

Yannan (Nellie) Wu

70 Pacific Street Apt. 729, Cambridge, MA, 02139
Personal website: <https://nellie-wu.github.io>

Email : nelliewu@mit.edu
Mobile : +1-607-379-2186

Education

Massachusetts Institute of Technology

Ph.D. in Computer Science and Electrical Engineering (GPA: 5.0/5.0)

M.S. in Computer Science and Electrical Engineering (GPA: 5.0/5.0)

Advisors: Prof. Joel Emer (emer@csail.mit.edu) & Vivienne Sze (sze@mit.edu)

Cambridge, MA

Aug. 2017 – present

Feb. 2020

Cornell University

B.S. in Electrical & Computer Engineering (GPA: 4.02/4.3)

Ithaca, NY

May. 2017

Summary and Objectives

I'm a Ph.D. candidate from MIT in computer architecture and computer systems, looking for full-time technical positions starting around June 2023. I have extensive 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 top-tier conference publications, a US patent application, and significant contributions to an open-source industrial code base.

Work Experience

- **NVIDIA Computer Architecture Research Intern** May 2021 - Aug. 2021
 - Investigated the limitations of the existing sparse tensor core accelerator in Ampere GPU.
 - Explored various approaches to extend the existing structured sparsity support in sparse tensor core.
 - 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 and area estimation backend (developed at MIT) to NVIDIA's DNN accelerator modeling tool.
 - Developed a statistical approach for analytically modeling the energy consumption of sparse tensor accelerators.
 - Contributed the proposed modeling flow 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 to allow user-friendly specification of the filter.

Research Experience

- **Software-Hardware Co-design with Novel DNN Sparsity Structures** *Under submission*
 - 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.
- **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 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* 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 Dense DNN Accelerators** *ISPASS20; tutorials at MICRO19, ISCA20, ISPASS21*
 - Participated in developing a dense DNN accelerator modeling framework by integrating Accelergy and Timeloop, an open-source infrastructure that analytically derives runtime activities of various hardware components.

- Participated in developing various *open-source* design specs to illustrate the flexibility of the framework.
- Studied the design characteristics for processing-in-memory-based DNN accelerator designs using the framework.
- **Modeling of Fused-Layer DNN Accelerators** *Under submission*
 - 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.

Selected Publications and Patent

- **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
- **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
- **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
Under submission
- **LoopTree: Enabling Exploration of Fused-layer Dataflow Accelerators**
Michael Gilbert, Yannan Nellie Wu, Angshuman Parashar, Vivienne Sze, Joel S. Emer
Under submission
- **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

Skills

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

Selected Courses

- Advanced Computer Architecture, Hardware Architecture for Deep Learning, Data Structures and Object-Oriented Programming, Digital Circuit Design, Operating Systems, Embedded Systems, Graph Analytics, Probabilities