

I'm a **Ph.D. candidate at the University of Michigan** (advised by Prof. Satish Narayanasamy), working on confidential computing and trusted hardware, with a strong background in GPU architecture and ML systems. I'm looking for a research or engineering internship position during the summer of 2024.

My research seeks to advance confidential computing solutions for enabling privacy-preserving data analytics solutions ranging from population-scale genomic analysis to generative AI. My approach is to develop trustworthy hardware, and use it to efficiently guarantee privacy from the rest of system components, including the operating system and system administrators.

I have made four specific contributions. First, I invented the **Toleo**. Today, trusted processors (Intel SGX) support only a few hundred MBs of secure memory space. Toleo is an innovative solution that expands trust to intelligent memory and scales secure memory space to tens of TeraBytes, which is a million times larger than what is feasible today (ASPLOS'24).

Secondly, in collaboration with Meta, I built FlexDecoding, the inference backend for **FlexAttention**. FlexDecoding addresses the lack of flexibility in supporting new attention variants in today's LLM infrastructure and combines the flexibility of the PyTorch compiler with the performance of expert-tuned decoding kernels. FlexAttention (with only training backend) was launched in August 2024 with 170k views on X. FlexDecoding is set to launch in October 2024.

Thirdly, I have helped build **SECRET-GWAS**, a privacy-preserving genome-wide association study platform on Microsoft Azure's confidential computing platform. SECRET-GWAS scales to over 1000 cores. We showed for the first time that we can perform regression analysis on large genomic datasets from multiple institutions in less than a few seconds, without revealing data to even the cloud service provider (under review for Nature Computational Science).


In addition to these thesis work, in collaboration with AMD, I have also developed new techniques to accelerate long-read genome sequencing (mm2-gb). It will soon be released as part of AMD Research Open-source Project and is accepted to ACM BCB'24.

Going forward, I plan to advance confidential computing to address privacy and safety concerns of generative AI. I'm also looking into providing safe storage against ransomware attacks through trusted hardware.

Recent release of NVidia's Hopper confidential computing feature brings GPUs into the trusted hardware family, and provides an opportunity to impact how ML models are trained and used. Current solution for GPU TEE heavily relies on software drivers that disallow important optimizations for CPU/GPU communication. I will seek solutions that integrate GPU and CPU into a unified TEE through low-level hardware primitives (such as through CXL-IDE feature (Integrity Data Encryption)) to allow finer granularity on data movement in order to construct a unified trusted CPU-GPU memory. I'm also open to explore other problems in integrating GPUs into the TEE system.

Please see next page for my CV and detailed background.

# JUECHU DONG

✉ [joydong@umich.edu](mailto: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