CCF推荐A类国际学术会议介绍

MICRO 2021
——拥抱更广阔的未来

张明喆[1] 路 航[2]
[1] 中国科学院信息工程研究所
[2] 中国科学院计算技术研究所

关键词：计算机体系结构 MICRO

2021年10月18~22日，2021年度国际微体系结构年会（Annual IEEE/ACM International Symposium on Microarchitecture，MICRO）在线上召开。受新冠疫情影响，本届会议连续第二年由希腊雅典大学承办。MICRO自1972年创办以来，至今召开过54届，收录论文1964篇，中国大陆的高校、科研机构和企业共有42篇论文被收录，可见其难度。

拥抱未来——泛加速器与安全

本届MICRO共收到投稿430篇，数量仅次于2020年的446篇，为历史上第二多。会议最终录用了94篇，录用率为22%，为历史上最多的一次。在录用的全部文章中，除了低功耗（energy efficiency & low power）、访存优化（memory）、并行化（parallelism）等传统话题外，加速器（accelerator）和安全（security）成为了最热门的话题。19个分论坛（session）中，共有6个与加速器相关。与往届会议多关注人工智能加速器不同，本届会议中加速器相关工作关注的应用领域更加广泛，例如，Session 4B的5篇论文关注加速器在云计算及大型数据中心的应用；Session 5A的5篇论文涉及硬件加速器在自动驾驶、虚拟现实和病毒检测等领域的应用。此外，安全与隐私保护是本届会议的另一个热门话题，共有3个分论坛与安全相关。其中，Session 3A的5篇论文关注硬件对隐私保护技术的支持；Session 10A关注硬件自身安全性的提升。会议论文涉及领域的日趋多样化，从侧面体现了计算机体系结构研究正在与越来越多的学科和场景相融合，这也为未来的学术研究提供了更广阔的空间。

特邀报告—来自工业界和学术界的思考

本届MICRO邀请了来自工业界和学术界的3位嘉宾作特邀报告。

来自AMD的首席架构师迈克尔·克拉克（Michael T. Clark）在报告中介绍了近年来备受瞩目的AMD Zen系列处理器的相关情况，Zen2与Zen3的对比如图1所示。迈克尔首先回顾了Zen系列处理器的发展历程以及每一代Zen处理器核设计过程中的思考。然后他详细介绍了Zen3架构的设计细节，并通过大量测试数据揭示了工程实践中的不同选择对处理器性能的影响。最后迈克尔从芯片集成的角度介绍了Zen3面临的挑战，并简要介绍了他对未来工作的思考和展望。该演讲展示了工业界在实际产品设计过程中对不同要素进行取舍的思考方式，这也完美契合了MICRO促进工业界与学术界沟通的初衷。

