Nathan Zhang, Ph.D.

nzhang32@gmail.com linkedin.com/in/nzhang32 github.com/pyprogrammer Google Scholar

Summary

Experienced systems and compiler engineer, specializing in **end-to-end compiler pipelines**, **abstraction design**, and **HW/SW co-design**. Proven track record of **bridging high-level semantics with low-level hardware constraints** to deliver foundational improvements. Recipient of the **SambaNova CEO Award** and **Google Silver Perfy Award** for technical excellence.

Professional Experience

SambaNova Systems | *Principal Compiler Engineer* *Palo Alto, CA* 2024 – Present

- **Compiler Architecture:** Lead the evolution of the **CUDA-equivalent** stack, including the **Intermediate Representation**, solver-based optimizations (**Z3**) for dynamic programs, and the lowering pipeline down to the assembler.
- **ML Framework Integration:** Lead the engineering of StableHLO and PyTorch integration into the proprietary compiler stack, including comprehensive correctness validation, and **co-lead** the design of performance annotation systems.
- **Hardware-Software Co-design:** Optimize critical interfaces and abstractions between the compiler stack, runtime, and next-generation hardware architectures.
- **Engineering Velocity:** Directed organization-wide **Bazel migration** (improving build times by $10\times$) and spearheaded **AI/LLM-powered tooling** for automated code cleanup and compiler bug triage.

Stanford University | *Graduate Student Researcher* *Stanford, CA* 2017 – 2024

- **Systems Architecture & Simulation:** Pioneered **Streaming Tensor Interfaces** for high-throughput, composable tensor libraries on FPGAs. Architected the **Dataflow Abstract Machine (DAM)** simulator, implementing a **high-performance dataflow overlay on multicore CPUs** via **communicating sequential processes (CSP)** that scales to 100s of cores with near-perfect strong scaling.
- **Kernel Fusion & IR Design:** Advised the architectural development of **FuseFlow** (ASPLOS '26) and **STeP** (ASPLOS '26). Contributed the core insight for FuseFlow that **topological sorts and inversions** on iteration graphs dictate intermediate memory materialization, and designed its **Fusion Tables** mechanism based on Streaming Tensor Interfaces. For STeP, designed the core IR abstraction utilizing **streams and stream operators** to naturally bound spatial and temporal state for dynamic parallelism. Enabled high-fidelity full-model simulation by architecting validation frameworks for both systems atop the **DAM** simulator.
- **Distributed Systems & Tensor Modeling:** Guided the development of **DFModel**, designing its core solver architecture to rigorously **model tensor lifetimes and collective communication costs** across varied data distribution schemes (e.g., sharding, replication) and storage (SRAM, DRAM, etc.). Developed solver-based mapping methodologies for large-scale accelerators (**SARA**) and contributed to probabilistic models for optimizing sparse tensor computations (**D2T2**).

Google | *Student Researcher / Software Engineering Intern (Multiple Teams)* *Mountain View, CA* 2016 – 2020

- **ML Compiler Backend & Code Generation:** Accelerated lattice regression model training by **12**$\times$ via **Cloud Dataflow**; optimized key kernels using **AVX2** for a **4**$\times$ **speedup** (2017). Later deployed a custom **MLIR-based** compiler backend performing optimizations such as branch-avoidance via sorting networks and weight compression – achieving **8-12**$\times$ **speedup** for microsecond-scale **batch-1 inference** (2020).
- **Custom Hardware Compilers:** Authored a pilot low-level **C compiler** for the Pixel Visual Core; its success **catalyzed** the formation of a dedicated engineering team (2016). Subsequently developed **C/C++ implementations** of lattice regression using the compiler, establishing its utility beyond image processing (2018).

Skills

Languages: Rust, Python, C++, Starlark (Bazel)
Domain Expertise: Programming Models, Dataflow Architectures, Satisfiability Modulo Theories (SMT), Constrained Optimization, MLIR

Education

Stanford University | *Ph.D. in Computer Science (Advisor: Kunle Olukotun)* *Stanford, CA* 2017 – 2024
*Thesis: **Scaling Dataflow: Programmability and Simulation***
University of California, Berkeley | *B.S. in Electrical Engineering and Computer Science* *Berkeley, CA* 2014 – 2017

Selected Presentations & Awards

- **SambaNova CEO Award (2025)**: Recognition for architectural contributions and technical roadmap development.
- **ISCA Distinguished Artifact Award (2024)**: For the Dataflow Abstract Machine Simulator framework.
- **Google Silver Perfy Award (2020)**: Recognition for technical vision, execution, and impact on production performance.

