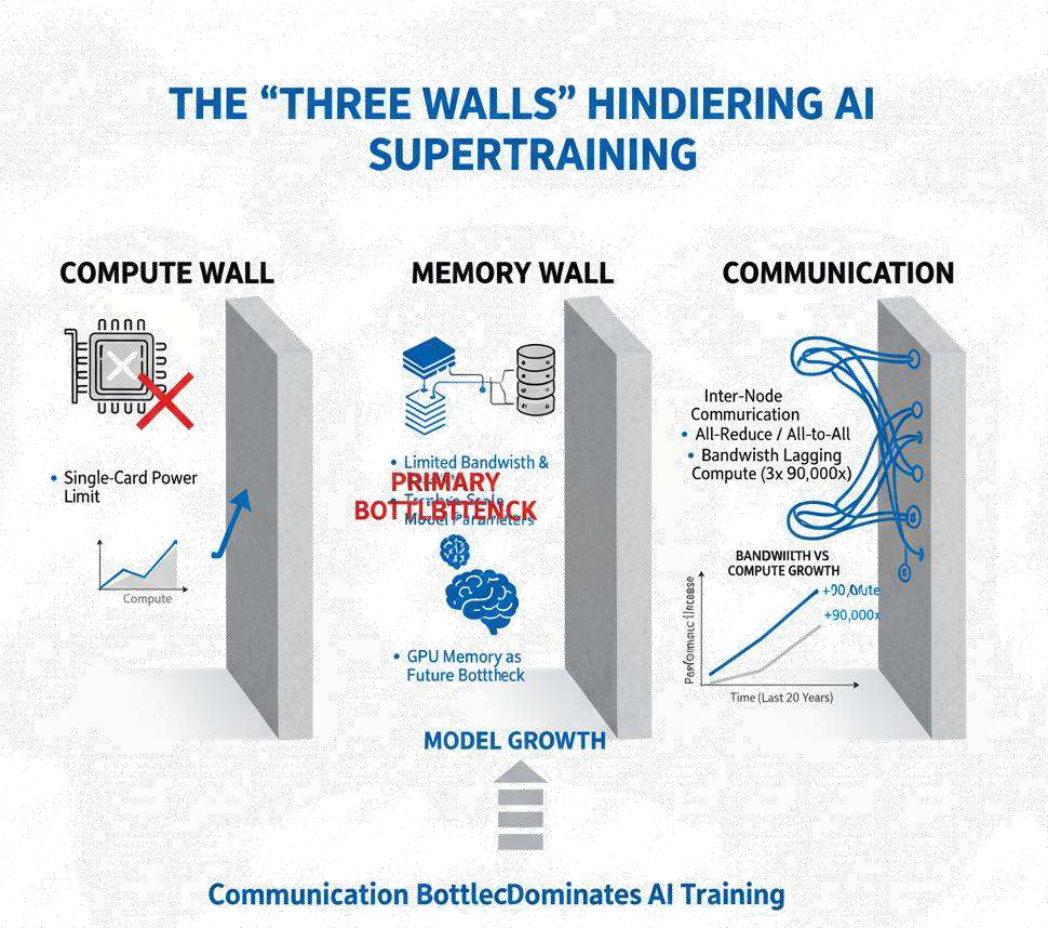
Major Bottenecks in Large Model Training

算力墙指的是单卡算力上限的限制，即使是最先进的 GPU 芯片，其计算能力也难以满足大模型训练的需求，必须通过大规模集群扩展算力。然而，随着模型规模的增长，单纯增加计算单元的效果递减，算力墙问题日益突出。

存储墙主要体现在内存带宽和容量的限制上。大模型参数量巨大，万亿参数模型需要数百 GB 到数 TB 的内存容量，而当前 AI 加速器的内存容量和带宽往往成为瓶颈。研究表明，AI 训练未来的瓶颈可能不是算力，而是 GPU 内存，内存墙问题已成为制约大模型发展的关键因素。数据加载、参数交换等内存密集型操作往往成为训练过程中的性能瓶颈。



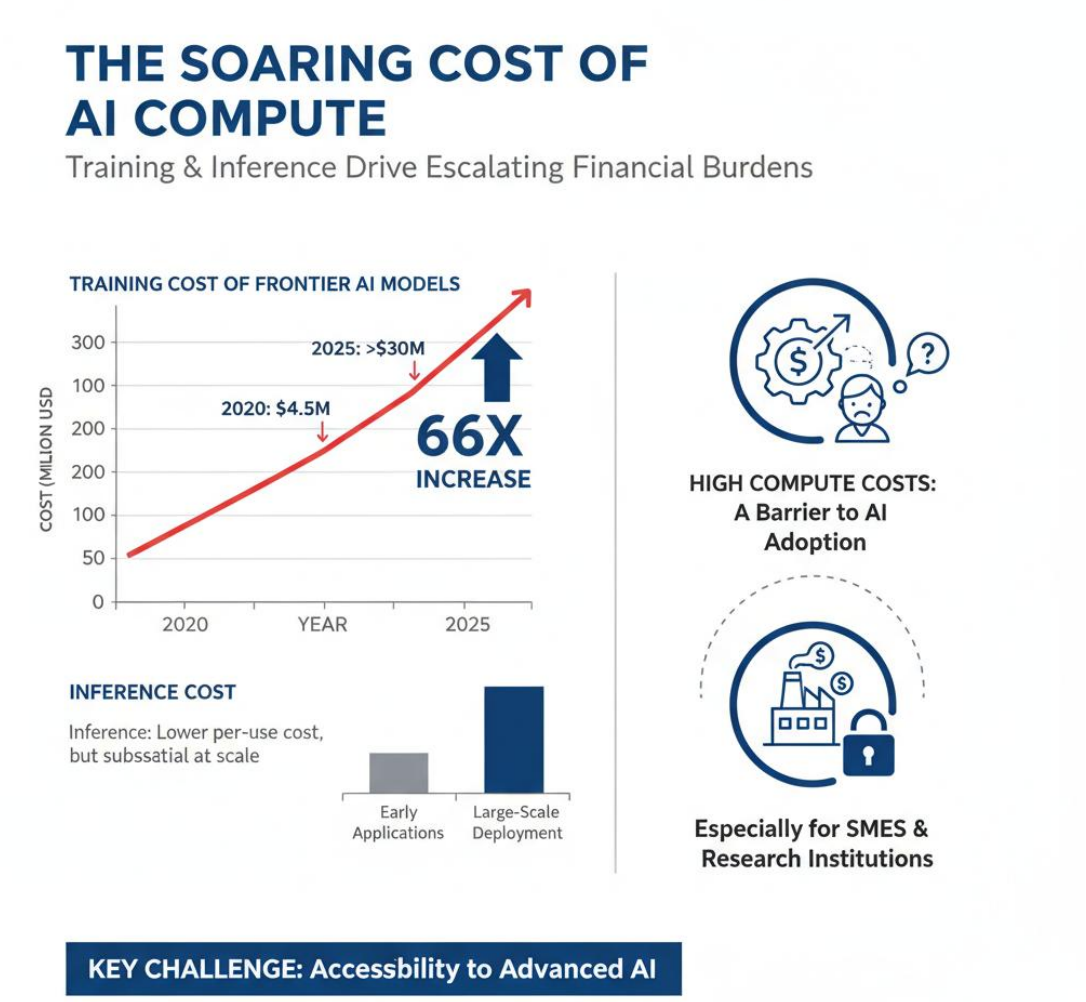
通信墙则是指集群网络通信开销的限制。大模型并行训练需要大量节点间通信，如 AllReduce 梯度同步、AlltoAll 参数交换等，通信性能直接决定训练效率。无论是芯片内部、芯片间，还是 AI 加速器之间的通信，都已成为 AI 训练的瓶颈。扩展带宽的技术难题尚未被完全攻克，过去 20 年间，运算设备的算力提高了 90,000 倍，而互连带宽仅提高了 30 倍，通信墙问题日益严峻。



"通信墙"作为大模型训练的限制因素，对比算力与带宽增长的速度差异

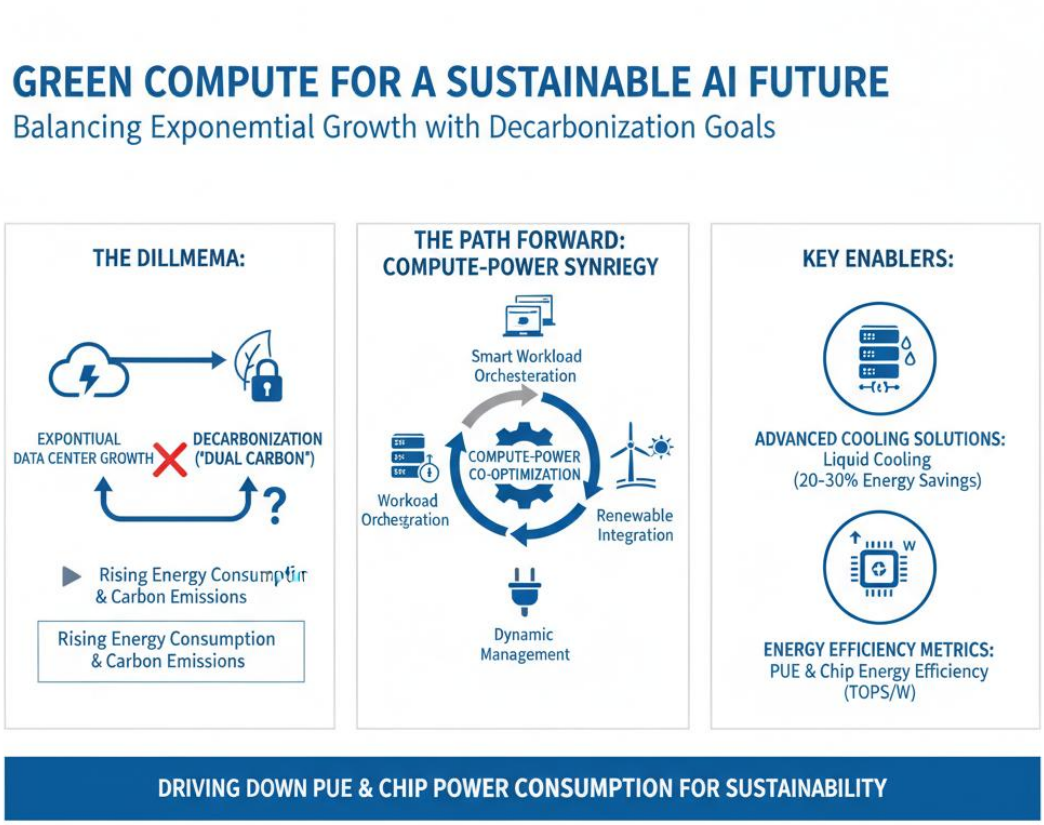
2.2.3 算力成本与能效挑战

大模型训练和推理的高算力需求带来了巨大的成本压力。前沿模型的训练成本从 2020 年的 450 万美元增长到 2025 年的 3 亿美元以上，增长了约 66 倍。推理成本虽然相对较低，但随着应用规模的扩大，总体成本仍然可观。高昂的算力成本成为大模型技术普及和应用落地的重要障碍，特别是对于中小企业和科研机构而言。



大模型训练和推理成本的快速增长，以及高算力需求带来的成本压力

数据中心能耗与"双碳"目标之间的矛盾日益凸显。算力需求呈指数级增长趋势，带来了数据中心能耗、成本以及碳排放的不断攀升。在"双碳"目标约束下，算电协同（算力与电力协同优化）正成为破解 AI 能耗困局、实现数据中心绿色可持续发展的关键路径。液冷技术作为降低数据中心能耗的重要手段，比传统电制冷节能 20%-30%，正得到广泛应用。

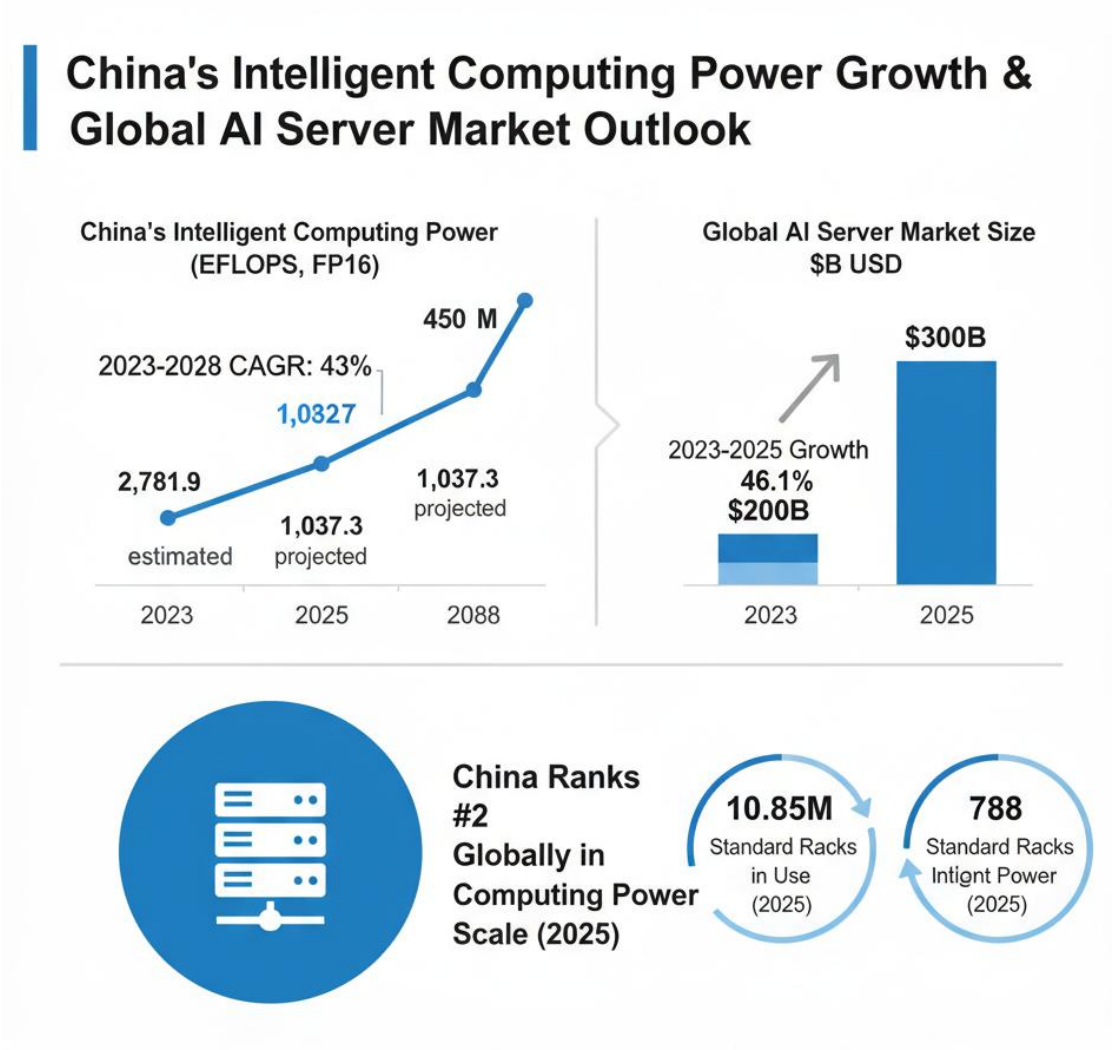


数据中心能耗与“双碳”目标之间的矛盾，以及算电协同、液冷技术和能效比提升实现绿色算力的解决方案

能效比成为衡量算力基础设施的重要指标。传统的以性能为中心的设计理念正在向以能效为中心转变，绿色算力成为行业发展的重要趋势。液冷技术、可再生能源应用、算力调度优化等节能技术得到广泛应用，数据中心 PUE（Power Usage Effectiveness）值不断降低。同时，芯片能效比（TOPS/W）的提升也成为 AI 芯片设计的重要目标，通过架构创新、制程工艺优化等手段，在提升算力的同时降低能耗。

2.3 国内外算力基础设施对比

2.3.1 全球算力规模与分布

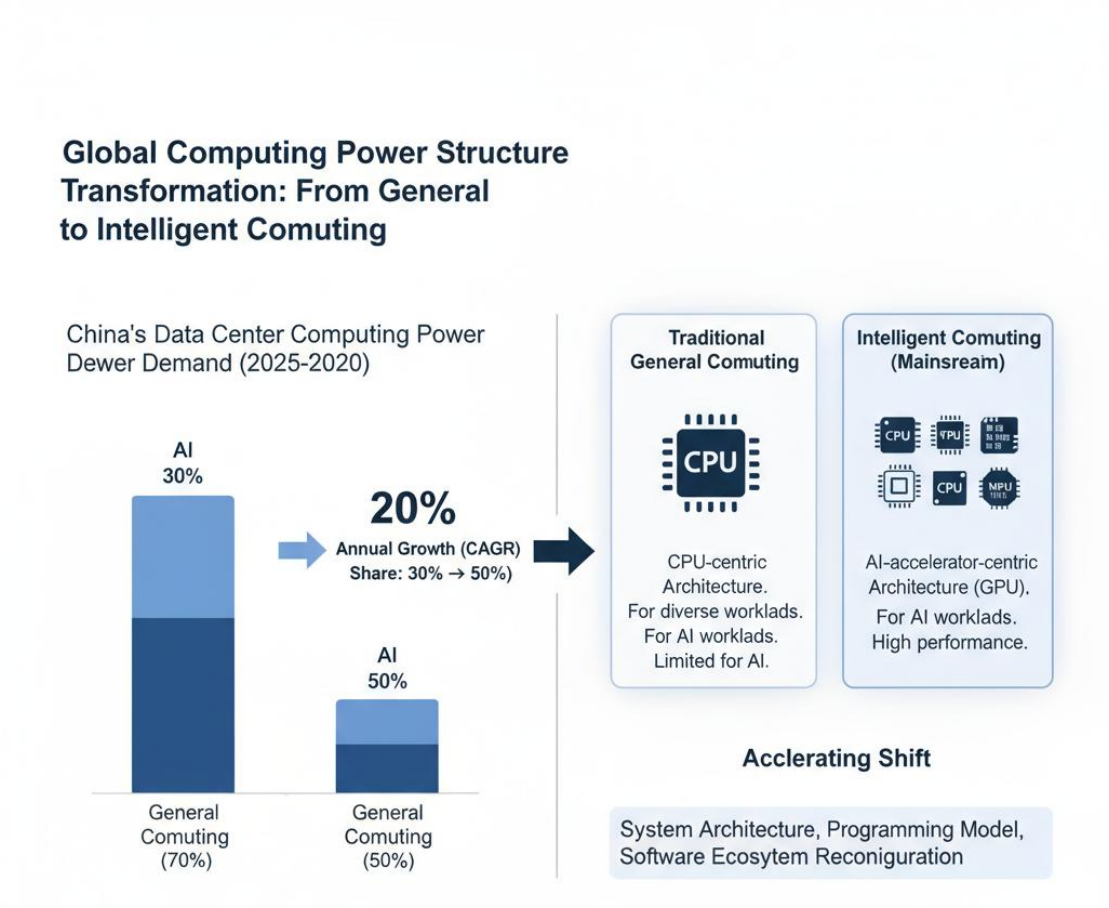


中国智能算力规模增长预测及全球 AI 服务器市场展望

根据最新数据，截至 2025 年 6 月，中国在用算力中心标准机架达 1085 万架，智能算力规模达 788EFLOPS（FP16 半精度），算力总规模位居全球第二。IDC 预测，2025 年中国智能算力规模将达到 1,037.3 EFLOPS，并在 2028 年达到 2,781.9 EFLOPS，2023-2028 年中国智能算力规模五年年复合增长率达 43%。全球 AI 服务器市场预计到 2025 年将达到 3,000 亿美元，年增长 46.1%。

全球算力地域分布与中国面临的挑战

从地域分布来看，美国在高端芯片和算力基础设施方面仍占据主导地位，拥有最先进的 AI 芯片制造能力和大规模的算力集群。按照近 6 年 AI 服务器算力总量估算，美国和中国算力全球占比分别为 34%和 33%。中国在算力规模上已位居全球第二，但在高端芯片、核心软件等方面与国际先进水平仍有差距。特别是在先进制程芯片方面，受制于技术封锁和供应链限制，中国在高端 AI 芯片领域面临挑战，这也促使中国加速自主研发和替代进程。

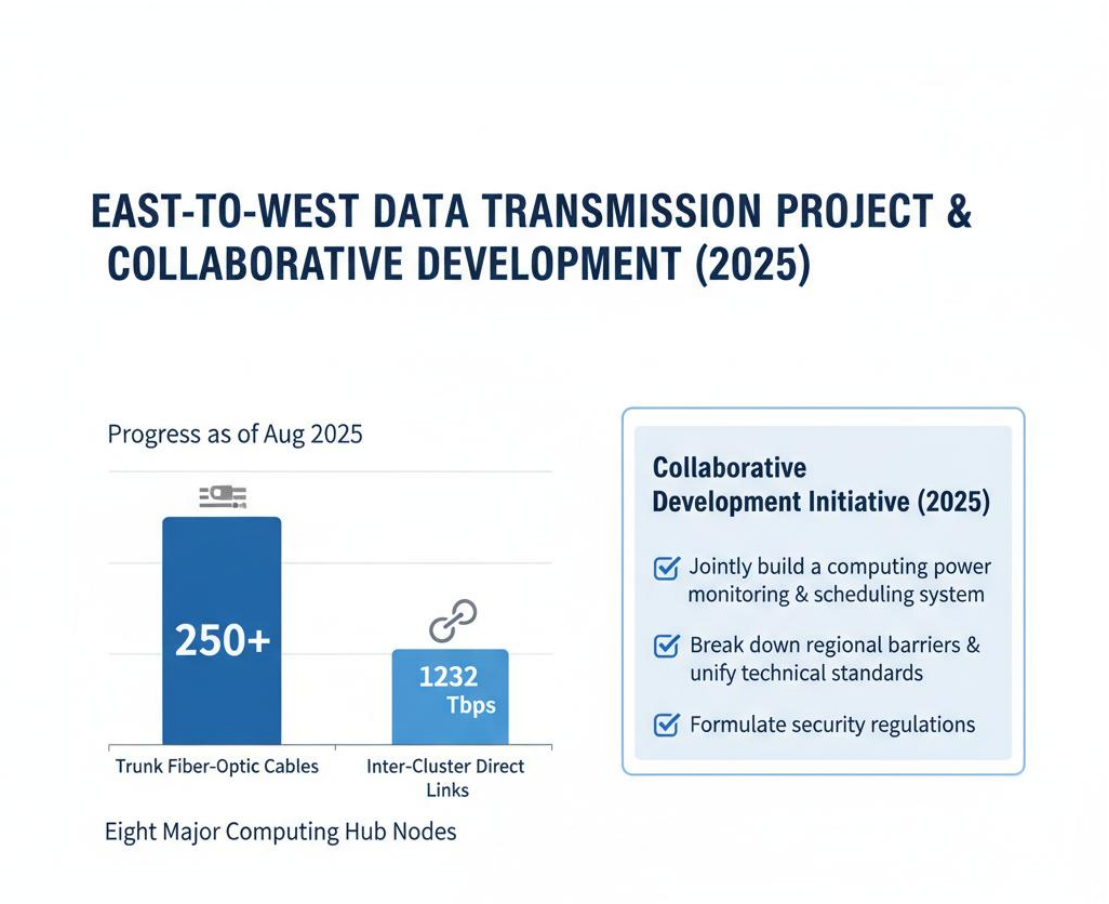


全球算力结构转型：从通用计算到智能计算

从算力结构来看，全球算力正从通用计算向智能计算加速转变。传统以 CPU 为中心的通用计算架构难以满足 AI 工作负载的需求，以 GPU、TPU、NPU 等专用 AI 加速器为核心的智能计算成为主流。这种转变不仅体现在硬件层面，也反映在系统架构、编程模型、软件生态等各个方面，推动整个计算产业的重构。

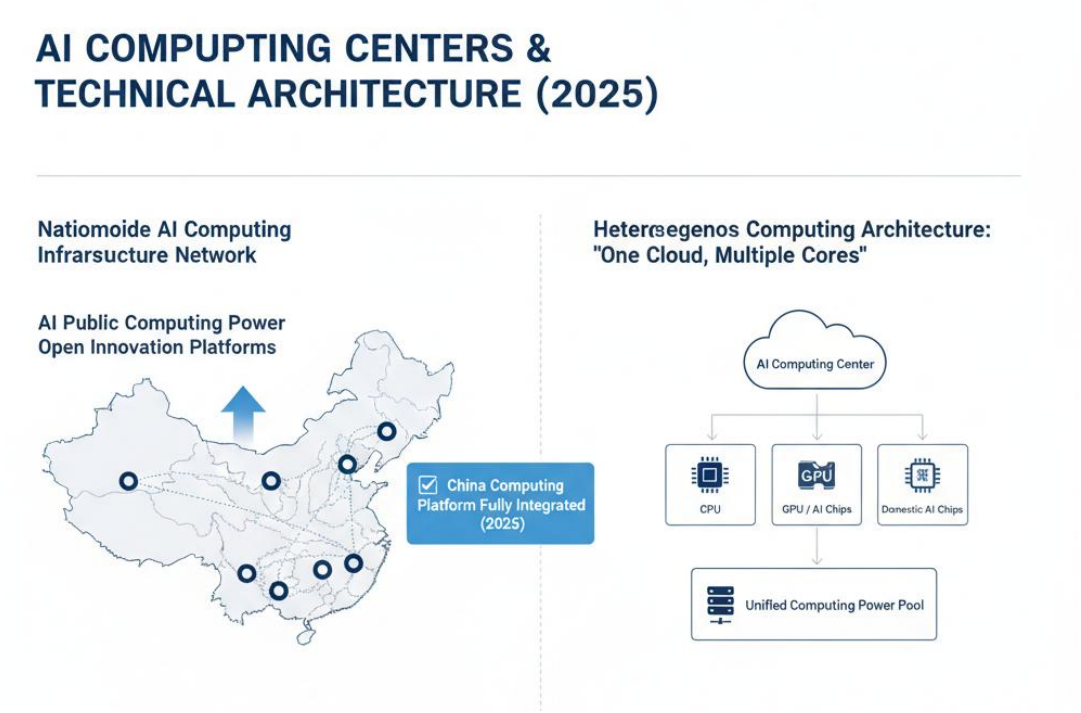
2.3.2 国内智算中心建设

我国智算中心建设近年来取得了显著进展。中国已初步形成 1ms 时延城市算力网、5ms 时延区域算力网、20ms 时延跨国家枢纽节点算力网，算力网络建设成效显著。



"东数西算"工程进展与协同发展

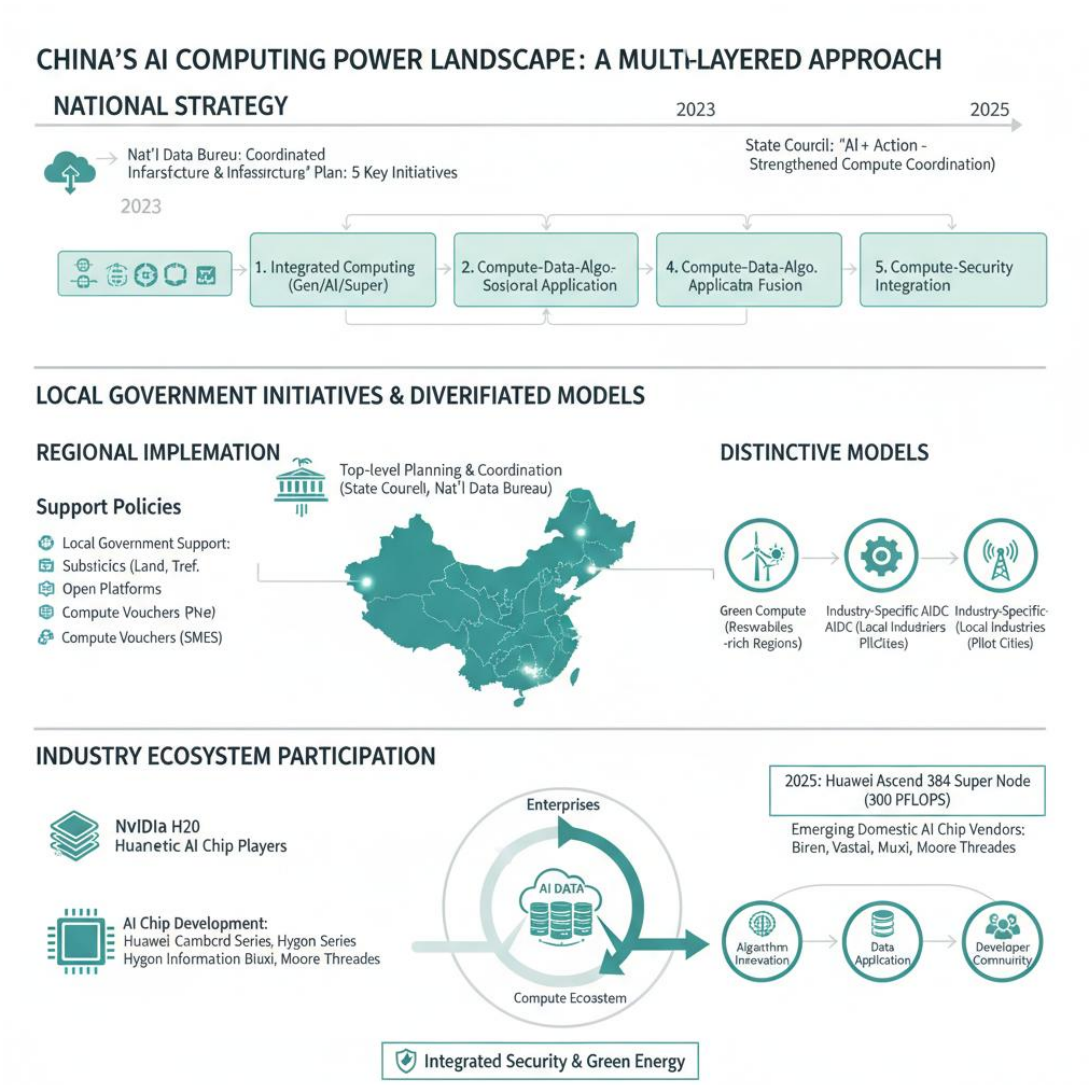
在智算中心建设方面，国家新一代 AI 公共算力开放创新平台相继建成，为 AI 研发和应用提供了强大的算力支撑。各地智算中心建设如火如荼，形成了覆盖全国的算力基础设施网络。2025 中国算力大会上，中国算力平台全面贯通，标志着一个国家级算力调度和管理体系的基本建成。从技术架构看，国内智算中心普遍采用异构计算架构，支持 CPU、GPU、国产 AI 芯片等多种计算单元，实现"一云多芯"的技术路线。



智算中心建设与技术架构

2.3.3 政策支持与地方实践

国家层面出台了一系列政策支持算力基础设施发展。2025 年，国务院《关于深入实施"人工智能+"行动的意见》提出强化智能算力统筹。国家数据局统筹推进算力基础设施建设，推动算力资源的优化配置和高效利用。2023 年 12 月，国家五部委联合印发《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》，从五大方面统筹推进算力网建设：通用算力、智能算力、超级算力一体化布局，东中西部算力一体化协同，算力与数据、算法一体化应用，算力与绿色电力一体化融合，算力发展与安全保障一体化推进。



地方政府也积极响应国家战略，出台了一系列支持政策。各地通过智算中心补贴、电价优惠、开放平台等措施，吸引算力相关企业和项目落地。例如，一些地区对新建智算中心给予土地、税收等方面的优惠；一些地区通过"算力券"等方式，降低中小企业使用算力的成本；一些地区则重点支持算力应用创新，推动算力与产业深度融合。

在地方实践中，形成了各具特色的发展模式。一些地区依托丰富的可再生能源资源，发展绿色算力；一些地区则结合本地产业特点，建设行业专用智算中心；一些地区注重算力与网络的协同发展，构建算力网络体系。这些多样化的实践探索，为中国算力基础设施的高质量发展提供了宝贵经验。

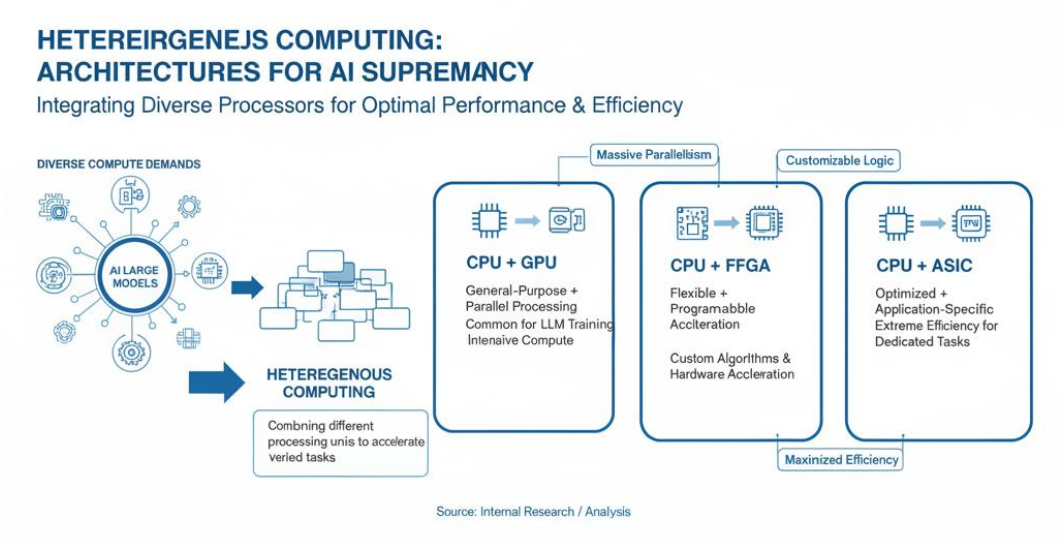
产业层面，企业、资本、开发者多方参与生态构建。国内 AI 芯片市场参与者主要有英伟达 H20、华为昇腾系列、寒武纪思元系列、海光信息 DCU 系列等。2025 年，华为首次展出昇腾 384 超节点真机，其算力总规模达 300PFLOPS，展

现了国产 AI 算力的技术实力。寒武纪、壁仞、燧原、沐曦和摩尔线程等国产 AI 芯片厂商也各具特色，共同推动国产异构算力生态的繁荣发展。

2.4 异构算力成为主流趋势

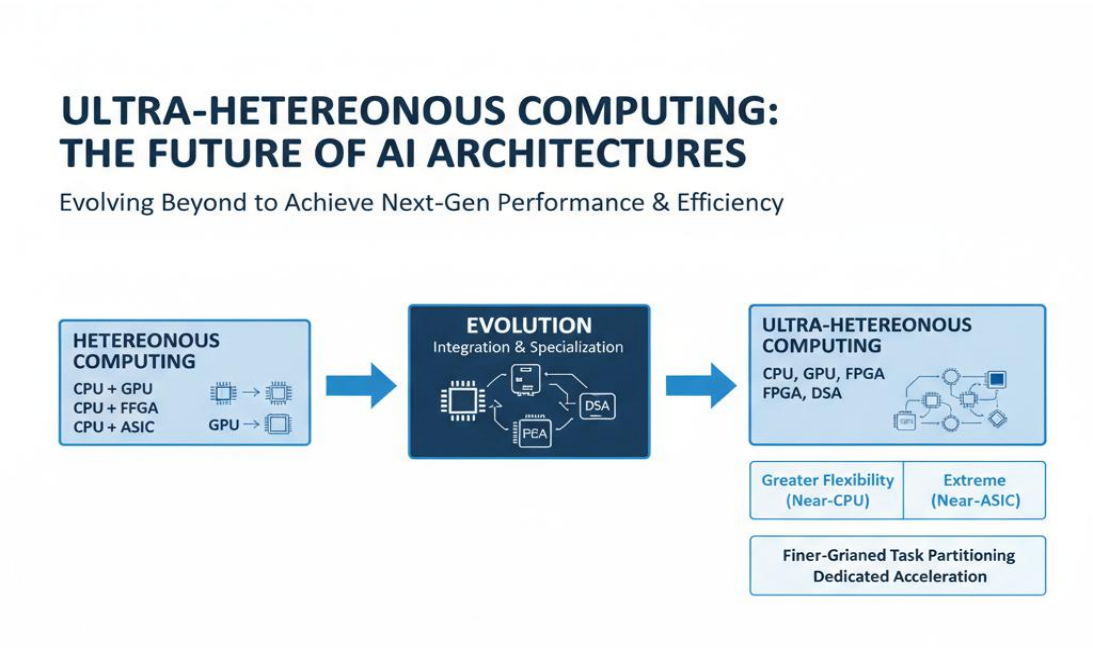
2.4.1 异构计算定义与分类

异构计算是指在同一计算系统集成不同类型或架构的处理单元，以便更有效地执行不同类型的任务。随着 AI 大模型对算力需求的多样化，单一架构的计算单元难以满足所有需求，异构计算通过组合不同特性的计算单元，实现整体性能的最优化。根据组合方式的不同，异构计算主要分为三类：CPU+GPU、CPU+FPGA 和 CPU+ASIC。



异构计算概述及其三大主要类型

CPU+GPU 是最常见的异构计算组合，CPU 负责通用计算和任务调度，GPU 负责大规模并行计算。这种组合充分利用了 GPU 在并行计算方面的优势，适合大模型训练等计算密集型任务。CPU+FPGA 组合则利用 FPGA 的灵活可编程特性，适合需要定制化加速的场景，如特定算法的硬件加速。CPU+ASIC 组合则针对特定应用进行深度优化，如 TPU（Tensor Processing Unit）专门用于加速 TensorFlow 计算，能效比极高。

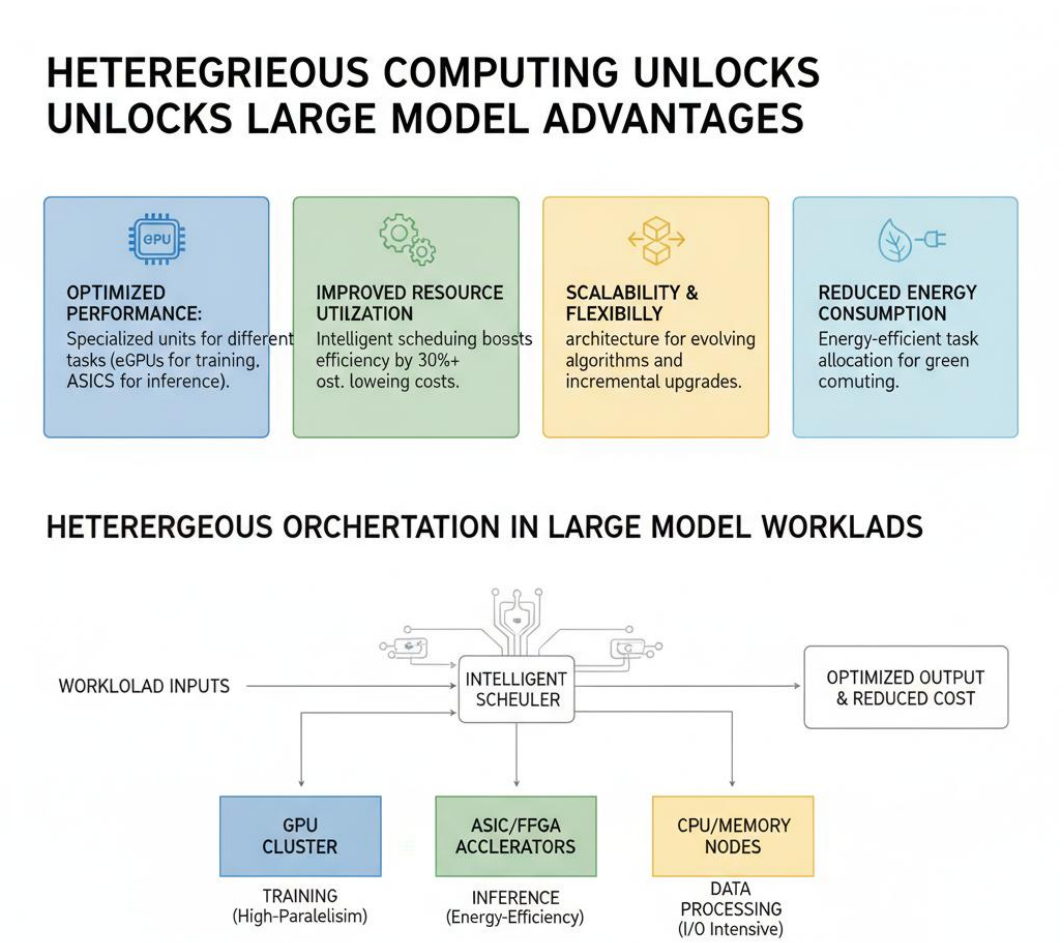


从异构计算到超异构计算的演进与优势

超异构计算是异构计算的进一步发展，由 CPU、GPU、FPGA 和 DSA（Domain-Specific Architecture）多架构处理器组成，目标是接近 CPU 的灵活性和 ASIC 的性能效率。超异构计算架构通过更加精细的任务划分和专用加速，实现更高性能和能效，成为未来计算架构的重要发展方向。

2.4.2 异构算力在大模型场景优势

异构算力在大模型场景中展现出显著优势：



一、不同类型的计算单元擅长大模型不同环节，GPU 在大规模并行计算方面表现优异，适合大模型训练；ASIC 在特定任务上能效比极高，适合推理加速；FPGA 则具有灵活可编程的特性，能够适应不断变化的算法需求。通过异构计算架构，可以将不同类型的计算任务分配给最适合的处理单元，从而实现整体性能的最优化。

二、异构调度能够显著提升资源利用率，降低总体成本。在实际应用中，大模型的工作负载往往呈现多样化特征，既有计算密集型的训练任务，也有延迟敏感型的推理任务，还有 IO 密集型的数据处理任务。异构算力通过智能调度，将不同类型的任务分配给最适合的计算资源，避免资源闲置和浪费，提高整体资源利用率。研究表明，合理的异构调度可以将资源利用率提升 30%以上，显著降低算力成本。

三、异构算力提供了更好的扩展性和灵活性。随着大模型技术的快速发展，新的算法和模型结构不断涌现，对算力的需求也在不断变化。异构算力架构通过多种计算单元的组合，能够更好地适应这种变化，为新算法和新模型提供支持。

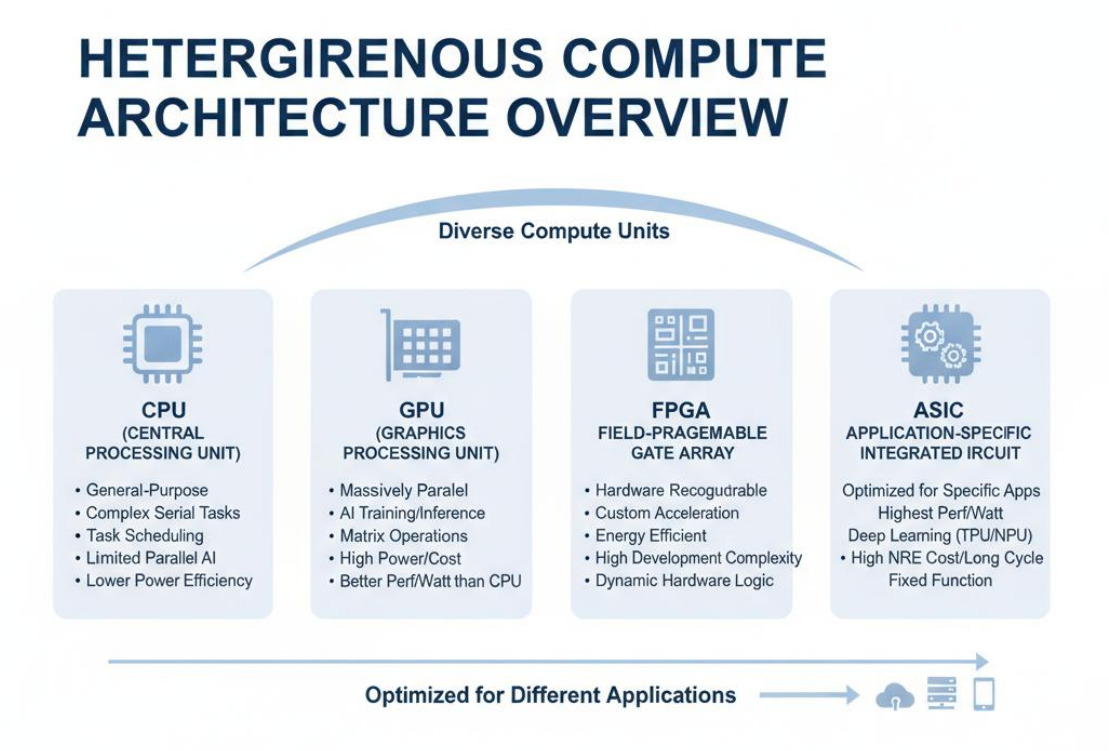
同时，异构算力也支持渐进式的升级和扩展，企业可以根据需求逐步增加或更新计算资源，降低技术升级的成本和风险。

四、异构算力有助于降低能耗，实现绿色计算。不同类型的计算单元在能效比方面各有优势，通过异构调度，可以将任务分配给能效比最高的计算单元，从而降低整体能耗。特别是在推理场景，ASIC 和 FPGA 等专用计算单元的能效比往往远高于通用计算单元，能够显著降低推理过程的能耗。在全球"双碳"目标下，异构算力的这一优势具有重要意义。

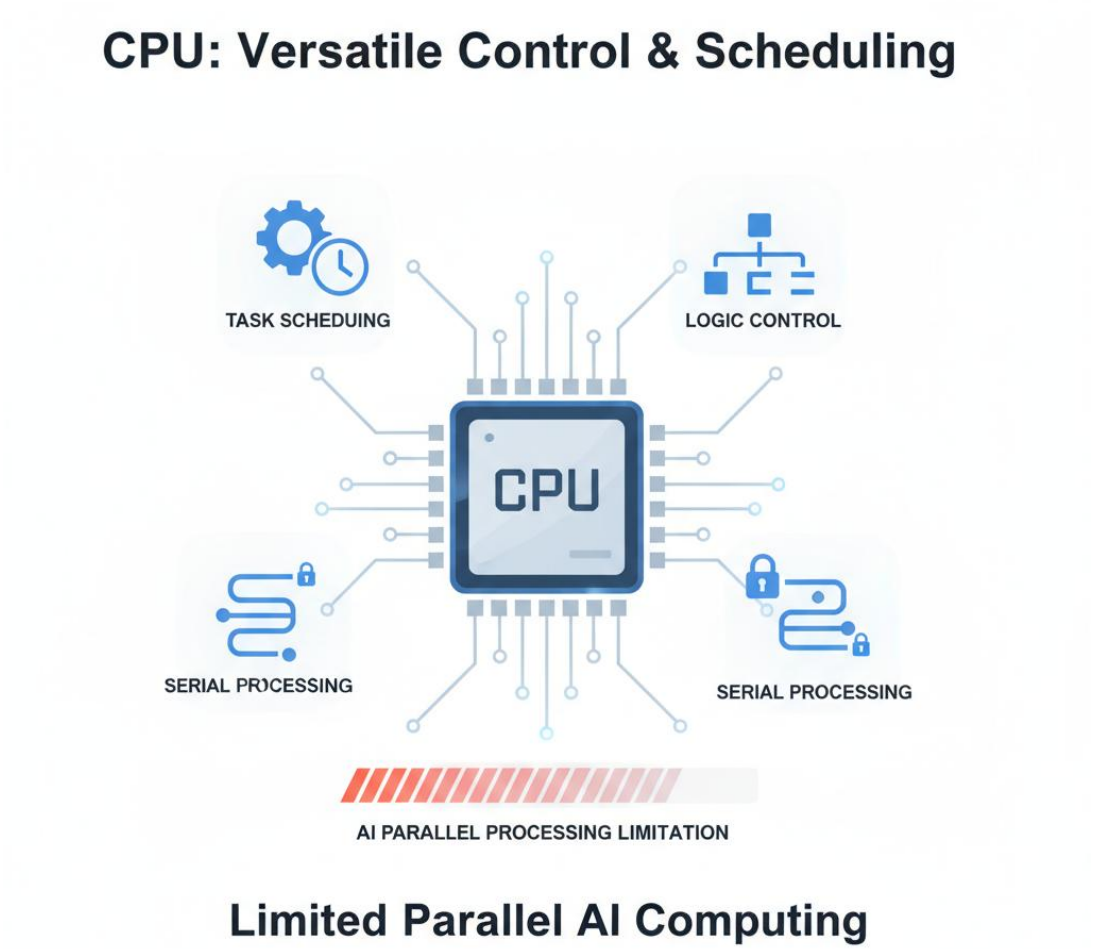
三、异构算力技术架构与核心组件

3.1 异构计算硬件体系

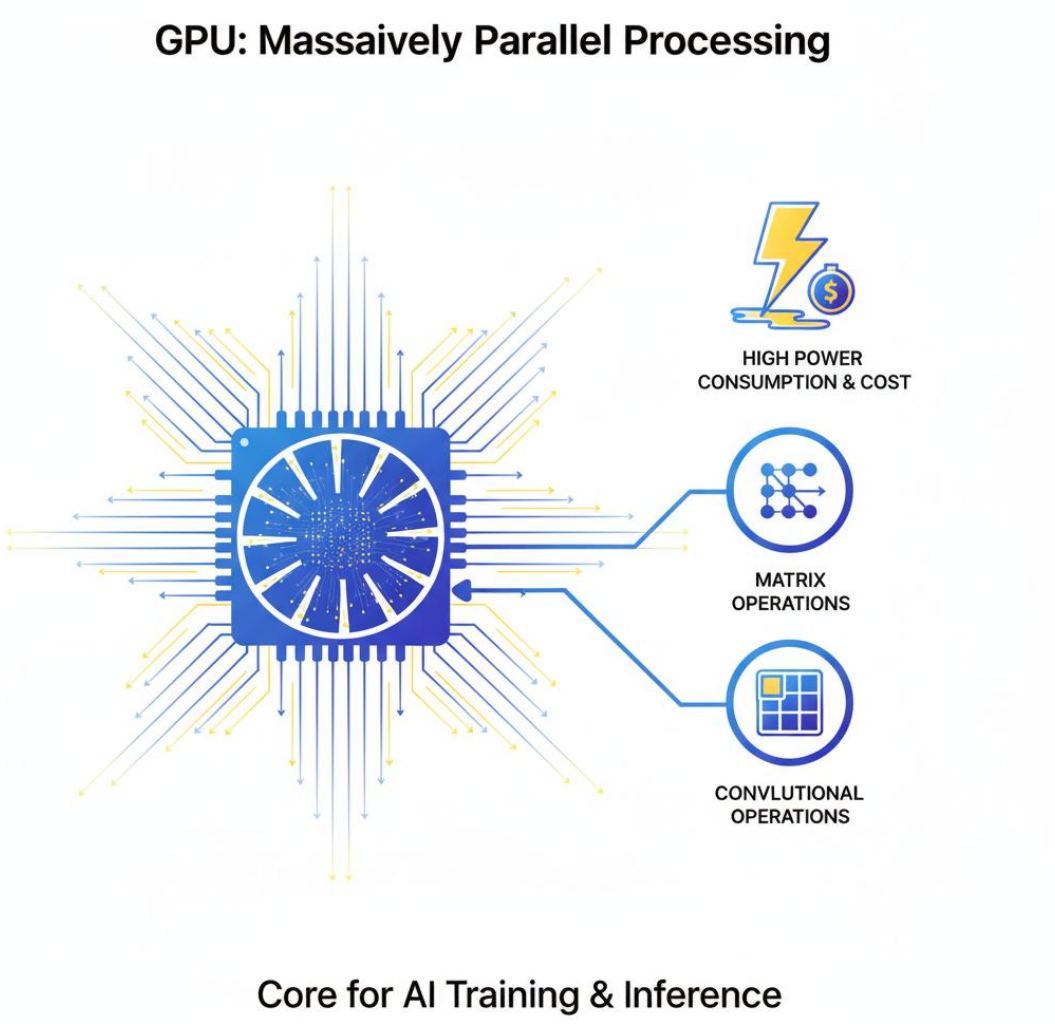
3.1.1 主流 AI 芯片对比



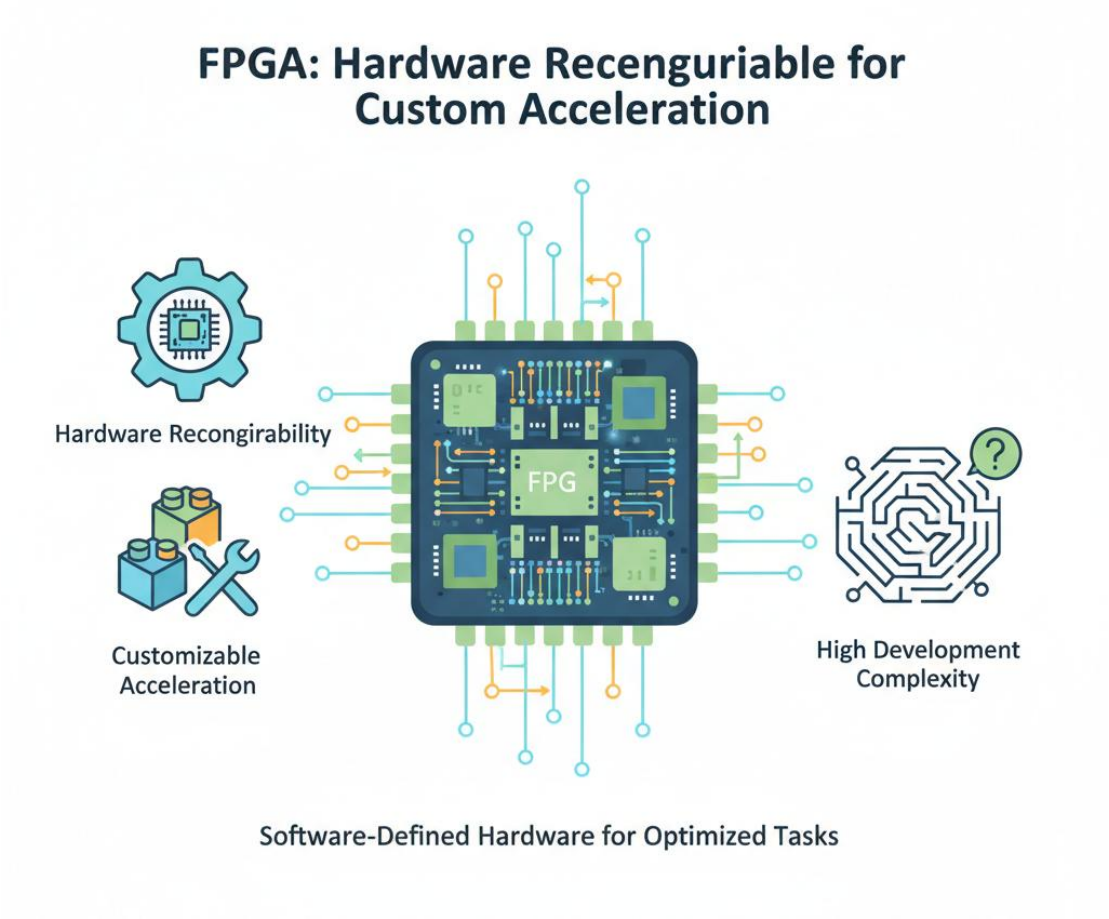
异构计算硬件体系由多种类型的计算单元组成，主要包括 CPU、GPU、FPGA、ASIC 等，各具特点，适用于不同的应用场景。



