# Jie You

Email : jieyou@umich.edu
Mobile: +1-734-834-7781

## Education

- **Ph.D. in Computer Science and Engineering** — Sept 2016 – Jun 2022
  *Advisor: Prof. Mosharaf Chowdhury* — *University of Michigan*
- **B.S. in Information Engineering** — Sept 2012 – Jun 2016
  *GPA: 3.86/4.00* *Ranking: 2/162* — *Shanghai Jiao Tong University*

## Publications

[1] **Jie You***, Jae-Won Chung*, Mosharaf Chowdhury (*Equal contribution). Zeus: Understanding and Optimizing GPU Energy Consumption of DNN Training. NSDI 2023.
[2] **Jie You**, Jingfeng Wu, Xin Jin, Mosharaf Chowdhury. Ship Compute or Ship Data? Why Not Both?. NSDI 2021.
[3] **Jie You**, Mosharaf Chowdhury. Terra: Scalable Cross-Layer GDA Optimizations. arXiv:1904.08480. 2019.
[4] Fan Lai, **Jie You**, Xiangfeng Zhu, Harsha V. Madhyastha, Mosharaf Chowdhury. Sol: Fast Distributed Computation Over Slow Networks. NSDI 2020.
[5] Mosharaf Chowdhury, Samir Khuller, Manish Purohit, Sheng Yang, **Jie You**. Near Optimal Coflow Scheduling in Networks. SPAA 2020.
[6] Qi Alfred Chen, Matthew Thomas, Eric Osterweil, Yulong Cao, **Jie You**, Z Morley Mao. Client-side name collision vulnerability in the new gtld era: A systematic study. CCS 2017.
[7] Fan Lai, Wei Zhang, Rui Liu, William Tsai, Xiaohan Wei, Yuxi Hu, Sabin Devkota, Jianyu Huang, Jongsoo Park, Xing Liu, Zeliang Chen, Ellie Wen, Paul Rivera, **Jie You**, and Chun-cheng Jason Chen. AdaEmbed: Adaptive Embedding for Large-Scale Recommendation Models. OSDI 2023.

## Work Experience

- **Meta, Inc.** — **Research Scientist**
  *Manager: Intaik Park* — *Aug 2022 - Now*
  - **AdaEmbed: Reduce DLRM Training Memory Footprint by Up to 60%**: Applied adaptive embedding pruning technique to improving ML training efficiency and quality. Developed an in-training embedding pruning framework to reduce the size of embeddings needed for the same DLRM accuracy.
  - **DNN Training Quality and Stability**: Optimization of Machine Learning (ML) training quality and stability. Worked on improving training stability by preventing and mitigating gradient explosion. Also worked on improving training quality with state-of-the-art training algorithm such as Sparse Weight Decay and Nesterov Momentum.

## Academic Experience

- **Zeus: DNN Training Energy Saving up to 75% (NSDI '23)** — **Python/C++**
  *Advisor: Prof. Mosharaf Chowdhury* — *Jan 2021 - Apr 2022*
  - **Motivation**: Common practices of DNN training often lead to inefficient energy usage.
  - **Solution**: Characterized the energy-time trade-off of DNN training on GPUs. Designed an online algorithm which automatically finds optimal configurations for recurring DNN training jobs without offline profiling, improving the energy efficiency of DNN training by 15.3%–75.8%. Implemented the framework Zeus, which transparently integrates into PyTorch.
- **Adaptive Compute and Storage Disaggregation (NSDI '21)** — **Rust**
  *Advisor: Prof. Mosharaf Chowdhury* — *Jan 2019 - Dec 2020*
  - **Motivation**: Adaptively combining server-side and client-side processing together yields higher throughput and overall resource utilization in dissagregated in-memory KV stores.
  - **Solution**: Designed an online optimization algorithm that maximizes the request processing throughput while meeting latency SLO constraints. Implemented a prototype system and improved the throughput by 32.5%-63.4% comparing to the state-of-the-art solution.
- **Cross-Layer Optimization for Geo-Distributed Analytics** — **J**ava
  *Advisor: Prof. Mosharaf Chowdhury* — *Apr 2017 - Dec 2018*
  - **Motivation**: Geo-distributed analytics (GDA) frameworks transfer large datasets over the wide-area network (WAN). Existing solutions decouple WAN routing and GDA application transfer scheduling, resulting in missed opportunities for cross-layer optimizations. We want to bridge this gap between application and infrastructure.
  - **Solution**: Designed an efficient heuristic to co-optimize the scheduling and multi-path routing for GDA job execution, reducing the average JCT by 1.55×-3.43×. Implemented a WAN transfer framework that supports application-layer multi-path routing, integrated with Apache Spark and OpenFlow SDN controller.

## Professional Service

**Reviewer**, Conference on Information and Knowledge Management (CIKM) 2023
**Reviewer**, IEEE/ACM Transactions on Networking 2023
**Reviewer**, IEEE Transactions on Cloud Computing 2023
**Reviewer**, IEEE Transactions on Parallel and Distributed Systems 2023

## Technical Skills

**Programming Languages:** C++, Java, Python, Rust

## Awards

**Outstanding Winner**, Interdisciplinary Contest in Modeling (**top 0.2% worldwide**) 2015
**China National Scholarship** (top 1% nationwide) 2013, 2014, 2015