DOI 10.15302/J-SSCAE-2025.08.028

大模型时代我国高密度算力安全及发展研究

王志红[1]，龚振炜[1*]，武丽丽[2]，胡春卉[1]，石李妍[1]，程大宁[3]，安达[1]

（1. 中关村实验室，北京 100092；2. 中国工程院战略咨询中心，北京 100088；3. 中国科学院计算技术研究所，北京 100190）

摘要：近年来，以大语言模型（LLMs）和多模态大模型（MLLMs）为代表的生成式人工智能取得了显著进展。这些超大规模和复杂的大模型对计算资源提出了极高的要求，推动了高密度算力发展的迫切需求。本文从人工智能大模型技术发展的视角，探讨了大模型开发阶段和计算优化技术及其对算力需求的特点。围绕大模型对算力的需求，进一步剖析高密度算力的内涵与特征、发展现状及关键组成，并识别了我国高密度算力发展面临的五大关键挑战，包括供应链安全风险、物理硬件层瓶颈、软件栈不完整及高度依赖性、算力功耗与能源安全、网络安全风险等方面。对此，研究提出了未来我国高密度算力的主要发展策略，包括强化产业链自主可控、坚持自研与标准并举、构建开放统一软硬件生态、完善绿色算力创新体系、优化“学研”一体化创新体系等，以为我国高密度算力安全及发展提供参考。

关键词：高密度算力；智能算力；大模型；计算优化；安全发展

中图分类号：F420　　**文献标识码：**A

Security and Development of High-Density Computing Power in China in the Era of Large Language Models

Wang Zhihong[1], Gong Zhenwei[1*], Wu Lili[2], Hu Chunhui[1], Shi Liyan[1], Cheng Daning[3], An Da[1]

(1. Zhongguancun Laboratory, Beijing 100092, China; 2. Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088, China; 3. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In recent years, generative artificial intelligence represented by large language models (LLMs) and multimodal large language models (MLLMs) has achieved remarkable progress. These ultra-large-scale and complex models impose extremely high demand on computational resources, driving an urgent need for the development of high-density computing power. From the perspective of the technological advancement of LLMs, this study explores the development stages of LLMs, computational optimization techniques, and their characteristics in terms of computing power requirements. Focusing on the demand of LLMs for computing power, this study analyzes the implications and characteristics of high-density computing power, its current development status, and key components. It further identifies five major challenges faced by the development of high-density computing power in China, including supply chain security risks, bottlenecks at the hardware layer, incomplete software stacks with high external dependency, the power wall and energy security, as well as cybersecurity risks. In response, this study proposes several strategies for high-density computing power development in China, including improving self-supporting of the supply chain, adhering to self-dependent innovation and standardization, building

收稿日期：2025-08-19；**修回日期：**2025-10-20

通讯作者：*龚振炜，中关村实验室高级工程师，研究方向为网络空间安全战略；E-mail: gongzw@zgclab.edu.cn

资助项目：中国工程院咨询项目“网络空间安全新技术新应用风险研究”(2023-JB-13)，“算力安全及其产业高质量发展战略研究”（2024-HYZD-02）

本刊网址：sscae.engineering.org.cn

an open and collaborative software-hardware ecosystem, improving green computing innovation, and optimizing an integrated research and innovation system.

Keywords: high-density computing power; intelligent computing power; large language model; computational optimization; safe development

一、前言

近年来，人工智能（AI）领域取得了突破性进展，特别是以大语言模型（LLMs）和多模态大模型（MLLMs）等为代表的AI大模型再度兴起，深刻重塑了全球计算格局。这些模型以庞大的参数量（从数十亿到数万亿）和海量的训练数据集为特征，在自然语言处理、内容生成、机器人技术等多个领域展现出卓越的能力。中国信息通信研究院在2025年世界人工智能大会上发布的数据显示，截至2025年7月，全球AI企业数量已超过3.5万家，已发布大模型达3755个，其中我国AI企业的数量已超过5100家、已发布大模型1509个，大模型数量位居全球首位[1]。近年来，我国AI大模型产业规模持续壮大，形成了覆盖基础大模型、行业大模型、场景大模型及大模型应用的完整体系。

然而，大模型的超大规模和复杂性对计算资源或算力提出了极高的要求。算力也称计算能力，通常是指设备通过处理数据实现结果输出的能力[2]，是一种融合信息计算力、网络运载力和数据存储力的新型生产力[3]；其中，以一种高效的方式支持AI工作负载和应用的物理硬件与软件基础设施堆栈称为智能算力，且根据大模型等生成式AI不同研发阶段还可以细分为训练算力和推理算力。经济合作与发展组织（OECD）指出，算力可以根据访问位置分为以数据中心为代表的集中式算力、以云计算技术为基础的远端算力及边缘算力[4]。为了有效满足AI负载和应用产生的算力需求，智算中心的计算集群规模正在不断扩大，这种规模的扩大不是芯片在物理空间上的简单堆叠，而是要求这些芯片紧密协作并高效完成大模型训练和推理等任务。例如，X公司搭建的Colossus系统包含20万个AI芯片，但是，超级计算集群硬件成本和电力需求等均会随着规模扩展而快速增加，计算集群扩展规模始终有限[5]。因此，为满足大模型的快速发展，算力将朝向高密度算力的方向发展，并对传统算力中心产生深刻影响。

本文旨在从技术视角系统地探究AI大模型演进与高密度计算需求之间的复杂关系，深入探讨大模型技术演进的特点，以及其对高密度算力基础设施的设计、开发及部署等多方面的影响。通过对当前高密度算力的概念、内涵及关键组成部分进行分析，识别未来高密度算力发展面临的关键挑战和问题，从而提出我国高密度算力发展的策略。

二、大模型计算优化技术及算力需求演变

（一）大模型生命周期及其对算力需求的差异

大模型等生成式AI系统的生命周期主要包括：规划与设计、数据收集和处理、模型构建和使用、模型确认和验证、部署、系统运维与监控[6]。具体来看，生成式AI系统的开发流程包括：数据准备、预训练、对齐、评估和部署五个阶段[7]。数据准备阶段包括准备预训练数据和对齐数据，后者通常需要进行高质量人工标注；预训练阶段主要是选择和配置模型并进行自监督训练，从而得到一个基础模型；对齐阶段也称为后训练，主要包括使用对齐数据集进行微调、利用人类反馈等进行强化学习、测试时扩展等方式，以适应下游任务并符合人类意图；评估阶段通常对模型质量、安全性等方面进行多方面评估；部署阶段也称为推理阶段，是将模型以对话等形式进行交互，满足任务需求。生成式AI系统开发阶段如图1所示。

生成式AI系统的计算需求会随着其生命周期的发展阶段呈现显著变化。在数据准备阶段，数据量通常达到Pb级，需要庞大的存储空间，采集后的数据需要进行去重、过滤、纠错、分词等一系列操作，涉及大规模的数据转换等，对计算系统中的输入／输出（I/O）和存储架构提出了严格要求。预训练是大模型生命周期中对计算资源要求最严苛的阶段，涉及大量密集的矩阵乘法运算且需要在数千个分布式计算节点之间进行同步通信以更新梯度，对计算、内存带宽和网络等要求极高，需要实施多种高度并行策略，最大化计算资源的利用率。

