BING LI

Email: bing.li.ece@duke.edu | Tel: (919) 353-4154

EXPERIENCE

Sep. 2017- present **Postdoctoral Scholar,**
Research interests: Accelerator design for Machine Learning Applications
Department of Electrical and Computer Engineering, Duke University

Jul. 2016 - Jun.2017 **Research Engineer in Cloud Computing,**
China Unicom Research Institute

EDUCATION

Sep. 2010- Jun. 2016 **PhD** in Computer Systems and Architecture,
University of Chinese Academy of Sciences, Institute of Computing Technology
Supervisor: Prof. Yu Hu
Thesis: "Optimization of Performance and Reliability for the DRAM+PCM Hybrid Memory System"

Sep. 2006- Jun. 2010 **B.Eng.** in Computer Technology and Science **(Rank: 1/100)**
Minzu University of China

PUBLICATIONS

[1] **Bing Li,** Mengjie Mao, Xiaoxiao Liu, Tao Liu, Zihao Liu, Wujie Wen, Yiran Chen and Hai Li. Thread Batching for High-performance Energy-efficient GPU Memory Design, submitted to ACM Journal on Emerging Technologies in Computing Systems. (Under review)

[2] **Bing Li,** Bonan Yan, Chenchen Liu, Hai Helen Li. Build Reliable and Efficient Neuromorphic Design with Memristor Technology, in 24th Asia and South Pacific Design Automation Conference (ASP-DAC), 2019. (accepted)

[3] Zichen Fan, Ziru Li, **Bing Li**, Hai Li. RED: A ReRAM-based Deconvolution Accelerator, in Design, Automation & Test in Europe Conference & Exhibition (DATE), 2019 (accepted, corresponding author)

[4] Biresh Kumar Joardar, **Bing Li,** Janardhan Rao Doppa, Hai Li, Partha Pratim Pande, Krishnendu Chakrabarty. REGENT: A Heterogeneous ReRAM/GPU-based Architecture Enabled by NoC for Training CNNs. DATE 2019. (accepted)

[5] Qingli Guo, Ye, Jing, **Bing Li,** Yu Hu, Xiaowei Li, Yazhu Lan, Guohe, Zhang. (2018). PUFPass: A password management mechanism based on software/hardware codesign. Integration. 10.1016/j.vlsi.2018.10.003.

[6] **Bing Li**, Fan Chen, Wang Kang, Weisheng Zhao, Yiran Chen and Hai Helen Li. Design and Data Management for Magnetic Racetrack Memory, in 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 2018, pp 1-4.

[7] **Bing Li,** Linghao Song, Fan Chen, Xuehai Qian, Yiran Chen, Hai Helen Li. ReRAM-based accelerator for deep learning, " in Design, Automation & Test in Europe Conference & Exhibition (DATE), 2018, pp 815-820.

[8] **Bing Li,** Wei Wen, Jiachen Mao, Sicheng Li, Yiran Chen, Hai Helen Li. Running sparse and low-precision neural network: When algorithm meets hardware, in Asia and South Pacific Design Automation Conference (ASP-DAC), 2018, pp 534-539.

[9] **Bing Li,** Yunyong Zhang, Xu Lei. "An MEC and NFV integrated network architecture." ZTE Communications 15, no. 2 (2017): 1.

[10] **Bing Li,** Yu Hu, Ying Wang, Jing Ye, and Xiaowei Li. Power-Utility-Driven Write Management for MLC PCM. ACM Journal on Emerging Technologies in Computing Systems (JETC) 13.3 (2017): 50.

[11] **Bing Li,** ShuChang Shan, Yu Hu, Xiaowei Li, A Dynamic Adjustment Design for Hybrid Fault Tolerant Code in Memory System (in Chinese), in Journal of Computer-Aided Design & Computer Graphics (JCAD), Volume 26 Issue 9, September 2014.

