

Guanghui Song

Lushan Road (S), Yuelu District, Changsha, 410082, China

✉ sheensong@hnu.edu.cn | 🏠 <https://sheensong.top/academic>

RESEARCH INTERESTS

Guanghui Song is a PhD student at [Hunan University](#), under the supervision of [Prof. Jie Zhao](#). His research sits at the intersection of compilation and high-performance computing, emphasizing polyhedral methods, dataflow optimization, high-level synthesis, mixed-precision and hardware-aware acceleration. The overarching goal is to make diverse accelerators easier to program while improving end-to-end performance and resource utilization.

EDUCATION

- **Hunan University** September 2024 - July 2028
PhD in Computer Science Changsha, China
 - CYCLE Lab at College of Computer Science and Electronic Engineering
 - Supervised by [Prof. Jie Zhao](#)
- **Information Engineering University** September 2020 - July 2023
Master in Computer Science Zhengzhou, China
 - Supervised by Prof. Shaozhong Guo and Assoc. Prof. Jinchun Xu
- **Henan University of Technology** September 2016 - July 2020
Bachelor in Computer Science Zhengzhou, China
 - Department of Computer Science and Technology
 - GPA: 3.51

RESEARCH & INDUSTRY EXPERIENCE

- **National University of Singapore** September 2025 - April 2026
Research Intern Singapore
 - Contributed to a CGRA mapping approach that reformulates compilation as coarse-grained spatio-temporal search guided by polyhedral permutable bands instead of fine-grained DFG placement and routing, mitigating combinatorial explosion and utilization loss as PE meshes scale.
 - Implemented the execution-atom abstraction (innermost permutable loop body) with dependence-preserving loop skewing over $O(d^3)$ candidates (d : loop dimensions in that body), and an analytical polyhedral latency model over iteration spaces to score mappings without cycle-accurate simulation, enabling second-scale end-to-end search.
 - Integrated mapping flow into a CGRA toolchain and evaluated on Polybench and DL inference vs. Morpher, RfC-graTrans, and ML-CGRA: over 60% $4 \times 4/8 \times 8$ PE utilization with less scale-induced loss than DFG-centric mappers; up to $18.92 \times / 9.35 \times$ Polybench speedups and $1.31 \times / 2.21 \times$ geometric-mean DL speedups vs. ML-CGRA.
- **EVAS Intelligence** February 2025 - October 2025
AI Compiler R&D Intern Hangzhou, China
 - Co-designed TISA, a tile-level ISA that encodes operator semantics, typed dependencies and resource intents, enabling the runtime to reorder tiles across tensor/vector/DMA units without losing correctness.
 - Contributed to the semantics-preserving compiler pass that carries operator boundaries and dependency metadata from high-level graphs down to TISA binaries, eliminating most manual synchronization barriers.
 - Co-authored the ISCA 2026 paper on TISA that quantifies the framework, showing 1.52–1.92× speed-ups on ResNet, BERT, GPT-J, and LLaMA2 and 26% higher utilization on FlashAttention-3 versus the state-of-the-art H100 kernel.
- **Li Auto Inc.** July 2023 - August 2024
AI Compiler R&D Engineer Shanghai, China
 - Collaborated with chip architects to assist in the design and optimization of chip architecture, contributing to the development of hardware-software co-design strategies for the targeted AI chips.
 - Contributed to the development and optimization of AI operator libraries, ensuring their compliance with both functional and performance requirements critical to algorithm execution on specialized hardware.
 - Engaged in the design and development of a cutting-edge AI compiler, facilitating the efficient compilation of algorithmic models and operator libraries into executable files compatible with AI chip architectures.
 - Provided ongoing maintenance for the AI compiler and operator libraries, addressing and resolving performance and functionality issues identified during Virtual Platform, RTL simulations, and FPGA verification.
- **Thewake Systems Co. Ltd** June 2022 - September 2022
Compiler Development Intern Beijing, China
 - Contributed to the porting, development, and testing of the self-developed Fiuggi Compiler Collection (FCC), with a focus on enhancing its automatic parallelization capabilities based on LLVM 13.
 - Conducted rigorous benchmarking of the FCC compiler, comparing its multi-threading performance against that of ICC, GCC, LLVM, and AOCC using the [Polybench benchmark suite](#).

