

Insu Jang

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

insujang@umich.edu
<https://insujang.github.io>

RESEARCH INTERESTS

Systems for ML, Efficient ML, Distributed ML, HW-SW Co-design

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. **Cornstarch: Distributed Multimodal Training Must Be Multimodality-Aware**
Insu Jang, Runyu Lu, Nikhil Bansal, Ang Chen, Mosharaf Chowdhury
arXiv Preprint 2025
2. **Mordal: Automated Pretrained Model Selection for Vision Language Models**
Shiqi He, **Insu Jang**, Mosharaf Chowdhury
arXiv Preprint 2025
3. **Reducing Energy Bloat in Large Model Training**
Jae-Won Chung, Yile Gu, **Insu Jang**, Luoxi Meng, Nikhil Bansal, Mosharaf Chowdhury
ACM SOSP 2024
4. **Oobleck: Resilient Distributed Training of Large Models Using Pipeline Templates**
Insu Jang, Zhenning Yang, Zhen Zhang, Xin Jin, Mosharaf Chowdhury
ACM SOSP 2023
5. **LineFS: Efficient SmartNIC Offload of a Distributed File System with Pipeline Parallelism**
Jongyul Kim, **Insu Jang**, Waleed Reda, Jaeseong Im, Marco Canini, Dejan Kostić, Youngjin Kwon, Simon Peter, Emmett Witchel
ACM SOSP 2021 – **Best Paper Award!**
6. **Heterogeneous Isolated Execution for Commodity GPUs**
Insu Jang, Adrian Tang, Taehoon Kim, Simha Sethumadhavan, Jaehyuk Huh
ACM ASPLOS 2019

RESEARCH EXPERIENCE

- **Resource Scheduling for Multimodal LLM** Jan 2024 – Present
University of Michigan
Cornstarch: A distributed multimodal LLM training framework. It focuses on unique characteristics of multimodal LLMs when it is distributed across multiple GPUs. Exploiting the characteristics, Cornstarch introduces novel scheduling insights: for balanced end-to-end execution, you need imbalance in parallelization.

- Fault Tolerant Distributed ML Training**

Sep 2021 – Oct 2023
 University of Michigan

Oobleck: An efficient fault tolerance in large scale distributed training. Oobleck introduces a groundbreaking way of fault tolerance ML; it exploits model states redundancy in data parallelism to recover lost states to avoid restart from the checkpoint, and utilizes every available GPUs by deploying heterogeneous pipeline parallel replicas. Oobleck introduces pipeline template for quick reconfiguration.
- Offloading Operations to RDMA NIC**

Jan 2020 – Jul 2020
 KAIST

LineFS: Reimplemented Hyperloop to use it as a baseline of LineFS, which offloads replicated transaction into Infiniband RDMA adaptors. Studied RDMA architecture and witnessed the benefits of offloading in reducing host CPU overload.
- Architectural Support for Trusted Heterogeneous Execution**

April 2017 – Oct 2018
 KAIST

HIX: Designed a HW-SW co-design 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.

WORK EXPERIENCE

- Autopilot Software Engineer Intern (ML Infra)**

May 2023 – Aug 2023
 Palo Alto, CA, USA

Tesla Inc.

 - **Straggler detection:** Design core algorithm of detecting stragglers in distributed ML training.
 - **Production deployment:** Implement, deploy, and integrate straggler detection algorithm into the infrastructure. Identified and helped fix several issues.
- System Software Engineer – Fulfillment of Military Obligations**

Feb 2018 – Jun 2021
 Seongnam, Republic of Korea

TmaxSoft Inc.

 - **Network subsystem:** Worked on implementing a network subsystem for TmaxOS.
 - **Virtualization:** Worked on researching virtualization technologies to improve I/O performance.
 - **Ceph & Kubernetes analysis:** Worked on analyzing Ceph distributed storage system to improve cloud storage performance.

TEACHING

- TA – CSE585 Advanced Scalable Systems for Generative AI, The University of Michigan Fall 2024
- TA – CS230 System Programming, KAIST Spring 2017

MENTORING

- **Runyu Lu:** PhD Student @ UM CSE
- **Kevin Xue:** PhD Student @ UM CSE
- **Minkyung Cho:** PhD Student @ UM CSE
- **Vatsal Joshi:** MS Student @ UM CSE
- **Luke Zhu:** MS Student @ UM CSE → Tesla

HONORS AND AWARDS

- Best Paper Award**

Oct 2021

“LineFS: Efficient SmartNIC Offload of a Distributed File System with Pipeline Parallelism”
- 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**
Korea Student Aid Foundation and Korea Ministry of Education

Mar 2014

- **Dean's List**
Department of Computer Engineering, Sungkyunkwan University

Oct 2014, Apr 2015

TECHNICAL SKILLS

- **Languages:** Python, C++, Rust, Triton, English (fluent), Korean (native)
- **Tools and Frameworks:** PyTorch, Cornstarch, Megatron-LM, DeepSpeed, ColossalAI, RDMA