Curriculum Vitae For Di Wu

Homepage: diwu1990.github.io
Phone: +1 608-886-2103
E-mail: di.wu@ece.wisc.edu

Research Interest

Approximate Computing, Unary Computing, Neural Network Optimization

Education

Sep 2017 – Now **Ph.D Student in Department of ECE** at University of Wisconsin–Madison, USA
- **Advisor**: Joshua San Miguel

Sep 2012 – Jan 2015 **M.E. in IC Engineering** at Fudan University, China
- **Ranking**: 3/67

Sep 2007 – Jun 2012 **B.S. in Micro-Electronics** at Fudan University, China

Honor

May 2020 **Gerald Holdridge Outstanding Teaching Assistant Awards** from UW–Madison
Jun 2019 **Chancellor's Opportunity Fellowship (COF)** from UW–Madison
Apr 2019 **ASPLOS 2019 SRC Grant** from ACM ASPLOS 2019, Providence, RI
Mar 2019 **Student Research Travel Grant–Conference** from UW–Madison
Mar 2019 **Student Research Competition (SRC) Candidate** in ACM ASPLOS 2019, Providence, RI
Mar 2019 **Final List** in Qualcomm Innovation Fellowship 2019, UW–Madison
Mar 2019 **Winner** in Foxconn SmartCity Competition 2019, UW–Madison
Feb 2019 **Winner of Student Travel Grant** in ACM ASPLOS 2019, Providence, RI
Oct 2018 **The Hiran Mayukh Award** on Computer Architecture Affiliate Meeting 2018, UW–Madison
Sep 2015 **Excellent New Employee Prize** in Hisilicon Technologies Co., Ltd
Mar 2014 **National Scholarship** for Excellent Graduate Student of Fudan University (1/67)
Sep 2007 **Third Prize Scholarship** for Excellent Freshman of Fudan University (3/45)
Jun 2007 **Ranking 88/450,000** in National College Entrance Examination in Anhui Province, China

Research

Feb 2018 – Now **Research Assistant in STACS Lab, Department of ECE, UW–Madison**

◇ **Approximate Computing**

- ***RAVEN***: A Reconfigurable Architecture for Varying Emerging Neural Networks. (*Qualcomm Innovation Fellowship 2019 Finalist*)
  - Support both linear and nonlinear functions based on MAC operations.
  - Support both dense and sparse matrix operations with dedicated NoC and sparse data format. (*in progress*)
  - Support different dataflows with dedicated PE array partition and scheduling. (*in progress*)
- ***Unified Nonlinear Operation (UNO)***: Design fixed-point UNO architecture to unify both linear and nonlinear operations with MAC operation based on Horner's rule. (*under review*)
  - Support popular nonlinear operations in deep learning and image processing.
  - Support run-time accuracy-energy scaling based on the order of Taylor series.
  - Validate benefits of UNO over counterparts in terms of area, power, and energy efficiency.
  - Apply UNO to Neural Machine Translation with negligible accuracy loss.

- ***Approximate Exponential Function***: Design a cross-layer optimization framework to achieve the desired accuracy-energy tradeoff for exponentiation based on Taylor series. (*C6, ISLPED 2019*)
  - Support both *static design-time* and *dynamic run-time* accuracy-energy scaling, using an approximate Taylor series to asymptote exponentiation, where division is substituted with shift and multiplication can be skipped.
  - At the algorithm level, optimize the approximate Taylor series of exponentiation with discrete gradient descent for run-time tradeoff.
  - At the circuit level, model the error bound of the approximate Taylor series based on an open source approximate circuit library for design-time tradeoff.

◇ **Unary Computing**

- ***UnarySim***: Construct a PyTorch-based library for unary computing, a computing paradigm based on bit streams. [Github] (*in progress*)
  - Support popular linear and nonlinear operations in unary computing.
  - Support varying unary computing schemes, including stochastic computing and race logic.
  - Provide end-to-end synthesis pipeline from PyTorch to Verilog HDL.
  - Provide demo of popular unary computing applications.
- ***In-stream Stochastic Multiplication***: Partition input bit streams into segments for fast and accurate multiplication, minimizing the impact of the bit distribution in the streams. (*in progress*)
- ***System Design Principle for Stochastic Computing***: Identify the propagation of normalized stability in stochastic computing system, with which the system design space can be reduced exponentially. (*C9, ASP–DAC 2021*)
- ***Unary GEneral Matrix Multiply (uGEMM)***: Design uGEMM architecture, featured with high parallelism, in-stream process, early termination, input insensitivity, resulting in high efficiency. (*C8, ISCA 2020*)
  - Unify the bit stream representation for different unary computing schemes, including both rate-coded stochastic computing and temporal-coded race logic, based on correlation and stability of bit streams.
  - Unify the linear unary computing operations of varying schemes with strict mathematical proof of the computing mechanism.
  - Apply uGEMM on deep neural networks with significantly higher efficiency and accuracy than existing implementations.
- ***Bit Stream Generation In DRAM***: Generate random bit streams for stochastic computing with off-the-shelf SDRAM on FPGA.
  - Design processing in memory (PIM) architecture for generating bit streams at maximum DRAM bandwidth.
  - Model the proposed architecture with an in-house simulator designed on our own.
  - Test the randomness of generated bit streams with NIST SP 800-22 statistical test suit.
- ***In-stream Stochastic Division and Square Root***: Utilize the correlation among bit streams to process the bit streams faster without buffering. (*C5, DAC 2019*, *J3, D&T 2020*)
  - Propose proper metrics to mitigate the fluctuation at the output due to feedback loops.
  - Build C++ based simulating platform towards fast and accurate stochastic computing simulation, which is the precedent of UnarySim. [Github]
  - Achieve higher hardware efficiency and accuracy than state-of-the-art.
- ***Stochastic CapsNet Architecture***: Optimize the nonlinear stochastic computing units, dataflow-level feedback loops, and logic sharing in Capsule Neural Network. (*ASPLOS 2019 SRC*)

