# Dedong Xie

Website: ddxxdd-code.github.io
Email: dedongx@cs.washington.edu
LinkedIn: dedong-xie-547a501a8
GitHub: github.com/ddxxdd-code

## Education

**University of Washington** Seattle, WA, USA
PhD Sep. 2023–Jun. 2028 (expected)

**University of Toronto** Toronto, ON, Canada
Honours Bachelor of Science Jun. 2023

- Majors in computer science and mathematics, minor in statistics
- Cumulative GPA (cGPA): 3.99/4.0

## Publication

[1] **Dedong Xie**, Theano Stavrinos, Jonggyu Park, Simon Peter, Baris Kasikci, and Thomas Anderson. "PASS: A Power-Adaptive Storage Server". *EuroSys 2026*. To appear. 2026.

[2] **Dedong Xie**, Theano Stavrinos, Kan Zhu, Simon Peter, Baris Kasikci, and Thomas Anderson. "Can Storage Devices be Power Adaptive?" *Proceedings of the 16th ACM Workshop on Hot Topics in Storage and File Systems*. HotStorage '24. Santa Clara, CA, USA: Association for Computing Machinery, 2024, pp. 47–54. DOI: 10.1145/3655038.3665945.

[3] **Dedong Xie**, Zhen Jia, Zili Zhang, and Xin Jin. "Optimizing Half Precision Winograd Convolution on ARM Many-Core Processors". *Proceedings of the 13th ACM SIGOPS Asia-Pacific Workshop on Systems*. APSys '22. Virtual Event, Singapore: Association for Computing Machinery, 2022, pp. 53–60. DOI: 10.1145/3546591.3547529.

[4] Kan Zhu, Yufei Gao, Yilong Zhao, Liangyu Zhao, Gefei Zuo, Yile Gu, **Dedong Xie**, Zihao Ye, Keisuke Kamahori, Chien-Yu Lin, et al. "NanoFlow: Towards Optimal Large Language Model Serving Throughput". *19th USENIX Symposium on Operating Systems Design and Implementation (OSDI 25)*. 2025, pp. 749–765.

## Research Experience

**Research Assistant** Sep. 2023–Present
Syslab & EfesLab.
Paul G. Allen School of Computer Science and Engineering, University of Washington.
Advisors: Prof. Simon Peter and Prof. Baris Kasikci

- **Research project 1:** power adaptive storage system.
- Conducted a power measurement study of power consumption patterns of datacenter storage devices including HDDs and SSDs.
- Proposed how the datacenter storage devices can work to make storage system power adaptive, the ability to be able to adjust to dynamically changing power budgets.
- Published a first-author paper in ACM Workshop on Hot Topics in Storage and File Systems (HotStorage 2024).
- slides of my presentation in HotStorage 2024.

- **Followup work:** Designing a power adaptive storage server.
- Designed and implemented a power adaptive storage all-flash server for disaggregated storage.

- Proposed design based on principles of maximizing usage of power budget and use of cross-layer management to achieve optimal performance and widest power dynamic range.
- Using a cross-layer control approach—integrating hardware and software controls for both CPUs and SSDs—we demonstrated significant advancements in perfrmance and power dynamic range.
- Thoroughly studied the system over various microbenchmarks with fio and multiple application benchmarks of file systems and database system (Filebench, db_bench, and YCSB).
- Results showed that the system achieved the widest power dynamic range (94% of marginal power over idle power) and superior performance (about $10 - 100\times$) compared with best performing baselines in the dimensions of power dynamic range and performance.
- To appear as a first-author paper in EuroSys 2026.

- **Research Project 2:** NanoFlow: throughput oriented LLM serving framework
- Designed and implemented KV-cache offload to disk for NanoFlow, a throughput-oriented LLM serving framework.
- Focused on asynchronous migration of KV-cache pages between host memory and NVMe SSDs.
- Published in OSDI '25.

- **Research Project 3:** Profiler for software energy waste in ML applications
- Designed a energy profiler that can trace ML application energy consumption on kernel/operator level and find root cause of software-induced energy inefficiencies.
- With energy difference between different implementations of same functionality, identify energy optimization opportunities in software.
- Designed and implemented energy profiler and backward call stack tracing for root-cause localization.
- Leveraged CUDA graph and Pytorch Dynamo graphs for kernel replay and power measurement. Integrated traces from Pytorch profiler, CUPTI, and instantaneous power readings for analysis.
- Applied energy profiler to 9 different ML framework and applications. e.g. Pytorch, TensorFlow, vLLM, SgLang, HuggingFace Transformers, HuggingFace Diffusers.
- Found 18 existing software energy wastes and discovered 8 unknown cases.
- Submitted to ASPLOS '26.