Selected Publications

[1] R. Lacouture, **N. Zhang**, R. Sharma, M. Siracusa, F. Kjolstad, K. Olukotun, and O. Hsu, "Fuseflow: A fusion-centric compilation framework for sparse deep learning on streaming dataflow," in *Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ser. ASPLOS '26, USA: Association for Computing Machinery, 2026, pp. 798–820, ISBN: 9798400723599. DOI: 10.1145/3779212.3790165. [Online]. Available: https://doi.org/10.1145/3779212.3790165.

[2] G. Sohn, G. Zhang, K. Hossfeld, J. Kim, N. Sobotka, **N. Zhang**, O. Hsu, and K. Olukotun, "Streaming tensor programs: A streaming abstraction for dynamic parallelism," in *Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ser. ASPLOS '26, USA: Association for Computing Machinery, 2026, pp. 1912–1932, ISBN: 9798400723599. DOI: 10.1145/3779212.3790229. [Online]. Available: https://doi.org/10.1145/3779212.3790229.

[3] R. Sharma, Z. Y. Xue, **N. Zhang**, R. Lacouture, F. Kjolstad, S. Achour, and M. Horowitz, "A probabilistic perspective on tiling sparse tensor algebra," in *Proceedings of the 58th IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO '25, Association for Computing Machinery, 2025, pp. 795–808, ISBN: 9798400715730. DOI: 10.1145/3725843.3756095. [Online]. Available: https://doi.org/10.1145/3725843.3756095.

[4] S. Ko, **N. Zhang**, O. Hsu, A. Pedram, and K. Olukotun, "Dfmodel: Design space optimization of large-scale systems exploiting dataflow mappings," 2024. arXiv: 2412.16432 [cs.AR]. [Online]. Available: https://arxiv.org/abs/2412.16432.

[5] **N. Zhang**, R. Lacouture, G. Sohn, P. Mure, Q. Zhang, F. Kjolstad, and K. Olukotun, "The Dataflow Abstract Machine Simulator Framework (Distinguished Artifact Award)," in *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, Jul. 2024, pp. 532–547. DOI: 10.1109/ISCA59077.2024.00046.

[6] P. Mure, **N. Zhang**, C. Trippel, and K. Olukotun, "Tags: A Framework for Distributed Event Ordering," in *PLArch Workshop @ ISCA*, Jun. 2023.

[7] **N. Zhang**, K. Canini, S. Silva, and M. Gupta, "Fast Linear Interpolation," *J. Emerg. Technol. Comput. Syst.*, vol. 17, no. 2, Apr. 2021, ISSN: 1550-4832. DOI: 10.1145/3423184. [Online]. Available: https://doi-org.stanford.idm.oclc.org/10.1145/3423184.

[8] **N. Zhang**, M. Feldman, and K. Olukotun, "High performance lattice regression on FPGAs via a high level hardware description language," in *2021 International Conference on Field-Programmable Technology (ICFPT)*, 2021, pp. 1–10. DOI: 10.1109/ICFPT52863.2021.9609893.

[9] Y. Zhang, **N. Zhang**, T. Zhao, M. Vilim, M. Shahbaz, and K. Olukotun, "SARA: Scaling a Reconfigurable Dataflow Accelerator," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, 2021, pp. 1041–1054. DOI: 10.1109/ISCA52012.2021.00085.

[10] **N. Zhang**, J. Liang, A. Tomlinson, F. Boensch, and A. Sahai, "Undergraduate-Led Survey Class to Improve CS Education for New Students," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE)*, ser. SIGCSE '20, Portland, OR, USA: Association for Computing Machinery, 2020, pp. 142–148, ISBN: 9781450367936. DOI: 10.1145/3328778.3366897. [Online]. Available: https://doi-org.stanford.idm.oclc.org/10.1145/3328778.3366897.

[11] R. Singhal, **N. Zhang**, L. Nardi, M. Shahbaz, and K. Olukotun, "Polystore++: Accelerated Polystore System for Heterogeneous Workloads," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 2019, pp. 1641–1651. DOI: 10.1109/ICDCS.2019.00163.

[12] **N. Zhang**, M. Driscoll, C. Markley, S. Williams, P. Basu, and A. Fox, "Snowflake: A Lightweight Portable Stencil DSL," in *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2017, pp. 795–804. DOI: 10.1109/IPDPSW.2017.89.