Phone: +86 136 6276 2745
Email: zhouxulin2023@iscas.ac.cn

Xulin ZHOU

Github: xlinsist
Life Blog: xlinsist.github.io

Education

Institute of Software, Chinese Academy of Sciences | Software Engineering | M.Eng. **Sep. 2023 — Present**
- Research interest lies in high-performance computing systems, with a specific focus on compiler optimization.
- Contributed to open-source compiler infrustructures and optimization frameworks, including BuddyMLIR, VentusLLVM, Flaggems, TritonAscend, BiSheng and MindSpore. Mentored BuddyMLIR projects in OSPP 2023 and 2024.

Jinan University | Artificial Intelligence | B.Eng. **Sep. 2019 — Jun. 2023**
- GPA: 3.92/5 (89.2/100), ranked 2/47 in major. Awarded National Scholarship and Outstanding Graduate Honor.
- Dedicated to open-source communities since undergraduate; recognized as MindSpore Senior Developer in 2022.

Academic Papers

- **Synergizing Thread and Data Level Parallelism with Auto-Tuning on Multi-Core CPUs**(Submitted). Xulin Zhou, Hongbin Zhang, Mingjie Xing. *IEEE International Conference on Parallel and Distributed Systems (ICPADS)*, 2025
- **HybridSIMD: A Super C++ SIMD Library with Integrated Auto-tuning Capabilities**. Haolin Pan, Xulin Zhou, Mingjie Xing, Yanjun Wu. *Advanced Software Engineering(ASE)*, 2025
- **AutoConfig: Automatic Configuration Mechanism for Deep Learning Compilation Optimization**. Hongbin Zhang, Xulin Zhou, Mingjie Xing, Yanjun Wu, Chen Zhao. *Journal of Software*, 2024
- **VectorIR: A Unified Dynamic Abstraction for SIMD and Vector Computation**(Submitted). Hongbin Zhang, Liutong Han, Xulin Zhou, Hanghang Cao, Shihao Gao, Junyi Mei, Haolin Pan, Mingjie Xing, Yanjun Wu. *IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, 2026

Project Experiences

Tiling Optimization Framework on multi-core SIMD CPU **Jan. 2025 — Aug. 2025**
- **Background**: As a key technique to enhance kernel performance, tiling optimization decomposes computational tasks into hardware-friendly data blocks to leverage multi-level parallelization. Current tiling optimization methods (e.g., Triton) primarily target GPUs, with CPU implementations achieving only 5%-30% of GPU performance.
- **Challenges**: (1) Overlook of thread-level parallelism configuration during tile-to-threads mapping; (2) Time-consuming exploration in large tiling search space; (3) Inefficient SIMD register usage when assigning tiling workloads.
- **Contributions**: Modeled coordinated search space in kernel generation, designed heuristic pruning and register-aware SIMD strategy, implemented framework based on TritonCPU. Experimental results on x86 and RISC-V show 17% average performance improvement over TritonCPU autotune, with only 5% tuning overhead compared to TVM Ansor.

OpenCL Compiler Support for Ventus RISC-V GPU **Nov. 2024 — Jun. 2025**
- Contributed to OpenCL compiler VentusLLVM. Developed backend code generation to support half-precision data types. Exposed to Ventus-specific calling conventions, branch handling, and register bit-width extension modules.
- Mentored interns in resolving 10+ issues and organizing 8 technical documents. Hosted 3 tutorial workshops.

Deep Learning Compilation and Vectorization for RISC-V CPU **Dec. 2022 — Oct. 2024**
- Contributed to AI compiler BuddyMLIR. Built an end-to-end pipeline for LLaMA2 by converting Torch Aten to MLIR Linalg via a custom frontend; optimized compute-bound kernels (e.g., mm, bmm) with tiling; generated vectorized code with a custom RISC-V Vector Extension dialect to exploit dynamic vector configuration in backend.
- Mentored BuddyMLIR projects in OSPP 2023 and 2024; delivered a talk at RISC-V Summit China 2024.

Intern Experiences

Efficient NPU Kernel Generation with BiSheng Compiler (Huawei, Beijing) **Jul. 2025 — Sept. 2025**
- Contributed to tiling optimization framework TritonAscend and its compiler BiSheng. Implemented backend code generation with custom MLIR dialects for NPU-specific optimizations, extending operator support of TritonAscend.

Compiler Support for MindSpore Computational Graphs (Huawei, Shenzhen) **Jul. 2022 — Nov. 2022**
- Contributed to deep learning framework MindSpore. Added support for Python syntactic sugar and built-in functions in static graph mode, improving user experience; recognized as MindSpore Senior Developer in 2022.

Skills and Abilities

- Proficient in C++ with solid algorithm background; awarded bronze medal at ICPC Yinchuan Regional 2021.
- Familiar with PyTorch for deep learning, with practical exposure to transformer-based model training and inference.
- Experienced in parallel programming on multi-core CPUs. Specialized in MLIR code generation and vectorization.
- Served as open-source project maintainer and intern mentor, demonstrating effective collaboration and leadership.