[12] **Bing Li**, Yu Hu, Xiaowei Li. Short-SET: An energy-efficient write scheme for MLC PCM, in IEEE Non-Volatile Memory Systems and Applications Symposium (NVMSA), 2014, pp. 1-6.

[13] **Bing Li,** ShuChang Shan, Yu Hu, Xiaowei Li. Partial-SET: Write speedup of PCM main memory," in Design, Automation & Test in Europe Conference & Exhibition (DATE), 2014, pp. 1-4.

[14] **Bing Li**, ShuChang Shan, Yu Hu, Xiaowei Li. Tolerating Noise in MLC PCM with Multi-Bit Error Correction Code, in IEEE 19th Pacific Rim International Symposium on Dependable Computing (PRDC), 2013.

TALKS AND PRESENTATIONS

Jul., 2018 Talk at Beihang University, Beijing, China, Hosted by Prof. Yuanqing Chen

May, 2018 Poster Presentation in WISE workshop, IEEE International Symposium on Hardware Oriented Security and Trust (HOST), 2018, Washington DC, US

Jan., 2016 Poster Presentation in ASPDAC-18, Macao SAR, China

Aug, 2014 Talk in NVMSA, Chongqing, China

Dec., 2013 Talk in PRDC, Vancouver, Canada

PATENTS

[1] **Bing Li,** ShuChang Shan, Yu Hu, Xiaowei Li. Writing acceleration method and system for phase change memory. (CN201410064466)

[2] **Bing Li,** Wei Xiong, Lei Xu, Zhijun Wang. Method, device and system for evaluating telecommunication operation process. (CN201710153426)

[3] **Bing Li,** ShuChang Shan, Xiang Gao, Yu Hu. Data storage method and device. (CN 104182292)

[4] Xiang Gao, **Bing Li,** ShuChang Shan, Yu Hu. A messaging type memory module memory access method and apparatus. (CN 104347122)

[5] ShuChang Shan, **Bing Li**, Yu Hu, Xiang Gao. Memory access method, equipment and system (CN 104346234)

[6] Xiang Gao, **Bing Li,** ShuChang Shan, Yu Hu. Memory access method and apparatus for message-type memory module. (US Patent 9,811,416)

[7] ShuChang Shan, **Bing Li**, Yu Hu, Xiang Gao. Memory access method, device, and system (US Patent 9,772,891)

PROJECTS

Apr.2018–Present **Exploration of 3D-based Heterogeneous Architecture for Deep Neural Networks Training**

- Proposed a novel architecture, 3D-ReG, which utilize 3D integration to vertically stack ReRAM-based processing-in-memory with GPU for efficient training DNNs.
- Proposed two different pipelined task-mapping schemes which allocate the executions to different processing units during the DNN training to maximizes the trade-off the computing efficient and resource utilization.

Jun.2018–Present **Exploration of Automatic Mixed-Precision Search for Deep Neural Networks**

- Proposed to relax the search space of quantization bit-width from discrete to continuous domain.
- Proposed a gradient descent-based search algorithm to automate the process of finding the optimal mixed-precision quantization design.
- The network model generated by this method remains the similar accuracy to the full-precision model while its compression rate is close to the binary counterpart.

Sep.2017–Present **ReRAM-based Processing-in-Memory to accelerate Convolutional Neural Network**

- Proposed the pipelined execution dataflow to accelerate the convolutional neural network.
- Proposed the high parallel execution architecture to exploit the inherent parallelism of neural network.
- Proposed the mapping scheme to implement the convolutional computation on the ReRAM-based execution engine.
- Proposed the ReRAM-based deconvolution accelerator—RED, orthogonally combines *the pixel-wise mapping scheme* and *zero-skipping data flow* to accelerate the deconvolution computation.

Jun.2016–Jun. 2017 **Container-based Cloud Platform to Support**

- Developed a container-based Cloud system in a physical cluster.
- Implemented one private image registry, which was serving the users in local area network.
- Designed customized images that provide services including website, online SNS and individual blog.
- Completed one review report (in Chinese) which concluded the service discovery of the prevalent open-source Cloud architectures--Mesos, Kubernetes and Docker swarm.

