计算机科学
COMPUTER SCIENCE

AI+HPC:“智能+”驱动下的超算系统软件及应用技术发展综述

谭政源，钟佳卿，陈娟

引用本文

谭政源，钟佳卿，陈娟. AI+HPC:“智能+”驱动下的超算系统软件及应用技术发展综述[J]. 计算机科学, 2025, 52(5): 1-10.

TAN Zhengyuan, ZHONG Jiaqing, CHEN Juan. AI+HPC:An Overview of Supercomputing System Software and Application Technology Development Driven by “AI+” [J]. Computer Science, 2025, 52(5): 1-10.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241100177

AI+HPC:“智能+”驱动下的超算系统软件及应用技术发展综述

谭政源　钟佳卿　陈　娟
国防科技大学计算机学院　长沙 410073
(tanzhengyuan@nudt.edu.cn)

摘　要　人工智能(AI)和高性能计算(HPC)是计算机领域的两大重要技术。随着计算机技术的飞速发展,二者的联系逐渐紧密,并呈现出互相依赖、互相促进的关系。一方面,高性能计算系统面临的各种新问题与新挑战,需要人工智能方法技术辅助解决(AI for HPC);另一方面,人工智能领域理论的突破,依赖于HPC提供的强大的计算能力(HPC for AI)。在这样的背景下,AI和HPC两领域交叉融合,深度发展。文中系统回顾了近年来AI和HPC两个领域各自技术的发展脉络,着重从以下几方面展开分析:1)AI技术在解决HPC硬件体系结构、操作系统资源管理、编译优化和软件开发等几个方面问题的贡献;2)HPC为AI在硬件基础设施及软件应用上的支持;3)AI和HPC领域融合的未来发展前景与挑战。

关键词:人工智能;高性能计算;领域融合;硬件体系;软件应用
中图分类号　TP302

AI+HPC:An Overview of Supercomputing System Software and Application Technology Development Driven by “AI+”

TAN Zhengyuan,ZHONG Jiaqing and CHEN Juan
College of Computer Science and Technology,National University of Defense Technology,Changsha 410073,China

Abstract　Artificial Intelligence(AI) and High Performance Computing(HPC) are two essential technologies in computer science. With the rapid development of computer science and technology,there has been a gradual trend of convergence and development of AI and HPC. On the one hand,new challenges in high-performance computing systems require AI-powered solutions (AI for HPC). On the other hand,breakthroughs in artificial intelligence demand the support of high-performance computing (HPC for AI). Consequently,the convergence of AI and HPC strikes the development of core technologies in their respective fields. In this paper,we systematically review the respective technological development in the fields of AI and HPC in the past decade,focusing on three aspects:1)the role of AI technology in HPC hardware architecture,operating system resource management,compilation optimization,and software development,etc;2)the support of HPC for AI in terms of system hardware solutions and software applications;3)prospects and challenges for the future development of AI and HPC convergence.

Keywords　Artificial intelligence,High performance computing,Domain convergence,Hardware architecture,Software application

1　引言

近年来,随着计算机信息技术的不断发展,数据规模不断膨胀,“大数据”成为各行各业频繁提及的概念。而AI和HPC作为大数据时代两大重要的科学研究工具,正面临各种新的机遇与挑战。“智能+”理念在这一背景下应运而生,旨在将以人工智能为代表的新一代信息技术与国家重要产业相结合,创造新的生产力与价值。回顾近年来AI和HPC领域的技术融合与发展,对在大数据、“智能+”背景下实现两大领域的新突破具有重要意义。

技术和算力的飞速发展,是AI和HPC领域融合程度不断提高的重要动力。图1给出了近年来AI和HPC领域结合发展与算力的关系曲线。AI和HPC领域的结合从21世纪初期开始布局,在近十年达到高峰,关键在于AI领域深度学习理论的奠基与发展,以及HPC底层算力架构的大转变。21世纪初期,图形处理器(Graphics Processing Unit,GPU)作为计算设备的潜能开始被人们关注,通用计算图形处理器(General-purpose Computing on Graphics Processing Units,GPGPU)的概念诞生。2006年,AI领域“深度学习”概念的提出大大加快了AI的发展速度;同一时期,在HPC领域,NVIDIA提出使GPU具备通用计算能力的CUDA框架,为高性能计算领域算力的提升和系统架构的革新提供了崭新的

到稿日期:2024-11-28　返修日期:2025-03-02
基金项目:并行与分布计算全国重点实验室基金项目(2023-KJWPDL-01)
This work was supported by the Open Fund of National Key Laboratory of Parallel and Distributed Computing(PDL)(2023-KJWPDL-01).
通信作者:陈娟(juanchen@nudt.edu.cn)

思路，是异构并行的起点。2010 年，在 HPC 领域，采用 CPU+GPU 异构并行架构的天河 1A 一举登顶 TOP500，从此全球的 HPC 系统开始大规模采用异构并行架构。2012 年，在 AI 领域，使用 NVIDIA 异构算力平台训练的深度神经网络 AlexNet 在 ImageNet 竞赛中取得了大幅超越传统方法的表现，开创了 AI 和 HPC 结合的先河。从此，AI 和 HPC 全面进入融合发展阶段，目前超算实力相对较强的美、日、中在 TOP500 排行世界领先的超算平台上，已有大量开展 AI+HPC 科研的相关项目。例如 Summit 计算机上开展的 AI for Science 项目[1]，使用 AI 方法解决生物学、计算机科学、核物理和材料科学等领域的疑难问题，其中一种用于解决电子密度重建问题的神经网络模型在 4 600 个节点上实现了 2.15 EFlops 的混合精度计算性能；Fugaku 计算机上开展的脑神经科学模拟项目[2]，使用 1 536 个进程、384 个节点对大脑工作方式进行模拟，可模拟 8 000 万个神经元；神威太湖之光上开发的复杂神经网络的加速库 swDNN[3]，为深度学习类应用提供了大规模并行优化，能够实现超过 1.6 TFlops 的双精度性能和超过 54%的计算效率；芬兰的 LUMI 超算上运行的智能虚拟筛选方法选择药物候选项目[4]，底层是 GPU 计算节点提供的强大性能和 AI+HPC 算力基础，配合编译器优化，单 GPU 即可达到 1.46 倍 NVIDIA A100 的性能。这些超级计算机平台及其配置如表 1 所列。随着大模型、通用人工智能的发展以及新型超算软硬件的深入研究，AI 和 HPC 融合必要性逐渐增强，各种新的问题也陆续出现。

