Peipei Zhou

OFFICE: ERC 328, 184 Hope Street, Providence, RI 02912, U.S.
EMAIL: peipei_zhou@brown.edu • **HOMEPAGE:** https://peipeizhou-eecs.github.io/
Google Scholar: https://scholar.google.com/citations?user=px_jwFgAAAAJ&hl=en

RESEARCH INTEREST

- Customized computer architecture, programming abstraction, and electronic design automation for applications including healthcare, e.g., precision medicine, and artificial intelligence
- Characterization, modeling, and optimization in performance, energy, carbon, and cost for reconfigurable and heterogeneous systems from edge to data center
- Sustainable computing and chiplet-based system

ACADEMIC POSITION

Brown University 2024/09 - Present
Tenure-Track Assistant Professor, School of Engineering
Brown Undergraduate Computer Engineering Concentration Advisor

University of Pittsburgh 2021/09 - 2024/08
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 CEDA Award Link and UCLA Computer Science Department Award Link

[1]Updated September 16, 2025

- **ACM/SIGDA FPGA Best Paper Nominee, 2025**, "ARIES: An Agile MLIR-Based Compilation Flow for Reconfigurable Devices with AI Engines", 2025 ACM/SIGDA International Symposium on Field Programmable Gate Arrays (FPGA 2025)
- **IEEE IGSC Best Viewpoint Paper, 2024**, "Amortizing Embodied Carbon Across Generations", 2024 15th IEEE International Green and Sustainable Computing (IGSC 2024)
- **IEEE/ACM IGSC Best Viewpoint Paper Finalist, 2023**, "REFRESH FPGAs: Sustainable FPGA Chiplet Architectures", 2023 IEEE/ACM International Green and Sustainable Computing (IGSC 2023)
- **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) ISPASS 2018 Program Link
- **Outstanding Recognition in Research in Computer Science**, awarded annually to recognize top 3 outstanding Ph.D. Researcher in UCLA CS department, 2019.

  UCLA Samueli School of Engineering Award Link and UCLA VAST Lab Award Link
- **Phi Tau Phi Scholarship**, awarded annually in recognition of academic achievements and scholarly contributions in West America, 2018

  Phi Tau Phi Scholastic Honor Society of America Award Link
- **Best Poster Award**, High Throughput Sequencing (HiTSeq 2015)
- **Honeywell Innovator Scholarship**, one of five recipients in Mainland China, 2011

PUBLICATIONS

40. [**SC'25**] Zhuoping Yang, Jinming Zhuang, Xingzhen Chen, Alex K Jones, **Peipei Zhou**, "AGILE: Lightweight and Efficient Asynchronous GPU-SSD Integration," Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, SC 2025, Nov. 16 - Nov. 21, 2025, St. Louis, MO, US.

39. [**GLSVLSI'25**] Shixin Ji, Xingzhen Chen, Jinming Zhuang, Wei Zhang, Zhuoping Yang, Sarah Schultz, Yukai Song, Jingtong Hu, Alex Jones, Zheng Dong, **Peipei Zhou**, "ART: Customizing Accelerators for DNN-Enabled Real-Time Safety-Critical Systems," Proceedings of the 2025 ACM Proceedings of the Great Lakes Symposium on VLSI 2025, 30 June 2025 - 2 July 2025, New Orleans, LA, USA.

38. [**FPGA'25**] Jinming Zhuang, Shaojie Xiang, Hongzheng Chen, Niansong Zhang, Zhuoping Yang, Tony Mao, Zhiru Zhang, **Peipei Zhou**, "ARIES: An Agile MLIR-Based Compilation Flow for Reconfigurable Devices with AI Engines," Proceedings of the 2025 ACM/SIGDA International Symposium on Field Programmable Gate Arrays (FPGA'25), Feb 27 – Mar 1, 2025, Monterey, CA, USA. Open-source: `https://github.com/arc-research-lab/Aries`. **(FPGA'25 Best Paper Nominee)**.

37. [**TCAD'25**] Yue Tang, Alex K. Jones, Jinjun Xiong, **Peipei Zhou**, Jingtong Hu, "MTrain: Enable Efficient CNN Training on Heterogeneous FPGA-based Edge Servers," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2025.

36. [**TCAD'24**] Peiyan Dong*, Jinming Zhuang*, Zhuoping Yang, Shixin Ji, Dongkuan Xu, Yanyu Li, Heng Huang, Jingtong Hu, Alex K. Jones, Yiyu Shi, Yanzhi Wang, **Peipei Zhou**, "EQ-ViT: Algorithm-Hardware Co-Design for End-to-End Acceleration of Real-Time Vision Transformer Inference on Versal ACAP Architecture," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2024 (CODES+ISSS 2024 in ESWEEK).

35. [**TCAD'24**] Yue Tang, Yukai Song, Naveena Elango, Sheena Ratnam Priya, Alex K. Jones, Jinjun Xiong, **Peipei Zhou**, Jingtong Hu, "CHEF: A Framework for Deploying Heterogeneous Models on Clusters with Heterogeneous FPGAs," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2024 (CODES+ISSS 2024 in ESWEEK).

34. [**ESWEEK'24**] Zhuoping Yang, Wei Zhang, Shixin Ji, **Peipei Zhou**, Alex K Jones, "Reducing Smart Phone Environmental Footprints with In-Memory Processing," 2024 International Conference on Hardware/Software Codesign and System Synthesis, 2024 (CODES+ISSS 2024 in ESWEEK).

