Peipei Zhou

OFFICE: Benedum Hall 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
- Characterization, modeling, and optimization in performance, energy, and cost for mobile and embedded systems, distributed systems, heterogeneous systems

ACADEMIC POSITION

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

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 and Computer 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 and Computer Engineering, GPA: 3.9/4.0, Rank: 1/90

HONORS AND AWARDS

- **IEEE Council on Electronic Design (CEDA) IEEE Transactions on Computer-Aided Design Donald O. Pederson Best Paper Award, 2019**, "Caffeine: Towards Uniformed Representation and Acceleration for Deep Convolutional Neural Networks", 2 out of 1000+ submissions, awarded annually to recognize the best paper published in the IEEE Transactions on CAD in the two calendar years preceding the award
- **IEEE/ACM ICCAD Best Paper Nominee, 2018**, "SODA: Stencil with Optimized Dataflow Architecture", 6 papers out of 400 submissions, 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD 2018)
- **IEEE ISPASS Best Paper Nominee, 2018**, "Doppio: I/O-Aware Performance Analysis, Modeling and Optimization for In-Memory Computing Framework", 4 papers out of 67 submissions, 2018 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2018)

[1]Updated May 15, 2023

- **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 Ph.D. Forum Travel Award**, 2018
- **IEEE Council on Electronic Design Automation Travel Award**, 2016
- **Best Poster Award**, High Throughput Sequencing (HiTSeq 2015)
- **Honeywell Innovator Scholarship**, one of five recipients in Mainland China, 2011

PUBLICATIONS

23. [**DAC'23**] Jinming Zhuang, Zhuoping Yang, Peipei Zhou, "High Performance, Low Power Matrix Multiply Design on ACAP: from Architecture, Design Challenges, and DSE Perspectives," 60th ACM/IEEE Design Automation Conference, San Francisco, California, USA, (DAC '23), acceptance ratio: 263/1157 = 22.7%, July 9–13, 2023, San Francisco, CA, USA. Open-source: https://anonymous.4open.science/r/AutoMM-main/README.md.

22. [**FPGA'23**] Jinming Zhuang, Jason Lau, Hanchen Ye, Zhuoping Yang, Yubo Du, Jack Lo, Kristof Denolf, Stephen Neuendorffer, Alex Jones, Jingtong Hu, Deming Chen, Jason Cong, Peipei Zhou, "CHARM: Composing Heterogeneous Accelerators for Matrix Multiply on Versal ACAP Architecture," Proceedings of the 2023 ACM/SIGDA International Symposium on Field Programmable Gate Arrays (FPGA '23), February 12–14, 2023, Monterey, CA, USA. Open-source: https://github.com/arc-research-lab/CHARM.

21. [**IEEE Micro'23**] Sebastien Ollivier , Sheng Li , Yue Tang , Stephen Cahoon , Ryan Caginalp , Chayanika Chaudhuri , Peipei Zhou , Xulong Tang , Jingtong Hu , Alex K. Jones, "Sustainable AI Processing at the Edge," IEEE Micro.

20. [**TCAD'22**] Yue Tang, Yawen Wu, Peipei Zhou and Jingtong Hu, "Enabling Weakly Supervised Temporal Action Localization From On-Device Learning of the Video Stream," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 41, no. 11, pp. 3910-3921, Nov. 2022, doi: 10.1109/TCAD.2022.3197536.

19. [**DAC'22**] Xinyi Zhang, Cong Hao, Peipei Zhou, Alex Jones and Jingtong Hu. "H2H: Heterogeneous Model to Heterogeneous System Mapping with Computation and Communication Awareness". 59th ACM/IEEE Annual Design Automation Conference 2022 (DAC'22).

18. [**TODAES'22**] Yue Tang, Xinyi Zhang, Peipei Zhou, Jingtong Hu. "EF-Train: Enable Efficient On-device CNN Training on FPGA Through Data Reshaping for Online Adaptation or Personalization". ACM Transactions on Design Automation of Embedded Systems (TODAES) Special Issue on Energy-Efficient AI Chips (TODAES'22).

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**] Michael Lo, Zhenman Fang, Jie Wang, Peipei Zhou, Mau-Chung Frank Chang, Jason Cong. "Algorithm-Hardware Co-design for BQSR Acceleration in Genome Analysis ToolKit". 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

ADVISEES

- Doctoral Advisor for:
  Jinming Zhuang (2021/08–Now)
  Zhuoping Yang (2022/08–Now)
- Doctoral Dissertation Committee Member for:
  Xin Xin (2021/07–Now),
  Yuqi Li (2022/02–Now),
  Yue Tang (2022/04–Now),
  Yanan Guo (2022/11–Now),
  Noah Perryman (2022/11–Now),
  Luke Kljucaric (2022/12–Now),
  Tyler Garrett (2022/12–Now),
  Vivswan Shah (2023/02–Now)
  Jieru Zhao (2023/04–Now)