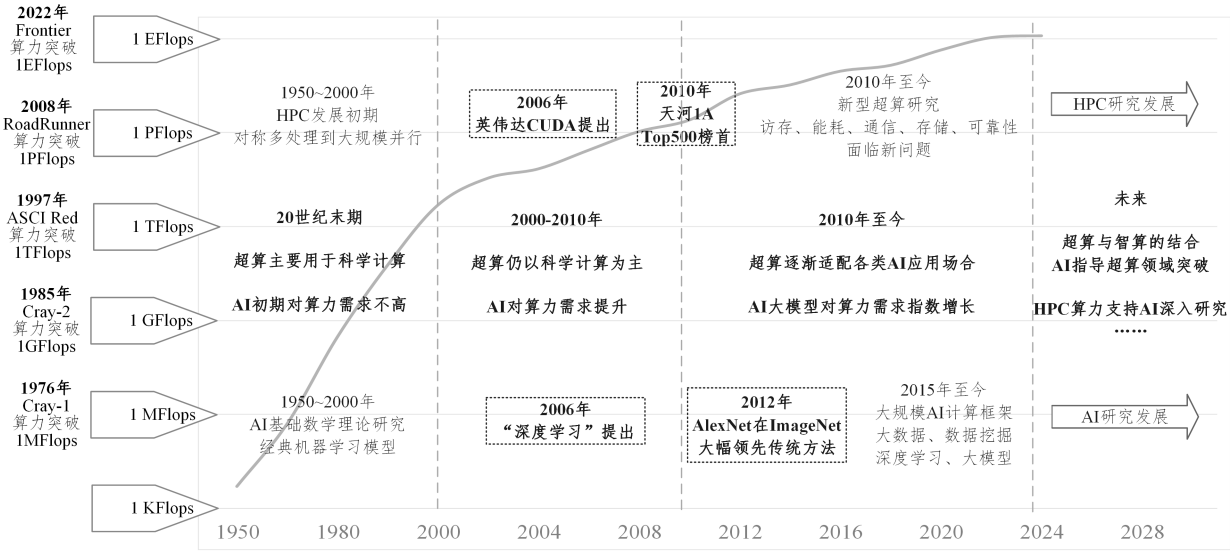

图 1 AI-HPC 发展曲线

Fig. 1 AI-HPC development curve

表 1 TOP500 典型超算平台及其配置

Table 1 Typical TOP500 supercomputing platforms and their configurations

超算名称	所属	硬件配置	理论性能（双精度，Rpeak）
Summit	美国	4 600 节点数 单节点 2×IBM POWER9，6×NVIDIA V100 GPU，512 GB DDR4 内存，96 GB HBM2 高带宽内存[1]	200 PFlops
Fugaku	日本	158 976 节点数 单节点 A64FX 处理器（48 核心），32 GB HBM2 高带宽内存[2]	537 PFlops
神威太湖之光	中国	20 480 计算节点数 单节点 2×SW26010 异构众核处理器[3]	125 PFlops
LUMI	欧洲	2 560 节点数 单节点 AMD EPYC Trento CPU，2×AMD MI250X GPU[4]	214 PFlops

在人工智能领域，深度学习、大数据、大模型的持续发展，对当代计算系统在算力、显存、通信等方面提出了更高要求。算力方面，小规模系统已无法满足当代大模型的算力需求，如使用 312 TFlops 的 A100 GPU 训练包含 300 万 tokens 的175 B GPT-3，在每输入一个 token 进行 6 到 8 次浮点运算的条件下，训练任务将带来长达 32 年的时间开销[5]；显存方面，单个模型副本中每个参数需要占据约自身大小 20 倍的显存量[6]，如 GPT-3 需要占据至少 3.5 TB 的显存，对应 44 块显存为 80 GB 的显卡；通信方面，在大模型的训练和推理过程中，由于大模型对多卡多节点的刚需，以及模型自身的流水线并行、张量并行等策略，节点间需要进行频繁的数据交换，造成极大的互连压力。AI 领域的研究逐渐呈现规模膨胀的特点，这决定了 AI 未来的发展必须依赖 HPC 系统在各个方面的全面保障。

在高性能计算领域，体系结构、系统软件、应用程序等方面正面临许多新问题。体系结构方面，HPC 系统发展至今，从对称多处理结构到大规模并行同构计算，再到大规模并行异构计算，随着核数的增加和规模的增大，能耗和互连等问题成为限制系统稳定性和性能的重要因素[7]，制约了 HPC 系统规模的进一步扩大升级[8]；系统软件方

面，面对大规模的硬件配置，HPC 操作系统需要合理分配调度各类资源，在计算、访存、通信等方面提升系统的性能，同时也需要通过编译优化等系统级方法提升程序运行的性能；应用程序方面，随着 HPC 系统在各个领域投入使用，并行程序逐渐复杂，这对程序开发者提出了更高的标准要求，如何加速程序开发并减轻程序开发者的负担成为重要问题。这些 HPC 领域新兴的难题，需要 AI 方法的指导与创新。

为迎接 AI＋HPC 领域面临的挑战，近年来研究人员开展了许多 AI 和 HPC 领域融合的工作，使用 AI 技术解决 HPC 领域面临的各种问题，同时在 HPC 系统上开展针对 AI 的适配与优化。相关工作的详细分类如图 2 所示。

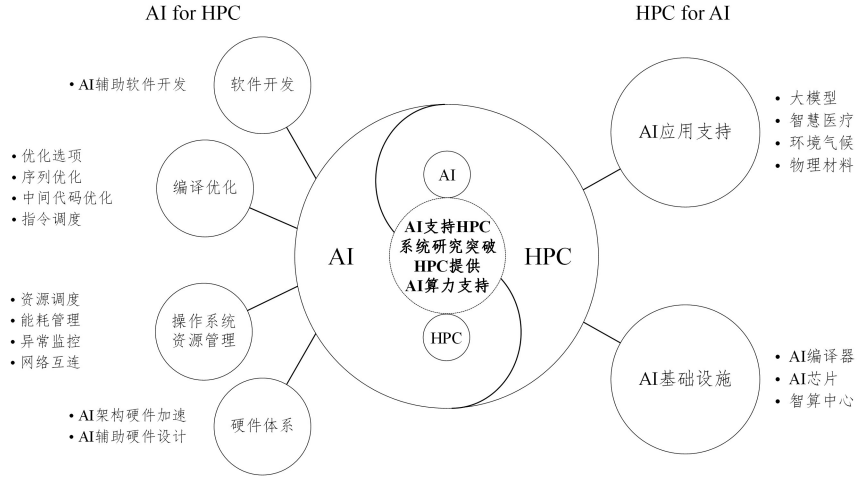

图 2 AI＋HPC 超算系统软件及应用最新技术领域

Fig. 2 AI＋HPC supercomputing system software and application of the latest technology areas

本文后续章节组织如下，所有内容汇总如表 2 所列。

第 2 章：AI for HPC。本章调研了 AI 解决 HPC 领域问题的相关工作，从硬件体系、操作系统资源管理、编译优化和软件开发 4 个方面进行阐述，表明了 AI 方法对于解决 HPC 领域各类疑难问题的贡献。

第 3 章：HPC for AI。本章对 HPC 适配 AI 相关领域应用，满足 AI 领域以及 AI for Science 需求的相关工作进行了总结，从 AI 基础设施和应用支持两个方面进行分析，表明了 HPC 对于 AI 理论与应用深入研究不可或缺的地位。

第 4 章：未来展望与挑战。本章分析了 HPC 和 AI 领域未来的发展趋势，并给出了 AI 和 HPC 的融合发展面临的挑战。

最后，总结了上述章节的内容，并对未来二者的结合趋势与方向进行了展望。

表 2 AI＋HPC 超算系统软件应用技术领域一览表

Table 2 Table of AI＋HPC supercomputing system software application technology areas

