Kangqi Zhang

zhkangqi@purdue.edu | zhkangqi@umich.edu | Google Scholar | GitHub | kangqizhang.com | LinkedIn
Ph.D. Student in ECE, Purdue University | B.S. Computer Engineering, University of Michigan

Education

Purdue University 2026/08 – Present
- Ph.D. in Electrical and Computer Engineering, Advisor: Jinghan Sun

University of Michigan 2024/08 – 2026/04
- Bachelor of Science in Computer Engineering, Summa Cum Laude, GPA: 4.0/4.0, Advisor: Nathaniel Bleier

Shanghai Jiao Tong University 2022/08 – 2026/08
- Bachelor of Science in Mechanical Engineering, Core GPA: 3.65/4.0

Coursework: Computer Architecture, Parallel Computer Architecture, Microarchitecture, Parallel GPU Programming, Operating Systems, Scalable Systems, Digital Integrated Circuit, Data Structures and Algorithms

Intellectual Property & Publications (*equal contribution)

H. Jin*, **K. Zhang***, J. Yang, B. Lyu, Q. Ding, R. Gao, N. Bleier. *Precision As You Need: Stochastic Computing Is a Dense Adaptive Quantizer.* NeurIPS 2026, accepted.

K. Zhang*, J. Wang*, R. Dreslinski, N. Bleier. *TurnRL: Reinforcement-Learning Driven Turn-Restricted Adaptive Routing for Chiplet-Based System.* ICCAD 2026, accepted.

W. Ni, **K. Zhang**, Z. Yu, O. Nelson, M. Lee, H. Cai, F. Porikli, J. Kim, Z. Liu, J. Zhao. *Double-P: Hierarchical Top-P Sparse Attention for Long-Context LLMs.* ICLR 2027, under review.

H. Jin*, B. Lyu*, **K. Zhang**, J. Yang, S. Walther, J. Postelnik, Q. Ding, J. Shin, C. Chiu, A. Xiong, N. Bleier. *Precision As You Need: A Stochastic-Computing Mixture-of-Precision Neural Network Accelerator.* HPCA 2027, accepted.

H. Jin, J. Yang, Z. Zhang, J. Shin, B. Lyu, **K. Zhang**, Y. Liu, N. Bleier. *Fengshui: Demystifying Chiplet Ecosystem and Bespoke Neural Network Accelerator Codesign.* MICRO 2026, accepted.

B. You, W. Ke, ..., **K. Zhang**, et al. *FrontierPhysics: Evaluating Agents for End-to-End Physics Research.* arXiv:2611.06910, 2026.

Y. Sun, X. Han, ..., **K. Zhang**, et al. *Agents' Last Exam.* arXiv:2606.05405, 2026.

PyBatSim V1.0: Multi-physics Simulation Software for Li-ion Cells
Co-developer, Software Copyright (China National Copyright Administration) 2024/10

Research Experience

Precision As You Need: Stochastic Computing for Mixed-Precision Inference, Advisor: **Nathaniel Bleier** 2026/01 – Present
- Reframed stochastic computing (SC) as a dense adaptive quantizer whose precision is set by bit-stream length rather than a fixed datapath, reducing each multiplication to a single AND/XNOR gate.
- Built a GPU library emulating SC matrix multiplication at scale and a retraining-free per-row mixed-precision policy, staying competitive with fixed-format INT quantization across vision tasks (NeurIPS 2026).
- Developed PaYN, an SC algorithm–hardware co-design with 128 selectable stream lengths per datapath, allocated per computation group by an offline-calibrated threshold table with no retraining (HPCA 2027).
- Restructured the SC systolic array to amortize stream-staging registers and per-stream accumulators; in TSMC 22 nm, PaYN cuts stream compute by 60–74% within 1.10× FP16 perplexity and reduces EDP by 2.0× over an INT8 array.

Chisel-Verilog Aligner (Undergraduate Thesis), Advisor: **Nathaniel Bleier** 2026/01 – 2026/05
- Developed a formally verified Chisel-Verilog Aligner leveraging Yosys and intermediate representations (FIRRTL) to precisely map optimized Verilog signals back to original Chisel source code with signal-level accuracy.
- Resolved traceability bottlenecks in agile hardware design, outperforming ML-driven methods and enabling downstream applications such as automated HDL debugging, semi-formal verification, and AI-driven testbench generation.

TurnRL: RL-Driven Turn-Restricted Adaptive Routing for Chiplet-Based Systems, Advisors: **Nathaniel Bleier, Ron Dreslinski** 2024/09 – 2025/11
- Developed fast turn-addition algorithms and baselines for arbitrary topologies and traffic patterns.
- Applied Deep Q-Learning with GNN to learn optimal turn sets under different traffic patterns.
- Evaluated with gem5 Garnet and PARSEC workloads; demonstrated robust latency reduction and scalability.
- Achieved 35.9% and 33.3% improvement in saturation injection rate on HexaMesh and MI300A-like topologies, and reduced queuing latency and end-to-end latency by 20.1% and 8.8% on congested NAS-NPB workloads.

