刚刚，华人再获IEEE/ACM微体系顶会唯一最佳论文奖！陈怡然组提出APOLLO全新方法

Original 新智元 新智元 2021-10-22 11:51

新智元报道

来源：IEEE/ACM

编辑：好困 小咸鱼

【新智元导读】昨日，第54届IEEE/ACM计算机体系结构顶会MICRO 2021开奖！杜克大学的谢知遥、陈怡然教授团队获得唯一最佳论文奖。该研究可能是首次实现了用AI技术对芯片全生命周期进行管理。

又拿大奖了！

北京时间21日晚，杜克大学的谢知遥荣获第54届IEEE/ACM微体系结构国际研讨会最佳论文奖（Best Paper Award）。

恭喜！

他的导师陈怡然教授也非常高兴地发了一条微博表示祝贺。

IEEE/ACM微体系结构国际研讨会是展示和讨论先进计算和通信系统的创新微体系结构思想和技术的首要论坛。

接收论文的主题包括硬软件的很多领域，其中就包括新兴应用领域的架构，如深度学习、机器学习、关系计算、神经形态、量子计算，还有加速器和异构架构设计等。

本次研讨会汇集了与微体系结构、编译器、芯片和系统相关领域的研究人员，就传统微体系结构主题和新兴研究领域进行技术交流。

2021年，MICRO将作为全球在线活动举办，由主办城市希腊雅典转播。主研讨会将于10月19日星期二至10月21日星期四举行，专题研讨会的时间为10月18日星期一和10月22日星期五。

在此之前，中科院计算所的陈云霁于2014年获得过该奖。

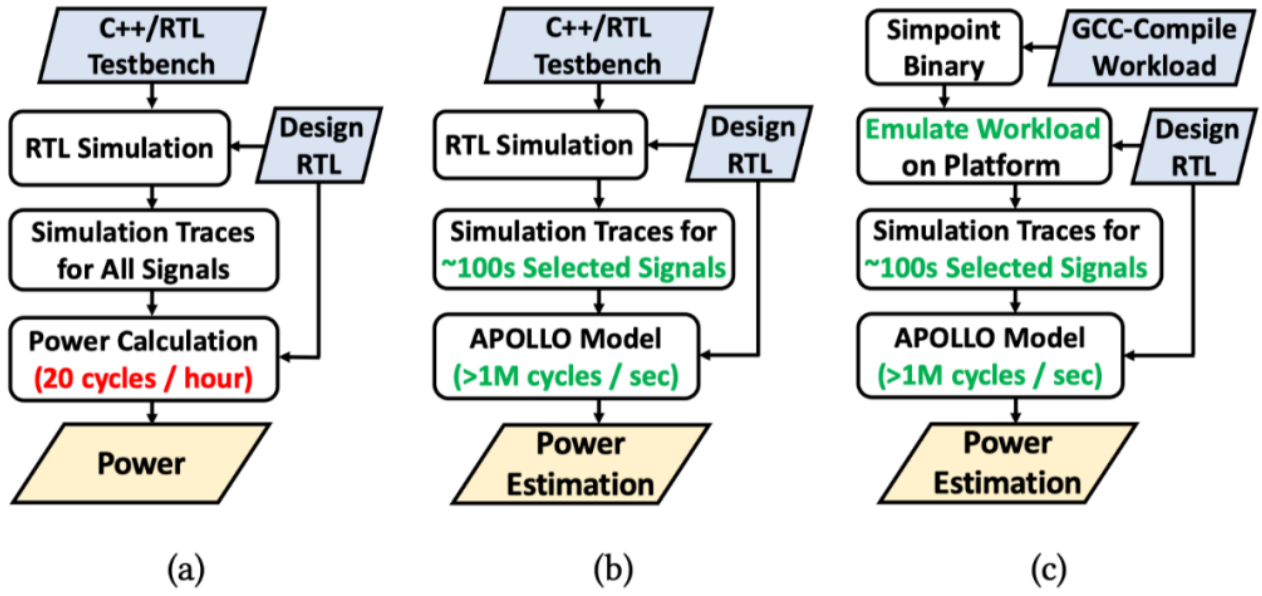
获奖论文

作为计算机的核心部件， CPU的性能每一年都在不断提高 。而随着摩尔定律日益失效，获得每一代CPU性能提升变得越来越困难。

为了获得更好的性能，CPU设计师们不得不集成更多的晶体管，并且使用更多的并行计算模块。这导致CPU内的功耗与电流需求不断增加，CPU的功耗成为一个日益严重的问题。