AI for HPC		HPC for AI	
硬件体系结构	AI 指导芯片设计[9-10]；AI 架构辅助 HPC 加速[11-12]	AI 软件基础设施	AI 编译器[39-41]
系统资源管理	资源调度优化（神经网络、强化学习、图挖掘）[13-18]；能耗管理（监督学习）[19]；异常监控（VAE、图神经网络、随机森林）[20-23]；网络互连（深度强化学习）[24-25]	AI 硬件基础设施	AI 芯片[42-47]；智算中心[48]
编译优化	优化选项组合（机器学习、贝叶斯网络）[26-28]；序列优化（聚类、监督学习）[29-30]；中间代码优化（强化学习）[31]；指令调度（强化学习）[32]	AI 前沿应用支持	AI 大模型[49-51]
软件开发	AI 辅助 HPC 软件开发[36-38]	AI 科学应用支持	智慧医疗[52-55]；环境气候[56-60]；物理学和材料学[61-62]

2 AI for HPC

AI 领域的高速发展为 HPC 领域提供了各类先进的研究方法与工具，使用 AI 解决 HPC 领域前沿问题已经成为目前高性能计算领域研究的主流。目前，人工智能方法被广泛用于解决 HPC 领域硬件体系、操作系统资源管理、编译优化、软件开发等各个方面的困难问题。这些方面从底层硬件到操作系统，乃至编译器及上层软件，囊括了 HPC 系统的各个层级。

2.1 硬件体系

硬件体系是 HPC 系统的物理组成，在 AI 方法的助力下，HPC 系统硬件在芯片设计、体系架构等方面呈现出多样性。

芯片设计方面，Mirhoseini 等[9]提出一种使用深度强化学习进行芯片布局的方法，用于辅助开发谷歌下一代加速器芯片，能够在 6h 内自动生成与人工设计水平相当的芯片布局，大大加速了算力芯片的研发；Jeffrey[10]总结了机器学习方法在硬件芯片设计方面的应用，如使用机器学习来自动化和优化 ASIC 设计流程中的某些步骤，包括布局和布线等，以及检查芯片流片质量，及时发现质量较差的芯片产品。

体系架构方面，Ltaief 等[11]提出了一种使用 AI 硬件体系代替传统 HPC 硬件体系的思路，通过批处理低秩矩阵运算、数据稀疏性压缩等方法对 HPC 基础计算进行优化，并采用适配 AI 架构的线性代数运算 API，在地震成像、无线通信和

气候预测等 HPC 应用上取得了 10 倍以上的性能提升；An 等[12]结合 AI 计算的特点指导 HPC 硬件架构的设计，实现了 AI-HPC 融合架构，使用 10 000 个 NVIDIA A100 GPU 设计并构建了 Fire-Flyer 2 体系结构，实现了接近 DGX-A100 的性能，同时将其他成本减半，理论能耗降低可达 40%。

2.2 操作系统资源管理

操作系统资源是 HPC 系统实现算力的底层基础。在 AI 方法指导下，HPC 的操作系统资源管理在资源调度、能耗管理、异常监控、网络互连等方面取得了很大进展和突破。

2.2.1 资源调度

一个高性能的 HPC 系统，需要利用有限的计算资源实现尽可能高的性能。资源调度的目的是通过对计算机硬件资源的管理和协调，对 HPC 系统作业性能、功耗等方面进行优化，为实际应用提供高质量的服务与体验。

对于 HPC 在云计算、雾计算环境中的任务调度问题，Alsadie 等[13]总结了目前常见的任务调度方法，揭示了使用可解释的人工智能方法解决此类问题的潜能；Iftikhar 等[14]在已有的资源管理系统基础上，采用优化的门控图卷积神经网络估计 HPC 上作业的服务质量参数，并将其作为资源分配和调度的依据，与现有方法相比，其能源效率提升了 17%，作业完成率提升了 14%。

对于高性能计算集群中一般任务的调度问题，使用强化学习方法也是领域研究者的常见选择。Fan 等[15]提出一种名为 DRAS 的强化学习 Agent，该方法使用两个层级的神经网络，以 HPC 系统的历史日志作为训练数据集，分别对立即执行或预留的作业以及需要回填的作业进行调度，在作业调度性能上比现有方法提升了 45%。Narantuya 等[16]提出一种使用多 Agent 深度强化学习进行计算资源分配的方法，该方法使用分布式多 Agent，每个 Agent 负责各自分区的资源分配；与传统方法相比，该方法在作业时长上减少了 20%，能量额外损耗降低了 40%。Zhang 等[17]提出一种使用强化学习的任务调度器，并对用于训练的作业日志轨迹集合进行了筛选，保证训练作业涵盖多个特征，以提升学习稳定性。该方法在实际的作业场景中表现良好，对未在训练集中的作业也能够较好地完成调度任务。

对于多用户共享资源的场景，如数据中心的资源调度问题，Ranganath 等[18]提出了一种基于图模式匹配的方法，将加速节点抽象为应用程序图和硬件图，使用图挖掘方法寻找可能的匹配模式，并根据每种模式的带宽特征进行评分，选择最为合适的模式进行资源分配，使程序执行时间缩短 12.4%，且最坏情况下的执行时间可减少 35%。

2.2.2 能耗管理

高性能的 HPC 系统往往伴随着高功耗的问题，从而增加了 HPC 系统运营的成本，同时降低了系统稳定性，造成不可预料的错误。能耗管理通过对系统运行时状态进行分析，采取合适的电源管理动作，保证系统功耗处在稳定状态，确保用户作业稳定运行。Jung 等[19]提出了一种基于监督学习的动态电源管理框架，通过贝叶斯分类器对即将进入队列的用户作业自动分配最佳的电压和频率设置，减轻了传统 DVFS 动态调频调压方法对系统造成的额外负担，且与随机策略和全局电源管理策略相比分别节省了 20.5% 和 11.5% 的系统能量。

2.2.3 异常监控

对于 HPC 集群而言，大规模的硬件配置和操作系统在叠加性能的同时，也会叠加发生错误的频率。HPC 系统出现各类错误已成为常态，这些错误对系统的稳定性造成了不可忽视的影响。异常监控需要准确掌握系统运行的实时状态，及时对出现的错误做出反应，减少异常对系统性能和稳定性的影响，以及造成的用户损失。Aksar 等[20]搭建了一个名为 Prodigy 的基于变分自动编码器(Variational Autoencoder, VAE)的无监督 HPC 系统异常检测系统，从正常和异常的 HPC 应用中提取运行时特征，并基于正常的应用特征训练 VAE，根据异常应用特征编解码的重构损失设置判断阈值，以此判断是否存在异常。该方法在 Eclipse 系统和 HPC 实际系统上分别取得了 0.95 和 0.88 的 F1 分数。Isakov 等[21]分析了 HPC 系统的 I/O 效率问题以及现有模型表现欠佳的原因，并对这些原因进行了分类，包括应用级建模误差、系统级建模误差、数据集未覆盖的应用或系统特征、I/O 竞争和噪声，同时开发了一系列鉴别这些分类的方法，为 I/O 效率监控提供了重要参考。Li 等[22]提出了一种基于图神经网络检测数据中心应用程序内存效率问题的方法，对不必要的内存操作进行识别，在 SPEC CPU 2017 Benchmark 上以高达 96% 的准确率捕捉到内存效率问题，检查开销仅为现有技术的 17.7%。Boixaderas 等[23]提出一种预测 DRAM 未校正错误的方法，使用随机森林分类器，并基于实际超算平台的错误日志进行训练，能够减少高达 57% 因硬件故障丢失的计算时间，每年能节省约 21 000 节点小时的计算资源。

