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

Masssachusetts Institute of Technology — Cambridge, MA
Ph.D.** in Computer Science* — ***Aug. 2017 – present
M.S.** in Computer Science (GPA: 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: 4.0)* — ***May 2017

Research Interests and Objectives

Im 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

- **Graduate Research Assistant** — Massachusetts Institute of Technology
  Advised by: Professors Joel Emer & Vivienne Sze — *Aug. 2017 - Present*
  - **Flexible Energy and Area Estimation for Accelerator Designs *[ICCAD19; ISPASS20]***
    - Proposed a systematic and flexible methodology to describe various accelerator designs.
    - 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.
  - **Integrated Modeling Framework for Dense DNN Accelerators *[ISPASS20; ISPASS21; Tutorials at MICRO19, ISCA20, ISPASS20]***
    - Participated in developing a DNN accelerator modeling framework by integrating Accelergy and Timeloop, an open-source infrastructure that analytically derives runtime activities of 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.
    - Participated in studying the characteristics of heterogeneous compute systems using the framework.
  - **Design Space Classification and 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 sparse tensor accelerators.
    - Proposed a decoupled modeling methodology for sparse tensor accelerators by recognizing the orthogonality between several important design aspects.
    - Proposed a statistical modeling approach for sparse tensor accelerators' data-dependent behaviors.

    - Developed an **open-source** fast, flexible and accurate modeling framework, **Sparseloop**, to enable design space exploration of sparse tensor accelerators.
  - **Software-Hardware Co-design with Novel DNN Sparsity Structures** ***[Under Submission as the 1st Author]***
    - Proposed a systematic way to carefully define various structured sparsity patterns used in DNN pruning. Based on the taxonomy, proposed a new class of structured sparsity patterns to represent a variety of sparsity degrees.
    - Proposed a novel hardware design methodology to support the class of proposed structured sparsity patterns in activations and/or weights with light hardware overhead.
    - Developed pruning/fine-tuning procedures using PyTorch to realize the target sparsity structures.
    - Prototyped the important components in the hardware design with RTL to analyze the energy and area overhead.
  - **Modeling Fused Layer Processing of DNN** ***[Under Submission as the 2nd Author]***
    - Mentored a master student to understand the design space of fused layer processing.
    - Participated in developing a methodology to systemically describe the fused layer processing schedules.
    - Participated in developing an analytical modeling framework that analyzes the hardware component runtime activities with fused layer scheduling.
  - **Storage Idiom for Efficient Processing of Workloads with Long-tail Density Distribution** ***[In Progress as Project Lead]***
    - Proposed a new storage idiom for sparse tensor accelerators to better accommodate for sparse tensor workloads, whose density distributions have long tails, such as power law distribution.

- **Undergraduate Research Assistant** Cornell University
  Advised by: Professors Jose Martinez, Rachit Agarwal & Christina Delimitrou *Aug. 2014 - May 2017*
  - Reinforcement Learning Aided Garbage Collection for FLASH SSD
  - User-Centric Energy-Efficient Scheduling on Multi-Core Mobile Device
  - Analysis of Disaggregated Datacenter Performance

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), October 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
- **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
  IEEE 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
  - *ACM/IEEE International Symposium on Computer Architecture (ISCA)*, June 2020
  - *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, April 2020
  - *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Oct. 2019

Awards and Horners

- IEEE/ACM International Symposium on Microarchitecture (MICRO) Distinguished Artifact Award Oct. 2022
- MIT Jacob's Presidential Fellowship Sept. 2017 - May. 2018
- Member of IEEE-HKN Eta Kappa Nu Honor Society Oct. 2015 - May. 2017
- Cornell University College of Energineering Dean's List All Semesters (2013 - 2017)
- Cornell Undergraduate Research Funding June 2015 - Aug. 2015
- 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

Conference/Journal Reviewer Services

- *IEEE Journal of Solid-State Circuits (JSSC)*
- *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*
- *IEEE Transactions on Very Large Scale Integration Systems (VLSI)*
- *ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*

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

Leadership

- MIT Sidney Pacific Residence Hall Publicity Chair May. 2019 - May. 2020
- MIT Sidney Pacific Residence Hall Social Chair May 2018 - May 2019
- Cornell Society of Women Engineers General Body Chair Aug. 2014 - May 2015