邹苏南

zousunan@pku.edu.cn suenar.github.io sunanzou +86-18511791898

个人简介

邹苏南是北京大学计算机学院计算机系统结构专业的四年级直博生，导师为罗国杰长聘副教授，隶属于高能效计算与应用中心（CECA）的智能异构系统（DASYS）实验室。他于 2022 年获得北京大学信息科学与技术学院的计算机科学与技术专业学士学位。当前邹苏南的研究兴趣主要包括模型压缩与适配算子/架构、领域定制加速器编译框架、AI 编译器。他在电子设计自动化尤其是逻辑综合优化领域也有丰富的经验。其博士论文将围绕在 Sketch 算法的系统级应用、硬件优化与自动生成展开讨论。

教育背景

计算机系统结构专业 博士 **2022 年 9 月 – 2027 年 6 月（预计）**
北京大学计算机学院 中国，北京
导师：罗国杰教授

计算机科学与技术专业 学士 **2018 年 9 月 – 2022 年 7 月**
北京大学信息科学与技术学院 中国，北京

研究经历

日本国立信息学研究所 **2025 年 10 月 - 2026 年 4 月（预计）**
NII 交换实习生，指导老师：Junichi Yamagishi 教授 日本，东京
研究主题：考虑演奏法的小提琴生成模型

卡耐基梅隆大学电子与计算机工程系 **2021 年 3 月 - 2021 年 9 月**
研究实习生，指导老师：Lawrence Pileggi 教授 远程
研究主题：图神经网络指导 FPGA 逻辑锁隐蔽设计。

北京大学信息科学技术学院 **2019 年 7 月 - 2022 年 1 月**
研究实习生，指导老师：罗国杰教授 中国，北京
研究主题：强化学习辅助的逻辑优化算法。

发表论文

[C8] Xiao Tan, Chenyue Li, **Sunan Zou**, and Guojie Luo. "RTLPilot: Skill-Driven Multi-Agent System for Repository-Level RTL Code Migration". *2026 International Symposium of EDA (**ISEDA**), 2026. (To Appear)*

[C7] **Sunan Zou**, Xueting Sun, Ziyun Zhang and Guojie Luo. "UltraSketchLLM: Sub-1-Bit LLM Compression via Sketch and Hardware-Friendly Operators". *63st Design Automation Conference (**DAC**), 2026. (To Appear)*

[C6] **Sunan Zou**, Bizhao Shi, Ziyun Zhang and Guojie Luo. "MuSA: Multi-Sketch Accelerator with Hybrid Parallelism and Coalesced Memory Organization". *IEEE 42nd International Conference on Computer Design (**ICCD**), 2024.*

[C5] Xinming Wei, Ziyun Zhang, **Sunan Zou**, Kaiwen Sun, Jiahao Zhang, Jiaxi Zhang, Ping Fan, and Guojie Luo. "AceRoute: Adaptive Compute-Efficient FPGA Routing with Pluggable Intra-Connection Bidirectional Exploration". *43th IEEE/ACM International Conference on Computer-Aided Design (**ICCAD**), 2024.*

[C4] Bizhao Shi, Tuo Dai, **Sunan Zou**, Xinming Wei, and Guojie Luo. "ImageMap: Enabling Efficient Mapping from Image Processing DSL to CGRA". *30th International European Conference on Parallel and Distributed Computing (**Euro-Par**), 2024.*

[C3] **Sunan Zou**, and Guojie Luo. "PONO: Power Optimization with Near Optimal SMT-based Sub-circuit Generation". *61st Design Automation Conference (**DAC**), 2024.*

[C2] **Sunan Zou**, Jiaxi Zhang, and Guojie Luo. "Incremental SAT-based Exact Synthesis", *34th Great Lakes Symposium on VLSI (**GLSVLSI**) 2024*. [**Best Paper Reward**]

[C1] **Sunan Zou**, Jiaxi Zhang, Bizhao Shi, and Guojie Luo. "BESWAC: Boosting Exact Synthesis via Wiser SAT Solver Call", *27th Design Automation and Test in Europe (**DATE**), 2024*.

[J2] Lei Chen, Yiqi Chen, Zhufei Chu, Wenji Fang, Tsung-Yi Ho, et al, and **Sunan Zou (Alphabetical Order)**. "Large circuit models: opportunities and challenges". *Science China Information Sciences Volume 67, 2024*.

[J1] **Sunan Zou**, Jiaxi Zhang, Bizhao Shi, and Guojie Luo. "PowerSyn: A logic synthesis framework with early power optimization", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems Volume: 43 (**TCAD**), 2023*.

