Yannan (Nellie) Wu

606 Antioch Terrace, CA, 94085
Personal website: *https://nellie-wu.github.io*

Email : nellieywu@gmail.com
Mobile : +1-607-379-2186

Summary

I am a machine learning accelerator modeling engineer at Google. Before joining Google, I obtained my Ph.D. from MIT in computer architecture and systems. I have extensive research experience modeling and designing energy-efficient hardware accelerators for data and computation-intensive applications (such as deep neural networks), in both academic and industrial settings. My works have led to significant contributions to open-source industrial code bases, publications/tutorials at top-tier conferences (e.g., MICRO, ISCA), and a US patent application.

Education

Massachusetts Institute of Technology — Cambridge, MA
Ph.D. *in Electrical Engineering and Computer Science (GPA: 5.0/5.0)* — ***June 2023***
M.S. *in Electrical Engineering and Computer Science (GPA: 5.0/5.0)* — ***Feb. 2020***
Advisors: Prof. Joel Emer (emer@csail.mit.edu) & Vivienne Sze (sze@mit.edu)

Cornell University — Ithaca, NY
B.S. *in Electrical & Computer Engineering (GPA: 4.02/4.3)* — ***May. 2017***

Skills

- C++, Python, C, Verilog, FPGA, PyTorch, MATLAB, HTML, Linux, Git, Docker, Synopsys Design Compiler

Work Experience

- **Google TPU Modeling Engineer** — June, 2023 -
  - Developed analytical modeling infrastructure for various components in TPU chips.
  - Developed large language model training benchmarks in Tensorflow to aid TPU performance modeling.
  - Performed co-design studies to improve TPU compute performance for large language models.
- **NVIDIA Computer Architecture Research Intern** — May 2021 - Aug. 2021
  - Explored and analytically modeled various approaches to extend the existing Ampere GPU's sparse tensor core sparsity support to accelerate more structured sparsity patterns and structured sparsity degrees.
  - Developed pruning and fine-tuning procedures using PyTorch to realize various sparsity structures.
  - Filed a patent on a hardware-friendly and novel sparsity structure (US patent application number: 63/236,629).
- **NVIDIA Computer Architecture Research Intern** — May 2020 - Aug. 2020
  - Integrated an energy estimation backend (developed at MIT) to NVIDIA's deep neural network accelerator models.
  - Developed a statistical approach to analytically model the energy consumption of various sparse workloads (e.g., pruned deep neural networks).
  - Implemented the proposed modeling flow in C++ and contributed to a large NVIDIA code base.
- **Goldman Sachs Summer Technology Analyst** — June 2016 - Aug. 2016
  - Developed filtering functionalities for querying a database of balanced sheets.
  - Developed a web front-end in JavaScript and HTML to allow user-friendly specification of the filter.

Teaching Experience

- MIT 6.825 Hardware Architecture for Deep Learning Lead TA — Jan. 2020 - May. 2020
- MIT 6.888 Hardware Architecture for Deep Learning TA (part-time) — Jan. 2019 - May. 2019
- Cornell ECE 3140 Embedded Systems TA — Jan. - May. 2016 & Aug.- Dec 2015
- Cornell ECE 2300 Digital Logic & Comp. Arch. TA — Jan. - May. 2015 & Aug. - Dec. 2014
- Cornell MATH 1920 Multivariable Calc. Course Assistant — Jan. - May. 2015 & Aug. - Dec. 2014
- Cornell CS 1112 MATLAB Programming Course Consultant — Jan. 2014 - May. 2014

Selected Research Experience

- **Software-Hardware Co-design with Novel DNN Sparsity Structures** — *MICRO23*
  - Proposed a systematic way to define various structured sparsity patterns used in DNN pruning and proposed a new class of structured sparsity patterns to represent a variety of sparsity degrees.
  - Proposed a novel hardware design methodology to support the proposed structured sparsity patterns with light hardware overhead. Characterized the energy and area of important components with synthesized RTL.
  - Developed pruning/fine-tuning procedures using PyTorch to realize the target sparsity structures in various DNNs.