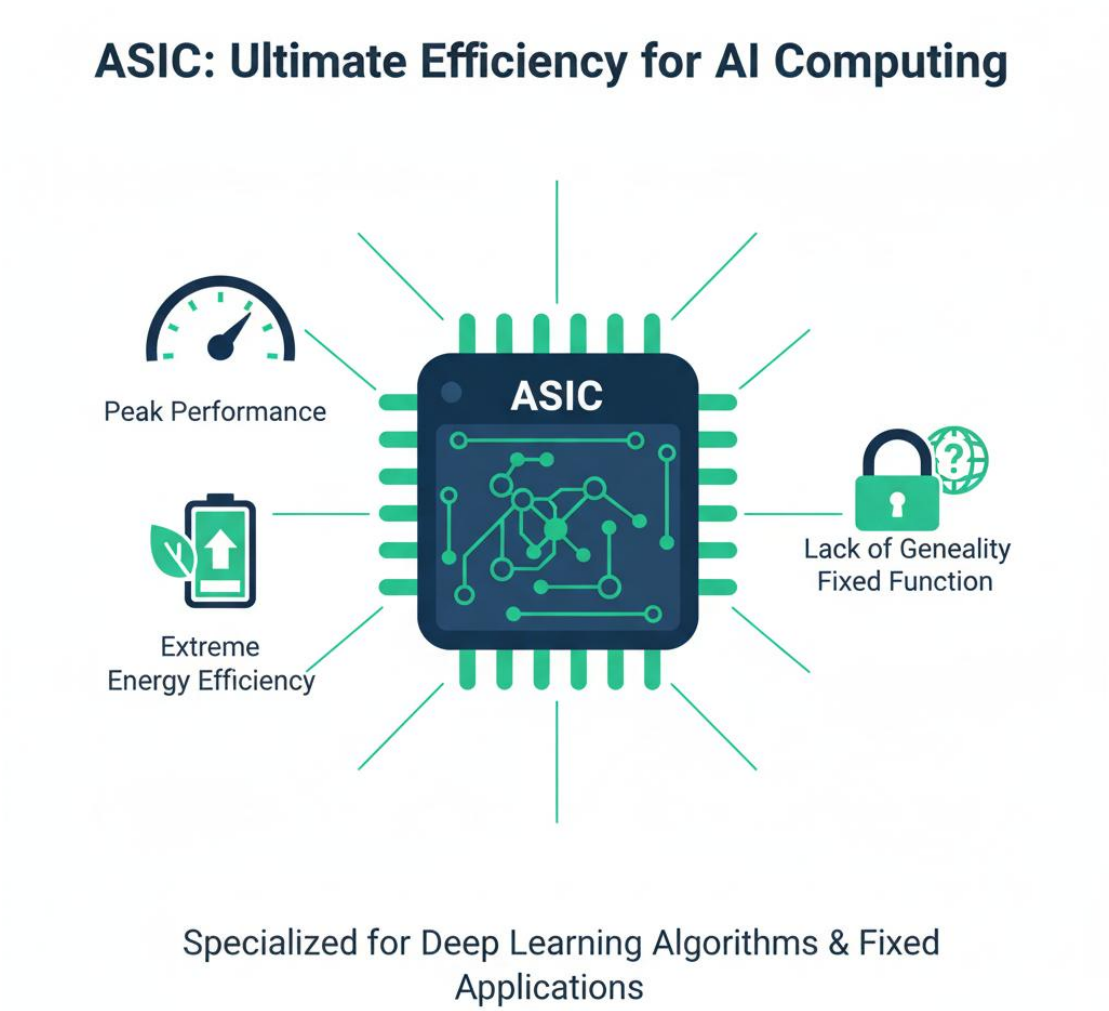
CPU（中央处理器）作为通用计算单元，具有强大的逻辑控制和任务调度能力，适合处理复杂的串行任务和多样化的工作负载。然而，在 AI 计算场景下，CPU 的并行计算能力相对有限，能效比较低。现代 CPU 通常集成多个核心，支持 SIMD（单指令多数据）指令集，如 AVX-512 等，在一定程度上提升了 AI 计算性能，但与专用 AI 加速器相比仍有差距。



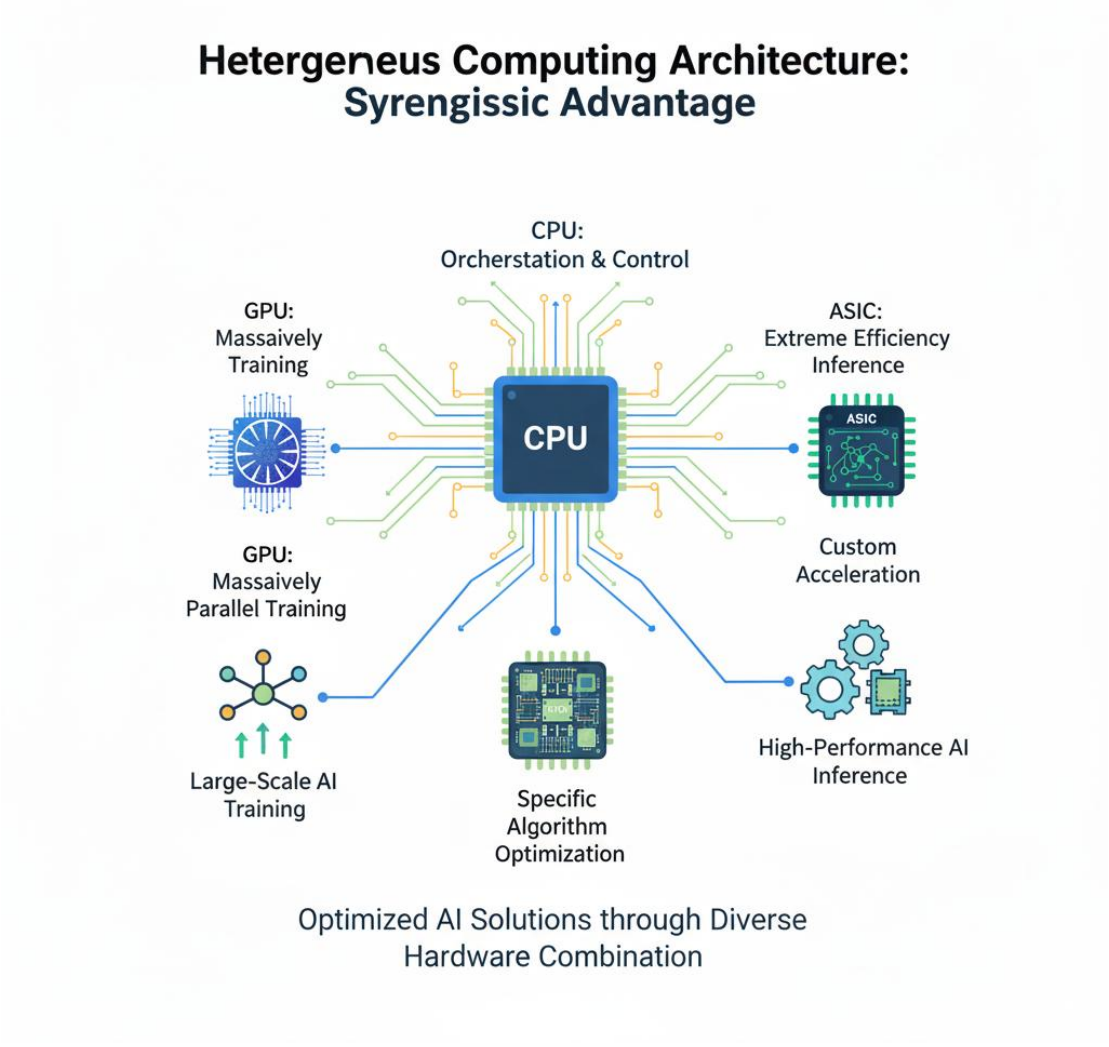
GPU（图形处理器）最初为图形渲染设计，因其强大的并行计算能力而成为AI 训练和推理的主流选择。GPU 拥有数千个计算核心，适合执行大规模并行计算任务，特别是在矩阵运算、卷积运算等 AI 核心算法上表现优异。然而，GPU 功耗较高，成本昂贵，且在某些特定算法上效率不如专用芯片。在能效比方面，GPU 优于 CPU 但不及 FPGA 和 ASIC。



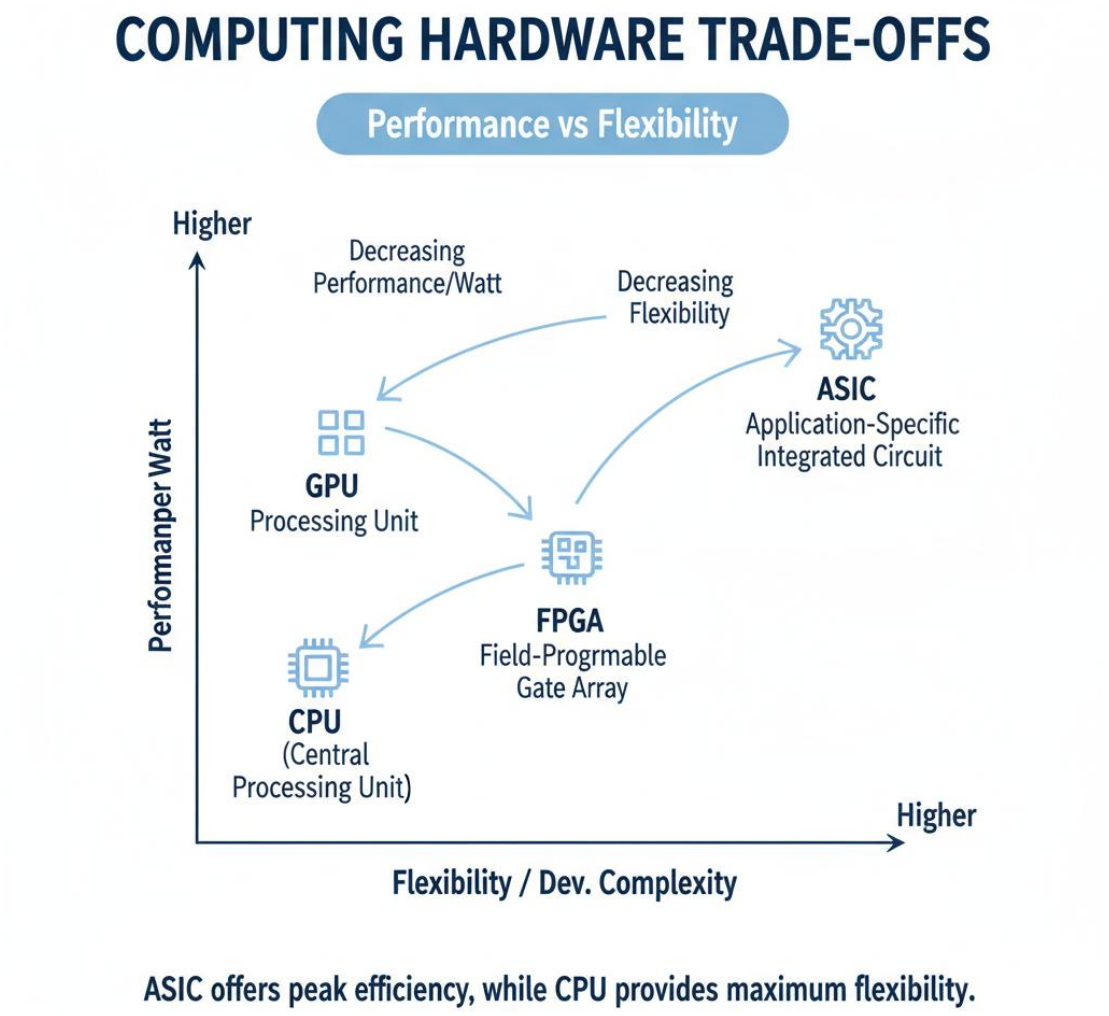
FPGA（现场可编程门阵列）具有硬件可重构的特点，用户可以根据特定应用需求定制硬件逻辑，实现高度优化的计算加速。FPGA 在能效比和灵活性方面具有优势，特别适合需要定制化加速的场景。然而，FPGA 开发复杂度高，需要专业的硬件设计知识，且运行频率相对较低，在大规模部署时面临挑战。与 GPU/CPU 相比，FPGA 采用软件定义的硬件架构，硬件逻辑可根据需求动态调整，而 GPU/CPU 硬件固定，其并行性设计是适应固定硬件的。



ASIC（专用集成电路）针对特定应用进行优化，在能效比和性能方面表现最佳。AI 领域的 ASIC 如 TPU、NPU 等，针对深度学习算法特点进行专门优化，实现了极高的计算密度和能效比。然而，ASIC 缺乏通用性，开发成本高，周期长，适合大规模、固定场景的应用。从能耗比方面来看，ASIC > FPGA > GPU > CPU，产生这样结果的根本原因是：对于计算密集型算法，数据的搬移和运算效率越高的能耗比就越高。



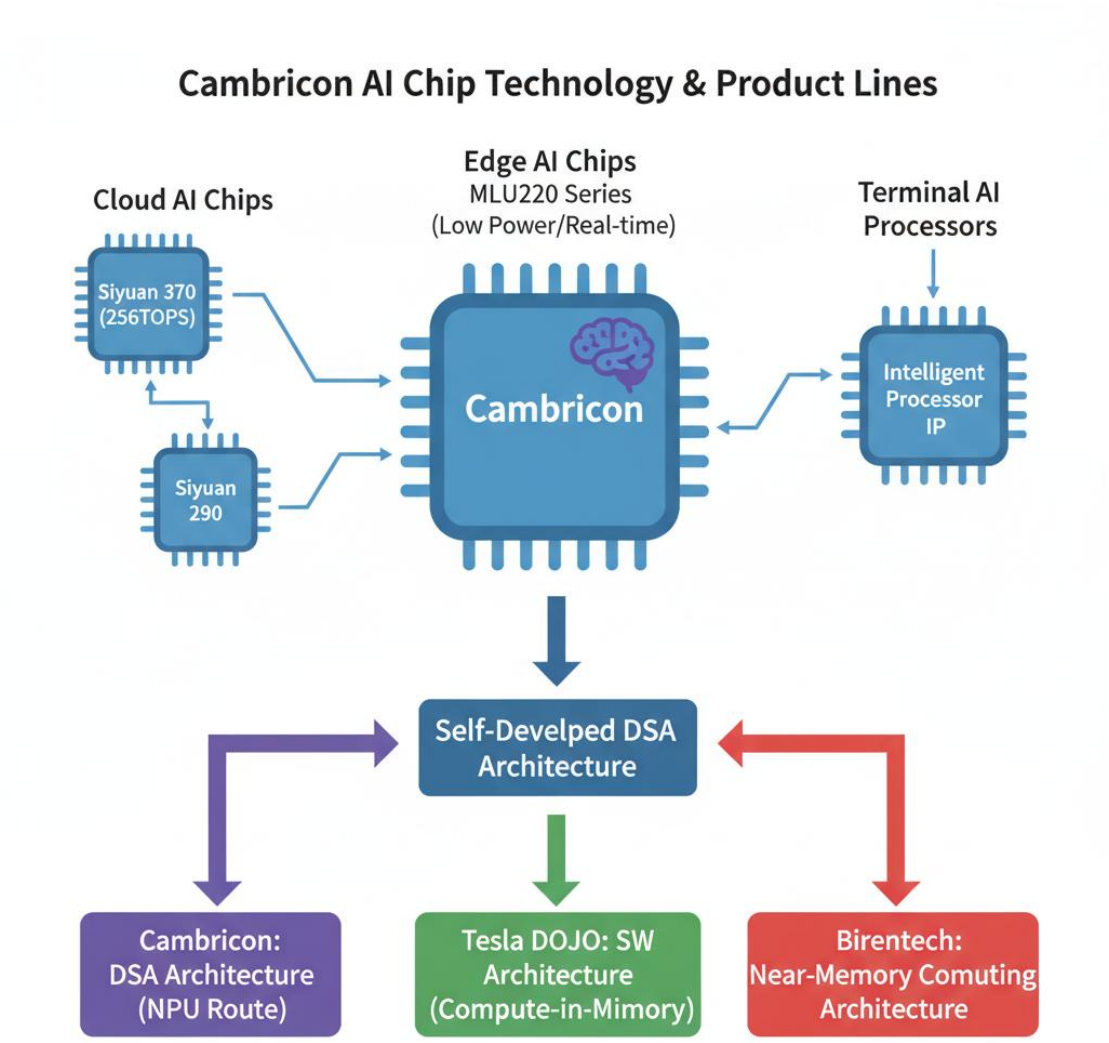
在大模型场景下，不同芯片各有所长：GPU 适合大规模并行训练，ASIC 适合高能效推理，FPGA 适合特定算法加速，CPU 适合任务调度和控制。异构计算架构通过合理组合这些不同类型的计算单元，可以充分发挥各自优势，实现整体系统性能的最优化。



从性能功耗比来看，ASIC 作为定制芯片表现最优，GPU 次之，FPGA 再次之，CPU 最低。但从灵活性和开发难度来看，则正好相反。在实际的异构计算系统中，通常采用 CPU+GPU 的组合用于通用 AI 训练，CPU+FPGA 的组合用于需要定制化加速的场景，CPU+ASIC 的组合则用于大规模推理部署。这种多样化的硬件组合，为不同场景下的 AI 计算提供了最优解决方案。

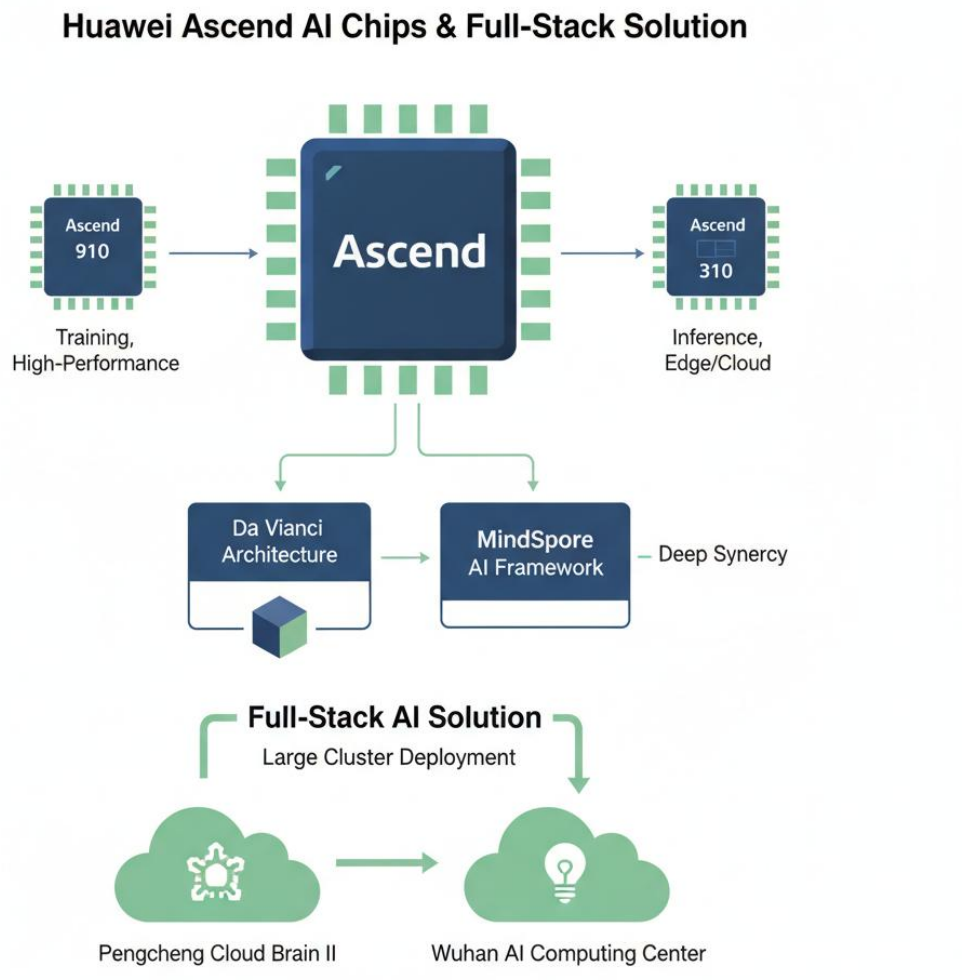
3.1.2 国产 AI 芯片技术路线

国产 AI 芯片近年来取得了显著进展，形成了多元化的技术路线和产品体系。主要厂商包括寒武纪、华为昇腾、海光、壁仞、燧原、沐曦、摩尔线程等，各自推出了具有特色的 AI 芯片产品。



寒武纪 AI 芯片技术路线

寒武纪作为国内 AI 芯片的领军企业，专注于人工智能芯片产品的研发与技术创新，提供云边端全场景 AI 芯片产品。云端产品线包括思元 290、思元 370 等，其中思元 370 达到 256TOPS INT8 算力；边缘端产品线包括 MLU220 系列，提供低功耗、高实时性的 AI 加速能力；终端产品线包括智能处理器 IP，授权给终端设备厂商使用。寒武纪采用自研 DSA 计算架构，与特斯拉 DOJO 的存算一体架构和壁仞科技的近存架构形成不同的技术路线。



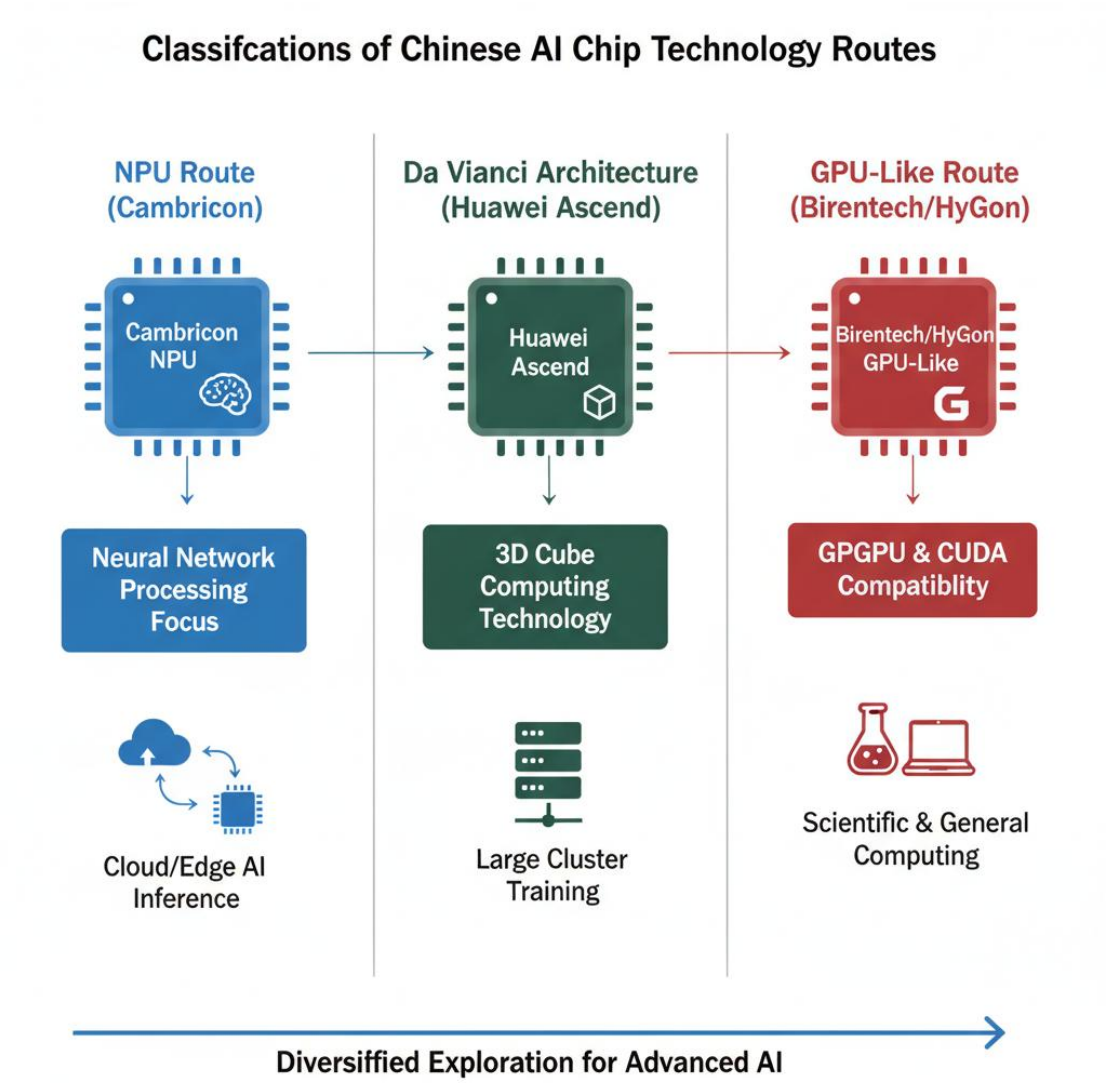
华为昇腾 AI 芯片与全栈解决方案

华为昇腾系列芯片包括昇腾 910 和昇腾 310 等，其中昇腾 910 是面向训练的高性能 AI 芯片，昇腾 310 主要面向推理场景。昇腾芯片采用达芬奇架构，支持 3D Cube 计算引擎，在 AI 计算性能方面具有竞争力。华为还推出了 MindSpore AI 框架，与昇腾芯片深度协同，形成了全栈 AI 解决方案。昇腾芯片在鹏城云脑 II、武汉人工智能计算中心等大集群实践中得到广泛应用。

海光 DCU 系列是基于 GPGPU 架构的 AI 加速器，兼容 CUDA 生态，降低了用户迁移成本。海光 DCU 产品深算一号在通用计算和 AI 计算方面表现均衡，特别适合科学计算与 AI 融合的应用场景。壁仞 BR100 系列采用近存计算架构，在计算密度和能效比方面具有创新，是国内高端 AI 芯片的代表之一。

燧原科技、沐曦集成电路、摩尔线程等新兴 AI 芯片企业也各具特色。燧原科技推出邃思系列 AI 芯片，采用自研的 GCU 架构；沐曦集成电路专注于高性能 GPU 研发；摩尔线程则面向图形计算和 AI 计算融合场景。这些企业的创新推

动着国产 AI 芯片技术的多元化发展。



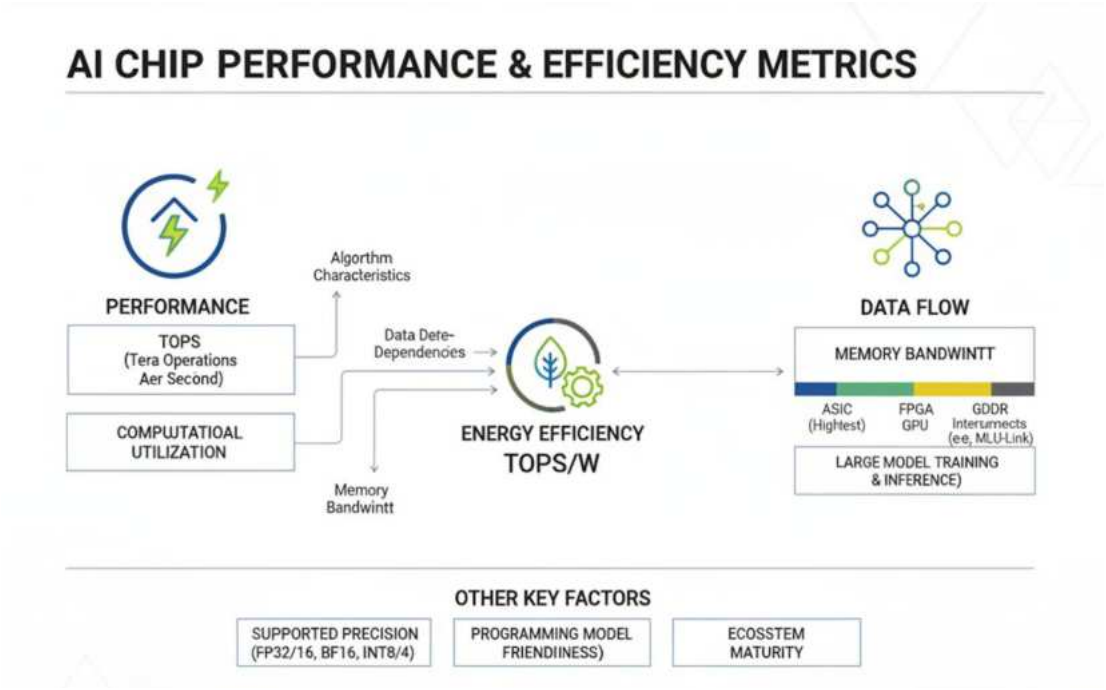
国产 AI 芯片技术路线分类

国产 AI 芯片在指令集、制程工艺、算力指标、生态兼容性等方面各有特点。在指令集方面，多数厂商采用自研指令集，以实现更好的性能优化；在制程工艺方面，普遍采用 7nm、5nm 等先进工艺；在算力指标方面，高端产品已接近国际领先水平；在生态兼容性方面，通过支持主流 AI 框架、提供迁移工具等方式，降低开发者使用门槛。

从技术路线来看，国产 AI 芯片主要分为三类：一是以寒武纪为代表的 NPU 路线，专注于神经网络处理；二是以华为昇腾为代表的达芬奇架构路线，强调 3D Cube 计算技术；三是以壁仞为代表的类 GPU 路线，兼容 CUDA 生态。这些不同的技术路线反映了国产 AI 芯片在追赶国际先进水平过程中的多元化探索。

3.1.3 芯片性能与能效评测

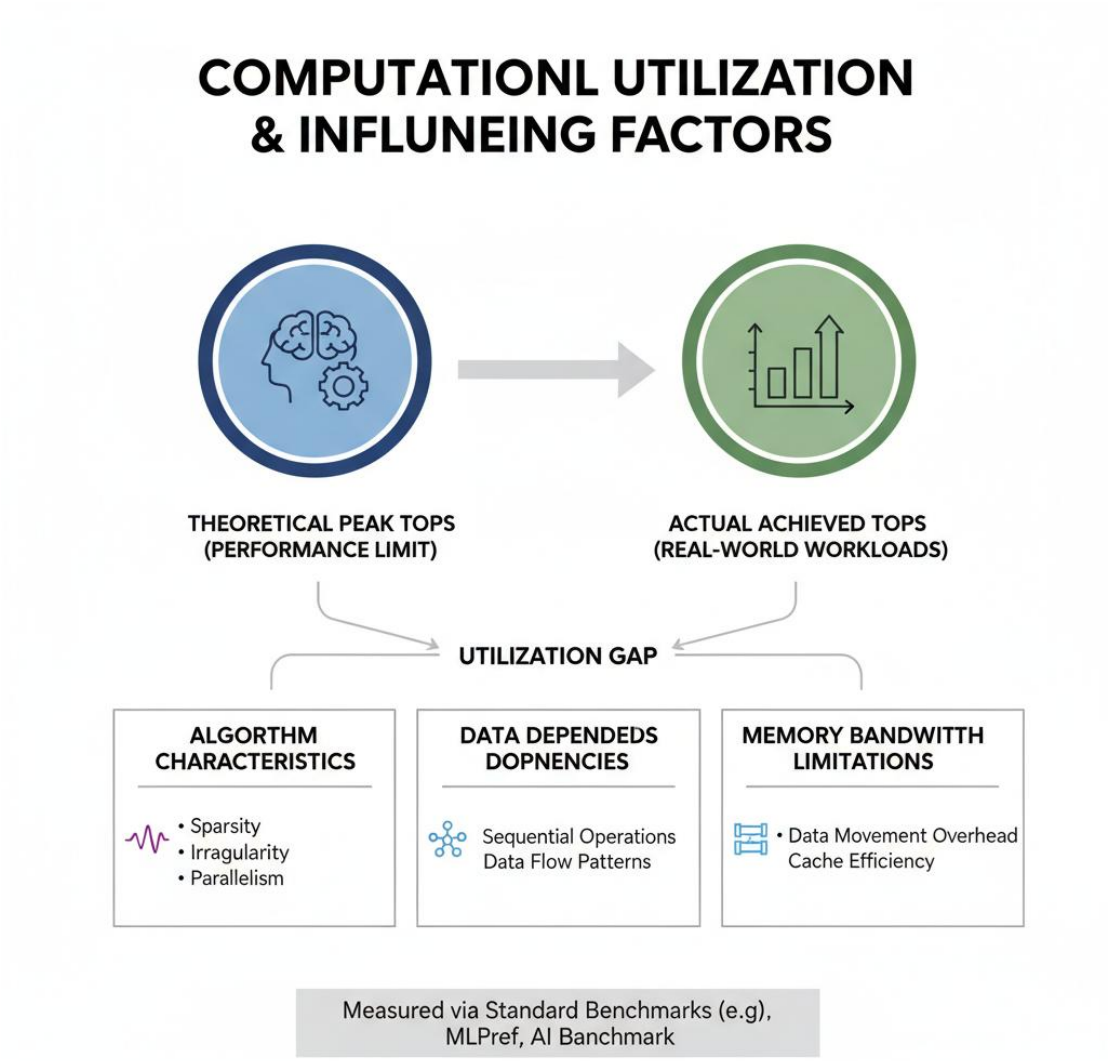
AI 芯片的性能和能效评测涉及多个关键指标，包括 TOPS/W（每瓦特万亿次运算）、算力利用率、内存带宽等，这些指标综合反映了芯片在实际应用中的表现。



TOPS（Tera Operations Per Second）是衡量 AI 芯片算力的核心指标，表示芯片每秒可执行的万亿次操作数。然而，理论 TOPS 值并不能完全反映实际性能，还需要考虑算力利用率，即实际达到的算力与理论峰值的比例。影响算力利用率的因素包括算法特性、数据依赖性、内存带宽限制等。在实际评测中，需要通过标准基准测试套件，如 MLPerf、AI Benchmark 等，来衡量芯片在典型 AI 任务上的实际性能。

能效比（TOPS/W）是衡量 AI 芯片能效的关键指标，表示每瓦特功耗可提供的算力。随着数据中心能耗问题的日益突出，能效比成为芯片设计的重要目标。不同类型芯片的能效比差异显著：ASIC 通常能达到最高的能效比，FPGA 次之，GPU 再次之，CPU 最低。在实际应用中，需要综合考虑性能和能效，选择最适合的芯片类型。

内存带宽是影响 AI 芯片性能的另一关键因素。大模型训练和推理涉及大量数据移动，内存带宽往往成为性能瓶颈。现代 AI 芯片普遍采用高带宽内存（HBM、GDDR 等）来提升内存带宽，如寒武纪 MLU370-X8 搭载 MLU-Link 多芯互联技术，每张加速卡可获得 200GB/s 的通讯吞吐性能。在实际评测中，需要关注理论内存带宽和有效内存带宽的差异，以及内存子系统对整体性能的影响。



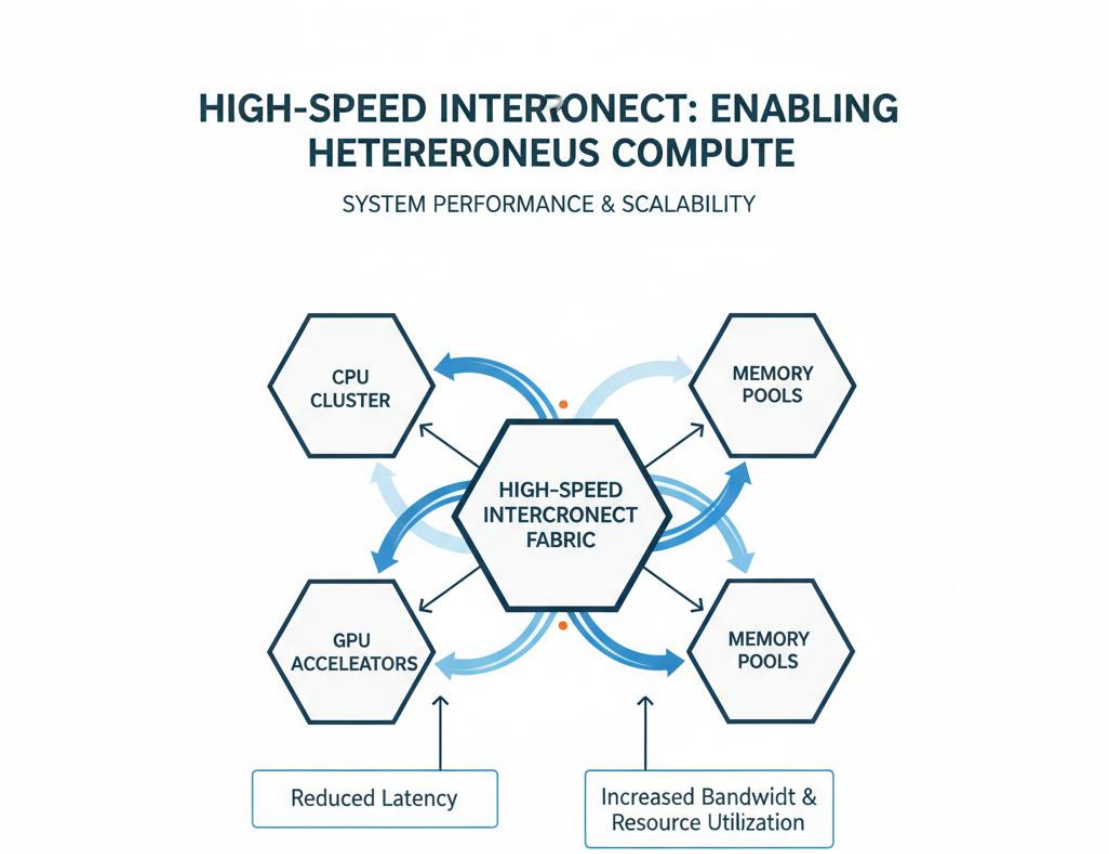
算力利用率及其影响因素

除了上述指标外，AI 芯片评测还需考虑支持精度（FP32/FP16/BF16/INT8/INT4 等）、编程模型友好度、生态成熟度等因素。支持精度决定了芯片在不同精度计算任务上的适用性；编程模型友好度影响开发效率；生态成熟度则关系到芯片的实际应用前景。

在国产芯片与国际标杆的对比中，寒武纪 MLU370、昇腾 910B 等国产芯片在算力指标上已接近 NVIDIA A100/H100 的水平，但在软件生态、编程模型等方面仍有差距。随着技术的不断进步和生态的持续完善，国产 AI 芯片的性能和能效将进一步提升，为大模型训练和推理提供强有力的硬件支撑。

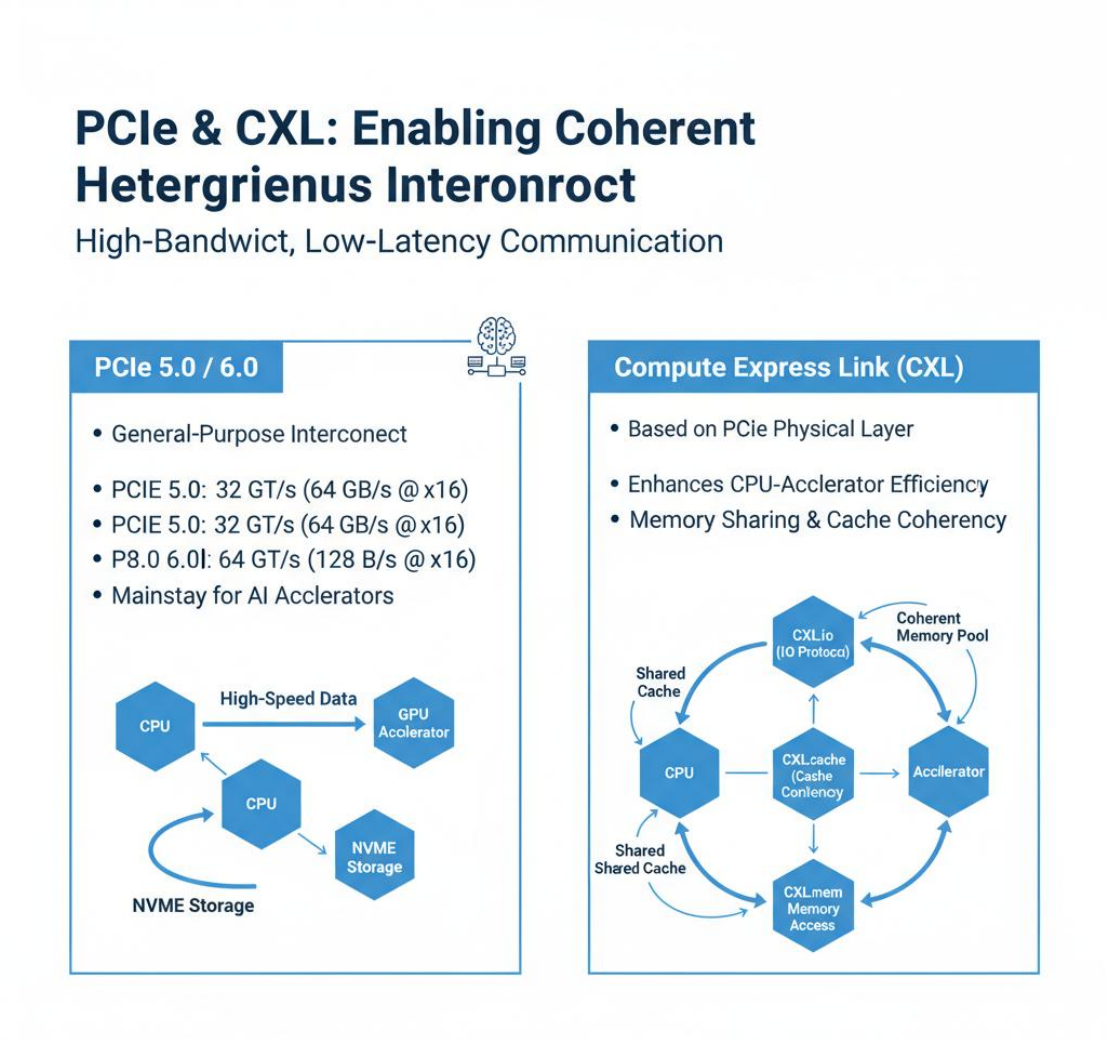
3.2 高速互联与网络架构

3.2.1 高速互联技术



高速互联技术在异构算力系统中的关键作用

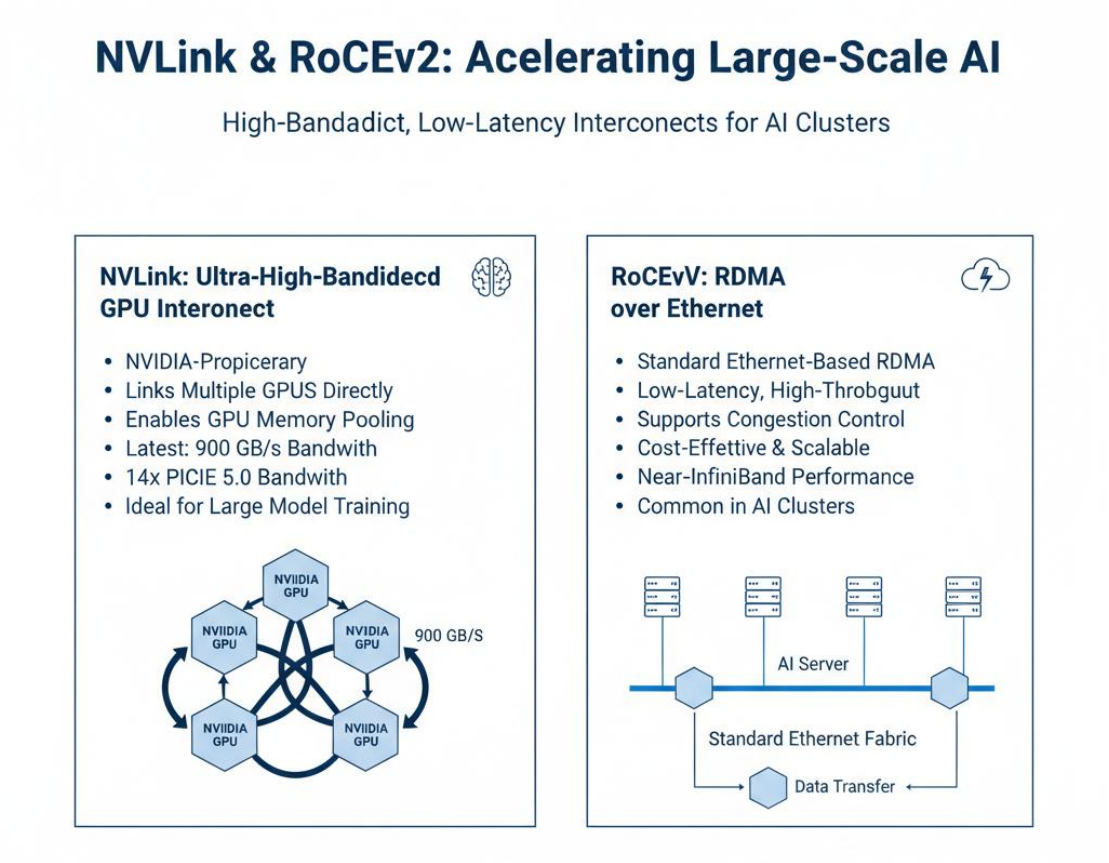
高速互联技术是异构算力系统的关键组成部分，直接影响系统的整体性能和扩展能力。在大模型训练和推理场景中，高效的高速互联技术能够显著提升系统性能，降低通信延迟，提高资源利用率。



PCIe 和 CXL 技术概览

PCIe（Peripheral Component Interconnect Express）是计算机系统中广泛使用的高速互联标准，目前主流的是 PCIe 5.0，正在向 PCIe 6.0 发展。PCIe 5.0 提供 32GT/s 的传输速率，x16 配置下可提供约 64GB/s 的带宽，满足大多数 AI 加速卡的互联需求。PCIe 6.0 进一步将传输速率提升至 64GT/s，并引入 PAM4 调制技术，在相同物理层下实现带宽翻倍。PCIe 5.0/6.0 已成为 AI 加速器与主机系统互联的主流选择，为 AI 计算提供高带宽、低延迟的数据传输通道。

CXL（Compute Express Link）是基于 PCIe 物理层的新型互联协议，旨在提高 CPU 与专用加速器之间的互联效率。CXL 协议保留并拓展了 PCIe 的兼容性，只要使用 PCIe 5.0 及以上版本且支持 CXL 的设备均可通过 CXL 实现高速互联。CXL 支持三种协议：CXL.io（基础 I/O 协议）、CXL.cache（缓存一致性协议）和 CXL.mem（内存访问协议），能够实现 CPU 与加速器之间的高效内存共享和缓存一致性，特别适合异构计算场景。



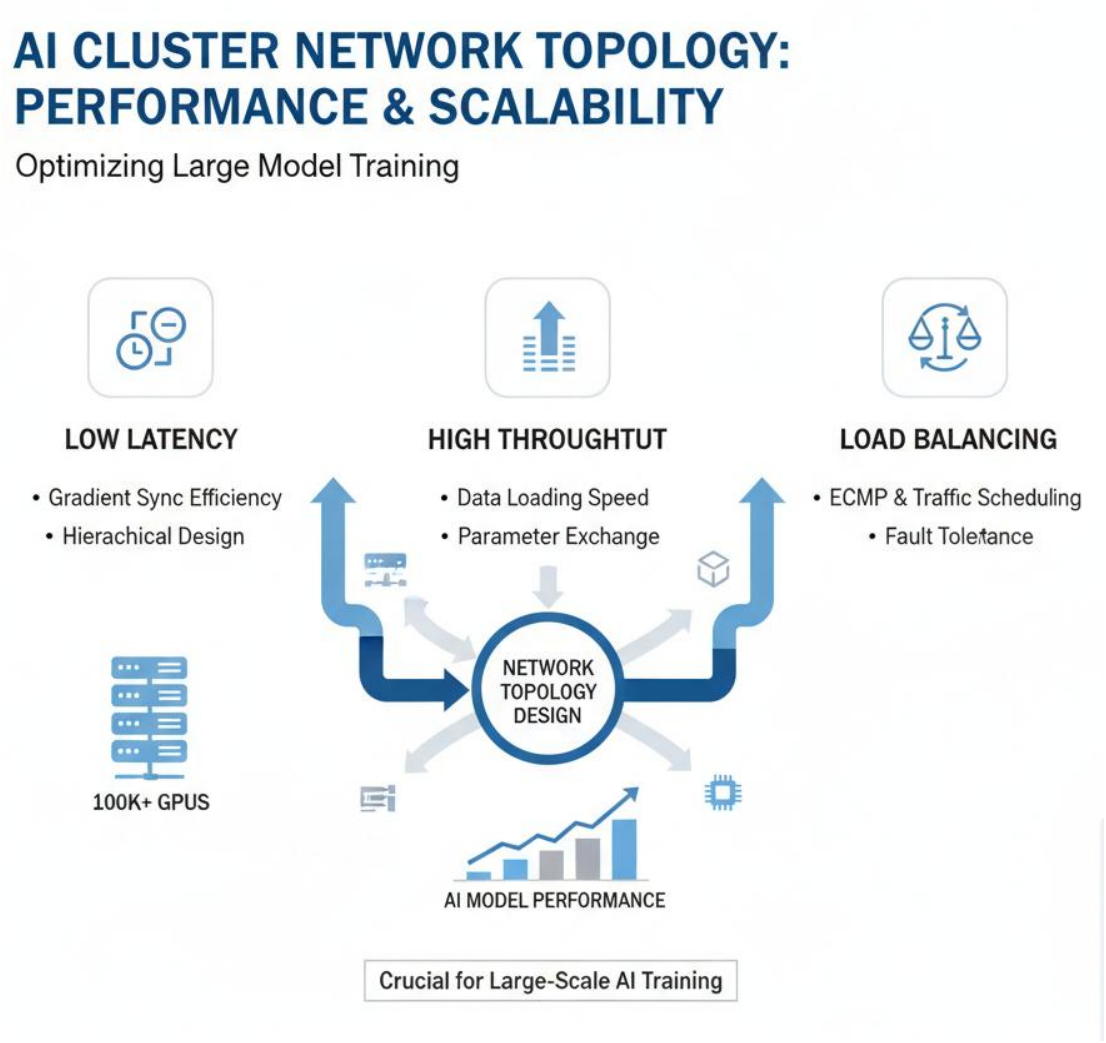
NVLink 和 RoCEv2 技术概览

NVLink 是 NVIDIA 专有的高速 GPU 互联技术，与传统的 PCIe 相比，能为更多 GPU 系统提供更快速的替代方案。NVLink 技术通过连接多个 NVIDIA 显卡，能够实现显存池化和高速数据交换，大幅提升多 GPU 系统的性能。最新的 NVLink 技术提供高达 900GB/s 的带宽，是 PCIe 5.0 的 14 倍以上，特别适合大模型训练等需要大量 GPU 间通信的场景。

RoCEv2（RDMA over Converged Ethernet version 2）是基于以太网的 RDMA（远程直接内存访问）技术，在标准以太网上实现低延迟、高吞吐的数据传输。RoCEv2 支持拥塞控制和流量控制，能够在不增加专用网络设备的情况下提供接近 InfiniBand 的性能。在大规模 AI 集群中，RoCEv2 因其成本优势和标准化特性，成为广泛选择的高速互联技术。

这些高速互联技术在带宽、延迟、扩展性等方面各有特点。PCIe 提供通用互联，CXL 增强内存一致性，NVLink 提供超高带宽 GPU 互联，RoCEv2 实现标准以太网上的 RDMA。在实际系统设计中，需要根据应用场景和性能需求，选择合适的高速互联技术，构建高效的异构算力系统。