2.2.4 网络互连

一个大规模的 HPC 集群，节点数可达数万至数十万，对于占据节点数目较多的作业，节点间通信开销极大，严重影响提交作业的性能，同时会增加系统的额外功耗。HPC 系统网络互连优化的目标是减少节点间的通信开销，提升作业性能表现，或优化通信造成的额外功耗损耗，降低系统功耗，提升系统稳定性，实现可靠的低延迟通信。Xiao 等[24]和 Gu 等[25]使用多 Agent 深度强化学习方法，来解决边缘计算、物联网环境下的超可靠低延迟通信(URLLC)问题，通过多 Agent 决策进行合理的任务划分和通信，对大规模 HPC 系统网络互连设计优化具有重要的借鉴意义。

2.3 编译优化

HPC 应用代码需要经过编译，形成可执行程序。在充足的硬件资源和合适的操作系统调度下，需要对编译过程进行优化，以得到更高质量的机器代码，提升 HPC 应用的表现。编译优化主要集中在优化选项组合、序列优化、中间代码优化和指令调度几个方面。

2.3.1 优化选项组合

优化选项是编译器在进行代码编译之前，根据用户添加的编译参数，在编译过程中对代码进行不同的处理，得到在性能、功耗等方面不同的程序。Ashouri 等[26]总结了近年来使用机器学习方法解决编译优化问题的相关工作，并在文献[27]中使用贝叶斯网络模型，综合各种静态程序特征和动态

程序特征来构建最佳编译优化选项模型。Fursin 等[28]提出一种机器学习驱动的自适应编译器优化框架，该框架使用机器学习技术预测优化设置，改善程序执行时间，并提供了可以稳定使用的开源工具。

2.3.2 序列优化

程序的编译过程涉及到多个步骤，如词法分析、语法分析、语义分析、中间代码生成、优化和目标代码生成等。这些步骤的排序属于 NP-Hard 问题，传统方式需要进行大量的枚举，复杂度极高，会极大影响编译器的性能。针对编译器序列优化的性能问题，Martins 等[29]提出了一种基于聚类的选择方法对编译序列进行优化，对相似的函数进行聚类，分组寻找特定的优化序列，对实际函数则以现有的类别为参考，根据与参考函数类的相似性确定优化序列。与遗传算法相比，其枚举空间减少到 1/18，使得编译得到的程序取得了 43％的性能提升。Ashouri 等[30]提出了一个使用监督学习方法自动完成序列优化的编译框架，该框架对 LLVM -O3 的所有编译优化进行聚类，形成若干子序列，并基于程序动态特征预测指定子序列相对 O3 优化的加速比，最终实现了在仅探索 0.001％序列优化空间的条件下使编译后的程序达到了 90％的理论最大性能。

2.3.3 中间代码优化

中间代码优化，是独立于目标系统架构的一种编译优化形式。使用 AI 方法对中间代码进行优化，能够在无需考虑目标硬件特性的条件下，对最终代码性能进行提升。Nazim 等[31]提出一种使用强化学习方法对 MLIR 中间代码进行优化的方法，对 MLIR 代码优化空间进行自动探索，经过优化的中间代码在转换为程序后，取得了比传统启发式方法更高的加速比。

2.3.4 指令调度

指令调度要求编译器在尽可能不破坏各类依赖关系的情况下，通过对指令的各种操作（如并行执行），提高程序的性能，或降低程序的无用功耗。McGovern 等[32]使用强化学习方法对指令调度进行优化，使用 Rollouts 调度器和强化学习调度器，取得了与商用指令调度器相似甚至更优的性能。

2.4 软件开发

高性能计算系统上的软件开发任务需要考虑大量和系统相关的因素，对软件工程师的要求极高。目前，TOP500 榜单领先的超级计算机系统提供了大量不同的基础软件开发环境，如 Frontier 超级计算机上的 AMD HIP[33]，富岳超级计算机上的自研优化编译器和通信库[34]，神威太湖之光上开发、优化并行程序的定制 OpenACC 工具[35]等。基础软件开发环境的多样性，要求工程师在软件开发和移植的过程中综合考虑软硬件架构因素，需要充足的经验和大量的测试，明显影响了软件开发过程的效率。

将 AI 技术融入软件开发，能够减少软件开发过程的负担，提升软件开发效率。Kousha 等[36-37]中设计并实现了用于将用户文字需求转换为 HPC 系统相关命令的智能 AI 接口，以方便开发人员进行研究与软件开发。Nichols 等[38]将大语言模型与 HPC 系统软件开发相结合，通过 HPC 程序数据集对大语言模型进行微调，使其能够自动对循环代码进行并行化，并对程序的并行性能进行分析。未来，随着人工智能、大模型的高速发展，AI 赋能的 HPC 平台软件开发工作流将获得全面加速，开发者能够更加方便地开发适合不同超级计算机平台的软件应用，并更加方便地进行软件性能和功耗的调优。

2.5 小结

人工智能技术在高性能计算领域的应用日益广泛，不断优化和完善的 AI 方法为 HPC 领域在硬件体系结构、操作系统资源管理、编译优化和软件开发等多方面的问题提供了更多的解决方案。目前，AI 方法在解决 HPC 领域问题时仍面临各种问题，尤其是在高质量训练数据不足时，需要通过数据增强、迁移学习、HPC 领域知识嵌入融合等策略来提高 AI 模型的泛化能力；同时，AI 在 HPC 领域的落地应用也存在各种瓶颈，包括软硬件兼容性、系统可靠性、低开销和快响应的需求等，需要通过对 AI 方法进行更加深入的研究来不断优化，最终构建能够有效解决问题的框架。

3 HPC for AI

HPC 算力的不断提升和人工智能领域对算力要求的不断提高相契合。目前，新型超级计算机系统正在积极为各类 AI 应用提供适配，并提供底层基础设施与上层应用的支持。

3.1 AI 基础设施

AI 基础设施包含了支撑人工智能应用开发、训练和部署的底层软硬件和工具。在 AI 算力需求日益增长的背景下，高算力 HPC 系统的软硬件逐渐成为 AI 基础设施的重要组成部分，新型 HPC 系统正在积极为各类 AI 应用开展适配工作。目前的领域前沿工作主要包含软件和硬件两个方面，其中软件主要是 AI 编译器，硬件主要分为 AI 芯片和智算中心。

3.1.1 AI 编译器

AI 编译器是一种特殊的编译器，其聚焦于对 AI 领域任务的专门优化，以提升模型推理性能，同时提供统一的编译框架，减少针对不同平台的代码修改与调试。Rotem 等[39]介绍了一种用于异构硬件的机器学习编译器，将传统的神经网络数据流图转换为两阶段的强类型中间表示，在两阶段分别进行特定领域优化和内存相关优化。该方法为多种硬件平台提供了支持。Chen 等[40]介绍的 TVM 框架，通过提供图级别和操作级别的优化，应对了深度学习特有的优化挑战，例如操作融合、映射到任意硬件原语，以及隐藏内存延迟等问题，并为不同硬件后端的高性能深度学习工作负载提供支持。Siemieniuk 等[41]介绍了一款名为 OCC 的 AI 编译器，其主要为基于忆阻器的特殊加速器设计，利用多级中间表示来逐步降低程序级别，使 ML 应用程序转换为到加速器硬件的映射，并通过实验证明了 OCC 能够可靠地识别和卸载常见的张量运算。

