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**
Rearch Scientist Intern

- Contribute to TorchDynamo, TorchInductor & TorchDistributed.
- 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**
GPU Deep Learning Architect Intern

- Model and analyze new memory features on next-gen GPUs such as distributed shared memory, asynchronous transaction barrier, etc.
- Analyze and optimize multi-GPU data movement for deep learning workloads using Tensor Memory Accelerator (TMA).

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 implement a higher-level DSL enabling efficient ML kernel authoring with minimal hardware expertise.
- Enable extensive automatic optimization space search (e.g., cache interleaving, persistence scheduling) for performance gains via concise code.
- Automate tensor memory layout management for developers using Python closure-based templating.

TEACHING

Instructional Aide & Graduate Student Instructor **2021 - 2024**

EECS470 Comp Arch; EECS471 Applied GPU Prog; EECS570 Parallel Comp Arch

SKILLS

Programming Languages: C/C++, CUDA, HIP, Triton, (system) verilog
Technologies/Frameworks:

ML Stack: PyTorch (TorchInductor, TorchDynamo)

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

Simulation: SniperSim, DRAMSim, pinplay

Confidential Computing: Open Enclave SDK, Intel SGX

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

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