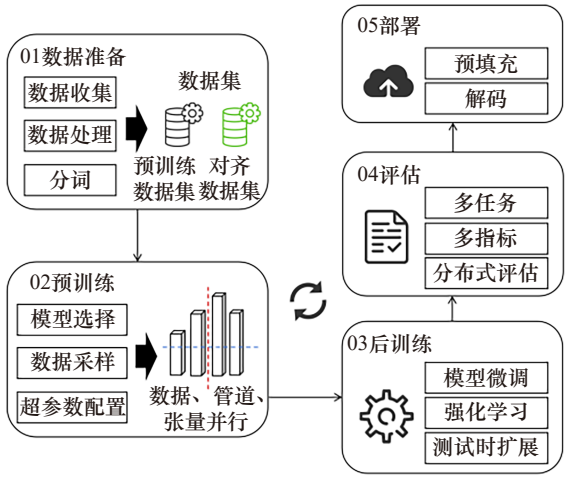

图1 生成式AI系统开发阶段

后训练阶段的算力需求根据技术方案不同而呈现较大差异。例如，用对齐数据集对模型进行全参数微调需要更新模型的所有权重，需要强大的算力支撑，而一些参数高效微调方法则可以显著降低算力成本。推理阶段包括预填充和解码两个子阶段，在预填充阶段，大模型需要处理输入文本并转换为键值缓存，涉及大量可并行化的矩阵运算，对计算要求高但对内存带宽要求较低；解码阶段需要根据键值缓存信息通过自回归方式进行推理并输出标记，对计算要求低但对内存带宽要求较高，以高效调取键值缓存。

（二）大模型扩展定律演进：从参数到数据再到推理时计算的算力需求升级

大模型扩展定律，也称神经网络扩展定律，是用来描述神经网络性能随着模型规模、数据集大小、计算量或训练时间等关键因素的增加而变化的统计定律[8]。目前，大模型扩展定律已成为指导全球AI竞争中国家和企业算力战略布局的关键工具，有助于有效分配算力资源，加速前沿AI模型的开发周期，甚至影响AI领域的市场领导地位和国家安全。

训练阶段的大模型扩展定律经历了从以参数为中心到以数据为中心的转变。大模型训练时扩展定律最早由美国开放人工智能研究中心（OpenAI）的Kaplan等人于2020年提出[9]。研究发现，语言模型的性能随着模型大小、数据集大小和训练算力大小呈现幂律分布特征，并建议模型参数与数据（以标记计）的比例大致为1∶1.7。2022年，Hoffmann等[10]发布Chinchilla扩展定律并建议该比例为1∶20，后由Epoch AI研究院进行复现，结果与此类似[11]。此后，不同的研究者提出了不同的扩展定律比例。例如，Sadana等[12]进一步考虑推理需求，认为在预期推理需求较大的情况下以更小模型进行更长时间的训练会更具成本效益，建议模型参数与数据的比例为1∶190。杭州深度求索人工智能基础技术研究有限公司（DeepSeek）[13]研究揭示了最佳模型与数据扩展分配策略，尤其是数据质量对该策略的影响，建议该比例为1∶30。清华大学[14]和Llama 3团队[15]进一步强化了数据的重要性，认为数据规模平均应为模型规模的192倍甚至1875倍。

在推理或部署阶段，随着“慢思考”范式的出现，大模型测试时扩展逐渐引发广泛关注。测试时扩展是指为大模型推理过程提供更多计算资源，即允许大模型进行更长时间的思考来输出更好的答案。有研究指出，优化大模型的测试时计算可能比简单地增加模型参数更有效[16]。测试时扩展通常涵盖多种技术，包括并行策略、逐步演化、搜索推理、内在优化等[17]。例如，通过输出多条可能的答案并采用投票机制选出其中一条答案，或者通过给出推理过程并使用过程奖励模型对这些过程进行打分，从而引导模型给出更好的输出。这些策略意味着大模型推理需要更多的算力资源支持，也意味着可以通过扩展测试时计算，使模型在达到同等性能时所需的规模更小，从而更便于部署在边缘设备上进行应用。

目前，大模型扩展定律的相关研究仍在持续开展，其关注点从模型参数到数据再到测试时计算的演进，导致大模型对算力需求也在不断发展变化。在以参数为中心的扩展定律指导下，业界认为模型参数的增加是模型性能提升的主要驱动因素，其次是数据规模。以数据尤其是高质量数据为中心的扩展，导致业界需重新评估和调整大模型的训练策略，使其对大规模、高质量数据集的需求显著增加，也对算力环境中数据处理、存储、网络等提出了更高的要求。测试时扩展的提出进一步对推理算力和边缘算力提出了更高的要求。未来，数据规模与质量、架构设计和算力三者之间的相互作用将继续推动大模型的相关研究。对于算力而言，这意味着在原始浮点运算能力的基础上，还需要优先考虑

数据高效摄取、高带宽存储以及计算资源灵活分配，以支持多样化的模型架构及其不断演进的数据需求。

（三）模型压缩技术：降低大模型部署门槛并推升推理算力需求

大模型的巨大规模和高计算需求对实际部署形成了严峻挑战，特别是在计算资源受限的环境中。因此，大模型压缩技术已成为一个关键的研究领域，旨在解决计算资源限制，并显著提升模型推理速度。当前的主要模型压缩技术包括模型量化、模型剪枝、知识蒸馏、混合专家（MoE）架构优化等类别[10]。

1. 量化技术通过降低精度来减少大模型内存和计算需求

量化是一种通过将模型的权重和激活值从高精度格式（如FP32）转换为低精度格式（如FP16、INT8、INT4等）来减少模型内存和计算需求的技术[18]。以一个70亿参数的模型为例，其内存需求在不同精度下差异显著，即FP32格式需要56 GB内存，FP16格式下可减少至28 GB，而INT8和INT4等低精度格式则可进一步压缩模型。主流的量化技术包括激活感知权重量化（如AWQ）[19]、后训练量化（如SmoothQuant）[20]以及高效后训练量化算法（如GPTQ）[21]等。量化技术显著降低了因大模型“内存墙”限制和高算力需求等带来的部署门槛，使其能够在除了大规模算力中心之外资源更加有限且多样化的环境中（如边缘设备和嵌入式系统）进行部署，并通过加速推理过程使大模型能够在延迟要求高等实时场景中进行广泛应用。此外，量化技术还将影响未来AI硬件的设计，推动越来越多的AI硬件集成用于低精度算术的专用单元。

2. 剪枝技术通过删除冗余参数来实现模型稀疏化以提升计算效率

剪枝技术主要通过移除模型中冗余或对性能影响较小的参数来实现稀疏性，通常包括非结构化剪枝、半结构化剪枝和结构化剪枝[22]。非结构化剪枝也称为权重剪枝，不遵循任何预定义的结构来移除模型中的单个权重，通常需要特殊的硬件支持才能实现实际加速[23]。半结构化剪枝则是删除符合某些结构约束的权重组，如每4个连续权重中删除2个。结构化剪枝是直接移除模型中的结构组件[24]，如神经元、注意力头等。已有研究还提出混合粒度剪枝技术[25]。结构化剪枝会产生更少的矩阵乘法和更低的内存需求，可以由标准的深度学习库和硬件加速器执行，对通用设备硬件更加友好。

3. 知识蒸馏通过解耦高成本训练和高效推理过程，降低模型推理的算力需求

知识蒸馏技术涉及将一个复杂的大型“教师模型”压缩至一个更小、更高效且更易部署的“学生模型”[26]。知识蒸馏的主要策略包括离线蒸馏、在线蒸馏和自蒸馏。离线蒸馏策略最为常见，是利用预训练好的“教师模型”指导“学生模型”；在线蒸馏是将“教师模型”和“学生模型”在端到端训练过程中进行同步更新；自蒸馏是将“教师模型”和“学生模型”作为同一个模型，包括在不同层次深度神经网络或不同阶段进行自蒸馏。苹果公司和牛津大学团队提出了知识蒸馏扩展定律，并揭示了不同算力预算和蒸馏情况下的算力资源分配策略[27]。知识蒸馏是通过解耦高成本训练和高效推理，降低推理成本、延迟和能耗，有效缓解在实际应用中部署大模型所面临的计算需求和资源限制相关的挑战，目前已在业界和学术界得到广泛应用。

