Wentao Hou

PhD Student · University of Wisconsin-Madison

1210 W. Dayton Street, Madison, WI 53706

taoh@cs.wisc.edu | ThomasH1881.github.io | github.com/ThomasH1881 | linkedin.com/in/wentao-hou-55825a220

Education

University of Wisconsin-Madison — *Madison, United States*

PhD in Computer Sciences — *Sep 2022 - present*

- Advisor: Ming Liu

Tsinghua University — *Beijing, China*

BE in Electronic Engineering — *Sep 2018 - Jun 2022*

Publications

Xincheng Xie, **Wentao Hou**, Zerui Guo, and Ming Liu, "Building Massive MIMO Baseband Processing on a Single-Node Supercomputer", in USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2025

Wentao Hou, Jie Zhang, Zeke Wang, and Ming Liu, "Understanding Routable PCIe Performance for Composable Infrastructures", in USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2024

Kai Zhong, Shulin Zeng, **Wentao Hou**, Guohao Dai, Zhenhua Zhu, Xuecang Zhang, Shihai Xiao, Huazhong Yang, Yu Wang, "CoGNN: An Algorithm-Hardware Co-Design Approach to Accelerate GNN Inference with Mini-Batch Sampling", in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 4883-4896, 42,12, IEEE, 2023

Wentao Hou*, Kai Zhong*, Shulin Zeng, Guohao Dai, Huazhong Yang, Yu Wang, "NTGAT: A Graph Attention Network Accelerator with Runtime Node Tailoring", in Asia and South Pacific Design Automation Conference (ASP-DAC), 2023

Awards and Honors

2023 **Honor Summer RAShip for first-year graduate student**, University of Wisconsin-Madison

2020 **Scholarship for excellence in the academy (top 30%)**, Tsinghua University

Dec. 2018 **35th National College Physics Competition (group of non-physics major), First prize**, Beijing Institute of Physics

Research Experience

Tsinghua University - Department of Electronic Engineering — *Beijing, China*

Advisor: Prof. Yu Wang — *Oct 2021 - Jul 2022*

- Proposed a runtime node tailoring algorithm based on attention coefficients sorting to accelerate graph attention network.
- Proposed a hardware-efficient pipeline insertion sorting scheme for fast node tailoring.
- Designed an accelerator architecture and dedicated processing units for Graph Attention Convolution.

University of Virginia - Department of Computer Science — *Remote*

Advisor: Prof. Samira Khan — *Jun 2021 - Oct 2021*

- Conducted a breakdown performance evaluation of pre-processing in DNN with an image dataset.
- Measured the overheads of different pre-processing steps and looked for bottlenecks in certain scenarios.
- Profiled overheads of differential privacy in machine learning.

Tsinghua University - Department of Electronic Engineering *Beijing, China*

Advisor: Prof. Yu Wang *Dec 2020 - Oct 2021*

- Worked on a software and hardware co-designed GNN accelerator which optimizes loading scattered features.
- Wrote several modules of the GNN and implemented a data path between host, device and on-chip memory via openCL and AXI channel in Xilinx SDx environment. Measured the bandwidth and delay of reading features over different granularities.
- Found the bandwidth saturates at 1KB per addressing in continuous memory access.

Tsinghua University - Department of Electronic Engineering *Beijing, China*

Advisor: Prof. Yongpan Liu *Jun 2021 - Jul 2021*

- Simulated a CNN accelerator with gem5. Simulated sparse acceleration by rounding small weights and skipping all-zero weight groups.
- Simulated a binary neural network on RRAM array by modifying with gem5. Simulated the effects of random noise in RRAM, and measured the relation between noise amplitude and accuracy. Designed an algorithm to reuse weights on different layers when on-chip memory is enough.