Double-P: Hierarchical Top-P Sparse Attention for Long-Context LLMs, Advisors: **Jishen Zhao, Zhijian Liu** 2025/09 – 2026/01
- Proposed a hierarchical sparse attention framework jointly optimizing top-p accuracy, selection overhead, and attention cost.
- Performed coarse-grained top-p estimation at the cluster level using size-weighted centroids, then adaptively refined computation through a second top-p stage.
- Achieved near-zero accuracy drop with up to 1.8× attention computation reduction and 1.3× end-to-end decoding speedup over state-of-the-art methods.

Long-Context LLM Inference Acceleration with ANN Search, Advisors: **Jishen Zhao, Zhijian Liu** 2025/05 – 2025/10
- Proposed a CPU-GPU heterogeneous ANN pipeline for efficient KV-cache retrieval.
- Exploited inter-layer query similarity via cross-layer cluster reuse, forming a cross-layer search pipeline.
- Implemented an ILP-based scheduler to minimize end-to-end latency within system constraints.
- Achieved up to 3× end-to-end decoding speedup over PQCache at 128K context length while maintaining comparable accuracy.

Three-dimensional Multi-physics Modeling of Li-ion Batteries, Advisor: **Yanming Wang** 2023/02 – 2024/02
- Built a comprehensive 3D electrochemical-thermal-mechanical model of lithium-ion cells.
- Predicted battery aging behavior, enabling more accurate performance and longevity assessment.

Course Projects

EECS 470 Out-of-Order RISC-V Superscalar Processor 2025/02 – 2025/05
- Designed and implemented the RTL of an out-of-order RV32IM processor based on the MIPS R10K microarchitecture.
- Developed advanced features including N-way superscalar execution, early branch resolution, early tag broadcast, prefetching, non-blocking associative caches, a load/store queue with forwarding, and a cycle-accurate simulator.
- Integrated advanced branch prediction subsystems: TAGE, perceptron, and CNN-based predictors, along with a branch target buffer (BTB) and return address stack (RAS).
- Achieved an average CPI of 1.68 at a 10 ns clock period; the final design ranked top 2 in the course.

EECS 570 Directory-Based MSI/MESI Cache Coherence Protocol 2025/10 – 2025/11
- Designed latency-optimized three-hop MSI and MESI protocols with transient states for pending data and acknowledgments.
- Introduced Exclusive (E) state enabling silent upgrades to Modified without network traffic.
- Implemented out-of-order message receive support for robust handling of race conditions.
- Verified correctness using Murphi model checking with coherence, mutual exclusion, and value consistency invariants.

EECS 570 Parallel Global Router for VLSI 2025/11 – 2025/12
- Developed a CPU/GPU-accelerated parallel global router for VLSI designs with Steiner tree construction, maze routing, and 3-bend routing.
- Evaluated runtime scalability and routing quality on ISPD benchmark circuits.

EECS 573 Adaptive Feature Selection for Heterogeneous Core Scheduling 2025/09 – 2025/12
- Extended gem5/QEMU to extract comprehensive MICA features and modeled realistic migration costs.
- Designed an information-theoretic feature selection framework to find Pareto-optimal MICA subsets balancing accuracy and overhead.

EECS 471 GPU Optimization of CNN Forward Pass 2025/03 – 2025/04
- Implemented the forward pass of a convolutional neural network (CNN) layer in CUDA with performance-critical kernels.
- Applied shared memory tiling, register-level GEMM transformations, and loop unrolling, reducing runtime from 13.5s to 0.11s.

EECS 482 Operating Systems 2025/09 – 2025/12
- Built a preemptive user-level thread library with mutexes and condition variables, a demand-paging virtual memory pager with copy-on-write, and a concurrent multi-client network file server with a hierarchical file system.

Teaching

- Teaching Assistant for ECE 2150 (Intro to Electric Circuits) 2024/09 – 2024/12
- Grader for EECS 370 (Intro to Computer Organization) 2026/01 – 2026/05

Honors and Awards

- Student Travel Grant, ICCAD 2026 2026
- Summer Undergraduate Research in Engineering (SURE) Stipend, University of Michigan ($6,000) 2025/05
- Dean's Honor List at College of Engineering, University of Michigan 2024/12, 2025/04

Skills

Languages	C/C++, Python, CUDA, Verilog, SystemVerilog, Chisel, Mathematica, MATLAB, R
Tools	gem5 (Garnet), QEMU, Yosys, FIRRTL, Murphi
ML Frameworks	PyTorch, HuggingFace Transformers