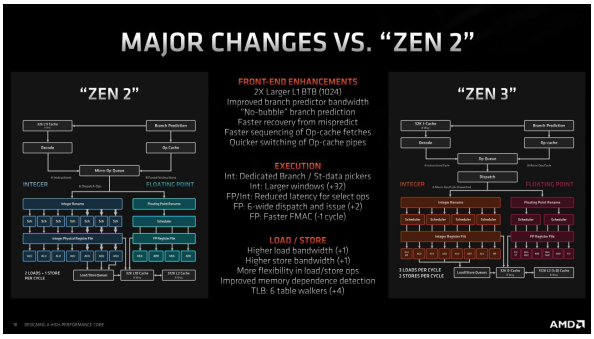


图1　通过ZEN3与ZEN2的对比，揭示未来发展趋势

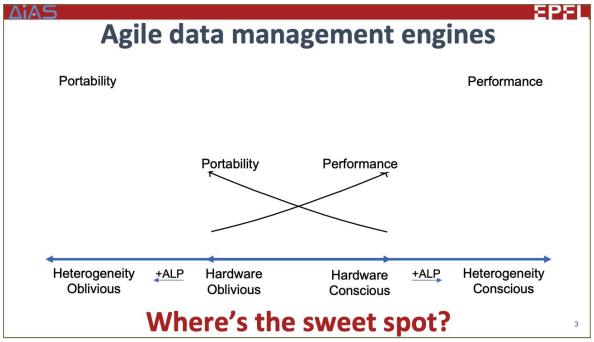


图2　实时数据处理的主要问题

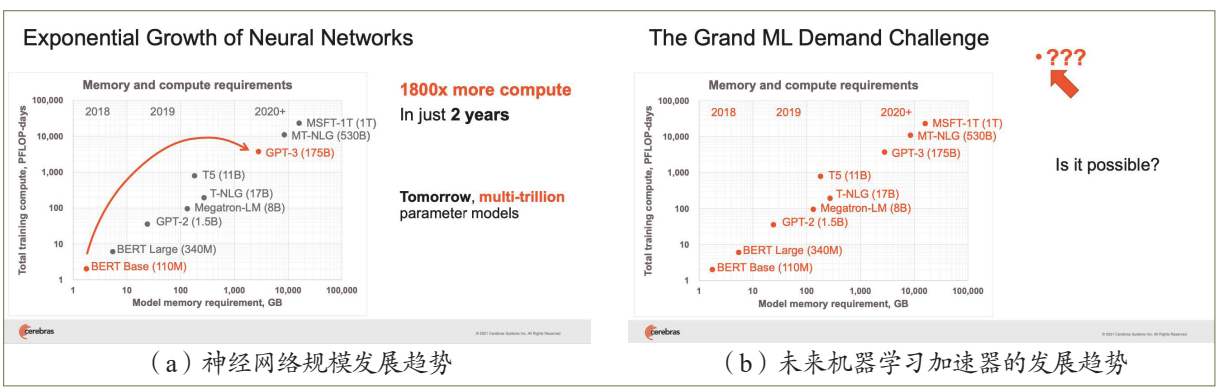
来自瑞士洛桑联邦理工学院（EPFL）的阿纳斯塔西娅·艾拉玛基（Anastasia Ailamaki）教授在报告中介绍了她的团队近年来基于实时处理技术加速数据管理系统的工作。艾拉玛基介绍了实时数据管理的发展历史及其面临的主要问题（见图 2）。她从硬件、负载、运行时数据等多个角度，介绍了如何利用异构硬件提升实时智能数据库系统的处理能力。

美国 AI 芯片创业公司 Cerebras 作为近年来人工智能加速器领域的新生力量，以其独有的晶圆级加速器芯片（wafer-scale chip）受到了全世界的关注。在本次会议上，Cerebras 的首席架构师肖恩·利（Sean Lie）详细介绍了其对晶圆级加速器设计的思考。他首先分析了神经网络计算的基本特征以及在传统体系结构上遇到的瓶颈，并介绍了 Cerebras 的团队如何利用软硬件协同设计的思路，从数据通路、片上存储分配、片上互连、软件映射和执行模型以及高效稀疏神经网络等不同角度进行创新，并将成果最终集成在晶圆级加速器芯片中。最后肖恩·利还介绍了在 Cerebras 架构设计空间探索过程中如何对不同要素进行取舍，并展望了未来机器学习加速器的发展（见图 3）。

MICRO颁奖

本届 MICRO 会议颁发了 2021 年度的罗摩克里希那·劳奖（B. Ramakrishna Rau Award）。该奖项是为了纪念著名计算机科学家 B. Ramakrishna Rau 而设立，每年颁发一次，以奖励在计算机体系结构领域作出杰出贡献的研究人员。已故著名计算机科学家高光荣教授曾于 2017 年获得过该奖。2021 年度的该奖项颁给了得克萨斯农工大学的丹尼尔·吉梅内斯（Daniel Jiménez）教授，以表彰其在神经分支预测领域的贡献。

本届会议的另一个重要奖项时间检验奖（Test of Times Award）颁发给了两篇发表于 2003 年的 MICRO 论文。第一篇是迪恩·图尔森（Dean Tullsen）教授的“Single-ISA Heterogeneous Multi-Core Architectures: The Potential for Processor Power