4. MoE架构通过专家网络和门控网络动态分配算力，显著提升训练和推理速度

MoE架构是一种能够根据输入特点动态选择和组合多个“专家”（通常为各种神经网络模型）来处理特定任务或输入的机器学习模型架构[28]。专家之间的动态协调需要通过门控网络（或称路由机制）来控制。MoE架构主要包括专家网络和门控网络，专家网络一般为Transformer架构的前馈神经网络[29]，门控网络则决定与特定输入数据最相关的专家。前馈神经网络是在层归一化之后执行的稠密网络，通常是全连接层，其中所有的参数都会被激活。在使用MoE架构替代之后，前馈神经网络层就会被分割为多个专家网络，且在特定情形下仅激活一部分专家网络。通过这种稀疏化，MoE架构能够使算力分配动态化、显著提升训练和推理速度以及降低算力成本。此外，MoE架构允许在算力预算一定的情况下训练更高参数量的大模型，从而提升模型性能。目前，多个主流大模型已经采用MoE架构，如Mixtral（8x7B，每层由8个专家组成，每标记激活2个专家）[30]、DeepSeek-V3[31]（总参数量为6710亿，每标记约激活8个专家、370亿参数）等。

从推理角度来看，MoE模型的有效部署和高效执行在模型层、系统层和硬件层等多方面面临独特挑战。在模型层，MoE架构设计需要有效平衡模型规模和计算效率；在系统层，专家动态激活和负载均衡需要复杂的调度算法与高效的内存管理来处理这些专家参数的加载、卸载；在硬件层，传统为稠密模型进行优化的硬件架构难以匹配MoE模型特性，需要专门的加速技术进行处理，为动态专家切换等提供灵活的算力。因此，MoE模型的兴起，意味着需要在AI软硬件方面进行协同优化，包括新型硬件架构和加速策略、内存管理系统、适应负载均衡的智能中间件等[32]，以实现更灵活的算力供给。

（四）多模态大模型：复杂架构与数据需求驱动的算力多元化

与通用大模型或LLMs主要采用文本进行训练和部署不同，MLLMs使用文本、图像、音频和视频等多模态信息进行训练，旨在完成多模态、多样化的应用任务[33]。这意味着，MLLMs通常涉及更大规模和多样化的数据集以及更复杂的模型架构。例如，相比于GPT-3，GPT-4拥有高达1.8万亿和13万亿的参数规模和训练数据；OpenAI提出的CLIP模型为图像和文本设计了两个编码器[34]。这种多模态理解能力使大模型的应用范围得到不断拓展，尤其是在具身智能、人形机器人等技术的快速发展下，多模态大模型已经成为一个重要方向。

MLLMs训练和推理的复杂性从根本上推动了对底层算力基础设施的需求，并进一步推动算力基础设施向多元异构算力发展。一方面，MLLMs训练需要更大容量的存储、更高性能的计算加速卡以及更优化的并行工具和策略等，来处理规模不断增加的参数量和数据集；另一方面，MLLMs推理场景更加多样化，且对低延迟有更高要求，这对边缘算力、云服务等提出了更严苛的要求，进一步凸显了内存带宽、高速互连以及高效管理和处理多样化数据流的重要性。

综上来看，目前以大模型为代表的生成式AI发展呈现模型规模持续扩展、模型压缩及多模态等带来的架构复杂性和数据多样性，以及高度追求大模型研发与应用的低成本和高能效等特点，使传统算力基础设施难以有效满足当前需求。因此，未来算力基础设施需要朝向高密度算力发展。高密度算力需要通过对计算性能、内存、网络、功耗、散热等所有要素进行协同优化，实现单位空间、单位能耗和单位成本下的有效算力最大化，满足大模型研发中海量数据处理、超大规模参数计算和更新对智能算力的全方面需求，为大模型提供稳定、高效、低成本和可持续的算力支持。

三、高密度算力的内涵、特征及关键组成

（一）高密度算力的内涵与特征

当前，“算力密度”有多种定义，但其核心都指向计算资源的密集程度，用于衡量计算设备或数据中心的性能和效率。已有研究从多个维度对算力密度进行了阐释。

在空间维度，《AI大模型与异构算力融合技术白皮书》将算力密度定义为单位面积或单位体积内的计算能力，高算力密度意味着在有限空间内提供更强大的计算能力[35]。在“算力100问”中，算力密度是指在一定的物理空间或计算资源范围内，所能够提供的计算能力的大小[36]。在能耗维度，有研究认为，算力密度是单位面积或单位功耗下的算力，反映芯片设计的能效和集成度[37]，分别使用每平方米算力密度和单机柜功率密度来衡量。在硬件维度，高密度服务器旨在最小的物理空间内最大化计算能力和存储能力。美国芯片出口管制主要使用性能密度作为限制指标，通常为总处理性能除以芯片面积，目的是保证Transformer架构的大模型难以高效运行。有研究将算力密度定义为算力与显存容量的比值，并指出通用图形处理器（GPGPU）的单位显存算力相对有限，而应用集成电路（ASIC）以高算力密度在特定任务中优势凸显[38]。

现有对服务器密度和数据中心密度的定义通常会同时考虑空间和能耗两个维度。有研究指出，高密度服务器是指在最小的物理空间内最大化计算能力和存储容量的服务器硬件[39]，一般具备高密度处理器、内存和存储配置，以及高效的电源和冷却技术；数据中心密度是指单位面积内数据中心服务器机架及相应设施的功耗总和，与单机柜功率密度密切相关[40]。还有研究将新型数据中心总结为具有高技术、高算力、高能效和高安全4个特征的数据中心，即算力规模和密度逐步提高、绿色低碳技术应用逐步扩大等[41]。

综合来看，高算力密度意味着在有限的空间内可以提供更强的计算能力，从而提升整体性能和资源利用效率。这种每单位面积计算能力的指数级增长对互连、散热、配电等其他基础设施组件产生了连锁效应。因此，发展高密度算力不仅仅是简单地堆叠更多的服务器，而是要求对算力中心整体架构设计进行变革，涵盖从芯片设计与封装到电力输送、散热和网络等各个环节，都需要为AI工作负载进行专门优化。

（二）高密度算力的发展现状

在大模型等AI技术的快速发展及其对高密度算力的迫切需求之下，增加处理器核数、内存容量、存储空间及互连速度等方式提升单机柜功率密度已经成为实现高密度算力的一种常用策略[42]。2022年，中国信息通信研究院发布《数据中心白皮书（2022年）》，指出传统依靠增加空间、扩大机架及服务器规模来提供更多算力的做法在AI时代已经变得不可取，未来数据中心的变革趋势之一是高密度服务器研发部署加快，单位面积算力提升[43]。麦肯锡咨询公司指出，近年来单位机柜功率密度不断增加，类似ChatGPT规模的大模型训练每机架功耗超过80 kW，英伟达公司的最新芯片GB200及其服务器的机柜功率密度可以达到120 kW[44]。

当前，各大厂商正在纷纷围绕提升计算效率和降低能耗等进行高密度服务器产品的设计与布局。例如，浪潮电子信息产业股份有限公司为高密度数据中心设计的多节点模块化服务器i48M6，实现了在标准机架4U高度中部署8个计算节点和72块3.5″大容量硬盘，最高支持48个I/O扩展，并采用智能调控技术和先进分冷系统，保障系统的稳定运行[45]。百度在线网络技术（北京）有限公司于2025年宣布昆仑芯超节点支持1U4卡超高密度算力，单一机柜能放入64张卡[46]。值得一提的是，在2025年世界人工智能大会上推出的国产软硬件一体高密度算力机柜Shanghai Cube突破算力极限，其算力密度达单一标准机柜128张图形处理器（GPU）模组，包括芯片、存储、网络、管理节点等硬件，以及操作系统、计算平台、调度软件、AI平台等软件，实现了全国产自主可控[47]。

在互连技术方面，英伟达公司率先提出纵向扩展解决方案——超节点[48]，通过内部高速总线互连，建起低延迟、高带宽的统一算力实体，有效支撑并行计算任务。2025年，华为技术有限公司推出昇腾384超节点（Atlas 900 A3 SuperPoD）[49]，实现了384个AI芯片之间的大带宽低时延互联。同时，阿里云计算有限公司也在2025年的云栖大会上首次展示了支持144个节点的高密度AI服务器及高性能网络架构HPN8.0[50]。