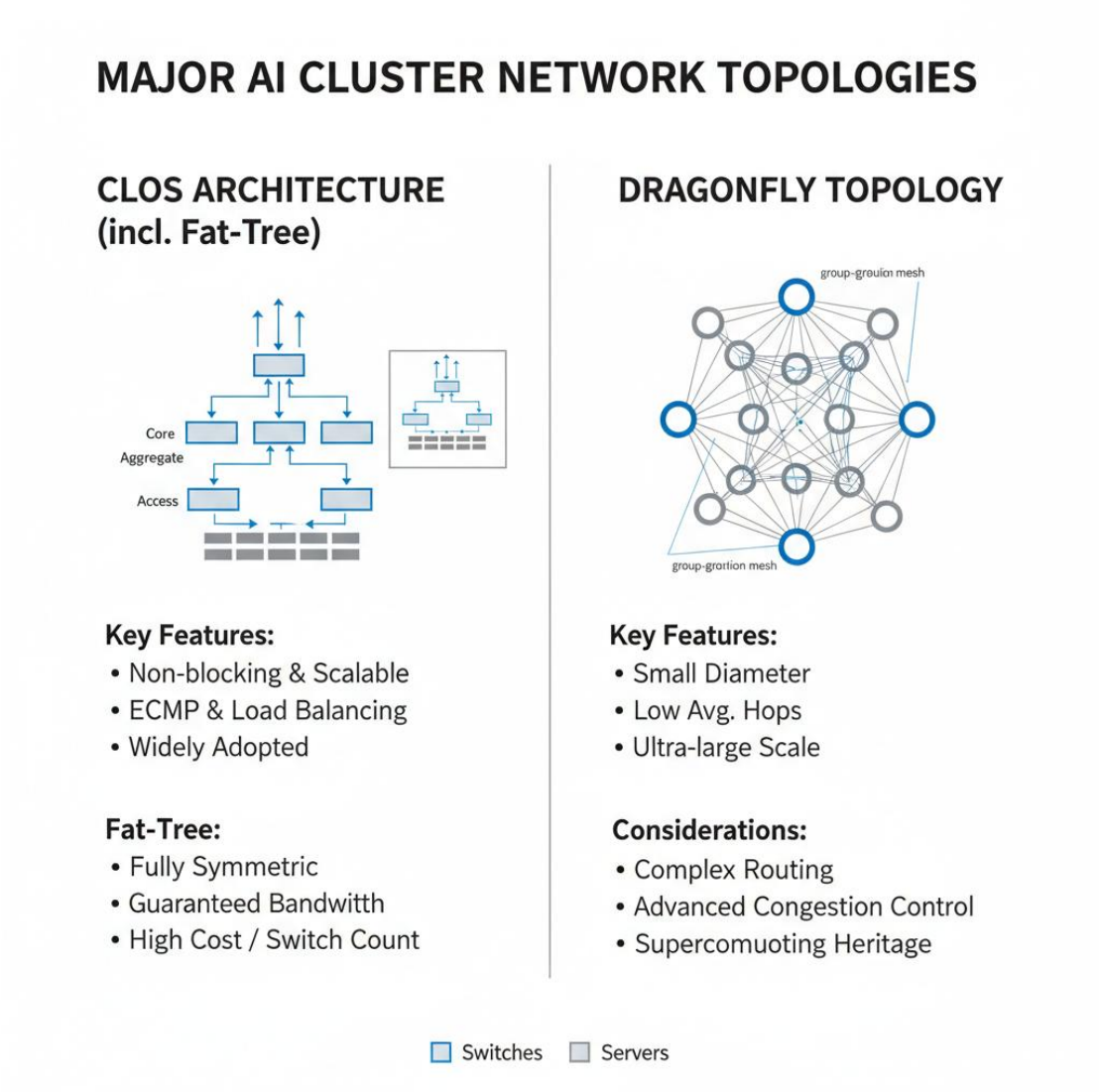
3.2.2 智算中心网络拓扑



大规模 AI 集群网络拓扑概述

智算中心网络拓扑设计直接影响大规模 AI 集群的性能和扩展能力。在大模型训练场景中，特别是万卡甚至十万卡集群，合理的网络拓扑设计对于降低通信延迟、提高网络吞吐、实现负载均衡至关重要。

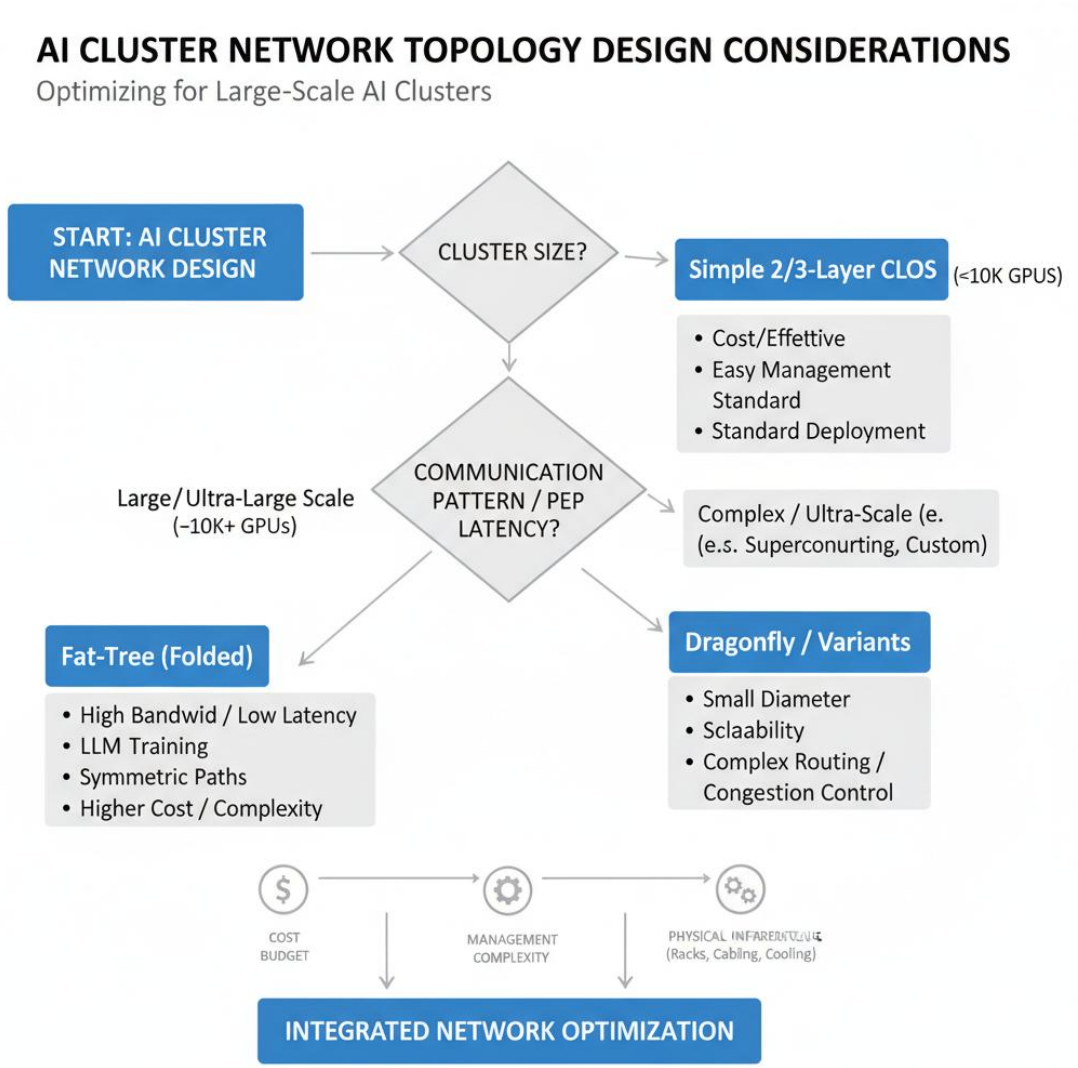
CLOS 三层架构是目前大规模数据中心网络的主流拓扑结构，包括核心层、汇聚层和接入层。CLOS 架构具有无阻塞、高可扩展性的特点，能够有效支持大规模服务器集群的互联。在 AI 集群中，CLOS 架构通常配合 ECMP（等价多路径）路由，实现负载均衡和故障容错。CLOS 架构的扩展性好，可以通过增加交换机数量和端口密度来线性扩展网络容量，适合大规模 AI 集群的部署。



主流网络拓扑结构对比

Fat-Tree 是 CLOS 架构的一种特例，采用完全对称的设计，所有路径具有相同的带宽和延迟。Fat-Tree 拓扑在 AI 集群中得到广泛应用，特别是在需要高带宽、低延迟通信的大模型训练场景中。Fat-Tree 网络的优点是带宽保证、无阻塞、易于管理，但缺点是成本较高，交换机数量多。在实际部署中，通常采用折叠式 Fat-Tree（Folded Fat-Tree）设计，减少交换机数量，降低成本。

Dragonfly 是一种高维网络拓扑，通过高维连接实现节点间的高效通信。Dragonfly 拓扑在超级计算机中得到广泛应用，近年来也开始应用于大规模 AI 集群。Dragonfly 网络的优点是直径小、平均跳数少、扩展性好，适合超大规模集群的部署。然而，Dragonfly 拓扑的路由和拥塞控制较为复杂，需要专门的算法支持。



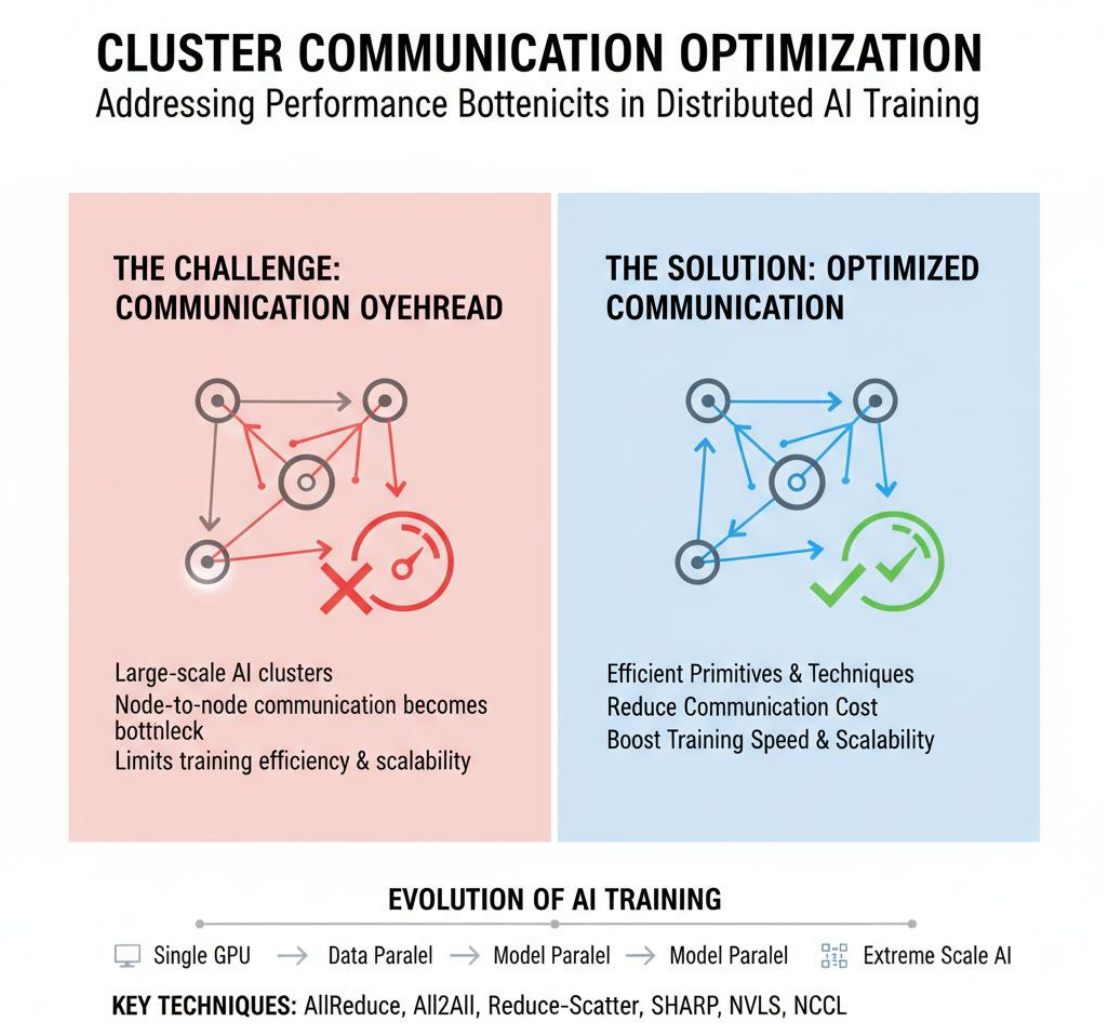
万卡集群网络设计中，如何综合考虑 P2P 延迟与吞吐优化，以及在实际部署中需要考虑的多种因素

在万卡集群网络设计中，需要综合考虑 P2P 延迟与吞吐优化。P2P 延迟直接影响大模型训练中的梯度同步效率，而吞吐则影响数据加载和模型参数交换的速度。为了优化 P2P 延迟，通常采用层次化网络设计，将物理上临近的节点组织在同一子网中，减少跨子网通信；为了优化吞吐，通常采用多路径负载均衡、流量调度等技术，充分利用网络带宽。

在实际部署中，智算中心网络拓扑设计需要考虑多个因素：集群规模、通信模式、成本预算、管理复杂度等。对于中小规模集群，通常采用简单的二层或三层 CLOS 架构；对于大规模集群，可能需要更复杂的拓扑结构，如 Dragonfly 或其变种。此外，网络拓扑设计还需要与机柜布局、线缆管理、散热设计等物理设施相协调，实现整体系统的最优化。

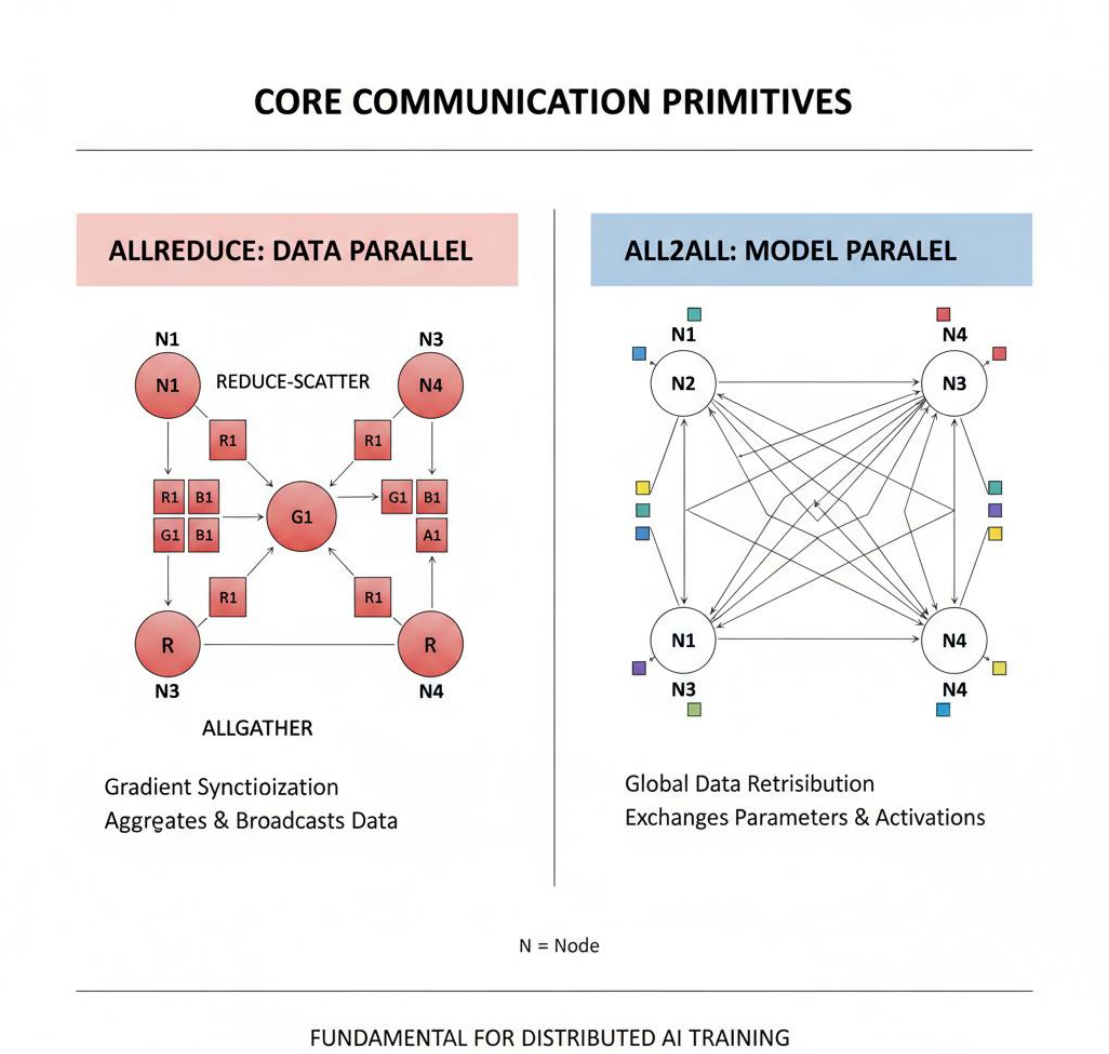
3.2.3 集群通信优化

集群通信优化是大模型分布式训练的关键技术，直接影响训练效率和扩展性。在大规模 AI 集群中，节点间的通信开销往往成为性能瓶颈，因此需要通过高效的通信原语和优化技术来降低通信开销，提高训练效率。



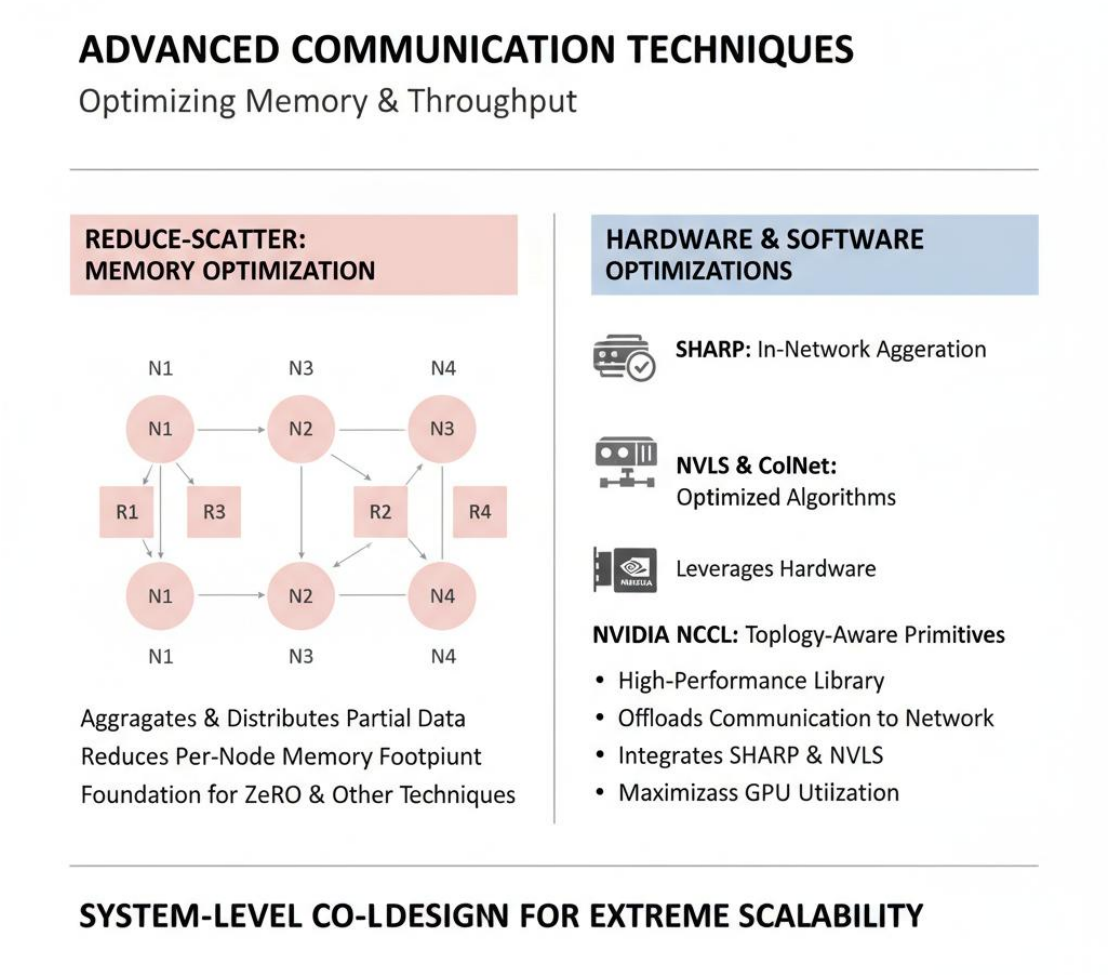
集群通信优化概览 - 分布式训练的性能瓶颈与解决方案

AllReduce 是最常用的通信原语之一，用于数据并行训练中的梯度同步。AllReduce 操作将所有节点的数据聚合后广播给所有节点，实现全局梯度的一致性。AllReduce 可以通过先进行 ReduceScatter 操作，然后进行 AllGather 操作来实现：ReduceScatter 操作首先聚合数据，然后将结果分散，这样每个成员仅持有聚合结果的一部分；AllGather 操作则将各部分结果收集到所有节点，形成完整的结果。AllReduce 通过组合操作，成为数据并行训练的核心通信原语。



核心通信原语 - AllReduce 与 All2All

All2All（All-to-All）是另一种重要的通信原语，在模型并行训练中广泛应用。All2All 操作实现全局数据的重新分布，每个节点向所有其他节点发送数据，同时从所有其他节点接收数据。在张量并行和流水线并行中，All2All 通信用于参数和激活值的交换，是实现模型并行的基础。



Reduce-Scatter 与通信优化技术

Reduce-Scatter 是平衡显存与通信的重要原语，为 ZeRO 等显存优化技术奠定基础。Reduce-Scatter 操作首先聚合数据，然后将结果分散到各个节点，每个节点仅持有聚合结果的一部分。这种操作可以有效减少单节点的内存占用，同时控制通信开销。

为了进一步提高通信效率，业界提出了多种优化技术。SHARP（Scalable Hierarchical Aggregation and Reduction Protocol）技术通过在网络设备中执行聚合操作，减少数据在节点间的传输量，显著提高分布式深度学习工作负载的可扩展性和性能。NVLS 与 CollNet 是专为优化 AllReduce 性能设计的特殊算法，其中 NVLS 还通过利用特定硬件能力支持 ReduceScatter 和 AllGather 操作。

在实际应用中，NVIDIA 集合通信库(NCCL)提供了高性能、拓扑感知型集合运算：AllReduce、Broadcast、Reduce、AllGather 和 ReduceScatter，这些运算已针对 NVIDIA 硬件进行了深度优化。NCCL 经过优化，可将关键的集合通信操作分流到网络，从而充分利用 SHARP，显著提高分布式深度学习工作负载的可

扩展性和性能。

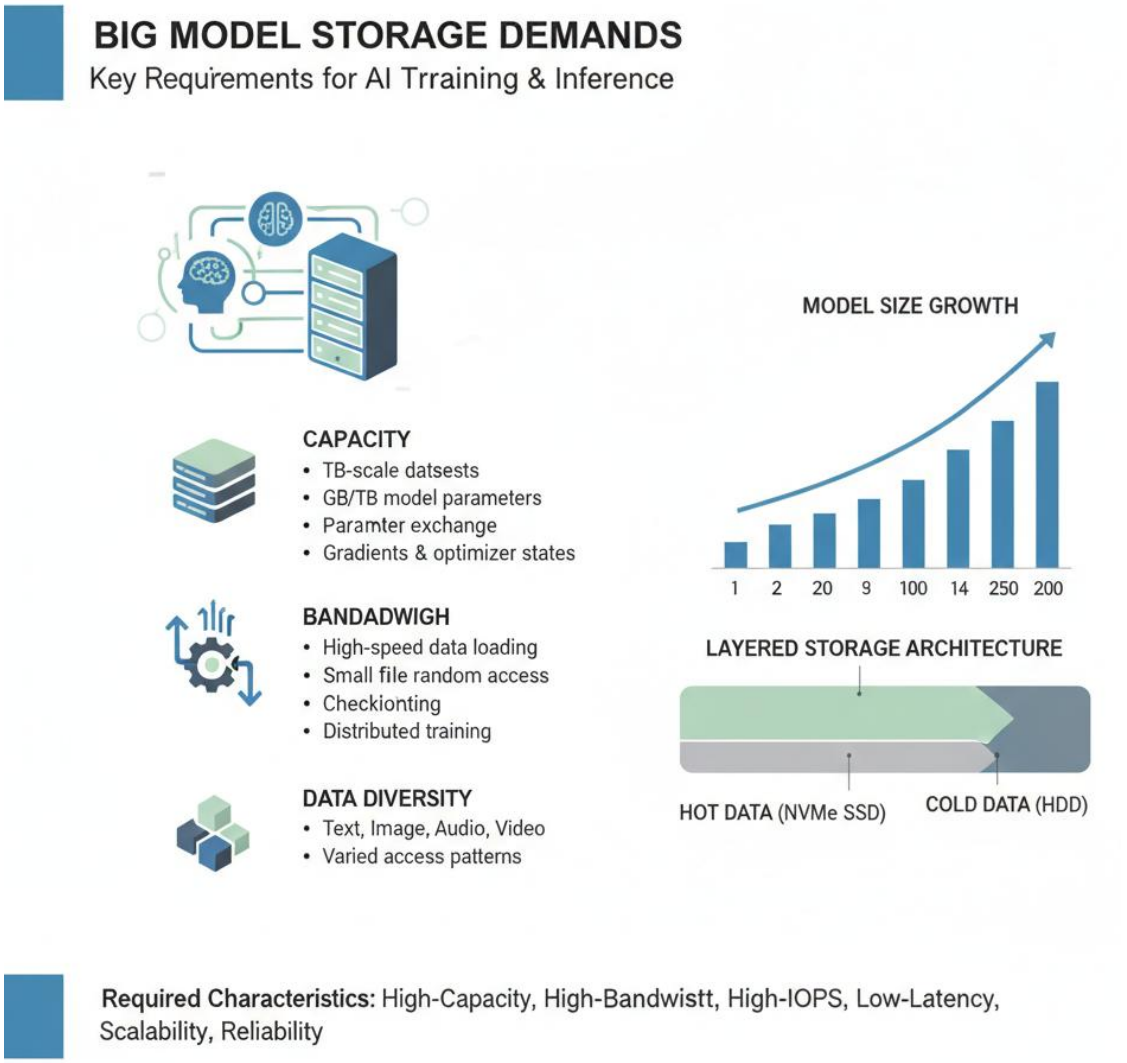
集群通信优化的系统性方法

集群通信优化还需要考虑网络拓扑感知、通信计算重叠、异步通信等技术。网络拓扑感知根据实际网络拓扑优化通信路径，减少跨交换机通信；通信计算重叠通过流水线技术，将通信与计算重叠执行，隐藏通信延迟；异步通信则允许计算任务在通信进行时继续执行，提高资源利用率。

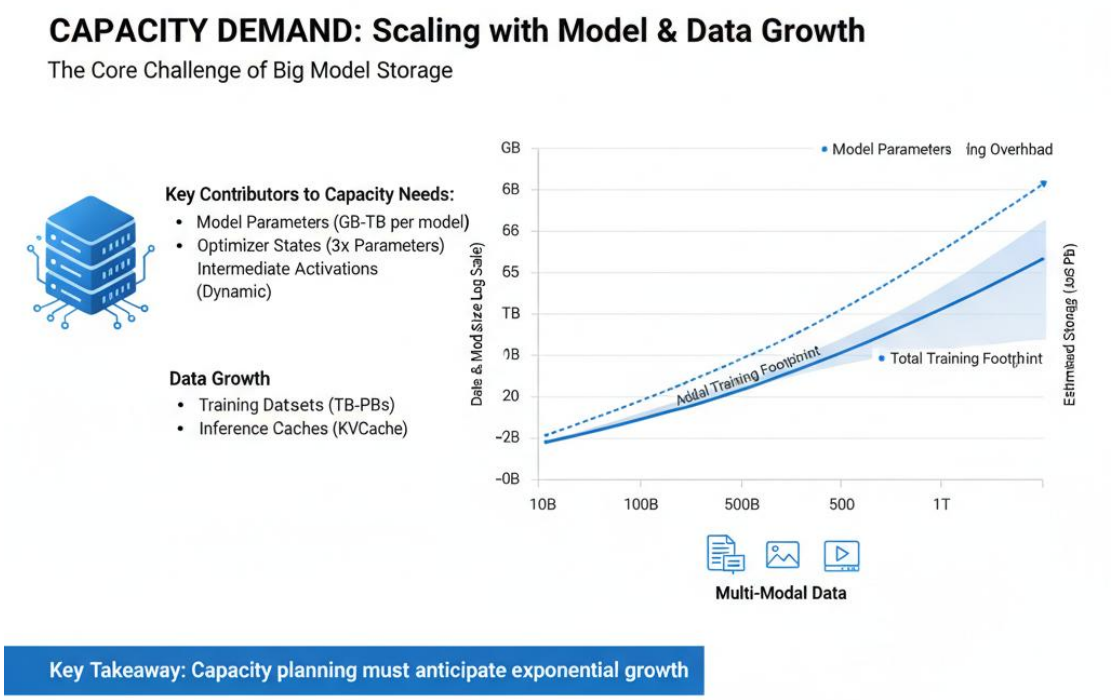
在大规模 AI 集群中，集群通信优化是一个系统工程，需要硬件、网络、软件、算法等多层面的协同优化。随着集群规模的不断扩大和模型复杂度的持续增加，集群通信优化将成为异构算力系统设计的关键挑战和研究热点。

3.3 存储与数据管理

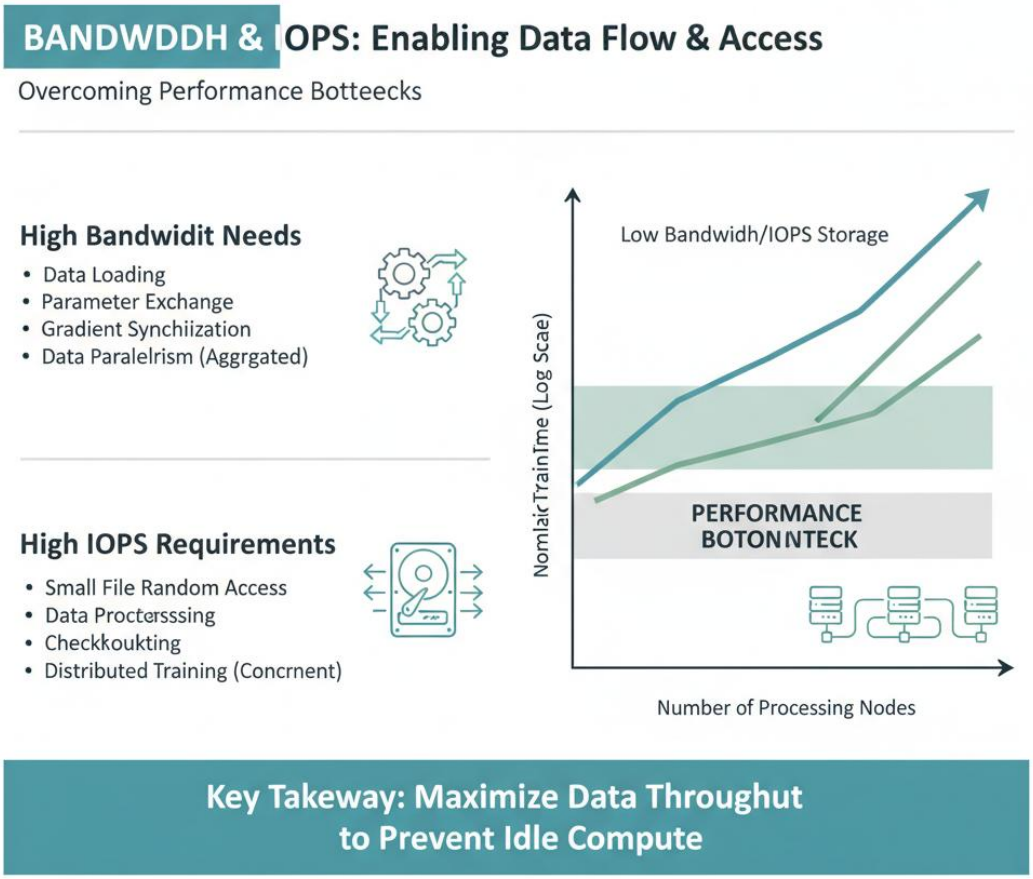
3.3.1 大模型存储需求



大模型训练和推理对存储系统提出了极高的要求，包括存储容量、带宽、IOPS 等多个方面。随着模型参数规模的不断扩大和数据量的爆炸式增长，存储系统已成为大模型训练的重要瓶颈之一。

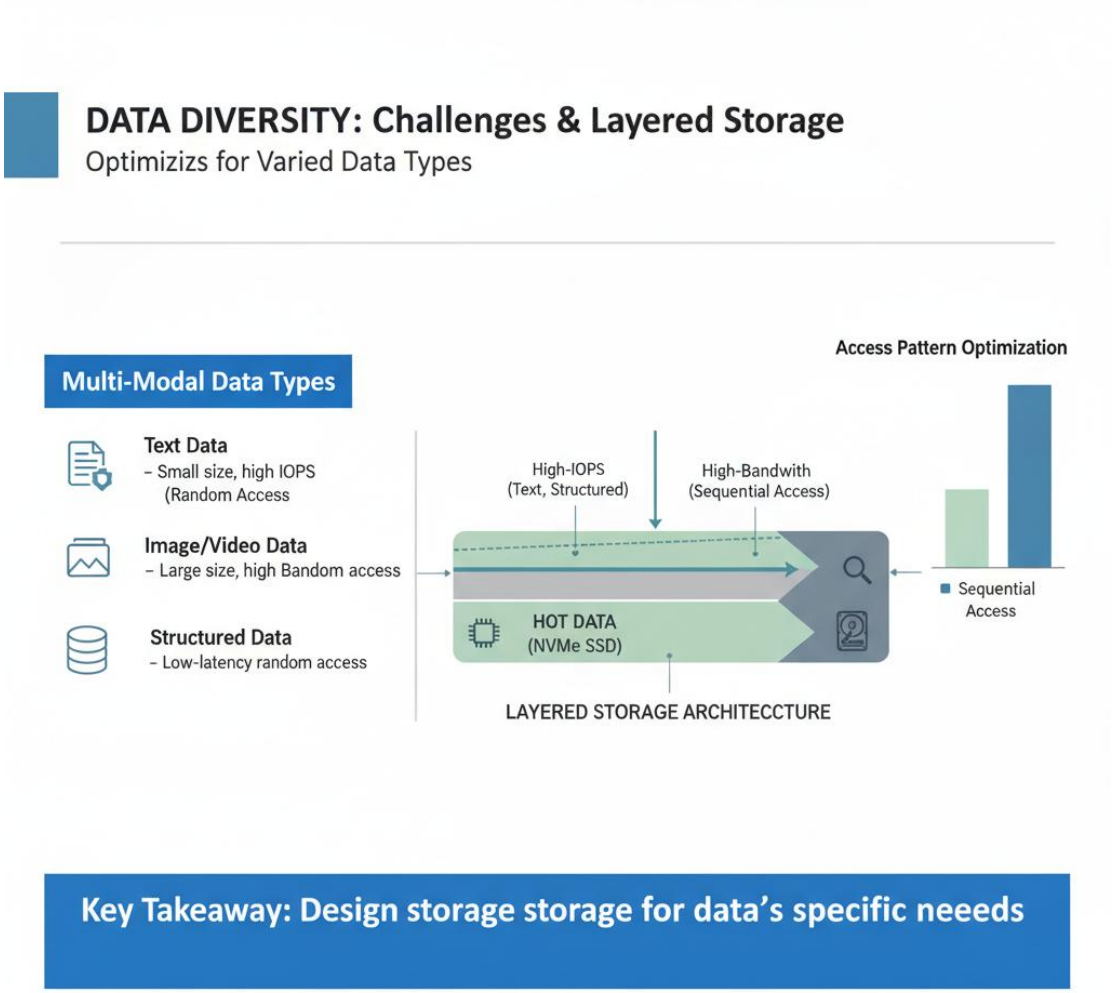


在存储容量方面，大模型训练涉及的数据集规模可达 TB 级，模型参数本身也需要 GB 级甚至 TB 级的存储空间。以千亿参数大模型为例，仅模型参数就需要数百 GB 的存储空间（假设每个参数为 16 位浮点数）。在训练过程中，还需要存储梯度、优化器状态、中间激活值等，进一步增加了存储需求。对于推理场景，虽然不需要存储训练相关数据，但模型参数和缓存（如 KVCache）仍需要大量存储空间。



在存储带宽方面，大模型训练需要高带宽的存储系统来支持数据的高效加载。训练过程中的数据加载、参数交换、梯度同步等操作都需要高存储带宽支持。特别是在数据并行训练中，每个节点都需要独立加载数据，对存储系统的聚合带宽要求极高。存储带宽不足会导致计算资源闲置，降低训练效率。

在 IOPS（每秒输入/输出操作数）方面，大模型训练通常涉及大量小文件的随机访问，如数据预处理、检查点保存/恢复等操作，需要高 IOPS 的存储系统支持。特别是在分布式训练中，多个节点同时访问存储系统，对 IOPS 的要求呈倍数增长。



大模型存储需求还体现在数据多样性上。训练数据通常包括文本、图像、音频、视频等多种模态，每种数据类型对存储系统的要求各不相同。文本数据通常体积小但数量多，需要高 IOPS；图像和视频数据体积大，需要高带宽；结构化数据则需要低延迟的随机访问能力。

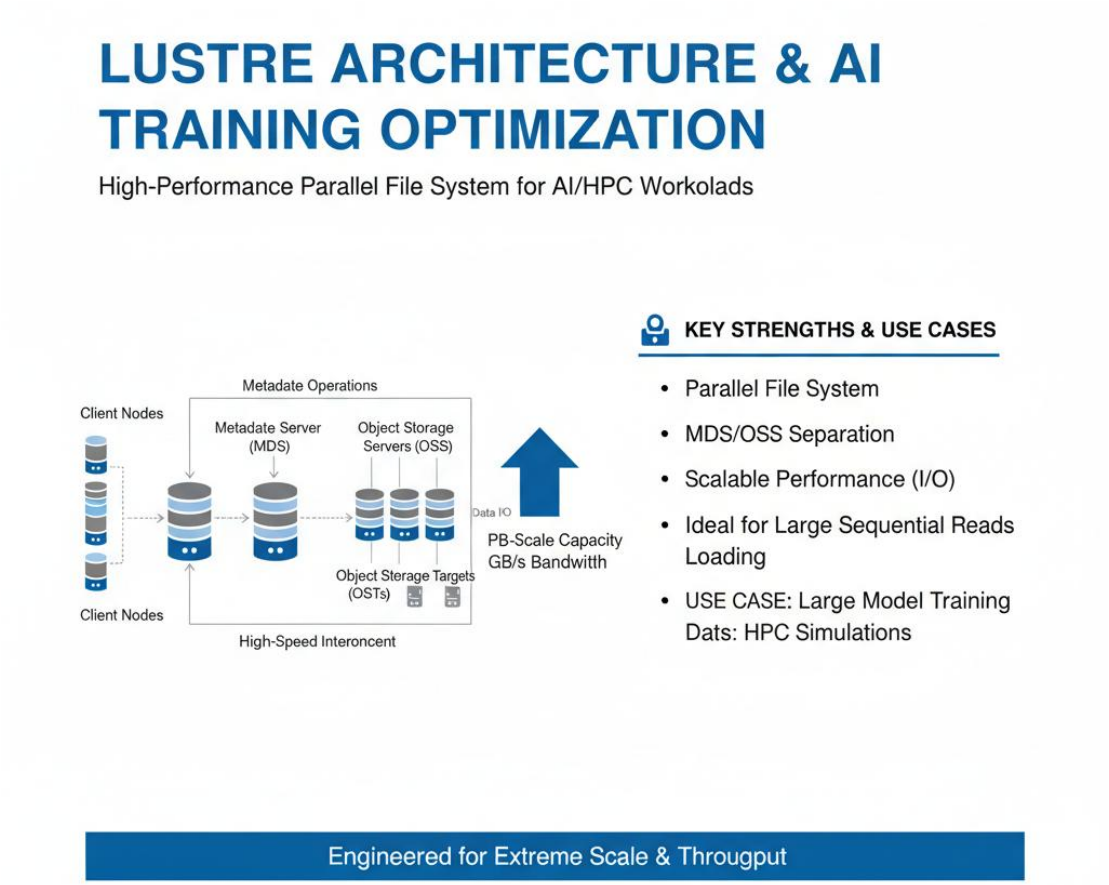
为了满足大模型存储需求，存储系统需要具备以下特性：高容量、高带宽、高 IOPS、低延迟、可扩展性、可靠性等。在实际系统设计中，通常采用分层存储架构，将热数据存储在高性能存储介质（如 NVMe SSD）上，冷数据存储在大容量存储介质（如 HDD）上，实现成本与性能的平衡。

3.3.2 分布式存储技术

分布式存储技术是满足大模型存储需求的关键，通过将数据分散存储在多个节点上，实现存储容量的线性扩展和性能的并行提升。在大模型训练场景中，分布式存储技术需要解决数据分片、缓存、预取等关键问题。

主流分布式存储技术对比，聚焦 Lustre、GPFS 和 Ceph 的特点和适用场景

Lustre 是一种高性能并行文件系统，广泛应用于 HPC 和大规模 AI 训练场景。Lustre 采用元数据服务器(MDS)和对象存储服务器(OSS)分离的架构，支持 PB 级存储容量和数百 GB/s 的聚合带宽。Lustre 的优势在于高性能、高可扩展性，特别适合大文件顺序读写场景，如大模型训练中的数据加载。然而，Lustre 在小文件处理和元数据操作方面相对较弱，需要配合其他技术使用。

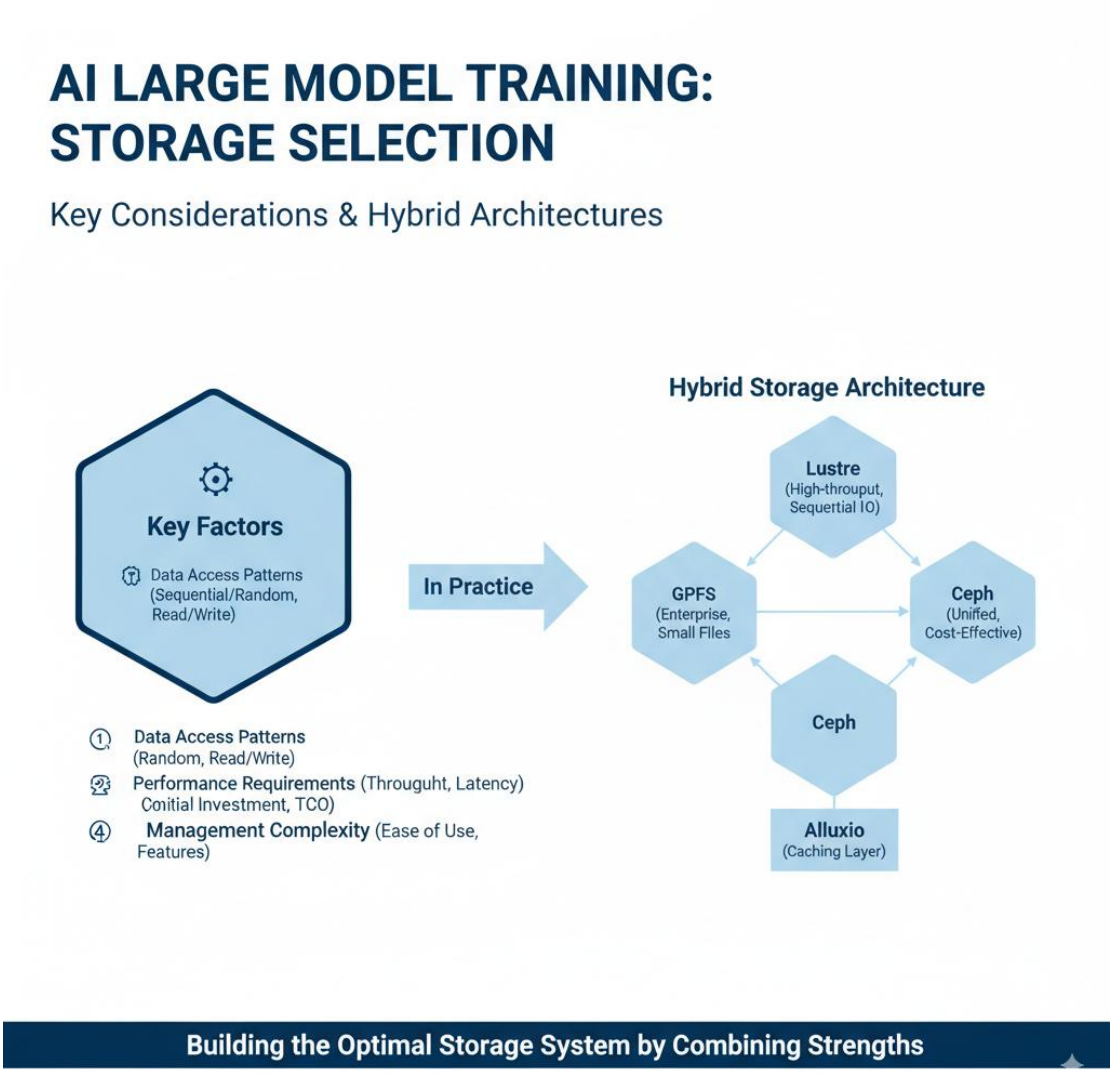


Lustre 的架构和优势

GPFS（General Parallel File System，现称 IBM Spectrum Scale）是 IBM 开发的高性能分布式文件系统，支持多种存储架构和访问协议。GPFS 采用共享磁盘架构，通过分布式锁管理机制实现数据一致性，支持高并发访问。GPFS 的优势在于全面的特性支持、良好的小文件性能和强大的管理功能，适合复杂的企业级 AI 训练环境。

Ceph 是一种统一的分布式存储系统，支持块存储、对象存储和文件存储三种接口，被称为统一存储。Ceph 采用 CRUSH 算法实现数据分布，无需中心化的元数据服务器，具有良好的可扩展性和容错性。Ceph 的优势在于统一架构、自修复能力和成本效益，适合需要多种存储接口的 AI 平台。然而，Ceph 在性能方面通常不如专用的并行文件系统，特别是在低延迟场景下。

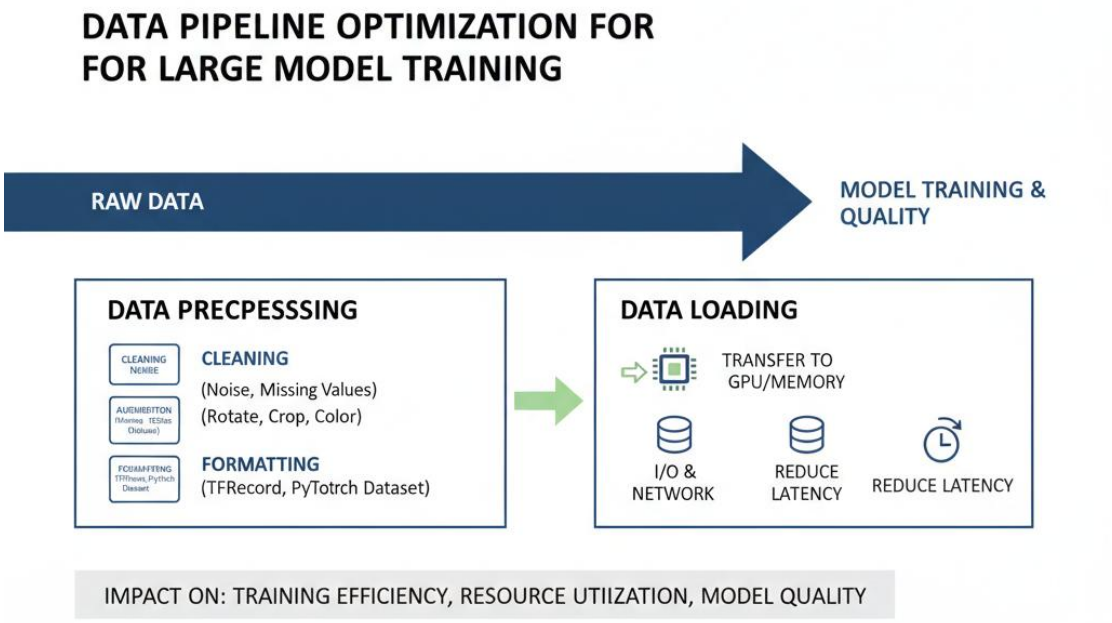
除了上述主流分布式存储技术外，还有一些针对 AI 场景优化的存储解决方案。例如，Alluxio 严格来说不是一个文件系统，而是构建在其他分布式文件系统之上的分布式缓存系统，在大数据领域使用非常广泛。Alluxio 通过内存缓存加速数据访问，特别适合多次迭代的 AI 训练场景。



分布式存储技术选择的关键考量因素和混合存储架构的理念

在大模型训练中，分布式存储技术的选择需要考虑多个因素：数据访问模式、性能需求、成本预算、管理复杂度等。对于大规模顺序访问为主的训练场景，Lustre 是理想选择；对于需要多种存储接口的复杂环境，Ceph 提供统一解决方案；对于企业级关键应用，GPFS 提供全面的特性支持。在实际部署中，通常采用混合存储架构，结合不同存储技术的优势，构建最适合大模型训练的存储系统。

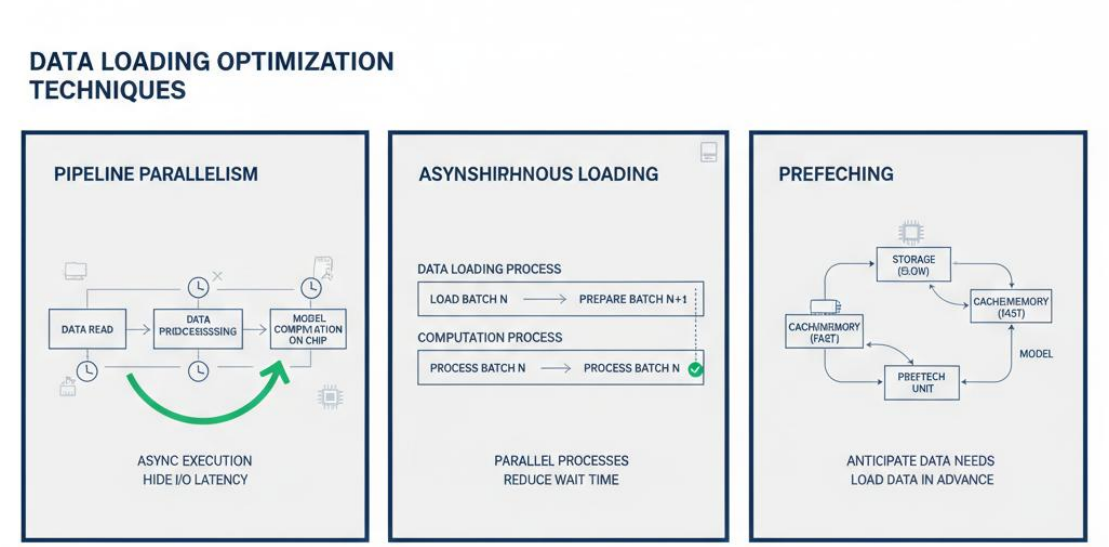
3.3.3 数据预处理与加载



数据预处理与加载是大模型训练流程中的重要环节，直接影响训练效率和模型质量。高效的数据预处理与加载技术能够最大化计算资源利用率，减少 I/O 等待时间，提高整体训练吞吐量。

数据预处理包括数据清洗、增强、格式转换等多个步骤。数据清洗主要处理原始数据中的噪声、缺失值、异常值等问题，确保训练数据的质量；数据增强通过旋转、裁剪、翻转、颜色变换等技术扩充训练数据集，提高模型的泛化能力；格式转换则将不同来源、不同格式的数据统一为模型训练所需的格式，如 TensorFlow 的 TFRecord、PyTorch 的 Dataset 等。

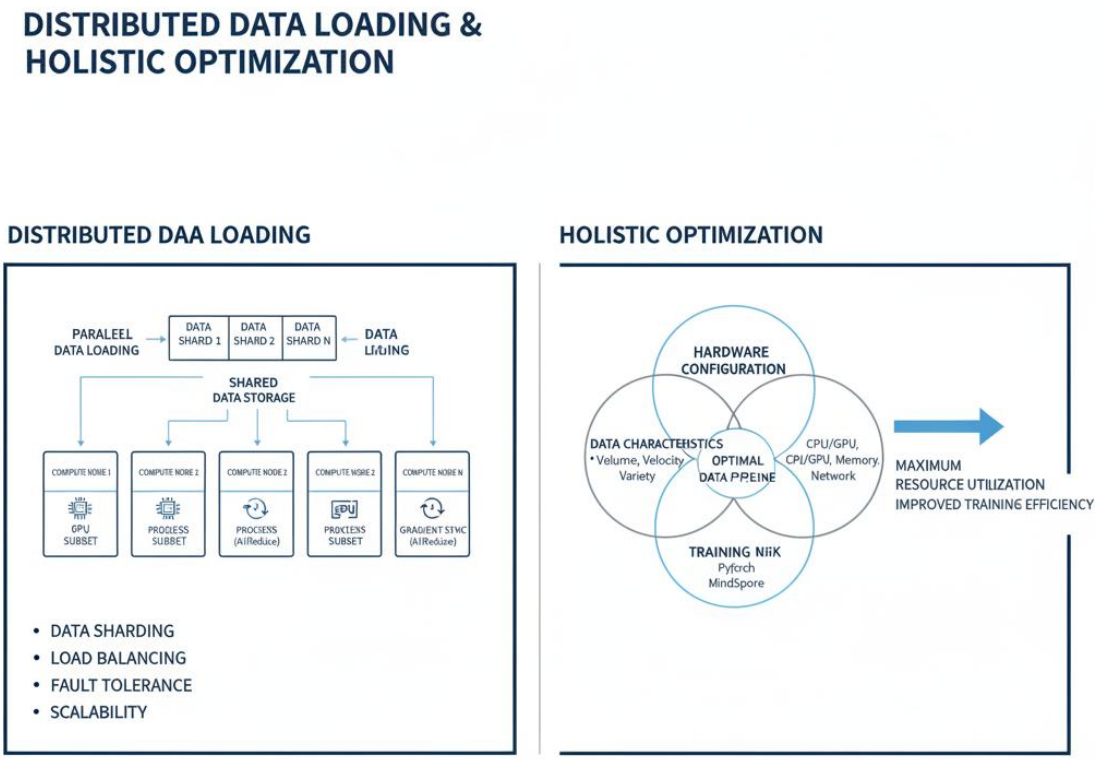
数据加载是将预处理后的数据高效传输到计算设备（如 GPU）内存中的过程。在大模型训练中，数据加载往往成为性能瓶颈，特别是在分布式训练场景中，多个计算节点同时加载数据，对存储系统和网络带宽提出极高要求。为了优化数据加载效率，通常采用以下技术：



