YINTAO HE

(+86)-188-1051-1960 ⋄ heyintao19z@ict.ac.cn ⋄ https://yintao-he.github.io/

EDUCATION

University of Chinese Academy of Sciences Sep. 2019 - Present
Ph.D. Student in Computer System and Architecture *Beijing, CN*

- Advisor: Prof. Huawei Li and Prof. Ying Wang
- Research Interest: Energy-Efficient Accelerators, In-Memory Computing, Deep Learning

Nankai University Sep. 2015 - June 2019
Bachelor of Electronic Science and Technology *Tianjin, CN*

PUBLICATIONS

- **Yintao He**, Songyun Qu, Ying Wang, Bing Li, Huawei Li, Xiaowei Li, "InfoX: An Energy-Efficient ReRAM Accelerator Design with Information-Lossless Low-Bit ADCs," in **DAC**, 2022.
- **Yintao He**, Ying Wang, Huawei Li, Xiaowei Li, "Saving Energy of RRAM-based Neural Accelerator through State-Aware Computing," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**).
- **Yintao He**, Ying Wang, Cheng Liu, Huawei Li, Xiaowei Li, "TARe: Task-Adaptive in-situ ReRAM Computing for Graph Learning," in **DAC**, 2021.
- **Yintao He**, Ying Wang, Xiandong Zhao, Huawei Li, Xiaowei Li, "Towards State-Aware Computation in ReRAM Neural Networks," in **DAC**, 2020.
- **Yintao He**, Ying Wang, Yongchen Wang, Huawei Li, Xiaowei Li, "An Agile Precision-Tunable CNN Accelerator based on ReRAM," in **ICCAD**, 2019.

RESEARCH EXPERIENCE

Task-adaptive in-situ ReRAM Computing for Graph Learning Oct. 2020 - Present

- We propose a ReRAM-based in-situ GL accelerator architecture. It supports both weight-static and data-static execution modes, and enables adaptive selection of in-situ computing modes for energy-efficient processing of various GL tasks. And we propose a novel sparse graph data mapping mechanism on ReRAM. This work is accepted by DAC 2021.
- Furthermore, for ultra-large graph learning tasks that exceed the available ReRAM capacity, we propose a ReRAM-based write-free processing engine to avoid expensive ReRAM writing. The extended work will be submitted to TC.

Energy-Efficient ReRAM Accelerator Design with Low-Bit ADCs Sep. 2021 - Dec. 2021

- We propose a ReRAM-based information-aware accelerator architecture. With the XB-wise ADC precision assignments and the hardware architecture, we reduce the ADC precision to 2.92-bit on average without accuracy loss on ResNet-18. This work is accepted by DAC 2022.

Towards State-Aware Computation based on ReRAM Jul. 2019 - Dec. 2020

- We propose a state-aware ReRAM accelerator architecture for neural networks, which could effectively reduce the number of ReRAM cells in a high-power low resistance state (LRS) without affecting the computation results. And for ReRAM-based BNNs, we propose a state-aware model training method, so that fewer weights are represented by LRS cells. These works are accepted by DAC 2020, IEEE TCAD.

A Fast Precision Tuning CNN Accelerator based on ReRAM Nov. 2018 - Apr. 2019

- We devise a novel NN training algorithm to train a mixed-precision neural network, which requires only one single set of parameters to work adaptively in different precision modes.
- We propose a ReRAM-based approximate accelerator design for the precision-tunable neural network, which could achieve on-line tradeoff of computation efficiency and accuracy as the system demands. These works are accepted by ICCAD 2019, IEEE TCAD.