更低开销CPU功耗分析，杜克陈怡然学生提出自动功耗建模，摘得微架构顶会MICRO最佳论文

机器之心 2021-10-24 00:41

机器之心报道

编辑：杜伟：陈萍

> 第 54 届 IEEE/ACM 微体系结构国际研讨会公布最佳论文奖，来自杜克大学、陈怡然教授团队的谢知遥摘得该奖项。

第 54 届 IEEE/ACM 微体系结构国际研讨会（IEEE/ACM International Symposium on Microarchitecture, MICRO-54）已于近日公布奖项，来自杜克大学的谢知遥荣获最佳论文奖，获奖论文题目为《 APOLLO: An Automated Power Modeling Framework for Runtime Power Introspection in High-Volume Commercial Microprocessors 》，他是该论文一作。

论文地址：：https://dl.acm.org/doi/pdf/10.1145/3466752.3480064

值得一提的是，谢知遥的导师是陈怡然教授。陈怡然教授是杜克大学电子与计算机工程系教授，计算进化智能中心主任，专注于新型存储器及存储系统，机器学习与神经形态计算，以及移动计算系统等方面的研究。在得知学生获奖时，陈教授发微博表示祝贺：

MICRO 大会是由 IEEE（Institute of Electricaland Electronics Engineers，电气和电子工程师协会）和 ACM（Association for Computing Machinery，国际计算机学会）共同举办的专业领域会议。

从 1968 年开始，MICRO 大会每年举行一次，到目前为止已经是第 54 届。

MICRO 致力于展示、讨论和辩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。本次研讨会汇集了微体系结构、编译器、芯片和系统相关领域的研究人员，就传统的微体系结构主题和新兴研究领域进行技术交流。

MICRO 社区一直保有学术研究人员和工业设计师之间的密切互动，其目标是在 MICRO-54 延续这一传统。2021 年，MICRO 将作为全球在线活动举办，主办城市是希腊雅典。

论文解读

从嵌入式应用、移动计算到数据中心，实现严格的能效要求驱动整个计算领域的设计决策。因此，无论是在 CPU 微体系架构设计期间，还是对于运行时功耗管理来说，准确的功耗估计对实施谨慎的工

程权衡至关重要。并且，对功耗估计的需求又因目标应用的不同而异。

最近用于快速功耗管理的电压增高的技术需要细粒度的时间分辨率，比如文献 [32] 中的完整电压增高操作发生在几十纳秒内。类似地，在现代高性能 CPU 中，电压噪声效应（如 Lid/dt）发生在 <10 个周期内。因此，量化快速电压噪声的影响以及自适应时钟等缓解功能的功效需要功耗追踪中出现细粒度时间分辨率，其中每个 CPU 周期都有一个 sample（即每周期时间分辨率）。

设计时功耗建模挑战。对于细粒度的功耗追踪，CPU 设计团队往往依赖行业标准的功耗分析工具（如文献[8]），以在 RTL 或者具有反向注释寄生程序的门级中重放模拟向量。功耗是根据单个信号网络的切换统计数据以及这些网络驱动的电容性负载所计算的。这种方法非常准确并可以作为验收标准，但遗憾的是计算成本也很高。

另一种替代方法依赖基于 FPGA 的网表模拟来解决功耗估计的速度影响。在这种方法中，模拟轨迹基于 FPGA 生成，然后使用功耗分析 EDA 软件处理提取的切换统计数据，以获得功耗轨迹。但是，由于现代计算机服务器存在严重的存储限制，因此使用这种方法来实施每周期功耗追踪依然任务繁重。

运行时功耗估计挑战。以往的工作已经展示了使用硬件监控事件计数器来指导 OS 编排的 DVFS 的运行时回归模型。这些模型在数千或数百万个 CPU 周期内对累积特定微架构事件（如 L2 缓存中丢失事件和退役指令）求取平均值。但是，这些事件与每周期微架构活动的相关性较差。并且，当需要细粒度功耗追踪时，对长 CPU 周期的求平均值过程会导致这种方法的准确性大打折扣。

最近已经有人提出基于片上功耗计（on-chip power meter, OPM）的 RTL 运行时功耗监控方法，以牺牲专用硬件电路为代价来提升时间分辨率。然而，现有相关方法无法同时实现高分辨率和低硬件面积开销。

