MinXuan Zhou

miz087@eng.ucsd.edu | +1 (858) 228 7085

EDUCATION BACKGROUND

University of California, San Diego (GPA: 3.72/4.0) — **San Diego, CA, US**
Ph.D in Computer Science — Sep. 2018 - present
Master of Science in Computer Science — Jun. 2017
Beihang University (GPA: 3.5/4.0) — **Beijing, China**
Bachelor of Engineering in Computer Science and Technology — Jul. 2015

AWARDS

- Excellent undergraduate thesis in school of computer science and engineering — 2015
- RANK #1 scholarship of discipline competition for the 2012-2013 academic year — 2013, 2014
- FIRST PRIZE in LanQiao Cup C/C++ programming contest national final (winner out of 3000+) — 2012

WORK EXPERIENCE

Machine Learning Research Intern (Apple Inc.) — Jun.2021 – Sep.2021
- Compiler optimizations for Apple Neural Engine

Ph.D. Software Engineering Intern (Facebook Inc.) — Jun.2020 – Sep.2020
- AI system infrastructure - software-hardware co-design

Research Intern (Alibaba Group U.S.) — Jun.2019 – Sep.2019
- Software-hardware co-design for accelerating deep neural networks on emerging architectures

Graduate Student Researcher (University of California, San Diego) — Jul. 2018 – present
- System and architecture support for processing in memory technology

Staff Research Associate (University of California, San Diego) — Aug. 2017 – Jun. 2018
- System modeling based thermal & power management for various computing platforms

PUBLICATIONS

- **Minxuan Zhou**, Weihong Xu, Jaeyoung Kang, and Tajana Rosing, "TransPIM: A Memory-based Acceleration via Software-Hardware Co-Design for Transformers", *The 28th IEEE International Symposium on High-Performance Computer Architecture (HPCA'2022), to appear*
- Yizhou Wei, **Minxuan Zhou**, Sihang Liu, Korakit Seemakhupt, Tajana Rosing and Samira Khan. "PIMProf: An Automated Program Profiler for Processing-in-Memory Offloading Decisions", *Design, Automation and Test in Europe Conference (DATE'2022), to appear*
- **Minxuan Zhou**, Lingxi Wu, Muzhou Li, Niema Moshiri, Kevin Skadron, and Tajana Rosing, "Ultra Efficient Acceleration for De Novo Genome Assembly via Near-Memory Computing", *International Conference on Parallel Architectures and Compilation Techniques (PACT), 2021*
- **Minxuan Zhou**, Guoyang Chen, Mohsen Imani, Saransh Gupta, Weifeng Zhang, and Tajana Rosing, "PIM-DL: Boosting DNN Inference on Digital Processing In-Memory Architectures via Data Layout Optimizations", *International Conference on Parallel Architectures and Compilation Techniques (PACT), 2021*
- **Minxuan Zhou**, Yunhui Guo, Weihong Xu, Bin Li, Kevin Eliceiri, and Tajana Rosing, "MAT: Processing In-Memory Acceleration for Long-Sequence Attention", *Design Automation Conference (DAC), 2021*
- **Minxuan Zhou**, Muzhou Li, Mohsen Imani, and Tajana Rosing, "HyGraph: Accelerating Graph Processing with Hybrid Memory-centric Computing", *Design, Automation and Test in Europe Conference (DATE), 2021*
- **Minxuan Zhou**, Mohsen Imani, Yeseong Kim, Saransh Gupta, and Tajana Rosing, "DPSim: A Full-stack Simulation Infrastructure for Digital Processing In-Memory Architecture", *26th Asia and South Pacific Design Automation Conference (ASP-DAC), 2021*
- **Minxuan Zhou**, Mohsen Imani, Saransh Gupta, and Tajana Rosing, "Thermal-Aware Design and Management for Search-based In-Memory Acceleration", *SRC TECHCON, 2019*
- **Minxuan Zhou,** Andreas Prodromou, Rui Wang, Hailong Yang, Depei Qian, Dean Tullsen. "Temperature-Aware DRAM Cache Management -Relaxing Thermal Constraints in 3D Systems". *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2019*
- Xiao Liu, **Minxuan Zhou**, Tajana Rosing, and Jishen Zhao. 2019. HR3AM: A Heat Resilient Design for RRAM-based Neuromorphic Computing. *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), 2019*
- Mohsen Imani, Saransh Gupta, Yeseong Kim, **Minxuan Zhou**, and Tajana Rosing. DigitalPIM: Digital-based