CONFERENCE PUBLICATIONS

*: EQUAL CONTRIBUTION, †: CORRESPONDING AUTHOR

- ISCA 2026** **Guanghui Song***, Xiaoqiang Dan*, Chengke Wang, Fei Liu, Wenyuan Lv, Zhongzhou Jiang, Jianjian Guan, Teng Lu, Lin Tao, Cheng Li, Weixing Pan, Wei Huang, Zirong Shen, Yi Yang, Hui Liu, and Jie Zhao†, **Dynamic Scheduling for AI Accelerators via TISA**. In *Proceedings of the 53rd Annual International Symposium on Computer Architecture (ISCA 2026)*, 27 June–01 July, 2026, Raleigh, NC, USA, to appear. <https://orcid.org/0009-0001-8089-1151>
- PPoPP 2024** Jinchen Xu*, **Guanghui Song***, Bei Zhou, Fei Li, Jiangwei Hao, and Jie Zhao†, **A Holistic Approach to Automatic Mixed-Precision Code Generation and Tuning for Affine Programs**. In *Proceedings of 29th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming (PPoPP 2024)*, 02–06 March, 2024, Edinburgh, United Kingdom, pages 55-67. <https://doi.org/10.1145/3627535.3638484>.
- ASE 2023** Zuoyan Zhang*, Bei Zhou, Jiangwei Hao, Hongru Yang, Mengqi Cui, Yuchang Zhou, **Guanghui Song**, Fei Li, Jinchen Xu, and Jie Zhao†, **Eiffel: Inferring Input Ranges of Significant Floating-point Errors via Polynomial Extrapolation**. In *Proceedings of 38th IEEE/ACM International Conference on Automated Software Engineering (ASE 2023)*, 11-15 September, 2023, Kirchberg, Luxembourg, pages 1441-1453. <https://doi.org/10.1109/ASE56229.2023.00139>

JOURNAL PUBLICATIONS

*: EQUAL CONTRIBUTION, †: CORRESPONDING AUTHOR

- 2026** Shihan Yuan*, Zuoyan Zhang*, **Guanghui Song**, Junhui Peng, Feng Wang, Zhuo Tang, Kenli Li, and Jie Zhao†, **A Decoupled Analytical Model for Tile Size Selection in Affine Programs**, *ACM Transactions on Architecture and Code Optimization*, to appear (accepted March 2026). <https://doi.org/10.1145/3806056>
- 2024** Fei Li*, Shaozhong Guo, Jiangwei Hao, Ming Hou, **Guanghui Song**, Jinchen Xu†, **Basic Math Library Implementation for RISC-V**, *Acta Electronica Sinica*, 2024, 52(5): 1633-1647 (in Chinese). <https://doi.org/10.12263/DZXB.20220375>
- 2023** **Guanghui Song***, Shaozhong Guo, Jie Zhao, Xiaohan Tao, Fei Li, Jinchen Xu†, **Automatic Mixed Precision Optimization for Stencil Computation**, *The Journal of Software*, 2023, 34(12): 5704-5723 (in Chinese). <https://doi.org/10.13328/j.cnki.jos.006757>
- Fei Li*, Shaozhong Guo, Bei Zhou, **Guanghui Song**, Jiangwei Hao, Jinchen Xu†, **Performance optimization of RISC-V basic math library**, *Computer Engineering & Science*, 2023, 45(09): 1532-1543 (in Chinese). <https://doi.org/10.3969/j.issn.1007-130X.2023.09.002>

