架起智能计算的“通用桥梁”

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/36932.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

架起智能计算的“通用桥梁”。



汪玉在课堂上。受访者供图

■本报记者 田瑞颖

“上大舞台，做大事。”这个信念早已融入清华大学教授汪玉团队的科研基因。

在国家迈向人工智能（AI）强国的进程中，算力已成为新一代关键基础设施。然而，一个严峻的现实摆在面前：AI算法对算力的需求每3至4个月翻一番，远快于硬件性能18至24个月的迭代周期。特别是在先进制程工艺受限的背景下，如何利用现有硬件资源突破算力瓶颈、提升计算能效，成为亟待攻克的核心难题。

国家自然科学基金企业创新发展联合基金项目“面向智能应用场景的大规模多源异构计算架构研究”（以下简称企业创新发展联合基金项目），为这一瓶颈的突破带来了希望。

汪玉团队联合中国科学院计算技术研究所研究员杜子东团队、中国电子科技集团公司第十五研究所（以下简称中电十五所）首席专家兼研究员杨军团队，开启了一段为期4年的协同攻关。他们要在纷繁复杂的智能应用与多样化的计算硬件之间，架设一座高效、统一的“桥梁”。

“ 通过企业创新发展联合基金项目，我们真正实现了企业‘ 出题 ’、科学家‘ 答题 ’。 ”汪玉在接受《中国科学报》采访时表示，“ 我们以国家智能算力能效瓶颈为牵引，聚焦多源异构计算架构创新，在智能应用的真实场景中进行架构级创新，回应了国家在智能计算领域提升自主可控能力的迫切需求。 ”

破解智能时代的“ 算力迷宫 ”

AI的世界，犹如一个巨大的“ 算力迷宫 ”。

“ 迷宫 ”的上层是千变万化的应用场景，从需要毫秒级响应的智能推理，到处理海量数据的大规模训练；从服务类机器人到智能车联网，任务需求复杂多样。而“ 迷宫 ”的下层是形态各异的硬件“ 军团 ”，诸如通用性强但能效偏低的中央处理器（CPU）、适合大规模并行计算但成本高昂的图形处理器（GPU）、灵活可编程能效较高的可编程门阵列（FPGA），以及为特定算法定制的极致能效的应用集成电路（ASIC）芯片等。

汪玉解释，提升计算能效通常有三条主要路径：一是缩小工艺和芯片尺寸，二是设计面向特定任务的定制化加速器，三是通过新器件与新模型协同发展突破现有架构瓶颈。

为了弥补AI算力需求与硬件性能之间日益扩大的差距，他们希望通过异构计算将上述三条路径的优势进行融合。“ 异构计算能够根据不同应用场景选取最匹配的硬件类型，并在具体场景中灵活结合上述能效提升手段，从而有效应对AI计算面临的能效挑战。 ”汪玉说。

2020年，清华大学等3家单位正式合作，启动了企业创新发展联合基金项目，旨在通过在多智能计算框架和多智能计算硬件平台中引入中间层抽象，实现对多框架多硬件的统一接口，并进一步实现智能计算任务在数据中心的快速与高能效调度。

“ 企业创新发展联合基金项目启动时，大模型还没有像今天这样席卷全球，但我们已经预见到，应用和硬件的双重多样性将使它们之间的‘ 匹配 ’变得异常困难。 ”汪玉说。

他进一步解释，如果有N种应用框架需要映射到M种硬件平台上，没有统一的中间层，就需要设计“ N×M ”种映射方法，这种复杂度和工作量是不可持续的。而这个问题在产业界表现得尤为突出，中电十五所的很多关键应用场景就因此受制。

多家单位联合攻关下，一种创新解决方案应运而生，那就是在应用与硬件之间构建一个“ 中间表示层 ”。这个中间层通过统一的算子抽象和硬件虚拟化，将复杂度从“ N×M ”降低至“ N+M ”，最终提高在异构硬件平台上的开发及计算效率。

“ 中电十五所作为重要的产业需求方，他们在信息化建设中遇到的真实问题为我们的研究提供了明确的方向和验证场景。 ”汪玉表示。

从“ 单兵作战 ”到“ 国家队 ”

企业创新发展联合基金项目的资助让3个团队实现了强强联合，优势互补，打造了“ 基础-应用-工程 ”贯通的研究体系，从源头上突破传统单一学科、单一机构的研究边界。

汪玉团队在低功耗电路与软硬件协同优化领域拥有多年研究经验，在机器学习加速器硬件结构与

软件优化方面处于国际领先水平，曾获中国计算机学会技术发明奖一等奖；杜子东团队在深度学习硬件加速器领域处于国际领先水平；杨军团队在高效分布式计算架构、大规模智能应用部署与中间件支持方面有很多积累，中电十五所还基于多种异构CPU构建了自主可控的云高效计算资源调度平台，具有实际的应用异构算力和应用场景。

在企业创新发展联合基金项目的资助下，科研团队经过4年的潜心研究与工程实践，在关键技术上取得了一系列进展。

在基础理论层面，团队在算子定义与抽象方面取得了重要进展。他们提出的面向神经网络的统一算子抽象，实现了稀疏/超稀疏算法的高效硬件加速。实验数据显示，这项技术可以将特定算法速度最高提升10倍，同时将存储需求最多压缩至十分之一。