（三）高密度算力的关键组成

建设高密度算力中心的关键是部署高密度服务器，包括在特定空间内集成更多的处理器和I/O扩展能力，提升存储密度、互连速度、按需内存扩容、提升冷负荷密度等[51]。高密度算力的核心组成包括基于先进封装的芯片制造技术、高速互联技术和高效散热技术等。

在先进封装技术方面，在摩尔定律放缓的背景下，以CoWoS为代表的2.5D封装技术成为延续芯片性能增长的关键技术之一。从单个芯片来看，AI芯片一般是专门针对AI工作负载进行特殊加速设计的芯片，可以分为GPU、现场可编程逻辑门阵列（FPGA）和ASIC等架构。以英伟达公司的GPU为例，从Pascal P100到Blackwell B100，其算力性能从19 TFLOPS上升到近20 000 TFLOPS，同时每单位功耗从170 000 J每标记下降到0.4 J每标记。从芯片封装来看，2.5D封装是指在芯片之间增加中介层来实现高密度互连的封装方式，具有多芯片集成及高密度的特点，其中CoWoS封装是将多颗芯粒通过晶圆上芯片（CoW）的封装制程连接至硅中介层，再通过基板（oS）与底层基板连接，构成整体CoWoS结构[52]。

在高速互连技术方面，由于大模型需同时跨越多个芯片和服务器集群进行分布式计算，互连带宽已成为关键瓶颈之一。从服务器集群来看，高速互连技术是将AI加速器集群连接起来，支持节点内和节点间的高速通信，实现模型高效运行。传统互连技术难以满足AI工作负载对高带宽和低延迟的需求，当前主要的高速互连技术包括NVLink、InfiniBand、PCIe、RoCE等。前两种都是英伟达公司的技术，极大提升了卡间互连的效率。

在高效散热技术方面，高密度算力带来了巨大的发热量，使散热问题成为硬件选择和基础设施布

局的关键因素之一。从算力中心来看，高密度算力的需求与发展为算力中心散热问题带来了极大的挑战，甚至决定了硬件选择、空间优化、电力输送等方面的整体布局。AI工作负载需要的高密度算力所产生的热量显著高于传统的算力服务器，如果没有有效的散热技术就会导致硬件过热从而产生故障。常见的散热技术包括风冷、水冷和液冷以及混合冷却等多种方式，液冷根据冷却方式又进一步细分为冷板式、浸没式和喷淋式冷却，后两者都属于浸没式[53]。在“双碳”目标下，传统风冷方式已经难以满足算力能耗要求的电源使用效率（PUE）值，液冷逐渐成为算力中心散热技术的优选。

四、我国高密度算力安全发展面临的关键挑战

（一）供应链安全风险仍然首当其冲

AI芯片供应链高度复杂且碎片化，集中在少数国家或地区，导致AI芯片供应链高度脆弱，极易受到地缘政治、突发事件等因素的影响，其安全性、韧性已成为国家安全和经济发展的当务之急。尤其是在当前的国际形势下，AI芯片领域供应链风险日益增长。2025年7月，美国发布《赢得人工智能竞赛：美国人工智能行动计划》，明确提出要审查我国的前沿AI模型，并在国际治理机构中对抗我国的影响力等。此前，美国更是提议安装“芯片定位”技术、联合盟友不断加强对我国芯片及相关设备和工具的出口管制等方式，持续加强对我芯片领域的打压，遏制我国前沿AI模型的发展。

当前，我国在高端制程芯片等领域存在多方面的挑战，包括以极紫外光刻机（EUV）及电子设计自动化（EDA）软件等为代表的软硬件设备、工艺技术、关键基础材料以及生产成本与稳定性等。7 nm及以下制程高端芯片的主要挑战在于在缩短制程、提高晶体管密度的同时还需要保持高效能和低功耗等。我国在芯片制造工艺上的研发起步较晚，技术储备不足，大量高端芯片需要依靠进口[54]，导致供应链安全存在极大的风险。

（二）物理硬件层瓶颈问题逼近极限

除了GPU等芯片外，内存和互联技术代表的存力和运力等广义算力基础设施的物理硬件层已成为制约当前AI发展的关键瓶颈，尤其是随着摩尔定律、登纳德缩放定律（Dennard scaling）和阿姆达尔定律（Amdahl’s law）等三大基本定律的逐渐放缓甚至消失，传统存算分离的冯·诺伊曼架构已经难以有效满足大模型等AI应用日益增长的性能和能效需求[55]。在内存方面，由于当前硬件制造工艺等正在逼近物理极限，“内存墙”问题凸显，成为制约大模型发展的根本性瓶颈之一。“内存墙”是指处理器速度与内存性能速度发展不平衡导致的一种存算失衡现象，最早于1994年由William. A. Wulf和Sally A. MaKee提出[56]。对于规模仍在不断扩展的大模型而言，“内存墙”会导致访存时延增加，严重降低计算效率。造成计算效率低下的另一根本原因是互联技术，包括芯片内部处理器与存储器以及芯片之间的网络带宽极大限制了数据的高效快速移动。通常，大模型训练需要多算力节点或算力集群进行分布式计算，而传统的网络系统无法满足AI工作负载对高带宽和低延迟的需求。如果互联速度慢或延迟高，计算资源就会空置等待，从而降低了大模型训练的效率，极大增加了训练时间和成本。

从存储芯片来看，我国存储芯片近年来取得了长足发展。国际知名机构Techinsight发布多个报告，通过逆向拆解我国长江存储科技有限责任公司和长鑫存储技术有限公司的产品，分析了我国存储芯片的发展现状。如分析长江存储科技有限责任公司232层QLC 3D NAND芯片，认为该芯片具有市售产品中最高的位密度19.8 Gb/mm^2，被认为是世界上最先进的3D NAND存储芯片[57]，对AI发展至关重要。长鑫存储技术有限公司也在动态随机存取存储器（DRAM）、高带宽内存（HBM）等领域取得了快速进步，突破了DDR5和HBM2技术[58]，但在HBM等技术与以美光科技有限公司的HBM3e为代表的国际先进技术仍存在较大的代差。在互联技术方面，单一集群的互联技术主要依赖于英伟达公司的NVLink、InfiniBand等技术，我国存在较大差距，但是近年来我国高度重视算力互联互通相关技术，正逐渐取得突破。如2025年上海人工智能实验室与中国联合网络通信集团有限公司联合实现千里算力互联训练千亿参数大模型。未来，如果能以算力网络互联弥补单颗芯片算力不足的短板，将有望降低我国智能算力存在的供应链风险。

（三）软件栈优化面临不完整及对外高依赖等多重挑战

高密度算力基础设施的瓶颈不仅局限于硬件，还延伸到基础设施软件栈的各个层面。这些基础设施软件栈也称为AI软件系统[59]，是连接硬件和AI工作负载的关键部分，包括操作系统及编排管理软件、AI编译器、分析器以及AI框架和软件库等，共同组成了AI软件生态。这些软件工具能够确保大模型充分利用底层硬件，减少训练模型所需的时间和资源，提升性能和能效[60]，是整个AI技术体系的核心，也是应对智能经济时代的技术利器。

目前，国际上正在积极开展AI框架软件开发，Cuda、MLIR、PyTorch等主流AI框架软件均由国外企业主导，我国正在这种国际垄断之下积极推动开源开放和自主可控。当前，我国智能算力软件生态仍面临着多重挑战，包括驱动软件的异构架构兼容与性能的折中、编程模型和语言高度依赖海外、加速库与工具链的完整性和效率不足等[61]。尤其是，当前我国国产软件的生态建设缺失，大量开源生态社区开发和使用人员较少，开源生态不成熟，导致我国国产AI软件栈优化速度慢，缺乏完整性并对外高度依赖。

