William Won

Ph.D. Candidate in Computer Science
Georgia Institute of Technology

266 Ferst Dr NW
Klaus Advanced Computing Building Rm 3305
Atlanta, GA 30332
william.won@gatech.edu
www.willjwon.com
willjwon
willjwon

Research Interests

Architecture, Systems, and Algorithms for Machine Learning

Topic: Software-Hardware Optimizations for Distributed Machine Learning

- Simulation Infrastructure for Distributed Machine Learning Architecture
- Software-Hardware Optimizations for Collective Communication
- Training, Inference, and Fine-tuning of Large-Scale Machine Learning Workloads
- General Machine Learning Algorithms and Workloads

Education

Aug. 2019 – Present **Georgia Institute of Technology**, *Ph.D. Candidate in Computer Science*, Atlanta, GA
Thesis (Proposed): Software-Hardware Optimizations for Efficient Collective Communications in Distributed Machine Learning Platforms
- Specialization: Computer Architecture
- Advisor: Tushar Krishna

Sep. 2023 – Oct. 2023 **Massachusetts Institute of Technology**, *Visiting Student*, Cambridge, MA
Laboratory: Computer Science and Artificial Intelligence Lab (CSAIL)
- Visiting Student in Electrical Engineering and Computer Science Department (EECS)
- Host: Manya Ghobadi

Aug. 2019 – May. 2022 **Georgia Institute of Technology**, *M.S. in Computer Science*, Atlanta, GA
Specialization: Machine Learning
- Advisor: Tushar Krishna

Mar. 2015 – Feb. 2019 **Seoul National University**, *B.S. in Computer Science and Engineering*, Seoul, South Korea
Thesis: A Deep Learning-Based Encrypted Data Detection Technique for Ransomware Defense
- Graduated with Honors (Summa Cum Laude)
- Advisor: Jihong Kim

Research Experience

May. 2024 – Present **AMD Research**, *Research Intern (Part-time)*, Remote
Topic: Developing Infrastructure to Simulate and Optimize Collective Communications
- Participating in developing infrastructure for collective communication optimization
- Mentor: Bradford Beckmann, Ruchi Shah, Vinay Ramakrishnaiah

May. 2023 – Aug. 2023 **AMD Research**, *Research Intern*, Austin, TX
Topic: Developing Infrastructure to Simulate Large-scale Distributed Deep Learning Systems
- Participated in updating ASTRA-sim for end-to-end distributed deep learning simulation
- Developed infrastructure related to the support of Chakra Execution Trace
- Mentor: Bradford Beckmann, Kishore Punniyamurthy

Jan. 2022 – Aug. 2022 **Intel**, *Research Intern*, Remote
Topic: Researching the Optimization of Collective Communications
- Participated in software-hardware co-designing of distributed deep learning platforms
- Developed automated topology-aware collective algorithm synthesizer
- Mentor: Sudarshan Srinivasan, Midhilesh Elavazhagan, Ajaya Durg

Aug. 2019 – Aug. 2024 **Georgia Institute of Technology**, *Graduate Research Assistant*, Atlanta, GA
Topic: Software-Hardware Co-Design of Distributed Machine Learning Architecture via ASTRA-sim
- Laboratory: Synergy Lab
- Advisor: Tushar Krishna

Jul. 2017 – Dec. 2018 **Seoul National University**, *Undergraduate Research Assistant*, Seoul, South Korea
Topic: Researching Machine Learning Techniques and Architecture for Ransomware-proof SSD Design
- Developed file fragment type classifier using deep neural networks
- Developed encrypted data detector using deep neural networks and and big data analytics
- Participated in developing FPGA-based data encryption detection system
- Participated in ransomware-proof SSD development
- Laboratory: Computer Architecture and Embedded Systems Lab

Research Projects

Aug. 2020 – Present **ASTRA-sim**, *Georgia Institute of Technology*, Atlanta, GA
Topic: Developing ASTRA-sim Distributed Machine Learning Simulator
- Lead developer and maintainer of ASTRA-sim simulation infrastructure
- Developing network simulation backends for large-scale cluster simulation
- Writing documentation and running tutorials and presentations
- 256 GitHub stars and used by multiple industry and academia groups
- https://astra-sim.github.io

Jul. 2023 – Present **Chakra Working Group**, *MLCommons*
Topic: Standardizing Distributed Machine Learning Execution Traces
- Updating ASTRA-sim simulator to leverage standardized Chakra execution traces
- Contributing to the development of toolchains generating Chakra execution traces
- Discussing standardized formats to capture distributed machine learning workloads