3.1.2 AI 芯片

AI 芯片是专门为人工智能计算加速设计的芯片，主要用于支持 AI 领域的常见运算，并对大规模加速进行支持。对于 HPC 系统而言，使用 AI 芯片，增加 HPC 系统对 AI 场景的支持，化“超算”为“智算”，符合 HPC 助力 AI 发展的要求。

目前常见的AI芯片主要分为如下几类[42]。

1)GPU芯片。GPU,即图像处理单元,其设计之初是为完成渲染计算机图形的任务,具有大量的光栅单元和纹理单元等。近年来,在GPU多核心的基础上逐渐发展出"流处理器"的概念,GPU开始在各类需要大规模并行计算任务的场景中使用。2010年,使用GPU加速的TH-1A登顶TOP500榜首,CPU+GPU异构超算系统开始全面取代传统的同构HPC系统[43]。2012年,在神经网络领域具有划时代意义的AlexNet网络问世,该网络不仅因在ImageNet数据集上表现突出而闻名,还因为只用两块GPU进行训练,让AI领域研究者发现了使用GPU加速AI计算的潜能[44]。如今,GPU成为HPC领域和AI领域广泛使用的异构加速器件之一,是各大超算、智算中心必需的基础硬件。

2)基于FPGA的加速芯片。FPGA,即现场可编程门阵列,是一种可以通过编程方式改变自身结构和功能的特殊器件,被广泛用于模拟实现各种功能的数字逻辑组件。FPGA具有灵活性和较高的性能,能够实现高算力和低延迟,满足AI对算力和大规模数据处理的要求。Qiu等[45]提出了一种使用FPGA平台为CNN网络计算提供加速的方法,除专门为CNN设计的FPGA结构外,还引入了数据量化步骤,使目标数据量转换为在FPGA上操作效率更高的定点数,对于典型应用VGG16-SVD,最终在精度只损失0.4%的情况下取得了比经典的FPGA加速方法明显更高的性能。

3)全定制的ASIC。全定制的ASIC(应用特定集成电路)是专门为特定应用和功能定制设计的集成电路。与FPGA相比,定制的ASIC虽不具有FPGA的灵活性,但能够以更低的功耗、更快的速度、更低的成本支持AI应用,是小型算力、数据中心的重要选择。Song等[46]提出一种名为C-Brain的深度学习加速器,用于解决现有的CNN加速器面对不同参数网络的性能不稳定问题。该加速器运行在ASIC上,在较大规模的CNN网络部分层上取得了4~8.3倍的加速,并较先前领域的先进加速器实现了28.04%的计算单元节能和90.3%的片上内存节能。

4)类脑芯片。类脑芯片是一种模仿生物神经元架构设计的芯片,这种芯片不使用经典的冯·诺依曼体系结构,而是采用存算一体结构,存储模仿突触,CPU模仿神经元,通信部件模仿轴突。这类芯片具有高性能、低功耗、可编程的特点,在支持AI领域研究方面具有极大潜能。IBM TrueNorth芯片是一种典型的类脑芯片[47],其将100万个模拟的神经元和2.56亿个模拟的突出集成在长宽仅数厘米的芯片中,在NeoVision2 Tower数据集上实现了高达80%的识别准确率,而能耗仅70mW。

3.1.3 智算中心

智算中心是AI+HPC融合理念下的新型算力基础设施,是专门用于处理运行人工智能任务的数据中心。这类数据中心主要配备适合进行AI计算的高性能计算系统,常使用GPU和TPU等适用于AI计算的加速器,以此支持大规模模型训练、推理等业务[48]。

目前,全球有大量知名IT企业参与到智算中心的建设当中,如国内的华为、百度、腾讯、商汤科技等,国外的谷歌、微软、亚马逊、IBM、Meta等,这些智算中心推动了IT企业新型业务的开展,为各行各业提供服务与支持,并进一步促进了AI和HPC领域的深入研究与发展。智算中心推动AI+HPC融合,助力各传统领域产业升级和转型,创造更多新兴产业、就业机遇和人才需求,推动IT技术形成新型社会生产力。

3.2 应用支持

HPC系统对AI类应用的支持,为AI方法在各个领域的大规模使用奠定了算力基础,其中既包括AI领域自身的应用,如大模型;也包括其他科学领域的相关应用,如智慧医疗、环境气候、物理学、材料学等。

3.2.1 大模型

大模型是当下最热门的AI研究之一,一个大模型的参数量可达数十亿,训练和微调过程需要调用大量算力,需要HPC系统的支持。Dash等[49]提出了一种在Frontier超级计算机上进行大语言模型分布式训练优化的方法,其整合了张量并行、管道并行、数据并行等技术,在大语言模型训练过程中能够达到GPU总吞吐量的31.9%以上,强可扩展性和弱可扩展性均表现突出。Narayanan等[50]提出一种在GPU集群上高效训练大语言模型的方法,主要聚焦于通过优化管道并行提升吞吐量,在使用3072块GPU对参数量为1×10^{12}的模型进行训练时,每块GPU可达到52%的吞吐量上限。Isaev等[51]提出了一种性能分析模型,通过收集软硬件特征以及模型自身特征,探索在给定约束条件下的最优模型训练配置;并在此基础上开发了一套能够寻找更优模型训练系统配置的开源工具,得到了新的系统配置设计方案。

3.2.2 智慧医疗

医疗是AI的重要应用领域之一,AI方法被广泛应用在医学影像分析、药物研发等场景,这些场景往往有着规模庞大的数据,需要使用AI方法,在HPC系统的协助下从大量的数据集中提取关键信息。Jain等[52]提出了一种增加模型在GPU上的并行度的方法,其取得了与现有模型并行框架更佳的性能;并将其应用于数字病理学,对图片影响进行分析。Stevenson等[53]、Sukumar等[54]以及Kadioglu等[55]在新冠肺炎疫情期间,使用AI方法,根据病毒的结构和候选药物的数据,对能够抑制新冠肺炎病毒的药物进行筛选;以HPC为算力支持,最终筛选出若干候选药物。实验证明,这些药物对病毒具有良好的抑制作用,加速了对抗新冠肺炎药物的研发。

3.2.3 环境气候

环境气候领域存在大量的建模和检测任务,对算力的需求较高,与AI方法的结合进一步提高了对算力的要求。Sood等[56]构建了一个物联网、大数据分析和HPC系统相结合的智能洪水检测和预报体系系统,将物联网设备广泛部署在各个地理区域,收集环境数据,并使用AI方法和HPC进行数据分析,以实现快速预测。Ichimura等[57]提出了使用AI方法和混合精度计算进行地震城市模拟的方法,并在Summit超级计算机上进行测试,其取得了25.3倍于基础方法的性能和更高的可扩展性。Bi等[58]提出一种基于AI方法的高精度全球天气预报方法,使用三维深度神经网络模型,基于39年的全球气象数据进行训练。他们开发的Pangu-Weather天气