（四）算力功耗面临“功耗墙”及能源安全双重挑战

智能算力的尽头可以认为是能源，算力与能源之前存在高度相关关系。尤其是，高密度算力的显著特点是单位面积内功率密度极高，这意味着在更小的空间内产生更多的热量，带来了高功耗及高散热需求。这种对空间密度的极致追求与功耗及能源需求之间的矛盾已成为限制算力密度持续提升的关键。这种巨大的能耗需求导致AI芯片的功耗过高，限制了AI技术的可持续发展和应用，也被称为“功耗墙”。解决“功耗墙”问题需要从能源政策、电力等基础设施规划、芯片工艺及AI算法设计等各方面进行提升和优化。国际能源署在2025年发布的《能源与人工智能》[62]报告中指出，一个典型智能算力中心的耗电量相当于10万户家庭的用电量，到2030年其耗电量将增加一倍达到945 TW·h，甚至高于全日本的总电力消耗，未来AI及算力中心的发展将加剧全球能源安全风险。

从能源基础设施方面来看，虽然我国目前的电力仍由煤电主导，但近年来正在大力发展太阳能、风能等绿色能源技术，并在光伏、电池等领域获得了重大进展，已跃升为全球领先的国家。截至2024年年底，我国可再生能源发电装机达到1.889×10^{9} kW，2024年发电量达3.46×10^{12} kW·h，占全部发电量的35%，这些绿电资源主要分布在我国广袤的西部地区[63]。当前，我国电力市场由国家电网有限公司和南方电网有限责任公司进行统筹，未来我国将建设全国统一电力市场，有利于对电力进行统一调度与使用。与此同时，我国出台“东数西算”等多项政策高度重视算电协同发展，但在绿电使用等方面仍面临一些技术、产业、政策等方面的问题，如算力用电集聚程度高、密度大对电网带来较大压力，绿电供应不足以及绿电就近由智算中心消纳存在障碍等。

（五）网络安全风险形势更加严峻

智能算力网络安全风险贯穿基础硬件、软件及模型部署与管理工具等层面。2025年8月，美国国家标准与技术研究院（NIST）发布了《用于保护人工智能系统控制覆盖》（SP 800-53 Control Overlays for Securing AI Systems）概念文件，指出AI系统的安全性与运行AI系统的基础设施安全性密切相关，为应对不同AI系统应用场景下的网络安全挑战提供操作性指南[64]。

基础硬件层面面临的风险主要来源于硬件设计缺陷、密码算法实现漏洞以及物理侧信道攻击等。以GPU为例，由于GPU拥有独特的内存系统，相比CPU会遭受更复杂的缓冲区溢出漏洞[65]，包括未实现地址空间布局随机化、未在释放内存后将其清零以及未初始化新分配的内存等[66]。例如，2025年1月，英伟达公司确认了7个影响其GPU的漏洞，导致内存损坏、代码执行、拒绝服务攻击、信息泄露或数据篡改等风险[67]。在基础软件层面，操作系统和AI框架是调度硬件资源的重要软件，其安全性与可靠性是AI系统稳定运行的基础。2024年，开源AI框架Ray被爆存在一个有争议的影子漏洞，攻击者可利用该漏洞接管相关组织的算力并泄漏敏感数据，导致全球数千台Ray服务器受到攻击[68]。2022年，知名深度学习框架PyTorch被发现有一个与框架“torchtriton”库同名、可运行恶意二进制文件的恶意依赖包，被上传到Python包索引（PyPI）

代码库[69]。此外，模型部署与管理工具是连接算力基础设施与大模型之间的重要桥梁，其安全问题也不容忽视。2025 年 3 月，知名开源大模型工具 Ollama 暴露存在文件泄漏和模型安全等多个严重漏洞，存在未授权访问与模型窃取等安全风险，对部署 AI 大模型的组织或个人构成严重风险[70]。

当前，我国 AI 算力基础设施网络安全风险主要来自于基础硬件的高度不可控带来的后门植入漏洞等，如近期英伟达公司的 H20 芯片被爆存在后门并被我国相关部门约谈；基础软件高度依赖国外开源软件框架，开源易导致软件质量低且对漏洞信息不公开等问题，加剧我国 AI 相关基础软件的网络安全风险[71]。

五、我国高密度算力安全发展策略

（一）强化产业链自主可控，保障算力供给安全

针对芯片供应链脆弱性等风险，应同时加大 AI 芯片硬件和软件关键核心技术研发力度。在硬件制造层面，需集中攻破先进制程工艺、互联、存储及存算一体等新兴架构的关键核心技术。在先进制程工艺上，应集中资源突破 7 nm 以下工艺，重点攻关高密度互联、EUV 等关键技术；在存储技术上支持加速三维堆叠技术研发，实现从 HBM2e 到 HBM3 的迭代跨越；在新兴架构上应支持大规模高密度存算一体介质、异构架构集成方法与工艺[72]等技术研发。在软件工具层面，应大力支持 EDA 软件工具链的开发和优化，建立 EDA 工具开源社区，推动 EDA IP 供应商与高校共建设计流程验证平台，形成覆盖芯片全生命周期的自主工具链。

（二）坚持自研与标准并举，强化安全保障能力

智能算力的基石仍是算力基础设施，其安全保障举足轻重，应坚持强化自研及开源技术研发和严格境外硬件准入标准双措并举。一是强化自主研发，持续加大对自研硬件的投入力度，并强制关键部件国产化率指标，构建多层级备份体系应对地缘政治风险。在金融、能源、国防等关键核心部门，积极推广基于开源 RISC-V 以及自研 LoongArch 等自主可控指令架构的芯片应用，坚决摆脱对国外核心技术的依赖，从根源上有效防止后门植入。二是严格准入标准，完善产业安全审查机制，建立智算中心分级准入制度，并延伸至供应链安全评估。针对现阶段仍无法被替代的英特尔公司、高通公司、英伟达公司等境外厂商芯片，应构建一套全面且严谨的硬件安全评测体系，将芯片设计、生产制造直至设备组装的各关键环节均纳入其中，进行严苛检测，确保进入市场应用的硬件均符合高等级安全要求。

（三）构建开放统一软硬件生态，释放全栈创新活力

为破解智能算力生态碎片化难题，应通过标准引领及开源策略构建体系性软硬件协同生态。在硬件层面，可以 RISC-V 为基础建立向量指令扩展国家标准作为兼容基础指令集，打破国产芯片厂商“各自为战”的分散局面。在软件层面，由头部企业牵头组建开源联盟，加大对跨架构 AI 编译器的开发和推广，大幅提升算子库覆盖率，解决框架适配问题；成立开源联盟并建立开源贡献激励机制，形成相关补助机制，对参与开源项目的企业和个人提供资金支持和技术指导，推动软硬件协同发展。

（四）完善绿色算力创新体系，赋能产业低碳化转型

为应对能耗挑战，需从架构革新及智慧能源供给等多个角度形成多维解决方案。一是从芯片设计等方面进行架构革新。发展存算一体、类脑计算等非冯·诺依曼架构，如阿里巴巴达摩院（杭州）科技有限公司研发的计算芯片通过近存计算降低数据搬运能耗达 60%。二是从散热、调度等技术和政策多方面实现智慧能源供给。在散热技术方面，加大浸没式液冷介质的研发力度，建立氟化液制备工艺中试平台，构建智慧冷却系统；在能源调度方面，借鉴“源网荷储”协同机制，在西部算力枢纽部署“风光储”一体化电站，通过智能调度大幅提升智能算力中心的清洁能源消纳率；此外，推行算力碳足迹追踪制度，研究算力–供暖解决方案，形成绿色算力认证体系，并实现能量有效再利用；在能源政策方面，为智能算力中心研究提高电费相关补助方案，尤其是针对绿电的补助方案，减少能耗负担，推动液冷、光伏等绿色技术在算力基础设施中的应用。

（五）优化“学研”一体化创新体系，打通成果转化“最后一千米”