流水线并行（Pipeline Parallelism）是一种常用的数据加载优化技术，将数据读取、数据预处理计算、以及芯片上的模型计算三个步骤异步并行执行。这三步构成了典型的数据生产者和数据消费者的上下游关系，通过流水线技术可以隐藏 I/O 延迟，提高资源利用率。MindSpore 等框架提供了灵活的数据集加载方法、丰富的数据处理操作，以及自动数据增强、动态批处理等功能，支持高效的数据流水线。

异步加载是另一种重要的优化技术，使用多个进程来并行加载和预处理数据，通过流水线处理减少数据等待时间。在异步加载模式下，数据加载进程与计算进程并行执行，计算进程在处理当前批次数据时，数据加载进程已经在准备下一批次数据，从而隐藏数据加载延迟。PyTorch 的 DataLoader、TensorFlow 的 tf.data 等 API 都支持异步加载模式。

预取（Prefetching）技术通过预测未来需要的数据，提前将其加载到内存或缓存中，减少数据访问延迟。预取技术可以与缓存技术结合使用，将频繁访问的数据保存在高速存储介质中，进一步提高数据访问效率。在大模型训练中，常用的预取策略包括基于访问模式的预取、基于训练进度的预取等。



分布式数据加载是大规模分布式训练中的关键技术，通过将数据分片存储在多个节点上，实现数据加载的并行化。在数据并行训练中，每个节点负责加载和处理数据的一个子集，通过 AllReduce 等通信原语实现梯度同步。分布式数据加载需要解决数据分片、负载均衡、容错等问题，确保每个节点都能高效获取所需数据。

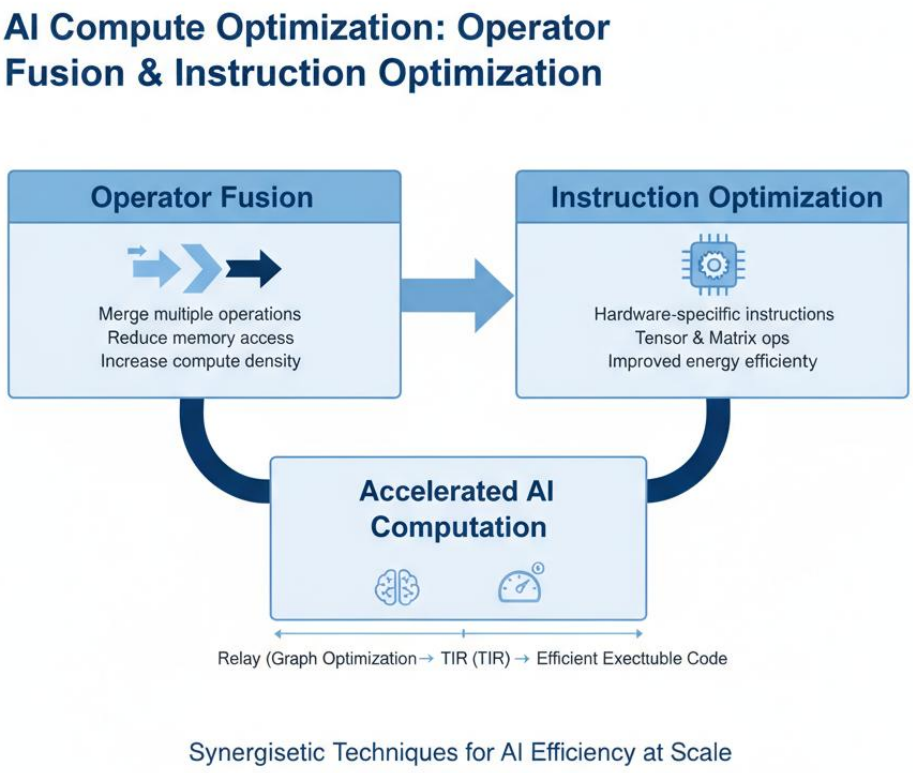
在实际应用中，数据预处理与加载的优化需要综合考虑数据特性、硬件配置、训练框架等多个因素。通过合理选择和组合上述技术，可以构建高效的数据流水线，最大化计算资源利用率，提高大模型训练的整体效率。

四、大模型与异构算力融合关键技术

4.1 软硬件协同优化

4.1.1 算子融合与指令优化

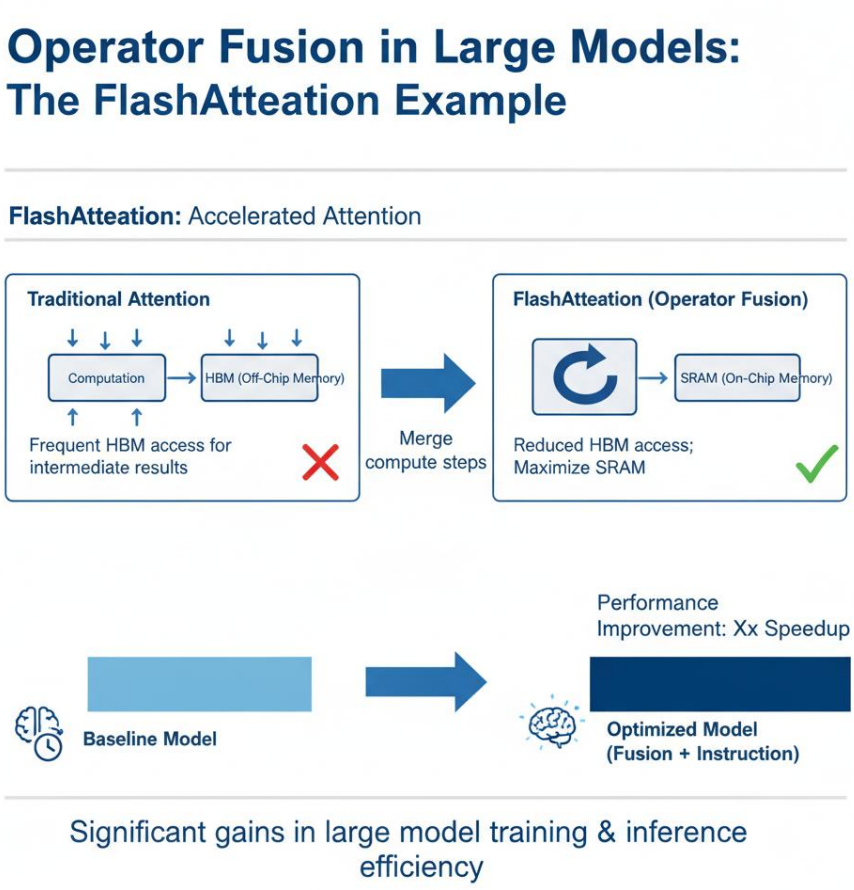
算子融合与指令优化是软硬件协同优化的核心技术，通过将多个计算操作合并为一个更大的操作，减少内存访问次数，提高计算密度，从而显著提升 AI 计算效率。在大模型训练和推理场景中，算子融合技术已成为性能优化的关键手段。



算子融合和指令优化

算子融合的核心思想是将多个连续的算子合并为一个更大的融合算子，减少中间结果的内存读写，降低内存带宽压力。以 FlashAttention 为例，其加速原理非常简单，就是更多地去利用带宽更高的上层存储单元，减少对低速下层存储单元的访问频率，从而达到加速的目的。在传统的注意力计算中，需要多次读写显存来存储中间结果，而 FlashAttention 通过算子融合技术，将多个计算步骤合并，大幅减少了内存访问次数，显著提升了计算效率。

指令优化是另一种重要的软硬件协同优化技术，通过针对特定硬件架构设计高效的指令集，提高计算密度和能效比。寒武纪 MLU 指令集就是专门针对 AI 计算优化的指令集，支持张量运算、矩阵运算等 AI 核心操作，相比通用指令集具有更高的计算效率。指令优化需要深入理解硬件架构特点，设计能够充分利用硬件计算能力的指令序列，同时考虑数据局部性和内存访问模式，实现最优的性能表现。



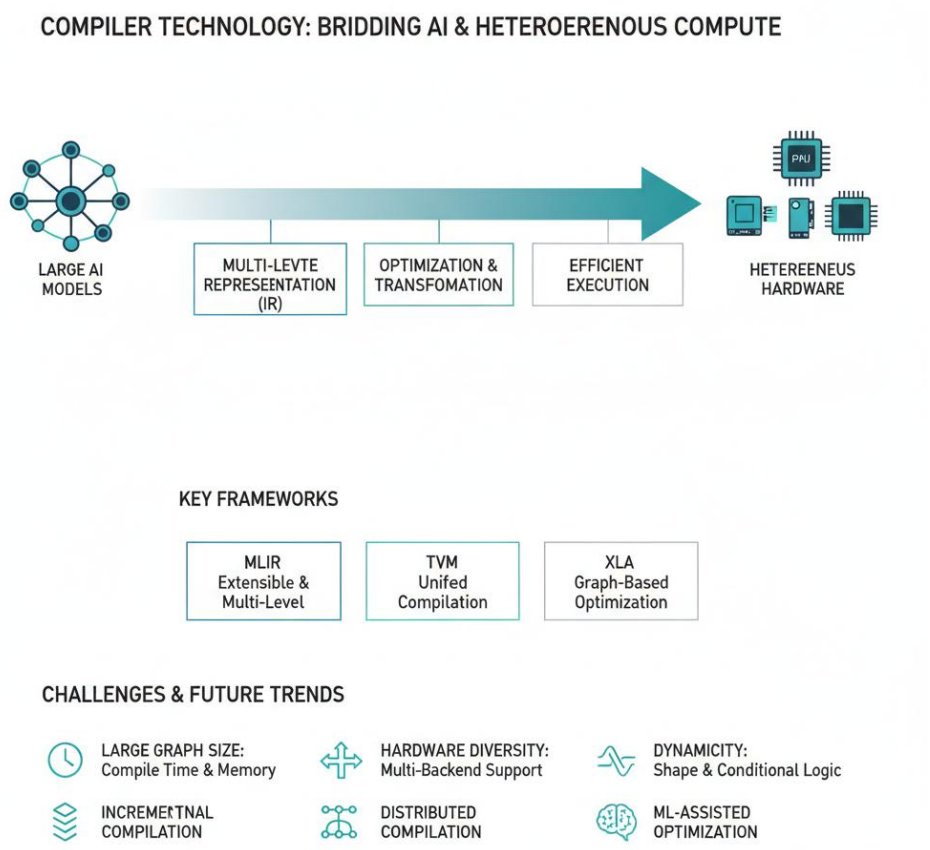
算子融合细节

在实际应用中，算子融合与指令优化通常结合使用，形成完整的优化方案。以寒武纪 BANG 算子库为例，它不仅提供了丰富的融合算子，还针对 MLU 硬件架构进行了深度指令优化，实现了算子层面的极致性能。通过 Relay 导入推理模型，进行算子融合等图层优化，通过 TIR 生成融合算子，最终形成针对特定硬件的高效执行代码。

算子融合与指令优化的效果在大模型场景中尤为显著。大模型通常包含大量的矩阵运算、注意力计算等操作，这些操作通过算子融合可以大幅减少内存访问，提高计算效率。同时，大模型对计算资源的需求极高，通过指令优化可以充分利用硬件计算能力，降低单位计算的成本。在实际案例中，经过算子融合和指令优化的模型，其性能可提升数倍甚至数十倍，能效比也有显著改善。

4.1.2 编译器与中间表示

编译器与中间表示技术是连接 AI 模型与异构硬件的桥梁，通过多层次的中间表示和优化转换，实现模型在不同硬件平台上的高效执行。随着异构算力的普及，编译器技术在大模型与异构算力融合中扮演着越来越重要的角色。



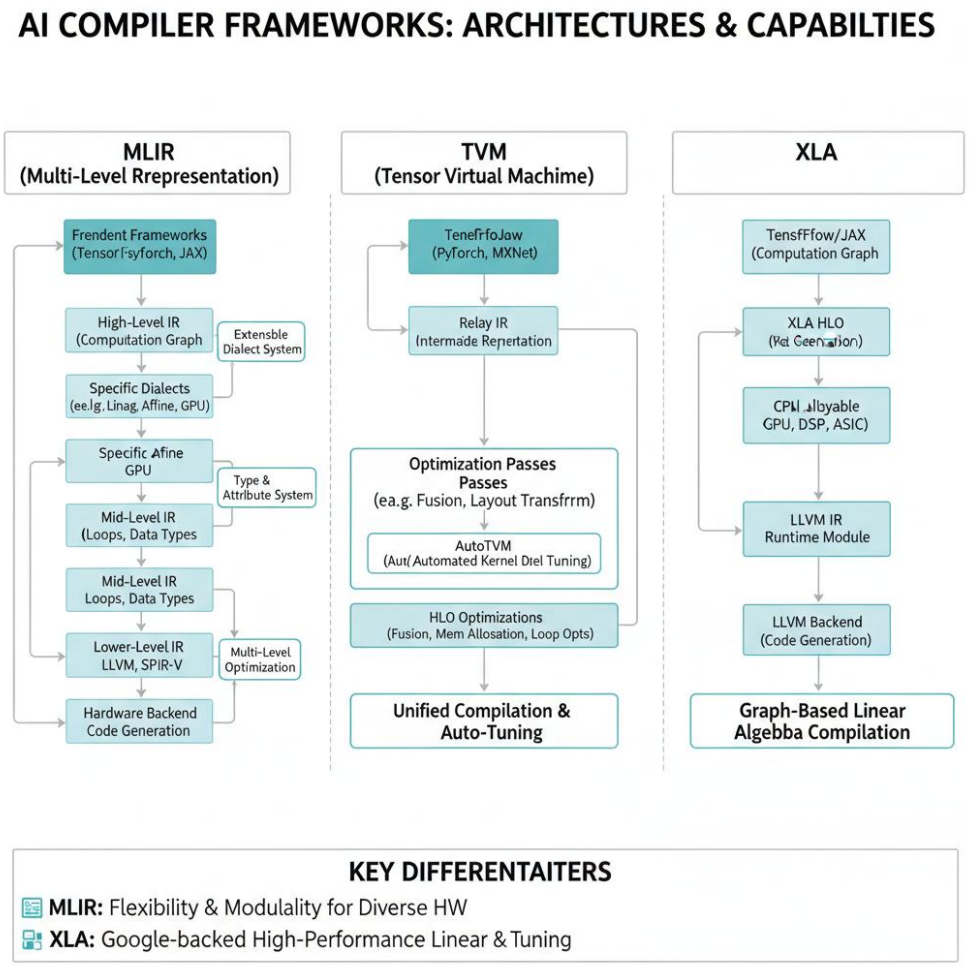
编译器技术在大模型与异构算力融合中的核心作用和发展趋势

MLIR（Multi-Level Intermediate Representation）是一种新兴的多级中间表示框架，支持不同抽象层次的 IR 定义和转换。MLIR 的架构设计原理包括可扩展的方言系统、类型系统、属性系统等，允许在不同抽象层次上定义和优化计算图。MLIR 不仅是一种中间表示，更是一个编译器框架，支持从高级计算图到底层硬件指令的全流程优化。在大模型编译中，MLIR 可以实现从计算图优化到硬件代码生成的无缝衔接，为异构算力提供统一的编译支持。

TVM（Tensor Virtual Machine）是面向深度学习的模型编译器，用户可直接获得编译/优化模型为推理 blob 的能力，可以看做机器学习时代的 GCC。TVM 支持多种前端框架（TensorFlow、PyTorch、MXNet 等）和多种后端硬件（CPU、GPU、AI 加速器等），通过统一的中间表示（Relay IR）和优化 passes，实现模型的高效编译和部署。TVM 的自动调优功能（AutoTVM）可以针对特定硬件自动生成最优的计算算子，大幅提升模型执行效率。

XLA（Accelerated Linear Algebra）是 Google 开发的线性代数编译器，最初

旨在加速 TensorFlow 模型，现已被 JAX 等框架采用。XLA 将计算图编译为高效的机器代码，通过算子融合、内存分配优化、循环优化等技术，提升计算效率。整个编译流程先将 TensorFlow 的图转化为 XLA HLO，即一种类似高级语言的图的中间表达形式，可以基于此进行一些 High-Level 的优化。接着将 XLA HLO 翻译为 LLVM IR，进行底层优化和代码生成。



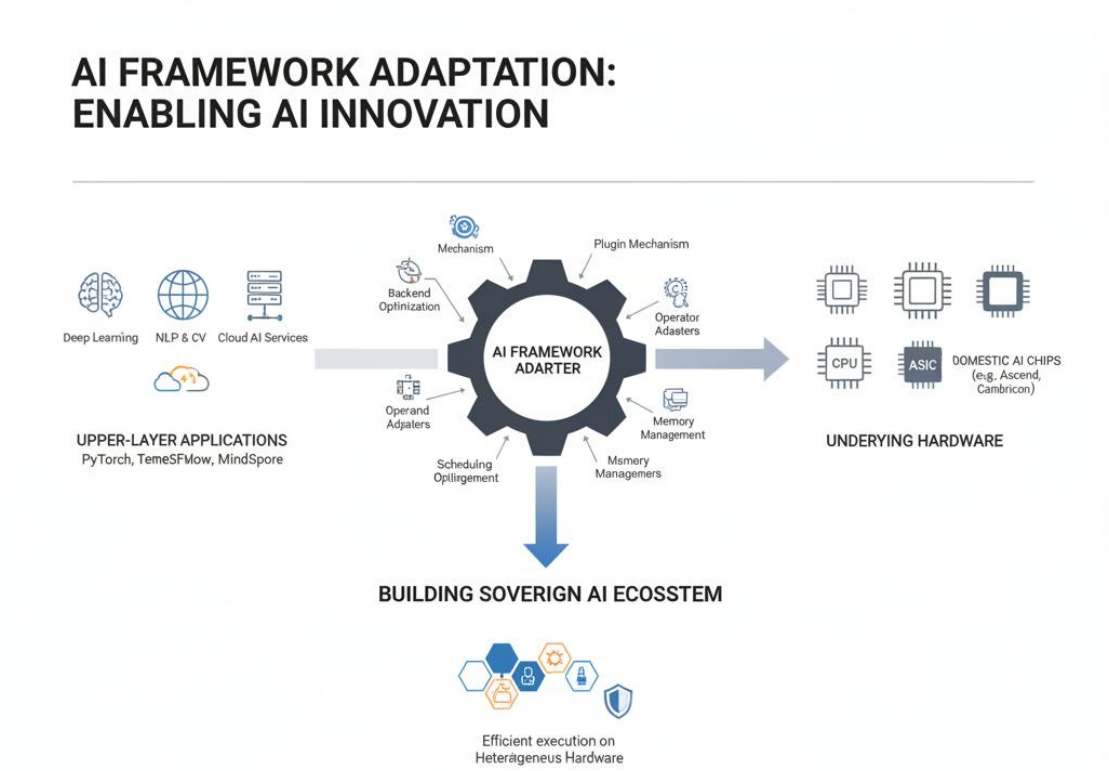
MLIR、TVM、XLA 这三个主流编译器框架的架构和功能特点

除了上述主流编译框架外，还有针对特定硬件的编译器，如 NVIDIA 的 NVCC（NVIDIA CUDA 编译器），仅适用于 CUDA；华为的昇腾编译器，针对昇腾芯片优化等。这些编译器通常与硬件深度绑定，能够充分发挥特定硬件的性能潜力。

在大模型与异构算力融合中，编译器技术面临诸多挑战：一是大模型的计算图规模庞大，编译时间和内存消耗成为瓶颈；二是异构硬件的多样性要求编译器支持多种后端；三是大模型的动态特性（如动态形状、条件计算等）增加了编译优化的复杂度。为应对这些挑战，编译器技术正在向更高效、更通用、更智能的

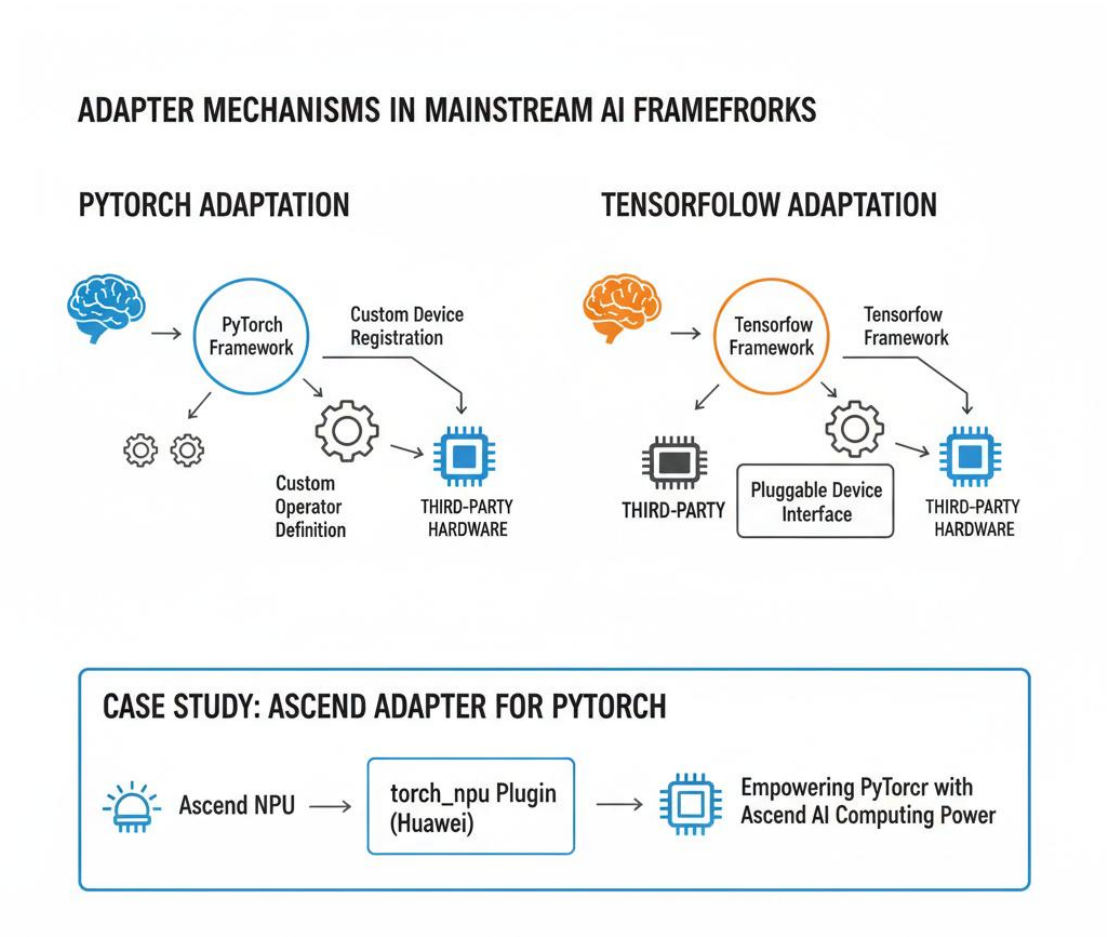
方向发展，如增量编译、分布式编译、机器学习辅助优化等。

4.1.3 AI 框架适配



AI 框架适配是连接上层应用与底层硬件的关键环节，通过插件机制、后端优化等方式，使主流 AI 框架能够高效运行在异构硬件上。随着国产 AI 芯片的快速发展，AI 框架适配技术成为构建自主可控 AI 生态的重要组成部分。

PyTorch 和 TensorFlow 是目前最主流的 AI 框架，它们都提供了插件机制，支持第三方硬件的接入。在 PyTorch 框架中，可以通过注册自定义设备、算子等方式实现硬件适配；在 TensorFlow 框架中，可以通过 Pluggable Device 接口支持新的硬件设备。以昇腾 NPU 为例，华为开发了名为 torch_npu 的 Ascend Adapter for PyTorch 插件，使得昇腾 NPU 可以与 PyTorch 框架兼容，为使用 PyTorch 框架的开发者提供了强大的昇腾 AI 处理器算力支持。

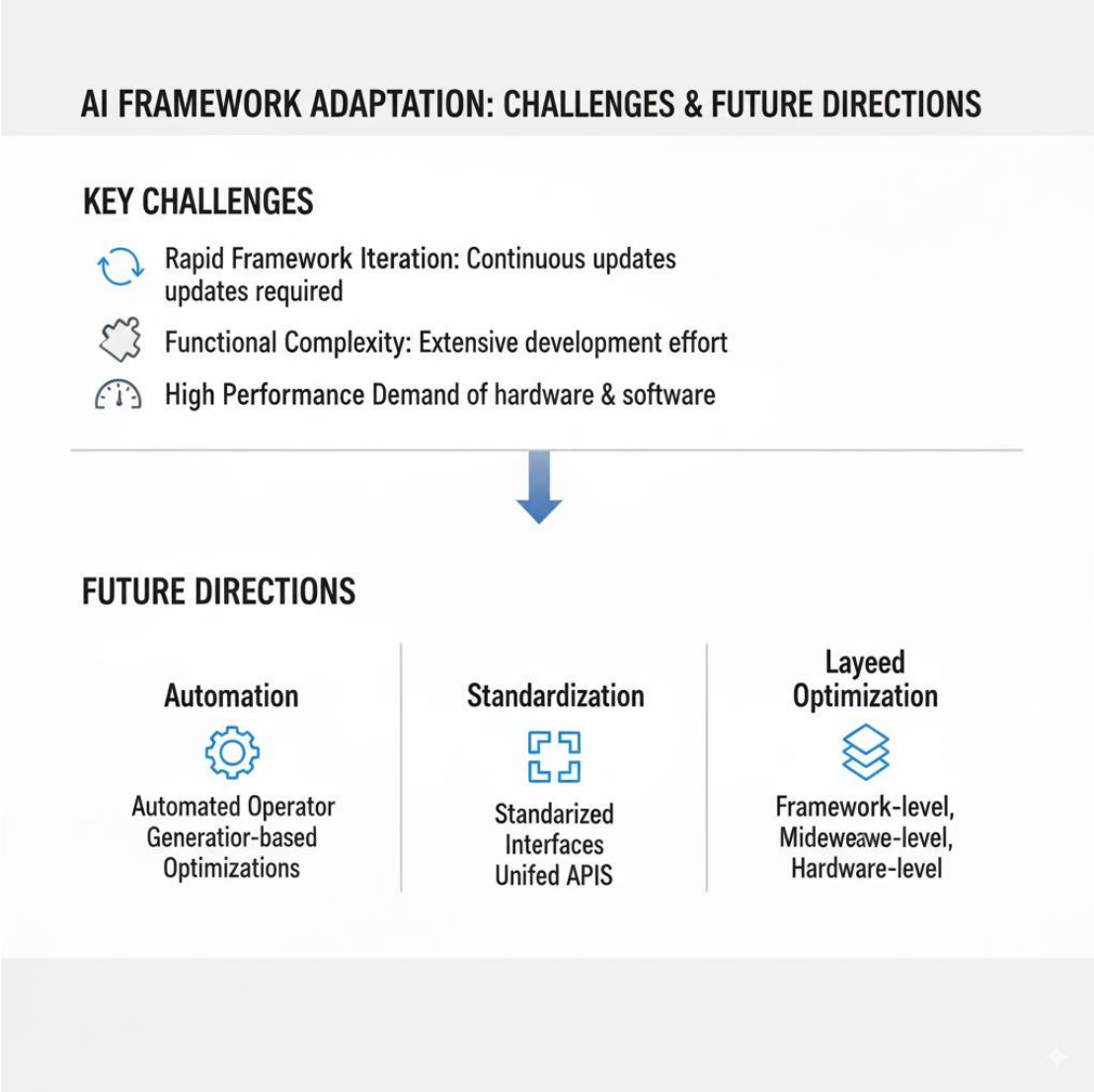


主流 AI 框架的适配机制与实例

适配插件开发是 AI 框架适配的核心工作，主要包括算子适配、内存管理、调度优化等方面。算子适配是将框架中的算子映射为硬件支持的操作，通常需要实现算子的前向计算、反向传播、形状推导等功能。内存管理包括内存分配、释放、复用等，需要考虑硬件的内存层次结构和访问特性。调度优化则涉及算子执行顺序、并行策略等，需要充分利用硬件的并行计算能力。

自动混合精度训练是 AI 框架适配中的重要优化技术。通过自动将模型中的部分操作转换为低精度计算（如 FP16、BF16），可以显著减少内存占用和计算量，提高训练效率。现代 AI 框架如 PyTorch、TensorFlow 都提供了自动混合精度训练的支持，硬件适配层需要实现相应的低精度算子和转换逻辑。

寒武纪 BANG 算子库是国产 AI 芯片框架适配的典型案例。BANG 算子库提供了丰富的 AI 计算算子，支持 TensorFlow、PyTorch、MindSpore 等主流框架，通过高效的算子实现和内存管理，充分发挥寒武纪芯片的计算能力。BANG 算子库不仅包含基础算子，还提供了针对大模型的优化算子，如注意力计算、矩阵乘法等，为大模型训练和推理提供高性能支持。



AI 框架适配的挑战与未来方向

AI 框架适配面临的挑战包括：一是框架版本迭代快，适配工作需要持续更新；二是框架功能复杂，全面适配工作量大；三是性能优化要求高，需要深入理解框架和硬件的内部机制。为应对这些挑战，框架适配技术正在向更自动化、更标准化、更高效的方向发展，如自动算子生成、标准化接口、分层优化等。

4.2 大模型并行训练技术

4.2.1 数据并行

数据并行是最常用的大模型并行训练方式，通过将训练数据分割到多个计算设备上，实现训练过程的并行化。在数据并行训练中，数据集被分割成几个碎片，每个碎片被分配到一个设备上。这相当于沿批次（Batch）维度对训练过程进行并行化。每个设备将持有一个完整的模型副本，并独立计算本地数据的梯度，然

后通过AllReduce等通信原语实现梯度同步，最终所有设备获得一致的模型更新。

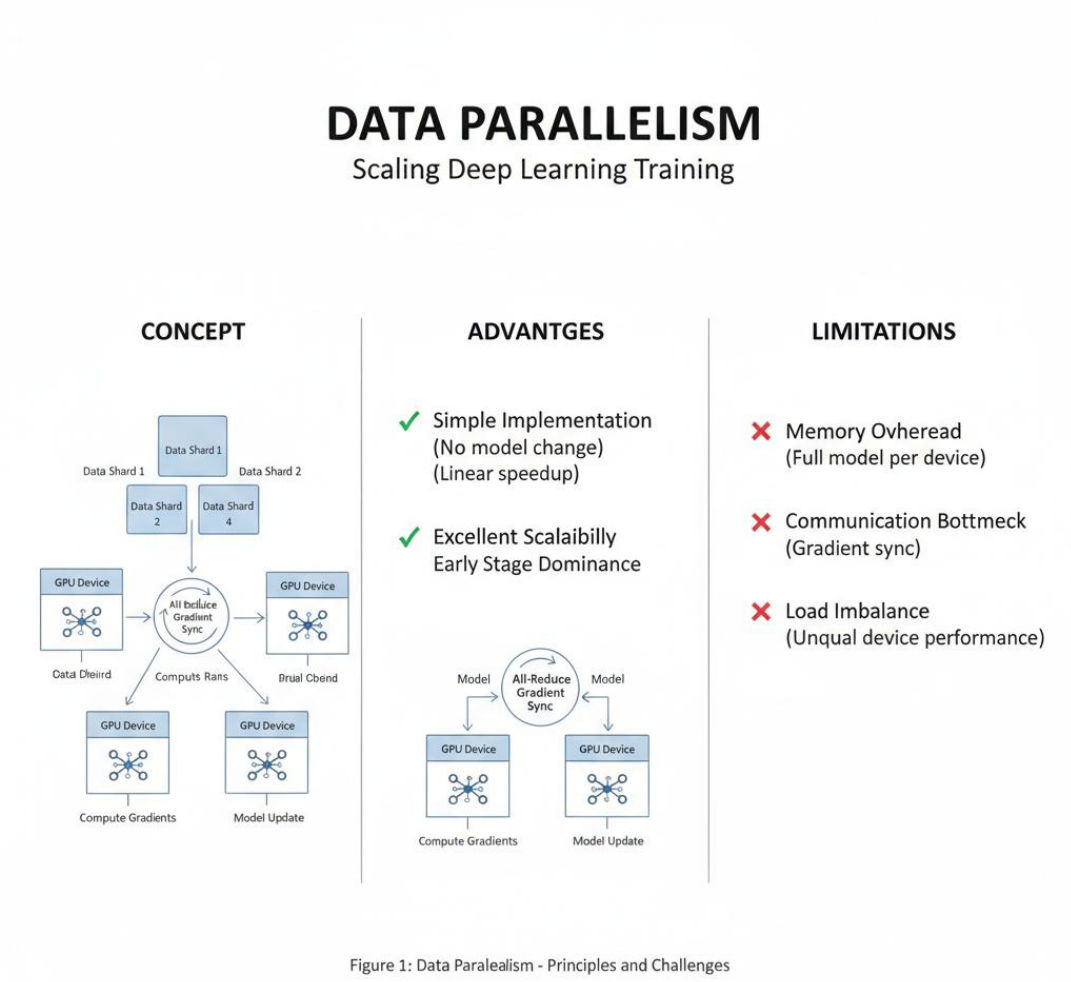


Figure 1: Data Paralealism - Principles and Challenges

数据并行的概念、优势和局限性

数据并行的核心优势在于实现简单、扩展性好。由于每个设备都维护完整的模型副本，不需要对模型结构进行修改，因此实现起来相对简单。同时，数据并行可以线性扩展到大量计算设备上，理论上训练速度可以随设备数量线性提升。在大模型训练的早期阶段，数据并行是最主要的并行方式。

然而，数据并行也存在明显的局限性。首先是内存占用问题，每个设备都需要存储完整的模型参数、梯度和优化器状态，对于大模型而言，单设备内存往往无法容纳。其次是通信开销问题，每个训练步骤都需要进行梯度同步，当模型规模增大或设备数量增多时，通信开销会成为性能瓶颈。最后是负载均衡问题，当计算设备性能不一致时，容易出现负载不均衡的情况，影响整体训练效率。

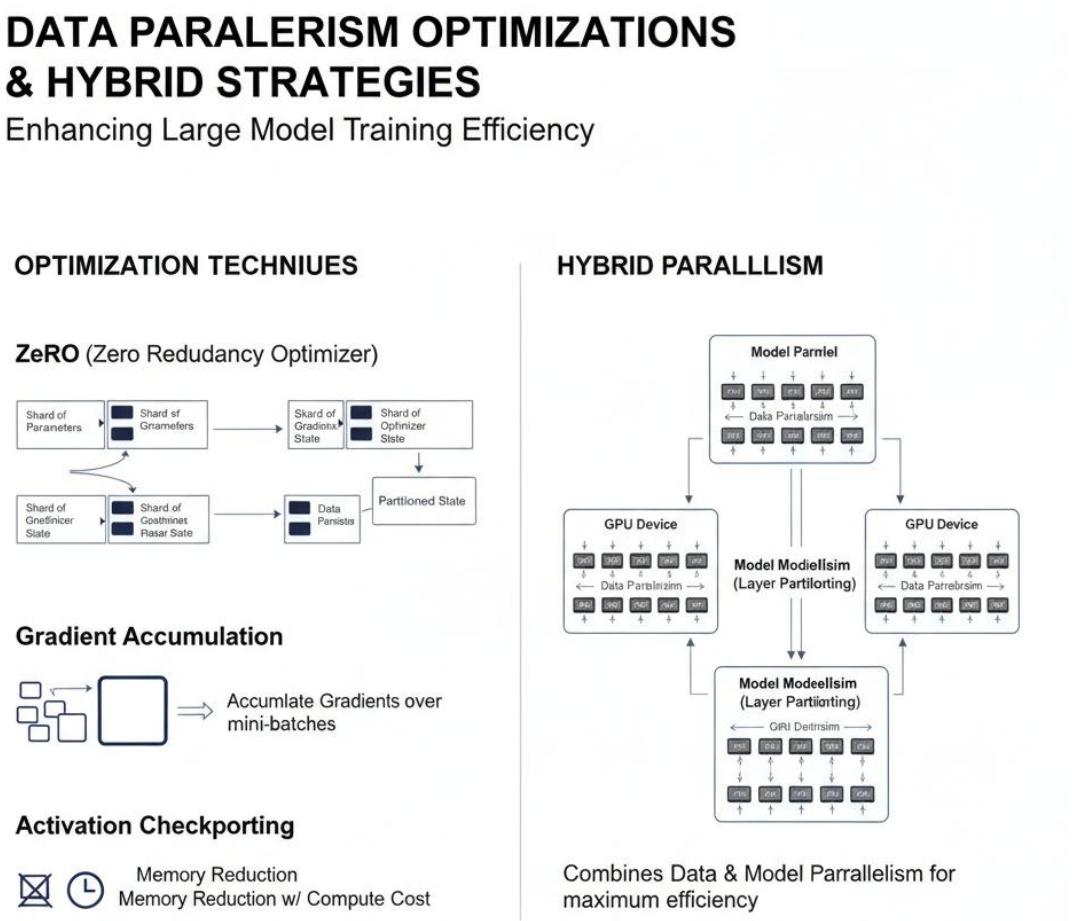


Figure 2: Data Paralelism - Optimization Techniques and Hybrid Strategies

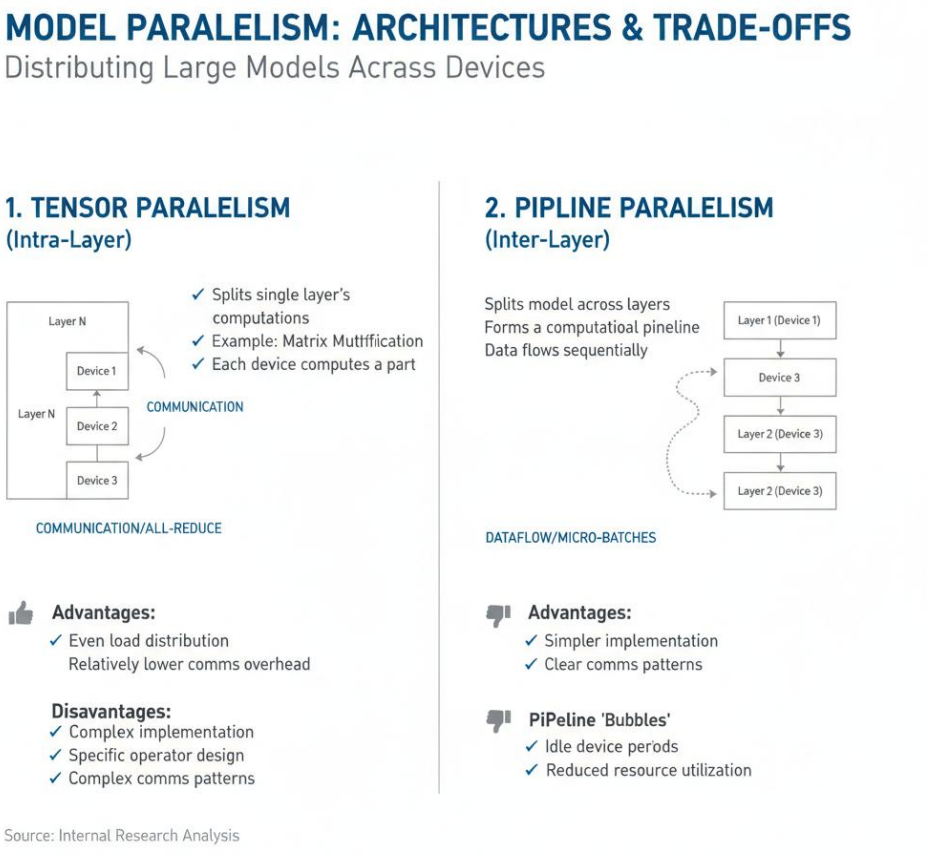
数据并行的优化技术和混合并行策略

为了缓解数据并行的内存压力，业界提出了多种优化技术。ZeRO（Zero Redundancy Optimizer）技术通过将优化器状态、梯度和参数分区存储在多个设备上，显著减少了单设备的内存占用。梯度累积技术通过累积多个小批次的梯度，模拟大批次训练的效果，可以在不增加内存占用的情况下使用更大的有效批次大小。激活检查点技术通过选择性存储和重新计算激活值，减少内存占用，但会增加额外的计算开销。

在实际应用中，数据并行通常与其他并行技术结合使用，形成混合并行策略。例如，可以将模型的不同层分配到不同的设备组上，每个设备组内部采用数据并行，设备组之间采用模型并行。这种混合并行策略可以充分发挥不同并行技术的优势，实现更高效的训练。

4.2.2 模型并行

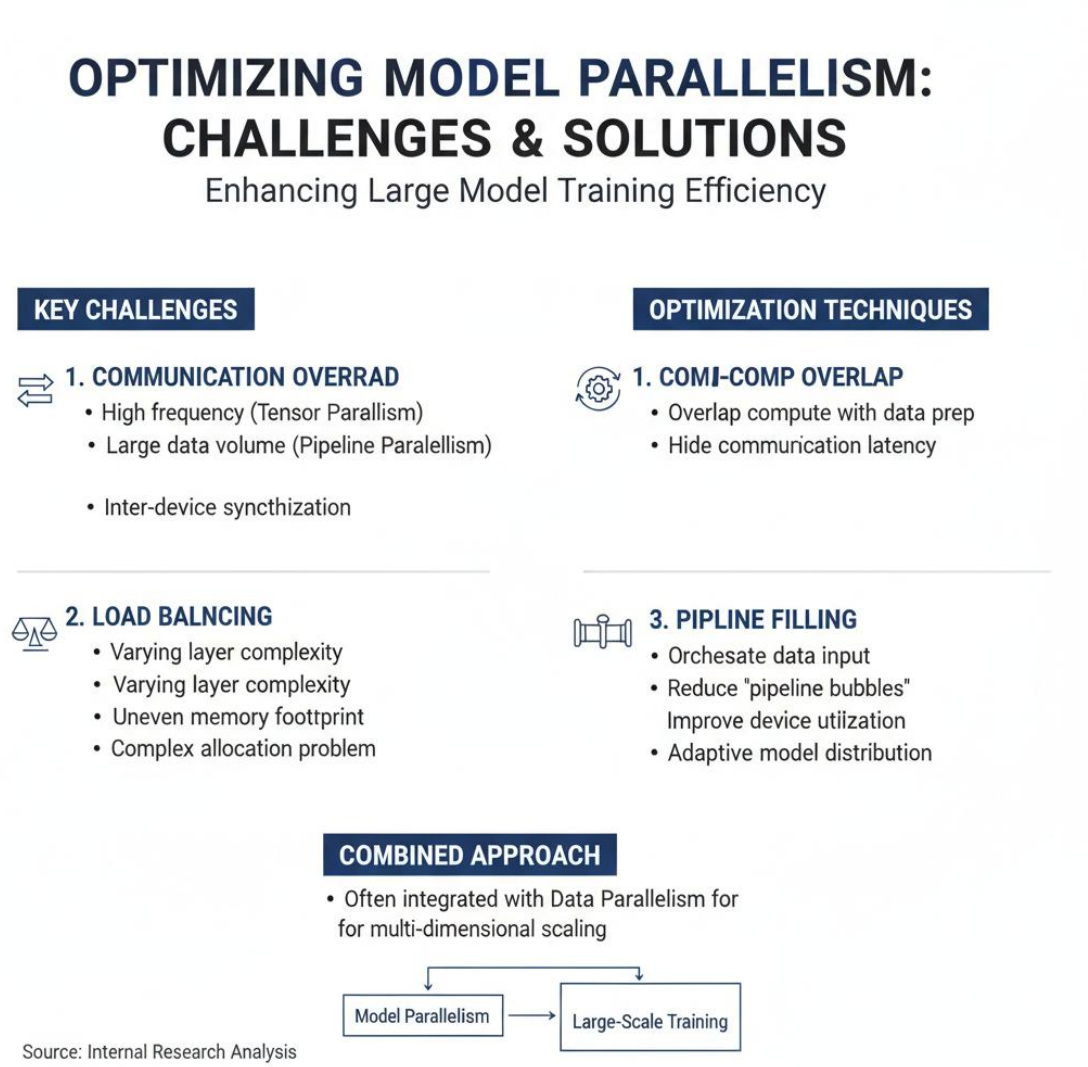
模型并行是另一种重要的大模型并行训练技术，通过将模型的不同部分分配到不同的计算设备上，解决单设备无法容纳完整大模型的问题。模型并行主要分为张量并行（Tensor Parallelism）和流水线并行（Pipeline Parallelism）两种形式。



模型并行的概述

张量并行(也称为层内并行)将模型单层内的参数和计算分割到多个设备上。以矩阵乘法为例，可以将权重矩阵按行或列分割，每个设备负责一部分计算，然后通过通信合并结果。张量并行的优势是可以均匀分配计算负载，通信开销相对较小；劣势是实现复杂，需要针对不同算子设计分割策略，且通信模式复杂。

流水线并行（也称为层间并行）将模型的不同层分配到不同的设备上，形成计算流水线。数据依次流经各个设备，每个设备负责计算模型的一部分层。流水线并行的优势是实现相对简单，通信模式清晰；劣势是存在流水线气泡(Bubble)，即部分设备在某些时间步处于空闲状态，影响资源利用率。



模型并行面临的挑战及主要的优化技术

模型并行面临的主要挑战是通信开销和负载均衡。在张量并行中，每个前向和反向传播步骤都需要进行设备间的通信，通信频率高；在流水线并行中，虽然通信频率较低，但单次通信的数据量可能较大。负载均衡方面，不同层的计算复杂度和内存占用可能差异很大,如何合理分配模型各层到设备上,实现负载均衡,是一个复杂的问题。

为了优化模型并行的性能，业界提出了多种技术。通信计算重叠是一种常用技术，通过在计算进行的同时准备通信数据，隐藏通信延迟。流水线填充技术通过精心设计数据输入顺序，减少流水线气泡，提高设备利用率。动态负载均衡则根据实际运行时的性能数据，动态调整模型分配策略，实现更好的负载均衡。

在实际的大模型训练中，模型并行通常与数据并行结合使用。例如，可以将模型按层分割到多个设备组上（流水线并行），每个设备组内部再对单层进行张

量并行，同时每个设备内部还可以采用数据并行。这种多维并行的策略可以充分利用大规模计算集群的资源，实现高效的大模型训练。

4.2.3 混合并行与 4D 并行

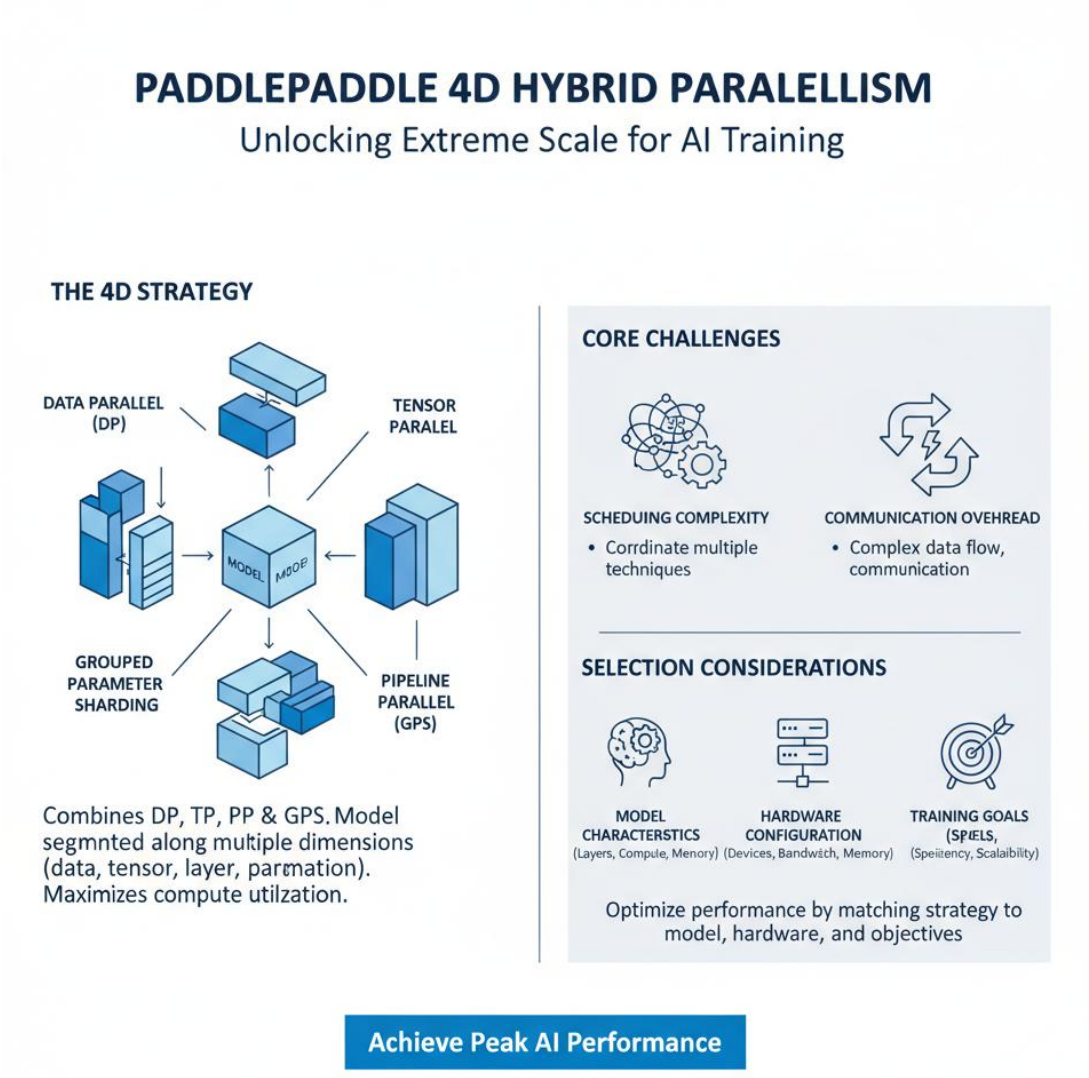
混合并行技术是指同时使用多种并行技术，比如数据并行和模型并行，或者数据并行和流水线并行，或者数据并行和张量并行。在大模型训练中，由于模型规模巨大、计算资源有限，单一并行技术往往无法满足需求，需要采用混合并行策略，充分发挥不同并行技术的优势。

混合并行技术的核心概念，以及 DP+PP 和 DP+TP 这两种常用策略

DP+PP（数据并行+流水线并行）是一种常用的混合并行策略。在这种策略中，模型被分割成多个阶段，每个阶段分配到一个设备组上，形成流水线并行；同时，每个设备组内部采用数据并行，处理不同的数据子集。这种策略可以同时利用数据并行的简单性和流水线并行的内存效率，适合中等规模的大模型训练。

DP+TP（数据并行+张量并行）是另一种常用的混合并行策略。在这种策略中，模型的每一层被分割到多个设备上，形成张量并行；同时，不同设备组之间采用数据并行，处理不同的数据子集。这种策略可以同时利用数据并行的扩展性

和张量并行的计算均衡性，适合计算密集型的大模型训练。



飞桨 4D 混合并行策略的复杂性及其带来的挑战

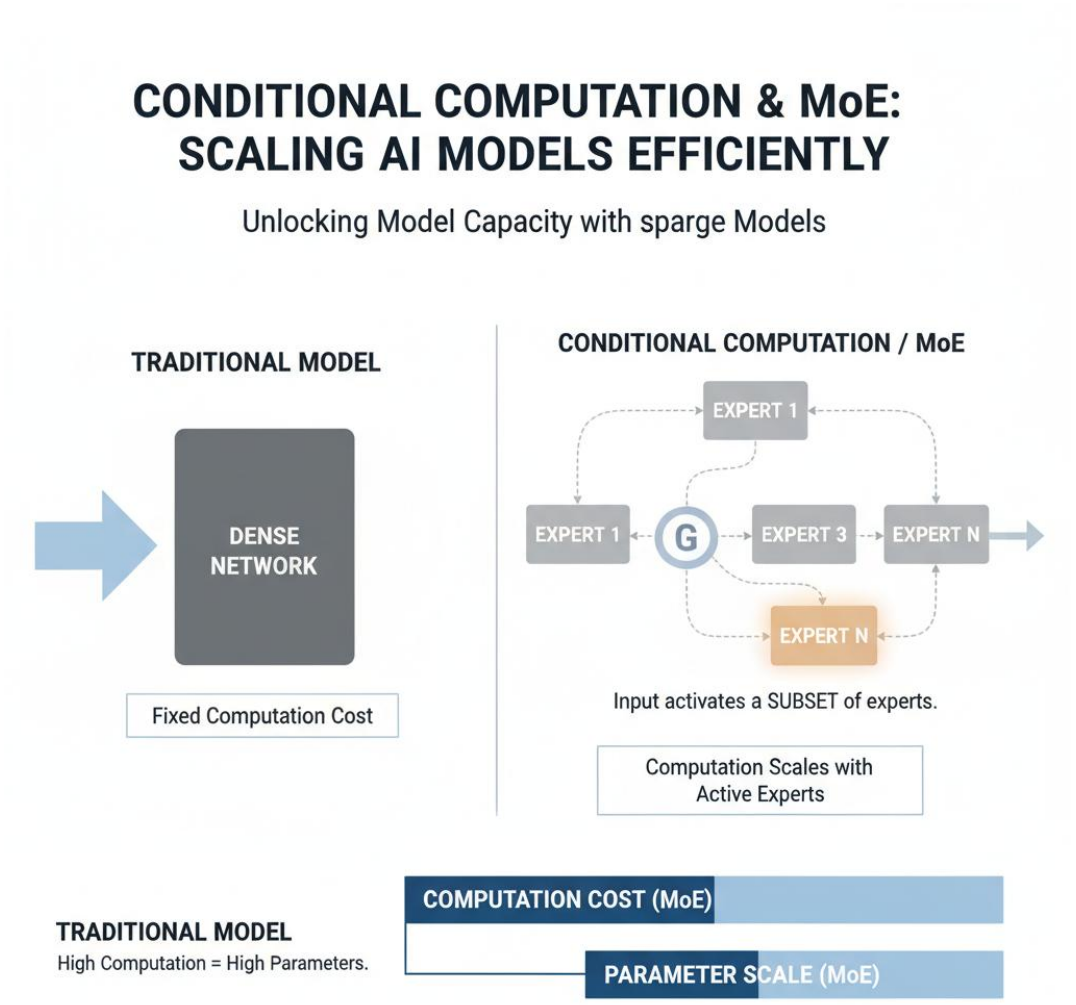
飞桨 4D 混合并行是一种更复杂的混合并行策略，结合了数据并行、张量并行、流水线并行和分组参数切片四种并行技术。在这种策略中，模型被同时沿多个维度进行分割：数据维度（数据并行）、张量维度（张量并行）、层维度（流水线并行）和参数维度（分组参数切片）。这种多维并行的策略可以最大化计算资源的利用率，适合超大规模的大模型训练。