预报模型，有效降低了传统数值模拟方法的开销，并保持了较高的预报精度。Maulik 等[59]建立了一种用于大气和海洋领域建模的模型，使用了基于正交分解的 LSTM 网络，并在 Theta 超级计算机上使用 NOAA 海表温度数据集进行验证，所建立的模型表现出较传统方法更高的性能和可扩展性。Li 等[60]提出了一种大气数据同化的快速模拟方法，使用深度学习技术加速大气数据同化过程和快速模拟小时尺度的大气现象；同时在新一代神威超算上建立了一套 DIDA 模拟原型系统，其具有快速、低开销、可扩展等优点。

3.2.4 *物理学和材料学*

物理学、材料学面临大量的模拟类问题与研究，HPC 系统既能够为这类场景提供模拟的算力，又能为 AI 融合的方法提供支持。Zhao 等[61]构建了一套基于 AI 方法来解决量子多体问题的框架，其整合了蒙特卡洛方法与基于卷积神经网络的波函数表示，并被部署在神威超级计算机上，在可扩展性和解的精度上取得了显著提升。Das 等[62]提出的在量子精度下进行大规模材料建模的新方法，结合了量子多体方法的准确性与密度泛函理论的计算效率，并在 DFT 中引入了机器学习方法，解决了长期以来存在的准确性和长度尺度之间的权衡问题，并在 Frontier 超级计算机上达到了 659.7 PFlops 的持续性能。

3.3 小结

HPC 为 AI 提供了强大的硬件算力基础，推动了大模型训练、智慧医疗、环境气候模拟等领域的发展。未来，HPC 将帮助 AI 在各个领域前沿研究中开展大规模应用，从而实现更广泛的科学突破和产业创新。需要注意的是，日益复杂的 AI 方法正不断对 HPC 系统提出更高的要求，这与 AI 方法被用于解决 HPC 自身问题中的需求相辅相成，需要 HPC 领域专家不断提升 HPC 系统的性能与兼容性，为 AI 的大规模应用提供稳定的支持。

4 未来展望与挑战

AI 和 HPC 领域的结合与发展，既符合两领域研究深入发展的需要，又满足了新兴产业对 AI 和 HPC 领域技术的需求。未来，AI 研究将逐渐追求多模态，逐渐向通用人工智能（Artificial General Intelligence，AGI）发展，并与更多生产生活领域相融合，对算力的要求呈现指数增长的趋势，需要 HPC 系统提供“智算”算力和新型应用程序的支持。对于 HPC 系统而言，随着算力突破 E 级，未来超算系统在性能、功耗、可靠性等方面会面临更多前所未有的问题与挑战，需要 AI 方法提供新的问题解决思路。与此同时，AI＋HPC 正逐渐成为各领域科学研究中不可或缺的思维，并在 IT 行业创造广泛的就业机会。

目前，AI 和 HPC 两领域的结合仍面临各种挑战，主要体现在以下几个方面。

1）AI＋HPC 融合研究需要良好的软硬件生态支持。AI 和 HPC 领域的融合，需要高性能的软件应用和硬件配置，而不同厂商提供的硬件配置和软件开发框架存在明显的差别，且软硬件生态的发展水平与领域研究的难易程度密切相关。目前，由 NVIDIA 提供的软硬件生态成为多数提供 AI＋HPC 服务的公司的选择，主要原因在于其 CUDA 生态的高成熟度：诸如 Tensorflow 和 Pytorch 等知名 AI 框架，以及 HPC 领域的 BLAS 和 FFT 等数学库，均对 CUDA 有良好的适配与支持；且 CUDA 本身对 C，C＋＋和 Python 等各种编程语言具有良好的兼容性。未来 AI＋HPC 深度融合的发展，需要供应商提供良好的软硬件生态，满足高效的硬件资源利用和简易的软件开发模式需求，高速推进领域研究和交叉学科的研究进程。

2）AI＋HPC 融合对系统的兼容性和可靠性要求极高。兼容性方面，AI 和 HPC 领域的计算任务特征存在明显的区别，AI 领域计算多使用单精度、半精度浮点运算，以牺牲精度来保证高吞吐量和低能耗；而 HPC 领域计算任务多使用双精度浮点运算，进行高精度的计算模拟，但运算速度较低，功耗较高。目前，前沿科学研究既需要使用低精度但快速的 AI 类计算高效解决问题，又需要高精度的 HPC 数值模拟对实际情况精准还原，AI＋HPC 计算融合发展的重要性与日俱增，混合精度计算的研究不断提上日程[63]，这需要计算系统综合考虑不同计算方式的特点，提供更加综合全面的系统兼容。可靠性方面，使用高性能计算系统开展前沿领域研究时，经常需要执行时间跨度较长的 AI 和 HPC 任务，如大模型训练、数值模拟等，这对计算系统的稳定运行和容错恢复能力提出了较高要求。目前，在大规模计算系统上部署的应用框架能够通过收集程序运行期间的硬件事件信息、定时的心跳机制等判断程序执行期间系统是否存在问题，并在遇到错误时使用检查点方法进行恢复[64]。

3）AI＋HPC 融合带来高昂的研究成本。AI 和 HPC 领域的融合研究，创造了大量的计算任务，在基础设施和系统运维等方面造成大量的成本开销。目前，智算中心、云计算中心基础设施建设的成本高达数十亿甚至数百亿元，每年能耗可达数百亿千瓦时，每年在运维上的开销已与建设成本相当[65]。AI＋HPC 算力需要的巨大成本，不仅需要国家在经济上、政策上予以支持，更需要 AI 和 HPC 各自领域的进一步突破，采用更加高效、低复杂度的 AI 算法，同时对 HPC 系统的能耗、散热等进行优化，以减少巨量算力消耗带来的高额成本，以及 AI＋HPC 融合研究对财力和物力的消耗。

4）AI＋HPC 融合研究需要领域交叉人才。无论是 AI 和 HPC 各自领域的突破，还是 AI＋HPC 融合在其他领域的应用，都需要 AI 和 HPC 领域乃至其他前沿研究领域的全能型人才。目前，国内开展 AIGC 研究的企业，如华为、百度、阿里、曙光、字节跳动等，提供大量的 AI＋HPC 相关新岗位，如推理引擎优化工程师、AI 平台优化工程师、人工智能解决方案架构师、高性能计算研究员等。同时，AI for Science 的发展，对包含医疗、环境科学、物理学、材料学等在内的各学科行业提出了更高的 HPC＋AI 程序开发要求，相关企业也在积极与 AI 企业合作，并招募 AI 和 HPC 专业人才，开展前沿领域交叉研究工作。AI＋HPC 融合研究对人才的能力水准提出了更高的要求，这些全能型人才将逐渐在各行各业发挥极其重要的作用，未来会创造更多 AI＋HPC 融合的相关岗位。

AI 与 HPC 领域融合，符合大数据、“智能＋”时代背景下各行各业前沿领域研究的算力需要，是两大领域实现新突破

的关键所在，也是国家战略中重要的一环。面对 AI 与 HPC 融合的必然趋势与挑战，需要各学科之间的紧密合作，使用 AI+HPC 融合方法解决学科疑难问题，实现科技自主创新。同时，还需要 AI 与 HPC 领域专家在 AI 与 HPC 核心技术上实现国产化的突破，以实现全面自主可控，进而推进国产多元异构 AI+HPC 服务器的研发部署，构建强大的 AI+HPC 融合算力基础设施，为科研、产业和社会的数字化转型提供坚实的技术基础。

结束语 本文回顾与总结了近十年 AI 和 HPC 技术在大数据和“智能+”背景下的发展与融合趋势，分析了 AI 技术在 HPC 硬件体系结构、操作系统资源管理、编译优化和软件开发几个方面的作用，揭示了 HPC 系统为 AI 在基础设施和软件应用层面的支持，并对 AI+HPC 领域融合的未来发展趋势和挑战进行了分析。AI+HPC 深度融合，是“智能+”背景下的必然趋势，不仅对两领域技术发展至关重要，更是众多新兴产业的动力核心。

参 考 文 献

[1] JOUBERT W, MESSER B, ROTH P C, et al. Learning to Scale the Summit: AI for Science on a Leadership Supercomputer [C]// 2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). 2022: 1246-1255.

