

Jae-Won Chung

+1 (734) 496-1803 | jwnchung@umich.edu | jaewonchung.me | jaywonchung | jae-won-chung-cs

Summary

I am a third year PhD candidate in CSE at the University of Michigan, working with [Professor Mosharaf Chowdhury](#). I build **cost-efficient software systems for deep learning**, with a recent focus on the efficient management of not only time, but also energy. I lead the [ML Energy initiative](#) and created [Zeus](#), an open-source library for deep learning energy measurement and optimization.

Education

University of Michigan

PH.D. CANDIDATE IN COMPUTER SCIENCE AND ENGINEERING

Ann Arbor, MI, USA

Sep 2021 - present

University of Michigan

M.S. IN COMPUTER SCIENCE AND ENGINEERING

Ann Arbor, MI, USA

Sep 2021 - Apr 2023

Seoul National University

B.S. IN ELECTRICAL AND COMPUTER ENGINEERING

Seoul, South Korea

Mar 2015 - Aug 2021

- GPA: 4.04/4.3 (overall) 4.15/4.3 (major), Summa Cum Laude. Period includes two years of military service.

Publications

- Andes: Defining and Enhancing Quality-of-Experience in LLM-Based Text Streaming Services**, Jiachen Liu, Zhiyu Wu, [Jae-Won Chung](#), Fan Lai, Myungjin Lee, Mosharaf Chowdhury, Preprint, 2024
- Toward Cross-Layer Energy Optimizations in Machine Learning Systems**, [Jae-Won Chung](#) and Mosharaf Chowdhury, Preprint, 2024
- Perseus: Removing Energy Bloat from Large Model Training**, [Jae-Won Chung](#), Yile Gu, Insu Jang, Luoxi Meng, Nikhil Bansal, Mosharaf Chowdhury, Preprint, 2023
- Chasing Low-Carbon Electricity for Practical and Sustainable DNN Training**, Zhenning Yang, Luoxi Meng, [Jae-Won Chung](#), Mosharaf Chowdhury, **ICLR Workshop: Tackling Climate Change with Machine Learning**, 2023
- Zeus: Understanding and Optimizing GPU Energy Consumption of DNN Training**, Jie You*, [Jae-Won Chung](#)*, Mosharaf Chowdhury, Symposium on Networked Systems Design and Implementation (**NSDI**), 2023 (Acceptance rate = 18.38%)
- ShadowTutor: Distributed Partial Distillation for Mobile Video DNN Inference**, [Jae-Won Chung](#), Jae-Yun Kim, Soo-Mook Moon, International Conference on Parallel Processing (**ICPP**), 2020 (Acceptance rate = 28.99%)

* Equal contribution

Research Experience

Energy-Efficient Systems for Machine Learning

SymbioticLab, UMich

ADVISOR: MOSHARAF CHOWDHURY

Sep 2021 - Present

- [Zeus](#): Discovered the trade-off between DNN training time and energy. Designed a Multi-Armed Bandit solution for time-energy optimization.
- [Perseus](#): A system for energy-efficient large model training. Cuts up to 30% energy without slowdown. Open-sourced as part of Zeus.
- [ML.ENERGY Leaderboard & Colosseum](#): The first systematic benchmark and interactive comparison service for LLM energy consumption.

Software Systems for Machine Learning

Software Platform Lab, SNU

ADVISOR: BYUNG-GON CHUN

Apr 2020 - Jun 2021

- [Crane](#): A GPU cluster manager for AutoML workloads. Built a Kubernetes backend that scaled to 288 GPUs. Contributed core features such as automatic bootstrapping on Docker Swarm and Kubernetes and log streaming through the EFK (Elasticsearch - Fluent Bit - Kibana) stack.

Online Model Specialization for Edge Video DNN Inference

Virtual Machine and Optimization Lab, SNU

ADVISOR: SOO-MOOK MOON

Dec 2019 - Jun 2020

- [ShadowTutor](#): Knowledge distillation from the server to the edge device reduced network data transfer by 95% and increased throughput by 3x.

Few-Shot Learning with Meta-Learning

Computer Vision Lab, SNU

ADVISOR: KYOUNG MU LEE

Jun 2019 - Dec 2019

- Designed improved meta-initialization methods for Model-Agnostic Meta-Learning (MAML) with neural memory modules and convex programs.

- Designed and implemented a full deep learning pipeline for QSM, a vision task for medical diagnostics with 3D MRI field data, including preprocessing (background removal, phase unwrapping, and patch slicing), augmentation (adding fake calcifications) and modeling ([CAD-QSMNet](#)).

