Peng Gu

Software Engineer, System Infrastructure, Google
Email: gupeng@google.com
Personal Website: miglopst.github.io

Professional Experience

2021 - now **Google System Infrastructure Team** **Sunnyvale, CA**
Software Engineer for Data Center Performance Optimization

2015 - 2021 **University of California, Santa Barbara** **Santa Barbara, CA**
Student Researcher in Scalable Energy-efficient Architecture (SEAL) Laboratory

Summer 2019, 2018, 2017 **Samsung Semiconductor** **San Jose, CA**
Memory System Accelerator Architecture Research Internship in Memory Solution Lab

07-09/2016 **Hewlett Packard Labs** **Palo Alto, CA**
Accelerator Architecture Research Internship in Platform Architecture Group

07-09/2014 **University of California, Los Angeles** **Los Angeles, CA**
Student Researcher in Design Automation Laboratory

07-09/2013 **Intel Asia-Pacific Research and Development Center** **Shanghai, P.R.China**
Technical Internship in Mobile Computing Group (MCG)

2013 - 2015 **Tsinghua University** **Beijing, P.R.China**
Student Researcher in Nanoscale Integrated Circuits and Systems (NICS) Laboratory

Education

2017 - 2021 **University of California, Santa Barbara** **Santa Barbara, CA**
Ph.D. Student in Electrical Computer Engineering
Advisor: Yuan Xie

2015 - 2017 **University of California, Santa Barbara** **Santa Barbara, CA**
M.S. in Electrical Computer Engineering
Advisor: Yuan Xie

2011 - 2015 **Tsinghua University** **Beijing, P.R.China**
B.S. in Electronic Engineering
Advisor: Yu Wang

Research Summary

My Ph.D. research focuses on near-data-processing / process-in-memory architecture, memory sub-system, and domain-specific accelerator design. In the past, I also participated several projects related to secure hardware design and cost-driven IC design.

Near-Data-Processing / Process-in-Memory Architecture
Memory-centric architecture has shown a great potential to tackle the "memory wall" challenge of the traditional compute-centric accelerator. From the technology perspective, I explored emerging RRAM technology[J7,6,5][C14,15,16,17], mature DRAM DIMM technology [C3,4], in-situ DRAM computing technology [C6], and 3D-stacking memory technology [J1][C1,2,5][P5,6,7,8,9].

Memory Sub-system Design
I helped build up a circuit-level model to enable evaluation of device/circuit innovations for emerging NVM [C11] as well as Neuromorphic computing systems [C14]. Also, I designed a transaction command based simulator for system architects to evaluate the performance of emerging NVM solutions [J3].

Domain-specific Accelerator Design
Customized computing architecture is becoming a promising approach to improve applications' performance and energy-efficiency. I explored accelerator designs for several emerging data-intensive application domains, including deep learning [J1,4,7,5,6][C14,15,16,17][P6,7,8,9], image processing [C2], graph analytic [C5], and bioinformatic [C3,4].

Cost-driven and Secure Design for 2.5D/3D Technology
2.5D/3D technology enables high-density and heterogeneous integration of multiple dies, thus allowing flexible designs. In this project, I explored (1) thermal-aware design utilizing die-stacking architecture for side-channel prevention [C9,C10]; (3) cost-efficient 3D integration for secure split-manufacturing [C7,C8]; (3) analytical cost model with 3D and interposer-based 2.5D die integration for IP reuse [C12,C13].

Awards and Honors

2020	ACM Student Research Competition First Place (Graduate), MICRO
2016	A. Richard Newton Young Student Fellowship, Design Automation Conference
2015	Holbrook Foundation Fellowship, The Institute for Energy Efficiency, UC Santa Barbara
2015	Excellent Undergraduate Thesis Award, Tsinghua University
2014	Academic Scholarship, Department of Electronic Engineering, Tsinghua University

Academic Service

2017	Web Chair, 24th IEEE International Symposium on High-Performance Computer Architecture (HPCA)

Publications (Google Citation 937, h-index 15)

Journal Publications

[J1]. **Peng Gu**, Xinfeng Xie, Shuangchen Li, Krishna T. Malladi, Dimin Niu, Hongzhong Zheng, Yuan Xie. "DLUX: a LUT-based Near-Bank Accelerator for Data Center Deep Learning Training Workloads." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2020.

[J2]. Xinfeng Xie, Xing Hu, **Peng Gu**, Shuangchen Li, Yu Ji, and Yuan Xie. "NNBench-X: A Benchmarking Methodology for Neural Network Accelerator Designs." *ACM Transactions on Architecture and Code Optimization (TACO)*, 2020.