混合并行的核心挑战是调度复杂性和通信开销。在混合并行中，需要协调多种并行技术的调度，确保计算和通信的高效进行。同时，多种并行技术的叠加会导致通信模式的复杂化，增加通信开销。为了应对这些挑战，混合并行系统通常需要精心设计的调度算法和通信优化技术。

在实际应用中，混合并行的选择需要考虑多个因素：模型特性（如层数、每

层计算量、内存占用等）、硬件配置（如设备数量、设备间带宽、内存容量等）、训练目标（如训练速度、内存效率、扩展性等）。通过合理选择和配置混合并行策略，可以实现大模型训练的最优性能。

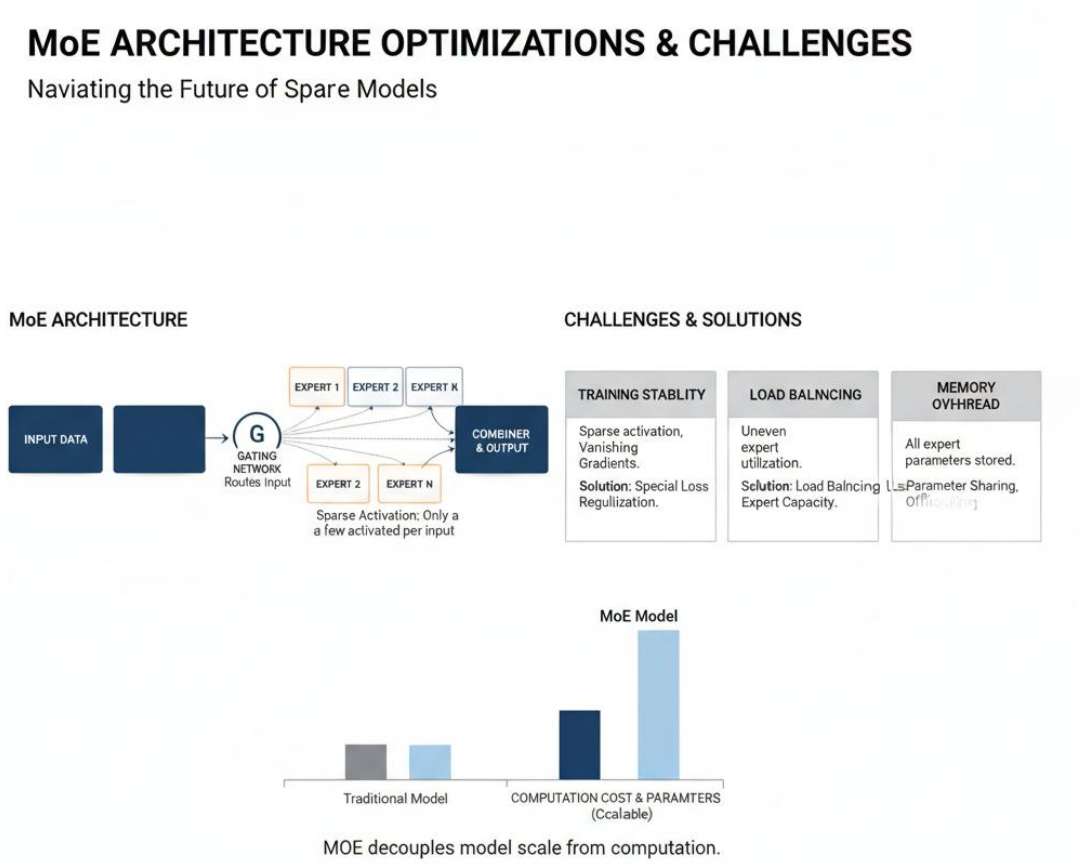
4.2.4 条件计算与 MoE



条件计算（Conditional Computation）和 MoE（Mixture of Experts）是大模型训练中的新兴技术，通过稀疏激活机制，在不增加计算负担的情况下扩展模型规模。条件计算的概念即仅在每个样本的基础上激活网络的不同部分，使得在不增加额外计算负担的情况下扩展模型规模成为可能。

稀疏激活（Sparse Activation）是指在神经网络中，某一层的激活值中只有一小部分是非零值，而大部分值为零或接近零。这种稀疏性可以减少计算量和内存需求，从而加速推理并降低能耗。稀疏激活通常出现在特定类型的神经网络或激活函数中，特别是在模型压缩和优化场景中。

MoE（Mixture of Experts）是一种实现条件计算的具体架构，通过门控网络动态选择少数专家进行计算。在 MoE 层中，输入数据被路由到多个"专家"网络中的少数几个，只有被选中的专家才会参与计算，其他专家则处于空闲状态。这种稀疏激活机制使得模型可以在不增加计算成本的情况下大幅增加参数规模，提高模型容量。



MoE 架构的核心组件包括专家网络、门控网络和路由机制。专家网络通常是前馈神经网络，负责具体的计算任务；门控网络负责根据输入数据决定激活哪些专家；路由机制则实现了数据到专家的分配。在实际实现中，MoE 层通常替换传统 Transformer 中的前馈网络层，形成稀疏激活的 Transformer 架构。

MoE 架构的优势在于可以实现模型规模和计算成本的解耦。通过增加专家数量，可以线性增加模型参数规模，而计算成本仅与被激活的专家数量相关，保持相对恒定。这种特性使得 MoE 模型在参数规模远超传统模型的情况下，仍能保持合理的训练和推理成本。

然而，MoE 架构也面临一些挑战。首先是训练稳定性问题，稀疏激活可能导致训练不稳定，需要特殊的训练技巧和正则化方法。其次是负载均衡问题，如果路由机制设计不当，可能导致某些专家过载而其他专家闲置，影响训练效率。

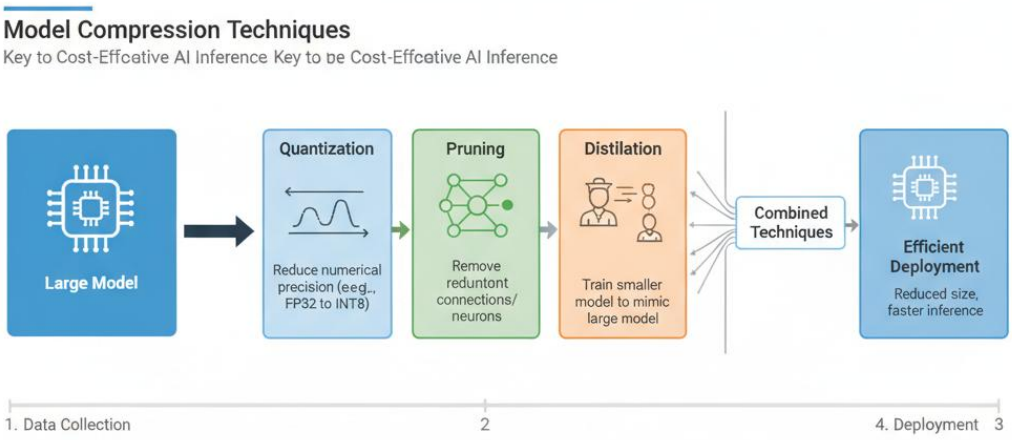
最后是内存开销问题，虽然计算是稀疏的，但所有专家参数都需要存储，内存占用仍然很大。

为了优化 MoE 架构的性能，业界提出了多种技术。负载均衡损失是一种常用的正则化方法，通过在损失函数中添加负载均衡项，鼓励门控网络均匀使用各个专家。专家容量限制则通过设置每个专家的最大处理批次大小，防止某些专家过载。通信优化技术则针对 MoE 特有的通信模式进行优化，减少专家间的数据传输开销。

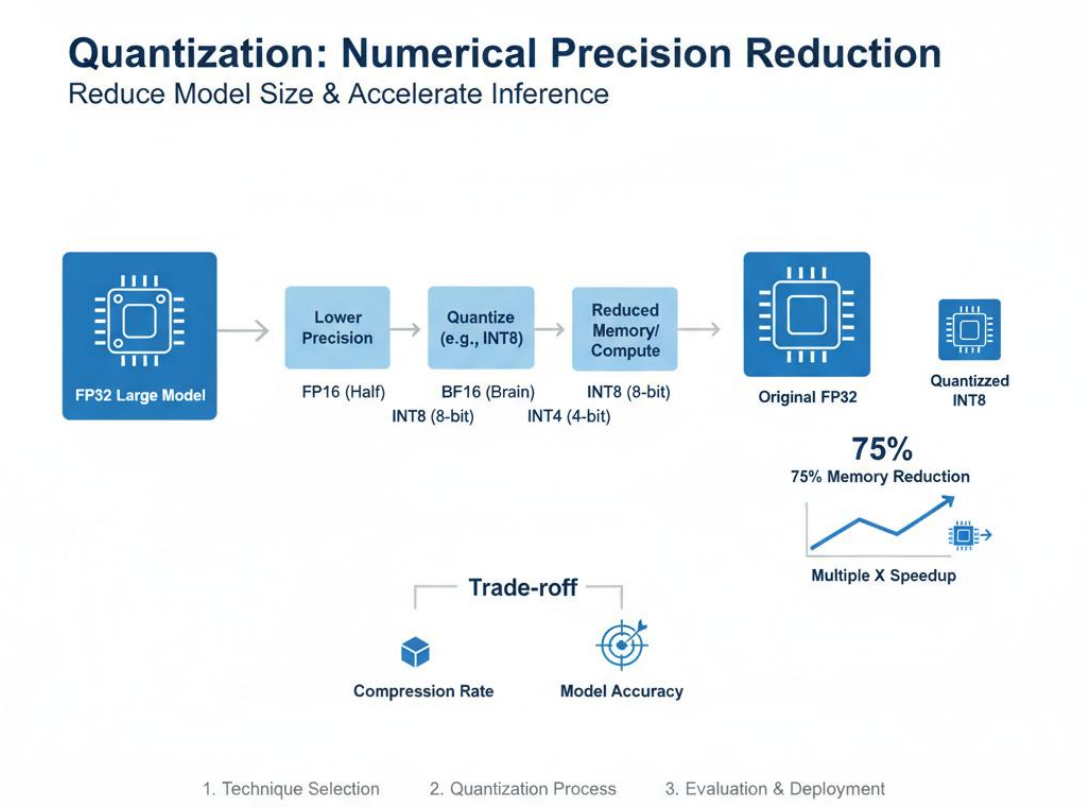
在实际应用中，MoE 架构已在大模型训练中取得显著成功。如 Google 的 Switch Transformer、Mixtral 8x7B 等模型都采用了 MoE 架构，在保持合理计算成本的同时实现了巨大的参数规模和优秀的性能表现。随着技术的不断发展，MoE 架构有望在大模型领域发挥更重要的作用。

4.3 推理加速与部署优化

4.3.1 模型压缩技术

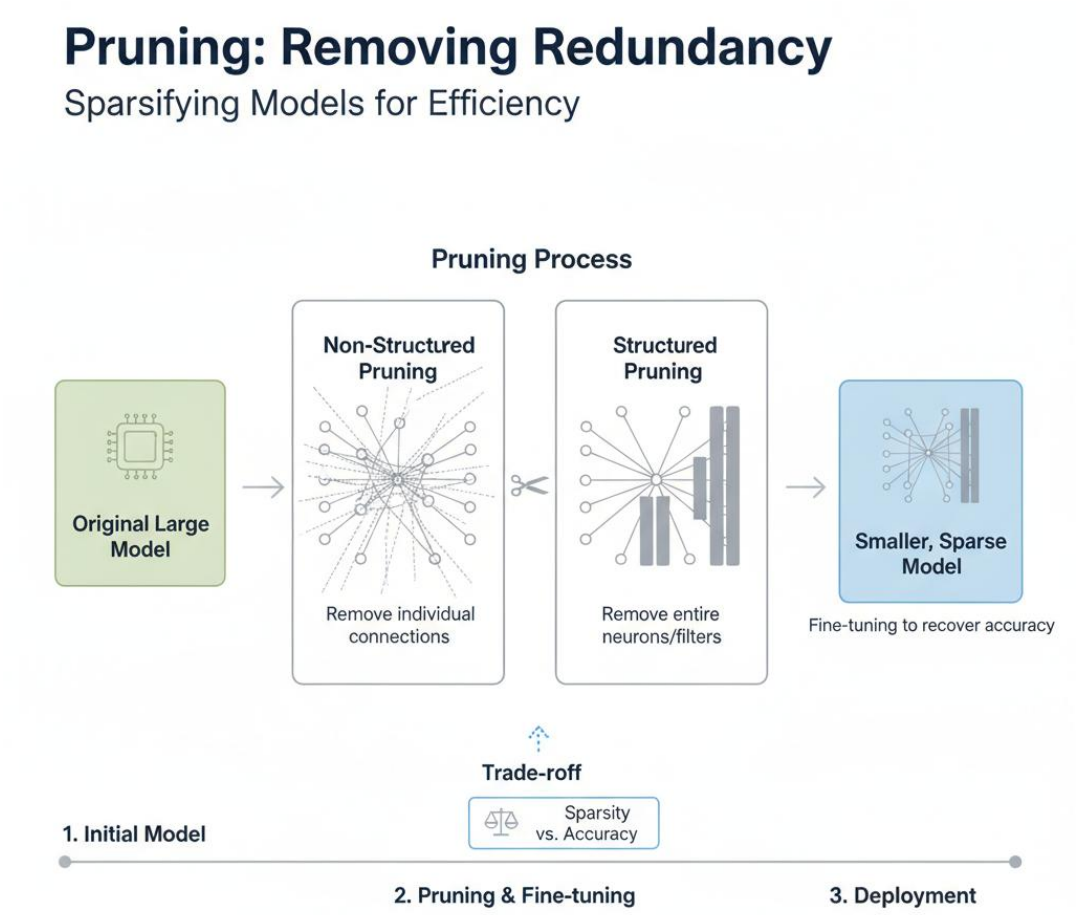


模型压缩技术是降低大模型推理成本的关键手段，通过减少模型参数量和计算复杂度，实现更高效的推理部署。主要的模型压缩技术包括量化、剪枝、蒸馏等，这些技术可以单独使用，也可以组合使用，形成综合的压缩方案。



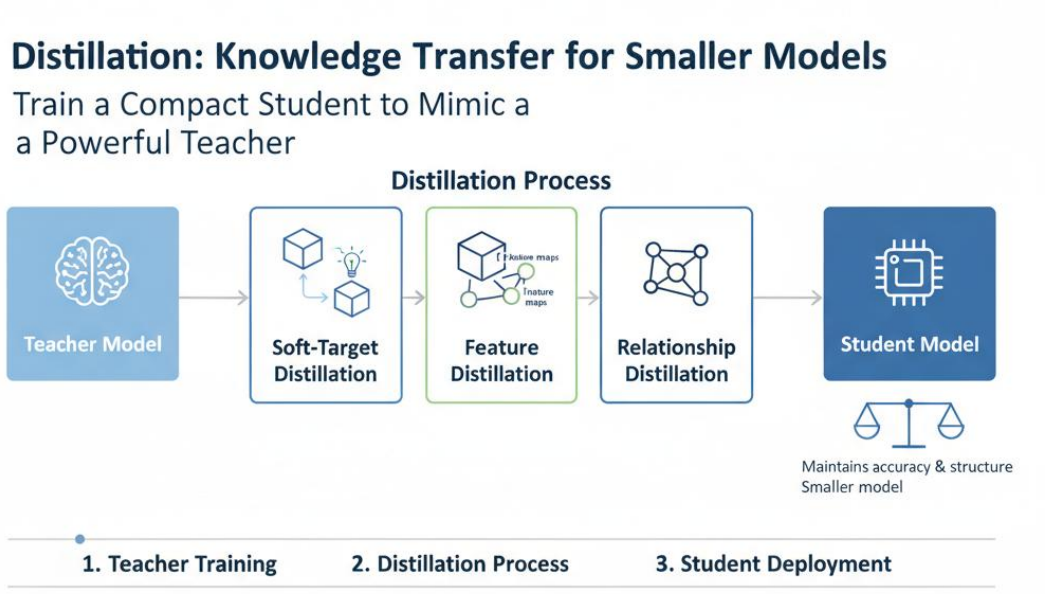
量化技术详解

量化（Quantization）是通过降低模型参数和激活值的数值精度来减少模型大小和计算量的技术。常见的量化方案包括 FP16（半精度浮点）、BF16（脑浮点）、INT8（8 位整数）、INT4（4 位整数）等。量化可以显著减少内存占用和计算量，同时利用现代 AI 硬件的低精度计算加速能力，提高推理速度。例如，将 FP32 模型量化为 INT8，可以减少 75%的内存占用，并在支持 INT8 计算的硬件上获得数倍的加速比。然而，过度量化可能导致模型精度下降，需要在压缩率和精度之间找到平衡。



剪枝技术详解

剪枝（Pruning）是通过移除模型中不重要的连接或神经元来减少模型参数量的技术。剪枝的核心思想是在尽量保持模型精度不受影响的前提下减少网络的参数量，例如减少网络中连接或神经元的数量。剪枝通常包括结构化剪枝和非结构化剪枝两种形式。结构化剪枝移除整个神经元或卷积核，保持模型结构规整，便于硬件加速；非结构化剪枝移除单个连接，压缩率更高但可能导致不规则的计算模式，难以直接加速。剪枝后的模型通常需要微调来恢复精度。



蒸馏技术详解

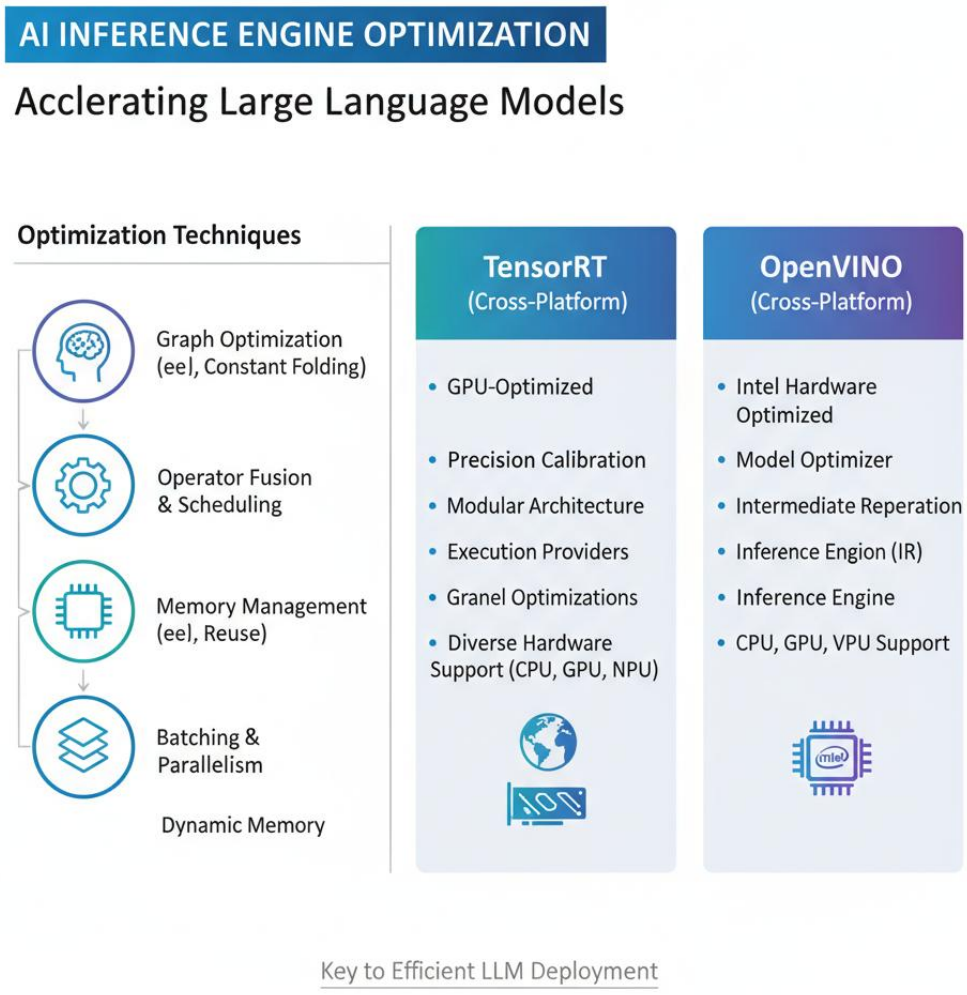
蒸馏（Distillation）是通过训练一个小模型（学生模型）来模仿大模型（教师模型）的行为，实现知识转移的技术。蒸馏不仅可以减少模型大小，还可以将多个大模型的知识集成到一个小模型中，提高小模型的性能。蒸馏通常包括软目标蒸馏、特征蒸馏、关系蒸馏等多种形式，分别针对模型的输出、中间特征、样本关系等进行知识转移。蒸馏的优势是可以保持模型结构规整，便于硬件加速；劣势是训练过程复杂，需要精心设计蒸馏策略。

除了上述主要技术外，还有一些其他的模型压缩方法，如二值化（将参数量化为 1 位）、低秩分解（将权重矩阵分解为多个小矩阵）、参数共享（多个参数共享相同值）等。这些技术通常与量化、剪枝、蒸馏等技术结合使用，形成综合的压缩方案。

在实际应用中，模型压缩技术的选择需要考虑多个因素：硬件特性（如支持的精度、计算能力等）、应用场景（如延迟要求、精度要求等）、模型特性（如结构、敏感度等）。通过合理选择和组合不同的压缩技术，可以在满足应用需求的前提下，最大化压缩效果，实现高效的大模型推理部署。

4.3.2 推理引擎优化

推理引擎优化是大模型推理加速的重要手段，通过图优化、算子调度、内存管理等技术，充分发挥硬件计算能力，实现高效的推理执行。主流的推理引擎包括 TensorRT、ONNX Runtime、OpenVINO 等，它们各自具有不同的特点和适用场景。



大模型推理引擎优化概述与主流引擎对比

TensorRT 是 NVIDIA 开发的高性能深度学习推理引擎，针对 NVIDIA GPU 进行了深度优化。TensorRT 的核心优化技术包括：精度校准（自动选择最佳精度）、层和张量融合（减少内存访问和 kernel 启动开销）、内核自动调整（针对特定 GPU 选择最优实现）、动态张量内存（最小化内存占用并重复使用内存）等。TensorRT 特别适合在 NVIDIA GPU 上部署大模型，可以显著提升推理速度和能效。

ONNX Runtime 是一个跨平台的开源推理引擎，支持多种硬件平台和 AI 框架。ONNX Runtime 的核心优势在于其模块化架构和可扩展性，通过执行提供程序（Execution Providers）机制支持不同的硬件后端，如 CPU、GPU、NPU 等。ONNX Runtime 还提供了丰富的图优化和内存管理功能，如常量折叠、死代码消除、内存规划等，可以在多种硬件平台上实现高效的推理执行。

OpenVINO 是 Intel 开发的开源推理工具包，针对 Intel 硬件（CPU、GPU、VPU 等）进行了优化。OpenVINO 的核心组件包括模型优化器（Model Optimizer）和推理引擎（Inference Engine）。模型优化器将训练好的模型转换为 OpenVINO 的中间表示（IR），进行图优化和精度校准；推理引擎则针对 Intel 硬件进行深度优化，实现高效的推理执行。OpenVINO 特别适合在 Intel 平台上部署大模型，可以充分利用硬件的加速能力。

推理引擎优化关键环节与大模型挑战

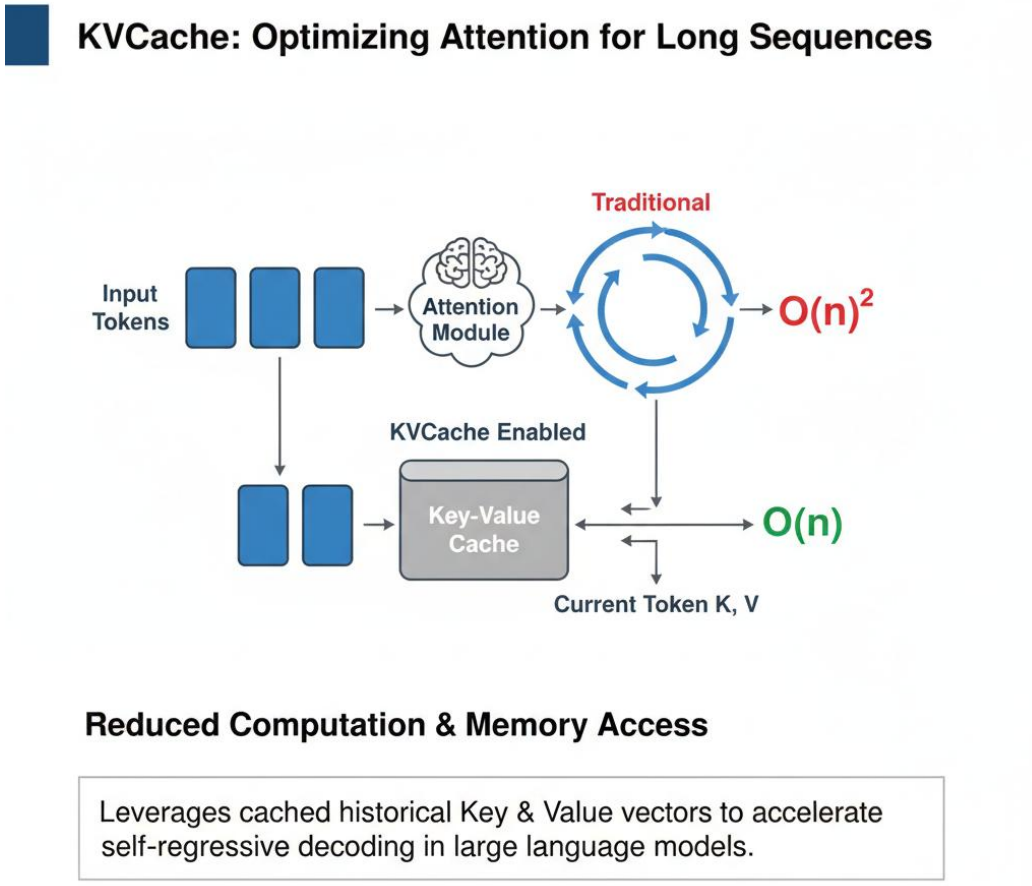
推理引擎优化通常包括以下几个关键环节：图优化、算子优化、内存管理、批处理优化等。图优化通过常量折叠、死代码消除、算子融合等技术，简化计算图结构，减少计算量；算子优化针对特定硬件实现高效的算子库，充分利用硬件的加速能力；内存管理通过内存复用、预分配等技术，减少内存分配和释放的开销；批处理优化则通过动态批处理、批处理大小自适应等技术，提高硬件利用率。

在大模型推理中，推理引擎优化面临一些特殊挑战。一是大模型的计算图规模庞大，图优化时间和内存消耗成为瓶颈；二是大模型的动态特性（如动态形状、

条件计算等）增加了优化难度；三是大模型的内存占用巨大，需要特殊的内存管理策略。为应对这些挑战，推理引擎技术正在向更高效、更智能、更灵活的方向发展，如增量编译、自适应批处理、内存分片等。

4.3.3 KVCache 与分离式推理

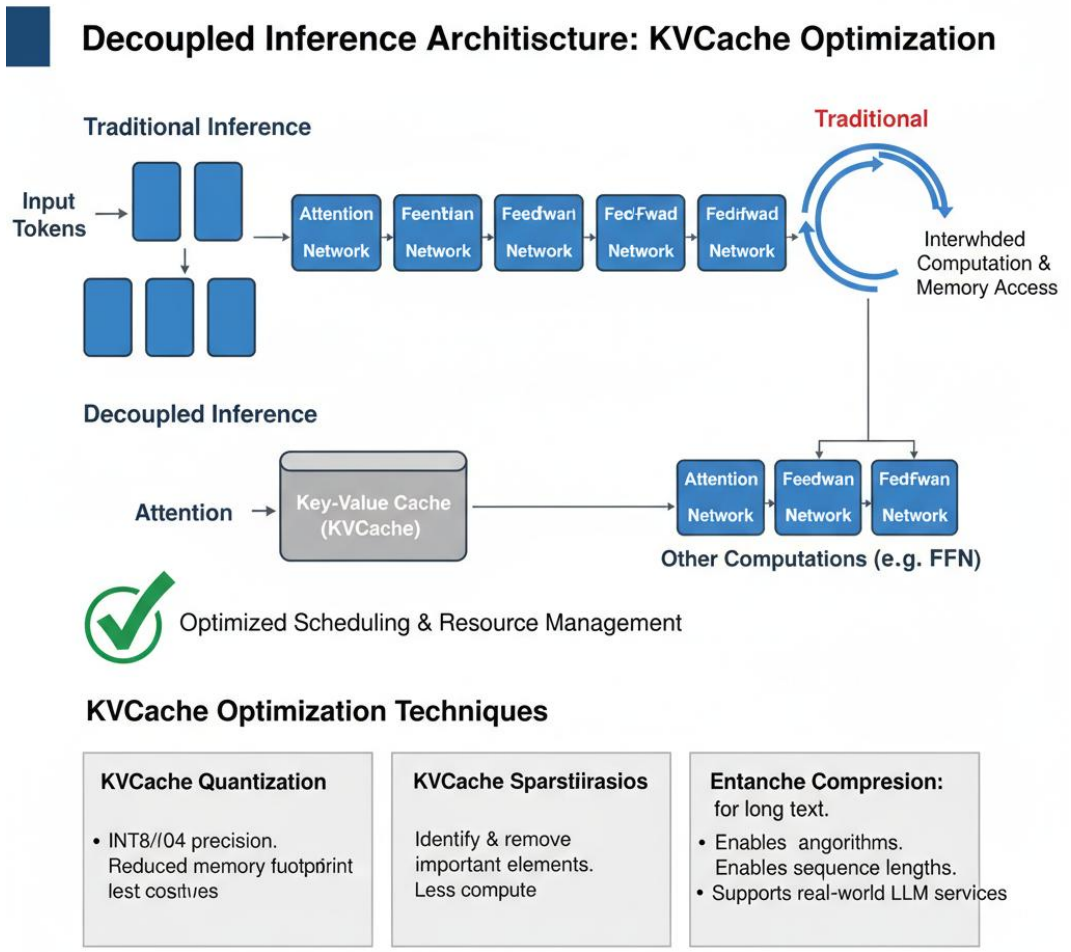
KVCache 与分离式推理是大模型推理优化的重要技术，通过优化注意力机制的计算和内存管理，显著提升长文本场景下的推理效率。这些技术特别适用于自回归生成任务，如文本生成、代码生成等场景。



KVCache 与注意力机制优化

KVCache（Key-Value Cache）是一种优化注意力计算的技术，通过缓存和复用历史计算的 Key 和 Value 向量，避免重复计算。在大模型的自回归生成过程中，每个新生成的 token 都需要与之前所有 token 进行注意力计算，如果不使用缓存，计算复杂度会随序列长度平方增长。KVCache 技术将历史计算的 Key 和 Value 向量存储在缓存中，新生成 token 时只需计算当前 token 的 Key 和 Value，然后与缓存中的历史向量进行注意力计算，将计算复杂度从 $O(n^2)$降低到 $O(n)$，其中 n 是序列长度。

分离式推理是一种以 KVCache 为中心的推理架构，将注意力机制的计算与其他计算分离，实现更高效的内存管理和计算调度。在传统的大模型推理中，注意力计算和其他计算（如前馈网络）是交织在一起的，难以独立优化。分离式推理将模型分为注意力模块和其他模块，分别进行优化和调度，可以更灵活地管理 KVCache，提高内存利用率和计算效率。



分离式推理架构与 KVCache 优化

KVCache 优化是分离式推理的核心环节，主要包括 KVCache 量化、KVCache 稀疏化、KVCache 压缩等技术。KVCache 量化通过将 KVCache 量化为低精度格式（如 INT8、INT4），减少内存占用和带宽需求；KVCache 稀疏化通过识别和移除 KVCache 中的不重要元素，减少存储和计算开销；KVCache 压缩则通过编码压缩等技术，进一步减少 KVCache 的内存占用。

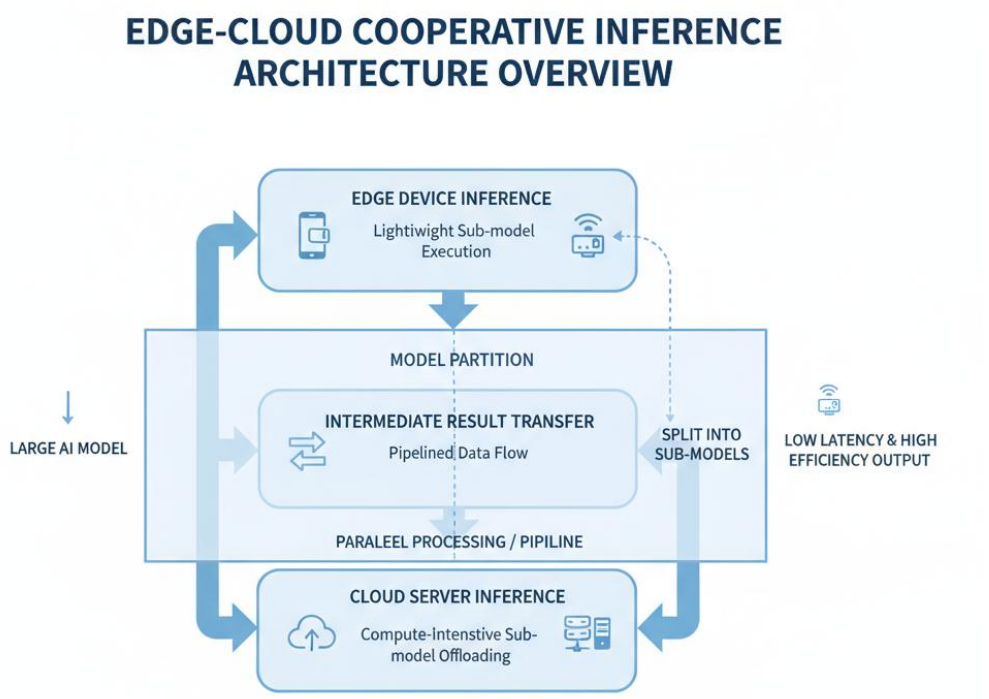
在实际应用中，KVCache 与分离式推理技术已在大模型服务中得到广泛应用。例如，清华大学的 KVCache.AI 项目针对长文本大模型推理进行了深度优化，

通过 KVCache 管理和分离式推理架构，显著提升了长文本场景下的推理效率。这些技术使得超长文本（如百万 token 级别）的大模型推理成为可能，为长文本应用场景提供了技术支撑。

KVCache 与分离式推理技术面临的挑战主要包括：一是 KVCache 的内存占用随序列长度线性增长，长序列场景下内存压力巨大；二是 KVCache 的管理和调度复杂，需要高效的内存分配和回收策略；三是分离式推理的实现需要对模型结构进行修改，增加了开发和维护的复杂性。为应对这些挑战，相关技术正在向更高效、更智能、更自动化的方向发展，如自适应 KVCache 管理、动态 KVCache 压缩、自动模型分割等。

4.3.4 边缘-云协同推理

边缘-云协同推理是一种分布式推理范式，通过将大模型分割为多个部分，分别部署在边缘设备和云端服务器上，实现低延迟、高效率的推理服务。这种技术特别适用于对实时性要求高、计算资源有限的边缘场景，如移动设备、物联网设备等。



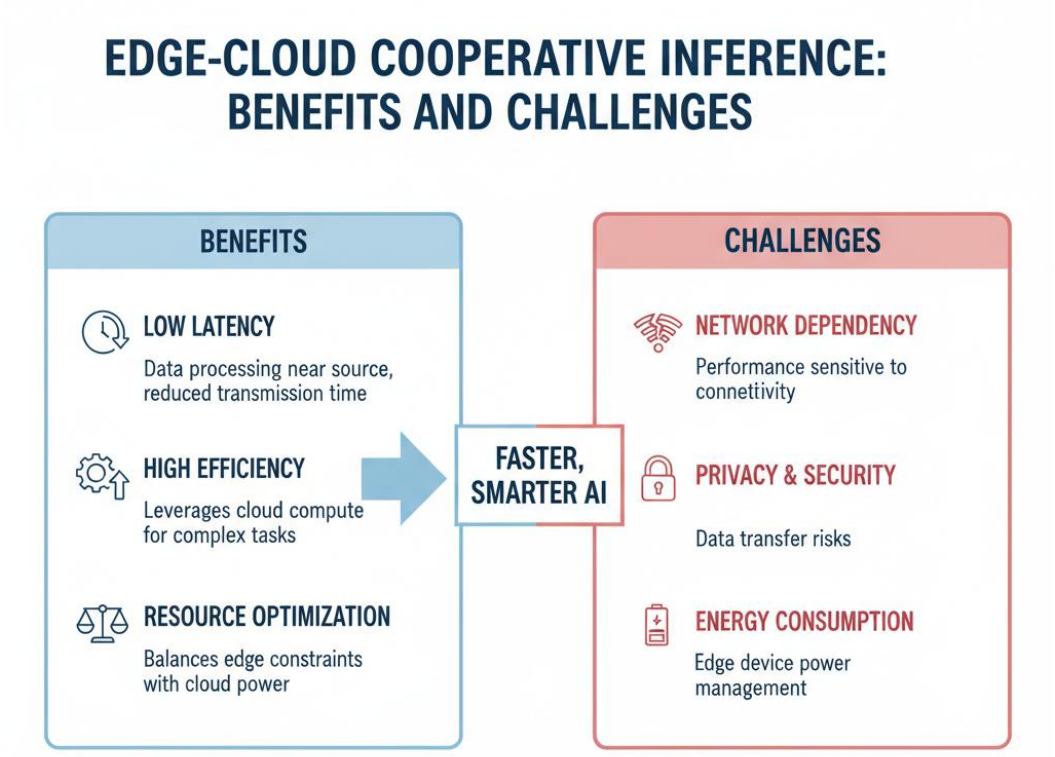
边缘-云协同推理架构概览

边缘-云协同推理的核心思想是模型分割，即将大模型分割为多个子模型，根据计算复杂度和延迟要求，将不同子模型分配到边缘设备和云端服务器上。通常，计算密集型和能耗密集型的子模型被卸载到云端服务器进行计算，而轻量级

的子模型则在边缘设备上执行。这种分割策略可以在保证推理质量的同时，满足边缘场景的低延迟要求。

边缘-云协同推理可以分为三个部分：边缘设备推理、中间结果传输、云服务器推理。这三部分可作为三个进程，在推理过程中并行处理。即云端在推理当前视频帧的同时，边缘设备可以推理下一帧，形成流水线式的处理流程，进一步提高整体效率。

边缘-云协同推理的优势在于可以充分利用边缘和云端的各自优势。边缘设备靠近数据源，可以提供低延迟的数据采集和预处理；云端服务器拥有强大的计算能力，可以处理复杂的计算任务。通过合理的模型分割和任务调度，边缘-云协同推理可以实现比纯边缘推理更高的性能，比纯云端推理更低的延迟。



边缘-云协同推理的优势与挑战

然而，边缘-云协同推理也面临一些挑战。首先是网络依赖性，边缘设备与云端之间的通信质量直接影响协同推理的性能，网络不稳定或带宽不足可能导致性能下降。其次是隐私安全问题，数据需要在边缘和云端之间传输，可能涉及隐私泄露风险。最后是能耗问题，边缘设备通常电池供电，需要考虑能耗优化。

为了优化边缘-云协同推理的性能，业界提出了多种技术。动态模型分割是一种关键技术，根据网络状况、设备状态、任务特性等因素，动态调整模型分割策略，实现最优的性能。自适应传输技术则根据网络条件动态调整数据传输策略，

如数据压缩、增量传输等，减少网络开销。隐私保护技术通过联邦学习、差分隐私、数据加密等手段，保护数据隐私和安全。

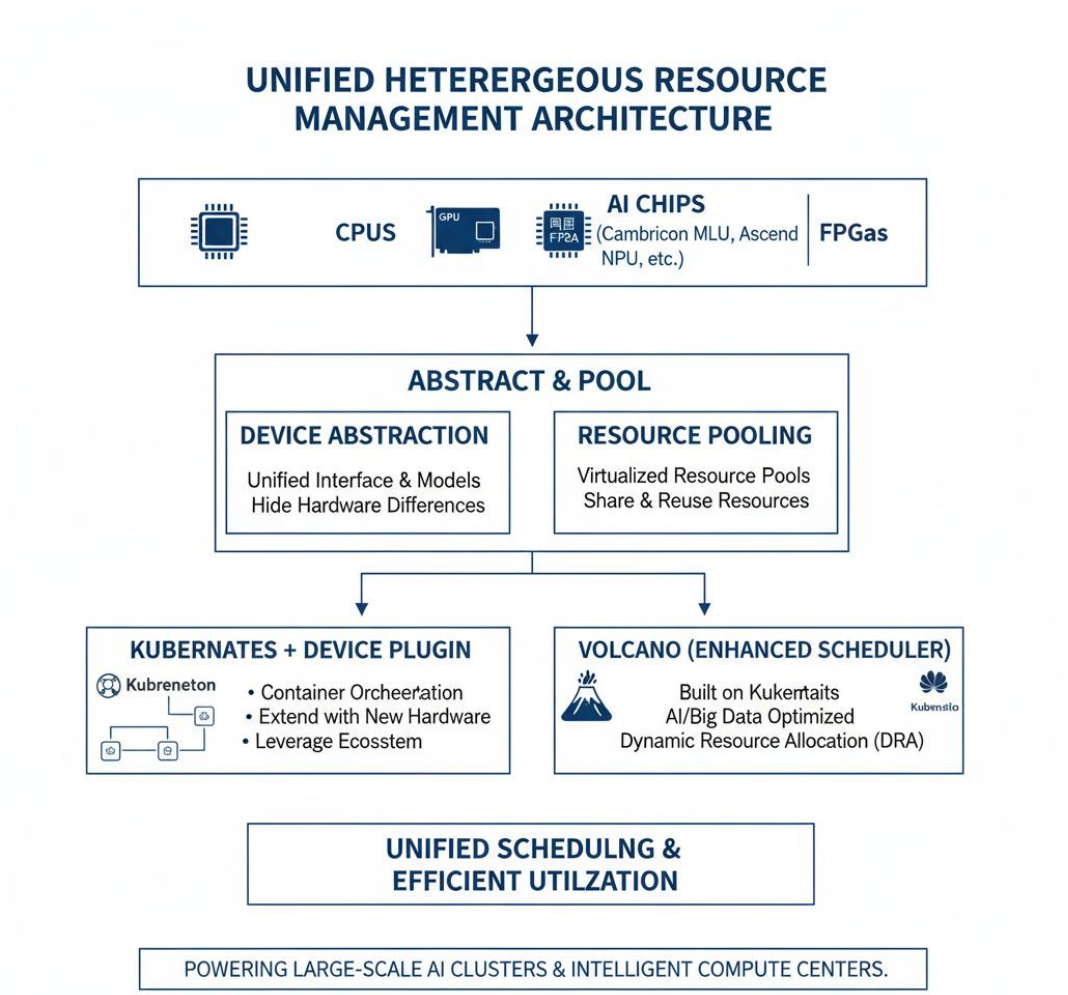
在实际应用中，边缘-云协同推理已在多个领域得到成功应用。例如，在智能监控场景中，边缘设备负责视频采集和预处理，云端服务器负责复杂的视频分析任务；在智能医疗场景中，边缘设备负责医学影像采集，云端服务器负责复杂的影像分析和诊断；在自动驾驶场景中，边缘设备负责实时感知和决策，云端服务器负责高精度地图更新和模型训练。

随着 5G、边缘计算、大模型等技术的发展，边缘-云协同推理将在更多场景中发挥重要作用，为 AI 技术的普及和应用提供新的技术路径。

4.4 异构资源调度与编排

4.4.1 资源统一管理

异构资源统一管理是构建高效异构算力系统的基础，通过抽象和池化不同类型的计算资源，实现资源的统一调度和高效利用。在 AI 大模型场景中，异构资源包括 CPU、GPU、国产 AI 芯片（如寒武纪 MLU、昇腾 NPU 等）、FPGA 等多种计算单元，如何实现这些资源的统一管理是一个重要挑战。



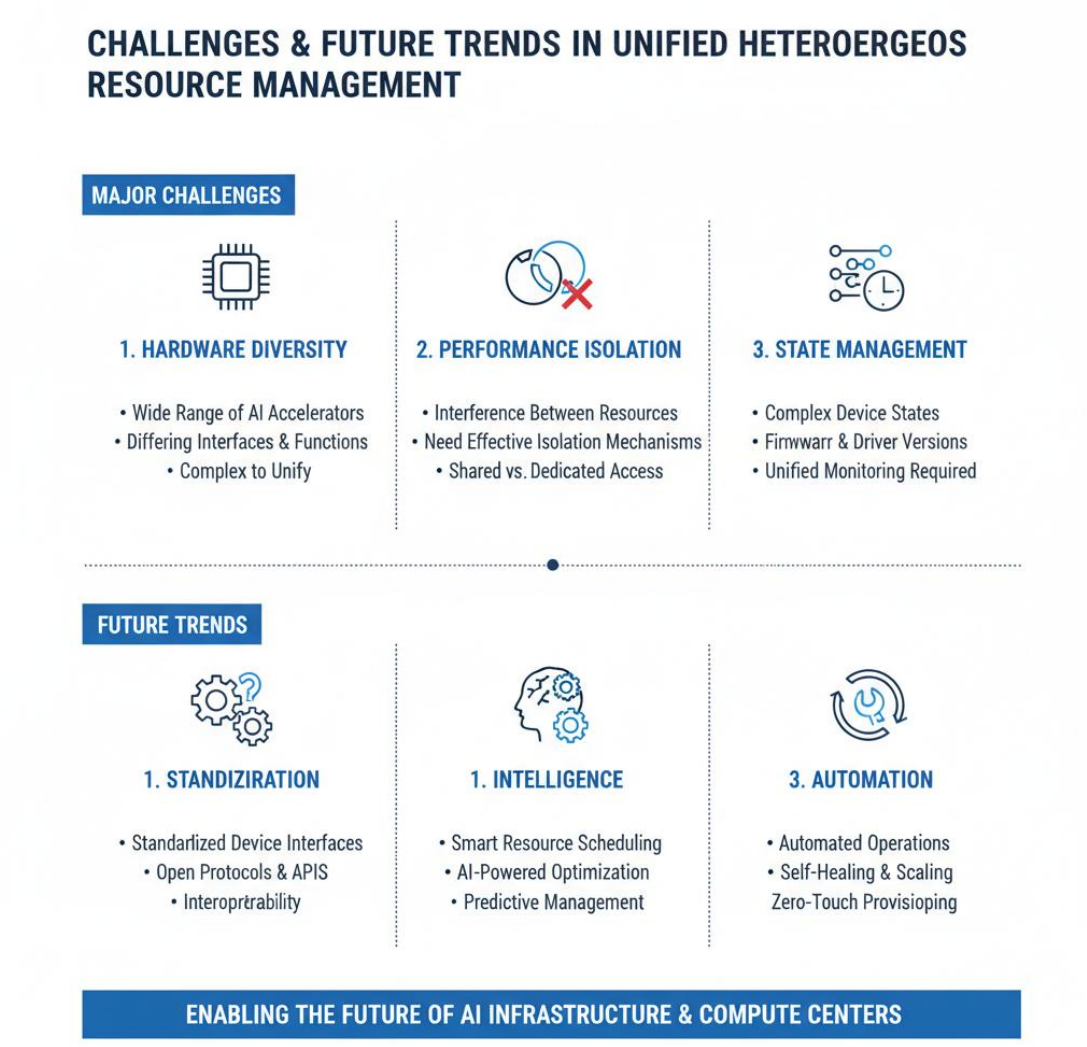
异构资源统一管理架构概览

Kubernetes+Device Plugin 是目前主流的异构资源管理方案。Kubernetes 作为容器编排平台，提供了强大的资源管理和调度能力；Device Plugin 机制则允许第三方设备厂商扩展 Kubernetes，支持新型硬件资源。通过 Device Plugin，各种 AI 加速器可以被抽象为 Kubernetes 的可调度资源，与 CPU、内存等资源一样进行管理和分配。这种方案的优势是可以利用 Kubernetes 成熟的生态和工具链，降低异构资源管理的复杂性。

Volcano 是面向 AI、大数据等高性能计算场景的增强型调度器，构建在 Kubernetes 之上，提供了更强大的异构资源管理能力。Volcano 提供了高性能任务调度引擎、高性能异构芯片管理、高性能任务运行管理等通用计算能力，通过接入 AI、大数据、基因、渲染等诸多行业计算框架服务终端用户。Volcano v1.12 增加了对 DRA（Dynamic Resource Allocation）的支持，允许集群动态分配和管

理外部资源，增强了与异构硬件的集成能力。

异构设备抽象与池化是资源统一管理的核心技术。异构设备抽象通过统一的接口和描述模型，将不同类型的硬件资源抽象为标准化的资源对象，隐藏硬件差异，简化上层应用的开发。异构设备池化则将分散的硬件资源汇聚成虚拟的资源池，实现资源的共享和复用，提高资源利用率。在实际实现中，通常采用分层抽象的策略，底层针对特定硬件提供专用驱动，中层提供统一的资源抽象，上层提供标准化的 API 接口。



异构资源统一管理面临的挑战与发展趋势

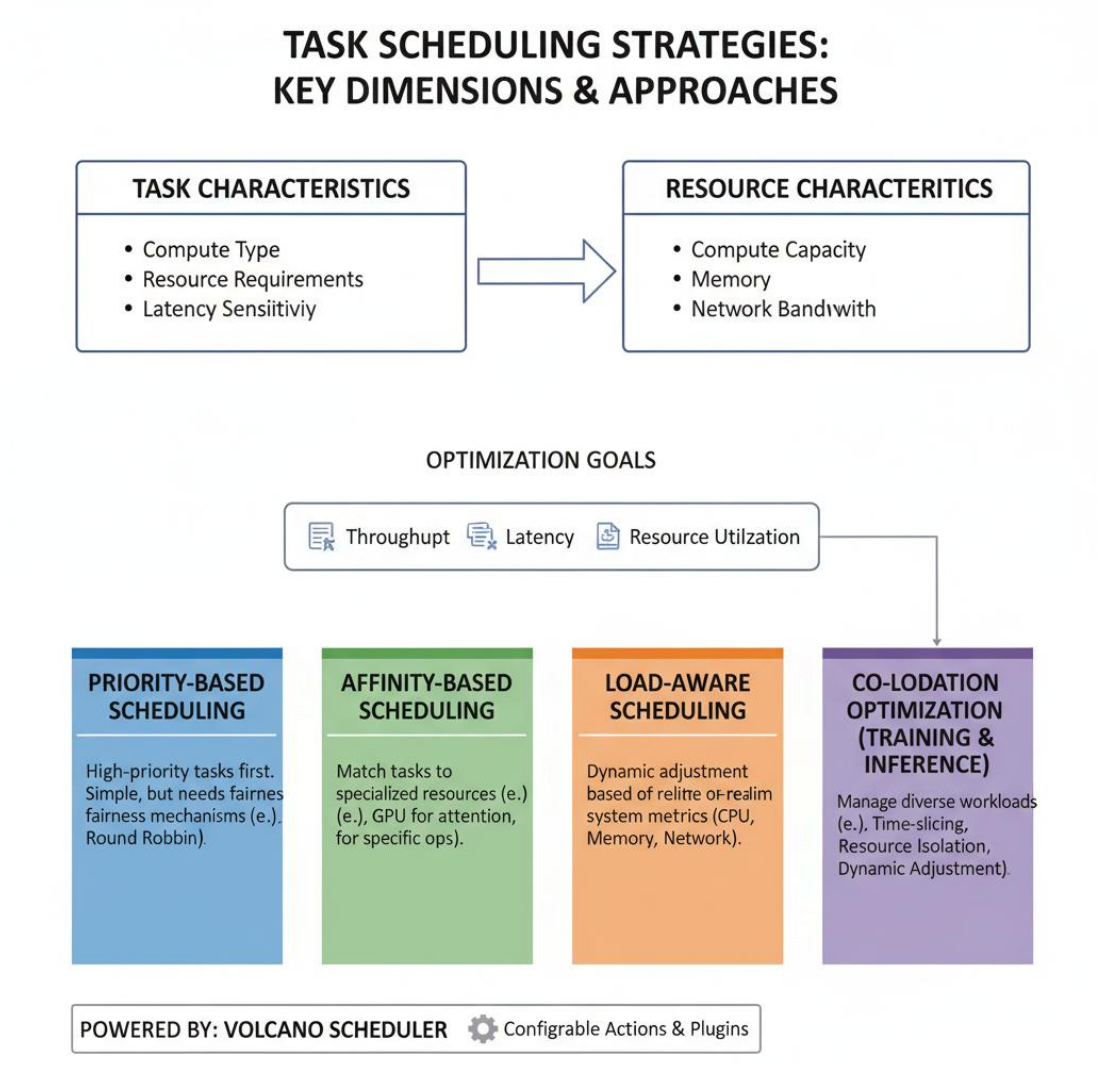
资源统一管理面临的挑战主要包括：一是硬件多样性，不同厂商、不同类型的 AI 加速器在接口、功能、性能等方面差异巨大，统一抽象难度高；二是性能隔离，不同类型的资源可能存在性能干扰，需要有效的隔离机制；三是状态管理，异构设备通常有复杂的状态（如固件版本、驱动版本等），需要统一的状态管理

机制。为应对这些挑战，资源统一管理技术正在向更标准化、更智能化、更自动化的方向发展，如标准化设备接口、智能资源调度、自动化运维等。

在实际应用中，异构资源统一管理已在大规模 AI 集群中得到广泛应用。例如，在智算中心中，通过 Kubernetes+Volcano 的架构，实现了对 CPU、GPU、国产 AI 芯片等多种资源的统一管理和调度，为大模型训练和推理提供了高效的算力支撑。随着异构算力的普及，资源统一管理技术将在 AI 基础设施中发挥越来越重要的作用。

