Peipei Zhou

OFFICE: Benedum Hall Office 1209, 3700 O'Hara Street, Pittsburgh, PA 15213, U.S.
EMAIL: peipei.zhou@pitt.edu • **HOMEPAGE:** https://peipeizhou-eecs.github.io/

RESEARCH INTEREST

Customized Computer Architecture and Programming Abstraction for Applications including Healthcare, e.g., Precision Medicine and Artificial Intelligence.

ACADEMIC POSITION

University of Pittsburgh 2021/09 - Present
Tenure-Track Assistant Professor of Electrical and Computer Engineering Department

EDUCATION

University of California, Los Angeles 2014/06 – 2019/08
Ph.D., Computer Science, GPA: 4.0/4.0 Advisor: Prof. Jason Cong
Dissertation: Modeling and Optimization for Customized Computing: Performance, Energy and Cost Perspective.

University of California, Los Angeles 2012/09 - 2014/06
M.S., Electrical Engineering, GPA: 3.8/4.0 Advisor: Prof. Jason Cong
Thesis: A Fully Pipelined and Dynamically Composable Architecture of CGRA.

Southeast University, Chien-Shiung Wu Honors College 2008/08 - 2012/06
B.S., Electrical Engineeringg, GPA: 3.9/4.0

HONORS AND AWARDS

- **Donald O. Pederson Best Paper Award**, Caffeine: Towards Uniformed Representation and Acceleration for Deep Convolutional Neural Networks, awarded annually to recognize the best paper published in the IEEE Transactions on CAD in the two calendar years preceding the award, 2019
- **Outstanding Recognition in Research in Computer Science**, awarded annually to recognize top 3 outstanding Ph.D. Researcher in UCLA CS department, 2019
- **Phi Tau Phi Scholarship**, awarded annually in recognition of academic achievements and scholarly contributions in West America, 2018
- **ACM Design Automation Conference PhD Forum Travel Grant**, 2018
- **Best Paper Nominee**, 6 papers out of 400 submissions, 2018 IEEE/ACM International Conference on Computer Aided Design (ICCAD 2018)
- **Best Paper Nominee**, 4 papers out of 67 submissions, 2018 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2018)
- **IEEE Council on Electronic Design Automation Travel Grant**, 2016

[1]Updated November 29, 2021

- **Best Poster Award**, High Throughput Sequencing (HiTSeq 2015)
- **Honeywell Innovator Scholarship**, one of five recipients in Mainland China, 2011

PUBLICATIONS

17. [**TECS'21**] Xinyi Zhang, Yawen Wu, Peipei Zhou, Xulong Tang, Jingtong Hu. "Algorithm-hardware Co-design of Attention Mechanism on FPGA Devices". ACM Transactions on Embedded Computing Systems (TECS'21).

16. [**FPGA'21**] Peipei Zhou*, Jiayi Sheng, Cody Hao Yu, Peng Wei, Jie Wang, Di Wu, Jason Cong. "MOCHA: Multinode Cost Optimization in Heterogeneous Clouds with Accelerators". 2021 ACM/SIGDA International Symposium on Field Programmable Gate Arrays (FPGA).

15. [**FCCM'20**] Algorithm-Hardware Co-design for BQSR Acceleration in Genome Analysis ToolKit Michael Lo, Zhenman Fang, Jie Wang, Peipei Zhou, Mau-Chung Frank Chang, Jason Cong IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM), May 2020

14. [**ICCAD'18**] Yuze Chi, Jason Cong, Peng Wei, Peipei Zhou. "SODA: Stencil with Optimized Dataflow Architecture" (Best Paper Nominee). IEEE/ACM International Conference on Computer Aided Design (ICCAD), November 2018, best paper nominee ratio: 6/400 = 1.5%

13. [**TCAD'18**] Chen Zhang, Guangyu Sun, Zhenman Fang, Peipei Zhou, Peichen Pan, Jason Cong. "Caffeine: Towards Uniformed Representation and Acceleration for Deep Convolutional Neural Networks" (Donald O. Pederson Best Paper Award 2019). IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), October 2018

12. [**FCCM'18**] Jason Cong, Peng Wei, Cody Hao Yu, Peipei Zhou*. "Latte: Locality Aware Transformation for High-Level Synthesis". IEEE International Symposium on Field Programmable Custom Computing Machines (FCCM), short paper, May 2018, acceptance ratio: 7/48 = 14.6%

11. [**FCCM'18**] Zhenyuan Ruan, Tong He, Bojie Li, Peipei Zhou, Jason Cong. "ST-Accel: A High-Level Programming Platform for Streaming Applications on FPGA". IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM), May 2018, acceptance ratio: 22/106 = 20.7%

10. [**ISPASS'18**] Peipei Zhou*, Zhenyuan Ruan, Zhenman Fang, Megan Shand, David Roazen, Jason Cong Doppio: I/O-Aware Performance Analysis, Modeling and Optimization for In-Memory Computing Framework (Best Paper Nominee). IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), April 2018, best paper nominee ratio: 4/67 = 5.9%