更糟糕的是，相比之下 CPU的输电（power delivery）技术进展依然缓慢。例如在先进的5纳米制程上，CPU输电线（power delivery network）上的电阻会非常大。另外现有的封装技术无法提供当代CPU需要的快速变化的电流。

这些都会导致最终CPU获得的电压低于设计电压，从而降低CPU运算速度，甚至导致功能发生错误。这些严峻的挑战直接限制了CPU的进一步性能提升。



设计时每周期功率分析流程。(a) 商业应用；(b) 基于APOLLO；(c) APOLLO与仿真器辅助相结合

要解决这些问题，首先在CPU设计阶段设计师需要充分考虑各种复杂的功耗场景，而在CPU运行阶段则需要实时避免过高的功耗与过快的电流需求变化。

那么，无论是在CPU设计还是运行阶段，都需要对功耗进行准确、快速、低开销和高分辨率的分析。而在此之前几乎没有工作能够同时满足这些优点。

由于很多功耗模型是资深设计师们针对每一款CPU人工调试而成的，这将带来巨大的人力成本，同时随着CPU设计日趋复杂，想要人工设计准确的功耗模型变得越来越困难。

即使功耗模型在过去二十年里已经被反复充分研究，但准确、快速、低开销、自动化的功耗分析方法也一直没有被实现。

杜克大学的团队与TAMU和工业界的Arm公司合作，共同提出了一套完整的解决方案。

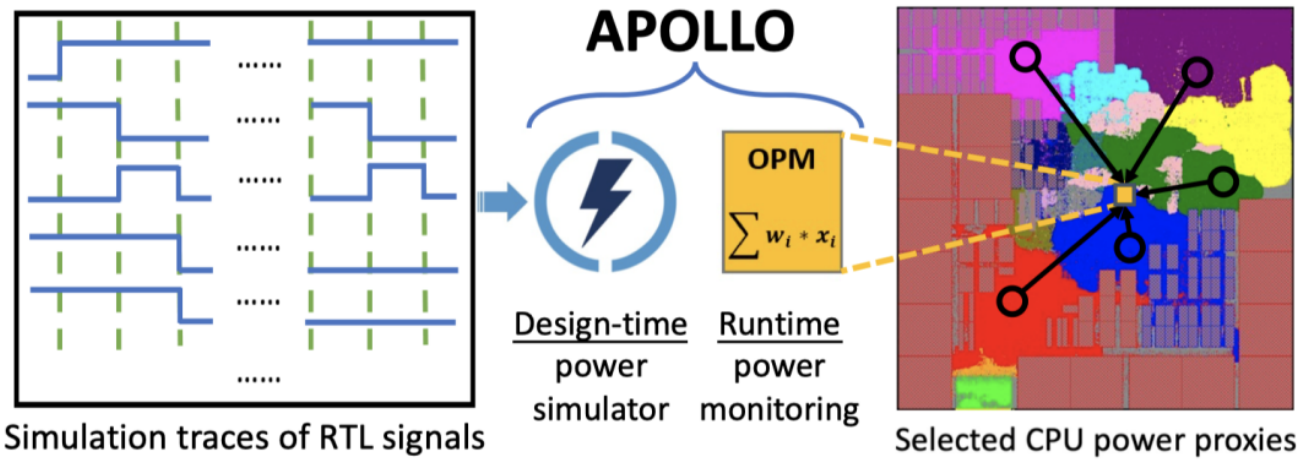
团队的研究论文《可用于大规模商业化处理器的全自动化功耗模拟架构》（APOLLO: An Automated Power Modeling Framework for Runtime Power Introspection in High-Volume Commercial Microprocessors）获得了2021年MICRO唯一的最佳论文奖。


APOLLO: An Automated Power Modeling Framework for Runtime Power Introspection in High-Volume Commercial Microprocessors

Authors: Zhiyao Xie, Xiaoqing Xu, Matt Walker, Joshua Knebel, Kumaraguru Palaniswamy, Nicolas Hebert, Jiang Hu, Huanrui Yang, Yiran Chen, Shidhartha Das Authors Info & Claims