在核心技术层面，硬件资源抽象方面的创新尤为突出。研究团队构建的快速系统仿真框架，将多任务分布式异构硬件仿真速度提升了1到2个数量级，同时将算法迁移性能下降控制在1%以内。这项技术的突破，使得大规模异构计算系统的调试和优化效率实现了质的飞跃。

在系统优化层面，他们在异构任务调度方面实现了“ 架构-调度-映射 ”的联合优化。通过创新性的调度算法，他们将调度开销降至十分之一以下，能耗延时积最低降至九分之一。这项技术创新显著提升了系统能效比，为绿色计算提供了技术支撑。

在平台建设层面，研究团队搭建的原型系统支持PaddlePaddle、PyTorch等多种AI框架，以及国产CPU、ASIC、FPGA和英伟达GPU等硬件，确保统一中间层技术的可落地、可验证。

“ 我们从‘ 单兵作战 ’变成了协同共进的‘ 国家队 ’。 ”汪玉说。

“ 上大舞台，做大事 ”

“ 科研成果不应只满足于‘ 上书架 ’，要确保成果能够转化为实际应用，实现‘ 上货架 ’的目标。 ”汪玉说。

在开展企业创新发展联合基金项目期间，研究团队承接了一项极具挑战性的任务，为某单位智能计算基础平台竞赛搭建支撑系统。这意味着他们要在2个月内，将研究成果转化为能稳定支撑近200台服务器、180余支队伍、近千名选手在线竞技的实战平台。

所有人员都进入了“ 战时状态 ”，平均每天近16小时的工作是常态。在这支队伍中，不乏刚步入科研生涯的学生和青年科研人员，没有人感到倦怠，反而斗志昂扬。“ 大家想的就是上大舞台，做大事，机会来了，那就抓住干！ ”汪玉说。

最终，他们顺利地完成了这次“ 大考 ”，获得了参赛队伍对系统稳定性的一致好评。“ 那段时间虽然披星戴月地工作，但看着系统一天天稳定起来，最终完美地支撑那么多选手同台竞技，这种成就感是无可替代的。 ”对于这段经历，汪玉团队成员曾书霖感慨道。

这些成果迅速走出了实验室，形成了既“ 上书架 ”又“ 上货架 ”的良性循环，展现出巨大的应用价值和战略意义。

企业创新发展联合基金项目中孕育出的技术已深度应用于中电十五所的“ 小可 ”大语言模型等产

品中，致力于在国产芯片上实现高效的训练与推理。“ 在一些特定环境下，数据无法公开，必须在一个私域内用国产芯片进行高效计算。我们的技术让这成为可能，并在重点领域得到了应用。 ”汪玉说。

基于上述技术的启发，汪玉希望推动科技成果转化走向产业化应用。2023年5月，在清华大学的支持和推动下，汪玉团队成立了上海无问芯穹智能科技有限公司（以下简称无问芯穹）。该公司专注于解决大模型在多元异构芯片上的高效部署问题，支撑包括大语言模型、科学计算在内的多种智能计算应用，旨在大力推动企业创新发展联合基金项目成果实现产业化应用。

成立第二年，无问芯穹就获得了“ 2024年上海市重点服务独角兽（潜力）企业 ”“ 2024年中国AI算力层创新企业 ”“ 2024年中国人工智能行业高科技高成长企业 ”等荣誉。截至目前，无问芯穹已获得超10亿元的融资，市场前景广阔。

汪玉告诉《中国科学报》，通过参与企业创新发展联合基金项目，青年科研人员和研究生培养了系统思维和工程能力。“ 学生在这个过程中能看到整个闭环，能感受到产业是什么样的，而不仅仅停留在实验室里。 ”

无问芯穹的联合创始人戴国浩就是从这个项目中走出来的青年代表。“ 参与项目的过程，让我深刻理解了从实验室技术到产业产品的完整路径。这种经验对于我们在企业推进技术创新至关重要。 ”

《中国科学报》：你和团队在企业创新发展联合基金项目的科学研究中，对协同攻关有哪些切身体会？

汪玉：在企业创新发展联合基金项目的实施中，我们有了更高的使命感和责任感，必须解决产业的真问题，啃下“ 硬骨头 ”。

这一系列成果的取得得益于在企业创新发展联合基金项目支持下，团队共享平台、算力、数据和测试资源，直接推动了科研资源流动与成果转化。

《中国科学报》：请问你对企业创新发展联合基金项目未来部署有什么具体建议？

汪玉：着眼于联合基金的未来发展，我建议进一步强化“ 科研+中试 ”的联动，打破科研与中试阶段的衔接壁垒，推动技术研发与工程化验证的深度融合，让实验室成果更顺畅地走向产业化前端；设立更大力度的后续滚动支持机制，让那些在联合基金项目中表现出色、展现出巨大应用潜力的团队，获得持续、稳定的支持。

《中国科学报》(2025-11-24 第4版 自然科学基金)
作者：田瑞颖 来源：中国科学报

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发