Haiyu Mao

Homepage: https://hybol1993.github.io/
Phone: +86 135-5273-6012
Email: maohaiyu1993@gmail.com

OVERVIEW I am a 5th-year Ph.D. candidate at Tsinghua University. My research interests primarily include emerging Non-Volatile Memories (e.g. Racetrack Memory, Phase Change Memory, and Resistive Random-Access Memory) and Processing In Memory (PIM). I focus on solving technical problems relating to 1) integrating NVMs into traditional memory hierarchy, 2) designing efficient PIM, taking advantages of the characteristics of NVMs, 3) managing the data in PIM to match its computational ability, and 4) protecting NVMs from security hazards when they are applied in both memory system and PIM.

EDUCATIONAL BACKGROUND

Tsinghua University Beijing, China
Ph.D. of Computer Science and Technology; Aug. 2015 – July. 2020 (expected)
Advisor: Jiwu Shu; GPA: 3.7/4.0 (Rank:24/95)

Northeastern University Shenyang, China
Bachelor of Software Engineering; Aug. 2011 – July. 2015
GPA: 90/100 (**Rank: 1/201**)

PUBLICATIONS

Haiyu Mao, Jiwu Shu, Jiaqi Zhang, Fan Yang, Tao Li, and Youyou Lu, "Title is not shown for the double-blind peer review", under submission to *The International Symposium on Computer Architecture* (**ISCA**), Valencia, Spain, June 2020.

Haiyu Mao, Jiwu Shu, Mingcong Song, and Tao Li, "LerGAN: A Compact PIM-based GAN Architecture with Low Energy Consumption", under submission to *IEEE Transactions on Computers* (**TC**), 2020.

Fan Yang, Youmin Chen, **Haiyu Mao**, Youyou Lu, and Jiwu Shu, "Libra: An Efficient and Fast Recoverable System for Secure Non-Volatile Memory", under submission to *ACM Transactions on Storage* (**TOS**), 2020.

Haiyu Mao, and Jiwu Shu, "3D Memristor Array Based Neural Network Processing in Memory Architecture", *in Journal of Computer Research and Development,* (In Chinese), 2019.

Fan Yang, Youyou Lu, Youmin Chen, **Haiyu Mao**, and Jiwu Shu, "No Compromises: Secure NVM with Crash Consistency, Write-Efficiency and High-Performance", *in Design Automation Conference* (**DAC**), Las Vegas, NV, June 2019.

Haiyu Mao, Mingcong Song, Tao Li, Yuting Dai, and Jiwu Shu, "LerGAN: A Zero-Free, Low Data Movement and PIM-Based GAN Architecture", *in International Symposium on Microarchitecture* (**MICRO**), Fukuoka, Japan, October 2018.

Haiyu Mao, Xian Zhang, Guangyu Sun, and Jiwu Shu, "Protect Non-Volatile Memory from Wear-Out Attack Based on Timing Difference of Row Buffer Hit/Miss", *in Conference on Design, Automation & Test in Europe* (**DATE**), Lausanne, Switzerland, March 2017.

Haiyu Mao, Chao Zhang, Guangyu Sun, and Jiwu Shu, "Exploring Data Placement in Racetrack Memory Based Scratchpad Memory", *in Non-Volatile Memory System and Applications Symposium* (**NVMSA**), Hong Kong, China, August 2015.

MAJOR PROJECTS

Data Integrity Protection of NVM-based PIM 05/2019-Present

Key idea: Ensure data integrity in the NVM-based PIM for neural network applications.

- Propose a data integrity attack to the NVM-based PIM for neural networks so as to destroy the result of inference or training progress.
- Propose a countermeasure to defend data integrity attack by isolating the verification of data integrity and the computation in NVM-based PIM.
- In order to cooperate with the proposed countermeasure, we modify the micro-architecture of NVM-based PIM for NN to support isolation between verification and computation.

Enhance The Endurance of NVM-based PIM Devices 04/2018-08/2019

Key idea: Utilize the characteristics of both NVM cells and neural networks to prolong the lifetime of PIM device.

