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