### Research Intern

Jul. 2022–Feb. 2023

RISELab. University of California, Berkeley.
Supervisor: Prof. Ion Stoica

- Migrated BigScience BLOOM Large Language Model to use the parallel interface provided by Alpa, a LLM training and serving framework.
- Contributed to Alpa's compliance with the latest version of transformers package.
- Made a guide on serving Alpa on Slurm clusters.

### Research Assistant

May. 2022–Dec. 2022

SysNet Lab. Department of Computer Science. University of Toronto.
Supervisor: Prof. Eyal de Lara

- Participated in IBM CAS Canada project 1153 - Reducing JVM memory costs in the cloud https://www-40.ibm.com/ibm/cas/canada/projects?projectId=1153
- Sole developer of the run-time memory profiler of OpenJ9 JVM JIT-compiler.
- Proposed instrumenting dynamic memory allocation logger in OpenJ9's memory allocator.
- Proposed visual illustration of memory usage over time to find source of peak usage.
- Implemented the memory allocation logger, post-process pipeline, and visualizer with 3,000 lines of code in C++ and Python.
- Found external fragmentation and late release of memory to be main causes of memory inefficiencies.
- Proposed using program slicing to identify memory that could have a shorter lifetime.

  - Currently working on identifying the scope of each allocated memory.
  - Video of my presentation, and slides of the presentation.

**Research Intern** Jun. 2021–Jul. 2022

AI Lab. Amazon Web Services (AWS).
Supervisors: Dr. Zhen Jia (AWS) and Prof. Xin Jin (Peking University)

  - Sole developer of HAWC, a half-precision Winograd convolution system for Amazon Graviton-2 ARM architecture chips.
  - Proposed customized memory layout for Amazon Graviton-2 chips, ARM-specific matrix multiplication kernel generator, and minimal multi-threading scheduler to accelerate Winograd convolution.
  - Implemented 3000 lines of code using C++ and ARM assembly.
  - Studied and adapted baseline systems for comparison.
  - Extensively tested on a variety of representative convolution layers against state-of-the-art solutions.
  - Achieved on average 11× and up to 28× speedup.
  - Generated matrix multiplication kernels exploited up to 89% of theoretical maximum TFLOPS of the hardware.
  - Published a first-author paper in ACM Asia-Pacific Workshop on Systems (APSys 2022).
  - Video and slides of my presentation in APSys 2022.

## Coursework and Projects

- **CSE599K Systems for ML**
  **Theme:** Modern large language model (LLM) training and serving systems
  **Topics:** Transformers, GPU basics, Serving frameworks (vLLM, DistServe, etc.), Training frameworks (ZeRO), optimization techniques (parallelism paradigms, quantization, kernel fusion), MoE and RAG
  **Course assignments:** Implement attention with KV cache in pyTorch, profile prefill and decode attention performance on AMD MI200 GPU, RAG pipeline with Gmail data
- **CSE582 Course project**
  **Course project:** Life-cycle-assessment of LLM carbon footprint
  **Project description:** Analyzed checkpointing and different parallelization paradigms (data parallelism, tensor parallelism, pipeline parallelism) in training and serving and quantitative calculated their usage of resources including compute, memory, network communication and disk I/O. Then attribute the utilization of computer components like GPU, CPU, NIC, SSD to model for operational energy cost and thus operational and embedded carbon.

## Awards and Achievements

- **Dean's List Scholar** Jun. 2021, Jun. 2022, Jun. 2023
  - Faculty of Arts and Science, University of Toronto
- **Dr. James A. & Connie P. Dickson Scholarship In Science & Mathematics** Sept. 2022, Sept. 2023
  - "Given to the best students enrolled in science and mathematics programs."
  - University College, University of Toronto
- **Department of Computer Science Undergraduate Research Award** May 2022
  - Department of Computer Science, University of Toronto
- **Galois Awards in Mathematics** Oct. 2021
  - "Given to the best students enrolled in a mathematics specialist program."
  - University College, University of Toronto

- **The Faculty of Engineering Dean's Award** 2020
  - "For the best performance in year 1, 2 or 3."
  - University of New South Wales
- **COMP1511 (Programming Fundamentals) Hall of Fame** Sept. 2019
  - "A list of students who have achieved great distinction and honour by completing large amounts of extra work."
  - http://web.cse.unsw.edu.au/~cs1511/hall_of_fame/
  - COMP1511 Teaching Team, University of New South Wales

## Skills

- **Programming languages:** C/C++, Python, Java, Racket, and Haskell.
- **Assembly programming:** MIPS, ARM Aarch64, Intel x86 instruction sets.
- **Database management systems:** Microsoft Access, MySQL.
- **Mathematical computation and data analysis:** R, Mathematica, MATLAB.

## Languages

- **English:** proficient
  - **IELTS:** Overall 8.0 (Aug. 2019)
- **Chinese:** native speaker