Processing In-Memory for Big Data Acceleration. *ACM Proceedings of the 2019 on Great Lakes Symposium on VLSI*

- **Minxuan Zhou**, Mohsen Imani, Saransh Gupta, and Tajana Rosing, "Thermal-Aware Design and Management for Search-based In-Memory Acceleration", *Design Automation Conference (DAC), 2019.*
- **Minxuan Zhou**, Mohsen Imani, Saransh Gupta, Yeseong Kim, and Tajana Rosing, "GRAM: Graph Processing in a ReRAM-based Computational Memory", 24th Asia and South Pacific Design Automation Conference (ASP-DAC), 2019
- **Minxuan Zhou**, Mohsen Imani, Saransh Gupta, Yeseong Kim, and Tajana Rosing, "GP^3: Graph Processing in a Parallel Processing-in-Memory Architecture", *SRC TECHCON, 2018*
- **Minxuan Zhou**, Mohsen Imani, Saransh Gupta, and Tajana Rosing, "GAS: A Heterogeneous Memory Acceleration for Graph Processing", *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), 2018.*
- Cheng, Kun, Yuebin Bai, Yongwang Zhao, Yao Ma, Duo Lu, Yuanfeng Peng, and **Minxuan Zhou**. "HV 2 M: A novel approach to boost inter-VM network performance for Xen-based HVMs." *Journal of Systems and Software* 114 (2016): 54-68.

RESEARCH PROJECTS

Acceleration of emerging applications on processing in-memory (PIM) architectures **C++**
University of California, San Diego Dec. 2018– present

- Proposed a new processing paradigm on digital PIM architectures to solving the long-sequence issues in emerging attention-based machine learning models
- Designed compiler-level and architectural optimizations of data layout for various DNNs running on emerging processing in-memory architectures
- Software-hardware co-design which accelerates bio-informatics applications.
- Proposed an innovative programming model and ReRAM-based digital PIM architecture which accelerates graph processing by orders of magnitude over conventional systems

System and architecture support for processing in-memory (PIM) architectures **C++**
University of California, San Diego Dec. 2018 – present

- Developing a profiling and automatic offloading toolchain to analyze and accelerate general programs using PIM technology
- Developed a simulation infrastructure which simulates general applications on the PIM architecture with software-hardware co-design

System modeling based thermal/power management for mobile devices **C/Python**
University of California, San Diego Feb. 2018 – Jan. 2019

- Investigated innovative thermal/power management policies in state-of-the-art Android devices based on accurate system models in terms of performance, power, and temperature
- The error rates of power and thermal models are within 8.6% and 2.5% respectively
- The implemented policy achieved 10.7% reduction in power consumption without a performance degradation

Thermal Management in 3D-Stacked DRAM Cache **C++/C/Python**
University of California, San Diego Jul. 2016 – Dec. 2017

- Implemented a simulation infrastructure based on Sniper and Ramulator for simulating multi-core systems with large 3D-stacked DRAM caches
- Proposed and implemented three cache management mechanisms for 3D DRAM cache which improve the performance of 3D systems by 26.1% in average

Hypervisor based, and Micro-Kernel based Embedded Virtualization **C**
Beihang University May. 2014 – Jul. 2015

- Implement *memory ballooning* in OKL4 micro-kernel to optimize the process memory management
- Add system-thread based, and system-call based checkpoint-restart mechanisms in seL4 micro-kernel

SKILLS

- Research interests: computer architecture, memory system, domain-specific acceleration, software-hardware co-design for emerging applications, thermal and power management
- Programming languages (advanced): C, C++, Python

- Open source tools/projects experience: Intel Pin-tool, PyTorch Glow compiler, Linux kernel, LLVM, L4 microkernels, Android Open Source Project
- Solid background in algorithms and data structures (1st prize in National Olympiad in Informatics in Provinces)