- **Analytical Modeling of Sparse Tensor Accelerators** *ISPASS21, MICRO22, tutorial at ISCA21*
  - Proposed a taxonomy to systematically describe the previously unstructured and confusing design space of sparsity-related hardware optimizations proposed by existing sparse tensor accelerators.
  - Proposed a decoupled methodology to statistically model sparse tensor accelerators by recognizing the orthogonality between several important design aspects.
  - Developed an ***open-source*** fast, flexible and accurate modeling framework, **Sparseloop**, to enable design space exploration of sparse tensor accelerators. Contributed >40,000 lines of C++ code to an NVIDIA codebase.
- **Flexible Energy and Area Estimation for Accelerator Designs** *ICCAD19, ISPASS20*
  - Proposed a systematic and flexible methodology to describe accelerator architecture organizations.
  - Based on the methodology, developed **Accelergy**, an ***open-source*** Python-based infrastructure for architecture-level (pre-RTL) energy and area estimation of accelerator designs.
  - Developed several ***open-source*** prototype energy and area estimation plug-ins for Accelergy to showcase Accelergy's flexibility to understand user-provided, process-dependent data.
- **Modeling of Fused-Layer DNN Accelerators** *ISPASS23, TCASAI24*
  - Mentored a master student to understand the design space of fused-layer DNN processing.
  - Participated in developing a methodology to systemically describe various fused-layer dataflows.
  - Participated in developing an analytical modeling framework that analyzes the runtime activities of the hardware components in fused-layer accelerators.

Publications and Patent

- **LoopTree: Exploring the Fused-layer Dataflow Accelerator Design Space**
  Michael Gilbert, Yannan Nellie Wu, Angshuman Parashar, Vivienne Sze, Joel S. Emer
  IEEE Transactions on Circuits and Systems for Artificial Intelligence (TCASAI), Sep. 2024
- **HighLight: Efficient and Flexible DNN Acceleration with Hierarchical Structured Sparsity**
  Yannan Nellie Wu, Po-An Tsai, Saurav Muralidharan, Angshuman Parashar, Vivienne Sze, Joel S. Emer
  IEEE/ACM International Symposium on Microarchitecture (MICRO), Oct. 2023
- **Accelerating Sparse Tensor Algebra by Overbooking Buffer Capacity**
  Fisher Xue, Yannan Nellie Wu, Joel S. Emer, Vivienne Sze
  IEEE/ACM International Symposium on Microarchitecture (MICRO), Oct. 2023
- **LoopTree: Enabling Exploration of Fused-layer Dataflow Accelerators**
  Michael Gilbert, Yannan Nellie Wu, Angshuman Parashar, Vivienne Sze, Joel S. Emer
  IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), April 2023
- **Sparseloop: An Analytical Approach to Sparse Tensor Accelerator Modeling**
  Yannan Nellie Wu, Po-An Tsai, Angshuman Parashar, Vivienne Sze, Joel S. Emer
  IEEE/ACM International Symposium on Microarchitecture (MICRO), Oct. 2022 ***(Distinguished Artifact Award)***
- **Architecture-Level Energy Estimation for Heterogeneous Computing Systems**
  Francis Wang, Yannan Nellie Wu, Matthew Woicik, Vivienne Sze, Joel S. Emer
  IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), March 2021
- **Sparseloop: An Analytical, Energy-Focused Design Space Exploration Methodology for Sparse Tensor Accelerators**
  Yannan Nellie Wu, Po-An Tsai, Angshuman Parashar, Vivienne Sze, Joel S. Emer
  IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), March 2021
- **An Architecture-Level Energy and Area Estimator for Processing-In-Memory Accelerator Designs**
  Yannan Nellie Wu, Vivienne Sze, Joel S. Emer
  IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), April 2020
- **A Systematic Approach for Architecture-Level Energy Estimation of Accelerator Designs**
  Yannan Nellie Wu
  Master Thesis, Massachusetts Institute of Technology, Feb. 2020
- **Accelergy: An Architecture-Level Energy Estimation Methodology for Accelerator Designs**
  Yannan Nellie Wu, Joel S. Emer, Vivienne Sze
  IEEE/ACM International Conference on Computer-Aided Design (ICCAD), Nov. 2019
- **Pruning and Accelerating Neural Networks with Hierarchical Structured Sparsity**
  Yannan Wu, Po-An Tsai, Saurav Muralidharan, Joel S. Emer
  US Patent Application Number: 63/236,629

Conference Tutorials

- **Sparse Tensor Accelerators: Abstraction and Modeling**
  Yannan Nellie Wu with Joel S. Emer, Vivienne Sze, Po-An Tsai, and Angshuman Parashar
  International Symposium on Computer Architecture (ISCA), June 2021
- **Tools for Evaluating Deep Neural Network Accelerator Designs**
  Yannan Nellie Wu with Joel S. Emer, Vivienne Sze, Angshuman Parashar, and Po-An Tsai
  IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Aug. 2020
  International Symposium on Computer Architecture (ISCA), June 2020
  IEEE/ACM International Symposium on Microarchitecture (MICRO), Oct. 2019

Selected Awards

- MICRO22 Distinguished Artifact Award — Oct. 2022
  Awarded to ONE paper accepted to MICRO22 based on reproducibility of experimental results
- MIT Jacob's Presidential Fellowship — Sept. 2017 - May. 2018
- Cornell ECE Early Career Scholarship — June. 2014 - Aug. 2014