Juechu Dong

✉ joydong@umich.edu joydddd.github.io

SUMMARY

Juechu (Joy) Dong is a Ph.D. candidate at the University of Michigan, advised by Prof. Satish Narayanasamy. She studies emgering technologies in computer architecture and systems, with a focus on confidential computing and GPU kernel optimizations. Her research seeks to democratize kernel customization by building flexible and adaptive infrastructure for mapping novel algorithms to GPU hardware.

EDUCATION

University of Michigan - Ann Arbor **(exp.) 2027**
***Ph.D.**, Computer Science and Engineering*
Advisor: Prof. Satish Narayanasamy

Shanghai Jiao Tong University **2022**
***B.S.**, Computer Engineering*

University of Michigan - Ann Arbor **2022**
***B.S.E.**, Computer Engineering, Summa Cum Lauda*
GPA: 3.99/4.00

SELECTED HONORS

MLCommons ML and Systems Rising Star **2025**
Selected as one of 38 junior researchers worldwide fostering potential in ML and Systems research.

Meta 2024 Internship Project Spotlight: FlexDecoding **2024**
Awarded as one of 3 outstanding internship projects each year

Rackham Doctoral Intern Fellowship **2025**

Rackham International Student Fellowship (12,990 USD) **2023-24**

INDUSTRY EXPERIENCE

PyTorch group, Meta Inc. **2024,25**
Research Scientist Intern

- Contribute to TorchInductor & Helion DSL.
- Develop new techniques in PyTorch compiler with a focus on GPU performance optimization.
- Design GPU programming language for fast, flexible, and easy-to-use ML kernel authoring.
- Research new techniques for high-performance distributed GPU communication.
- Engage in the open source community to identify user needs and promote new features.

NVIDIA **2022,25**
Deep Learning Architect Intern

- Modeled and analyzed next-gen GPU memory features, including distributed shared memory, asynchronous transaction barrier, and Tensor Memory Accelerator (TMA).
- Designed and improved compiler abstractions and programming interfaces for GPU domain-specific languages.
- Developed modular framework for CuTe DSL decoupling algorithmic logic from GPU scheduling, enabling flexible kernel composition and ML framework integration.

PUBLICATIONS

[1] Juechu Dong*, Boyuan Feng*, Driss Guessous*, Yanbo Liang*, Horace He. ”Flex Attention: A Programming Model for Generating Optimized Attention Kernels”. In *Proceedings of Machine Learning and Systems 7.* **(MLSys ’25)** 2025.

- Develop a novel compiler-driven programming model that allows implementing the majority of attention variants in a few lines of idiomatic PyTorch code.

- Optimize customizable attention kernels to provides 1.1x - 1.3x speedup compared to FlashAttn2 by lowering customizable attention into a fast Triton kernel + taking advantage of sparsity.
- Adapt FlexAttention to efficiently support decoding, GQA and PagedAttention.

[2] Juechu Dong, Jonathon Rosenblum, Satish Narayanasamy. ”Toleo: Scaling Freshness to Tera-scale Memory Using CXL and PIM”. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 4.* **(ASPLOS ’24)** 2024.

- Scale trusted memory size from hundreds of MB to tens of TB by expanding the span of trusted from a single trusted processor to an entire platform including intelligent memories.
- Design a new scheme of freshness protection that reduces the space requirement by 50x.
- Reduce deployment cost by spacing sharing one intelligent memory device among multiple CPUs.

[3] Juechu Dong*, Xueshen Liu*, Harisankar Sadasivan, Sriranjani Sitaraman, Satish Narayanasamy. ”mm2-gb: GPU Accelerated Minimap2 for Long Read DNA Mapping”. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics.* **(BCB ’24**[1]**)** 2024.

- Accelerate computational intensive chaining step in the state-of-art long sequence mapping tool minimap2 using AMD GPU by 2.57-5.33x.
- Optimize towards ultra long reads of 100k+ to accommodate genome sequencing technology trend.
- Develop adaptive GPU scheduling algorithm to balance highly heterogeneous workload.

[4] Jonathon Rosenblum, Juechu Dong, Satish Narayanasamy. ”SECRET-GWAS: Confidential Computing for Population-Scale GWAS”. In *Nature Computer Science.* 2025.

- Develop a thousand-core platform on Azure Confidential Computing to conduct multi-institutional GWAS on millions of patients in less than a minute.
- Adapt Spark-based Hail genomic analysis framework to run on TEE under obliviousness requirement.
- Parallelize GWAS computation on 1k cores to achieve near linear speedup.

PROJECTS (Work in Progress)

Helion: **Python-embedded Domain-Specific Language (DSL) for High-Performance ML Kernels** *2025 – Present*

- Design and extend higher-level DSL for efficient distributed ML kernels authoring.
- Investigate into in-kernel communication and computation fusion for GPUs (PyTorch conf poster: Kraken)
- Implement cross-device communication/sync abstraction and lowering path in Helion.

GCnC: **GPU Kernel DSL Abstraction Briding Framework** *2025 - present*

- Designed abstraction framework bridging domain expertise and low-level GPU optimization in CuTe DSL.
- Developed compiler transformation preserving scheduling semantics for high-level kernel customization.
- Implement abstraction lifting data flow graph generator for CuTe DSL, enabling visualization & sync sanitinzation.
- *Targetting submission to ASPLOS 2027*

SKILLS

Programming Languages: C/C++, CUDA, python, Triton, verilog

Technologies/Frameworks:

ML Framework: PyTorch (TorchInductor, TorchDistributed)

GPU DSL: Helion, CuTeDSL, cuTile, Triton, ThunderKittens, CUDA, PTX

GPU Tuning: nsight-compute/nsight-sys, omniperf/omnitrace/rocprof

Simulation Software: SniperSim, DRAMSim, pinplay

Confidential Computing: Open Enclave SDK, Intel SGX

Architectures: AMD CDNA2 Instinct GPU, NVIDIA Hopper/Blackwell, Intel Xeon Phi, Out-of-order CPU

Compiler Design: MLIR, CUTLASS IR, TileIR, Helion IR

[1] ACM-BCB is the flagship conference of the ACM SIGBio.