Insu Jang

4828 BBB, 2260 Hayward Street, Ann Arbor, MI 48109

insujang@umich.edu

https://insujang.github.io

Research Interests

System Architecture, Distributed Systems, Heterogeneous Computing, Systems for ML

Education

- **The University of Michigan** — Aug 2021 – Present
  Ph.D. Candidate in Computer Science and Engineering — Ann Arbor, MI, USA
  Advisor: Prof. Mosharaf Chowdhury
- **Korea Advanced Institute of Science and Technology (KAIST)** — Mar 2016 – Feb 2018
  M.Sc. in Computer Science — Daejeon, Republic of Korea
  Advisor: Prof. Jaehyuk Huh
- **Sungkyunkwan University (SKKU)** — Mar 2011 – Feb 2016
  B.Sc. in Computer Engineering — Seoul, Republic of Korea

Publications

1. Jae-Won Chung, Yile Gu, **Insu Jang**, Luoxi Meng, Nikhil Bansal, and Mosharaf Chowdhury. **"Perseus: Reducing Energy Bloat in Large Model Training."** *ACM Symposium on Operating Systems Principles **(SOSP)***, November 2024.

2. **Insu Jang**, Zhenning Yang, Zhen Zhang, Xin Jin, and Mosharaf Chowdhury. **"Oobleck: Resilient Distributed Training of Large Models Using Pipeline Templates."** *ACM Symposium on Operating Systems Principles **(SOSP)***, October 2023.

3. Jongyul Kim, **Insu Jang**, Waleed Reda, Jaeseong Im, Marco Canini, Dejan Kostić, Youngjin Kwon, Simon Peter, and Emmett Witchel. **"LineFS: Efficient SmartNIC Offload of a Distributed File System with Pipeline Parallelism."** *ACM Symposium on Operating Systems Principles **(SOSP)***, October 2021. **Best Paper Award.**

4. **Insu Jang**, Adrian Tang, Taehoon Kim, Simha Sethumadhavan, and Jaehyuk Huh. **"Heterogeneous Isolated Execution for Commodity GPUs."** *International Conference on Architectural Support for Programming Languages and Operating Systems **(ASPLOS)***, April 2019.

Research Experience

- **Resource Scheduling for Multimodal LLM** — Jan 2024 – Present
  Studying efficient resource scheduling for large scale multimodal large language model (MLLM). — University of Michigan

- **Fault Tolerant Distributed Training** — Sep 2021 – Oct 2023
  Studied efficient fault tolerance in large scale distributed training. Implemented Oobleck, a distributed training framework with pre-generated pipeline templates that can recover from failures fast by quickly reinstantiating a pipeline, instead of fully restarting the entire job. Oobleck has been published to SOSP'23. — University of Michigan

- **Offloading Replicated Storage Transactions to RDMA NIC** — Jan 2020 – Jul 2020
  Reimplemented Hyperloop to use it as a baseline of LineFS, which offloads replicated transaction into Infiniband RDMA adaptors. Studied Infiniband RDMA architecture and witnessed the benefits of offloading in reducing host CPU overload. LineFS paper has been published to SOSP'21 and won the best paper award. — KAIST

- **Architectural Support for Trusted Heterogeneous Execution** Mar 2016 – Feb 2018
  Designed a HW-SW codesigned architecture for GPU trusted execution environment. To realize it, studied the PCIe interconnect architecture and Intel SGX architecture. It focuses on providing protection in the path between the GPU and the CPU to support commodity GPUs for practicality. HIX paper has been published to ASPLOS'19. KAIST

Work Experience

- **Autopilot Software Engineer Intern (ML Infra)** May 2023 – Aug 2023
  Tesla Inc. Palo Alto, CA, USA
- **System Software Engineer – Fulfillment of Military Obligations** Feb 2018 – Jun 2021
  TmaxSoft Inc. Seongnam, Republic of Korea
- **Research Intern** Jan 2016 – Feb 2016
  Electronics and Telecommunications Research Institute (ETRI) Daejeon, Republic of Korea
- **Research Intern** Jul 2015 – Aug 2015
  Advanced Institute of Convergence Technology (AICT) Suwon, Republic of Korea
- **Student Member** Jan 2013 – Apr 2014
  Samsung Software Membership (Student Program of Samsung Electronics) Suwon, Republic of Korea

Teaching Experience

- **Graduate Student Instructor (GSI)** Fall 2024
  CSE585: Advanced Scalable Systems for Generative AI University of Michigan
- **Teaching Assistant** Spring 2017
  CS230: System Programming KAIST

Honors and Awards

- **Best Paper Award** Oct 2021
  "LineFS: Efficient SmartNIC Offload of a Distributed File System with Pipeline Parallelism"
  The 28th ACM Symposium on Operating Systems Principles (SOSP)
- **Richard H. Orenstein Fellowship in Memory of Murray Orenstein** Aug 2021
  Department of Electrical Engineering and Computer Science, The University of Michigan
- **Korea National Scholarship** Mar 2016
  KAIST and Korea Ministry of Science and ICT
- **Korea National Scholarship for Science and Engineering** Mar 2014
  Korea Student Aid Foundation and Korea Ministry of Education
- **Dean's List** Oct 2014, Apr 2015
  Department of Computer Engineering, Sungkyunkwan University