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 seeking a research internship position for the spring or summer of 2025.**

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 involves developing trustworthy hardware to efficiently guarantee privacy across system components, excluding the operating system and system administrators from the trust base.

I have made four key 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 (initially launched with only the training backend in Aug 2024) received 170k views on X. FlexDecoding is set to launch in Oct 2024.

Thirdly, I contributed to the development of **SECRET-GWAS**, a privacy-preserving genome-wide association study platform built on Microsoft Azure's confidential computing platform. SECRET-GWAS scales to over 1000 cores and we demonstrated for the first time that regression analysis on large genomic datasets from multiple institutions can be performed in a few seconds, without exposing data to cloud service provider (under review for Nature Computational Science).

Lastly, in collaboration with AMD, I developed new techniques to accelerate long-read genome sequencing (**mm2-gb**, published in ACM BCB'24). mm2-gb advances computing for genome mapping and alignment, removing computational bottlenecks in the sequencing pipeline to keep up with the ultra-long read trend. Our artifact has gained significant community attention and will soon be released as part of AMD Research Open-source Project.

Additionally, I am contributing to **Timelocked Storage**, a project aimed at minimizing the Trusted Code Base (TCB) for ransomeware defense and excluding human administrators, operating systems, even the main processor from the TCB.

Looking ahead, I plan to further advance confidential computing to address privacy and safety concerns of generative AI.

The release of NVidia's Hopper confidential computing feature brings GPUs into the trusted hardware family, and offers an exciting opportunity to impact how ML models are trained and used. Current GPU TEE solutions heavily relies on software drivers and limits key optimizations, such as DMA for CPU/GPU communication. I aim to explore solutions that integrate GPU and CPU into a unified TEE, leveragin finer granularity on data movement based by low-level hardware primitives (such as through CXL-IDE feature (Integrity Data Encryption)) to construct a unified trusted CPU-GPU memory. I'm also open to exploring other challenges related to integrating GPUs into the TEE system.

Please see the next page for my CV and detailed background.

– Joy

# 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. Her research focuses on confidential computing, computing for biotechnologies and GPU architecture. 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 Architecture, Confidential Computing, Computing for Biotechnologies
**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 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*

## 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.

## PROJECTS IN PROGRESS

**FlexAttention: Flexibility of PyTorch with the Performance of FlashAttn** ***PyTorch Blog***

*Joy Dong, Driss Guessous, Yanbo Liang, Boyuan Feng, Horace He*

- Flexible: An API that allows implementing many attention variants in a few lines of idiomatic PyTorch code.
- Performant: Provides 1.1x - 1.3x speedup compared to FlashAttn2 by lowering customizable attention into a fast Triton kernel + taking advantage of sparsity.
- Self Adaptive: optimized to efficiently support decoding, GQA and PagedAttention.

**Timelocked Storage for Ransomware Defense**

*J. Rosenblum, J. Dong, S. Narayanasamy*

- Shrink the Trusted Code Base (TCB) for randomware defense to a formally verified timelock gate-keeper in front of disks.
- Compatible with different flavors of versioning file system designs.
- Achieve near zero disk access overhead.

## INDUSTRY EXPERIENCE

**PyTorch group, Meta Inc.** **May 2024 - Aug. 2024**

*Rearch 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/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