Methods (Hardware Overhead in Area %)	Demonstrated Application	Model Type	Temporal Resolution	PC / Proxy Selection	Cost or Overhead
[20, 35, 43, 48, 61]	Design-time software model	Analytical	>1K cycles	N/A	Low
[78]		Proxies	>1K cycles	Automatic or no selection	High
[17, 64]					Medium
[79]			Per-cycle		High
[19, 42, 44, 72, 76]					Medium
[22] **(300% overhead)**	Design-time FPGA emulation	Proxies	Per-cycle	Automatic	High
[75] **(16% overhead)**					Medium
[40]			~100s cycles		
[66]			Per-cycle	Hybrid manual/auto	
[10, 11, 16, 24, 26, 33, 34, 36, 52, 58, 62, 63, 65, 68]	Runtime monitor	Event Counters	>1K cycles	Manual	Low
[38]			~100s cycles		
[23] **(2-20%)**, [51] **(1.5-4%)**, [53] **(7%)**		Proxies	>1K cycles	Automatic	Medium
[80] **(4-10%)**, [81] **(7%)**			~100s cycles		
APOLLO **(0.2% overhead)**	Design-time model Runtime monitor	Proxies	Per-cycle	Automatic	Low

Table 1: Comparison among various power modeling approaches. The percentage numbers are area overheads.

不同功耗建模方法的比较。

在本文中，研究者提出一个统一的 RTL 级功耗建模框架 APOLLO，它在一致性的模型结构中同时解决了设计时和运行时的挑战，如下图 1 所示。APOLLO 的核心是一个基于最小最大凸惩罚（MCP）回归的全新功耗代理选择方法，并且可以为数百万个 CPU 周期内执行的基准实施功耗追踪。对于运行时监控，APOLLP 以 0.2% 的面积开销实现了每周期准确的功耗估计。

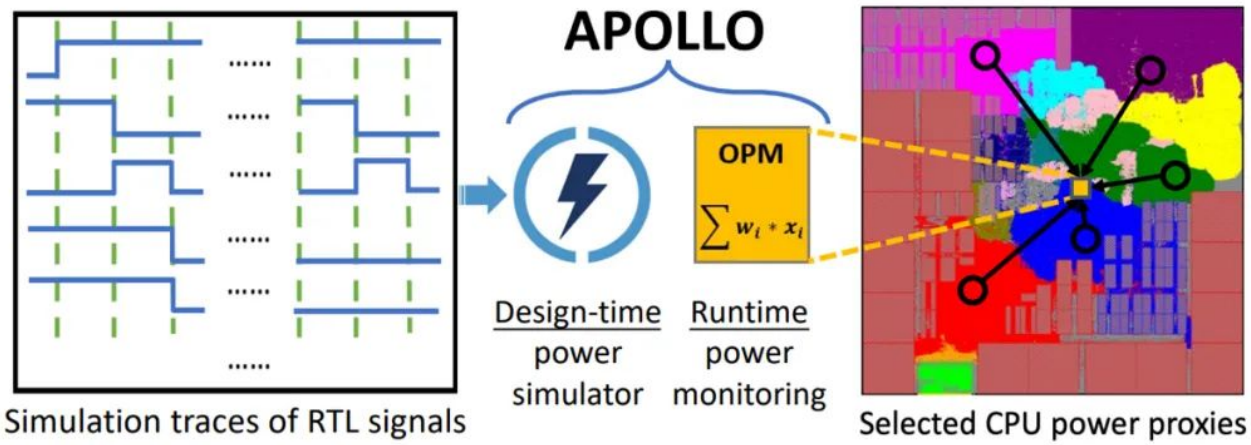


Figure 1: APOLLO provides a design-time power simulator and a runtime on-chip power meter (OPM) based on a consistent model, as an example, for Neoverse™ N1.