4.4.2 任务调度策略

任务调度策略是异构资源管理的核心环节，通过合理的任务分配和资源调度，实现系统性能的最优化。在大模型与异构算力融合场景中，任务调度需要考虑多种因素，如任务特性、资源特性、网络状况等，是一个复杂的优化问题。

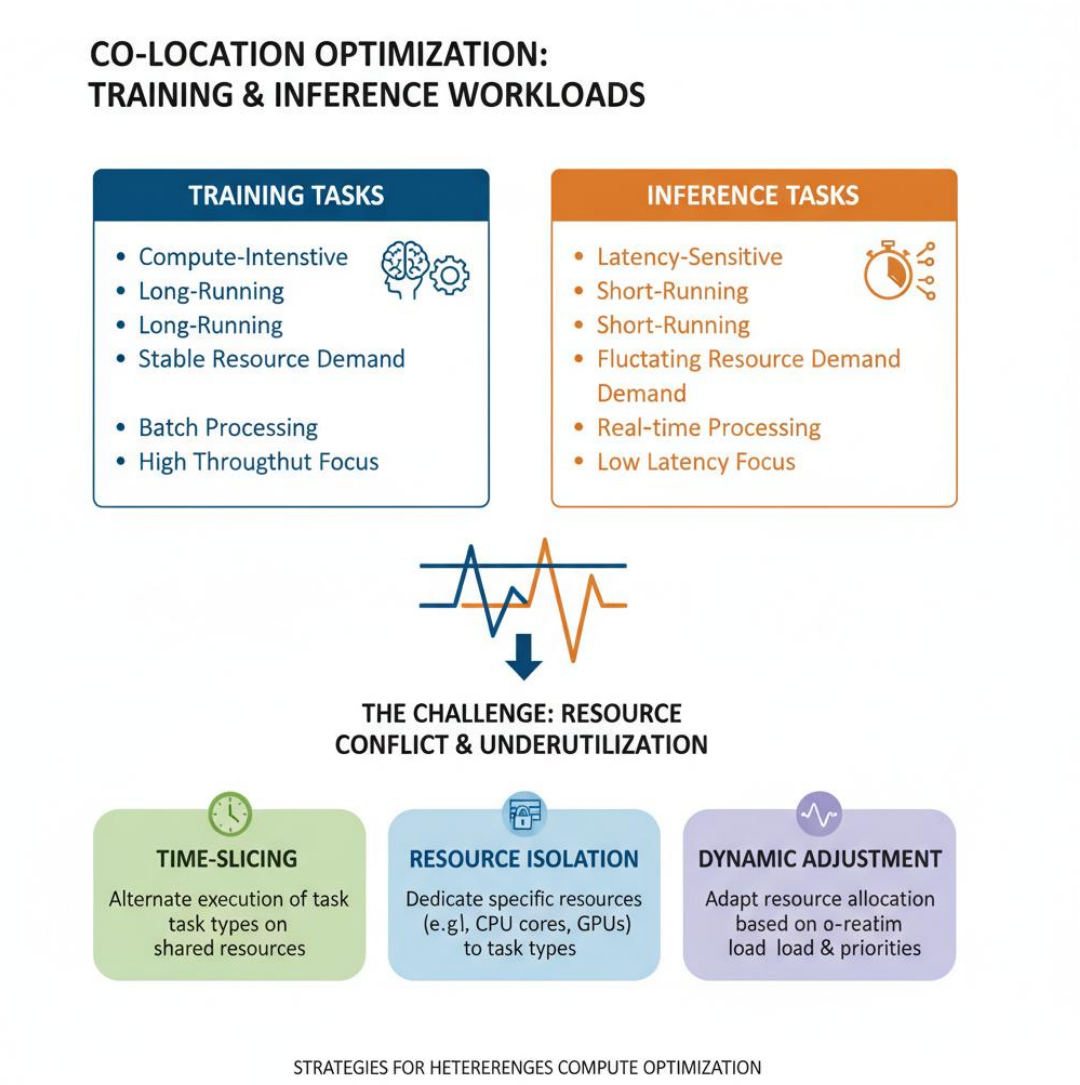


任务调度策略的维度

基于优先级的调度是最常用的调度策略之一。根据任务的重要性和紧急程度，为任务分配不同的优先级，高优先级任务优先获得资源。这种策略简单直观，适合有明显重要性差异的任务场景。然而，简单的优先级调度可能导致低优先级任务饥饿，需要结合其他机制（如时间片轮转、优先级衰减等）来保证公平性。

资源亲和性调度是另一种重要的调度策略，根据任务与资源之间的亲和关系，将任务分配到最适合的资源上。在大模型场景中，不同的模型层或算子可能对不同类型的硬件有不同的亲和性，如注意力计算适合在 GPU 上执行，而某些特定的算子可能在专用 AI 芯片上更高效。资源亲和性调度可以充分利用硬件特性，提高任务执行效率。

负载感知调度是一种动态调度策略，根据系统的实时负载情况，动态调整任务分配策略。负载感知调度需要监控系统的各项指标，如 CPU 利用率、内存使用量、网络带宽、设备温度等，基于这些信息做出调度决策。这种策略可以适应系统负载的动态变化，实现更均衡的资源利用。



训练与推理任务混部优化

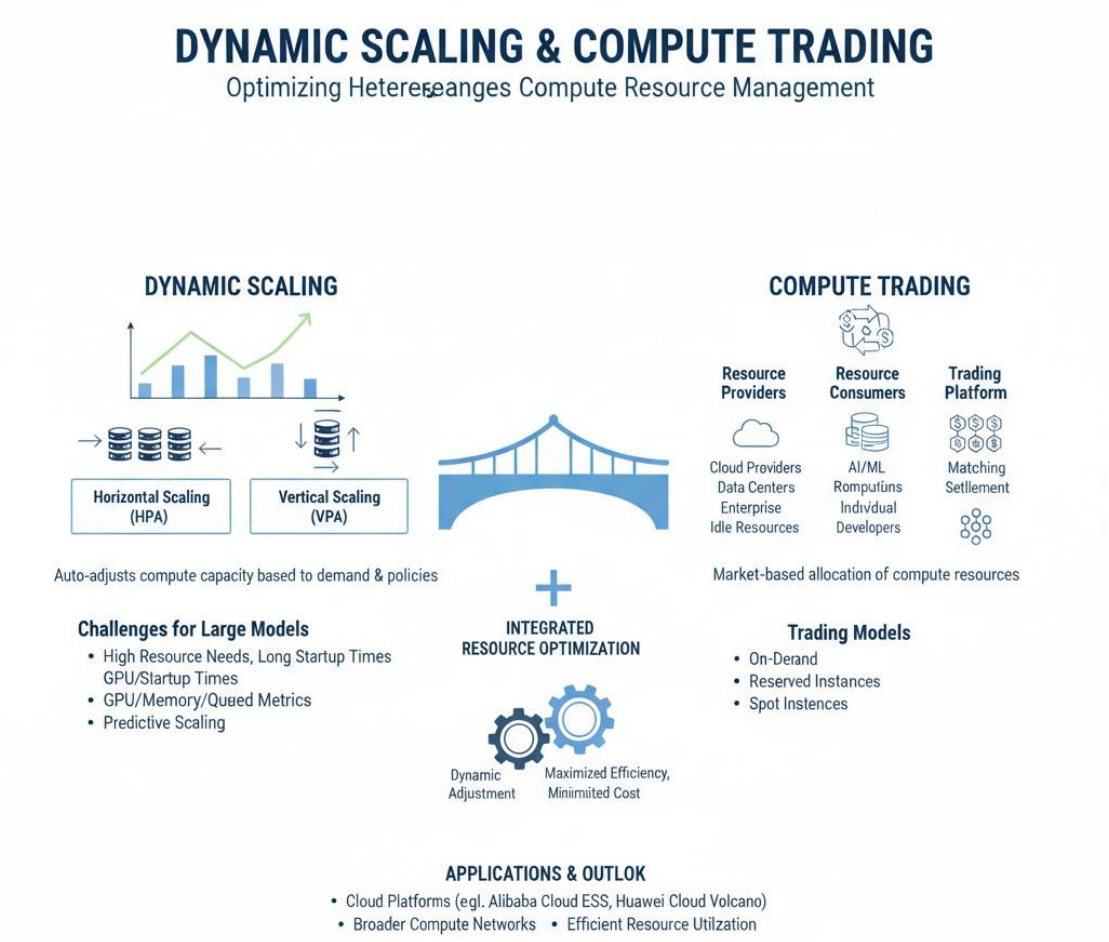
训练与推理任务混部优化是异构算力调度中的特殊挑战。训练任务通常计算密集、长时间运行、资源需求稳定；推理任务则通常延迟敏感、短时间运行、资源需求波动大。如何将这两种不同特性的任务合理混部，提高资源利用率，是一个复杂的问题。常见的策略包括时间分片（不同时间段运行不同类型任务）、资源隔离（为不同类型任务分配专用资源）、动态调整（根据负载情况动态调整资源分配）等。

Volcano 调度器提供了丰富的任务调度策略支持。Volcano Scheduler 由一系列 action 和 plugin 组成，action 定义了调度各环节中需要执行的动作；plugin 根据不同场景提供了 action 中算法的具体实现细节。Volcano 支持节点负载感知调度与重调度，支持多样化的监控系统，可以根据实际需求配置不同的调度策略。

在实际应用中，任务调度策略的选择需要考虑多个因素：任务特性（如计算

类型、资源需求、延迟要求等）、资源特性（如计算能力、内存容量、网络带宽等）、系统目标（如吞吐量、延迟、资源利用率等）。通过合理选择和配置任务调度策略，可以实现异构算力系统的高效运行，为大模型训练和推理提供强大的算力支撑。

4.4.3 弹性伸缩与算力交易



弹性伸缩与算力交易是异构算力资源管理的高级特性，通过动态调整资源供给和实现算力的市场化交易，提高资源利用效率，降低使用成本。这些技术在大模型与异构算力融合场景中具有重要意义，可以帮助用户更灵活、更经济地使用算力资源。

弹性伸缩是指根据业务需求和策略自动调整计算能力的服务。在 Kubernetes 环境中，弹性伸缩主要包括水平伸缩（HPA，Horizontal Pod Autoscaler）和垂直伸缩（VPA，Vertical Pod Autoscaler）两种形式。HPA 主要通过增加或减少 Pod 数量来实现伸缩，适合无状态服务的扩展；VPA 则通过调整 Pod 的资源请求和限制来实现伸缩，适合需要调整资源配额的场景。HPA 伸缩算法相对保守，如

果某个 Pod 获取不到资源指标或者资源没有准备好的情况下，在进行扩容操作时，该 Pod 的资源指标均不会加入计算，确保伸缩的稳定性。

弹性伸缩在大模型场景中面临特殊挑战。大模型任务通常资源需求大、启动时间长，传统的基于 CPU 利用率的伸缩策略可能不够准确。针对这些特点，大模型弹性伸缩需要考虑更多因素，如 GPU 利用率、内存使用量、队列长度等，并结合预测性伸缩（Predictive Scaling）技术，提前预判资源需求，避免资源不足导致的性能下降。

算力交易是算力资源的市场化配置方式，通过将算力资源商品化，实现供需双方的高效匹配。算力交易可以采用多种形式，如按需付费（Pay-as-you-go）、预留实例（Reserved Instances）、竞价实例（Spot Instances）等。按需付费适合短期、不确定的资源需求；预留实例适合长期、稳定的资源需求，可以享受折扣价格；竞价实例则适合可中断、弹性大的任务，价格更低但可能被中断。

算力交易市场通常包括资源提供方、资源需求方、交易平台等参与者。资源提供方包括云服务商、算力中心、企业闲置资源等；资源需求方包括 AI 企业、研究机构、个人开发者等；交易平台则提供资源发布、匹配、交易、结算等功能。区块链技术可以用于构建去中心化的算力交易平台，通过智能合约实现自动化的交易执行和结算，提高交易的透明度和可信度。

弹性伸缩与算力交易的结合可以形成更智能的算力资源管理体系。通过弹性伸缩技术，可以根据实际需求动态调整资源规模；通过算力交易技术，可以在不同资源提供商之间选择最优的资源组合。这种结合可以实现算力资源的最优配置，在满足性能需求的同时，最小化使用成本。

在实际应用中，弹性伸缩与算力交易已在多个云平台和算力网络中得到实现。例如，阿里云弹性伸缩服务（ESS）支持根据业务需求和策略自动调整计算能力，支持 ECS 实例和 ECI 实例；华为云 Volcano 调度器支持多种弹性伸缩策略，可以适应不同类型的 AI 工作负载。随着算力网络的不断发展，弹性伸缩与算力交易技术将在更广泛的场景中发挥作用，推动算力资源的高效利用和市场化配置。

五、国内企业实践与案例分析

5.1 华为昇腾：异构算力与大模型融合实践

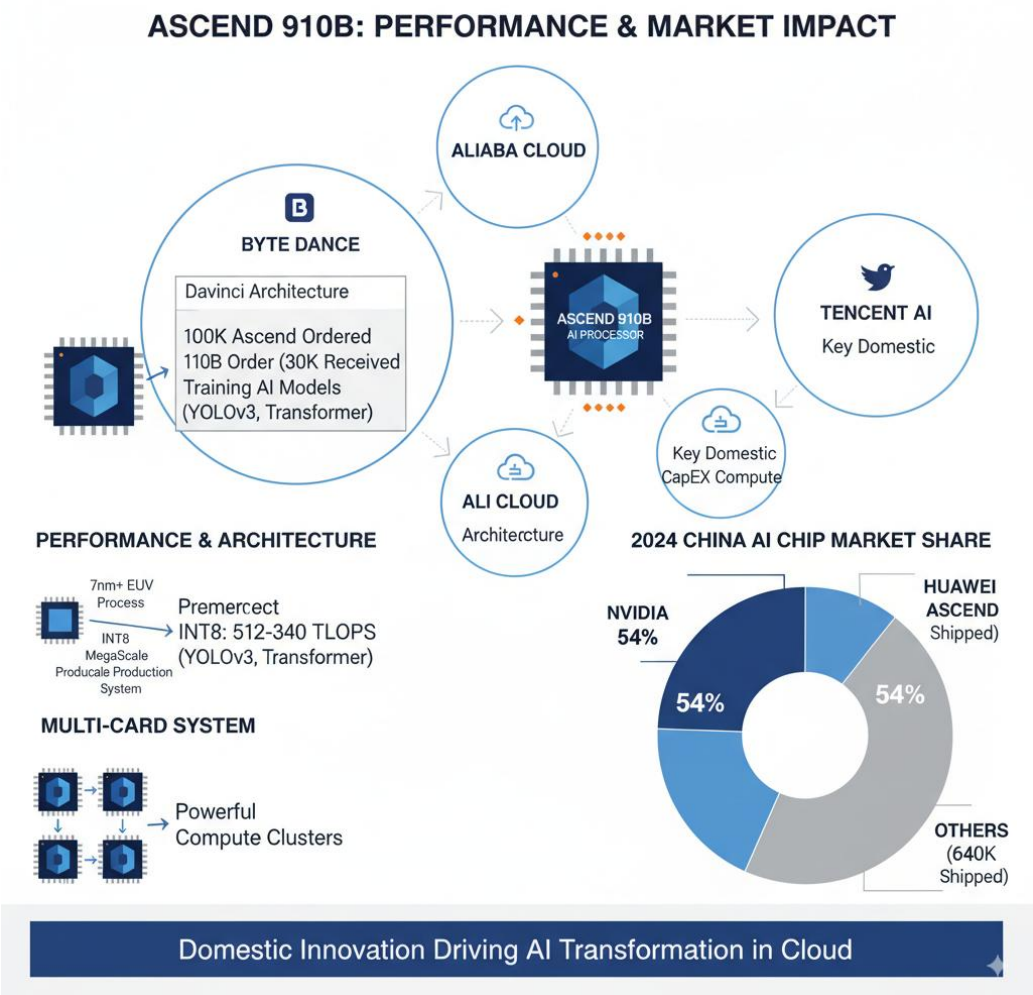
5.1.1 云端芯片在互联网大厂部署

华为昇腾作为国内 AI 芯片领域的领军企业，其昇腾 910 系列芯片已在多家

互联网大厂实现规模化部署，展现出国产 AI 芯片在大模型场景下的实用价值。昇腾 910B 是华为面向云端训练的高性能 AI 处理器，采用 7nm+EUV 工艺制造，拥有 32 核自研达芬奇架构，其半精度（FP16）算力达到 256-320 TFLOPS，整数精度（INT8）算力达到 512-640 TOPS，功耗 310W，被视为业界算力最强的 AI 处理器之一。

在字节跳动，华为昇腾芯片已成为大模型训练的重要算力支撑。据最新消息，字节跳动已向华为订购了多达 10 万颗昇腾 910B 芯片。昇腾 910B 的性能、能效都优于 NVIDIA A100，字节跳动计划使用昇腾 910B 芯片来训练新的 AI 模型。华为与字节跳动的合作不仅限于硬件供应，还包括软件栈的深度适配和优化，确保在大规模生产环境中的稳定运行。

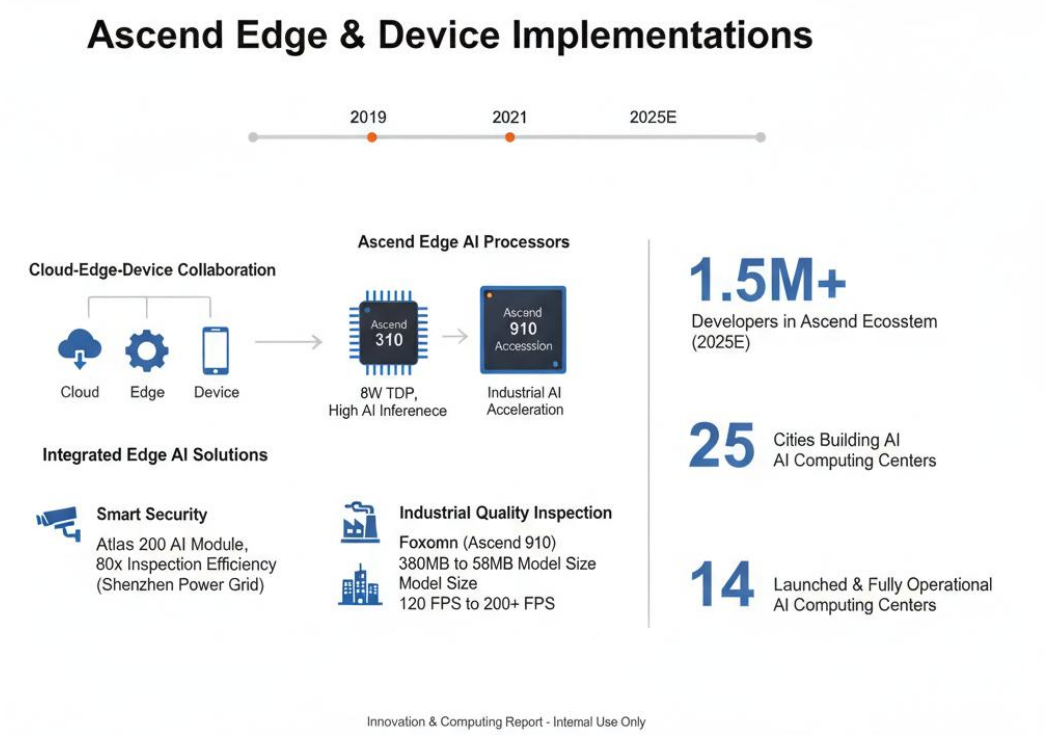
腾讯与字节跳动的 AI 资本开支也将显著增长，昇腾芯片在其中扮演重要角色。在业界应用广泛的 YOLOv3、Transformer 等训练任务中，多卡昇腾 910B 计算系统展现出优异的性能表现。华为为多卡系统专门设计了 HCCS 互连技术，可实现高速互联，形成强大的计算集群，满足大规模 AI 训练和推理需求。



华为昇腾与互联网大厂的合作模式不仅限于硬件供应，还包括联合研发、场景适配、性能优化等多个层面。通过深度合作，华为不断优化产品设计和软件生态，而互联网大厂则获得了更加适合自身业务需求的 AI 算力解决方案。这种互利共赢的合作模式，推动了国产 AI 芯片在实际应用中的快速迭代和成熟。据市场数据显示，2024 年华为昇腾出货 64 万片，在国内 AI 芯片市场占据 23%～28%的份额，排名第二，仅次于英伟达。

5.1.2 边缘与端侧落地案例

除云端部署外，华为昇腾芯片在边缘和端侧场景也有广泛应用。昇腾 310 系列是华为面向边缘计算场景的 AI 处理器，采用华为自研的达芬奇架构，在功耗仅为 6.5W 的条件下，提供强大的 AI 推理能力。Atlas 200 AI 加速模块集成了昇腾 310 处理器，可在边端侧实现目标识别、图像分类等 AI 应用加速，广泛用于智能边缘设备、机器人、无人机、智能工控等边端侧 AI 场景。



在智能安防领域，华为昇腾 Atlas 200 AI 加速模块被广泛应用于各类智能摄像头中。南方电网深圳供电局与华为携手，在边缘侧部署输电视频监控终端，集成 Atlas 200 AI 加速模块，运行 AI 推理算法进行就地图像视频分析，使巡检效率提升了 80 倍。通过在边缘设备本地完成 AI 计算，不仅减少了数据传输延迟，也保护了用户隐私，同时降低了对网络带宽的依赖。在实际部署中，搭载昇腾 310 的智能摄像头能够在复杂环境下稳定运行，满足 7×24 小时不间断工作的需

求。

在工业质检场景，华为昇腾边缘芯片与机器视觉技术结合，实现了产品质量的自动检测。富士康采用华为昇腾 910 芯片+动态量化方案，显著提升了检测效率。通过将 AI 推理能力下沉到生产现场，可以实时发现产品缺陷，及时调整生产工艺，提高产品质量和生产效率。昇腾边缘芯片的低功耗特性，使其能够直接集成到工业设备中，无需额外的散热和供电设施，大大简化了部署复杂度。

在智慧城市领域，华为昇腾边缘计算解决方案已在全国多个城市落地。昇腾 AI 边缘智能已经广泛应用到工业质检、高速收费稽核、智慧营业厅等场景，极大地加速了行业智能升级。例如，在高速收费稽核场景，昇腾边缘设备能够实时分析车辆信息，自动识别违规行为；在智慧营业厅，昇腾边缘设备能够提供智能客服、人脸识别等服务，提升用户体验。

华为昇腾边缘与端侧产品与云端产品形成了完整的算力梯度，支持从云端到边缘再到终端的全场景 AI 计算需求。这种云边端一体化的产品布局，使得用户可以根据实际需求选择最适合的产品形态，构建灵活高效的 AI 计算系统。截至 2025 年，已有超过 100 万的开发者加入昇腾生态，有 25 个城市基于昇腾构建人工智能计算中心，其中 14 个已经上线并饱和运营。

5.2 国内企业布局

5.2.1 寒武纪

寒武纪作为国内 AI 芯片的重要企业，其思元系列芯片在性能和技术创新方面表现突出。寒武纪成立于 2016 年，专注于人工智能芯片产品的研发与技术创新，致力于打造人工智能领域的核心处理器芯片。2023 年，寒武纪发布最新一代云端高算力芯片产品——思元 590 芯片，该芯片方便兼容主流 AI 大模型，综合性能对标英伟达 A100，实力处于国内领先水平。

思元 590 是寒武纪最新一代云端智能训练芯片，该产品性能相比思元 370 有翻倍以上的提升。根据测试数据，寒武纪 590 单卡性能测试接近 A100，达到 A100 80%-90%的程度，目前 MLU-Link 的片间互联速度 512GB/s（A800 是 400GB/s），集群互联目前性能发挥大概在 A100 8090%之间。思元 590 采用寒武纪自研的 MLUarch05 架构，能够提供更大的内存容量和带宽，IO 和片间互联接口也较上代实现大幅升级，主要面向训练任务。

在智能计算中心建设方面，寒武纪取得了显著进展。南京智能计算中心是寒

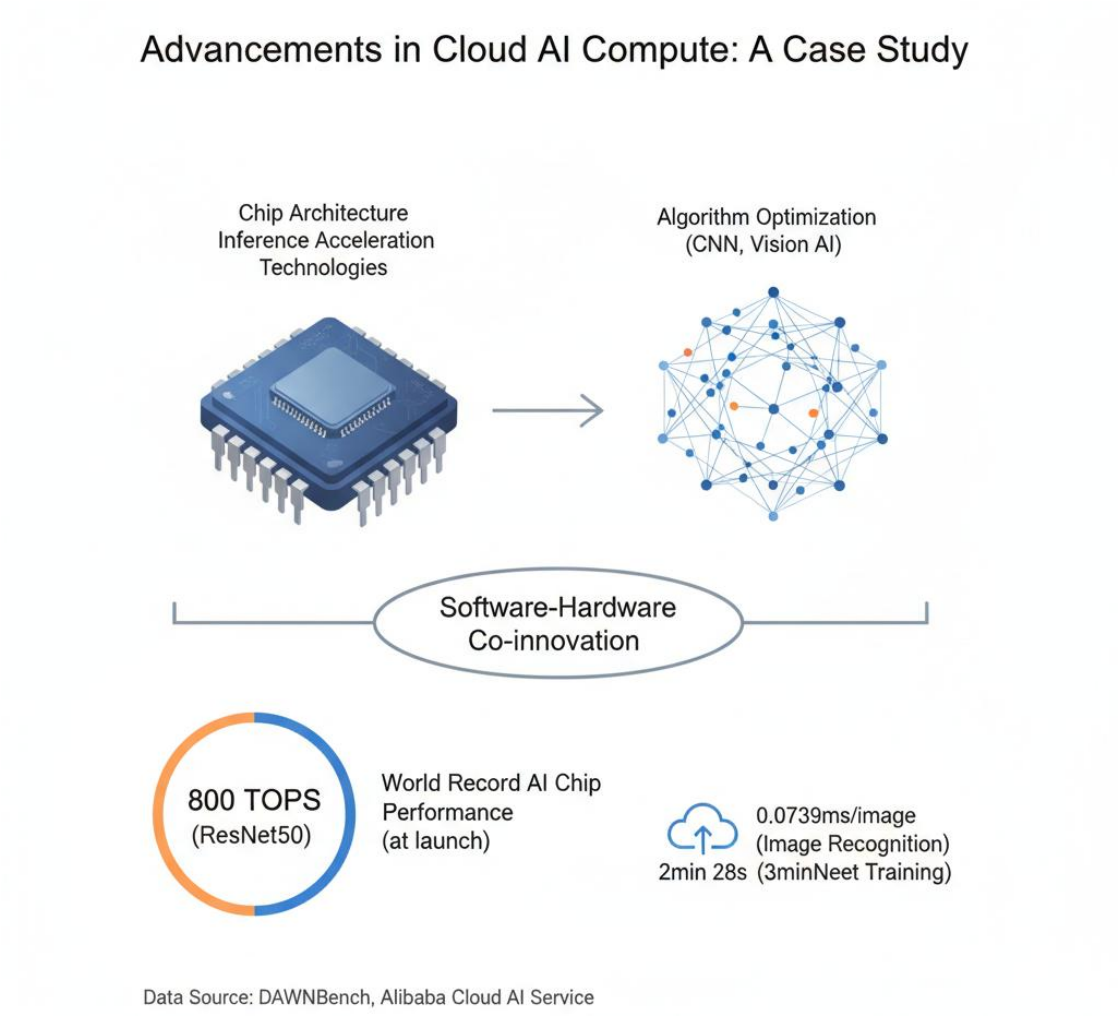
武纪的重要落地案例，该中心由 7280 块搭载国产芯片的 AI 智能加速卡提供智能算力，采用全国产化硬件和软件体系，算力达到每秒 180 亿亿次（1800P FLOPS）。2025 年 2 月，南京智算中心宣布联合国产芯片厂商寒武纪，用全国产设备运行国产大模型 DeepSeek，为苏宁易购提供全国产化模型推理服务，成为南京首例全国产算力版 DeepSeek 商业化应用案例。

寒武纪的智能计算中心布局不仅限于南京，在全国范围内都有重要项目。中心通过提供智能算力、通用算力、行业应用等，服务包括中科院计算所、中国科学技术大学、南京大学、寒武纪行歌、中汽创智在内的近百家科研院所、高校机构和企业。寒武纪在 2024 年上半年持续发力智能计算集群系统的部署效率，其训练软件平台开发了集群分析工具，完善了故障判断逻辑，同时优化了故障处理流程，进一步提升了产品竞争力。

在技术路线上，寒武纪主要采用 ASIC 架构，劣势是通用性会比较差，优势是某些特定应用场景下，算力可以做到比 GPU 更高。寒武纪思元 590 与英伟达的差距主要体现在通用性方面，但在特定应用场景下具有优势。百度内部的测试结果显示，590 在某些任务上表现优异，展现了寒武纪芯片的技术实力。

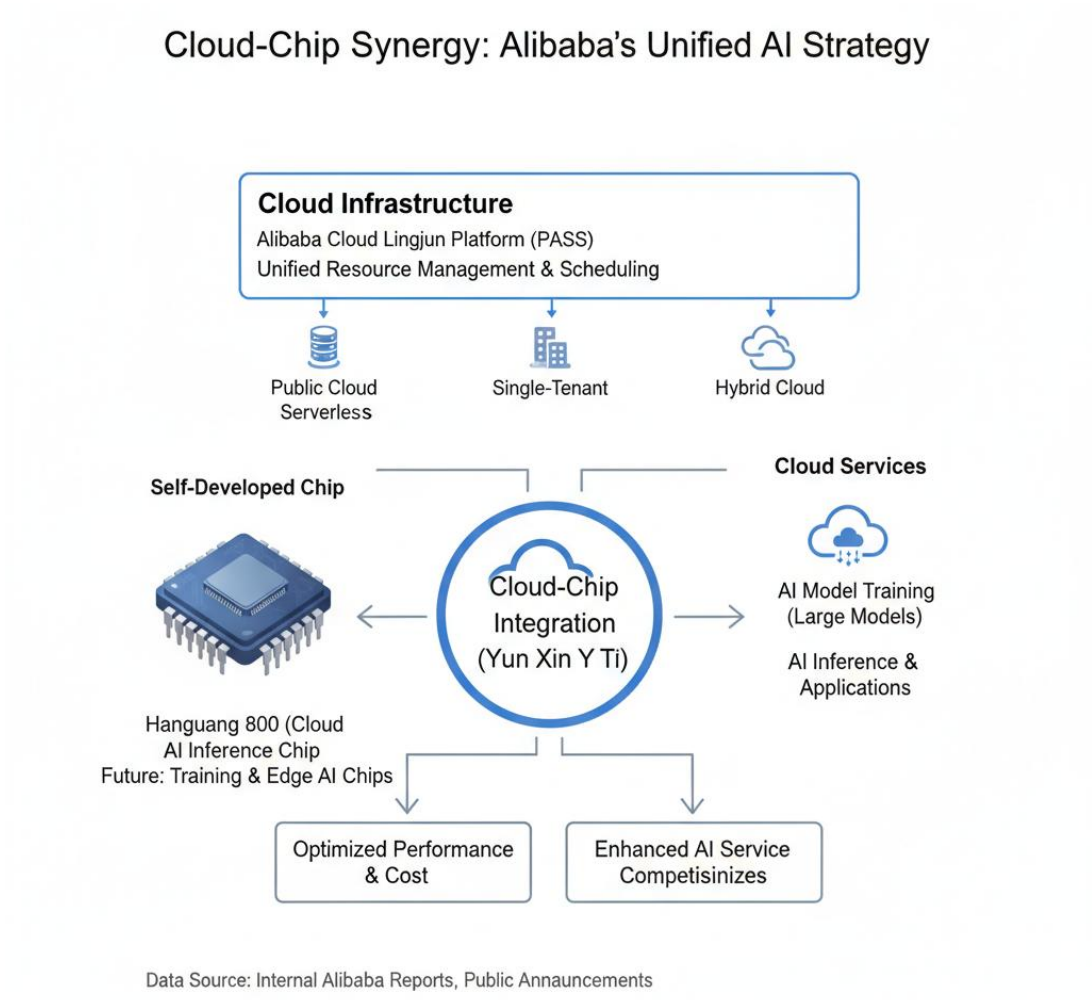
5.2.2 阿里平头哥与含光芯片

阿里平头哥是阿里巴巴旗下的半导体公司，其含光 800 芯片是面向云端推理的高性能 AI 芯片。含光 800 性能的突破得益于软硬件的协同创新：硬件层面采用自研芯片架构，通过推理加速等技术有效解决芯片性能瓶颈问题；软件层面集成了达摩院先进算法，针对 CNN 及视觉类算法深度优化。按照 ResNet50 需要的算力反推，含光 800 的算力达到 820TOPS，在当时创造了全球 AI 芯片性能的新纪录。



阿里平头哥含光 800 芯片的性能优势与技术创新

在 AI 场景中，含光 800 是异构计算的很好补充。阿里云基于含光 800 的 AI 服务识别一张图片仅需 0.0739ms，同时在训练成本和推理效率方面都有显著提升。DAWNBench 官方显示，阿里云异构计算服务训练 ImageNet 128 万张图片仅需 2 分 38 秒，展现了含光 800 在实际应用中的优异性能。未来，平头哥的产品形态还会进一步完善，包括云端 AI 训练芯片和端侧 AI 推理芯片，形成完整的产品矩阵。



阿里云灵骏平台与“云芯一体”战略

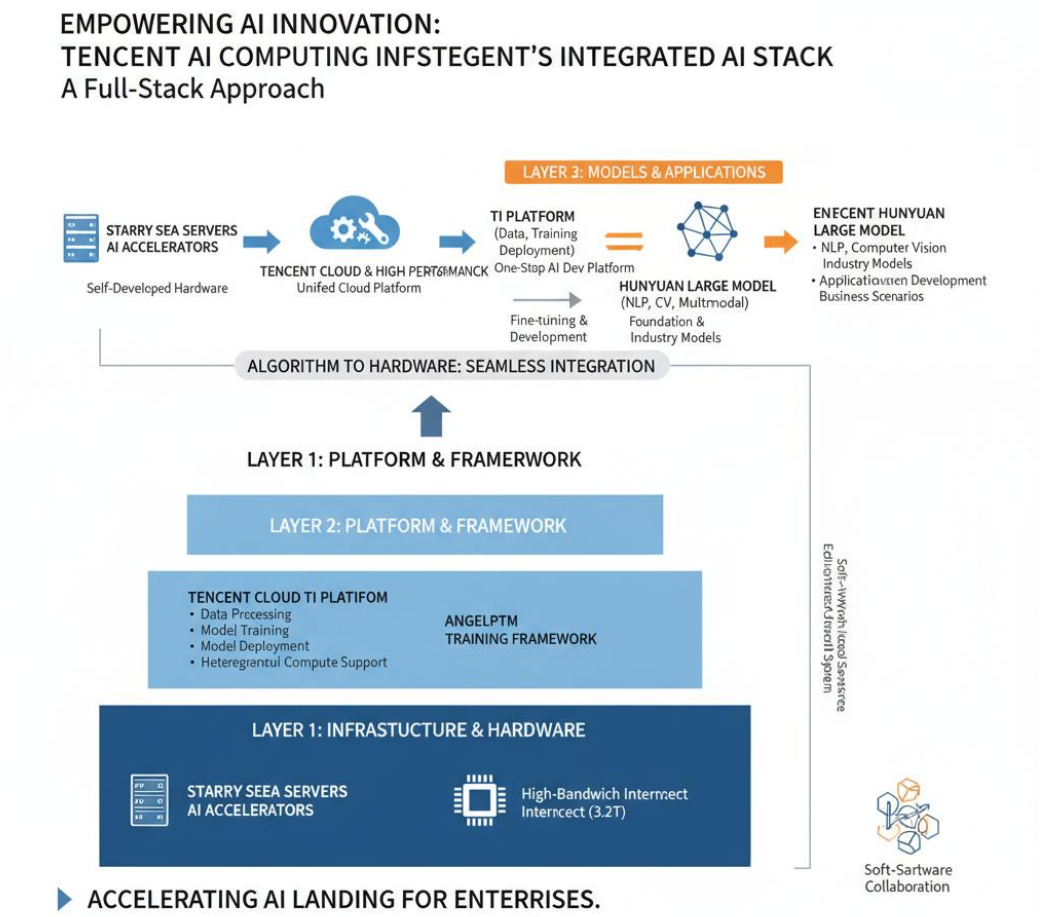
阿里云灵骏是面向大规模深度学习及融合智算的 PaaS 产品，支持公共云 Serverless 版、单租版以及混合云形态，基于软硬件一体优化技术，构建高性能异构算力底座。灵骏平台整合了含光 800 等异构计算资源，通过统一的资源管理和调度系统，为 AI 大模型训练和推理提供高效的算力支持。在实际应用中，灵骏平台已经支持了阿里巴巴内部众多 AI 业务，并通过阿里云对外提供服务，助力企业 AI 创新。

阿里巴巴在 AI 芯片领域的布局体现了"云芯一体"的战略思想，通过自研芯片与云服务的深度结合，实现性能和成本的最优化。含光 800 与阿里云的协同，不仅提升了阿里云 AI 服务的竞争力，也为阿里巴巴自身的 AI 业务提供了强大的算力支撑，形成了技术与业务的良性循环。

5.2.3 腾讯星星海与 AI 加速卡

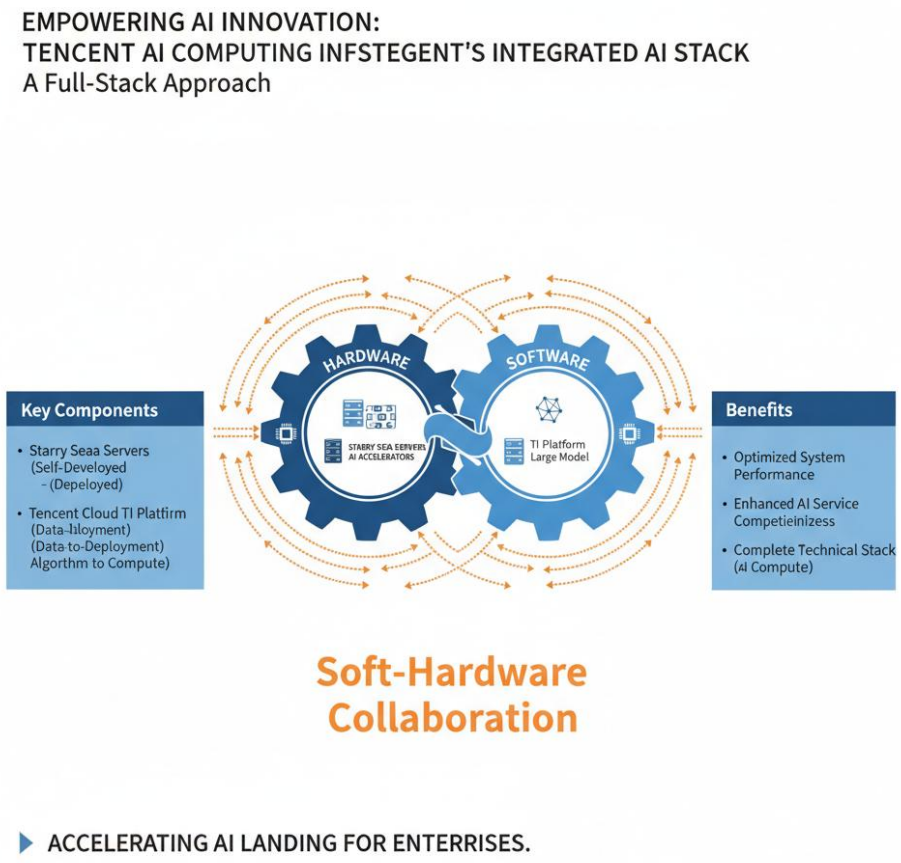
腾讯在 AI 算力领域的布局主要体现在星星海自研服务器和 AI 加速卡上。腾

讯云结合星星海自研服务器，自研高性能智能网络提供的 3.2T 超高互联带宽，以及腾讯自研 AI 加速卡，构建了强大的 AI 算力基础设施。在腾讯云上，企业基于 TI 平台的大模型能力和工具箱，可结合自身场景数据，进行大模型的精调和应用开发，加速 AI 落地。



腾讯混元大模型是腾讯 AI 技术的重要成果，已经覆盖了自然语言处理、计算机视觉、多模态等基础模型和众多行业、领域模型。混元大模型背后的训练框架 AngelPTM，也已通过腾讯云对外提供服务，帮助企业加速大模型落地。腾讯混元 AI 大模型与腾讯云的算力基础设施深度结合，形成了从算法到算力的完整技术栈。

腾讯云 TI 平台是腾讯云面向 AI 开发的一站式平台，提供了从数据处理、模型训练到模型部署的全流程支持。TI 平台与腾讯云的异构算力基础设施深度集成，支持多种 AI 框架和硬件平台，为开发者提供灵活高效的开发环境。在实际应用中，TI 平台已经支持了腾讯内部众多 AI 业务，并通过腾讯云对外提供服务，助力企业 AI 创新。

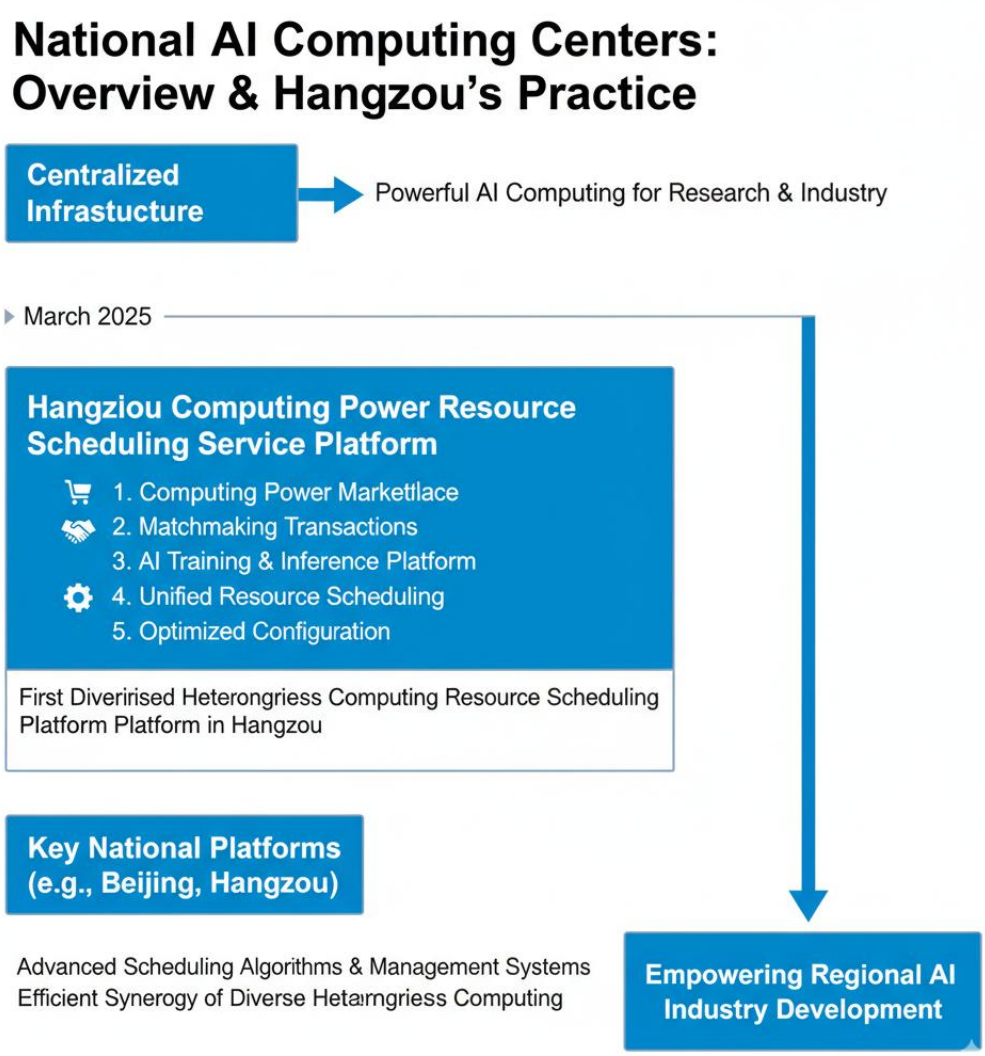


腾讯在 AI 算力领域的布局体现了"软硬协同"的理念，通过自研硬件与软件平台的深度结合，实现系统性能的最优化。星星海服务器与 AI 加速卡、TI 平台、混元大模型的协同，不仅提升了腾讯云 AI 服务的竞争力，也为腾讯自身的 AI 业务提供了强大的技术支撑，形成了从底层硬件到上层应用的完整技术体系。

5.3 智算中心与云服务商实践

5.3.1 国家级智算中心

国家级智算中心是中国 AI 算力基础设施的重要组成部分，通过集中化的建设和运营，为科研机构和企业提供强大的 AI 计算能力。北京、杭州等地的智算中心在异构算力资源调度方面进行了积极探索，形成了各具特色的实践案例。



杭州市算力资源调度服务平台于 2025 年 3 月正式启用，是杭州首个多元异构力资源调度服务平台。该平台首批接入 5 家数据中心，整合了通用算力、智能算力等多元资源，具有算力超市、撮合交易、AI 训推一体化平台等五大重点功能。通过统一的资源调度和管理，杭州市算力资源调度服务平台实现了算力资源的高效利用和优化配置，为区域 AI 产业发展提供了强大支撑。

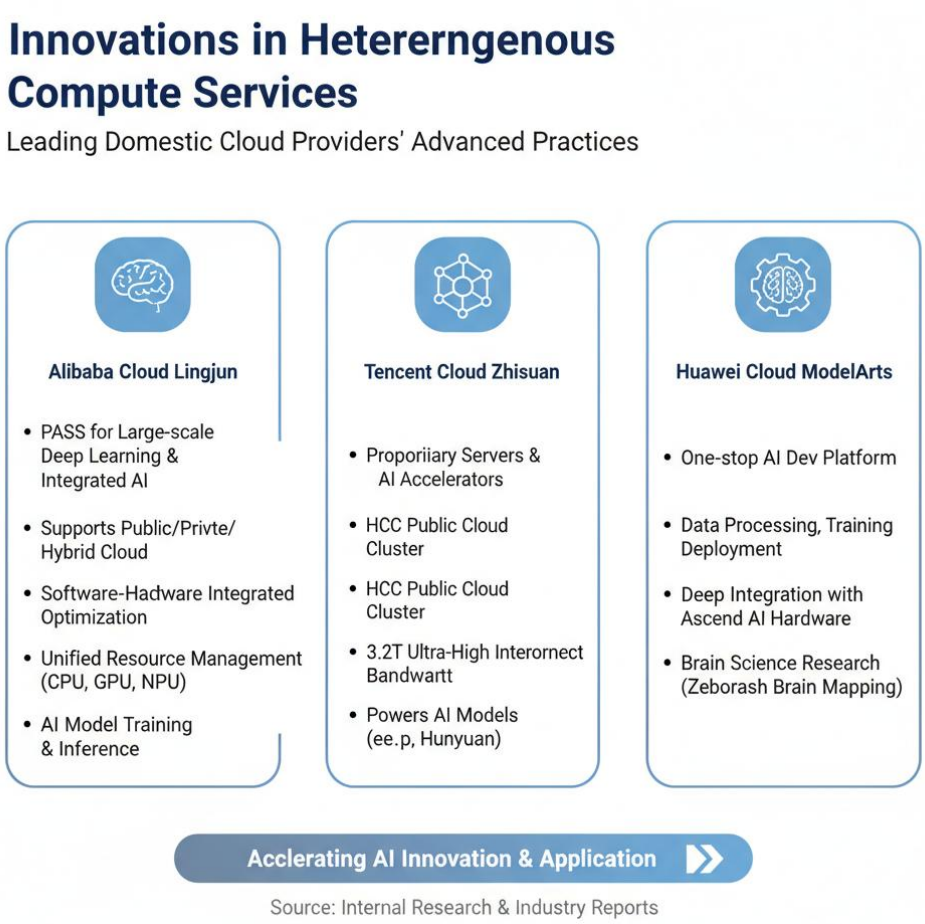
国家新一代 AI 公共算力开放创新平台（北京、杭州等）是国家级智算中心的代表，这些平台不仅提供强大的算力资源，还构建了完整的技术生态和服务体系。在这些平台中，异构算力资源调度是核心技术挑战，需要解决不同架构硬件的统一管理、任务调度、负载均衡等问题。通过先进的调度算法和管理系统，这些平台实现了多元异构算力的高效协同，为各类 AI 应用提供了强大的算力支持。

智算中心的建设面临着核心供给不足与结构错配、通信连接瓶颈、算力调度复杂以及异构算力生态融合难等挑战。为应对这些挑战，智算中心建设要以开放

硬件和开源软件为主，融合多元算力，实现算力的聚合、调度、释放，让智算中心"用起来、用得好"。同时，要朝着标准化、集约化方向发展，提高建设和运营效率，降低使用成本。

5.3.2 商业云服务商

商业云服务商在异构算力服务方面进行了大量创新，阿里云灵骏、腾讯云智算、华为云 ModelArts 等平台代表了国内云服务商在异构算力领域的先进实践。

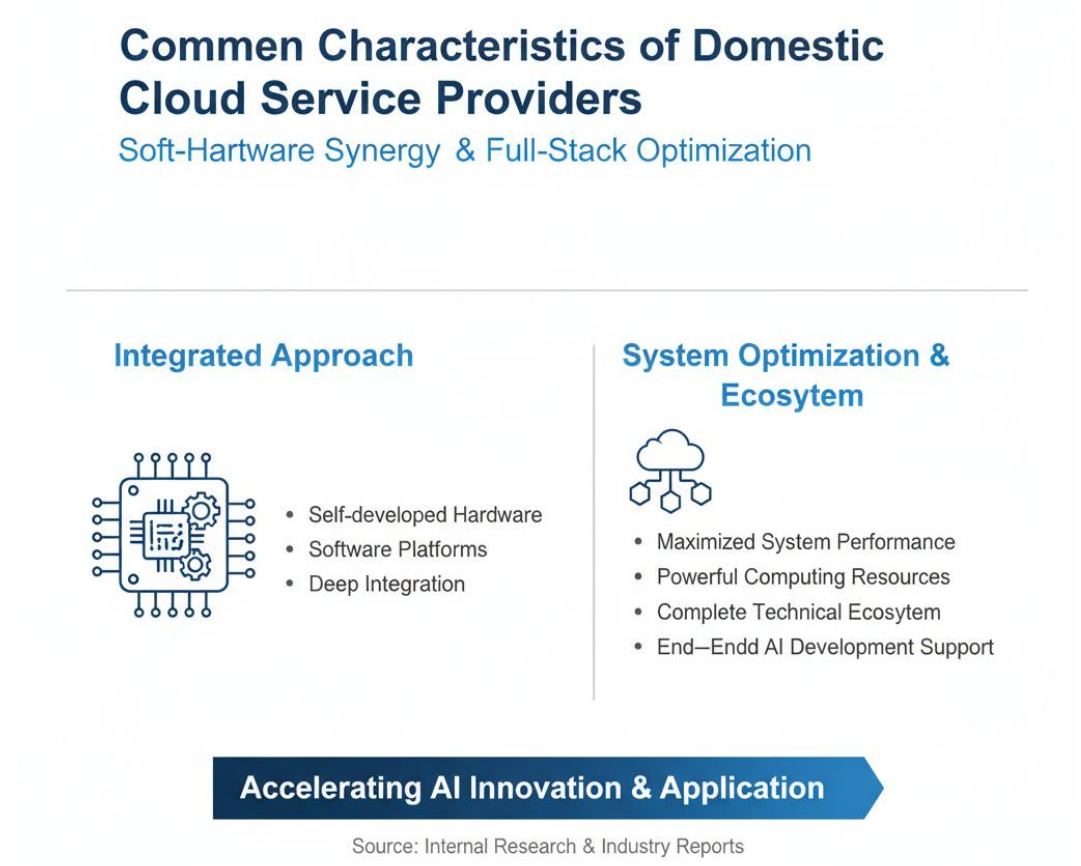


阿里云灵骏是面向大规模深度学习及融合智算的 PaaS 产品，支持公共云 Serverless 版、单租版以及混合云形态。灵骏基于软硬件一体优化技术，构建高性能异构算力底座，整合了 CPU、GPU、NPU 等多种计算资源，通过统一的资源管理和调度系统，为 AI 大模型训练和推理提供高效的算力支持。灵骏平台支持多种 AI 框架和硬件平台，为开发者提供灵活高效的开发环境。

腾讯云智算结合腾讯星星海自研服务器和 AI 加速卡,构建了强大的 AI 算力基础设施。腾讯云在行业率先发布了大模型公有云算力集群 HCC，该集群结合腾讯云星星海自研服务器，腾讯云自研高性能智能网络提供的 3.2T 超高互联带

宽，以及腾讯自研 AI 加速卡，为大规模 AI 训练和推理提供了强大支撑。在实际应用中，腾讯云智算已经支持了混元大模型等众多 AI 业务，展现出优异的性能和稳定性。

华为云 ModelArts 是华为云面向 AI 开发的一站式平台，提供了从数据处理、模型训练到模型部署的全流程支持。ModelArts 与华为昇腾 AI 硬件深度集成，通过软硬件协同优化，实现了高性能的 AI 训练和推理。在脑科学研究方面，华为云 ModelArts 也有所作为：中科院脑智卓越中心通过完整解析斑马鱼的透明大脑来揭示大脑的工作原理，对接类脑智能。ModelArts 平台的易用性和高性能，使其成为科研机构和企业 AI 开发的重要工具。



国内云服务商在异构算力服务方面的共同特点是"软硬协同、全栈优化"，通过自研硬件与软件平台的深度结合，实现系统性能的最优化。这些平台不仅提供强大的算力资源，还构建了完整的技术生态和服务体系，为 AI 开发者提供从底层硬件到上层应用的全方位支持，加速了 AI 技术的创新和应用。