代表性项目

[编译器] PTO 驱动的 Unsloth 高性能 NPU 微调框架 **2026 年 1 月 – 至今**

内容： 本项目将聚焦于 Unsloth 兼容 NPU 后端的优化实现，打造 PTO 与 PyPTO 接入主流生态的实例。首先，本项目将基于 Unsloth 后端抽象嵌入 NPU Runtime API；结合训推 Infra 打通流程。之后，本项目将基于 PyPTO/PTO-ISA，进行 Unsloth 中关键算子开发和调优，识别训推框架中的关键算子并优化，延展 PTO 的作用域。项目还将深度结合微调场景和 NPU 架构，进行 PyPTO Runtime 的调优设计。

角色： 主要贡献者 & 代码维护者。

[大语言模型算法] 少样本乐器音频生成语言模型构建 **2025 年 10 月 – 至今**

内容： 本项目聚焦于高质量少样本乐器音频生成模型的两阶段迁移训练框架和数据生成范式设计。首先，本项目设计了考虑演奏法的乐谱词元化方法，并依此构建乐谱合法化-力度拟合采样-演奏法自动标记的数据生成范式，解决了基于语言模型的少样本乐器生成数据短缺问题。在使用生成数据对语言模型进行全量迁移训练的基础上，使用少量真实数据对语言模型和编码器模型进行联合局部微调，以提升音频生成保真度。该项目为日本国立信息学研究所交换期间的研究课题，计划投稿至 ACM Multimedia 2026。

角色： 独立项目。

[软硬件协同设计] 基于 Sketch 的模型参数压缩与适配算子融合 **2025 年 5 月 – 至今**

内容： 本项目使用专用 Sketch 算法压缩大语言模型权重，并开发解压缩计算融合算子以降低额外开销。我们首先使用针对大模型权重分布特征的 Sketch 算法对模型权重进行压缩，在保证性能的同时突破了 1 比特压缩率限制。其次，使用重要性度量分配 Sketch 压缩空间，进一步减少压缩带来的性能损失。为减少解压缩带来的推理开销和训练性能损失，我们将 Sketch 中的顺序哈希操作映射为等价矩阵乘操作，开发压缩解压缩操作和后续矩阵乘操作融合算子，有效降低额外开销。

角色： 独立项目。 **发表：** [C7]

[编译器] Sketch 加速器生成与优化框架 **2024 年 2 月 – 至今**

内容： 本项目聚焦于适配多种 Sketch 算法与不同后端硬件的加速器生成优化框架设计。首先，我们定义了可扩展的领域定制语言，支持表示层次化、多 Sketch 等应用场景和算法设计。其次，提出了高效分析性模型预估所需的硬件资源、运行性能、Sketch 预估准确率等关键指标，为后续架构探索提供指引。项目涵盖 FPGA、GPU、CPU、等多种后端，结合实际编译结果与分析性模型预估指标进行架构探索，以优化设计关注的吞吐量或时延指标。项目将与不同 Sketch 算法系统级应用如网络监测、冷热页检测、工作集预估等应用结合，优化整体运行效率。项目后续工作预计投稿至 ASPLOS 2027。

角色： 独立项目。 **发表：** [C6]

[设计自动化] FPGA 布线并行加速 **2023 年 10 月 – 2024 年 3 月**

内容： 本项目聚焦于使用粗细粒度结合的并行方式大幅提升 FPGA 布线搜索速度。一方面，本项目使用信号间并行，递归划定并行区域，利用有向无环图解决了信号间潜在冲突问题。另一方面，我们使用源节点到终节点搜索过程的双向并行，结合信号间并行进一步提升了搜索效率。

角色： 主要贡献者 & 代码维护者。 **发表：** [C5]

[设计自动化] 逻辑精确综合流程加速 **2023 年 1 月 – 2025 年 1 月**

内容： 本项目聚焦于精确综合的运行时间的减少和可解问题规模的增加。精确综合依赖 SAT 后端生成最优电路，导致运行时间长且不可预估，且存在可解问题规模小造成后续工艺优化空间有限的问题。观察到不同布尔电路适合不同编码方式之后，本项目利用机器学习方法布尔函数特征预测适合编码以提升 SAT 求解效率。本项目还利用部分最优电路具有最优子结构的特点，设计经验性辅助函数抽取待解问题

子问题，并利用子问题设计基于拓扑的和基于子电路的两种增量式 SAT 求解方式。后续将工作开发问题特定 SAT 求解器，预计投稿至 TCAD。
角色： 独立项目。　　**发表：** [C1, C2]

[设计自动化] 强化学习辅助的逻辑综合功耗优化 **2021 年 7 月 – 2023 年 7 月**
内容： 本项目使用强化学习建模逻辑综合优化命令搜索问题，通过引入高效功耗建模作为强化学习奖励函数，自动优化设计的功耗-面积-性能帕累托最优前沿。本项目首先通过考虑了毛刺和动态信号的功耗建模方法，利用 SMT 表示扩充了精确综合优化目标，生成最优功耗局部电路，并将其融入传统逻辑优化命令。其次，本项目手动设计电路表示特征作为强化学习环境，以功耗-面积-性能为优化目标训练命令决策模型，生成了特定电路的优化路径，在不影响面积和性能的情况下有效提升了逻辑综合的功耗表现。
角色： 独立项目。　　**发表：** [J1, C3]

荣誉奖项

- 北京大学三好学生 2024 年
- 北京大学比亚迪奖学金 2024 年
- International Symposium on FPGA 布线比赛第三名 2024 年
- 集成电路 EDA 设计精英挑战赛三等奖 2022 年
- 北京大学本科生科研训练校长基金（理科） 2021 年
- 北京大学新生奖学金 2018 年
- 北京大学信息科学技术学院新生院长奖学金 2018 年

服务

- **课程助教：** 并行与分布式计算导论，北京大学 2022 年，2023 年
- **官方审稿：** IEEE TCAD

技能

- **编程：** C/C++, Python(PyTorch), FPGA HLS, Verilog, CUDA
- **工具：** LATEX, Git, Docker/Apptainer
- **语言：** 中文、英语、日语
- 前北京大学学生交响乐团成员