（a）神经网络规模发展趋势　　（b）未来机器学习加速器的发展趋势

图3　神经网络规模发展趋势和未来机器学习加速器的发展

Reduction”。在该论文中，作者深入分析了单一指令集的异构多核架构在降低处理器能耗方面的表现。在 20 年后的今天，ARM、英特尔、苹果等行业巨头纷纷推出 big.LITTLE 架构的单指令集异构多核处理器，并在移动端和桌面处理器市场取得了巨大成功，足以证明该篇论文对学术界和工业届的影响力。另一篇获得时间检验奖的论文是托德·奥斯汀（Todd Austin）教授的“Razor: A Low-Power Pipeline Based on Circuit-Level Timing Speculation”。该篇论文介绍了一种动态监测并纠正由于电压调节导致的电路错误的方法，这种方法已被应用于多款商用处理器中。

作为每届会议最受关注的奖项，本届 MICRO 的最佳论文奖由杜克大学的谢知遥获得。其论文“APOLLO: An Automated Power Modeling Framework for Runtime Power Introspection in High-Volume Commercial Microprocessors”介绍了一种用于分析商用处理器功耗的自动化分析模型，可以对商用 CPU 上的瞬时功耗进行快速预测，准确率可达 90%~95%，而其硬件开销仅为 0.2%。

MICRO与中国大陆科研机构

在 MICRO2021 录用的全部论文中，第一作者单位在中国大陆的论文共有 5 篇，分别来自清华大学、北京大学、中科院计算所、天津大学和腾讯量子实验室，论文内容涉及存内计算、稀疏数据处理、自动驾驶、量子计算等多个领域。笔者统计了中国大陆地区高校、科研机构和企业在过去 54 届 MICRO 会议中的论文发表情况，发现可大致分为两个阶段：1985—1993 年，中国大陆地区一共发表 9 篇论文，全部来自清华大学，研究内容集中在指令系统和流水线优化，其中 7 篇论文的第一作者为苏伯珙教授（1987 年两篇，1985 年、1986 年、1988 年、1990 年、1991 年各一篇）；2009 年以后，随着国内计算机体系结构领域研究水平的提高，大陆地区共计发表了 34 篇论文，内容涉及控制流优化、神经网络加速器、硬件安全等多个领域。目前国内在 MICRO 上发表论文最多的单位是清华大学（16 篇），其次是中科院计算所和北京大学（各 4 篇）。然而，截至 2021 年，中国大陆科研机构作为第一作者在 MICRO 上也仅发表了 43 篇文章，占过去 54 届 MICRO 总发文量（1964 篇）的 2.2%。这从侧面说明中国大陆地区的计算机体系结构研究仍有很大的提升空间。

基于比特级稀疏并行性的通用深度学习加速方法

笔者参加本届 MICRO 的主要任务是在大会上报告论文《基于比特级稀疏并行性的通用深度学习加速方法》。由于新冠疫情，所有报告均为线上，组织方在雅典向全世界从事体系结构研究的学者直播所有大会报告环节。

为了达到更高的精度，深度学习模型规模不断增大，而与之对应的，深度学习加速器的性能和能效也应当随着模型规模的增大逐渐提高。然而，由于电池寿命、功耗预算以及成本等的限制，特别是在以机器人和智能手机为代表的人工智能物联网（AIoT）设备上，硬件设计师并不愿意由于深度神经网络复杂度的增加而投入更多的计算资源。因此，提高加速器的能效在高性能（云端 AI 芯片）和低功耗（边 / 终端）的场景中都是非常必要的。