◇ **Deep Learning**

- ***Security Side of Homomorphic Inference***: Perform security analysis on the inference process

of plaintext deep learning model and ciphertext data based on homomorphic encryption, and conclude that adversarial attacks can be detected under such threat model.
- ***Style Transfer for Medical Images***: Design a neural network with both residual and skip connections, and train it with MS-SSIM for style transfer from Hematoxylin-Eosin (H&E) image to Second Harmonic Generation (SHG) image for collagen fiber extraction. [slides]
- ***Automatic Quantization Platform***: Develop Tensorflow library to fast reach the best fixed-point quantization for Deep Neural Networks.

Sep 2017 – Jan 2018 Research Assistant in WiCIL Lab, Department of ECE, UW–Madison
◇ **Hardware Architecture for Convolutional Neural Networks**
- ***Sparse Neural Network***: Prune deep neural network in a hardware-aware manner based on network-on-chip theories to minimize the gap between dense and sparse deep neural network.
  - Analyze the hardware consumption of a channel-level pruned VGG16 model on ImageNet.
  - Model the performance for the channel-level pruned VGG16 model.

Aug 2012 – Jan 2015 State Key Laboratory of ASIC and System, Fudan University
◇ **High Throughput Low-Density Parity-Check (LDPC) Architecture Design**
- ***Stochastic LDPC Architecture***: Design the hardware architecture with better latency for 10GBASE-T with 45nm technology. (*J2, TCASII 2016*)
- ***Latency Reduction in Stochastic LDPC decoding***: Optimize latency by introducing a parameter initialization procedure and combining the original stochastic algorithm with a simplified Bit-Flipping algorithm, as well as utilizing precise posterior information. (*C2, ISCAS 2015*)
- ***Dual clock edge triggered LDPC decoding***: Design Min-Sum LDPC architecture working on both clock edges for WiMAX. (*C1, AISCON 2013*)

INDUSTRIAL EXPERIENCE

May 2020 – Aug 2020 Research Intern in Cerebras Systems.
◇ **Numerical Optimization for Deep Neural Network**
- Explore a noval floating-point format for training both CV and lanuage models on wafer-scale engine, CS-1, with FP32 and FP16 as reference.
- Simulate the training process on the real hardware with boundary casting in software simulator.

May 2019 – Aug 2019 Research Intern in Facebook Inc.
◇ **Lossless Quantization for Video Model**
- Perform post training quantization for ResNet-50 with Min/Max quantization and distribution-based quantization.
- Perform quantization-aware training for ResNext-50 with channel-wise 3D convolution.

Mar 2015 – May 2017 Digitial Circuit Engineer in Hisilicon Technologies Co., Ltd.
◇ **Algorithm Analysis for SoC**
- Analyze DSP resource consumption, covering parallelism and ISA.
- Optimize task scheduling for both DSP and SoC.
- Analyze the bandwidth for SoC workload.

◇ **Hardware Design for DSP**
- Design customized ARM AXI-4 bus for DSP cores in SoC;
- Design reconfigurable microarchitecture for multiple functions;

- Perform unit test with formal verification using the Cadence Jaspergold;
- Analyze power consumption for multiple DSP systems.

◇ **Design Automation**
- Automate post-placement-and-route simulation with customized platform;
- Design CAD report analyzing tool for entire hardware design flow;
- Design execution monitoring and analyzing tool for different DSP SDKs;
- Optimize ISA code length using discrete gradient descent.

PUBLICATION

Journal:

J3. **D. Wu**, etc., "In-Stream Correlation-Based Division and Bit-Inserting Square Root in Stochastic Computing", *under review*.

J2. **D. Wu**, Y. Chen, Q. Zhang, Y. Ueng and X. Zeng, "Strategies for Reducing Decoding Cycles in Stochastic LDPC Decoders," in *TCASII*, 2016.