[2] LYU T, SATO M, AOKI S, et al. CORTEX: Large-Scale Brain Simulator Utilizing Indegree Sub-Graph Decomposition on Fugaku Supercomputer[J]. arXiv: 2406.03762, 2024.

[3] FANG J, FU H, ZHAO W, et al. swdnn: A library for accelerating deep learning applications on sunwaytaihulight[C]// 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS). 2017: 615-624.

[4] ACCORDI G, GADIOLI D, PALERMO G, et al. Unlocking performance portability on LUMI-G supercomputer: A virtual screening case study[C]// Proceedings of the 12th International Workshop on OpenCL and SYCL. 2024: 1-4.

[5] LIU X, MCDUFF D, KOVACS G, et al. Large Language Models are Few-Shot Health Learners[J]. arXiv: 2305.15525, 2023.

[6] SMITH S, PATWARY M, NORICK B, et al. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model[J]. arXiv: 2201.11990, 2022.

[7] WANG Z, TANG Y, CHEN J, et al. Energy wall for exascale supercomputing[J]. Computing and Informatics, 2016, 35(4): 941-962.

[8] WANG R, LU K, CHEN J, et al. Brief introduction of tianhe exascale prototype system[J]. Tsinghua Science and Technology, 2020, 26(3): 361-369.

[9] MIRHOSEINI A, GOLDIE A, YAZGAN M, et al. A graph placement methodology for fast chip design[J]. Nature, 2021, 594(7862): 207-212.

[10] JEFFREY D. The Deep Learning Revolution and Its Implications for Computer Architecture and Chip Design[C]// ISSCC 2020. 2020.

[11] LTAIEF H, HONG Y, DABAH A, et al. Steering Customized AI Architectures for HPC Scientific Applications[C]// International Conference on High Performance Computing. 2023: 125-143.

[12] AN W, BI X, CHEN G, et al. Fire-Flyer AI-HPC: A Cost-Effective Software-Hardware Co-Design for Deep Learning[J]. arXiv: 2408.14158, 2024.

[13] ALSADIE D. Advancements in heuristic task scheduling for IoT applications in fog-cloud computing: challenges and prospects [J]. PeerJ Computer Science, 2024, 10: e2128.

[14] IFTIKHAR S, AHMAD M M M, TULI S, et al. HunterPlus: AI based energy-efficient task scheduling for cloud-fog computing environments[J]. Internet of Things, 2023, 21: 100667.

[15] FAN Y, LAN Z, CHILDERS T, et al. Deep reinforcement agent for scheduling in HPC[C]// 2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS). 2021: 807-816.

[16] NARANTUYA J, SHIN J S, PARK S, et al. Multi-Agent Deep Reinforcement Learning-Based Resource Allocation in HPC/AI Converged Cluster[J]. Computers, Materials & Continua, 2022, 72(3): 4375-4395.

[17] ZHANG D, DAI D, HE Y, et al. RLScheduler: an automated HPC batch job scheduler using reinforcement learning[C]// SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. 2020: 1-15.

[18] RANGANATH K, SUETTERLEIN J D, MANZANO J B, et al. MAPA: Multi-Accelerator Pattern Allocation Policy for Multi-Tenant GPU Servers[J]. arXiv: 2110.03214v1, 2021.

[19] JUNG H H, PEDRAM M P. Supervised Learning Based Power Management for Multicore Processors[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2010, 29(9): 1395-1408.

[20] AKSAR B, SENCAN E, SCHWALLER B, et al. Prodigy: Towards unsupervised anomaly detection in production hpc systems[C]// Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2023: 1-14.

[21] ISAKOV M, CURRIER M, DEL ROSARIO E, et al. A taxonomy of error sources in HPC I/O machine learning models[C]// SC '22: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. 2022.

[22] LI P, GUO Y, LUO Y, et al. Graph neural networks based memory inefficiency detection using selective sampling [C]// SC'22: International Conference for High Performance Computing, Networking, Storage and Analysis. 2022: 1-14.

[23] BOIXADERAS I, ZIVANOVIC D, MORÉ S, et al. Cost-aware prediction of uncorrected DRAM errors in the field[C]// SC'20: International Conference for High Performance Computing, Networking, Storage and Analysis. 2020: 1-15.

[24] XIAO Y, SONG Y, LIU J. Collaborative multi-agent deep reinforcement learning for energy-efficient resource allocation in heterogeneous mobile edge computing networks [J]. IEEE Transactions on Wireless Communications, 2024, 23(6): 6653-6668.

[25] GU B, ZHANG X, LIN Z, et al. Deep multiagent reinforcement-learning-based resource allocation for internet of controllable

things[J]. IEEE Internet of Things Journal，2020，8(5)：3066-3074.

[26] ASHOURI A H，KILLIAN W，CAVAZOS J，et al. A survey on compiler autotuning using machine learning[J]. ACM Computing Surveys(CSUR)，2018，51(5)：1-42.

[27] ASHOURI A H，MARIANI G，PALERMO G，et al. Cobayn：Compiler autotuning framework using bayesian networks[J]. ACM Transactions on Architecture and Code Optimization(TACO)，2016，13(2)：1-25.

[28] FURSIN G，KASHNIKOV Y，MEMON A W，et al. Milepost gcc：Machine learning enabled self-tuning compiler[J]. International journal of parallel programming，2011，39：296-327.

[29] MARTINS L G，NOBRE R，DELBEM A C，et al. Exploration of compiler optimization sequences using clustering-based selection [C]// Proceedings of the 2014 SIGPLAN/SIGBED Conference on Languages，Compilers and Tools for Embedded Systems. 2014：63-72.

[30] ASHOURI A H，BIGNOLI A，PALERMO G，et al. Micomp：Mitigating the compiler phase-ordering problem using optimization sub-sequences and machine learning[J]. ACM Transactions on Architecture and Code Optimization(TACO)，2017，14(3)：1-28.

[31] BENDIB N. Automatic Code Optimization in the MLIR Compiler Using Deep Reinforcement Learning[J/OL]. (2024-07-27) [2024-11-01]. http://dx. doi. org/10. 13140/RG. 2. 2. 17390. 42569.

[32] MCGOVERN A，MOSS J. Scheduling straight-line code using reinforcement learning and rollouts[C]// Advances in Neural Information Processing Systems. 1998.