5.4 开源社区与开发者生态

5.4.1 国内 AI 开源平台

国内 AI 开源平台在推动异构算力与大模型融合方面发挥着重要作用，OpenI 启智、ModelScope、飞桨开源社区、算网 AI 平台等平台为开发者提供了丰富的资源和支持。

OpenI 启智是由新一代人工智能产业技术创新战略联盟（AITISA）组织运作的开源社区，旨在培育高水平的开源技术，汇聚国家从事开源项目的开发者和各个单位的力量。OpenI 启智社区旗下的一站式 AI 开发协作平台，汇聚人工智能开源项目，涵盖自动建模、算力容器、自动参数调优、模型部署、数据标注工具等功能。通过社区建设，OpenI 启智希望建立一个从底层芯片到上层应用的技术体系，推动 AI 技术的开源共享和协同创新。

飞桨（PaddlePaddle）是百度自主研发的产业级深度学习平台，集深度学习核心框架、基础模型库、端到端开发套件、工具组件和服务于一体。飞桨同时支持动态图和静态图，兼顾灵活性和效率；精选应用效果最佳算法模型并提供官方支持；真正源于产业实践，经过大规模业务验证。据统计，依托飞桨，产学研用共建技术和产业生态，已累计培养超过百万 AI 人才。

ModelScope 是阿里巴巴达摩院推出的 AI 模型开源平台，提供了大量预训练模型和开发工具，支持开发者快速构建 AI 应用。ModelScope 平台整合了阿里巴巴在 AI 领域的技术积累，涵盖自然语言处理、计算机视觉、语音识别等多个领域，为开发者提供了丰富的模型资源和开发支持。通过开源共享，ModelScope 促进了 AI 技术的普及和创新，推动了异构算力在更多场景中的应用。

算网平台（https://sumw.com.cn/）是中科算网旗下推出的面向 AI 开发者的平台，集异构算力、模型、数据为一体的一站式 AI 开发平台，涵盖了从数据到模型训练微调服务、模型部署服务、按需租赁算力等服务。涵盖了国内外主流的模型库，如 DeepSeek、阿里千问系列等模型，可一键部署和使用。同时基于该平台构建的算泥开发者社区，为广大的 AI 开发者、高校学生/老师/科研团队提供了成体系的 AI 大模型相关课程，市场报告，专业的内容输出，AI 项目的开发与交流平台。

这些国内 AI 开源平台的共同特点是"开放共享、生态共建"，通过开源代码、模型、工具等资源，降低 AI 技术门槛，促进技术创新和应用落地。同时，这些平台也积极支持国产异构算力，通过优化适配和性能调优，推动国产 AI 芯片在实际应用中的普及和成熟。

5.4.2 开发者工具链与支持

完善的开发者工具链和支持体系是异构算力与大模型融合的关键保障。国内企业和开源社区在开发者工具链建设方面进行了大量投入，为开发者提供了丰富的资源和支持。

华为昇腾社区提供了完整的开发资源和支持体系，包括社区版和商业版CANN（Compute Architecture for Neural Networks）计算架构。昇腾开发资源下载中心提供社区版和商业版下载，其中社区版快速提供新特性的体验版，供开发者提前试用；商业版满足商用标准的稳定版本。昇腾开发指南用于指导开发者如何基于昇腾平台进行模型开发、应用开发、算子开发，并提供常见故障处理指导以及日志参考等，为开发者提供全方位的技术支持。

算泥社区（https://c.sumw.com.cn/）作为 AI 大模型开发者平台，在异构算力与大模型融合中扮演着重要角色。算泥社区提供"AI 大模型开发服务+模型+算力"的三位一体服务，通过 API、镜像、教程等丰富的开发者资源，降低开发者使用异构算力的门槛。算泥社区不仅提供技术支持，还通过社区活动、培训认证、开发者大赛等形式，促进技术交流和人才培养，推动异构算力与大模型融合技术的普及和应用。

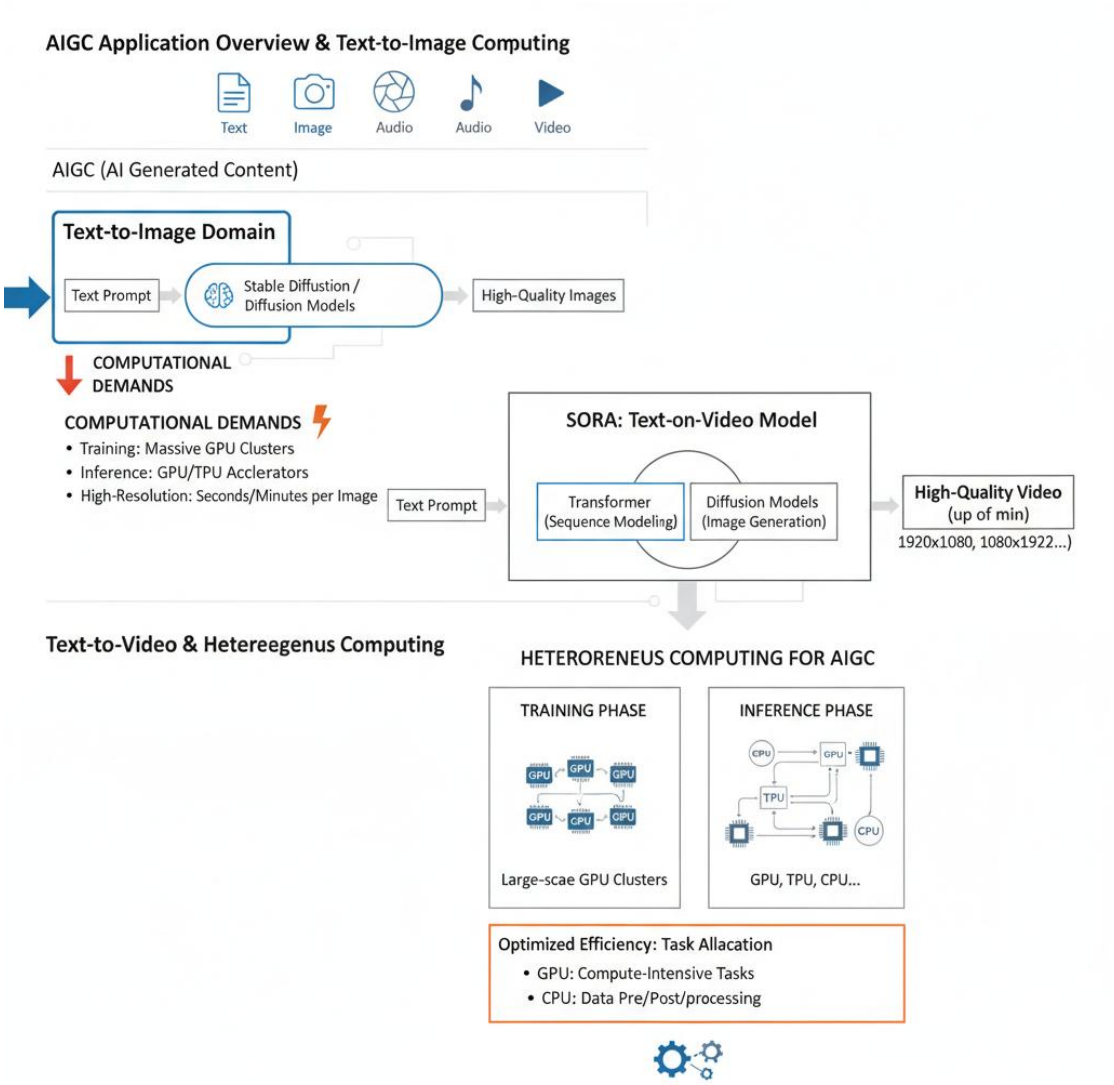
国内企业和开源社区在开发者工具链建设方面的共同特点是"全栈支持、生态共建"，通过提供从底层硬件到上层应用的完整工具链和支持体系，降低开发者使用异构算力的门槛，加速技术创新和应用落地。这些工具链和支持体系不仅提高了开发效率，也促进了异构算力在更多场景中的应用，推动了整个 AI 产业的健康发展。

六、行业应用与场景落地

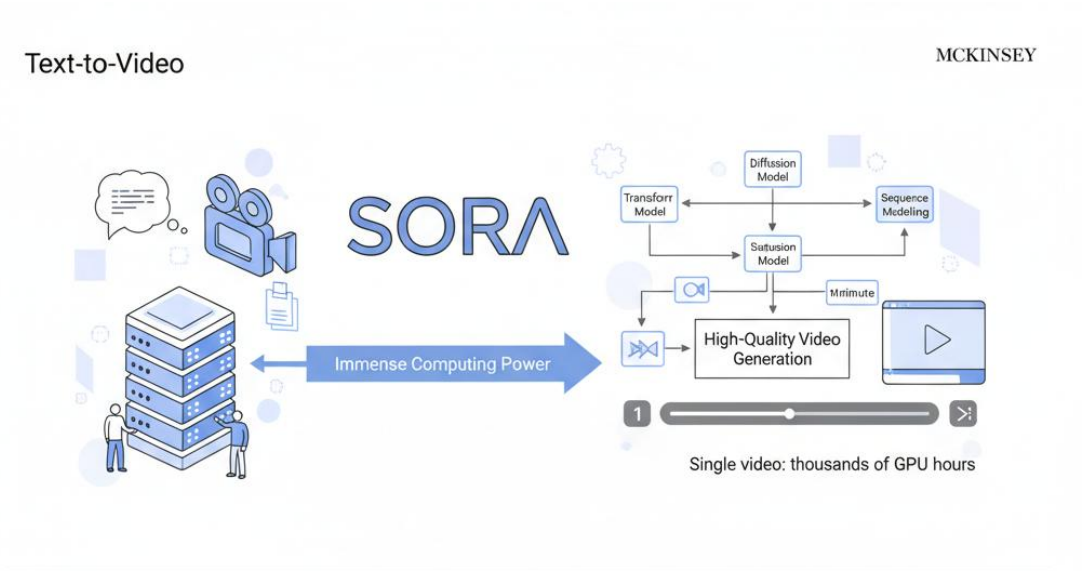
6.1 互联网与内容生成

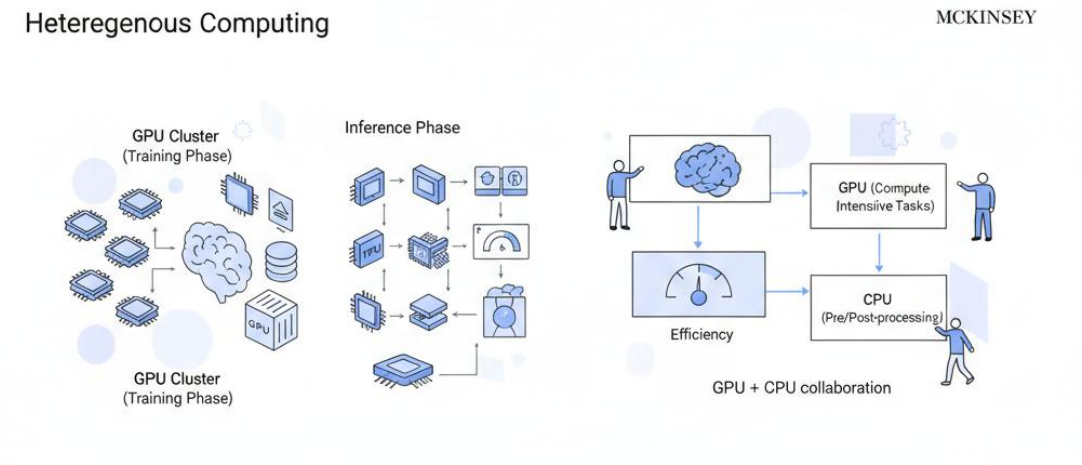
6.1.1 AIGC 应用

AIGC（AI Generated Content）是当前大模型技术最具代表性的应用场景，涵盖了文本、图像、音频、视频等多种内容形式的自动生成。在文生图领域，Stable Diffusion 等模型通过扩散模型技术，能够根据文本描述生成高质量图像，广泛应用于创意设计、广告制作、游戏开发等领域。这些模型的训练和推理需要大量算力支持，特别是在高分辨率图像生成场景下，单次生成可能需要数秒到数分钟的计算时间，对算力的实时性和稳定性提出了高要求。



文生视频是 AIGC 领域的新兴方向，Sora 能根据文本生成最长 20 秒的高质量视频，理论上支持任意分辨率，如 1920x1080、1080x1920 等。从技术原理上看,Sora 可以理解成是一种融合 Transformer 模型与 Stable Diffusion 的混合模型，通过 Transformer 原理的序列建模能力，结合扩散模型的图像生成能力，实现了高质量的视频生成。Sora 的出现彻底颠覆了文生视频领域，但其背后是巨大的算力需求，单次视频生成可能需要数千 GPU 小时的计算量。



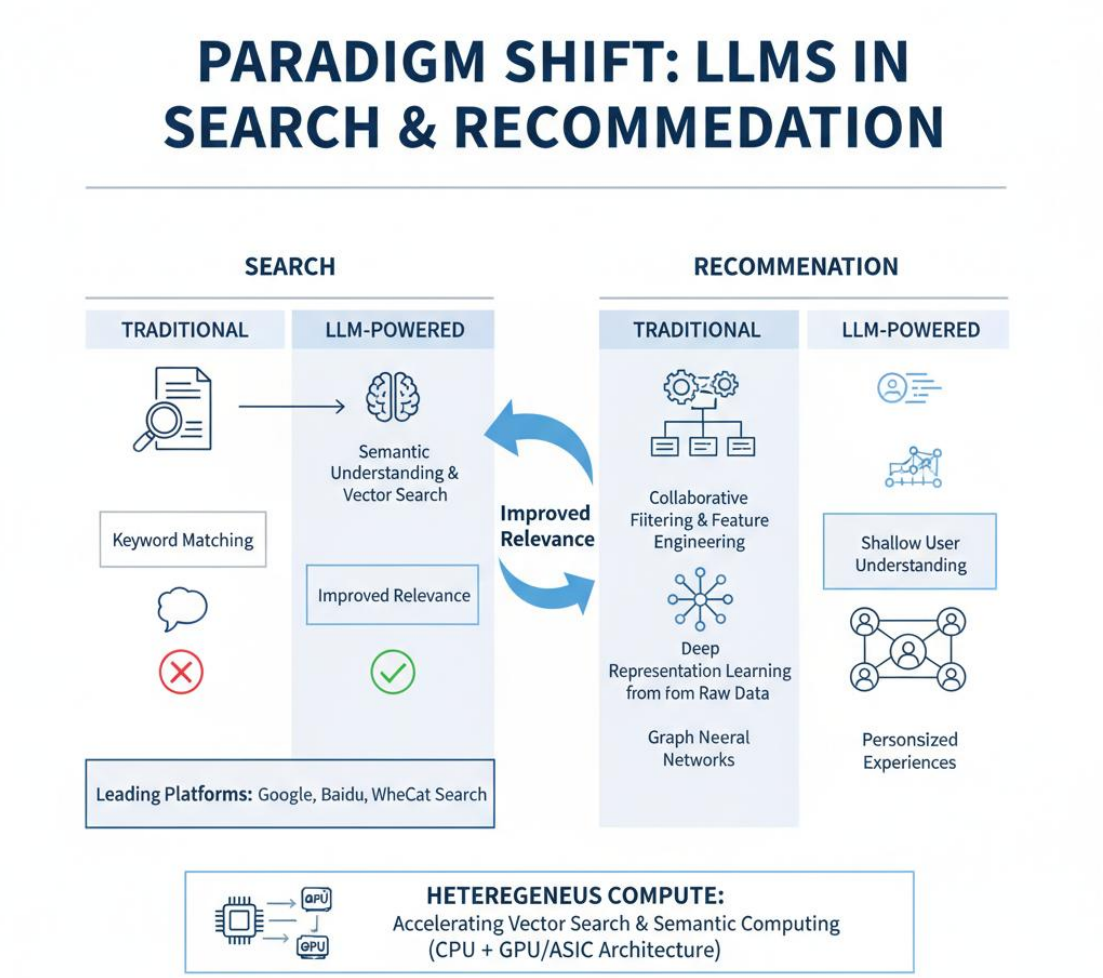


异构算力在 AIGC 应用中发挥着关键作用。在训练阶段，大规模 GPU 集群提供了必要的计算能力，支持模型在海量数据上的训练；在推理阶段，GPU、TPU 等专用加速器则提供了高效的推理性能，满足实时生成需求。特别是在视频生成等高计算复杂度场景，异构算力通过不同类型处理器的协同工作，实现了计算效率的最优化。例如，可以使用 GPU 处理主要的计算密集型任务，而使用 CPU 处理数据预处理和后处理等任务，通过合理的任务分配，实现整体性能的提升。

6.1.2 大模型搜索与推荐

大模型技术在搜索与推荐领域的应用正在深刻改变传统的信息获取方式。传统搜索主要基于关键词匹配，难以理解用户的真实意图；而基于大模型的搜索则通过向量检索和语义理解技术，能够更准确地把握用户需求，提供更相关的搜索结果。向量检索技术通过将文本转换为高维向量，计算向量间的相似度来实现语

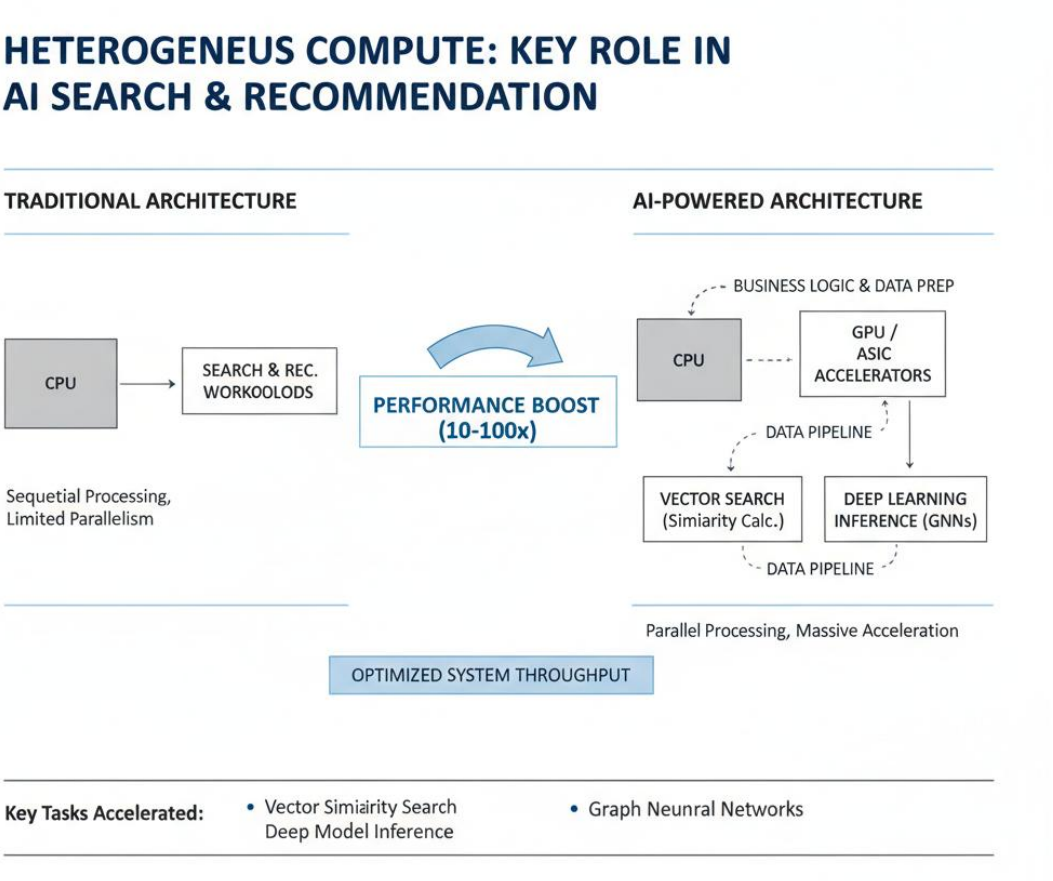
义匹配，能够处理语义关系、上下文和数据的丰富语义信息，适用于处理图像、音频、视频等多种数据类型。



大模型在搜索和推荐领域的应用范式转变

在推荐系统领域，大模型通过深度理解用户行为和内容特征，实现了更精准的个性化推荐。传统的推荐系统主要依赖协同过滤和特征工程，而大模型推荐则能够直接从原始数据中学习用户和内容的深层表示，捕捉更复杂的关联关系。例如，通过构建用户与内容的交互图，利用图神经网络学习节点表示，可以实现更精准的推荐效果。

目前全球主要的搜索厂商，百度、谷歌均对原始的搜索方式与大模型进行了整合，搜索结果页面除了传统的网页索引外，还在搜索结果顶部给出了大模型的直接结果供用户参考。腾讯在微信内部搜一搜也集成了大模型搜索结果，预计未来会更进一步增加大模型搜索权重。



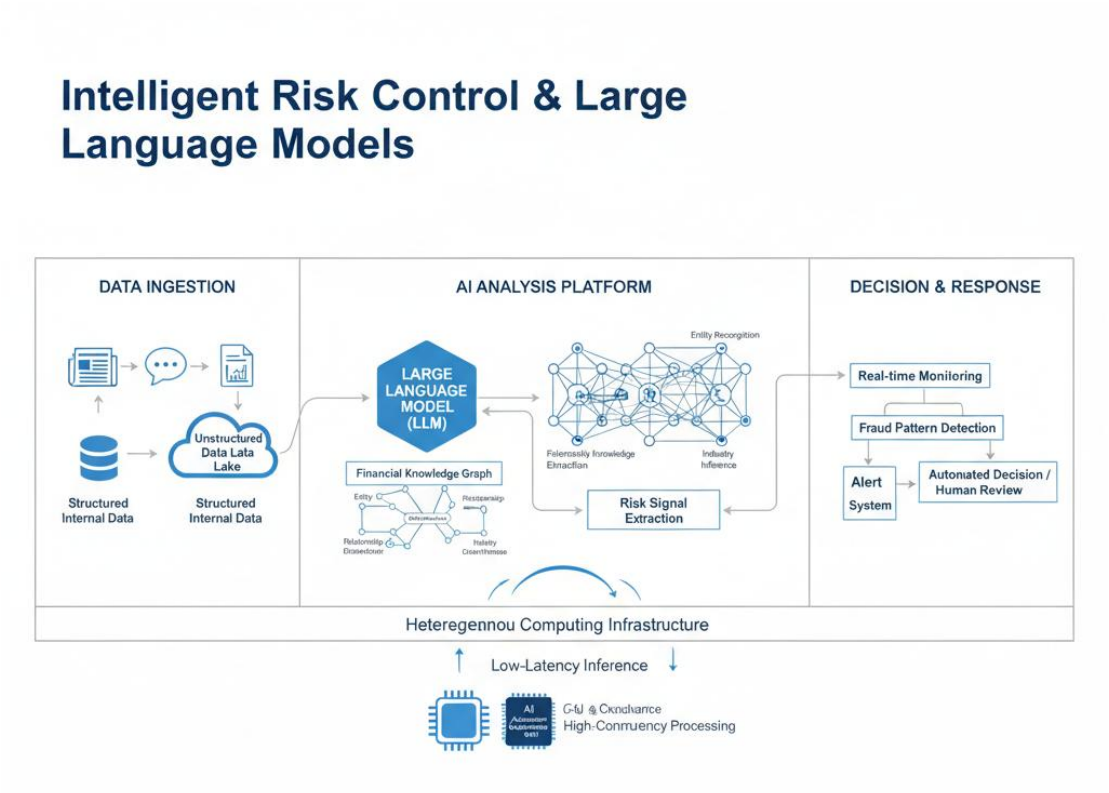
异构算力在大模型搜索与推荐中的关键作用

异构算力在大模型搜索与推荐系统中主要支持向量检索和语义计算等密集型任务。GPU/ASIC 加速推荐系统推理，特别是在向量相似度计算、图神经网络推理等场景下，能够提供数十倍甚至上百倍的加速效果。在实际部署中，通常采用 CPU+GPU 的异构架构，CPU 负责业务逻辑和数据预处理，GPU 负责向量计算和模型推理，通过合理的任务划分和数据流水线，实现系统整体性能的最优化。

6.2 金融与医疗

6.2.1 智能风控与投研

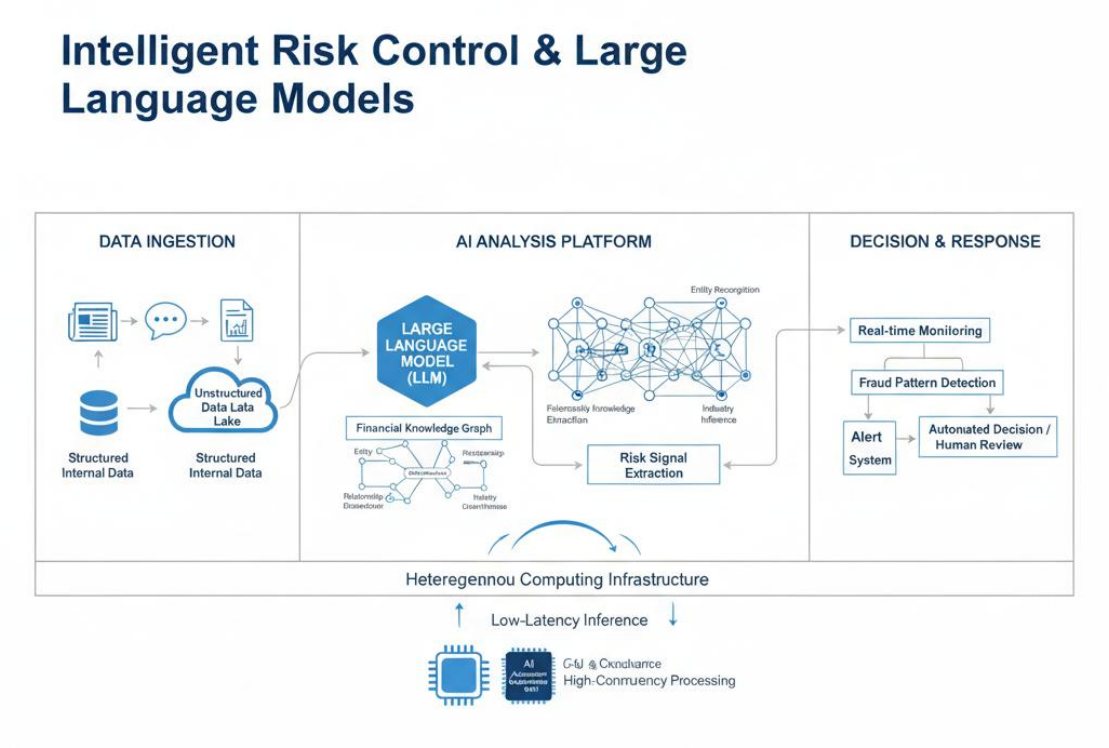
金融行业是大模型技术的重要应用领域，智能风控和智能投研是两个典型场景。在智能风控方面，大模型与知识图谱的结合展现出强大能力。金融机构通过构建金融知识图谱来进行市场数据及基本面分析，通过自然语言处理、关联关系分析、行业产业知识推理等为动态、多维度的基本面分析做支持。在信贷风控的业务实现中，通常需要搭建知识图谱分析平台，梳理现有一方数据，整合必要的三方数据后导入知识图谱数据库，建立知识图谱风控流程和预警体系。



智能风控流程与大模型

大模型在金融风控中的优势在于能够处理非结构化数据，如新闻、公告、社交媒体等信息，从中提取风险信号。通过深度学习技术，大模型能够识别复杂的欺诈模式和异常行为，实现实时风险监测和预警。在实际应用中，金融机构通常将大模型与传统风控系统结合，形成多层次的风控体系，既利用大模型的语义理解能力，又保持传统系统的稳定性和可解释性。

异构算力在金融风控场景中主要支持低延迟推理和高并发处理。寒武纪 MLU 等国产 AI 芯片在金融客户案例中表现出色，特别是在实时风险监测、交易反欺诈等对延迟敏感的场景。通过异构计算架构，可以实现毫秒级的风险评估和决策响应，满足金融业务对实时性的高要求。同时，异构算力的高并发处理能力，使得系统能够同时处理大量交易和用户行为数据，实现全方位的风险覆盖。

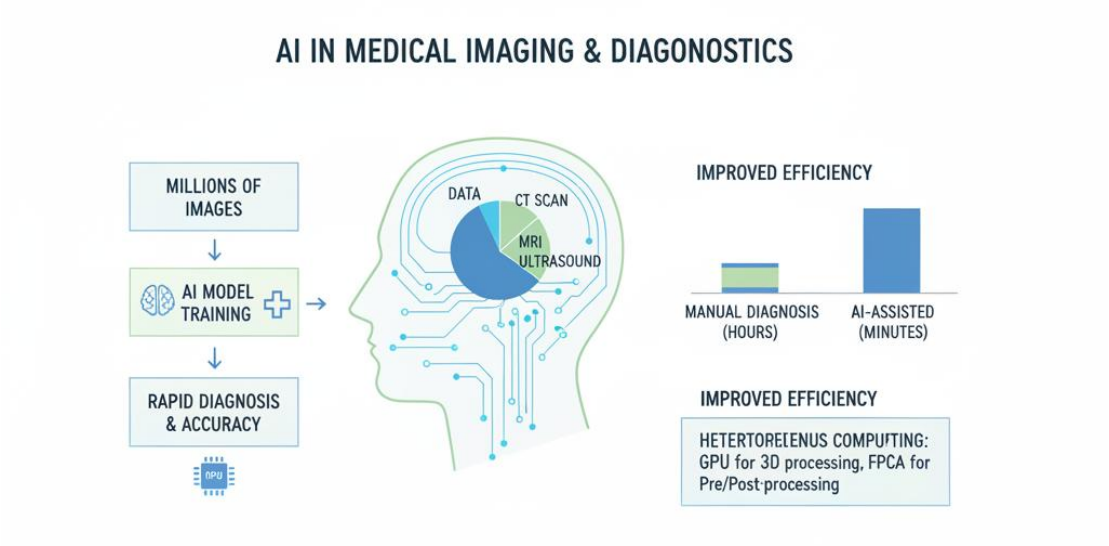


大模型在智能投研中的应用

在智能投研方面，大模型通过提供具有针对性的提示，能够更加深入地分析市场数据，为投资决策提供支持。大模型可以快速处理海量财经新闻、公司公告、行业报告等信息，提取关键观点和趋势，辅助投资分析师进行决策。在实际应用中，大模型通常与量化模型结合，形成"AI+量化"的投资策略，既利用 AI 的信息处理能力，又保持量化模型的纪律性和系统性。

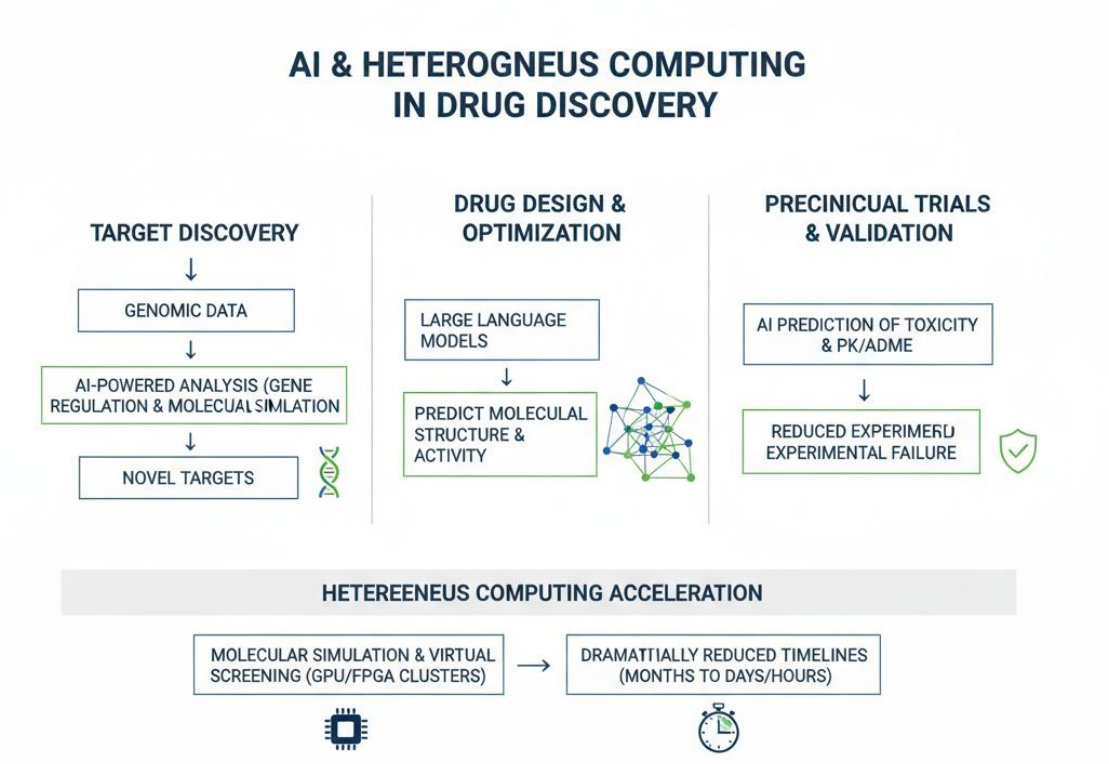
6.2.2 医学影像与药物研发

医疗领域是大模型技术的另一个重要应用场景，医学影像和药物研发是两个代表性方向。AI 医学影像是人工智能在医疗领域应用最为广泛的场景，率先落地、率先应用、率先实现商业化。在 GPU 的加持下，智能医学影像平台能支持数百万的医学影像数据的训练；同时基于训练的人工智能模型，可快速实现脑部、心脏以及身体各器官疾病的辅助诊断，大大提高了诊断效率和准确性。



医学影像与 AI 辅助诊断

在医学影像处理中，异构算力主要支持多模态数据处理和复杂模型推理。医学影像通常包括 CT、MRI、X 光等多种模态，每种模态的数据特点和诊断需求各不相同。异构计算架构通过不同类型处理器的协同工作，能够高效处理这些多样化的数据类型。例如，可以使用 GPU 处理 3D 卷积等计算密集型任务，而使用 FPGA 处理数据预处理和后处理等任务，通过合理的任务分配，实现整体处理效率的最优化。



AI 与异构算力驱动药物研发

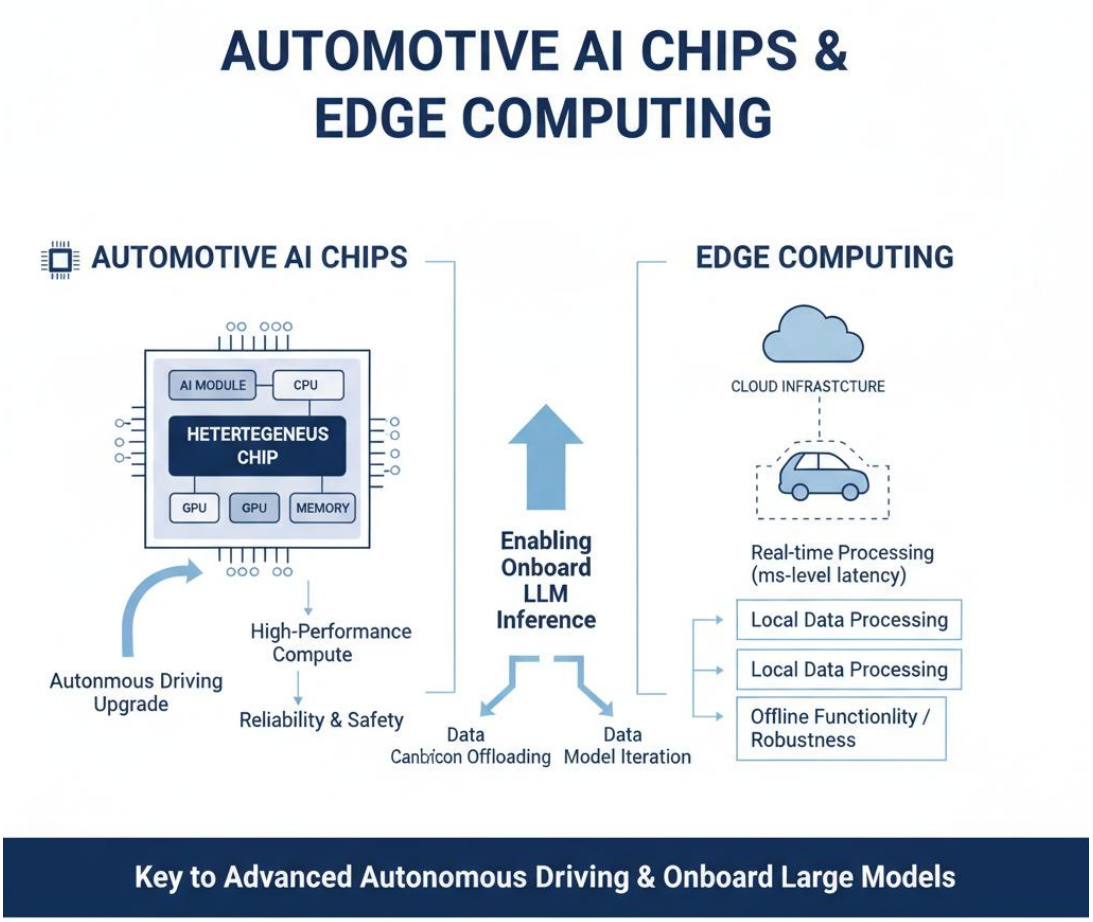
药物研发是另一个受益于大模型和异构算力的领域。人工智能与药学的交叉融合是重塑传统新药研发路径和范式的重要驱动力。首先，人工智能与药学的交叉融合有助于形成智慧化靶点发现系统，通过跨物种基因调控网络分析和分子模拟，加速潜在药物靶点的发现。其次，大模型可以预测分子结构与生物活性之间的关系，指导药物分子的设计和优化。最后，大模型还可以预测药物的毒副作用和药代动力学特性，减少实验失败的风险。

异构算力在药物研发中主要支持分子模拟和虚拟筛选等计算密集型任务。分子模拟需要计算分子间的相互作用力，预测分子的三维结构和动态行为，计算量巨大。异构计算架构通过 GPU 等专用加速器，可以显著加速分子动力学模拟和量子化学计算，将原本需要数月甚至数年的计算任务缩短到几天或几小时。在实际应用中，药物研发机构通常构建大规模异构计算集群，支持多个研发项目的并行计算需求，大大加速了新药研发进程。

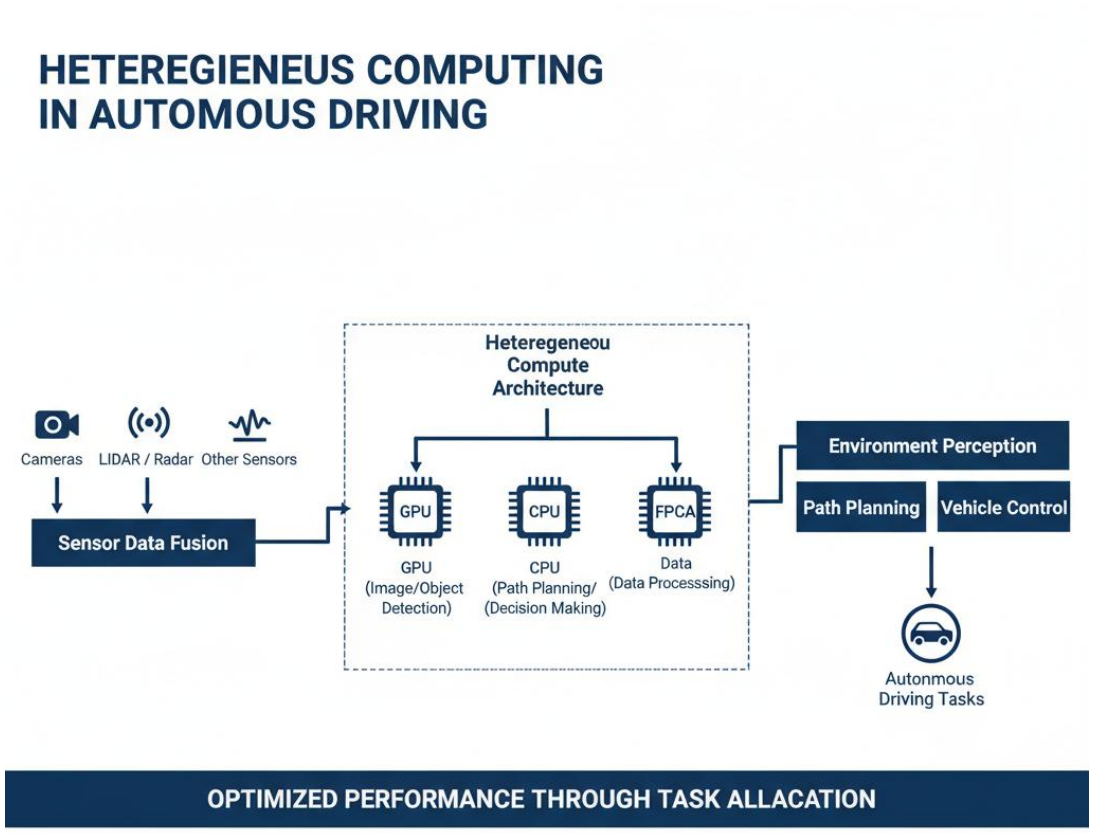
6.3 自动驾驶与智能制造

6.3.1 车规级 AI 芯片与边缘计算

自动驾驶是 AI 技术最复杂的应用场景之一，对算力、能效、可靠性等方面都提出了极高要求。近年来，大模型推理从云端走向边缘侧，已成为人工智能落地的重要趋势。相比传统规则式或轻量模型算法，车载大模型具备更强泛化能力和语义理解力，但其高算力需求一直是制约其广泛应用的主要因素。



车规级 AI 芯片是支持车载大模型推理的关键硬件。寒武纪行歌是寒武纪切入车载智能芯片的主体，致力于成为安全可靠的智能车载芯片引领者，用 AI 芯片支撑自动驾驶更快升级。寒武纪行歌提供的车载智能芯片是一个异构芯片，不仅包括 AI 模块，还包括 CPU、GPU 等多种计算单元，形成完整的异构计算架构。通过车云协同，能够将车端的数据快速回传，实现 AI 模型的快速迭代升级。



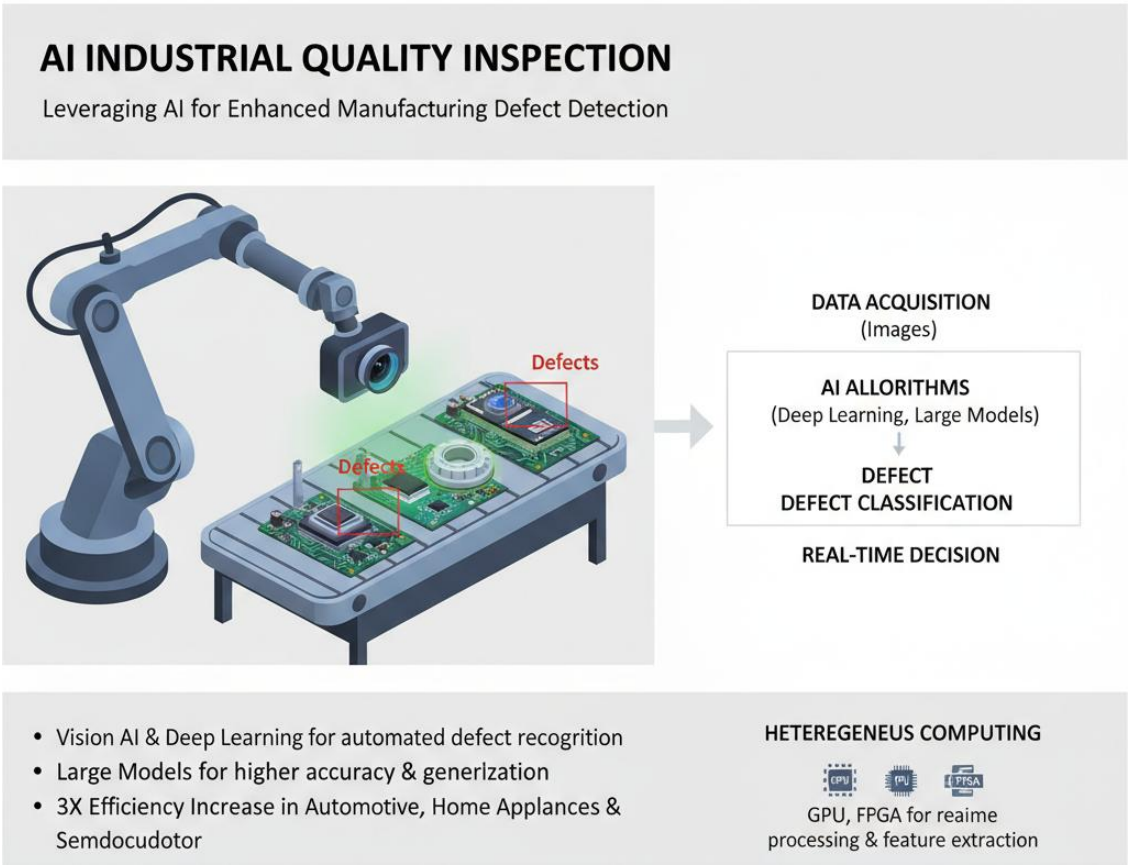
边缘计算在自动驾驶中扮演着重要角色。自动驾驶系统需要在毫秒级时间内处理大量传感器数据，做出驾驶决策，这对计算延迟提出了极高要求。边缘计算通过在车辆本地部署计算能力，避免了数据传输到云端再返回的延迟，满足了实时性要求。同时，边缘计算也可以在网络连接不稳定或断开的情况下保持基本功能，提高了系统的可靠性。

异构算力在自动驾驶中主要支持传感器数据处理、环境感知、路径规划等任务。自动驾驶系统需要处理摄像头、激光雷达、毫米波雷达等多种传感器的数据，每种数据类型和处理需求各不相同。异构计算架构通过不同类型处理器的协同工作，能够高效处理这些多样化的计算任务。例如，可以使用 GPU 处理图像识别和目标检测等并行计算任务，使用 CPU 处理路径规划和决策等串行任务，使用 FPGA 处理传感器数据预处理等专用任务，通过合理的任务分配，实现系统整体性能的最优化。

6.3.2 工业质检与数字孪生

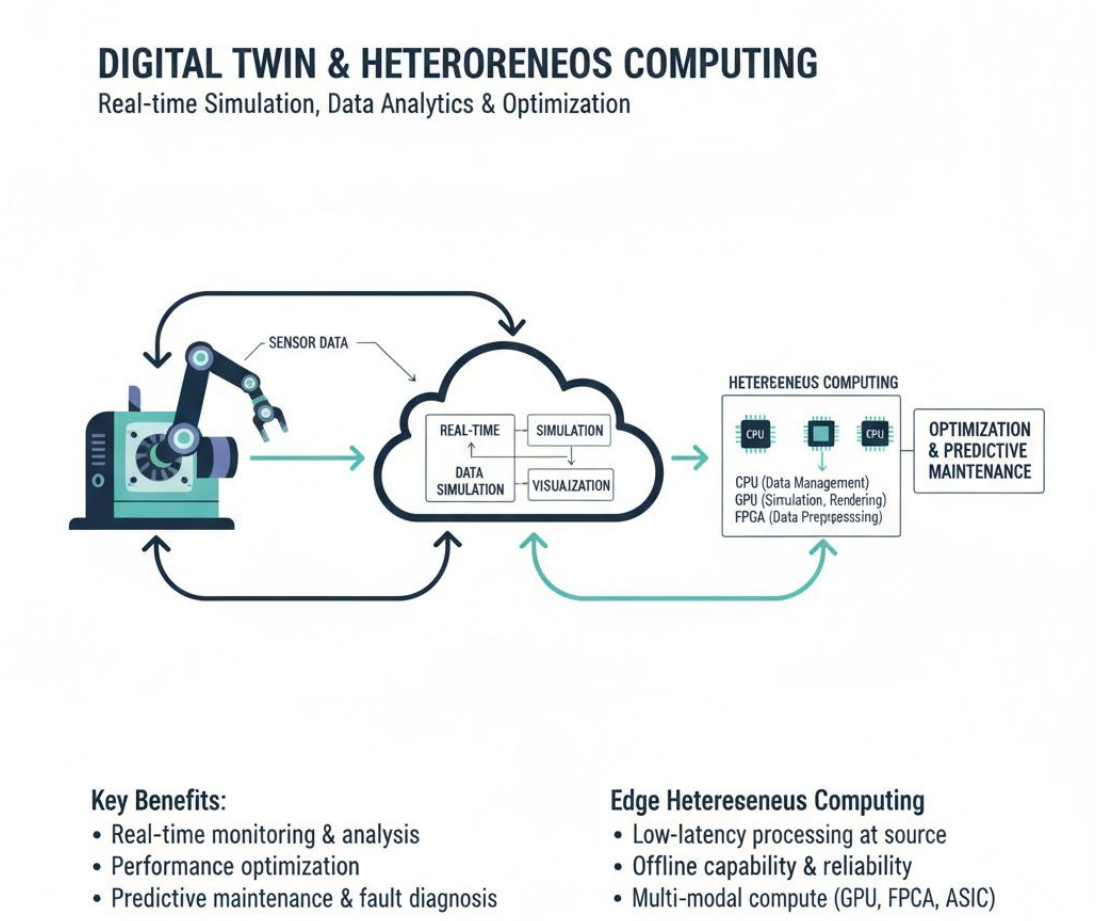
工业质检是 AI 技术在制造业中的重要应用，通过视觉检测技术自动识别产品缺陷，提高质检效率和准确性。AI 工业检测是利用基于深度学习、大模型等 AI 技术的视觉检测技术，在工业生产过程中对产品图像进行视觉检测，从而帮

助发现和消除缺陷。通过大模型技术，工业质检智能化已成数字化转型的核心战场，通过 3D 视觉+AI 算法实现检测效率提升 300%，在汽车零部件、家电、半导体等行业都有成功应用。



工业质检与 AI 视觉检测

视觉质检大模型是工业质检的最新发展方向。与传统的小模型相比，大模型具有更强的泛化能力和更高的检测精度，能够适应更复杂的产品缺陷和更多的生产场景。在实际应用中，视觉质检大模型通常需要针对特定行业和产品进行微调，以适应不同的检测需求。异构算力在视觉质检中主要支持图像预处理、特征提取、缺陷分类等计算密集型任务，通过 GPU 等专用加速器，可以实现实时的检测速度，满足生产线的高节拍要求。



数字孪生与异构算力

数字孪生是智能制造的另一项关键技术。数字孪生是指充分利用物理模型、传感器、运行历史等数据，集成多学科、多尺度的仿真过程。通过构建产品数字孪生模型，通过实时采集来分析产品运行、工况和环境数据，监控物理产品运行状态，以及进行功能、性能衰减分析，从而对产品效能分析、寿命预测、故障诊断等提供支持。

异构算力在数字孪生中主要支持实时仿真、数据分析和可视化等任务。数字孪生系统需要实时处理大量传感器数据，更新仿真模型，并进行可视化展示，计算量巨大。异构计算架构通过不同类型处理器的协同工作，能够高效处理这些多样化的计算任务。例如，可以使用 GPU 进行物理仿真和渲染计算，使用 CPU 进行数据管理和业务逻辑处理，使用 FPGA 进行传感器数据采集和预处理，通过合理的任务分配，实现系统整体性能的最优化。

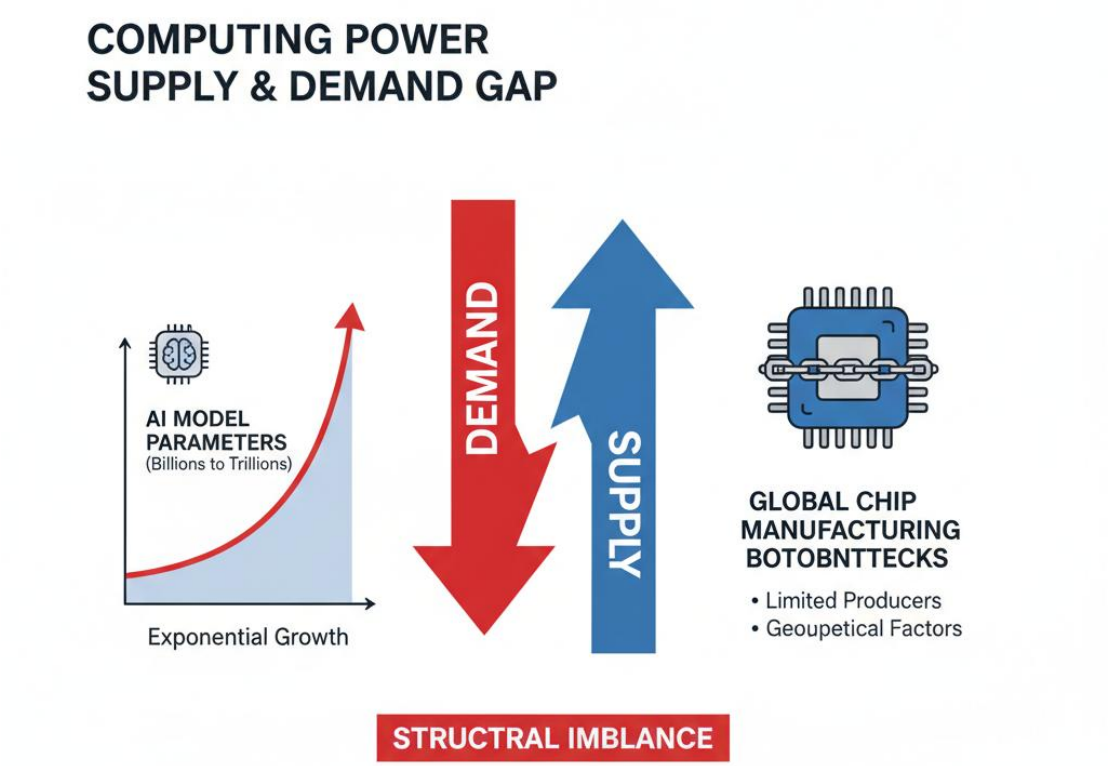
边缘异构算力在工业场景中具有特殊价值。工业场景通常对实时性、可靠性和安全性有高要求，边缘计算通过在工业现场部署计算能力，避免了数据传输到云端再返回的延迟，满足了实时性要求。同时，边缘计算也可以在网络连接不稳