一方面，以剪枝方法为代表的“值级别”稀疏加速方法的应用空间越来越小，从算法优化的角度来看，如果无损精度是第一设计原则，无论采用何种剪枝方法，都需要花费大量的时间来探索这种压缩比和精度的最佳权衡，以平衡模型精度和体量。另一方面，从硬件实现的角度来看，利用值的稀疏性也不可避免地引入更为复杂的加速器架构和数据流设计。针对这些现有问题，我们创新性地提出利用神经网络模型天然暴露出的比特级稀疏性及其“并行性”来加速神经网络计算。我们提出的 bitlet 加速器具备两个主要特性：(1) 通过独特的架构设计，将多种数据精度类型的乘加计算（Multiply Accumulate, MAC）集成于同一加速单元，包括 32/16

比特以及 bfloat16 类型浮点数、1~24 比特定点数和整数；此通用性可以将各种智能计算场景所需要的计算统一起来，使用相同架构的计算单元，从而可以使用统一的指令集、数据流架构甚至软件栈，从而极大地提高了用户设计智能芯片的效率并降低设计成本。(2) 高效利用操作数的比特级稀疏性加速模型计算，此特性不但可以在边 / 终端设备上提升定点数的推理计算速度，还可以在云端场景中加速训练过程，从而使得终端用户可以方便地定制自己的智能芯片以适配智能超算、智能手机、智能机器人乃至 AIoT 等诸多场景。

总结

MICRO 作为计算机体系结构领域的顶级会议之一，其关注微架构和硬件实现的特色吸引了来自学术界和工业界的广泛关注。在过去的 50 多年中，学术界与工业界的交流与碰撞促进了一大批新技术的诞生和发展，并对当今计算机产业产生了深远影响。从 MICRO 的发展历史来看，我国在体系结构领域的研究曾经与世界领先水平接近，但由于历史原因，导致出现了近 20 年的断层。但令人欣喜的是，我们通过对近 10 年来 MICRO 论文发表趋势的分析看到，中国大陆的科研机构和企业在体系结构领域的研究水平不断提升，来自大陆的研究人员的研究视野也在不断拓展。笔者希望未来能够有更多的中国大陆研究人员参与到体系结构研究中，并在 MICRO 这样的国际顶级学术会议上展示自己，与国际同行交流，也希望能有更多来自中国的体系结构研究成果对业界和未来产生深远的影响。 ■

张明喆

CCF 专业会员。中国科学院信息工程研究所副研究员。主要研究方向为新型存储器结构优化、近数据计算加速器结构和硬件安全。

zhangmingzhe@iie.ac.cn

路　航

CCF 专业会员。中国科学院计算技术研究所副研究员，中国科学院青年创新促进会会员，中科院计算所“新百星计划”入选者。主要研究方向为可定制计算。

luhang@ict.ac.cn

（本文责任编委：**杜子东　王　斌**）

关于CCF会费调整的补充说明

2022 年 7 月 1 日零时起，CCF 专业会员会费调整至 360 元 / 年 / 人，学生会员会费仍为 50 元 / 年 / 人。相关事宜补充说明如下：

关于终身会员（Life Member）

满足下列条件之一的 CCF 会员可申请成为终身会员，终身会员权益与其他类型会员相同。

1. 会员男性年满 65 周岁、女性年满 60 周岁，且连续会龄不少于 15 年，可成为终身会员，免缴会费；
2. 会员男性满 55 周岁、女性满 50 周岁，参照上述第 1 条年龄差一次性缴纳足额会费，可成为终身会员。

关于边远地区会费减免

现工作或学习地址位于以下省份或自治区的 CCF 会员可申请减免会费：甘肃、贵州、西藏、云南、青海、新疆、广西。会费减免至调整之前的标准：200 元 / 年 / 人（暂定，每年根据各地居民人均可支配收入进行调整）。

1. 兰州、贵阳、昆明、西宁、乌鲁木齐、南宁已成立 CCF 会员活动中心，会员可以分部为单位申请；
2. 同一单位申请人数超过 10 人的，可由所在单位申请；
3. 如不符合以上条件，可由一位召集人组织申请，申请人数不少于 10 人；
4. 符合条件者可在规则生效后登录 CCF 会员系统 – 会员中心申请。