AWARDS AND SCHOLARSHIPS

- **PLMW@PLDI'25 Scholarship** May 2025
- **EuroLLVM Developer's Meeting Student Travel Grant** March 2025
- **Li Auto Inc. Proactive "Excellent Individual"** March 2024
- **Information Engineering University First Class Academic Scholarship** June 2023
- **Information Engineering University Second Class Academic Scholarship** June 2022
- **Information Engineering University Second Class Academic Scholarship** June 2021
- **Henan University of Technology "Excellent Recent Graduates"** June 2020
- **Henan University of Technology Innovation and Entrepreneurship Scholarships** October 2019
- **Henan University of Technology "Excellent Student Assistant"** December 2018
- **Ministry of Education of China National Inspiration Scholarships** December 2017

LANGUAGES

- **Mandarin** : Mothertongue (Daily use)
- **English** : Fluent (Con conversationally fluent)

ACADEMIC ACTIVITIES

- **Participant & Scholarship Recipient** June 2025
PLDI 2025 + PLMW 
 - Attended Programming Language Design and Implementation conference
 - Participated in mentoring workshops for programming languages research
- **Session Moderator** April 2025
EuroLLVM 2025 
 - Moderated Session 6 (Student Talks and Technical Talk) at European LLVM conference
 - Coordinated Q&A sessions and facilitated discussions between speakers and attendees

- **Presenter**
CCF Chip 2024
◦ Discussed hardware/software co-design with industry practitioners and researchers
◦ Received peer feedback for improving mixed-precision methodologies
- **Presenter**
PPOPP 2024
◦ Delivered live talk at ACM SIGPLAN Symposium in Edinburgh
◦ Engaged with international experts in parallel programming and compilers

July 2024



March 2024



SELECTED PROJECTS

- **AutoPoly** 2025 - now
An architecture-aware polyhedral scheduling optimization tool based on MLIR (ongoing; code not yet public)
◦ Delivered an MLIR-based "autopoly-scheduling" pass that lifts affine kernels into polyhedral SCoPs, builds call-tree-aware dependence graphs, and maps user-selected targets to backend scheduling options with rich debug knobs.
◦ Engineered a CGRA-focused scheduler atop ISL/PPCG that constructs validity/coincidence/proximity constraints, applies wavefront + tiling reshaping, and emits PE-mesh-friendly temporal/mesh marks for downstream codegen.
◦ Built a unified CLI/Python tool (autopoly) that auto-detects LLVM/MLIR toolchains, runs scheduling, lowers MLIR to LLVM IR, and drives end-to-end build/test flows (run, check, build, test) with configurable pass pipelines.
- **PrecTuner** 2021 - 2023
A Holistic Approach to Automatic Mixed-Precision Code Generation and Tuning for Affine Programs
◦ Developed an approach to holistically generate mixed-precision code and predict its optimal performance, avoiding the need to evaluate all code variants.
◦ Implemented an automatic code generation and tuning framework, significantly reducing the burden of users to benefit from the reduced-precision optimization.
◦ Integrated mixed-precision code generation with various loop transformations, outperforming the state of the art and exhibiting a good scalability to parallel execution on both CPU and GPU.

SKILLS

- Proficient in programming languages such as C/C++, Shell, and Python
- Solid foundation in polyhedral compilation optimization theories and extensive experience in programming and debugging on Linux systems
- Familiar with end-to-end compilation based on MLIR targeting domestic hardware platforms
- Experienced in operator development and optimization for autonomous driving algorithms, such as UniAD, on both general-purpose and domestic platforms
- Skilled in using compilation frameworks and tools, including LLVM, ISL library, and PPCG
- Proficient in loop optimizations and commonly used techniques for performance testing
- Familiar with typical algorithms for implementing transcendental functions and precision tuning

REFERENCES

1. **Jie Zhao**
Professor
Hunan University
Email: jiezhao@hnu.edu.cn
Relationship: PhD Supervisor
2. **Jinchen Xu**
Associate Professor
Information Engineering University
Relationship: Master Co-Supervisor