为破解人才与协同困境，需加强推动平台共享及革新培养模式等多种举措。在平台共享方面，组建国家级AI算力创新平台，设立开放算力池与大模型训练沙盒环境，降低高校团队在大模型研发中的算力成本，提升算力利用的便利程度。在培养模式方面，推动高校与企业之间的联合培养及联合研发机制。一方面，推动高校优化芯片设计、框架开发、模型训练等课程，提升人才培养的实践能力；另一方面，推动头部企业开放工具链接口，促进高校成果在产业级平台的研发和验证，并完善知识产权共享机制，为校企联合攻关项目提供税收减免等优惠政策。

六、结语

本文深入探讨了大模型技术发展与高密度算力需求之间的相互关系，揭示了当前AI时代计算范式的深刻变革。首先，LLMs扩展定律的演进从根本上重塑了模型开发策略。未来大模型的性能提升将不仅依赖于原始计算能力，更日益受限于海量高质量训练数据的可获得性，从而推动数据工程和数据基础设施成为高密度算力环境中的关键组成部分。同时，MoE等高效模型架构的兴起，通过稀疏激活实现了计算效率和专业化的提升，改变了传统的扩展定律，预示着未来硬件和软件栈将向更灵活、自适应的方向发展。其次，量化、剪枝、蒸馏等大模型轻量化技术为其在资源受限环境中的部署提供了关键解决方案，使AI能力能够更广泛地触达各类应用场景。同时，MLLMs的发展带来了对异构计算的复杂需求。不同模态融合机制和训练范式的多样性，反映了计算效率、知识保留与跨模态交互深度之间的精妙权衡，推动了未来计算架构向高度灵活、异构且动态可重构的方向发展。最后，本研究深入分析了高密度算力面临的关键挑战。“功耗墙”和“散热瓶颈”构成了计算密度的根本性物理限制，“内存墙”和互连带宽限制则制约了LLMs的效率和可扩展性；软件栈优化，尤其是操作系统层面的挑战，凸显了软硬件协同设计的重要性。特别是，AI芯片供应链的脆弱性已上升为国家安全和经济发展的战略性问题。

展望未来，高密度算力的发展将与大模型技术的演进更加紧密地融合，并呈现出以下多个关键趋势和方向，包括软硬件协同设计的深度融合、光学互联与近存计算的加速突破和普及、可持续性与能源效率的优先考量、供应链韧性与地缘战略布局以及安全与隐私的内生设计等。当前，大模型技术与高密度算力的融合正处于一个关键的转折点。未来的发展将是多维度、跨学科的系统性工程，需要硬件、软件、算法、材料科学和能源策略的全面创新和紧密协作，以共同应对前所未有的技术挑战，并推动AI迈向更广阔、更可持续的未来。

利益冲突声明

本文作者在此声明不存在任何利益冲突或财务冲突。

Received date: August 19, 2025; **Revised date:** October 20, 2025

Corresponding author: Gong Zhenwei is a senior engineer from Zhongguancun Laboratory. His research field is cyberspace security strategy. E-mail: gongzw@zgclab.edu.cn

Funding project: Chinese Academy of Engineering project “Research on Risks of New Technologies and Applications in Cyberspace Security” (2023-JB-13), “Research on Computing Power Security and Its High-Quality Development Strategy in the Industry” (2024-HYZD-02)

参考文献

[1] 新华网. 我国大模型数量超1500个 [EB/OL]. (2025-07-27)[2025-08-07]. http://www. news. cn/tech/20250727/97930c6826c147349fc068894ac6bb96/c.html.
Xinhuanet. The number of large language models in China exceeds 1500 [EB/OL]. (2025-07-27)[2025-08-07]. http://www.news.cn/tech/20250727/97930c6826c147349fc068894ac6bb96/c.html.

[2] 2022中国算力大会. 中国算力白皮书(2022年) [R]. 济南: 中国算力大会, 2022.
2022 China Computational Power Conference. White paper on China’s computing power [R]. Jinan: China Computational Power Conference, 2022.

[3] 中国政府网. 算力基础设施高质量发展行动计划 [EB/OL]. (2023-10-10)[2025-09-17]. https://www. gov. cn/zhengce/zhengceku/202310/P020231009520949915888.pdf.
China Goverment Website. High-quality development action plan for computing power infrastructure [EB/OL]. (2023-10-10)[2025-09-17]. https://www.gov.cn/zhengce/zhengceku/202310/P020231009520949915888.pdf.

[4] OECD. A blueprint for building national compute capacity for artificial intelligence [EB/OL]. (2023-02-28)[2025-09-17]. https://www. oecd. org/content/dam/oecd/en/publications/reports/2023/02/a-blueprint-for-building-national-compute-capacity-for-artificial-intelligence_c22fbbee/876367e3-en.pdf.

[5] EPOCH AI. Trends in AI supercomputers [EB/OL]. (2025-04-23)[2025-09-17]. https://epoch.ai/blog/trends-in-ai-supercomputers.

[6] OECD. OECD framework for the classification of AI systems [EB/OL]. (2022-02-22)[2025-09-17]. https://www.oecd.org/content/dam/oecd/en/publications/reports/2022/02/oecd-framework-for-the-classification-of-ai-systems_336a8b57/cb6d9eca-en.pdf#page=21.40.

[7] Hu Q, Sun P, Zhang T. Understanding the workload characteristics of large language model development [EB/OL]. (2024-03-19)[2025-09-17]. https://www.usenix.org/publications/loginonline/understanding-workload-characteristics-large-language-model-development.

[8] Bahri Y, Dyer E, Kaplan J, et al. Explaining neural scaling laws [J]. Proceedings of the National Academy of Sciences, 2024, 121(27): e2311878121.

[9] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models [EB/OL]. (2020-01-23)[2025-07-16]. https://arxiv.org/abs/2001.08361.

[10] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models [EB/OL]. (2022-05-29)[2025-07-16]. https://arxiv.org/abs/2203.15556.

[11] Besiroglu T, Erdil E, Barnett M, et al. Chinchilla scaling: A replication attempt [EB/OL]. (2024-04-15)[2025-07-16]. https://arxiv.org/abs/2404.10102.

[12] Sardana N, Portes J, Doubov S, et al. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws [EB/OL]. (2023-12-31)[2025-07-16]. https://arxiv.org/abs/2401.00448.

[13] DeepSeek-AI, Bi X, Chen D L, et al. DeepSeek LLM: Scaling open-source language models with longtermism [EB/OL]. (2024-01-05)[2025-07-16]. https://arxiv.org/abs/2401.02954.

[14] Hu S D, Tu Y G, Han X, et al. MiniCPM: Unveiling the potential of small language models with scalable training strategies [EB/OL]. (2024-04-09)[2025-07-16]. https://arxiv.org/abs/2404.06395.

[15] Meta. Llama 3 [EB/OL]. [2025-11-07] https://www.llama.com/models/llama-3/.

[16] Snell C, Lee J, Xu K, et al. Scaling LLM test-time compute optimally can be more effective than scaling model parameters [EB/OL]. (2024-08-06)[2025-08-07]. https://arxiv.org/abs/2408.03314.

[17] Zhang Q Y, Lyu F Y, Sun Z X, et al. A survey on test-time scaling in large language models: What, how, where, and how well? [EB/OL]. (2025-03-31)[2025-08-07]. https://arxiv.org/abs/2503.24235.

[18] Lang J D, Guo Z H, Huang S Y. A comprehensive study on quantization techniques for large language models [R]. Xiamen: 2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC), 2025.

[19] Lin J, Tang J M, Tang H T, et al. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration [J]. GetMobile: Mobile Computing and Communications, 2025, 28(4): 12–17.

[20] Xiao G X, Lin J, Seznec M, et al. SmoothQuant: Accurate and efficient post-training quantization for large language models [EB/OL]. (2022-11-18)[2025-08-07]. https://arxiv.org/abs/2211.10438.

[21] Frantar E, Ashkboos S, Hoefler T, et al. GPTQ: Accurate post-training quantization for generative pre-trained transformers [EB/OL]. (2022-10-31)[2025-08-07]. https://arxiv.org/abs/2210.17323.

