


JUECHU DONG

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

SUMMARY

Juechu (Joy) Dong is a 3rd year Ph.D. student at the University of Michigan, advised by Prof. Satish Narayanasamy. Her research focuses on confidential computing and parallel optimization for GPUs. Her work seeks to advance confidential computing solutions for enabling privacy-preserving data analytics solutions ranging from population scale genomic analysis to generative AI.

EDUCATION

- University of Michigan - Ann Arbor** (exp.) 2027
Computer Science and Engineering, PhD
Topics: Computer Systems and Architecture, GPU Architecture, Confidential Computing
Advisor: Prof. Satish Narayanasamy
- 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
Selected Coursework: Comp. Architecture A, Compiler A+, Operating System A
GPA: 3.99/4.00

SELECTED HONORS

- Meta 2024 Internship Project Spotlight: FlexDecoding** 2024
Awarded to 3 Internship project picked 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
- James B. Angell Scholar** 2020-2023
Achieving "A Record" for 6 consecutive terms

PUBLICATION

- Toleo: Scaling Freshness to Tera-scale Memory Using CXL and PIM** ASPLOS'24
J. Dong, J. Rosenblum, S. Narayanasamy Accepted
 - 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.
- mm2-gb: GPU Accelerated Minimap2 for Long Read DNA Mapping** ACM BCB'24
J. Dong, X. Liu, H. Sadasivan, S. Sitaraman, S. Narayanasamy Accepted
 - 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.
- 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.

INDUSTRY EXPERIENCE

PyTorch group, Meta Inc.

May 2024 - Aug. 2024

Research Scientist Intern

- Project: FlexAttention: The Flexibility of PyTorch with the Performance of FlashAttention
- Build the fast, efficient and flexible attention API for inference with GQA support.
- 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

2021FA, 2022WN

EECS470 Computer Architecture

Graduate Student Instructor

2023FA

EECS471 Applied Parallel Programming with GPUs

Graduate Student Instructor

2024WN

EECS570 Parallel Computer Architecture

SERVICE

Computer Engineering Lab Reading Group

2022 - present

Coordinator

- Organize weekly paper reading presentations and discussions.
- Host talks from visiting researchers and professors.

UM-SJTU Joint Institute Alumni Association

2022 - present

Founder & Vice President

- **Alumni Engagement:** Organize alumni and student gatherings.
- **Relationship Building:** Involve in expanding SJTU - UM collaborations, connecting to JI sponsors, and building industry relationships.
- **Career Advising:** Organize students career development workshops.
- **Welcoming:** Host new student orientation events, organize airport pickups, and offer settle down help.
- **Student Support:** Support students during the stressful transition to start in a new university in a new country, and during urgent crisis.

SKILLS

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

Technologies/Frameworks:

ML Stack: Pytorch (TorchInductor, TorchDynamo), Triton

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

Formal Verification: Murphi,

SIMD: avx512, avx2 on Xeon Phi

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