


JUECHU DONG

✉ joydong@umich.edu  [joydddd.github.io](https://github.com/joydddd)

SUMMARY

Juechu (Joy) Dong is a Ph.D. candidate at the University of Michigan, advised by Prof. Satish Narayanasamy. She studies emerging technologies in computer architecture and systems, with a focus on confidential computing and GPU kernel optimizations. Her work 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

Computer Science and Engineering, PhD

Topics: Computer Architecture, Confidential Computing, Computing for Biotechnologies

Advisor: Prof. Satish Narayanasamy **GPA:** 3.92/4.00

University of Michigan - Shanghai Jiao Tong University Joint Institute Aug 2022

Computer Engineering, Bachelor of Science

University of Michigan - Ann Arbor Apr 2022

Computer Engineering, Bachelor of Science in Engineering, Summa Cum Lauda

GPA: 3.99/4.00

SELECTED HONORS

Meta 2024 Internship Project Spotlight: FlexDecoding 2024

Awarded to 3 Internship projects selected by CEO Mark Zuckerberg as spotlight of the year

Rackham International Student Fellowship 2023-2024

Awarded to 25 outstanding students among all international PhD and MS students in the university

PUBLICATIONS

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. 2024. **ASPLOS '24**

- 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.

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. 2024. **BCB '24¹**

- 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.

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. 2025. **MLSys '25**

- 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.

¹ACM-BCB is the flagship conference of the ACM SIGBio

PROJECTS IN PROGRESS

SECRET-GWAS: Confidential Computing for Population-Scale GWAS

Nature Comp. Sci.

J. Rosenblum, J. Dong, S. Narayanasamy

under review

- 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.

Timelocked Storage for Ransomware Defense

J. Rosenblum, J. Dong, S. Narayanasamy

under submission

- Propose a new randomware defence mechanism that adds a strong layer of protection on top of conventional user credential based security via a trusted storage system.
- Design simple yet efficient interface between the disk and the operating system to provide safe rollback.
- Optimize the secure storage system to achieve near zero access overhead.

INDUSTRY EXPERIENCE

PyTorch group, Meta Inc.

May 2024 - Aug. 2024

Research Scientist Intern

- Build the fast, efficient and flexible attention API for decoding and GQA.
- Develop new techniques in PyTorch compiler with a focus on GPU performance optimization.
- Conduct performance analysis and optimizations on attention kernels.

NVIDIA

May 2022 - Aug. 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).

TEACHING

Instructional Aide & Graduate Student Instructor

2021 - 2024

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

SERVICE

University of Michigan - Shanghai Jiao Tong University Dean Search Committee

2024

Committee Member, Alumni Representative

University of Michigan Computer Engineering Lab Reading Group

2022 - 2024

Coordinator

UM-SJTU Joint Institute Alumni Association

2022 - present

Founder & Vice President

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/rocpf

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