[22] Xu Z H, Xu Y, Xu H L, et al. Lightweight and post-training structured pruning for on-device large lanaguage models [EB/OL]. (2025-01-25)[2025-08-07]. https://arxiv.org/abs/2501.15255.

[23] Wang Y X, Ma M H, Wang Z K, et al. CFSP: An efficient structured pruning framework for LLMs with coarse-to-fine activation information [EB/OL]. (2024-09-20)[2025-08-07]. https://arxiv.org/abs/2409.13199.

[24] Wang Z H, Wohlwend J, Lei T. Structured pruning of large language models [EB/OL]. (2019-10-10)[2025-08-07]. https://arxiv.org/abs/1910.04732.

[25] Geng X, Gao J X, Zhang Y H, et al. Complex hybrid weighted pruning method for accelerating convolutional neural networks [J]. Scientific Reports, 2024, 14: 5570.

[26] Gou J P, Yu B S, Maybank S J, et al. Knowledge distillation: A survey [J]. International Journal of Computer Vision, 2021, 129(6): 1789–1819.

[27] Busbridge D, Shidani A, Weers F, et al. Distillation scaling laws [EB/OL]. (2025-02-12)[2025-08-07]. https://arxiv.org/abs/2502.08606.

[28] 史宏志, 赵健, 赵雅倩, 等. 大模型时代的混合专家系统优化综述 [J]. 计算机研究与发展, 2025, 62(5): 1164–1189.
Shi H Z, Zhao J, Zhao Y Q, et al. Survey on system optimization for mixture of experts in the era of large models [J]. Journal of Computer Research and Development, 2025, 62(5): 1164–1189.

[29] Lepikhin D, Lee H, Xu Y Z, et al. GShard: Scaling giant models with conditional computation and automatic sharding [EB/OL]. (2020-06-30)[2025-08-07]. https://arxiv.org/abs/2006.16668.

[30] Jiang A Q, Sablayrolles A, Roux A, et al. Mixtral of experts [EB/OL]. (2024-01-08)[2025-08-07]. https://arxiv.org/abs/2401.04088.

[31] DeepSeek-AI, Liu A X, Feng B, et al. DeepSeek-V3 technical report [EB/OL]. (2024-12-27)[2025-08-07]. https://arxiv.org/abs/2412.19437.

[32] Liu J C, Tang P, Wang W F, et al. A survey on inference optimization techniques for mixture of experts models [EB/OL]. (2024-12-18)[2025-08-07]. https://arxiv.org/abs/2412.14219.

[33] 郭园方, 余梓彤, 刘艾杉, 等. 多模态大模型安全研究进展 [J]. 中国图象图形学报, 2025, 30(6): 2051–2081.
Guo Y F, Yu Z T, Liu A S, et al. Recent progress of the security research for multimodal large models [J]. Journal of Image and Graphics, 2025, 30(6): 2051–2081.

[34] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [R]. Online: International Conference on Machine Learning, 2021.

[35] 中科算网, 中国科大高研院. AI大模型与异构算力融合技术白皮书 [EB/OL]. (2025-10)[2025-11-07]. https://pdf.dfcfw.com/pdf/H3_AP202510141762072518_1.pdf?1760514880000.pdf.
Zhong Ke Suan Wang, Suzhou Institute For Advanced Research University of Science and Technology in China. White paper on AI large model and heterogeneous computing power fusion technology [EB/OL]. (2025-10)[2025-11-07]. https://pdf.dfcfw.com/pdf/H3_AP202510141762072518_1.pdf?1760514880000.pdf.

[36] 算力 100 问. 第 60 问: 什么是算力密度? [EB/OL]. (2025-05-23)[2025-08-07]. https://mp.weixin.qq.com/s/I9tIKVmtylf371k9juyP-A.
100 Questions on Computing Power. Question 60: What is computing density [EB/OL]. (2025-05-23)[2025-08-07]. https://mp.weixin.qq.com/s/I9tIKVmtylf371k9juyP-A.

[37] 范特科技有限责任公司. 智算未来: 范特科技大模型与智算发展白皮书 [R]. 无锡: 范特科技有限责任公司, 2025.
Fantec. The future of AI computing power: Fantech's big model and the development of AI computing power [R]. Wuxi: Fantec, 2025.

[38] 东吴证券. 从 GPGPU 与 ASIC 之争——算力芯片看点系列 [EB/OL]. (2025-03-12)[2025-08-07]. https://pdf.dfcfw.com/pdf/H3_AP202503121644311455_1.pdf 1741816606000.pdf.
Soochow Securities. From the battle between GPGPU and ASIC-the focus of computing chips series [EB/OL]. (2025-03-12)[2025-08-07]. https://pdf. dfcfw. com/pdf/H3_AP202503121644311455_1.pdf 1741816606000.pdf.

[39] FS. High-density servers: Maximizing efficiency and performance in data centers [EB/OL]. (2024-03-28)[2025-08-07]. https://www.fs.com/blog/highdensity-servers-maximizing-efficiency-and-performance-in-data-centers-7070.html.

[40] 李泽林. 高密度数据中心建设浅析 [J]. 智能建筑与智慧城市, 2025 (S1): 153–155.
Li Z L. Brief analysis of the construction of high density data center [J]. Intelligent Building & Smart City, 2025 (S1): 153–155.

[41] 徐建, 郑伟, 郭晓春, 等. 新型数据中心网络安全体系研究 [J]. 信息安全与通信保密, 2022, 20(7): 123–132.
Xu J, Zheng W, Guo X C, et al. Research on next-generation data center security system [J]. Information Security and Communications Privacy, 2022, 20(7): 123–132.

[42] 中国信通院. 算力中心冷板式液冷发展研究报告 [EB/OL]. (2024-05)[2025-08-07]. http://www. caict. ac. cn/kxyj/qwfb/ztbg/202405/P020240523566116859176.pdf.
China Academy of Information and Communications Technology. Research report on the development of cold plate liquid cooling in data centers [EB/OL]. (2024-05)[2025-08-07]. http://www.caict.ac.cn/kxyj/qwfb/ztbg/202405/P020240523566116859176.pdf.

[43] 中国信通院. 数据中心白皮书 [EB/OL]. (2022-04)[2025-08-07]. http://www.caict.ac.cn/kxyj/qwfb/bps/202204/P020220422707354529853.pdf#page=33.43.
China Academy of Information and Communications Technology. Data center white paper [EB/OL]. (2022-04)[2025-08-07]. http://www. caict. ac. cn/kxyj/qwfb/bps/202204/P020220422707354529853.pdf#page=33.43.

[44] McKinsey & Company. AI power: Expanding data center capacity to meet growing demand [EB/OL]. (2024-10-29)[2025-08-07]. https://www. mckinsey. com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand.

[45] 浪潮信息. i48M6 [EB/OL]. [2025-08-07]. https://www.ieisystem.com/product/server/8335.html.
IEIT SYSTEMS.i48M6 [EB/OL]. [2025-08-07]. https://www.ieisystem.com/product/server/8335.html.

[46] 腾讯网. 百度智能云: 昆仑芯超节点支持 1U4 卡的超高密度算力交付形式 [EB/OL]. (2025-05-29)[2025-09-25]. https://news.qq.com/rain/a/20250529A08EZK00.
Tencent News. Baidu Intelligent Cloud: Kunlun chip super node supports ultra-high-density computing power delivery in 1U4 card form factor [EB/OL]. (2025-05-29)[2025-09-25]. https://news.qq.com/rain/a/20250529A08EZK00.

[47] 中国新闻网. 国产软硬件一体高密度算力机柜——Shanghai Cube亮相WAIC2025 即将进入量产 [EB/OL]. (2025-07-28)[2025-08-07]. https://www.toutiao.com/article/7532007542914941490/.
ChinaNews. Domestic hardware and software integrated high-density computing cabinet-Shanghai Cube debuted at WAIC2025 about to enter mass production [EB/OL]. (2025-07-28)[2025-08-07]. https://www.toutiao.com/article/7532007542914941490/.