9. [**DAC'17**] Jason Cong, Peng Wei, Cody Hao Yu, Peipei Zhou. "Bandwidth Optimization Through On-Chip Memory Restructuring for HLS" 54th Annual Design Automation Conference (DAC), June 2017, acceptance rate: 161/676 = 24%

8. [**ICCAD'16**] Chen Zhang, Zhenman Fang, Peipei Zhou, Peichen Pan, Jason Cong. "Caffeine: Towards Uniformed Representation and Acceleration for Deep Convolutional Neural Networks". IEEE/ACM International Conference on Computer Aided Design (ICCAD), November 2016, acceptance rate: 97/408 = 23.8%

7. [**FCCM'16**] Peipei Zhou*, Hyunseok Park, Zhenman Fang, Jason Cong, André DeHon. "Energy Efficiency of Full Pipelining: A Case Study for Matrix Multiplication". IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM), May 2016, acceptance rate: 32/133 = 24.1%

6. [**FCCM'14**] Jason Cong, Hui Huang*, Chiyuan Ma, Bingjun Xiao*, Peipei Zhou*. "A Fully Pipelined and Dynamically Composable Architecture of CGRA". IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM), May 2014, acceptance rate: 22/134 = 16.4%

PREPRINTS AND CONFERENCE POSTERS

5. [**arXiv'18**] Jason Cong, Zhenman Fang, Yuchen Hao, Peng Wei, Cody Hao Yu, Chen Zhang, Peipei Zhou. "Best-Effort FPGA Programming: A Few Steps Can Go a Long Way". arXiv:1807.01340 [cs.AR], July 2018

4. [**FPGA'18**] Yuze Chi, Peipei Zhou, Jason Cong. "An Optimal Microarchitecture for Stencil Computation with Data Reuse and Fine-Grained Parallelism (Abstract Only)". ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA), February 2018

3. [**arXiv'16**] Yu-Ting Chen, Jason Cong, Zhenman Fang, Bingjun Xiao, Peipei Zhou. "ARA-Prototyper: Enabling Rapid Prototyping and Evaluation for Accelerator-Rich Architecture". arXiv:1610.09761 [cs.AR], October 2016

2. [**FPGA'16**] Yu-ting Chen, Jason Cong, Zhenman Fang, Peipei Zhou. "ARAPrototyper: Enabling Rapid Prototyping and Evaluation for Accelerator-Rich Architecture (Abstract Only)". IEEE ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA), February 2016

1. [**HitSeq'16**] Yu-Ting Chen, Jason Cong, Jie Lei, Sen Li, Myron Peto, Paul Spellman, Peng Wei, and Peipei Zhou. "CS-BWAMEM: A fast and scalable read aligner at the cloud scale for whole genome sequencing (Abstract Only)" High Throughput Sequencing, Algorithms,& Applications (HiTSeq), an ISMB/ECCB 2015 special interest group (SIG) satellite conference, July 2015

TALKS

- 2021/10, Duke ECE SEMINAR: "Practice on Performance Autotuning in AI Compute Chip"
- 2020/10, IEEE Pittsburgh Section: "Women in ECE Panel Discussion"
- 2018/12, "Customizable Domain Specific Computing (3-Minute Lightning Talk)". National Science Foundation (NSF) 10-Year Expeditions Anniversary PI Meeting: 10 Years of Transforming Science & Society, Washington. D.C.

- 2019/02, "Mocha: Multinode Optimization of Cost in Heterogeneous Cloud with Accelerators (Technology Review Talk)". Falcon Computing Solutions Technology Review, Los Angeles
- 2018/11, "Cost Optimization Engine for Heterogeneous Public Cloud Featuring Genomics Workloads (Technology Review Talk)". Falcon Computing Solutions Technology Review, Los Angeles
- 2017/06, "Analyzing and Accelerating Genome Analysis Toolkit GATK4". CDSC/InTrans Project Annual Review, Los Angeles. The Center for Domain-Specific Computing (CDSC) was established in 2009. The initial funding for CDSC was provided by the National Science Foundation under the Expedition in Computing Program. Intel Corporation become the first industrial partner to provide financial support of CDSC under the Innovation Transitions (InTrans) Program. Participating universities are University of California, Los Angeles, Rice University, Ohio State University, Oregon Health and Science University (OHSU).
- 2016/10, "Quantifying the I/O Impact on Apache Spark: with Application to Computational Genomic". CDSC/InTrans Project Annual Review, Los Angeles.
- 2016/10, "Fully Pipeline or Not? An Energy Perspective". CDSC/InTrans Project Annual Review, Los Angeles
- 2016/01, "Parallelization of CAVIAR and Identification of Causal Variants in Breast Cancer". CDSC/InTrans Project Annual Review, Los Angeles.
- 2014/05, "Prototype of A Fully Pipelined Configurable Array FPCA". CDSC Year 5 Semi-Annual Review, Los Angeles.
- 2013/10, "FPGA Prototyping of Accelerator-Rich Architectures". CDSC Year 4 Semi-Annual Review, Los Angeles.