Jun.2016–Jun. 2017 **WoCloud Interoperability Test**

- The Wocloud bases on the Openstack Liberty, has been in production for 2 years, provides ~200 compute nodes and ran ~1K compute instances.
- Completed the WoCloud interoperability test using Refstack toolset.
- Completed the test report with >2k words (in Chinese), and detailed introduced the test toolset, the test environment, the premise and the test results.
- We ran 83 official-certificated tests, and the pass rate reached 91.6%.

Mar. 2014–Jun. 2016 **Performance and Power Optimization for PCM System**

- Investigated the PCM write scheduling issues induces by both the memory/bank-level parallelism restriction and the power limitation.
- Designed a power-aware scheduling algorithm to improve the performance of PCM-based system by 24%.
- Designed an ultra-fine-grained power management scheme to improve the throughput of PCM by 27%.
- Verified our proposals on the full-system simulator *MarssX86+DRAMsim2*. (Published in JETC)

Jun.2013-Mar.2014 **Write Acceleration for PCM**

- Proposed and implemented the fast write policies to improve the PCM memory access performance by 47% and 91% respectively for the SLC and MLC. (Published in NVMSA 2014 and DATE 2014)
- Proposed the full-programing schemes to preserve the reliability for PCM.
- Verified our proposals on memory simulator *DRAMsim2*.

Sep.2011–May 2013 **Reliability Improvement for PCM System**

- Designed and implemented a multi-bit error correction code for MLC PCM which reduced the system's error rate by 16 orders of magnitude error rate with less than 1% performance overhead. (Published in PRDC 2013)
- Proposed a lifetime model to estimate the failure rate of MLC PCM induced by noise.
- Evaluated the reliability improvement and the overhead with modified *Msim*.

Mar.2011– Mar. 2013 **Hybrid Fault-tolerance Codes Tuner for Main Memory System**

- Designed a dynamic adjustment scheme for hybrid fault-tolerant code in memory system, which alleviated the performance overhead by 53%. (Published in JCAD 2014)
- Modified the model and simulated a main memory system with DRAMsim2.
- Realized the hardware design with Design Compiler and evaluated the hardware overhead.

AWARDS

2018 National Academy of Science Associate Fellowship Award, US

2014 Langchao Scholarship, University of Chinese Academy Sciences, China

2013, 2014 Outstanding Student Award, State Key Laboratory of Computer Architecture, China

2011 Excellent Student Award, University of Chinese Academy Sciences, China

2010 Outstanding Graduate Award in the top universities of Beijing, Beijing Municipal Bureau of Education, China

2009 National Scholarship, Ministry of Education (MOE) of China **(13/700)**

2008 National Motivational Scholarship, The Ministry of Education (MOE) of China

2007-2009 Top-level Student Scholarship, Minzu University of China **(awarded for consecutive three years)**

SERVICES

Session Chairs

2018 23th ASP-DAC Session Chair

2018 32nd ACM International Conference on Supercomputing Financial Chair

Reviewers

- IEEE Transactions on Multi-Scale Computing Systems
- IEEE Transactions on Embedded Computing Systems
- IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems
- IEEE Transactions on Very Large-Scale Integration Systems
- IEEE Journal on Emerging and Selected Topics in Circuits and Systems
- Elsevier, Integration, the VLSI Journal
- Elsevier, Microelectronics Journal
- Elsevier, Neurocomputing
- Army Research Office (ARO) Young Investigator Program
- International Conference on Embedded Software (EMSOFT)
- Design, Automation and Test in Europe Conference

SKILLS

Systems CPU Microarchitecture, Memory System

Programming C, C++, Python

Tools GDB, nvprof, NVSim, GPGPUsim, Marssx86, DRAMsim simulator, TensorFlow