Peijing Li

Email: peli@stanford.edu
Phone: +1 (515) 715-3989
LinkedIn: https://www.linkedin.com/in/peijing-li
Portfolio: https://peijli.github.io

EDUCATION

Stanford University Stanford, CA
Ph.D. and M.S. in Electrical Engineering Sep 2024 – Present

- M.S. expected December 2026. Advisor: Keith Winstein (Fall 2026)
- Selected Coursework: Computer Systems Architecture, Parallel Computing, Emerging Non-Volatile Memory Devices and Circuit Design, Formal Methods for Computer Systems, Engineering Design Optimization, Medical Imaging Systems

University of Michigan Ann Arbor, MI
B.S.E. in Computer Science Aug 2020 – May 2024

- Minor in Civil Engineering; graduated *summa cum laude*
- Selected Coursework: Computer Architecture, Advanced Compilers, Computer Security, Computer Vision, Software Engineering, Transportation Engineering, Linear Optimization

RESEARCH INTERESTS

Compilers and program analysis; formal methods for compiler and systems optimization; computer architecture and systems; performance measurement and profiling; hardware-software co-design

RESEARCH EXPERIENCE

Stanford Differentiated Access Memories Project Sep 2024 – Mar 2026
Compiler infrastructure, formal methods, and architectural support for heterogeneous on-chip memory systems in AI accelerator hardware.

1. First author of the foundational positional paper introducing the StRAM (short-term RAM) and LtRAM (long-term RAM) classification for non-hierarchical, software-managed heterogeneous on-chip memory.
2. First author and lead developer of GainSight, an open-source profiler that extracts cycle-accurate data lifetimes from accelerator workloads and correlates them against emerging memory device models to synthesize optimal heterogeneous on-chip memory compositions. Measured that 64% of GPU L1 accesses and 79% of systolic-array scratchpad accesses are sub-microsecond-lived, motivating retention-limited on-chip memory in place of uniform SRAM. Adopted as the evaluation methodology for a subsequent gain-cell memory compiler study and as the empirical basis for a peer-reviewed workshop paper on memory specialization.
3. Originated and led Flan, a compiler framework that formulates data placement across heterogeneous memory as an Optimization Modulo Theories problem solved by Z3, integrated into the MLIR pipeline at the bufferization boundary. Averaged 19.8% energy and 21.0% latency reduction against the best greedy heuristic baseline across nine JAX linear-algebra kernels and 34 memory configurations.
4. Redesigned the embedded DRAM partition, design-for-test module, and refresh controller for GEMMA, a taped-out 16 nm gain-cell eDRAM-based accelerator targeting Mamba-class state-space model inference, including a refresh policy driven by data lifetimes profiled with GainSight. Implemented the MXFP6 quantization and arithmetic logic, authored the AXI and DMA transaction specifications, and designed a high-throughput SIMD compute unit using high-level synthesis. Silicon has returned from fabrication and is under test.

PUBLICATIONS

[1] Peijing Li, Muhammad Shahir Abdurrahman, Rachel Cleaveland, Sergey Legtchenko, Philip Levis, Ioan Stefanovici, Thierry Tambe, David Tennenhouse, Caroline Trippel, and H.-S. Philip Wong. 2025. Towards Memory Specialization: A Case for Long-Term and Short-Term RAM. In *Workshop on Disruptive Memory Systems (DIMES '25)*. Association for Computing Machinery, Seoul, Korea (South), (Oct. 2025), 10. doi:10.1145/3764862.3768175.

[2] Samuel Dayo, Shuhan Liu, Peijing Li, Philip Levis, Subhasish Mitra, Thierry Tambe, David Tennenhouse, and H.-S. Philip Wong. 2025. The Future of Memory: Limits and Opportunities. arXiv:2508.20425 [cs]. (Sept. 2025). doi:10.48550/arXiv.2508.20425.

[3] Peijing Li, Matthew Hung, Yiming Tan, Konstantin Hoßfeld, Jake Cheng Jiajun, Shuhan Liu, Lixian Yan, Xinxin Wang, Philip Levis, H.-S. Philip Wong, and Thierry Tambe. 2025. GainSight: A Unified Framework for Data Lifetime Profiling and Heterogeneous Memory Composition. arXiv:2504.14866 [cs]. (Aug. 2025). doi:10.48550/arXiv.2504.14866.

[4] Xinxin Wang, Lixian Yan, Shuhan Liu, Luke Upton, Zhuoqi Cai, Yiming Tan, Shengman Li, Koustav Jana, Peijing Li, Jesse Cirimelli-Low, Thierry Tambe, Matthew Guthaus, and H.-S. Philip Wong. 2026. Heterogeneous Memory Design Exploration for AI Accelerators with a Gain Cell Memory Compiler. arXiv:2602.21278 [cs]. (Feb. 2026). doi:10.48550/arXiv.2602.21278.