Jan. 2023 – Present **SRC JUMP 2.0**, *Georgia Institute of Technology*, Atlanta, GA
Center: ACE (Evolvable Computing for Next-generation Distributed Computer Systems)
- Theme 3: Fine-grained communication and coordination
- Studying distributed machine learning infrastructure, architecture, and network fabrics
- Developing simulation infrastructure for modeling distributed machine learning

Jan. 2020 – May. 2021 **DARPA DRBE**, *Georgia Institute of Technology*, Atlanta, GA
Project: DRBE (Digital RF Battlespace Emulator)
- Topic: Developing large-scale, high-performance, real-time system for RF simulation
- Studied on-chip interconnection network architecture

Publications

C.7 **William Won**, Midhilesh Elavazhagan, Sudarshan Srinivasan, Swati Gupta, and Tushar Krishna, "TACOS: Topology-Aware Collective Algorithm Synthesizer for Distributed Machine Learning," in *Proc. of the 57th IEEE/ACM International Symposium on Microarchitecture (**MICRO**)*, 2024. (*to appear*)

C.6 Jinsun Yoo, **William Won**, Meghan Cowan, Nan Jiang, Benjamin Klenk, Srinivas Sridharan, and Tushar Krishna, "Towards a Standardized Representation for Deep Learning Collective Algorithms," in *Proceedings of the 31st IEEE Hot Interconnects Symposium (**HotI**)*, 2024.
- (W.2) Also appears in *Workshop on Machine Learning for Computer Architecture and Systems (**MLArchSys**)*, 2024.

C.5 **William Won**, Saeed Rashidi, Sudarshan Srinivasan, and Tushar Krishna, "LIBRA: Enabling Workload-aware Multi-dimensional Network Topology Optimization for Distributed Training of Large AI Models," in *Proc. of the IEEE International Symposium on Performance Analysis of Systems and Software (**ISPASS**)*, 2024.

W.1 Taekyung Heo, Saeed Rashidi, Changhai Man, Divya Kiran Kadiyala, **William Won**, Sudarshan Srinivasan, Midhilesh Elavazhagan, Madhu Kumar, Alexandros Daglis, and Tushar Krishna, "Exploring Memory Expansion Designs for Training Mixture-of-Experts Models," in *Workshop on Hot Topics in System Infrastructure (**HotInfra**)*, 2023.

C.4 **William Won***, Taekyung Heo*, Saeed Rashidi*, Srinivas Sridharan, Sudarshan Srinivasan, and Tushar Krishna, "ASTRA-sim2.0: Modeling Hierarchical Networks and Disaggregated Systems for Large-model Training at Scale," in *Proc. of the IEEE International Symposium on Performance Analysis of Systems and Software (**ISPASS**)*, 2023. (**equal contribution*)

C.3 Saeed Rashidi*, **William Won***, Sudarshan Srinivasan, Srinivas Sridharan, and Tushar Krishna, "Themis: A Network Bandwidth-Aware Collective Scheduling Policy for Distributed Training of DL Models," in *Proc. of the 49th International Symposium on Computer Architecture (**ISCA**)*, 2022. (**equal contribution*)

C.2 Eric Qin, Geonhwa Jeong, **William Won**, Sheng-Chun Kao, Hyoukjun Kwon, Sudarshan Srinivasan, Dipankar Das, Gordon E. Moon, Sivasankaran Rajamanickam, and Tushar Krishna, "Extending Sparse Tensor Accelerators to Support Multiple Compression Formats," in *Proc. of the 35th IEEE International Parallel & Distributed Processing Symposium (**IPDPS**)*, 2021.

C.1 Jisung Park, Youngdon Jung, **Jonghoon Won**, Minji Kang, Sungjin Lee, and Jihong Kim, "RansomBlocker: a Low-Overhead Ransomware-Proof SSD," in *Proc. of the 56th Annual Design Automation Conference (**DAC**)*, 2019.

Preprints

P.1 Saeed Rashidi, **William Won**, Sudarshan Srinivasan, Puneet Gupta, and Tushar Krishna, "FRED: Flexible REduction-Distribution Interconnect and Communication Implementation for Wafer-Scale Distributed Training of DNN Models," in *arXiv:2406.19580 [cs.AR]*, 2024.