https://dl.acm.org/doi/pdf/10.1145/3466752.3480064

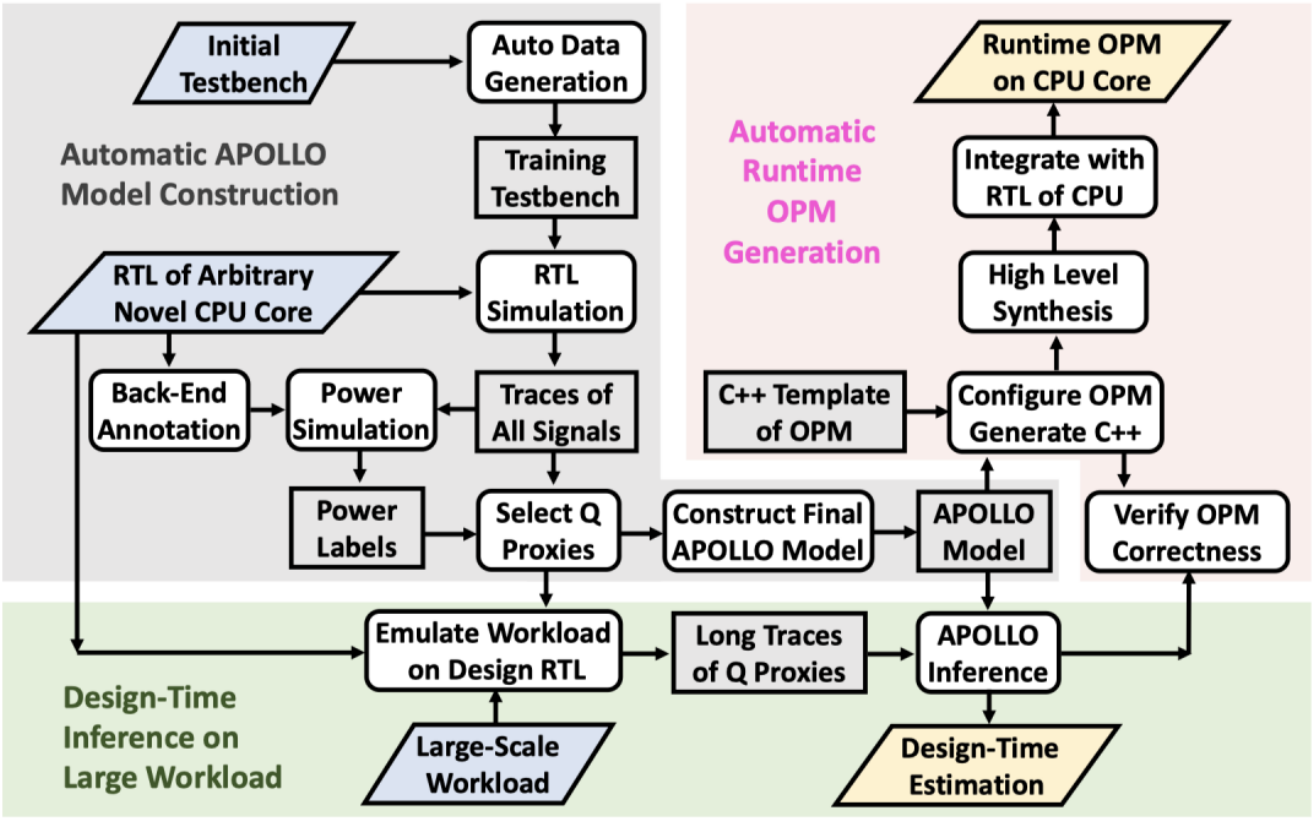
APOLLO使用一套统一的机器学习模型同时对设计和运行阶段的CPU功耗进行极低开销的快速实时计算。这个方法在商业化的Arm CPU设计Neoverse N1和Cortex-A77上得到了充分验证。这种前所未有的功耗计算能力可能会极大改变CPU的设计和使用方式，同时开启新的应用领域。



以Neoverse N1为例，APOLLO提供了一个设计时功率模拟器和一个基于一致模型的运行OPM

另外，APOLLO可能是第一个AI技术用来对芯片全lifetime做管理的应用，同时这个方法的整个流程是完全自动化的，不依赖任何工程师的经验。

理论上可以用于任何芯片设计，除了各种类型的CPU，还可用于GPU，NPU，和其他芯片，甚至是芯片的某一部分模块。团队认为这种方法以后可能会成为芯片设计的标准设计。