APOLLO 是第一个实现周期准确度和 1% 以下面积开销的功耗监控方法。此外，APOLLP 的代理选择过程是全自动化的，因而可以扩展到新设计。下图 2 为自动化的 APOLLO 框架示意图：

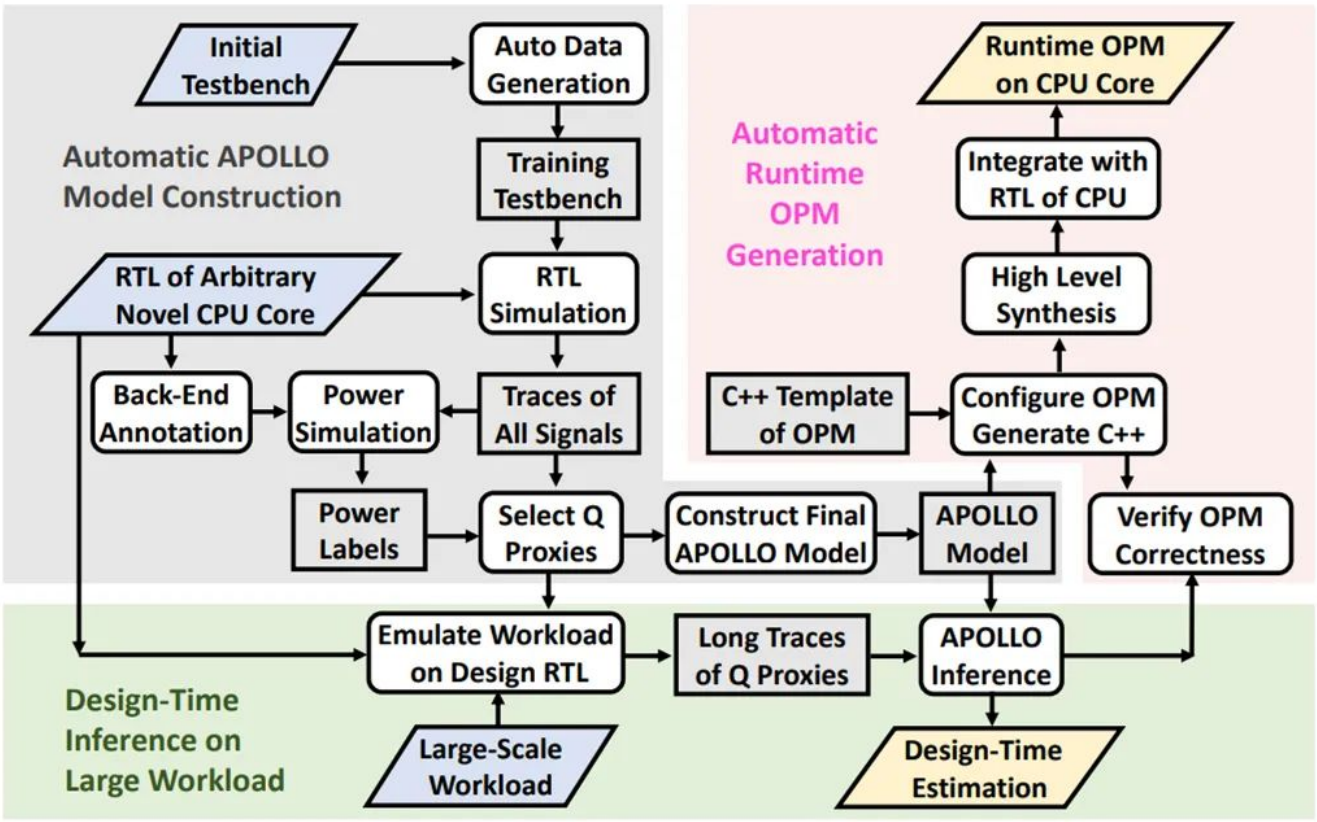


Figure 2: The automated APOLLO framework.

与最近的机器学习方法 PRIMAL 相比，APOLLO 达到相似的准确度，但速度快了数个量级。APOLLO 在准确度和计算速度方面还超越了另一 SOTA 方法 Simmani。不仅如此，与最近的 OPM 方法相比，APOLLO 实现了细粒度的时间分辨率和更低的硬件开销。下图为 APOLLO 与其他基线方法的一些比较结果：

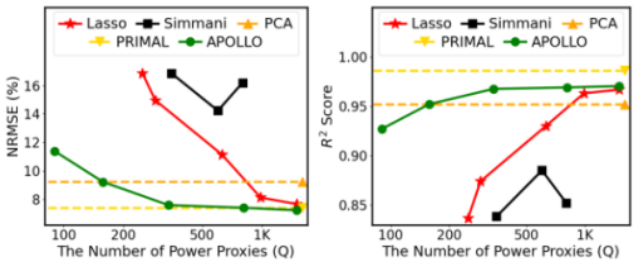


Figure 10: Per-cycle power accuracy vs. number of proxies for per-cycle power prediction (Neoverse N1).