[5] Yuheng Wu, Berk Gokmen, Zhouhua Xie, Peijing Li, Caroline Trippel, Priyanka Raina, and Thierry Tambe. 2026. LLM-FSM: Scaling Large Language Models for Finite-State Reasoning in RTL Code Generation. arXiv:2602.07032 [cs]. (Feb. 2026). doi:10.48550/arXiv.2602.07032.

FUNDING & AWARDS

Qualcomm Innovation Fellowship (North America) Apr 2026
Finalist

- Title: Refresh-Aware Scheduling and Memory Planning for AI Accelerators with Short-Term RAM
- Collaborators: Nathan Kaplan & Sharad Malik, Princeton University

Sunlin and Priscilla Chou Graduate Fellowship Sep 2024
First-year PhD fellowship at Stanford University

TEACHING EXPERIENCE

Stanford University Department of Computer Science Stanford, CA
Head Teaching Assistant, CS 217 Hardware Accelerators for Machine Learning Jan 2026 – Mar 2026

- Rebuilt the course from the ground up with two faculty instructors and three other teaching assistants after a four-year hiatus; enrollment reached roughly 100 students against a forecast of 40 to 60.
- Authored the SystemC and Verilog lab assignment sequence in which students build a large language model accelerator one component at a time, covering a systolic array, a memory hierarchy, and a compiler backend, culminating in an open-ended final project.
- Tuned DSP utilization throu gh Siemens Catapult and Xilinx Vivado to keep FPGA laboratory turnaround tractable at course scale.
- Designed the final-project rubric and gave in-person feedback on every project proposal, covering scoping, literature review, and implementation strategy.

Stanford University School of Engineering Stanford, CA
Research Mentor, Undergraduate Visiting Research Program Jun 2025 – Aug 2025

- Mentored a rising senior through a ten-week summer research project, defining the problem scope and onboarding the student to two active projects.

- Guided extension of the Synopsys ASIP Designer commercial toolchain with cycle-accurate memory-access logging, which the toolchain lacked, by instrumenting the RTL testbench memory module to emit load and store records with cycle numbers and addresses.
- Scoped a DMA-based DRAM subsystem for a VLIW and RISC-V AI core, including a three-instruction custom ISA extension and an HBM-parameterized latency model.
- Profiled four distinct processor designs across workloads from single CNN layers up to Llama-3-8B, plus quicksort, motion estimation, and FFT.

University of Michigan Department of Civil and Environmental Engineering Ann Arbor, MI
Instructional Aide, CEE 375 Sensors, Circuits, and Signals Jan 2023 – Apr 2023

- Ran weekly laboratory sections in which students built sensing and signal-processing circuits on breadboards using Arduino and MATLAB, supporting them on both the mathematics and the physical debugging.

PROFESSIONAL EXPERIENCE

TSMC North America San José, CA
Research Intern, Corporate Research Jun 2026 – Sep 2026

- Extended an in-house analytical performance model for ASIC and GPU inference to support workload-level profiling.
- Investigated embedded DRAM opportunities across activation, KV cache, and weight data types for large language model serving.
- Built an analytical cost model for multi-chip scaling behavior, unifying several interconnect topology families under a single latency and energy notation.

ASML Silicon Valley San José, CA
Software Test Engineer Intern May 2023 – Aug 2023

- Owned automation of nightly regression triage for focus-exposure modeling software.
- Replaced two to three hours of daily manual triage, distributed across a five-engineer team, with an unattended pipeline.
- Attributed each failure in a 300-case nightly suite, itself running roughly six hours, to its originating commit and owner in under ten minutes.
- Deployed the attribution scripts into the CI/CD pipeline with automated stakeholder reporting; the tooling remains in active production use three years later.

Dell Technologies Shanghai, China
Software Engineering Intern, Research & Development Jun 2021 – Aug 2021

- Automated lifecycle-management validation for VxRail hyperconverged infrastructure across the Kubernetes interface, the VMware ESXi hypervisor, and VMware Cloud Foundation using Python and Selenium.
- Cut a three-to-four-hour manual procedure to roughly thirty minutes running unattended, and eliminated most physical trips to the server room to check cluster status.
- Debugged physical server racks when experimental deployments failed to boot.

SKILLS

- **Software:** Python (PyTorch/JAX), C/C++, CUDA, LLVM/MLIR, Z3 SMT solver, RISC-V/ARM/x86 assembly
- **Hardware:** SystemVerilog, SystemC/TLM, AXI, HLS, logic synthesis
- **Tools:** Git, Linux administration, Slurm, Atlassian tools, Jenkins, qTest

SERVICE & LEADERSHIP

Leland Stanford Junior University Marching Band Stanford, CA
Props & Uniforms Staff Mar 2025 – Present

Tau Beta Pi, Michigan Gamma Chapter Ann Arbor, MI
Undergraduate Student Member Sep 2022 – May 2024

University of Michigan Engineering Student Government Ann Arbor, MI
Student Life Committee Member Sep 2020 – May 2022