ACADEMIC SERVICES

- **Conference Session Chair**, DAC 2021, IEEE/ACM CHASE 2021
- **Technical Program Committee**, DAC 2022, DAC 2021, FCCM 2022, MLSYS 2022, IEEE SOCC 2021, IEEE SOCC 2020, H2RC 2020, IEEE VLSI-DAT 2021
- **Reviewer**, IEEE TRETS, IEEE TII, IEEE JETCAS, IEEE TVLSI, IEEE TC, 2020
- **Reviewer**, IEEE Transactions on Parallel and Distributed Systems, 2019, 2020
- **Subreviewer**, IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM), 2017, 2018
- **Subreviewer**, IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2018
- **Subreviewer**, ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA), 2019
- **Subreviewer**, IEEE/ACM International Symposium on Microarchitecture (Micro), 2017
- **Subreviewer**, IEEE International Symposium on Workload Characterization (IISWC), 2017
- **Subreviewer**, IEEE International Conference on Parallel and Compilation Techniques (PACT), 2016

- **Subreviewer**, IEEE International Conference on Field-Programmable Logic and Applications (FPL), 2016, 2018, 2019

INDUSTRIAL EXPERIENCES

Enflame Technology 2019/08 – 2021/08
Staff Software Engineer
I worked on the biggest AI chip in China featured with 57.5mmx57.5mm package, tens of billions of transistors and hundreds of TOPs computation throughput till July 2021. My responsibilities include: 1) Pre-silicon Architecture Exploration and Modeling; 2) Post-silicon AI Software Stack Build and System Optimization; 3) Domain Specific Language for AI Accelerator including MLIR based Compilation and Optimization. I Led a group of 6-people team working on high-performance convolution neural network library for deep learning ASIC accelerator. I also coordinated weekly progress within software teams over 50 people and collaborated with hardware, platform, product system teams across the whole company. I was awarded with **Enflame CEO Award**, **Enflame COO Award**, the highest honors in the company for multiple times.

Falcon Computing Solutions, later acquired by Xilinx 2018/06 – 2019/04
Software Engineer Intern
I worked on heterogeneous computing for genomic application by orchestrating seas of datacenter scale resources including computing (CPUs+GPU+FPGA accelerators), storage (HDD, SSD, local disk) and etc.

Microsoft, Microsoft Headquarter, Washington 2017/06 – 2017/09
Software Engineer Intern Supervisor: Robert Rounthwaite
I worked in Office Machine Learning Team with my manager Robert Rounthwaite and my mentor Dr. Vincent Etter where I developed a scalable end-to-end tool to generate over 2M grammar fixed (preposition, article and etc.) sentences from Wikipedia. I also implemented neural network model for language engine tasks.

Microsoft Research, Microsoft Headquarter, Washington 2014/06 – 2014/09
Research Engineer Intern Supervisor: Doug Burger
I work in Computer Architecture Group directed by Dr. Doug Burger in Microsoft Research, Redmond Campus. My mentor Dr. Joo-Young Kim and I worked on image compression pipeline. I have implemented the advanced compression algorithm in C++(software reference code) and also implemented hardware accelerator on FPGA.

CONTRIBUTING OPEN-SOURCE SOFTWARE

- Latte: Latte is an automated framework to insert pipelined transfer controllers along data paths in HLS with minimal user efforts
  Github: https://github.com/AriesLL/Latte
  Publication: Latte: Locality Aware Transformation for High-Level Synthesis, FCCM 2018

- Doppio: Doppio is a framework that builds I/O-Aware Analytic Model for Apache Spark Applications
  Github: https://github.com/UCLA-VAST/Doppio
  Publication: Doppio: I/O-Aware Performance Analysis, Modeling and Optimization for In-Memory Computing Framework, ISPASS 2018
- Java-Fpga-Pipeline: Java-Fpga-Pipeline implements a fully pipelined data transfer stack that achieves efficient JVM-FPGA communication through extensive pipelining
  Github: https://github.com/UCLA-VAST/java-fpga-pipeline
  Publication: From JVM to FPGA: Bridging Abstraction Hierarchy via Optimized Deep Pipelining, HotCloud 2018
- HDLRevisit: HDLRevisit provides best-effort programming guideline and design templates for HLS FPGA accelerator design
  Github: https://github.com/peterpengwei/HDLRevisit
  Publication: Best-Effort FPGA Programming: A Few Steps Can Go a Long Way, arXiv:1807.01340 [cs.AR], 2018
- CS-BWAMEM: Cloud-scale BWAMEM (CS-BWAMEM) is an ultrafast and highly scalable aligner built on top of cloud infrastructures, including Spark and Hadoop distributed file system (HDFS)
  Github: https://github.com/ytchen0323/cloud-scale-bwamem
  Publication: CS-BWAMEM: A fast and scalable read aligner at the cloud scale for whole genome sequencing, HitSeq 2015