Jae-Won **Chung**

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

Summary

I am a fifth year PhD candidate in CSE at the University of Michigan, working with Professor Mosharaf Chowdhury. I build **efficient software systems for deep learning**, with a recent focus on the efficient management of not only time, but also **energy**.

I view power and energy as fundamental systems resources that are worth carefully optimizing and allocating, not only in hardware, but also from software. Doing so provides automatic downstream benefits, such as reducing operational expenses, alleviating power delivery pressure for datacenters, and allowing the hardware to truly max out on performance.

I lead the ML Energy initiative as part of my research and open-source work.

Education

University of Michigan — *Ann Arbor, MI, USA*

Ph.D. candidate in Computer Science and Engineering — *Sep 2021 - present*

University of Michigan — *Ann Arbor, MI, USA*

M.S. in Computer Science and Engineering — *Sep 2021 - Apr 2023*

Seoul National University — *Seoul, South Korea*

B.S. in Electrical and Computer Engineering — *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

- **The ML.ENERGY Benchmark: Toward Automated Inference Energy Measurement and Optimization,** Jae-Won Chung, Jeff J. Ma, Ruofan Wu, Jiachen Liu, Oh Jun Kweon, Yuxuan Xia, Zhiyu Wu, Mosharaf Chowdhury, **NeurIPS Datasets & Benchmarks track (spotlight)**, 2025 (Spotlight acceptance rate = 2.81%)
- **Perseus: Reducing Energy Bloat in Large Model Training,** Jae-Won Chung, Yile Gu, Insu Jang, Luoxi Meng, Nikhil Bansal, Mosharaf Chowdhury, **SOSP**, 2024 (Acceptance rate = 17.34%)
- **Toward Cross-Layer Energy Optimizations in AI Systems,** Jae-Won Chung, Nishil Talati, and Mosharaf Chowdhury, **DOE ASCR Energy-Efficient Computing for Science Workshop**, 2024
- **Andes: Defining and Enhancing Quality-of-Experience in LLM-Based Text Streaming Services,** Jiachen Liu, Jae-Won Chung, Zhiyu Wu, Fan Lai, Myungjin Lee, Mosharaf Chowdhury, Preprint, 2024
- **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**

Experiences

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 (e.g., LLM). 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 GenAI energy consumption, including LLMs, Multimodal LLMs, and Diffusion models.

MoE Training Support on MTIA Platforms

AI and Systems Co-Design Team, Meta

Manager: Mengchi Zhang

May 2025 - Aug 2025

- MoE (Mixture-of-Experts) training support on MTIA platforms. Fixed issues and closed gaps across the entire stack, including MTIA kernels, PyTorch MTIA backend, collective communication, and Meta's internal large model training framework.

Software Systems for Machine Learning

Software Platform Lab, SNU

Advisor: Byung-Gon Chun

Mar 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*: Online knowledge distillation from the server to the edge device that specializes the on-device model to the target inference video at the moment 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 for few-shot classification tasks. Provided mathematical intuition for the improvements.

Quantitative Susceptibility Mapping (QSM) with Deep Learning

Lab of Imaging Science and Technology, SNU

Advisor: Jongho Lee

Jun 2019 - Aug 2019

- Created a full deep learning pipeline for QSM, a 3D MRI medical imaging task, including training data preprocessing, augmentation, deep learning modeling, and loss design (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, TorchTitan, 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** (☆394 ⑂93), A PyTorch framework for recommendation model training, with abstract classes for pluggable model, dataset, and samplers. BERT4Rec and Netflix VAE models implemented.
- **Zeus** (☆293 ⑂36), A framework for deep learning energy measurement and optimization. **PyTorch ecosystem project.** Constantly leading a team of student contributors. Integrates best practices such as full type-annotation, auto-generated source code reference, Docker, Pytest, and examples.
- **Reason** (☆194 ⑂4), 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.
- **Pegasus** (☆31 ⑂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 September 25th, 2025.

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

Aug 2025 **GitHub Secure Open Source Fund,** $10,000 for the development of the Zeus project — *GitHub*

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*

Selected Talks

Dec 2025 **Energy and Power as First Class ML Design Metrics** — *NeurIPS 25 Tutorial*

May 2025 **Energy-Efficient Systems for Machine Learning** — *Harvard Power and AI Initiative*

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

- **CSE585: Systems for Generative AI (UMich CSE, Fall 25),** GSI, Three lectures on GenAI and GenAI systems fundamentals.
- **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!**