Insu Jang

4828 BBB, 2260 Hayward Street, Ann Arbor, MI 48109
insujang@umich.edu
https://insujang.github.io

Research Interests

Systems for ML, ML Parallelism, Large-scale ML Systems, Adaptive Resource Scheduling

Education

- **The University of Michigan** — Aug 2021 – May 2026 (Expected)
  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. **Efficient Distributed MLLM Training with Cornstarch**
   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

- **Adaptive Resource Scheduling for Multimodal LLM** — Jan 2024 – Present
  University of Michigan
  Cornstarch [source]: A distributed multimodal LLM training framework. It optimizes imbalance across GPUs in pipeline parallelism and context parallelism by exploiting unique characteristics of multimodal LLMs.
  MANDu: While Cornstarch addresses the imbalance within a batch, multimodal LLM training also introduces cross-batch imbalance due to heterogeneous variability of multimodal inputs. MANDu introduces data-driven dynamic model parallelism to balance the workload across GPUs.

- **Fault Tolerant Distributed ML Training** Sep 2021 – Oct 2023
  Oobleck [source]: 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. University of Michigan
- **Offloading Operations to RDMA NIC** Jan 2020 – Jul 2020
  LineFS [source]: 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. KAIST
- **Architectural Support for Trusted Heterogeneous Execution** April 2017 – Oct 2018
  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. KAIST

Work Experience

- **Software Engineering Intern** May 2025 – Aug 2025
  ML Networking Team, Google LLC Sunnyvale, CA, USA
  - **Straggler detection**: Design and implement a framework that systematically analyzes stragglers in distributed ML training.
  - **ML in Kubernetes**: Worked on distributed ML training in Google Kubernetes Engine (GKE).
- **Autopilot Software Engineer Intern** May 2023 – Aug 2023
  ML Infrastructure Team, Tesla Inc. Palo Alto, CA, USA
  - **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
  System Kernel Team, TmaxSoft Inc. Seongnam, Republic of Korea
  - **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
- **Minkyoung Cho**: PhD Student @ UM CSE
- **Vatsal Joshi**: MS Student @ UM CSE → Meta
- **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** Mar 2014
  Korea Student Aid Foundation and Korea Ministry of Education
- **Dean's List** Oct 2014, Apr 2015
  Department of Computer Engineering, Sungkyunkwan University

Technical Skills

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