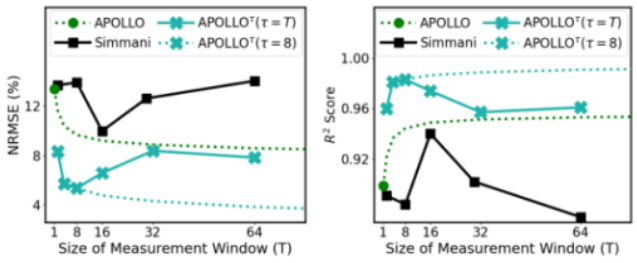
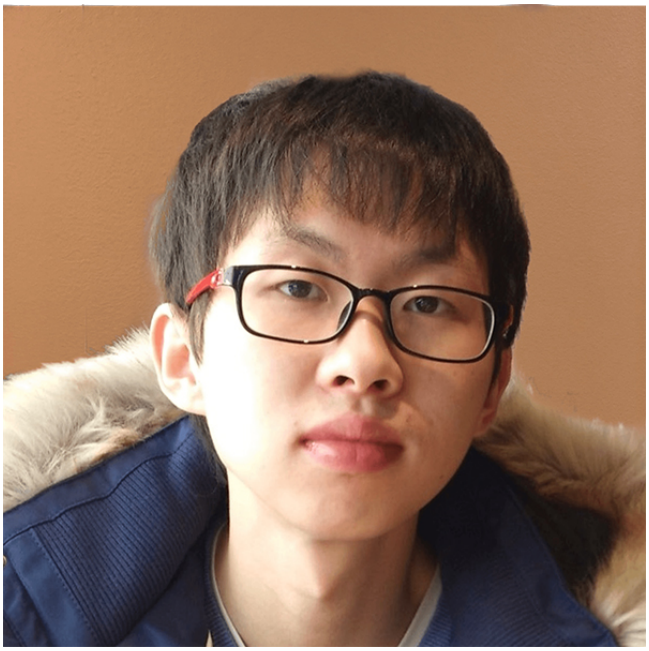


Figure 11: T-cycle accuracy vs. window size (T) for multi-cycle prediction (Neoverse N1). — $Q = 200$ for Simmani, $Q = 70$ for APOLLO methods.

一作简介



个人主页：https://www.linkedin.com/in/zhiyaoxie/

获奖论文一作是谢知遥，于 2017 年获得香港城市大学电子与通信工程学士学位，本科毕业后加入陈怡然和李海教授的实验室团队，成为杜克大学计算机工程专业博士生。

谢知遥主要研究领域包括机器学习、EDA、深度学习、VLSI 设计等。他曾在多家知名公司实习，包括 Arm、英伟达、Cadence、Synopsys。

参考链接：

https://m.weibo.cn/status/4694893102105307?

sourceType=weixin&from=10BA195010&wm=9006_2001&featurecode=newtitle

2021 NeurIPS MeetUp China

受疫情影响，NeurIPS 2021依然选择了线上的形式举办。虽然这可以为大家节省一笔注册、机票、住宿开支，但不能线下参与这场一年一度的学术会议、与学术大咖近距离交流讨论还是有些遗憾。

今年，我们将在NeurIPS官方支持下，再次于 12 月份在北京举办线下NeurIPS MeetUp China，促进国内人工智能学术交流。

2021 NeurIPS MeetUp China将设置 Keynote、圆桌论坛、论文分享和 Poster 等环节，邀请顶级专家、论文作者与现场参会观众共同交流。

欢迎 AI 社区从业者们积极报名参与，同时我们也欢迎 NeurIPS 2021 论文作者们作为嘉宾参与论文分享与 Poster 展示。感兴趣的小伙伴点击**「阅读原文」**即可报名。

Read more

People who liked this content also liked

节能1000倍！仿人脑神经芯片跑AI模型竟然这么省电

新智元

让机器学习“如何学习”！从零开始读懂MAML！

AI蜗牛车

注意力机制YYDS，AI编辑人脸终于告别P一处而毁全图

量子位