Talks and Presentations

ASTRA-sim Tutorials and Presentations

Oct. 2024 **ASTRA-sim and Chakra Demo**, *SRC ACE Annual Review Meeting*, Rosemont, IL
Aug. 2024 **ASTRA-sim and Chakra Tutorial**, *HotI 2024*, Remote
Aug. 2024 **ASTRA-sim and Chakra Poster Presentation**, *SRC AIHW Review Meeting*, Wilmington, MA
May. 2024 **ASTRA-sim and Chakra Poster Presentation**, *PRC IAB 2024*, Atlanta, GA
Feb. 2024 **ASTRA-sim and Chakra Poster Presentation**, *CRNCH Summit 2024*, Atlanta, GA
Oct. 2023 **ASTRA-sim and Chakra Poster Presentation**, *SRC ACE Annual Review Meeting*, Champaign, IL
Sep. 2023 **ASTRA-sim and Chakra Talk**, *SRC TECHCON 2023*, Austin, TX
Sep. 2023 **ASTRA-sim and Chakra Poster Presentation**, *MIT AI and Cloud Workshop*, Cambridge, MA
Sep. 2023 **ASTRA-sim and Chakra Talk**, *SRC ACE Liaison Meeting*, Remote

Mar. 2023 **ASTRA-sim Tutorial**, *ASPLOS 2023*, Vancouver, Canada
Aug. 2022 **ASTRA-sim Tutorial**, *MLSys 2022*, Santa Clara, CA
Jun. 2022 **ASTRA-sim Tutorial**, *ISCA 2022*, New York, NY
Feb. 2022 **ASTRA-sim Tutorial**, *ASPLOS 2022*, Remote

Presentations of Other Works

Aug. 2024 **LIBRA and TACOS Talk**, *SRC ACE Liaison Meeting*, Remote
Aug. 2024 **LIBRA Lightening Talk and Poster Presentation**, *ModSim 2024*, Seattle, WA
Aug. 2024 **TACOS Talk**, *AMD Research*, Remote
Feb. 2023 **Themis Poster Presentation**, *CRNCH Summit 2023*, Atlanta, GA
Jun. 2022 **Themis Poster Presentation**, *ISCA 2022*, New York, NY

Paper Presentation at Conferences

May. 2024 **LIBRA Talk**, *ISPASS 2024*, Indianapolis, IN
Apr. 2023 **ASTRA-sim2.0 Talk**, *ISPASS 2023*, Raleigh, NC

Teaching

Aug. 2024 – Present **Graduate Teaching Assistant**, *Georgia Institute of Technology*, Atlanta, GA
Course: CS 6290 High Performance Computer Architecture
- Graduate teaching assistant.

Sep. 2023 **Guest Lecture**, *Harvard University*, Cambridge, MA
Class: CS 243 (Advanced Computer Networks)
- Topic: Topology-aware collective algorithms and research trends in distributed training
- Led a 40-minute guest lecture
- Host: Minlan Yu

Jul. 2017 – Aug. 2017 **Computer Architecture Course Development**, *Seoul National University*, Seoul, South Korea
Topic: Developing Hands-on Assignments using Y86-64 Instruction Set Architecture
- Developed hands-on assignments regarding Y86-64 implementation in Bluespec System Verilog
- Updated course materials and a textbook (Building Y86-64 using BSV)

Services

Dec. 2024 **Program Committee**, *MLForSys 2024 Workshop (at NeurIPS 2024)*, Vancouver, Canada
Nov. 2024 **Program Committee**, *CAMS 2024 Workshop (at MICRO 2024)*, Austin, TX
Nov. 2024 **Artifact Evaluation Committee**, *MICRO 2024*, Austin, TX
Mar. 2022 **Artifact Evaluation Committee**, *MLSys 2022*, Santa Clara, CA

Honors and Awards

Feb. 2019 **Undergraduate Thesis Award**, *Seoul National University*, Seoul, South Korea
Jun. 2018 **Presentation Award**, *Korean Institute of Information Scientists and Engineers*
Mar. 2017 – Feb. 2019 **National Scholarship for Science and Engineering**, *Korean Student Aid Foundation*
Merit-based

Skills

Software	Python, C/C++
(knowledgeable)	Java, JavaScript, TypeScript, Swift, Objective-C, OCaml, Scala
Hardware	Bluespec System Verilog, Verilog
Machine Learning	PyTorch, ASTRA-sim, Chakra
Tools	Git, LaTeX, HTML/CSS, Matplotlib/Seaborn, ggplot2
Language	English (Proficient), Korean (Native)