Sunan Zou

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

Personal Profile

I am a fourth-year Ph.D Candidate in the School of Computer Science, Peking University, associated with the Center for Energy-Efficient Computing and Applications (CECA). I am a member of the PKU-DASYS Lab, advised by Professor Guojie Luo. I received my B.S. degree in Computer Science and Technology from the School of Electronics Engineering and Computer Science, Peking University in 2022. My current research interests include model compression and adaptive operators/architectures, domain-specific accelerator compilation frameworks, and AI compilers. He also has extensive experience in electronic design automation, particularly in logic synthesis optimization. His doctoral dissertation focuses on system-level applications, hardware optimization, and automatic generation of Sketch algorithms.

Education

Ph.D. in Computer Science **Sep 2022 – Jun 2027 (expected)**
School of Computer Science, Peking University *Beijing, China*
Advisor: Prof. Guojie Luo

B.S. in Computer Science and Technology **Sep 2018 – Jul 2022**
School of Electronics Engineering and Computer Science (EECS), Peking University *Beijing, China*

Research Experience

Yamagishi Lab, National Institute of Informatics (Japan) **Oct 2025 - Now**
Research Intern, advised by Prof. Junichi Yamagishi. *Tokyo, Japan*
Topics: Neural-Language Model for Violin Synthesis

Electrical and Computer Engineering, Carnegie Mellon University **Mar 2021 - Sep 2021**
Research Intern, advised by Prof. Larry Pileggi. *Remote*
Topics: Graph Neural Network-Based Logic Locking Detection

DASYS Lab, Peking University **Jul 2019 - Jan 2022**
Research Intern, advised by Prof. Guojie Luo. *Beijing, China*
Topics: Reinforcement Learning-Based Logic Optimization

Publications

[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.*

Representative Projects

[Compiler] PTO-Based Unsloth Framework for NPU **Jab 2026 - Now**

Contents: This project focuses on the optimized implementation of an Unsloth-compatible NPU backend, establishing a showcase for integrating PTO and PyPTO into mainstream ecosystems. First, the project embeds the NPU Runtime API within the Unsloth backend abstraction layer and integrates end-to-end training and inference infrastructure pipelines. Subsequently, based on PyPTO/PTO-ISA, the project will develop and tune critical operators within Unsloth. The project further incorporates fine-tuning scenarios and NPU architecture characteristics to guide the tuning design of the PyPTO Runtime.
Role: Main contributor & Code maintainer.

[LLM Algorithm] Violin Audio Model **Oct 2025 - Now**

Contents: This project focuses on the design of a two-stage transfer training framework and data generation paradigm for musical instrument audio generation. First, the project proposes a score tokenization method that accounts for playing techniques, upon which a data generation paradigm is constructed comprising score legalization, dynamic fitting sampling, and automatic articulation annotation, addressing the data scarcity problem. Building on full-parameter transfer training of the language model using generated data, a small amount of real data is employed for joint partial fine-tuning of both the language model and the encoder model to improve audio generation fidelity. This project was conducted as a research topic during an exchange visit at the National Institute of Informatics (NII), Japan, with a planned submission to ACM Multimedia 2026.

Role: Independent project.

[SW/HW Co-design] Sketch-Based LLM Compression & Hardware-Friendly Kernel **May 2025 - Now**

Contents: This project applies Sketch algorithms to compress LLM weights and develops fused compression and decompression kernels to reduce associated overhead. We first compress model weights using a Sketch algorithm tailored for LLMs, achieving sub-1-bit compression rates while maintaining model performance. To reduce the inference overhead and training performance loss introduced by decompression, we reformulate the sequential hashing operations in Sketch as equivalent matrix multiplication operations, and develop fused kernels that integrate compression/decompression with computation, effectively reducing additional overhead.
Role: Independent project. **Publication:** [C7]

[Compiler] Sketch Accelerator Design & Auto-Generation **Feb 2024 - Now**

Contents: This project focuses on the design of an accelerator generation and optimization framework adaptable to multiple Sketch algorithms and diverse backend hardware targets. First, we define an extensible domain-specific language (DSL) capable of expressing hierarchical and multi-Sketch application scenarios and algorithm designs. Second, we propose efficient analytical models to estimate key metrics, providing guidance for subsequent architecture DSE. The project spans multiple backend targets including FPGAs, GPUs, and CPUs. The framework will be integrated with system-level applications of various Sketch algorithms, including network monitoring, hot-cold page detection, and working set estimation, to optimize overall runtime efficiency. Follow-up work from this project is planned for submission to ASPLOS 2027.
Role: Independent project. **Publication:** [C6]

[Design Automation] Multi-Level Parallel FPGA Routing **Oct 2023 - Mar 2024**
Contents: This project focuses on significantly accelerating FPGA routing search through a combined coarse- and fine-grained parallelization approach. On one hand, the project employs inter-signal parallelism by recursively partitioning parallel regions and leveraging directed acyclic graphs (DAGs) to resolve potential inter-signal conflicts. On the other hand, we introduce bidirectional parallelism in the search process from source nodes to sink nodes, which, combined with inter-signal parallelism, further improves overall search efficiency.
Role: Main contributor & Code maintainer. **Publication:** [C5]

[Design Automation] SAT-Baed Incremental Synthesis Acceleration **Jan 2023 - Jan 2025**
Contents: This project focuses on reducing the runtime and increasing the solvable problem scale of exact synthesis. Upon observing that different Boolean circuits are better suited to different encoding schemes, this project employs machine learning methods to predict appropriate encodings from Boolean function features, thereby improving SAT solving efficiency. The project further exploits the optimal substructure property present in certain optimal circuits, designing heuristic auxiliary functions to extract subproblems from the target problem, and developing two incremental SAT solving strategies based on topology and a sub-circuit. Follow-up work will develop a problem-specific SAT solver, with a planned submission to TCAD.
Role: Independent project. **Publication:** [C1, C2]

[Design Automation] Reinforcement Learning for Logic Optimization **Jul 2021 - Jul 2023**
Contents: This project formulates the logic synthesis optimization command search problem as a reinforcement learning task, and optimizes circuit designs by incorporating an efficient power modeling method as the reinforcement learning reward function. First, the project extends the exact synthesis optimization objective using SMT formulations with a power modeling approach that accounts for glitches and dynamic signals, generating locally optimal low-power circuits and integrating them into conventional logic optimization commands. Second, the project manually engineers circuit representation features as the reinforcement learning environment, trains a command decision model with power, area, and performance as joint optimization objectives, and produces circuit-specific optimization sequences that effectively improve the power characteristics of logic synthesis without compromising area or performance.
Role: Independent project. **Publication:** [J1, C3]

Honors and Awards

- Honors for Merit Student (三好学生), Peking University 2024
- BYD Scholarship (北京大学比亚迪奖学金), Peking University 2024
- ISFPGA FPGA Routing Contest Third Place, ISFPGA Committee 2024
- Principal's Fund for Undergraduate Research Training (Natural Science), Peking University 2021
- Freshman Scholarship (新生奖学金), Peking University 2018
- Freshman Scholarship (院长新生奖学金), School of EECS, Peking University 2018

Services

- **Official Reviewer**: TCAD
- **Teaching Assistant**: Introduction to Parallel and Distributed Computing, Peking University 2022, 2023

Skills

- **Programming Languages**: C/C++, Python(Pytorch), FPGA HLS, Verilog, CUDA
- **Tools**: LATEX, Git, Docker/Apptainer
- **Languages**: Chinese, English, and Japanese.
- Former Member of Peking University Student Orchestra

[Last updated on March 27, 2026]