[33] BUDIARDJA R D，BERRILL M，EISENBACH M，et al. Ready for the Frontier：Preparing Applications for the World's First Exascale System[C]// International Conference on High Performance Computing. 2023：182-201.

[34] WATANABE K，NOSE T，SUZUKI K，et al. Application development environment for supercomputer fugaku[EB/OL]. fujitsu. com/global/documents/about/resources/publications/technicalreview/2020-03/article07. pdf.

[35] FU H，LIAO J，YANG J，et al. The Sunway TaihuLight supercomputer：system and applications[J]. Science China Information Sciences，2016，59：1-16.

[36] KOUSHA P，JAIN A，KOLLI A，et al. "Hey CAI"-Conversational AI Enabled User Interface for HPC Tools[C]// International Conference on High Performance Computing. 2022：87-108.

[37] KOUSHA P，JAIN A，KOLLI A，et al. SAI：AI-Enabled Speech Assistant Interface for Science Gateways in HPC[C]// International Conference on High Performance Computing. 2023：402-424.

[38] NICHOLS D，MARATHE A，MENON H，et al. HPC-Coder：Modeling Parallel Programs using Large Language Models [C]// ISC High Performance 2024 Research Paper Proceedings (39th International Conference). 2024：1-12.

[39] ROTEM N，FIX J，ABDULRASOOL S，et al. Glow：Graph Lowering Compiler Techniques for Neural Networks[J]. arXiv：1805. 00907，2019.

[40] CHEN T，MOREAU T，JIANG Z，et al. TVM：An automated End-to-End optimizing compiler for deep learning[C]// 13th USENIX Symposium on Operating Systems Design and Implementation(OSDI 18). 2018：578-594.

[41] SIEMIENIUK A，CHELINI L，KHAN A A，et al. OCC：An Automated End-to-End Machine Learning Optimizing Compiler for Computing-In-Memory[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems，2022，41(6)：1674-1686.

[42] BAVIKADI S，DHAVLLE A，GANGULY A，et al. A survey on machine learning accelerators and evolutionary hardware platforms[J]. IEEE Design & Test，2022，39(3)：91-116.

[43] YANG X J，LIAO X K，LU K，et al. The TianHe-1A supercomputer：its hardware and software[J]. Journal of computer science and technology，2011，26(3)：344-351.

[44] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. Imagenet classification with deep convolutional neural networks[C]// NIPS 2012. 2012.

[45] QIU J，WANG J，YAO S，et al. Going deeper with embedded FPGA platform for convolutional neural network[C]// Proceedings of the 2016 ACM/SIGDA International Symposium on Field-programmable Gate Arrays. 2016：26-35.

[46] SONG L，WANG Y，HAN Y，et al. C-Brain：A deep learning accelerator that tames the diversity of CNNs through adaptive data-level parallelization[C]// Proceedings of the 53rd Annual Design Automation Conference. 2016：1-6.

[47] DEBOLE M，TABA B，AMIR A，et al. TrueNorth：Accelerating From Zero to 64 Million Neurons in 10 Years[J]. Computer，2019，52(5)：20-29.

[48] SUN C，DU C，LI X，et al. Research on key technologies of the Smart Computing Center[J]. Communications management and technology，2024：33-37，52.

[49] DASH S，LYNGAAS I R，YIN J，et al. Optimizing distributed training on frontier for large language models[C]// ISC High Performance 2024 Research Paper Proceedings (39th International Conference). 2024：1-11.

[50] NARAYANAN D，SHOEYBI M，CASPER J，et al. Efficient large-scale language model training on gpu clusters using megatron-lm[C]// Proceedings of the International Conference for High Performance Computing，Networking，Storage and Analysis. 2021：1-15.

[51] ISAEV M，MCDONALD N，DENNISON L，et al. Calculon：a methodology and tool for high-level co-design of systems and large language models[C]// Proceedings of the International Conference for High Performance Computing，Networking，Storage and Analysis. 2023：1-14.

[52] JAIN A，AWAN A A，ALJUHANI A M，et al. GEMS：Gpu-enabled memory-aware model-parallelism system for distributed dnn training[C]// SC20：International Conference for High Performance Computing，Networking，Storage and Analysis. 2020：1-15.

[53] STEVENSON G A，JONES D，KIM H，et al. High-throughput virtual screening of small molecule inhibitors for SARS-CoV-2

protein targets with deep fusion models[C]// Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2021: 1-13.

[54] SUKUMAR S R, BALMA J A, RICKETT C D, et al. The convergence of HPC, ai and Big Data in rapid-response to the COVID-19 pandemic[C]// Smoky Mountains Computational Sciences and Engineering Conference. 2021: 157-172.

[55] KADIOGLU O, SAEED M, GRETEN H J, et al. Identification of novel compounds against three targets of SARS CoV-2 coronavirus by combined virtual screening and supervised machine learning[J]. Computers in Biology and Medicine, 2021, 133: 104359.

[56] SOOD S K, SANDHU R, SINGLA K, et al. IoT, big data and HPC based smart flood management framework[J]. Sustainable Computing: Informatics and Systems, 2018, 20: 102-117.

[57] ICHIMURA T, FUJITA K, YAMAGUCHI T, et al. A fast scalable implicit solver for nonlinear time-evolution earthquake city problem on low-ordered unstructured finite elements with artificial intelligence and transprecision computing[C]// SC18: International Conference for High Performance Computing, Networking, Storage and Analysis. 2018: 627-637.

[58] BI K, XIE L, ZHANG H, et al. Accurate medium-range global weather forecasting with 3D neural networks[J]. Nature, 2023, 619(7970): 533-538.

[59] MAULIK R, EGELE R, LUSCH B, et al. Recurrent neural network architecture search for geophysical emulation[C]// SC'20: International Conference for High Performance Computing, Networking, Storage and Analysis. 2020: 1-14.

[60] LI Y, JU X, XIAO Y, et al. Rapid simulations of atmospheric data assimilation of hourly-scale phenomena with modern neural networks[C]// Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2023: 1-13.

[61] ZHAO X, LI M, XIAO Q, et al. Ai for quantum mechanics: High performance quantum many-body simulations via deep learning[C]// SC'22: International Conference for High Performance Computing, Networking, Storage and Analysis. 2022: 1-15.

[62] DAS S, KANUNGO B, SUBRAMANIAN V, et al. Large-scale materials modeling at quantum accuracy: Ab initio simulations of quasicrystals and interacting extended defects in metallic alloys[C]// Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2023: 1-12.

[63] JHA S, PASCUZZI V R, TURILLI M. Ai-coupled hpc workflows[J]. arXiv: 2208.11745, 2022.

[64] JIANG Z, LIN H, ZHONG Y, et al. MegaScale: Scaling large language model training to more than 10,000 GPUs[C]// 21st USENIX Symposium on Networked Systems Design and Implementation(NSDI 24). 2024: 745-760.

[65] PANDEY S K, SINGH K P, DHAR P, et al. Green Computing: Importance, Approaches, and Practices[M]// 6G Connectivity-Systems, Technologies, and Applications. River Publishers, 157-186.

TAN Zhengyuan, born in 2002, postgraduate. His main research interests include high performance computing and so on.

CHEN Juan, born in 1980, Ph.D, professor. Her main research interests include high performance computing, low-power compiler and power management.

(责任编辑:李亚辉)