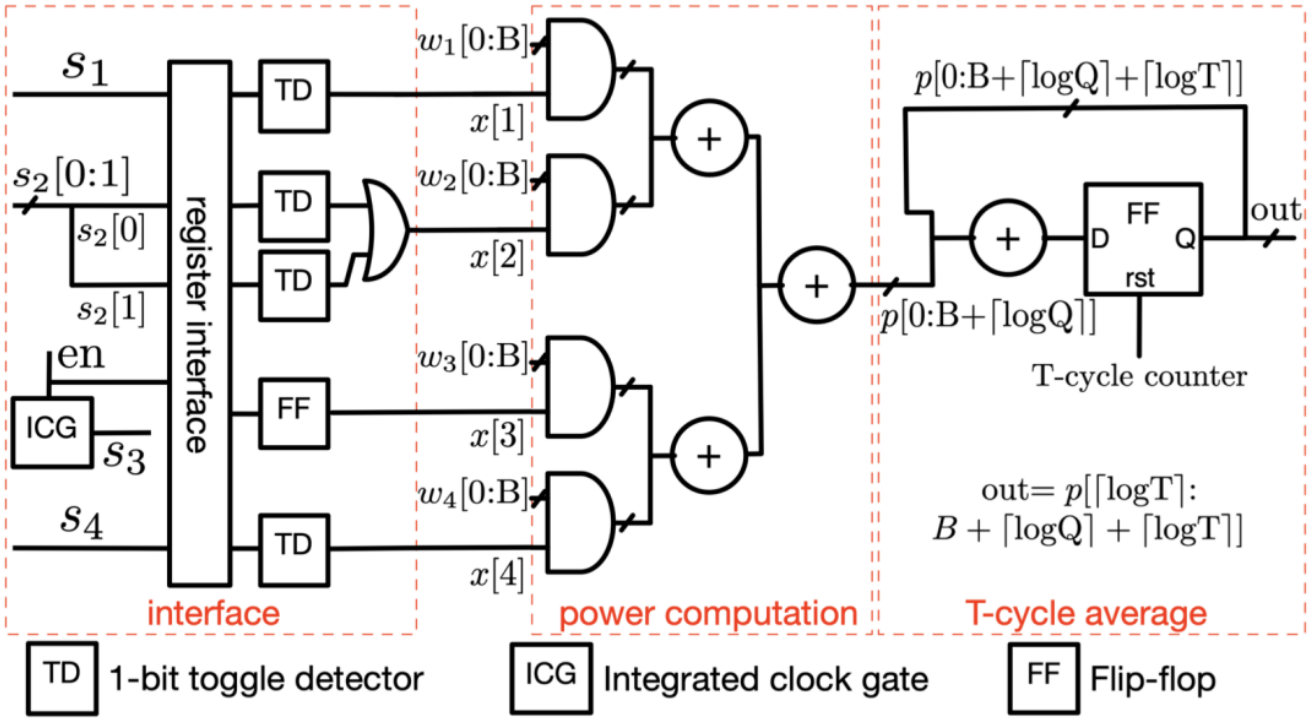


自动化APOLLO框架

在设计阶段， APOLLO可以在几分钟内获得几千万时钟周期（cycle）的功耗，而传统基于emulator的工业界方法需要多达两个星期。APOLLO的准确率非常之高，可以达到90%至

95%。

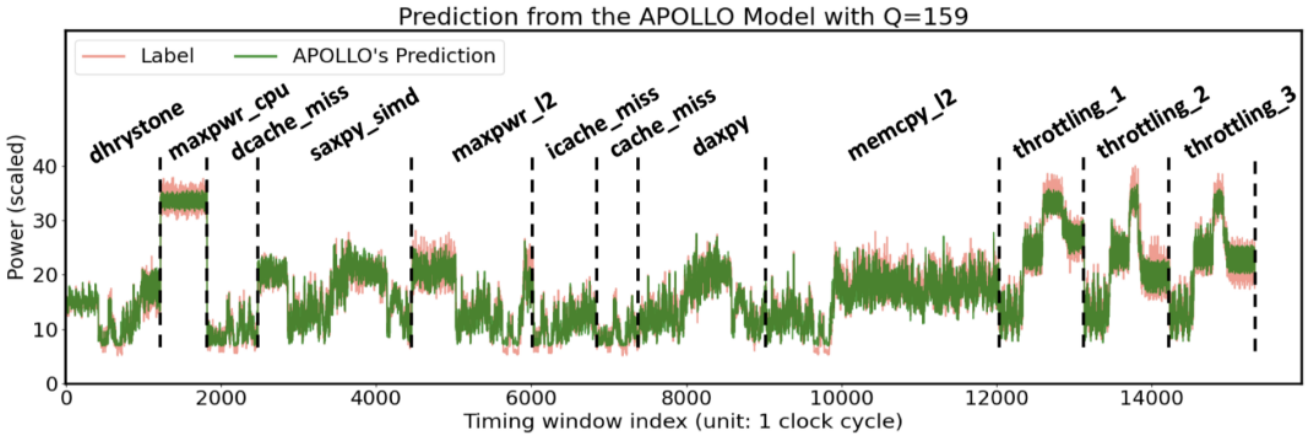
另外APOLLO的功耗分析可以精确到每个时钟周期（cycle），之前任何方法在这种速度下都无法获得这样的高分辨率 （temporal resolution） 的功耗分析。