[J3]. **Peng Gu**, Benjamin Lim, Wenqin Huangfu, Krishna T. Malladi, Andrew Chang, Yuan Xie. "NMTSim: Transaction-Command based Simulator for New Memory Technology Devices." *IEEE Computer Architecture Letters*, 2020.

[J4]. Xinfeng Xie, Xing Hu, **Peng Gu**, Shuangchen Li, Yu Ji, and Yuan Xie. "NNBench-X: Benchmarking and Understanding Neural Network Workloads for Accelerator Designs." *IEEE Computer Architecture Letters*, 2019.

[J5]. Lixue Xia, **Peng Gu**, Boxun Li, Tianqi Tang, Xiling Yin, Wenqin Huangfu, Shimeng Yu, Yu Cao, Yu Wang, Huazhong Yang. "Technological Exploration of RRAM Crossbar Array for Matrix-vector Multiplication." *Journal of Computer Science and Technology (JCST)*, 2016.

[J6]. Boxun Li, **Peng Gu**, Yu Wang, Huazhong Yang. "Exploring the Precision Limitation for RRAM-Based Analog Approximate Computing." *IEEE Design & Test, Volume 33*, 2016.

[J7]. Boxun Li, **Peng Gu**, Yi Shan, Yu Wang, Yiran Chen, Huazhong Yang. "RRAM-based Analog Approximate Computing." *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2015.

Refereed Conference Publications

[C1]. Xinfeng Xie, Zheng Liang, **Peng Gu**, Abanti Basak, Lei Deng, Ling Liang, Xing Hu, Yuan Xie "SpaceA: Accelerating Sparse Matrix Vector Multiplication with Processing-in-Memory Architecture." *International Symposium on High-Performance Computer Architecture (HPCA), 2021*

[C2]. **Peng Gu**, Xinfeng Xie, Yufei Ding, Guoyang Chen, Weifeng Zhang, Dimin Niu, Yuan Xie "iPIM: Programmable In-Memory Image Processing Accelerator Using Near-Bank Architecture." *International Symposium on Computer Architecture (ISCA), 2020*

[C3]. Wenqin Huangfu, Krishna T. Malladi, Shuangchen Li, **Peng Gu**, Yuan Xie "NEST: DIMM based Near-Data-Processing Accelerator for K-mer Counting." *International Conference On Computer Aided Design (ICCAD), 2020*

[C4]. Wenqin Huangfu, Xueqi Li, Shuangchen Li, Xing Hu, **Peng Gu**, Yuan Xie "MEDAL: Scalable DIMM based Near Data Processing Accelerator for DNA Seeding Algorithm." *International Symposium on Microarchitecture (MICRO), 2019*

[C5]. Mingyu Yan, Xing Hu, Shuangchen Li, Abanti Basak, Han Li, Xin Ma, Itir Akgun, Yujing Feng, **Peng Gu**, Lei Deng, Xiaochun Ye, Zhimin Zhang, Dongrui Fan, Yuan Xie "Alleviating Irregularity in Graph Analytics Acceleration: a Hardware/Software Co-Design Approach." *International Symposium on Microarchitecture (MICRO), 2019*

[C6]. Shuangchen Li, Alvin Oliver Glova, Xing Hu, **Peng Gu**, Dimin Niu, Krishna T. Malladi, Hongzhong Zheng, Bob Brennan, Yuan Xie. "SCOPE: A Stochastic Computing Engine for DRAM-based In-situ Accelerator." *International Symposium on Microarchitecture (MICRO), 2018*

[C7]. **Peng Gu**, Dylan Stow, Prashansa Mukim, Shuangchen Li, Yuan Xie. "Cost-efficient 3D Integration to Hinder Reverse Engineering During and After Manufacturing." *Asian Hardware Oriented Security and Trust Symposium (Asian HOST), 2018*

[C8]. Jaya Dofe, **Peng Gu**, Dylan Stow, Qiaoyan Yu, Eren Kursun, Yuan Xie. "Security Threats and Countermeasures in Three-Dimensional Integrated Circuits." *Proceedings of the 27th Great Lakes Symposium on VLSI (GLSVLSI), 2017.*

[C9]. **Peng Gu**, Dylan Stow, Russell Barnes, Eren Kursun, Yuan Xie. "Thermal-aware 3D Design for Side-channel Information Leakage." *Proceedings of the 34th IEEE International Conference on Computer Design (ICCD), 2016.*

