Juechu Dong

joydong@umich.edu | joydddd.github.io

juechu-dong | joydddd | Juechu Dong | Juechu Dong

Ann Arbor, MI, USA

Bio

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 GPU programming models and confidential computing hardware. 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

Technische Universität Berlin *2020*
***Visiting Student**, Virtual Reality and Game Design*

McGill University *2019*
***Visiting Student**, Communication and Interpersonal Skills in Business*

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.

Honors and Awards

Honors & Recognitions

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

James B. Angell Scholar *2020-23*

Paper & Project Awards

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

Fellowships & Scholarships

Rackham Doctoral Intern Fellowship *2025*

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

John Wu & and Jane Sun Outstanding Scholarship (100,000 CNY) *2018-22*

SJTU Outstanding Academic Performance Scholarship *2018-20*

Publications

*=Enqual Contribution

Conference Papers

[C.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.

[C.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. DOI: 10.1145/3622781.3674180

[C.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.

Journal Articles

SECRET-GWAS: Confidential Computing for Population-Scale GWAS Jonathon Rosenblum, **Juechu Dong**, Satish Narayanasamy Nature Computer Science 2025

WorkShops

[W.1] **Juechu Dong***, Xueshen Liu*, Harisankar Sadasivan, Sriranjani Sitaraman, Satish Narayanasamy. "mm2-gb: GPU Accelerated Minimap2 for Long Read DNA Mapping". In *The 1st Workshop on Emerging Computer Systems Challenges and Applications in Biomedicine*. (BioSys @ASPLOS '24) 2024.

Poster

[P.1] **Juechu Dong***, Surya Subramanian. "Kraken: Hackable Triton Kernels for Computation and Multi-GPU Communication Fusion.". In *PyTorch Conference*. 2025.

Technical Reports & Blogs

[T.1] **Joy Dong**, Boyuan Feng, Driss Guessous, Joel Schlosser, Yanbo Liang, Horace He. "FlexAttention Part II: FlexAttention for Inference". In *PyTorch Blogs*. Apr 2025.

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

[T.2] Team PyTorch: Driss Guessous, Yanbo Liang, **Joy Dong**, Horace He. "FlexAttention: The Flexibility of PyTorch with the Performance of FlashAttention". In *PyTorch Blogs*. Aug 2024.

Under Submission / Preprints

[PP.1] Jonathon Rosenblum, **Juechu Dong**, Satish Narayanasamy. "Timelock Drive: Isolated Time-Based Defense for Storage Systems". In *under submission*. 2025.

[PP.2] Jason Ansel, Oguz Ulgen, Will Feng, Jongsok Choi, Neil Dhar, **Juechu Dong**, Driss Guessous, Markus Hoehnerbach, Jie Liu, Alexey Loginov, Shane Nay, Manman Ren, Nick Riasanovsky, Karthick Panner Selvam, Hongtao Yu, Paul Zhang, Xuan Zhang, Peng Wu. "Helion: A High-Level DSL for Performance Portable Kernels". In *under submission*. 2026.

[PP.3] Pierre Abillama, Changwoo Lee, **Juechu Dong**, David Blaauw, Dennis Sylvester, Hun-Seok Kim. "Memory-Efficient Acceleration of Block Low-Rank Foundation Models on Resource Constrained GPUs". In *arXiv*. arXiv:2512.20861 [cs.LG] 2025.

Invited Talks & Guest Lectures

[P.1] Juechu Dong. "Programming Abstractions for GPUs: Control vs. Complexity". In *EECS570: Parallel Computer Architecture, University of Michigan*. Host: Prof. Satish Narayanasamy Nov 2025.

[P.2] Juechu Dong. "Navigating the "Software Lottery": Flexible and Adaptable Programming Framework for AI Innovation". In *MLCommons Machine Learning & Systems Rising Star Workshop*. @Meta, Menlo Park. May 2025.

[P.3] Juechu Dong. "Programming Modern GPUs: Tensor Core and Beyond". In *EECS471: Applied Parallel Programming with GPUs, University of Michigan*. Host: Prof. Reetu Das. Apr 2025.

[P.4] Juechu Dong. "Powered by torch.compile: Simple, Flexible & Performant LLM Models". In *EECS483: Compiler Construction, University of Michigan*. Host: Prof. Lingjia Tang. Nov 2024.

Projects

Helion
2025 Present

Python-embedded Domain-Specific Language (DSL) for High-Performance ML Kernels [ ]

- 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.

cuFolium
2025 - present

GPU Kernel DSL Abstraction Briding Framework [ ]

- 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*

Teaching

Instructional Aide: Computer Architecture (EECS470) *2021 FA, 2022 WN*

Graduate Student Instructor: Applied Parallel Programming with GPUs (EECS471) *2023FA*

Graduate Student Instructor: Parallel Computer Architecture (EECS570) *2024WN*

SERVICE

Conference Review

International Symposium on Computer Architecture Artifact Evaluation Committee *2025*

USENIX Symposium on Operating Systems Design and Implementation Artifact Evaluation Committee *2025*

Annual Conference on Machine Learning and Systems Artifact Evaluation Committee *2025*

Organization

University of Michigan Computer Engineering Lab Reading Group Coordinator *2022-24*

Search Committee

University of Michigan Dean Search Committee *2024*

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