Technical Skills

- Programming language proficiency**, Python (typing, sync and async), Rust (sync and async), Go, CUDA, C++, Zig, Verilog, Shell scripting
- Library/Framework familiarity**, PyTorch, Pandas, NumPy, Matplotlib, FastAPI, Pydantic, SQLAlchemy, Serde
- Tool familiarity**, Docker, Kubernetes, KubeFlow, Elasticsearch, Fluent Bit, Prometheus, Jaeger, OpenTelemetry, LaTeX
- Deep Learning systems optimization**, Experience in running and optimizing LLM training and inference serving. Small to medium code contributions to Text Generation Inference, vLLM, FastChat, and DeepSpeed.
- Deep Learning inference server deployment**, Publicly deployed an LLM chat service (The ML.ENERGY Colosseum) that can multiplex requests to multiple Text Generation Inference servers behind a NGINX reverse proxy. Contributed to Gradio in the process.
- Deep Learning and Computer Vision**, Experience in identifying and formulating deep learning problems and building up the data processing, training, and evaluation pipeline, including few-shot image classification, meta-learning, and medical imaging (MRI).

Open Source Projects

- BERT4Rec-VAE-Pytorch** (☆343 ♪ 80), [A PyTorch framework for recommendation model training](#), with abstract classes for pluggable model, dataset, and samplers. BERT4Rec and Netflix VAE models implemented.
- Reason** (☆190 ♪ 5), [A shell for managing research papers, written in Rust](#). Supports importing papers from file and URL, attaching markdown notes, and creating an HTML book with notes. Uses `serde` to persist data in human-readable and cloud sync-friendly format.
- Zeus** (☆183 ♪ 25), [A framework for deep learning energy measurement and optimization](#). **PyTorch ecosystem project**. Leading a team of two to three student contributors. Integrates best practices such as full type-annotation, auto-generated source code reference, Docker, Pytest, and examples.
- Pegasus** (☆30 ♪ 3), [An SSH command runner with a focus on simplicity, written in Rust](#). Runs multiple commands asynchronously using the `tokio` runtime and streams stdout and stderr back to the user. Battle-tested through multiple research projects and benchmarking.

Number of stars and forks are up-to-date as of July 20th, 2024.

Honors & Awards

Nov 2022 **Carbon Hack '22 Second Best Solution**, [Carbon-Aware DNN Training with Zeus](#), \$25,000

Green Software Foundation

Jul 2021 **Kwanjeong Overseas Scholarship**, \$25,000

Kwanjeong Educational Foundation

Mar 2019 **Kwanjeong Undergraduate Scholarship**, \$20,000 over two years

Kwanjeong Educational Foundation

Grants & Funding

Jan 2024 **Research grant**, \$20,000 for the development of the ML.ENERGY Initiative

Salesforce

Jan 2024 **Mozilla Technology Fund 2024**, \$50,000 for the development of the [Zeus](#) project

Mozilla

Invited Talks

Apr 2024 **Power and Energy Considerations in Machine Learning Systems**

University of Michigan (EECS 598)

Oct 2023 **Energy-Efficient Software Systems for Machine Learning**

Seoul National University

Oct 2023 **Energy-Efficient Deep Learning with PyTorch and Zeus**

PyTorch Conference

Sep 2023 **Energy-Efficient Deep Learning with Zeus**

Massachusetts Institute of Technology

Service

- Systems/Software Reading Group**, Paper reading group inside Michigan CSE, Organizer since Fall 2022

Teaching

- **Operating Systems (SNU, Spring 21)**, Lead TA, Managed Linux kernel hacking projects and led student team design reviews.
- **Computer Architecture (SNU, Fall 20)**, Peer tutor, Provided 30 hours of online lecture, **Best Tutor Award!**

Mentorship

- **Luoxi Meng**, Zeus open-source development, co-author of Perseus. Master's at UMich → PhD at UCSD
- **Yile Gu**, Co-author of Perseus. Master's at UMich → PhD at UW
- **Zhenning Yang**, Lead author of Chase, CarbonHack '22 second place award. Master's at UMich → PhD at UMich
- **Yong Seung Lee**, Zeus open-source development. Master's at UMich → Bloomberg
- **Yuxuan Xia**, ML.ENERGY Leaderboard contribution, Diffuserve. Master's at UMich.
- **Oh Jun Kweon**, Zeus open-source development. Master's at UMich.
- **Parth Raut**, Zeus open-source development. Master's at UMich.