Weihong Xu

9500 Gilman Dr, La Jolla, CA 92093, USA

wh-xu.github.io | wexu@ucsd.edu | +86 186-5183-3181

RESEARCH INTERESTS

- **Computer Architecture and Domain-specific Accelerator Design**
- **Processing in Memory and Near Storage Computation**

EDUCATION

University of California San Diego — **La Jolla, USA**
Ph.D. in Computer Science — *Oct. 2020 - Present*
- Advisors: Prof. **Tajana Šimunić Rosing**
- Major Courses: Parallel Computation and Embedded Systems

Southeast University — **Nanjing, China**
M.E. in Information and Communication Engineering — *Sept. 2017 - June. 2020*
- Thesis: Application of Neural Networks in Baseband Processing and their Efficient Implementations
- Advisors: Prof. **Chuan Zhang** and Prof. **Yair Be'ery** from Tel Aviv University, Israel
- Major Courses: Digital Signal Processing and Fundamentals of Information Theory

Southeast University — **Nanjing, China**
B.E. in Information Engineering — *Sept. 2013 - Jun. 2017*
- Thesis: Acceleration of Convolutional Neural Networks based on Fast Algorithms
- Outstanding Bachelor Thesis Award, Advisor: Prof. **Chuan Zhang**
- Major Courses: Digital Communications, Communication Network, Computer Architecture and ASIC Design

RESEARCH EXPERIENCE

Processing in Memory Computing System Design — **UC, San Diego**
Research Assistant, advised by Prof. ***Tajana Šimunić Rosing*** — *Oct. 2020 - Present*
- Designed energy-efficient in-memory architectures and accelerators for attention models.
- Developed processing in memory-based servers and clients for Fully Homomorphic Encryption (FHE).
- Near storage computation system for hyperdimensional computing.
- Related publications: [C7]

Energy-efficient Accelerator Design for Convolutional Neural Network — **Southeast University**
Research Assistant, advised by Prof. ***Chuan Zhang*** — *Feb. 2017 - Aug. 2019*
- Reduced the computational complexity of convolution layers by 44% on ResNet-50 through exploiting *fast Fermat number transform*.
- Developed low bit-width and logarithm quantization methods to compress CNN models by 5.3× and speed up inference tasks without multiplication.
- Designed and implemented reconfigurable hardware architectures on ASIC, and developed analytical models to optimize the energy efficiency of dataflow.
- Related publications: [J1], [C2], [C3]

Deep Learning Methods in Wireless Communication Systems — **Southeast University**
Research Assistant, advised by Prof. ***Chuan Zhang*** *and Prof.* ***Yair Be'ery*** — *Jun. 2017 - Mar. 2020*
- Applied gradient descent optimizations of deep learning to enhance the error-correction performance of decoder for polar codes and MIMO detector.
- Exploited convolutional neural networks to realize channel equalization for the cancellation of *intersymbol interference* (*ISI*) and non-linear distortion.
- Reduced complexity of *expectation propagation* (*EP*) MIMO detection for massive antenna arrays by exploiting approximate matrix inversion methods.
- Designed VLSI architectures with high throughput and low latency for MIMO detector and polar decoder, and implemented them on ASIC.
- Related publications: [J2], [J3], [J4], [C1], [C4], [C5], [C6]

PROJECT & INTERNSHIP

Intel Labs **Beijing, China**
*Research Intern, advised by **Sunny Zhang*** *Jun. 2019 - Oct. 2019*

- Developed flexible MIMO processor supporting various detection algorithms.
  - Designed fully pipelined arithmetic modules for *K-best sphere decoding*.
  - Designed systolic array for *minimum mean square error* (*MMSE*) detection.
  - Developed commercial IP core to automatically generate Verilog code for Intel Quartus FPGA.
  - Conducted simulations and experiments on 5G testbed.

Project: Neural Network based Wireless Vision Detection System **Sapporo, Japan**
Team Mentor *May 2019*

- Designed edge computing systems to realize real-time computer vision applications.
  - Implemented dual-camera sampling and H.264 encoder on FPGA.
  - Implemented 2×2 MIMO transceivers to improve transmit rate.
  - Fine-grained parallelism and multi-thread optimization on GPU.
- Project participated in *2019 IEEE Circuits and Systems Society Student Design Competition*.
  - Won the **1st place** in Asia and Pacific region, and was among the **top 4** teams from worldwide.
  - Link: *https://ieee-cas.org/2018-2019-cass-student-design-competition-world-and-regional-winners*

PUBLICATIONS Google Scholar | Citations: 205 | h-index: 8

Journal

[J1] **Weihong Xu**, Zaichen Zhang, Xiaohu You, and Chuan Zhang. "Reconfigurable and low-complexity accelerator for convolutional and generative networks over finite fields". *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (*TCAD*), 2020.

