Altan Haan / Resume

Professional Experience

Graduate Student Researcher, UC Berkeley EECS, Berkeley, CA. (Aug. 2022 - May 2025)

- Conducted **research in compilers and high-performance computing** as part of my PhD program in Computer Science at UC Berkeley.
- Optimized sparse matrix multiplication algorithms on NUMA shared-memory systems. Interned at the Lawrence Berkeley National Lab (LBNL) in 2023 as part of this work.
- Investigated **database query optimization techniques** in the context of sparse computing (especially graph algorithms).
- Built prototype **e-graph based optimizing compilers**, and worked on natively supporting SSA-style intermediate representations in e-graphs.

Senior Software Engineer, OctoML, Seattle, WA. (Sep. 2020 - Aug. 2022)

- Worked broadly on the **Apache TVM Deep Learning Compiler**.
- Fixed lots of bugs in the compiler!
- **Adapted and implemented core optimizing compiler passes**, including liveness analysis which reduced memory usage by 10x in the VM runtime.
- **Developed a prototype deep learning training library on top of TVM**, as part of a multi-million dollar contract with AMD. **Demonstrated feasibility of training models (BERT and DLRM) on prototype AMD hardware** (MI50 and MI100, unreleased at the time).

Education

PhD in Computer Science, UC Berkeley EECS, Berkeley, CA. (Aug. 2022 - on leave)

- Currently on leave, looking to work more directly in the industry.
- **Graduate-level coursework:** Reinforcement Learning, Parallel Computing, Database Theory, Database Systems, Compiler Optimization and Code Generation.

BS in Computer Science, University of Washington, Seattle, WA. (Sep. 2016 - Jun. 2020)

- Cum Laude, Departmental Honors, with a Minor in Mathematics.
- Enrolled after middle school as part of UW's Early Entrance Program.
- **Graduate-level coursework:** Formal Methods, Programming Languages.

Publications

A. Haan, D. T. Popovici, K. Sen, C. Iancu and A. Cheung, "**To Tile or not to Tile, That is the Question**," 2024 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), San Francisco, CA, USA, 2024, pp. 449-458, doi: 10.1109/IPDPSW63119.2024.00096.

- A deep dive into performance-tuning algorithms for sparse matrix multiplication on multicore NUMA systems.

C. Hong, S. Bhatia, A. Haan, S. K. Dong, D. Nikiforov and A. Cheung, "**LLM-Aided Compilation for Tensor Accelerators**," 2024 IEEE LLM Aided Design Workshop (LAD), San Jose, CA, USA, 2024, pp. 1-14, doi: 10.1109/LAD62341.2024.10691748.

- Experiments with LLMs for optimizing numerical programs running on accelerators. Prototyped an iterative LLM-driven compilation loop on top of the Exo compiler framework.

Arash Ardakani, Altan Haan, Shangyin Tan, Doru Thom Popovici, Alvin Cheung, Costin Iancu, and Koushik Sen. 2024. **SlimFit: Memory-Efficient Fine-Tuning of Transformer-based Models Using Training Dynamics**. In Proceedings of the 2024 Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6218–6236, Mexico City, Mexico. Association for Computational Linguistics.

- Reduce memory requirements when finetuning by adaptively freezing model layers.

Marisa Kirisame, Steven Lyubomirsky, Altan Haan, Jennifer Brennan, Mike He, Jared Roesch, Tianqi Chen, Zachary Tatlock. "**Dynamic Tensor Rematerialization**." International Conference on Learning Representations. 2021.

- Reduce memory requirements when training by dynamically deallocating and recomputing tensors as needed.

Miscellaneous

Cycled across the US, Astoria, OR to Washington D.C. (May 2025 - Aug. 2025)

- Took some time off to see the country on two wheels.
- Cycled a bit over 4000 miles in a bit less than 3 months.