[C10]. **Peng Gu**, Shuangchen Li, Dylan Stow, Russell Barnes, Liu Liu, Eren Kursun, Yuan Xie. "Leveraging 3D Integration Technologies to Improve Hardware Security: Opportunities and Challenges." Invited Paper - *Proceedings of the 26th Great Lakes Symposium on VLSI (GLSVLSI), 2016.*

[C11]. Shuangchen Li, Liu Liu, **Peng Gu**, Cong Xu, Yuan Xie. "NVSim-CAM: A Circuit-Level Simulator for Emerging Nonvolatile Memory based Content-Addressable Memory." *Proceedings of the 35th International Conference On Computer Aided Design (ICCAD), 2016.*

[C12]. Dylan Stow, Itir Akgun, Russell Barnes, **Peng Gu**, Yuan Xie. "Cost Analysis and Cost-Driven IP Reuse Methodology for SoC design Based on 2.5D/3D Integration." Invited Paper - *Proceedings of the 35th International Conference On Computer Aided Design (ICCAD), 2016.*

[C13]. Dylan Stow, Itir Akgun, Russell Barnes, **Peng Gu**, Yuan Xie. "Cost and Thermal Analysis of High-Performance 2.5D and 3D Integrated Circuit Design Space." *Proceedings of the IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2016.*

[C14]. Lixue Xia, Boxun Li, Tianqi Tang, **Peng Gu**, Xiling Yin, Wenqin Huangfu, Pai-yu Chen, Shimeng Yu, Yu Cao, Yu Wang, Yuan Xie, Huangzhong Yang. "MNSIM: Simulation Platform for Memristor-based Neuromorphic Computing System." *Proceedings of IEEE/ACM Design Automation and Test in Europe (DATE), 2016.*

[C15]. **Peng Gu**, Boxun Li, Tianqi Tang, Shimeng Yu, Yu Cao, Yu Wang, Huazhong Yang. "Technological Exploration of RRAM Crossbar Array for Matrix-vector Multiplication." *Proceedings of the 20th Asia and South Pacific Design Automation Conference (ASP-DAC), 2015.*

[C16]. Boxun Li, Lixue Xia, **Peng Gu**, Yu Wang, Huazhong Yang. "Merging the Interface: Power, Area and Accuracy Co-optimization for RRAM Crossbar-based Mixed-signal Computing System." *Proceedings of the 52nd Design Automation Conference (DAC), 2015.*

[C17]. Yu Wang, Tianqi Tang, Lixue Xia, Boxun Li, **Peng Gu**, Huazhong Yang, Hai Li, Yuan Xie. "Energy Efficient RRAM Spiking Neural Network for Real Time Classification." *Proceedings of the 25th Great Lakes Symposium on VLSI (GLSVLSI), 2015.*

Patents

[P1]. **Peng Gu**, Krishna T. Malladi, Hongzhong Zheng. "Computing mechanisms using lookup tables stored on memory." *US Patent 10,628,295.*

[P2]. **Peng Gu**, Krishna T. Malladi, Hongzhong Zheng, Dimin Niu. "Dataflow accelerator architecture for general matrix-matrix multiplication and tensor computation in deep learning." *US Patent App. 16/388,863.*

[P3]. Yu Wang, Boxun Li, **Peng Gu**, Tianqi Tang, Lixue Xia, Huazhong Yang "The method for parameter configuration of memristor crossed array." *CN Patent CN105,390,520 B.*

[P4]. Yu Wang, Boxun Li, Lixue Xia, **Peng Gu**, Tianqi Tang, Huazhong Yang "Digital-to-analogue mixed signal processing system for Imprecise computation." *CN Patent CN105,184,365 B.*

[P5]. Krishna Malladi, Hongzhong Zheng, Dimin Niu, **Peng Gu** "Scale-out High Bandwidth Memory System." *US Patent App. 16/194,219.*

[P6]. **Peng Gu**, Krishna Malladi, Hongzhong Zheng. "HBM Silicon Photonic TSV Architecture for Lookup Computing AI Accelerator." *US Patent App. 15/911,063.*

[P7]. **Peng Gu**, Krishna Malladi, Hongzhong Zheng. "Computing Accelerator Using a Lookup Table." *US Patent App. 15/916,196.*

[P8]. Krishna T. Malladi, **Peng Gu**, Hongzhong Zheng, Robert Brennan. "Memory Lookup Computing Mechanisms." *US Patent App. 15/913,758.*

[P9]. **Peng Gu**, Krishna T. Malladi, Hongzhong Zheng. "HBM-based Memory Lookup Engine for Deep Learning Accelerator." *US Patent App. 15/916,228.*