定或断开的情况下保持基本功能，提高了系统的可靠性。在工业场景中，异构算力通过整合 GPU、FPGAs、ASICs 等不同计算单元，形成多模态算力供给，满足工业智能化进程中多样化的计算需求。

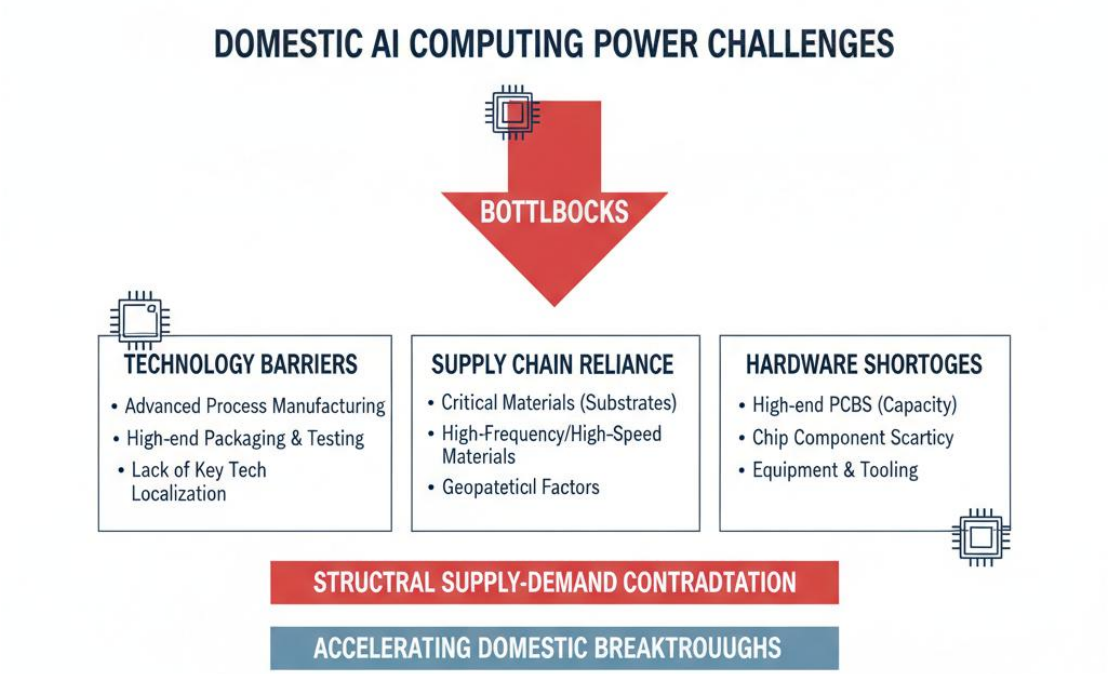
七、挑战、趋势与展望

7.1 主要挑战

7.1.1 算力供给与需求缺口



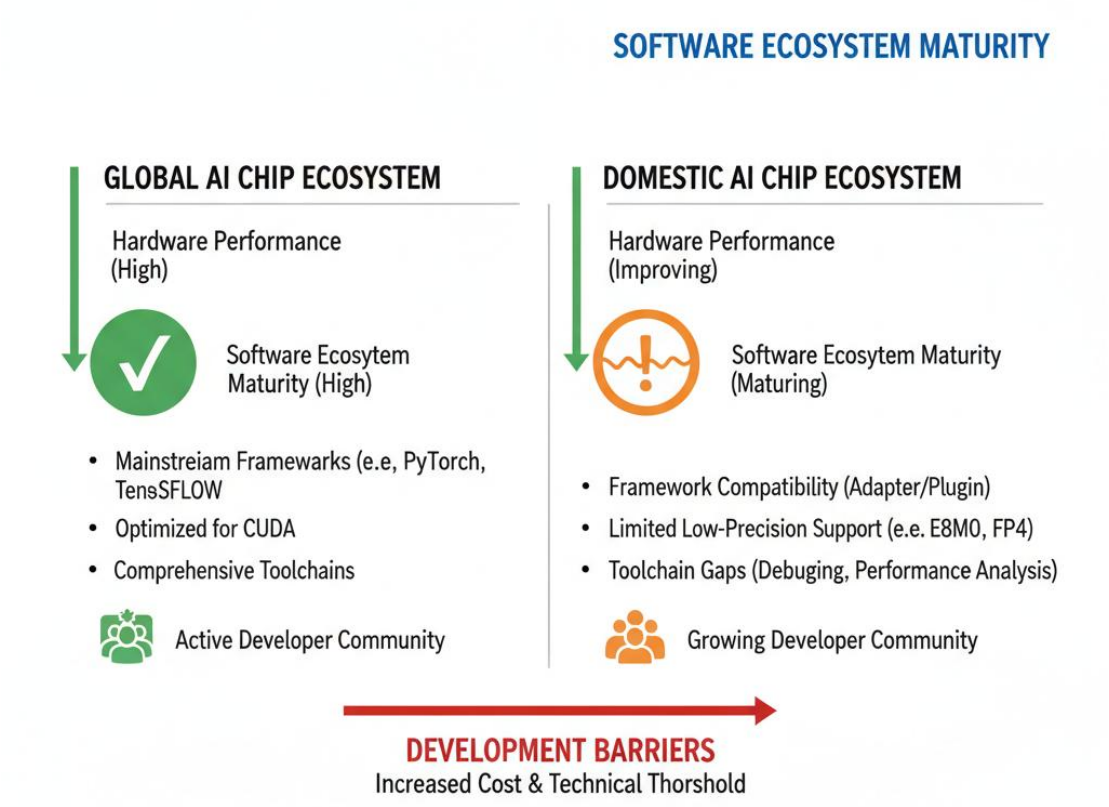
当前，AI 大模型与异构算力融合发展面临的首要挑战是算力供给与需求之间的巨大缺口。随着大模型参数规模从千亿级迈向万亿级，训练算力需求呈现指数级增长，而高端芯片产能却面临严重瓶颈。一方面，全球仅少数厂商具备稳定量产高端 AI 芯片的能力，技术壁垒导致产能短期难以填补需求真空；另一方面，地缘政治因素加剧了供应链不确定性，使得算力供给更加紧张。



国产替代进程中的技术壁垒是另一大挑战。尽管中国在 AI 芯片设计领域取得了显著进展，但在先进制程制造、高端封装测试等环节仍存在明显短板。现阶段，中国在算力供给方面仍存在关键技术国产化水平不足、应用支撑多样化能力欠缺等问题，亟待加快推动数据中心相关芯片的核心技术攻关，以逐渐形成产业自主可控能力。

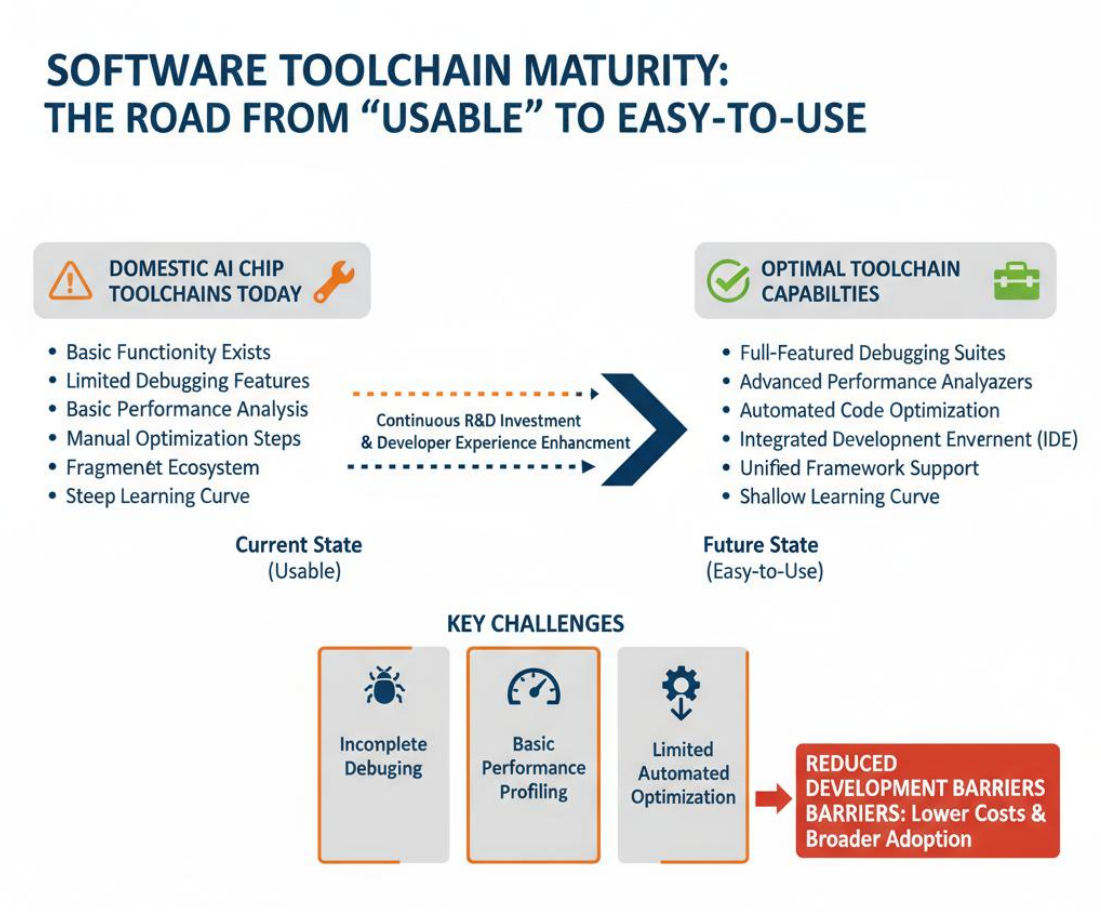
高端 PCB（印刷电路板）等配套材料的供给短缺也是制约因素。AI 硬件升级引爆了高端 PCB 需求，供应链已现缺口。供给端面临产能瓶颈，一是技术壁垒制约，高端产品全球仅少数厂商具备稳定量产能力；二是原材料供应受限，特殊基板材料、高频高速材料等关键材料对外依存度高。这些因素共同导致了算力供给与需求之间的结构性矛盾。

7.1.2 软件生态成熟度



软件生态成熟度是制约异构算力广泛应用的另一大挑战。尽管国产 AI 芯片在硬件性能上逐步追赶国际领先水平，但其软件生态与主流开发框架兼容性不足，开发者需针对不同芯片重新编写代码，这大大提高了开发成本和技术门槛。以昇腾、寒武纪等为代表的国产 AI 芯片，虽然性能不断提升，但在软件栈的完整性、工具链的易用性、开发社区的活跃度等方面仍与国际领先水平存在差距。

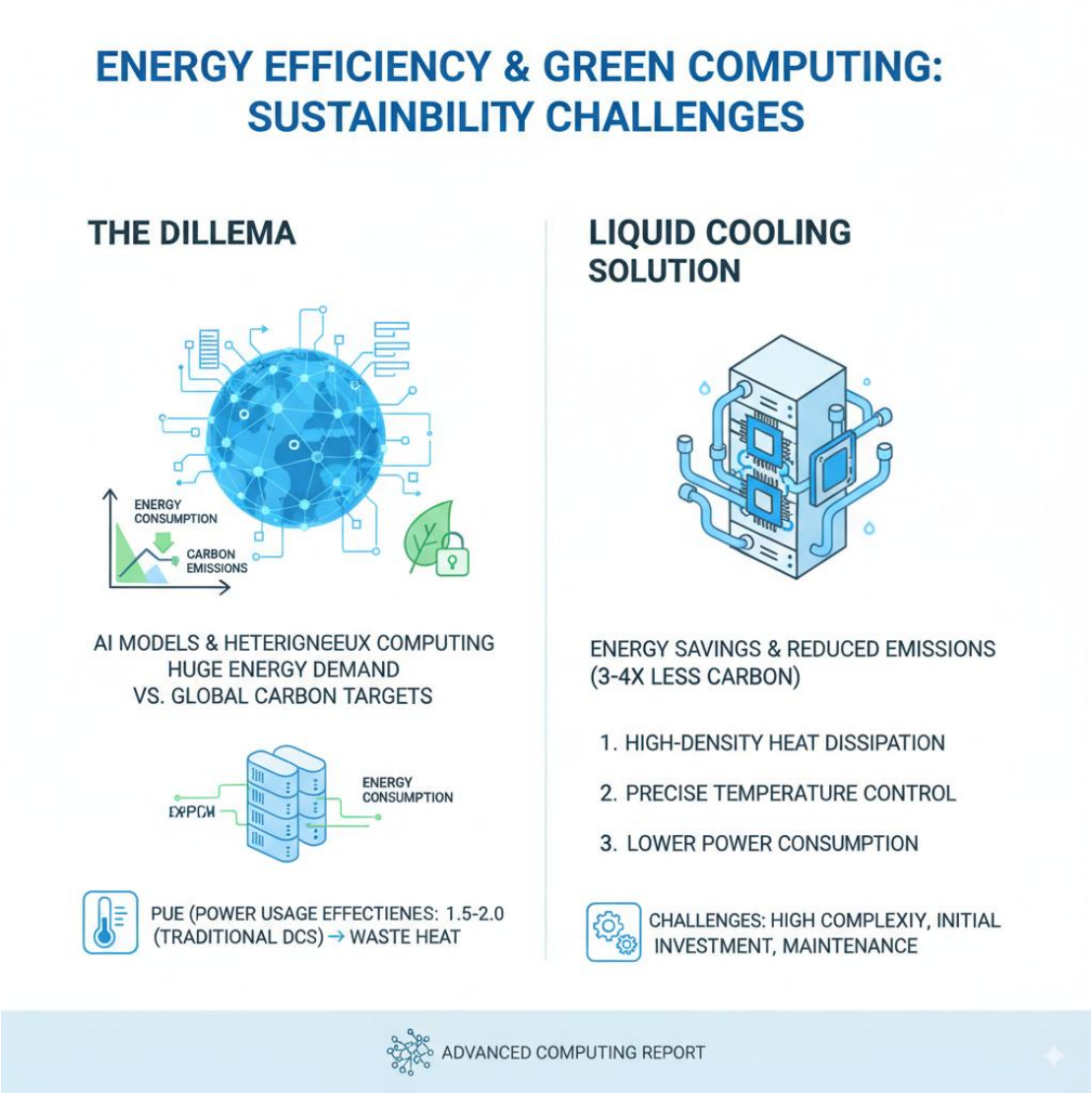
国产芯片软件栈兼容性问题尤为突出。目前，主流 AI 框架如 PyTorch、TensorFlow 等主要针对 NVIDIA CUDA 生态进行优化，国产芯片需要通过适配层或插件机制才能支持这些框架，这不仅影响了性能，也增加了开发复杂度。例如，框架 dtype 与编译工具支持未完全成熟：PyTorch 核心层面对某些基础类型（如 E8M0、FP4）的支持仍在推进中，这限制了新型低精度计算在国产芯片上的应用。



开发者工具链的完善度也是关键挑战。工具链的成熟度直接决定开发者的选择。虽然国产 AI 芯片软件生态从框架到工具链日趋完善，成熟度显著提升，华为昇腾 CANN 全面开源，但与国际领先水平相比，国产芯片的工具链在调试功能、性能分析、自动化优化等方面仍有不足。从"能用"到"好用"，国产芯片工具链还有很长的路要走，需要持续投入研发资源，降低开发门槛，提升开发者体验。

7.1.3 能效与绿色计算

能效与绿色计算是 AI 大模型与异构算力融合面临的可持续发展挑战。大模型训练和推理的巨大能耗与全球"双碳"目标形成矛盾，如何降低 AI 系统的能耗，实现绿色计算，成为行业必须解决的问题。数据中心电能利用效率（PUE）是衡量绿色计算水平的关键指标，传统数据中心的 PUE 值通常在 1.5-2.0 之间，意味着大量的能源被消耗在散热等非计算任务上。



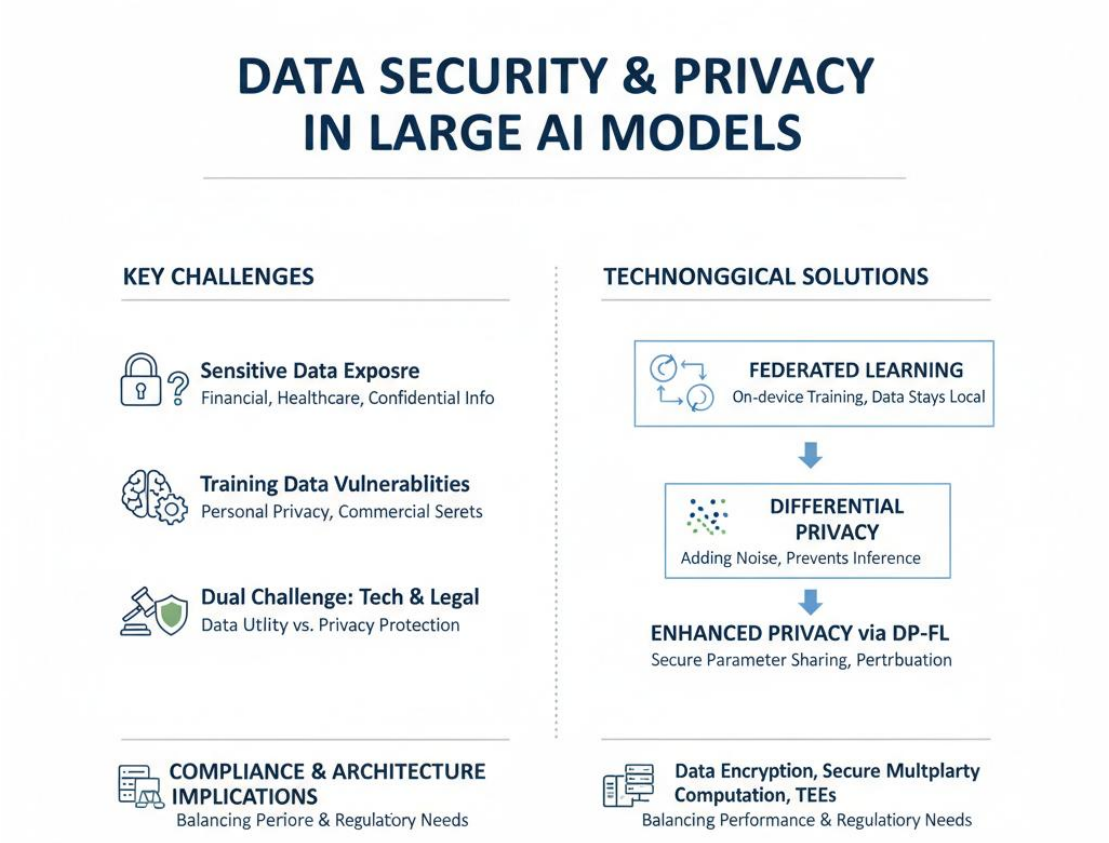
液冷技术是降低数据中心能耗的重要手段。相对于直接用电制冷散热，采用液冷技术不仅节约能源消耗，而且还有效地减少用电制冷过程中 3~4 倍的碳排放，从而达到算力运营的绿色化。液冷可提高处理高密度热量的效率，实现精确的温度控制并降低能耗。然而，液冷技术的推广面临技术复杂度高、初期投资大、运维难度增加等挑战，需要产业链各方共同努力推动技术成熟和成本下降。

可再生能源的应用是绿色计算的另一重要途径。绿色数据中心还经常使用太阳能和风能等可再生能源，通过提高 PUE 值和增加可再生能源比例，数据中心显著降低了碳足迹，并重视废旧电子设备的回收再利用。然而，可再生能源的间歇性和不稳定性给数据中心供电带来了新挑战，需要配备储能系统和智能能源管理系统，确保供电的稳定性和可靠性。

7.1.4 数据安全与隐私保护

数据安全与隐私保护是 AI 大模型应用中的关键挑战，特别是在金融、医疗

等敏感领域。大模型的训练需要大量数据，而这些数据往往包含个人隐私和商业机密，如何在利用数据价值的同时保护隐私安全，成为技术和法律层面的双重挑战。



联邦学习与差分隐私是应对这一挑战的重要技术手段。联邦学习允许在数据不出本地的情况下进行模型训练，有效保护了数据隐私；差分隐私则通过在数据或模型中添加噪声，防止个体信息被推断出来。在联邦学习的框架下，使用差分隐私技术可以进一步增强对数据隐私的保护。例如，在模型参数的共享过程中，可以采用差分隐私算法对参数进行加密和扰动，以防止攻击者通过分析参数反推出原始数据。

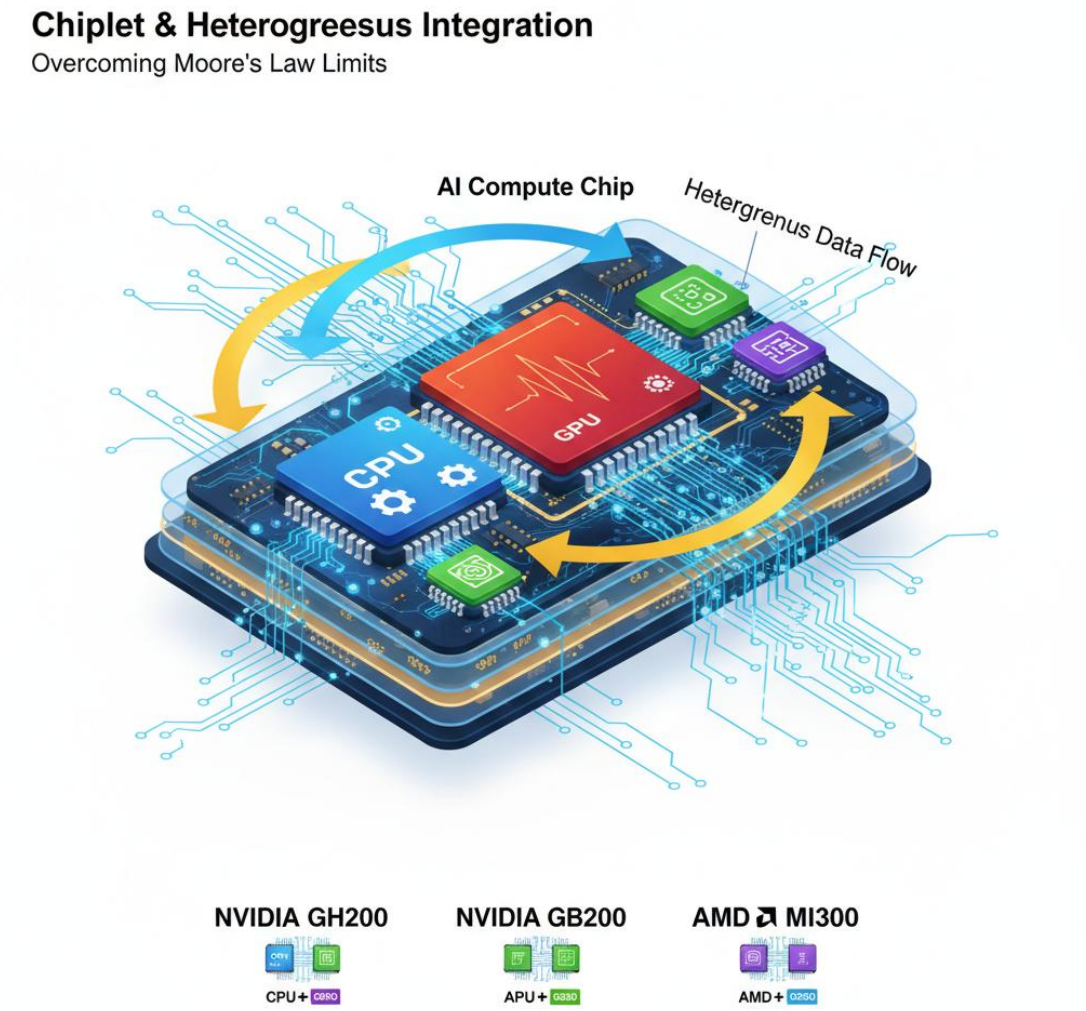
然而，联邦学习与差分隐私技术在实际应用中面临诸多挑战。联邦学习的通信开销大、收敛速度慢，特别是在大模型场景下更为明显；差分隐私则需要在隐私保护和模型精度之间做出权衡，过强的隐私保护可能导致模型性能下降。此外，这些技术的安全性也需要持续验证，新型的隐私攻击手段不断出现，需要相应的防御技术进行应对。

合规性要求对算力架构也提出了新要求。随着《数据安全法》《个人信息保护法》等法律法规的实施，AI 系统的数据处理和模型训练必须符合严格的合规

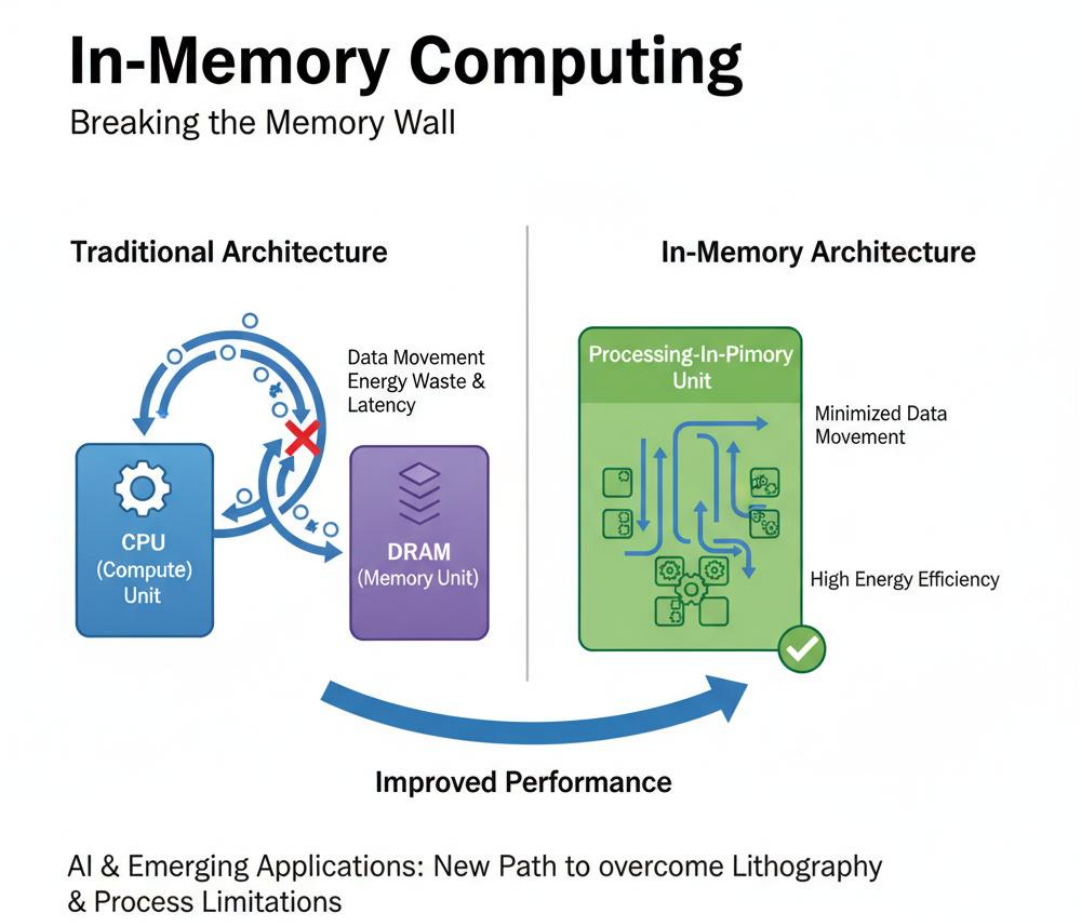
要求。这可能影响算力架构的设计，例如需要支持数据加密、安全多方计算、可信执行环境等功能，增加了系统复杂度和性能开销。如何在保证合规性的前提下维持系统性能，是算力架构设计面临的新挑战。

7.2 技术趋势

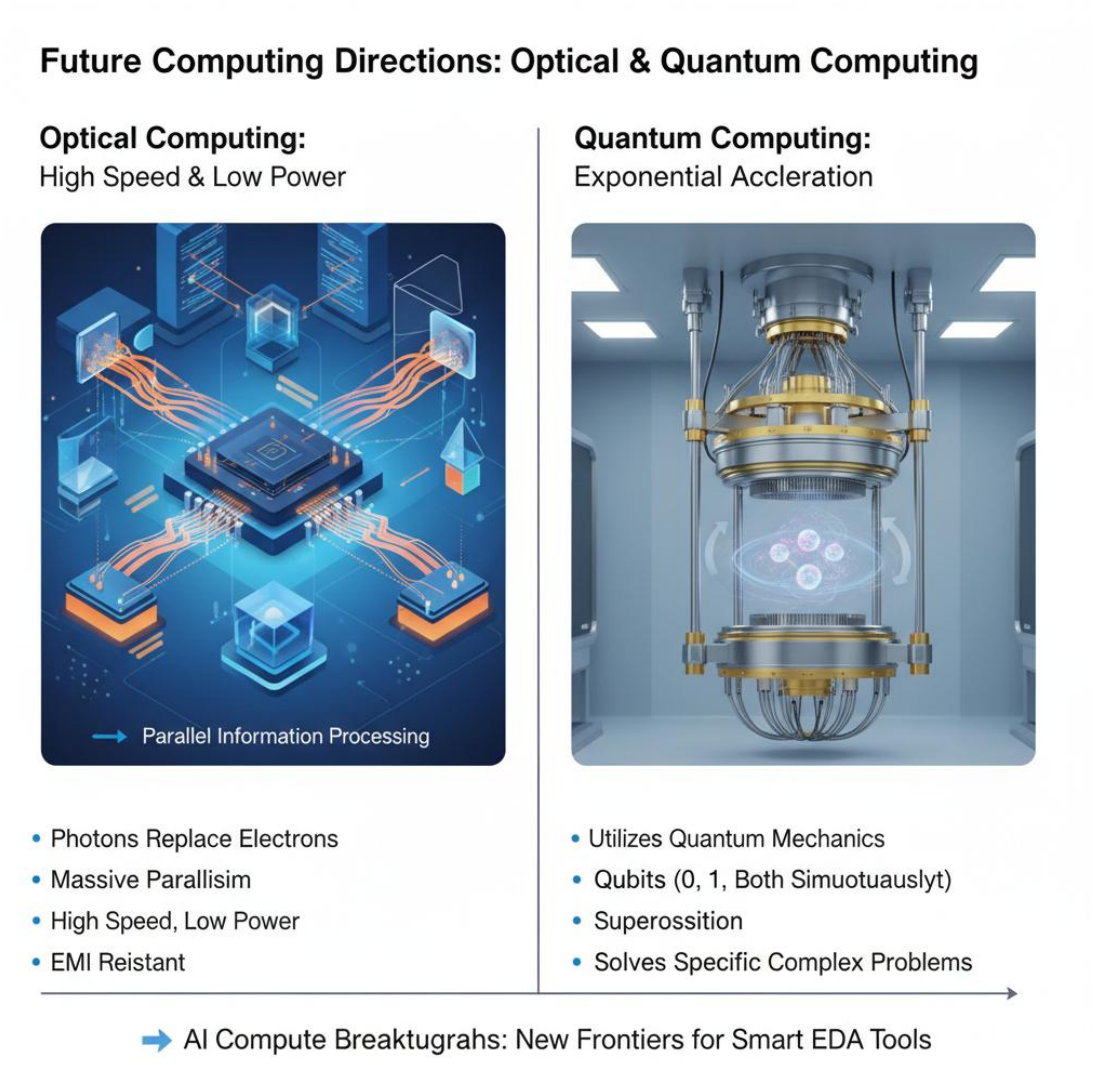
7.2.1 芯片与封装技术



Chiplet 与先进封装技术是应对摩尔定律放缓的重要趋势。Chiplet（芯粒）技术允许将不同功能、不同工艺制造的小芯片通过先进封装技术互联形成大芯片，将大面积芯片成本从晶圆制造环节转嫁到封装环节，提升大面积芯片良率。英伟达 GH200、GB200 和 AMD MI300 均采用 CPU+GPU Chiplet 异构方案，异构集成为算力芯片发展趋势。Chiplet 异构集成含有异构和异质两重含义，为 AI 算力芯片提供了新的发展路径。



存算一体技术是突破内存墙的关键方向。传统计算架构中，数据在存储单元和计算单元之间频繁移动，造成大量能耗和延迟。存算一体技术通过在存储单元中集成计算功能，大幅减少数据移动，提高能效比。近年来，面向人工智能等新兴领域，采用存算一体、模拟计算、数字化模拟射频电路、芯粒集成等新途径有望突破芯片光刻面积的极限和工艺制约，为算力提升开辟新道路。

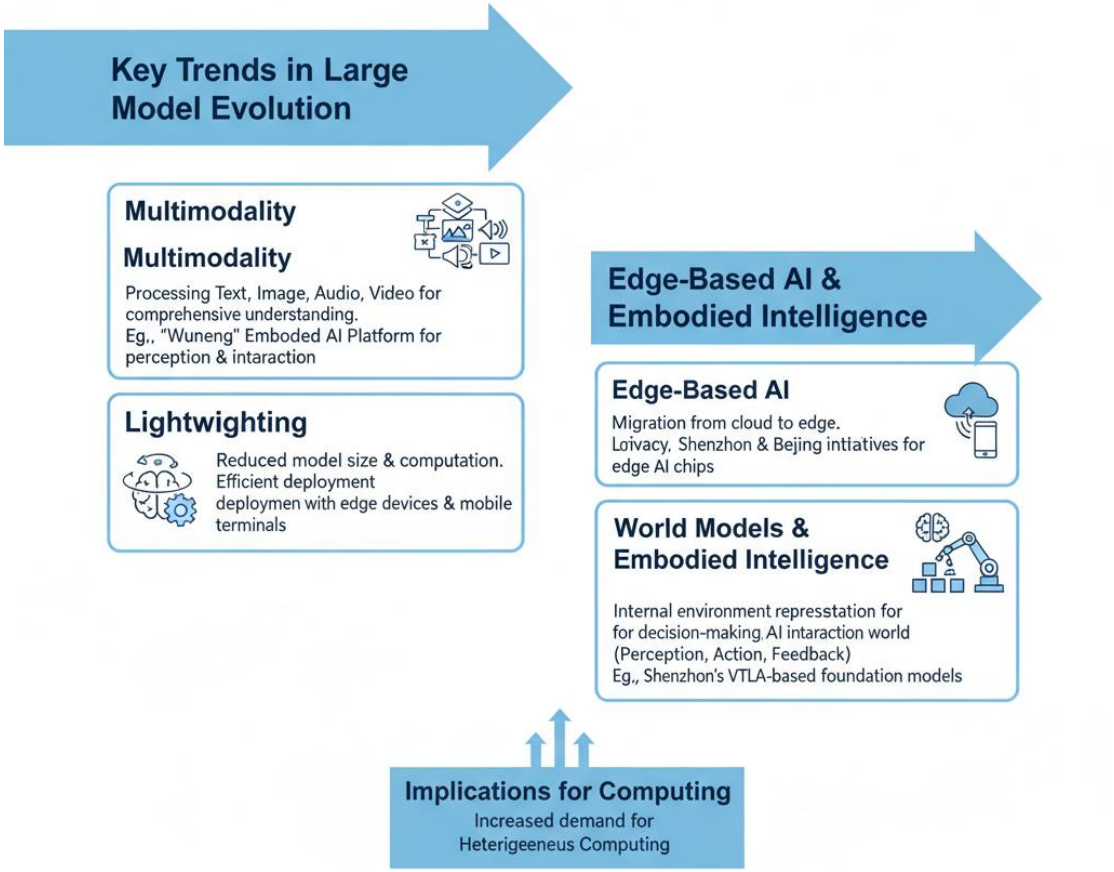


光计算与量子计算代表了更远期的技术方向。光计算利用光子代替电子进行信息处理，具有高速、低功耗、抗电磁干扰等优势，特别适合大规模并行计算；量子计算则利用量子力学原理，在特定问题上具有指数级加速能力。虽然这些技术目前仍处于早期发展阶段，但它们为解决 AI 算力瓶颈提供了全新的思路。发展面向先进集成技术，量子计算与光计算的智能 EDA 工具，将成为未来芯片设计的重要方向。

7.2.2 大模型技术演进

多模态与轻量化是大模型技术演进的重要趋势。多模态大模型能够同时处理文本、图像、音频、视频等多种类型的数据，实现更全面的理解和生成能力。商汤科技发布的"悟能"具身智能平台以商汤具身世界模型为核心引擎，依托商汤大装置提供端侧和云侧算力支持，能够为机器人、智能设备提供强大的感知、视觉导航及多模态交互能力。轻量化则关注如何在保持模型性能的同时减小模型规模

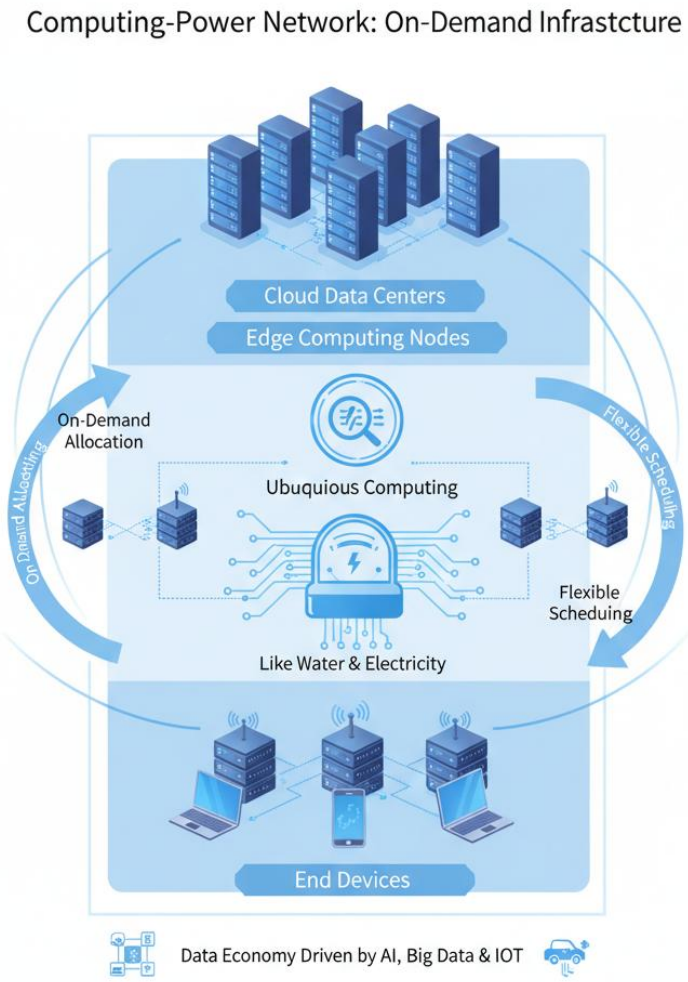
和计算需求，使大模型能够在边缘设备和移动终端上高效运行。



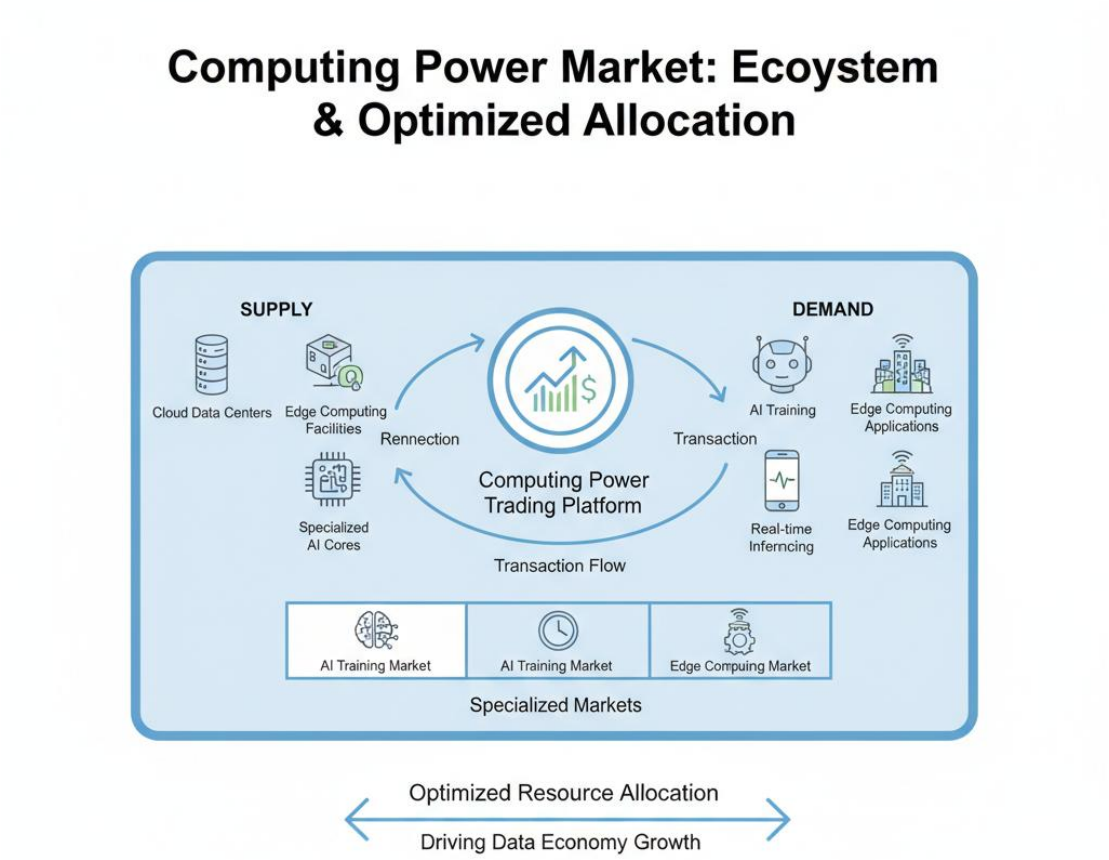
端侧化是另一重要趋势。随着边缘计算能力的提升，大模型正从云端向边缘端迁移，实现更低的延迟和更好的隐私保护。深圳和北京等地纷纷发布具身智能机器人技术创新与产业发展行动计划，强调研制机器人端侧计算芯片及模组，推进国产化替代。端侧大模型面临的主要挑战是计算资源和能耗限制，需要通过模型压缩、量化、剪枝等技术，以及专用 AI 芯片的硬件加速，实现高效部署。

世界模型与具身智能代表了更前沿的发展方向。世界模型旨在构建对环境的内部表征，实现更智能的决策和规划；具身智能则强调 AI 系统与物理世界的交互，通过感知、行动和反馈实现智能行为。深圳市科技创新局印发的《深圳市具身智能机器人技术创新与产业发展行动计划（2024-2026 年）》提出，基于世界模型及视觉-触觉-语言-动作（VTLA）等多模态输入输出，构建具备交互、预测与决策能力的具身智能基座大模型及其训练、推理技术体系，形成长序列推理能力。这些新兴方向将对算力提出更高要求，推动异构计算技术的进一步创新。

7.2.3 算力网络与交易



算力网络是未来算力基础设施的重要形态。算力网络就是一种在云、边、端之间按需分配和灵活调度计算资源、存储资源以及网络资源的新型信息基础设施。随着互联网、大数据、云计算、人工智能、区块链等技术创新，数据经济的发展将推动海量数据产生，数据处理需要云、边、端协同的强大算力和广泛覆盖的网络连接。算力网络的目标是让算力像水电一样成为按需调度的基础设施，为各类应用提供无处不在的计算能力。



算力交易市场的发展将促进算力资源的优化配置。随着算力需求的多样化，不同类型、不同规模的算力资源需要通过市场机制进行高效分配。算力交易平台将连接算力提供方和需求方，通过价格信号调节供需平衡，提高资源利用效率。未来，我们可能会看到更加专业化的算力交易市场，如针对 AI 训练的算力市场、针对实时推理的算力市场、针对边缘计算的算力市场等，形成更加精细化的算力交易生态。

算网平台（https://sumw.com.cn/）目前整合了国内外的各地区的算力资源，将单点的智算机房汇聚到平台形成算力网络，供个人和企业用户进行灵活购买和使用，最小可实现按小时级进行购买使用，让算力使用更加高效。同时个人和组织也可将闲置的算力接入到算网平台进行售卖，让算力高效分配。

7.3 产业与生态展望

7.3.1 国产异构算力产业链

国产异构算力产业链正在加速完善，形成从芯片设计、制造、封测到软件、应用的全链条生态体系。长三角、珠三角、成渝地区将形成三大数字 IC 产业集群，涵盖设计、制造、封测、设备、材料等全产业链环节。例如，某产业园区通

过"链主企业+配套企业"协同模式，构建了完整的产业生态，提高了整体竞争力。AI 算力芯片行业的核心为芯片设计和芯片制造，芯片设计工具厂商、晶圆代工厂商与封装测试厂商为 AI 算力芯片提供了研发工具和产业支撑。

在芯片设计环节，国内已涌现出一批具有竞争力的企业，如寒武纪、昇腾、海光、壁仞等，在 AI 芯片架构创新方面取得突破；在制造环节，虽然先进制程仍存在差距，但在成熟制程和特色工艺方面已具备一定能力；在封装测试环节，国产光学检测、离子注入等方面取得突破，晶圆级封装、3D 封装和测试设备已应用于先进封装产线；在软件和 EDA 工具方面，华大九天等企业已开发出部分 EDA 工具，覆盖了从设计到产业应用的完整链条。

政策支持是国产异构算力产业链发展的重要推动力。国家层面出台了一系列政策支持 AI 芯片和算力基础设施发展，地方政府也通过产业基金、人才政策、应用示范等方式支持本地产业发展。随着行业需求激增以及人工智能时代到来，2024 年我国半导体设备国产化率约为 50%，国产芯片自给率要达到 70%的目标正在稳步推进。

7.3.2 开发者生态繁荣

开发者生态是异构算力与大模型融合发展的关键支撑。随着国产异构算力产品的不断丰富，开发者生态建设日益重要。开源社区、培训认证、开发者大赛等形式是促进开发者生态繁荣的有效途径。华为昇腾社区、寒武纪开发者社区、算泥开发者社区等平台通过提供丰富的开发资源、技术支持和交流机会，吸引了大量开发者参与，形成了活跃的开发者社区。

培训认证体系是培养专业人才的重要手段。随着 AI 技术的快速发展，市场对 AI 开发人才的需求激增，但具备异构算力开发经验的专业人才相对稀缺。建立完善的培训认证体系，通过系统化的课程学习和实践项目，培养一批掌握异构算力开发技能的专业人才，对推动技术普及和应用落地具有重要意义。国内高校和企业已经开始合作开设 AI 芯片和异构计算相关课程，为产业输送人才。

开发者大赛是促进技术创新和生态建设的重要平台。通过举办面向异构算力和大模型的开发者大赛，可以激发创新活力，发掘优秀人才和项目，促进技术交流与合作。算泥社区等平台通过定期举办开发者大赛、技术沙龙、开源项目等活动，构建了开放、协作的开发者生态，推动了异构算力与大模型融合技术的创新和应用。未来，随着开发者生态的不断繁荣，我们将看到更多基于国产异构算力的创新应用涌现，推动整个产业的健康发展。

7.3.3 算力普惠与行业渗透

算力普惠是AI技术广泛应用的必要条件。当前，大模型训练和推理的高成本限制了技术的普及，特别是对中小企业和传统行业而言。推动算力普惠，降低AI技术的使用门槛，是实现AI技术广泛渗透的关键。算力普惠包括多个层面：一是降低算力成本，通过技术进步和规模效应降低算力价格；二是提高算力可及性，通过云服务、算力网络等方式让算力触手可及；三是简化使用难度，通过友好的开发工具和平台降低技术门槛。

行业渗透是算力普惠的具体体现。随着算力成本的降低和使用门槛的下降，AI技术正在向更多行业渗透，从互联网、金融等数字化程度高的行业，向制造、农业、医疗、教育等传统行业扩展。在制造业，AI技术被用于产品设计、生产优化、质量控制等环节；在农业，AI技术被用于精准种植、病虫害识别、产量预测等场景；在医疗领域，AI技术被辅助诊断、药物研发、健康管理等方面。这些传统行业的AI应用，往往对算力的成本、易用性、可靠性有更高要求，推动了异构算力技术的进一步优化。

算力普惠与行业渗透将形成良性循环。随着更多行业采用AI技术，算力需求将进一步增长，推动算力基础设施的规模扩张和技术进步，从而进一步降低算力成本，促进更广泛的应用。在这个过程中，异构算力通过提供多样化、高效率、低成本的算力选择，将发挥关键作用。未来，我们有望看到AI技术像电力、互联网一样，成为各行各业的基础设施，为经济社会发展提供强大动力。

八、附录

8.1 名词解释

8.1.1 异构计算

异构计算是指在同一计算系统集成不同类型或架构的处理单元，以便更有效地执行不同类型的任务。异构计算通过组合CPU、GPU、FPGA、ASIC等不同特性的计算单元，发挥各自的优势，实现更高的性能和能效。根据组合方式的不同，异构计算主要分为三类：CPU+GPU、CPU+FPGA和CPU+ASIC。异构计算的核心优势在于能够利用各类芯片的特点，针对不同计算任务选择最合适的处理单元，从而实现整体性能的最优化。

8.1.2 AI 大模型

AI 大模型是指参数规模巨大（通常在亿级以上）的人工智能模型，通过在海量数据上训练，具备强大的表示学习和泛化能力。大模型的参数规模从亿级到万亿级不等，训练算力需求每 6.2-10 个月翻一番，远超传统摩尔定律。大模型通常基于 Transformer 等深度学习架构，能够处理自然语言、图像、音频等多种模态的数据，在语言理解、内容生成、知识问答等任务上表现出色。代表性的大模型包括 GPT 系列、Llama 系列、Qwen、GLM 等。

8.1.3 训练与推理

训练与推理是 AI 模型的两个主要阶段。训练是指通过大量数据调整模型参数，使模型能够学习数据中的模式和规律的过程。大模型训练通常需要大规模计算集群支持，计算量大、耗时长、成本高。推理是指利用训练好的模型对新数据进行预测或生成结果的过程。推理更注重低延迟、高并发和能效比，在实际应用中的算力总需求已超过训练。训练和推理对算力的需求特点不同，需要采用不同的优化策略和硬件配置。

8.1.4 算力密度与能效

算力密度是指单位体积或单位面积内提供的计算能力，通常用 FLOPS/cm^3 或 FLOPS/cm^2 等单位表示。高算力密度意味着在有限空间内提供更强大的计算能力，对于数据中心和边缘计算场景尤为重要。能效是指计算设备在执行计算任务时的能源利用效率，通常用 TOPS/W（每瓦特功耗提供的万亿次运算次数）表示。高能效意味着在相同计算任务下消耗更少的能源，对于降低运营成本和实现绿色计算具有重要意义。PUE（Power Usage Effectiveness）是数据中心能效的重要指标，指数据中心总能耗与 IT 设备能耗之比，理想值为 1.0。

8.2 参考文献

8.2.1 国内外权威报告

[1] 中国信息通信研究院.《中国算力发展报告（2024 年）》

[2] 中国信息通信研究院.《先进计算暨算力发展指数蓝皮书》

[3] 中国信息通信研究院.《综合算力评价研究报告》

[4] 中国信息通信研究院.《中国绿色算力发展研究报告（2023 年）》

[5] 工信部等六部门.《算力基础设施高质量发展行动计划》

[6] 欧盟委员会.《欧洲芯片法案》（EU Chips Act）

[7] 美国国会.《CHIPS 法案》

[8] OpenAI.《GPT-4 技术报告》

[9] Meta.《Llama 2 模型报告》

[10] 寒武纪.《寒武纪年度报告》

8.2.2 学术论文与技术文档

[1] Sevilla, J., et al. "Compute Trends Across Three Eras of Machine Learning." arXiv preprint arXiv:2202.05924 (2022).

[2] Vaswani, A., et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[3] Raffel, C., et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." J. Mach. Learn. Res. 21 (2020): 1-67.

[4] Brown, T., et al. "Language models are few-shot learners." Advances in Neural Information Processing Systems 33 (2020): 1877-1901.

[5] Huawei. "Ascend CANN Developer Guide."

[6] Cambricon. "Cambricon Neuware Software Stack Documentation."

[7] NVIDIA. "NVIDIA TensorRT Documentation."

[8] TensorFlow. "TensorFlow Extended: An end-to-end platform for production machine learning."

[9] PyTorch. "PyTorch Documentation."

[10] Kubernetes. "Kubernetes Documentation."

8.3 致谢

8.3.1 行业专家与企业支持

本白皮书的编写得到了众多行业专家和企业的大力支持，在此表示衷心感谢。特别感谢寒武纪技术团队、华为昇腾团队在 AI 芯片和软件生态方面的宝贵资料；感谢阿里云、腾讯云等云服务商在智算中心建设和实践方面分享的经验；感谢开源社区在开发者生态建设方面的贡献。

同时，感谢中国信息通信研究院、中国电子技术标准化研究院等研究机构在行业数据和洞察方面的支持；感谢清华大学、北京大学、中国科学院等高校和科研院所的专家学者在技术理论方面的资料参考；感谢所有参与白皮书评审和提出宝贵意见的各位专家，你们的专业见解使本白皮书更加完善和权威。

8.3.2 开源社区与开发者

本白皮书也得益于众多开源社区和开发者的贡献。感谢 PyTorch、TensorFlow、MindSpore 等深度学习框架社区，为 AI 开发提供了强大的工具支持；感谢 MLIR、TVM、XLA 等编译器框架社区，推动了异构计算技术的发展；感谢 Llama、Qwen、ChatGLM 等开源模型社区，促进了大模型技术的普及和创新。

特别感谢算泥社区的贡献者和 MVP 专家们，你们在社区建设、技术分享、开发者支持等方面的辛勤工作，为 AI 大模型与异构算力融合技术的推广做出了重要贡献。感谢所有在 GitHub、Gitee 等平台上分享代码、文档和经验的开源开发者，你们的开放精神和协作态度是技术进步的重要动力。

最后，感谢所有关注和支持本白皮书的读者，你们的关注和反馈是我们持续改进的动力。我们期待与各方继续合作，共同推动 AI 大模型与异构算力融合技术的发展，为中国 AI 产业的自主可控和高质量发展贡献力量。