J1. Y. Chen, Q. Zhang, **D. Wu**, C. Zhou and X. Zeng, "An Efficient Multirate LDPC-CC Decoder With a Layered Decoding Algorithm for the IEEE 1901 Standard," in *TCASII*, 2014.

Conference:

C10. **D. Wu**, etc., "UNO: Virtualizing and Unifying Nonlinear Operations for Emerging Neural Networks," *under review*.

C9. **D. Wu**, R. Yin and J. San Miguel, "Normalized Stability: A Cross-Level Design Metric for Early Termination in Stochastic Computing," in *ASP–DAC*, 2021.

C8. **D. Wu**, J. Li, R. Yin, H. Hsiao, Y. Kim and J. San Miguel, "uGEMM: Unary Computing Architecture for GEMM Applications," in *ISCA*, 2020.

C7. Y. Kim, J. San Miguel, S. Behroozi, T. Chen, K. Lee, Y. Lee, J. Li, and **D. Wu**, "Approximate Hardware Techniques for Energy-Quality Scaling Across the System," in *ICEIC*, 2020. (Invited)

C6. **D. Wu**, T. Chen, C. Chen, O. Ahia, J. San Miguel, M. Lipasti and Y. Kim, "SECO: A Scalable Accuracy Approximate Exponential Function Via Cross-Level Optimization," in *ISLPED*, 2019.

C5. **D. Wu** and J. San Miguel, "In-Stream Stochastic Division and Square Root via Correlation," in *DAC*, 2019.

C4. Q. Zhang, Y. Chen, **D. Wu**, X. Zeng and Y. Ueng, "Convergence-optimized variable node structure for stochastic LDPC decoder," in *ICASSP*, 2016.

C3. Q. Zhang, Y. Chen, **D. Wu**, X. Zeng and Y. Ueng, "An area-efficient architecture for stochastic LDPC decoder," in *DSP*, 2015.

C2. **D. Wu**, Y. C., Q. Zhang, L. Zheng, X. Zeng and Y. Ueng, "Latency-optimized stochastic LDPC decoder for high-throughput applications," in *ISCAS*, 2015.

C1. **D. Wu**, Y. Chen, Y. Huang, Y. Ueng, L. Zheng and X. Zeng, "A high-throughput LDPC decoder for optical communication," in *ASICON*, 2013.

Patent (CN):

P1. "A high-throughput LDPC decoder for Optical Communication." Y. Chen, **D. Wu**, Y. Huang

P2. "A multi-standard high performance FEC decoder." Y. Chen, **D. Wu**, Y. Huang

SKILL

Programming language: Verilog HDL, C++, Python, Perl, MATLAB

Hardware: Modelsim, Design Compiler, VCS, Verdi, Quartus Prime, Vivado
Software: Linux, PyTorch, TensorFlow, Caffe2, Vim, LaTeX, MS Visual Studio, MS office

Academic Activity

Seq 2020 **External Reviewer** for HPCA 2021
Sep 2020 **Teaching Assistant** for ECE454 (Mobile Computing Lab), UW–Madison
Feb 2020 **Guest Lecture on Introduction to Unary Computing** on ECE757 (Advanced Computer Architecture II), UW–Madison
Jan 2020 **Teaching Assistant** for ECE554 (Digital Circuit Lab), UW–Madison
Jan 2020 **External Reviewer** for ISCA 2020
Jan 2020 **External Reviewer** for DAC 2020
Jan 2020 **External Reviewer** for ASPLOS AE 2020
Seq 2019 **Teaching Assistant** for ECE554 (Digital Circuit Lab), UW–Madison
Apr 2019 **External Reviewer** for MICRO 2019
Apr 2019 **ASPLOS 2019** attendee in Providence, RI, USA
Mar 2019 **Guest Lecture on Deep Neural Network and Stochastic Computing** on ECE752 (Advanced Computer Architecture I), UW–Madison
Feb 2019 **External Reviewer** for ISCA 2019
Jan 2019 **Research Assistant** with Prof. Joshua San Miguel
Jan 2019 **Teaching Assistant** for ECE554 (Digital Circuit Lab), UW–Madison
Seq 2018 **Teaching Assistant** for ECE552 (Introduction to Computer Architecture) and ECE554 (Digital Circuit Lab), UW–Madison
Jan 2018 **External Reviewer** for DAC 2018
Seq 2017 **Research Assistant** in WiCIL

Non-Academic Activity

Nov 2011 **Campus Computer Gaming Champion of DOTA** in Fudan University
May 2010 **Volunteer** for EXPO 2010 in Shanghai
Mar 2010 **Excellent Student Leader** for Excellent Student Union Leader of Fudan University
Sep 2009 **Minister of Sports** in the Student Union of Fudan University
2009,10,12 **Campus Basketball Champion** in Zhangjiang Campus of Fudan University
May 2008 **Volunteer** for helping disabled students of Guanxin School in Shanghai