- Analyze the write behavior of updating the weight matrix in the NVM-based PIM array when PIM is employed in neural network training.

- In order to simulate how the aging of cells influences the accuracy of neural networks, we modify Caffe framework to monitor the accuracy loss when changing weight values into given values during different training epochs.
- Propose a scheme for long-lived PIM by (a) leveraging the characteristics of NVM cells that 1) stuck-at-fault cells can still be used in analog computing, and 2) old cells can be rejuvenated into young cells through changing the reference of the sense amplifier; (b) combining inherent fault-tolerance characteristic of neural networks and their particular weight updating behaviors.
- According to the evaluation, the proposed scheme achieves 947× lifetime extension of the PIM device, as well as 1.24× speedup and 1.55× energy saving on average.

PIM-based High-performance/Low-power GAN Training 10/2016-04/2018

Key idea: Remove the structured zero insertion and shorten the interconnections when training a GAN in PIM.

- Analyze and then find that the zero-inserting operation incurs serious redundant storage and computation, which can not be solved by traditional compression since the data in PIM are all structured for both storage and computation.
- Observe that the interconnection is a bottleneck when training a GAN in PIM, since long routing paths hinder the performance of PIM.
- Propose a data reshaping scheme that removes inserted zeros, along with a structured data mapping scheme to save both storage capacity and communication bandwidth in PIM.
- Propose a 3D reconfigurable interconnection fabric in PIM to radically shorten the routing paths.
- The software-hardware co-design 3D-ReRAM based PIM achieves 7.46× speedup and 7.68× energy saving compared with the state-of-the-art PIM micro-architecture.

Demystify NVM Wear-out Vulnerability and Low-overhead Countermeasure 09/2015-10/2016

Key idea: Reveal particular information through the difference between row buffer hit and miss.

- According to the read latency difference between row buffer hit and miss, demystify that NVM is vulnerable to indirect information leakage about data location through side channels.
- Conduct an effective wear-out attack on physical data location, even though NVM is protected by the state-of-the-art wear-leveling scheme.
- Propose a countermeasure which prolongs the lifetime of NVM compared with the state-of-the-art wear-leveling scheme, while only introducing trivial hardware overhead.
- The proposed attack manages to wear out the PCM in 137 seconds and the corresponding countermeasure lengthen the lifetime of PCM to 4000 days.

Treat Racetrack Memory as A On-chip Scratchpad Memory 02/2015-08/2015

Key idea: Explore better data placement to minimize the movement of the read/write ports of Racetrack Memory.

- Characterize that the read/write ports consume most of the access time and energy when Racetrack Memory is used as on-chip cache.
- Propose a Scratchpad Memory based data placement scheme to reduce the movement of read/write ports in Racetrack Memory by leveraging the genetic algorithm.
- Optimize the data placement scheme by providing the initial genes assisted by *Fisrt Come First Store scheme*, *Most Access in The Middle scheme*, and *Most Access in The Front scheme*.

SELECTED AWARDS

National Scholarship for Ph.D. **(2.5%)** — Ministry of National Education of China, 2019
MICRO-51 Student Travel Award — ACM SIGMICRO, 2018
Second-Class Comprehensive Scholarship — Tsinghua University, 2017
Guanghua Scholarship — Tsinghua University, 2016
Scholarship Funded by The Mayor of The City of Shenyang **(Top 6)** — Mayor of Shenyang, 2015
Top 10 Excellent Undergraduates **(Top 10)** — Northeastern University, 2014
Outstanding Undergraduate in the City of Shenyang **(0.26%)** — Shenyang, 2014
Outstanding Pioneer Student **(0.5%, three times)** — Northeastern University, 2012/2013/2014
National Scholarship **(1%, three times)** — Ministry of National Education of China, 2012/2013/2014

TECHNICAL SKILLS

Programming Languages Proficient: C, C++, Java, Python
Used: Javascript, Go, MPI, OpenMP, CUDA, Matlab
Frameworks: Caffe, TensorFlow, PyTorch, Hadoop, Spark
Simulator: Gem5, DRAMSim, NVSim, NVmain, CACTI