COURSES

- 2021-2022 Spring: ECE2195 Reconfigurable Computing in Deep Learning (Graduate Course, attendees: 23), Teaching Score: 4.10 / 5.00
- 2022-2023 Fall: ECE1110 Computer Organization and Architecture (Undergraduate Course, attendees: 51), Teaching Score: 4.34 / 5.00
- 2022-2023 Spring: ECE2195 Reconfigurable Computing in Deep Learning (Graduate Course, attendees: 14), Teaching Score: 4.48 / 5.00
- 2022-2023 Spring: ECE1175 Embedded Systems Design (Undergraduate Course, attendees: 17), Teaching Score: 4.50 / 5.00

TALKS

- 2023/04/21, TAMU CESG Seminars, Invited Talk: "CHARM: Composing Heterogeneous Accelerators for Matrix Multiply on Versal ACAP Architecture"
  https://cesg.tamu.edu/2023/03/01/cesg-seminar-peipei-zhou/
- 2023/04/07, CMU Crossroads Seminars, Invited Talk: "CHARM: Composing Heterogeneous Accelerators for Matrix Multiply on Versal ACAP Architecture"
  Youtube link: https://youtu.be/Hl__s5WqSZA
- 2022/7, DAC-ROAD4NN 2022 International Workshop on Research Open Automatic Design for Neural Networks 2022 Invited Talk: "Practice on Performance Autotuning in AI Compute Chip"
- 2022/7, 31st International Workshop on Logic & Synthesis Keynote : "Practice on Performance Autotuning in AI Compute Chip",
  https://www.iwls.org/iwls2022/
- 2022/1, The Pittsburgh Center for AI Innovation in Medical Imaging (CAIIMI) SEMINAR: "Domain-Specific Computing for Artificial Intelligence",
  https://twitter.com/PittCAIIMI/status/1483443239344652291
- 2021/10, Duke ECE SEMINAR: "Practice on Performance Autotuning in AI Compute Chip",
  https://ece.duke.edu/about/events/3210
- 2020/10, IEEE Pittsburgh Section: "Women in ECE Panel Discussion",
  https://events.vtools.ieee.org/m/241997
- 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 becomes the first industrial partner to provide financial support to CDSC under the Innovation Transitions (InTrans) Program. Participating universities are the 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

- **Technical Program Committee**
  - ICCAD 2023
  - FCCM 2023
  - MLSYS 2023
  - DAC 2023
  - FPGA 2023
  - CHASE 2023
  - DATE 2023
  - ICCAD 2022
  - ESWEEK 2022
  - FCCM 2022
  - DAC 2022
  - MLSYS 2022
  - DAC 2021
  - IEEE SOCC 2021
  - IEEE SOCC 2020
  - H2RC 2020 2021 2022

- IEEE VLSI-DAT 2021

- **Organization Committee**
  - DAC 2023: University Demonstration Chair
  - NOCS 2023: Publicity Co-Chair
  - FCCM 2022: Hybrid Co-Chair
  - DAC 2022: University Demonstration Vice Co-Chair
- **Conference Session Chair**
  - DAC 2023
  - ICCAD 2022
  - MLSYS 2022
  - DAC 2022
  - FCCM 2022
  - DAC 2021
  - IEEE/ACM CHASE 2021
- **Reviewer**
  - IEEE Computer Architecture Letters (CAL)
  - IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems(TCAD)
  - ACM Transactions on Architecture and Code Optimization(TACO)
  - ACM Transactions on Reconfigurable Technology and Systems (TRETS)
  - IEEE Transactions on Industrial Informatics (TII)
  - IEEE Transactions on Very Large Scale Integration (VLSI) Systems
  - IEEE Transactions on Computers (TC)
  - IEEE Transactions on Parallel and Distributed Systems (TPDS)
  - IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)
  - The Computer Journal (COMPJ)
- **Subreviewer**
  - FPGA 2019
  - FCCM 2018, FCCM 2017
  - FPL 2019, FPL 2018, FPL 2016
  - HPCA 2018
  - IISWC 2017
  - MICRO 2017
  - PACT 2016

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 multi-level intermediate representation (MLIR)-based Compilation and Optimization. I Led a group of the 6-people team working on the high-performance convolution neural network library for deep learning ASIC accelerator. I also coordinated weekly progress within software teams of over 50 people and collaborated with hardware, platform, and product system teams across the whole company. **I was awarded Enflame CEO Award, and Enflame COO Award, the highest honors in the company 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. I published MOCHA (FPGA'23) paper from my intern work at Falcon Computing Solutions.

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 a 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 the image compression pipeline. I have implemented the advanced compression algorithm in C++(software reference code) and also implemented the hardware accelerator on FPGA.

CONTRIBUTING OPEN-SOURCE SOFTWARE

- CHARM: CHARM is an automated framework to compile matrix-multiply (MM) based applications and MM-based deep learning models onto AMD Versal adaptive compute acceleration platforms (ACAPs) hardware.
  Github: https://github.com/arc-research-lab/CHARM
  Publication: CHARM: Composing Heterogeneous Accelerators for Matrix Multiply on Versal ACAP Architecture, FPGA 2023

High Performance, Low Power Matrix Multiply Design on ACAP: from Architecture, Design Challenges, and DSE Perspectives, DAC 2023

- 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 guidelines 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 the whole genome sequencing, HitSeq 2015