[J2] **Weihong Xu**, Xiaosi Tan, Yair Be'ery, Zaichen Zhang, Xiaohu You, and Chuan Zhang. "Deep learning-aided belief propagation decoder for polar codes". *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* (*JETCAS*), 2020.

[J3] Xiaosi Tan, **Weihong Xu**, Yair Be'ery, Zaichen Zhang, Xiaohu You, and Chuan Zhang. "Improving massive MIMO message passing detectors with deep neural network". *IEEE Transactions on Vehicular Technology* (*TVT*), 2019.

[J4] Xiaosi Tan, **Weihong Xu**, Yaping Zhang, Xiaohu You, and Chuan Zhang. "Efficient expectation propagation massive MIMO detector with Neumann-series approximation". *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2019.

Conference

[C1] **Weihong Xu**, Zhizhen Wu, Yeong-Luh Ueng, Xiaohu You, and Chuan Zhang. "Improved polar decoder based on deep learning". *IEEE International Workshop on Signal Processing Systems* (*SiPS*), Lorient, France, Oct. 2017.

[C2] **Weihong Xu**, Xiaohu You, and Chuan Zhang. "Using Fermat number transform to accelerate convolutional neural network". *IEEE International Conference on ASIC* (*ASICON*), Guiyang, China, Oct. 2017.

[C3] **Weihong Xu**, Zaichen Zhang, Xiaohu You, and Chuan Zhang. "Efficient deep convolutional neural networks accelerator without multiplication and retraining". *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Calgary, AB, Canada, Apr. 2018.

[C4] **Weihong Xu**, Zhiwei Zhong, Yair Be'ery, Xiaohu You, and Chuan Zhang. "Joint neural network equalizer and decoder". *International Symposium on Wireless Communication Systems* (*ISWCS*), Lisbon, Portugal, Sept. 2018.

[C5] **Weihong Xu**, Xiaohu You, Chuan Zhang, and Yair Be'ery. "Polar decoding on sparse graphs with deep learning". *The 52nd Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, California, USA, Oct. 2018.

[C6] **Weihong Xu**, Xiaosi Tan, Xiaohu You, Chuan Zhang, and Yair Be'ery. "On the efficient design of neural networks in communication systems". *The 53rd Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, California, USA, Nov. 2019.

[C7] Minxuan Zhou, Yunhui Guo, **Weihong Xu**, Bin Li, Kevin W. Eliceiri, and Tajana Šimunić Rosing. "MAT: Processing In-Memory Acceleration for Long-Sequence Attention". submitted to *IEEE/ACM Design Automation Conference* (*DAC*), 2021.

[C8] Xiaofan Yu, **Weihong Xu**, Ludmila Cherkasova, and Tajana Šimunić Rosing. "Automating Reliable and Fault-Tolerant Design of LoRa-based IoT Networks". submitted to *IEEE International Conference on Distributed Computing Systems*, 2021.

AWARDS & ACHIEVEMENTS

- Fellowship Stipend of UC San Diego, 57918 USD — Oct. 2020
- Outstanding Master Graduate of Southeast University — Jun. 2020
- Travel Grant of IEEE Circuits and Systems Society for Student Design Competition — *May 2019*
- Graduate Scholarship in SEU (Top 3% students) — *Oct. 2018*
- Outstanding Bachelor Thesis Award in SEU (Top 3% students) — *Jun. 2017*
- Second Prize of National Undergraduate Electronic Design Competition — *Aug. 2016*
- Honorable Mention in Mathematical Contest in Modeling — *2015*

SKILLS & SERVICES

- **Independent Journal Reviewer**
  - IEEE Transactions on Signal Processing
  - IEEE Transactions on Cognitive Communications and Networking
- **Programming Languages and Skills**
  - Python, Tensorflow and Pytorch: Simulated and verified error-correction performance of deep learning-aided polar decoder and channel equalizer.
  - C++ and CUDA: Developed belief propagation decoder for polar codes and optimized CNN inference on NVIDIA GPU.
  - Verilog HDL: Implemented polar decoder, massive MIMO detector and CNN accelerator in publication papers and evaluated their performance on FPGA and ASIC platforms.

REFERENCES

Tajana Simunic Rosing
Professor
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA, USA
✉ tajana@ucsd.edu

Chuan Zhang
Professor
National Mobile Communications Research Laboratory
Southeast University
Nanjing, China
✉ chzhang@seu.edu.cn

Yair Be'ery
Professor
Department of Electrical Engineering
Tel Aviv University
Ramat Aviv, Israel
✉ ybeery@eng.tau.ac.il

Sunny Zhang
Director
Communication Computing Lab
Intel Labs China
Beijing, China
✉ sunny.zhang@intel.com