OPM与CPU设计的整合

在CPU运行阶段，APOLLO可以对CPU功耗和温度进行管理。为了实现这一点，作者以极低的开销将APOLLO整合进CPU芯片核心内部，而这仅仅占用CPU芯片0.2%的面积，远低于之前类似的工作。

另外APOLLO可以在极短时间内对CPU内部的功耗变化进行反馈，这样就保证即使CPU内部由于复杂的交互出现了电压的快速下降也不怕。这种快速反馈也是传统的功耗管理模块所无法做到的。

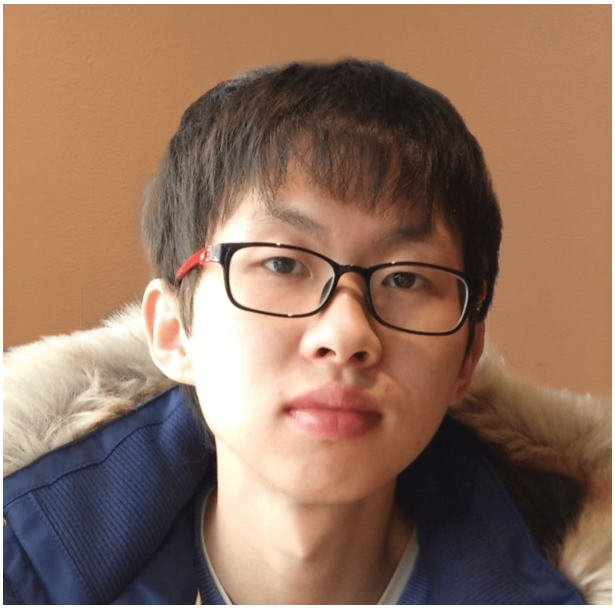


对Q=159的APOLLO模型的评估（Neoverse N1）

APOLLO的核心是通过剪枝(pruning)算法自动选取极少量与功耗最相关的CPU信号作为输入，然后建立一个快速的线性模型，用于对每个周期的功耗进行预测或监测。

除了信号选取和模型训练是自动完成之外，训练数据也是通过遗传算法自动生成的，另外APOLLO模型的硬件设计，也可以直接通过设置预先写好的C++ template并且进行高层次综合来简单实现。

作者简介



谢知遥，2017年在香港城市大学获得电子与通信工程学士学位。本科毕业后进入陈怡然和李海教授的实验室，成为杜克大学计算机工程专业博士生。

博士期间他曾在多家半导体公司实习，包括Cadence，Synopsys，NVIDIA，Arm。研究方向包括机器学习与芯片设计自动化，尤其是智能化的芯片设计方法。

陈怡然，杜克大学电子与计算机工程系教授、计算进化智能实验室联合主任。

目前，陈教授的研究聚焦新型存储系统、机器学习与神经形态计算、以及移动计算等方向的研究。陈怡然教授发表过一本专著及超过三百篇学术论文，获得过93项美国专利，并出任过多本IEEE和ACM期刊编委以及超过40个国际会议的组织与技术委员会主席和委员。

陈怡然教授曾经获得6次国际会议最佳论文以及12次最佳论文提名。他曾荣获美国国家自然科学基金委教授早期职业发展奖（NSF CAREER）和ACM电子自动化协会新教师奖。

今年1月，陈教授因为在非易失性存储技术领域做出杰出贡献，当选ACM Fellow。

今年4月29号，陈教授的项目「Privacy-preserving representation learning on graphs — a mutual information perspective」（互信息视角的图上保护隐私的表征学习）进入亚马逊研究奖（Amazon Research Awards, ARA）获奖名单。

当然，强将手下无弱兵，陈教授组里的最佳论文可不止这一篇。

2020年，陈怡然组就曾以一篇「TIPRDC: Task-Independent Privacy-Respecting Data Crowdsourcing Framework for Deep Learning with Anonymized Intermediate

Representations」（基于匿名中间表示的任务无关隐私的数据众包框架）荣获KDD 2020最佳学生论文奖。

文章的一作Ang Li是杜克大学电子和计算机工程系的一名在读博士，北京大学硕士毕业。

再次恭喜谢知遥和陈怡然教授！

参考资料：

https://www.microarch.org/micro54/

https://dl.acm.org/doi/pdf/10.1145/3466752.3480064

https://weibo.com/2199733231/KDOBo8PGz