[48] Nvidia. Nvidia DGX superPOD [EB/OL]. [2025-08-07]. https://www.nvidia.com/en-us/data-center/dgx-superpod/.

[49] Huawei. Atlas 900 A3 superPoD [EB/OL]. [2025-08-07]. https://info. support. huawei. com/computing/qrcode/atlas900a3superpod/index-cn.html.

[50] 新浪财经. 云栖大会今开幕, 阿里云将展示超大规模集群、分布式训练、推理加速等能力,首次展出高密度AI服务器和高性能网络架构 [EB/OL]. (2025-09-24)[2025-09-25]. https://finance.sina.com.cn/roll/2025-09-24/doc-infrpwtx6867998.shtml.
Finance Sina. The Yunqi Conference opens today, with Alibaba Cloud showcasing capabilities such as ultra-large-scale clusters, distributed training, and inference acceleration, while debuting high-density AI servers and high-performance network architectures for the first time [EB/OL]. (2025-09-24)[2025-09-25]. https://finance.sina.com.cn/roll/2025-09-24/doc-infrpwtx6867998.shtml.

[51] 中国信通院. 中国绿色算力发展研究报告 [EB/OL]. (2024-06)[2025-08-07]. https://www.caict.ac.cn/kxyj/qwfb/ztbg/202407/P020240711551514828756.pdf.
China Academy of Information and Communications Technology. Research report on the development of green computing power in China [EB/OL]. (2024-06)[2025-08-07]. https://www.caict.ac.cn/kxyj/qwfb/ztbg/202407/P020240711551514828756.pdf.

[52] 余佳, 马晓波. 半导体先进封装领域专利技术综述 [J/OL]. 电子与封装, 2025: 1–12[2025-07-22]. https://link. cnki. net/doi/10.16257/j.cnki.1681-1070.2026.0005.
Yu J, Ma X B. Review on patent technologies in advanced semiconductor packaging [J/OL]. Electronics & Packaging, 2025: 1–12 [2025-07-22]. https://link.cnki.net/doi/10.16257/j.cnki.1681-1070.2026.0005.

[53] 黄翔, 李潼, 褚俊杰. 算力时代数据中心液冷与蒸发冷的融合发展 [J/OL]. 制冷与空调, 1–10 [2025-08-07]. https://link.cnki.net/urlid/11.4519.tb.20250327.1407.002
Huang X, Li T, Chu J J. The integration development of evaporative cooling and liquid cooling in the era of computing power [J/OL].Refrigeration and Air-Conditioning, 1–10 [2025-08-07]. https://link.cnki.net/urlid/11.4519.tb.20250327.1407.002.

[54] 电子工程专辑. 海关公布中国2024年芯片进出口数据, 出口首破万亿元 [EB/OL]. (2025-02-05)[2025-09-25]. https://www.eet-china.com/news/202502059154.html.
EE Times China. Chinese Customs releases China's 2024 chip import and export data, with exports exceeding 1 trillion yuan for the first time [EB/OL]. (2025-02-05)[2025-09-25]. https://www. eet-china.com/news/202502059154.html.

[55] 陈聃. 基于近内存计算的图神经网络加速技术研究 [D]. 武汉: 华中科技大学(博士学位论文), 2024.
Chen D. Research on acceleration techniques of graph neural networks based on near-memory processing [D]. Wuhan: Huazhong

University of Science and Technology (Doctoral dissertation), 2024.

[56] Wulf W A, McKee S A. Hitting the memory wall: Implications of the obvious [J]. ACM SIGARCH Computer Architecture News, 1995, 23(1): 20–24.

[57] TechInsights. China does it againa NAND memory market first [EB/OL]. (2025-09-04)[2025-09-25]. https://www.techinsights.com/blog/china-does-it-again-nand-memory-market-first.

[58] 电子工程专辑. 长鑫存储 HBM2 内存获突破, DDR5 良率明年可达 90% [EB/OL]. (2024-12-30)[2025-09-25]. https://www.eet-china.com/news/202412303389.html.
EE Times China. Changxin Memory achieves breakthrough in HBM2 memory, DDR5 yield expected to reach 90% next year [EB/OL]. (2024-12-30)[2025-09-25]. https://www.eet-china.com/news/202412303389.html.

[59] ZOMI 酱, 苏统华. AI 系统—原理与架构 [M]. 北京: 科学出版社, 2024.
ZOMI Sauce, Su T H. AI Systems — Principles and architecture [M]. Beijing: Science Press, 2024.

[60] Medium. AI compilers demystified [EB/OL]. (2022-11-08)[2025-08-07]. https://medium.com/geekculture/ai-compilers-ae28afbc4907.

[61] 段柳成, 肖巧玲, 金怡, 等. 大模型时代国产大算力 GPU 的关键挑战与发展路径 [J]. 人工智能, 2025, 12(3): 8–21.
Duan L C, Xiao Q L, Jin Y, et al. Key challenges and development path of domestic large computing GPU in the age of large model [J]. AI-View, 2025, 12(3): 8–21.

[62] IEA. Energy and AI [R]. Paris: IEA, 2024.

[63] 中国信通院. 算力电力协同发展研究报告 (2025 年) [EB/OL]. (2025-05)[2025-09-25]. https://www.caict.ac.cn/kxyj/qwfb/ztbg/202505/P020250509511369626787.pdf#page=2.14.
China Academy of Information and Communications Technology. Research report on the collaborative development of computing power and electricity [EB/OL]. (2025-05)[2025-09-25]. https://www.caict.ac.cn/kxyj/qwfb/ztbg/202505/P020250509511369626787.pdf#page=2.14.

[64] NIST. SP 800-53 control overlays for securing AI systems [EB/OL]. (2025-08-14)[2025-09-18]. https://csrc.nist.gov/csrc/media/Projects/cosais/documents/NIST-Overlays-SecuringAI-concept-paper.pdf.

[65] Guo Y N, Zhang Z K, Yang J. GPU memory exploitation for fun and profit [R]. Philadelphia: The 33rd USENIX Conference on Security Symposium, 2024.

[66] Hoover J. Analysis of GPU memory vulnerabilities [D]. Fayetteville: University of Arkansas, Fayetteville (Undergraduate honors theses), 2022.

[67] Forbs. Nvidia security warning—Act now as 7 new GPU vulnerabilities confirmed [EB/OL]. [2025-09-17]. https://www.forbes.com/sites/daveywinder/2025/01/28/nvidia-security-warning-act-now-as-7-new-gpu-vulnerabilities-confirmed/.

[68] Oligo. ShadowRay: First known attack campaign targeting AI workloads actively exploited in the wild [EB/OL]. (2024-03-26)[2025-09-17]. https://www.oligo.security/blog/shadowray-attack-ai-workloads-actively-exploited-in-the-wild.

[69] PyTorch. Compromised PyTorch-nightly dependency chain between December 25th and December 30th, 2022 [EB/OL]. (2022-12-31)[2025-08-07]. https://pytorch.org/blog/compromised-nightly-dependency/.

[70] Ridge Security Research Team. Securing your AI: Critical vulnerabilities found in popular Ollama framework [EB/OL]. (2025-03-20)[2025-08-07]. https://ridgesecurity.ai/blog/securing-your-ai-critical-vulnerabilities-found-in-popular-ollama-framework/.

[71] Federal Register. Information security controls: Cybersecurity Items [EB/OL]. (2022-05-26)[2025-09-26]. https://www.federalregister.gov/documents/2022/05/26/2022-11282/information-security-controls-cybersecurity-items.

[72] 康旺, 寇竞, 赵巍胜. 存算一体芯片发展现状、趋势与挑战 [J]. 中国科学: 信息科学, 2024, 54(1): 16–24.
Kang W, Kou J, Zhao W S. In-memory computing technology: Development status, trends and challenges [J]. Scientia Sinica (Informationis), 2024, 54(1): 16–24.