33. [**IGSC'24**] Shixin Ji, Jinming Zhuang, Zhuoping Yang, Alex K Jones, **Peipei Zhou**, "Amortizing Embodied Carbon Across Generations," 15th International Green and Sustainable Computing Conference (IGSC'24).
(IGSC'24 Best Viewpoint Paper).

32. [**TRETS'24**] Jinming Zhuang, Jason Lau, Hanchen Ye, Zhuoping Yang, Shixin Ji, Jack Lo, Kristof Denolf, Stephen Neuendorffer, Alex Jones, Jingtong Hu, Yiyu Shi, Deming Chen, Jason Cong, **Peipei Zhou**, "CHARM 2.0: Composing Heterogeneous Accelerators for Deep Learning on Versal ACAP Architecture," Accepted at ACM Transactions on Reconfigurable Technology and Systems, 2024. Open-source: https://github.com/arc-research-lab/CHARM.

31. [**TRETS'24**] Yuqi Li, Kehao Zhao, Jieru Zhao, Qirui Wang, Shuda Zhong, Nageswara Lalam, Ruishu Wright, **Peipei Zhou**, Kevin P. Chen, "FiberFlex: Real-time FPGA-based Intelligent & Distributed Fiber Sensor System for Pedestrian Recognition," Accepted at ACM Transactions on Reconfigurable Technology and Systems, 2024.

30. [**ISVLSI'24**] Shixin Ji, Zhuoping Yang, Xingzhen Chen, Stephen Cahoon, Jingtong Hu, Yiyu Shi, Alex Jones, **Peipei Zhou**, "SCARIF: Towards Carbon Modeling of Cloud Servers with Accelerators," 2024 IEEE Computer Society Annual Symposium on VLSI, ISVLSI 2024. Open-source: https://github.com/arc-research-lab/SCARIF.

29. [**DAC'24**] Ruiyang Qin, Jun Xia, Zhenge Jia, Meng Jiang, Ahmed Abbasi, **Peipei Zhou**, Jingtong Hu, Yiyu Shi, "Enabling On-Device Large Language Model Personalization with Self-Supervised Data Selection and Synthesis," 61st ACM/IEEE Design Automation Conference, San Francisco, California, USA, (DAC'24).

28. [**FPGA'24**] Jinming Zhuang, Zhuoping Yang, Shixin Ji, Heng Huang, Alex K. Jones, Jingtong Hu, Yiyu Shi, **Peipei Zhou**, "SSR: Spatial Sequential Hybrid Architecture for Latency

Throughput Tradeoff Design Space Exploration," Proceedings of the 2024 ACM/SIGDA International Symposium on Field Programmable Gate Arrays (FPGA'24), March 3-5, 2024, Monterey, CA, USA. Open-source: https://github.com/arc-research-lab/SSR.

27. [**ASPDAC'24**] Zhuoping Yang, Shixin Ji, Xingzhen Chen, Jinming Zhuang, Weifeng Zhang, Dharmesh Jani, **Peipei Zhou**, "Challenges and Opportunities to Enable Large-Scale Computing via Heterogeneous Chiplets," Proceedings of the 28th Asia and South Pacific Design Automation Conference, ASPDAC'24, Incheon Songdo Convensia, South Korea.

26. [**IGSC'23**] **Peipei Zhou**, Jinming Zhuang, Stephen Cahoon, Yue Tang, Zhuoping Yang, Xingzhen Chen, Yiyu Shi, Jingtong Hu, Alex K. Jones, "REFRESH FPGAs: Sustainable FPGA Chiplet Architectures," 14th International Green and Sustainable Computing Conference (IGSC'23).
(IGSC'23 Best Viewpoint Paper Finalist)

25. [**ICCAD'23**] Zhuoping Yang, Jinming Zhuang, Jiaqi Yin, Cunxi Yu, Alex K. Jones, **Peipei Zhou**, "AIM: Accelerating Arbitrary-precision Integer Multiplication on Heterogeneous Reconfigurable Computing Platform Versal ACAP," IEEE/ACM International Conference on Computer-Aided Design, (ICCAD'23), acceptance ratio: 21%, October 29 - November 2, 2023, San Francisco, CA, USA. Open-source: https://github.com/arc-research-lab/AIM.

24. [**Book Chapter**] Yawen Wu, Yue Tang, Dewen Zeng, Xinyi Zhang, **Peipei Zhou**, Yiyu Shi, Jingtong Hu, "Efficient Hardware and Software Design for On-device Learning," Embedded Machine Learning for Cyber-Physical, IoT, and Edge Computing: Software Optimizations and Hardware/Software Codesign, Springer Nature Switzerland, 2024.

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://github.com/arc-research-lab/CHARM.

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.
No.1 Downloaded Paper in 2023 ACM/SIGDA FPGA Proceeding

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, 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).
Student Yue won TODAES Rookie Author of the Year (RAY) Award 2023

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". IEEE/ACM International Conference on Computer-Aided Design (ICCAD), November 2018, best paper nominee ratio: 6/400 = 1.5%
(ICCAD'18 Best Paper Nominee)

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". IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), October 2018
(IEEE TCAD 2019 Donald O. Pederson Best Paper Award)

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. IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), April 2018, best paper nominee ratio: 4/67 = 5.9%
(ISPASS'18 Best Paper Nominee)

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**. "ARAPrototyper: 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)
  Shixin Ji (2023/08–Now)
  Xingzhen Chen (2023/08–Now)
  Wei Zhang (2024/08–Now)
  Sarah Schultz (2025/01–Now)
- Master Advisor for:
  Sarah Schultz (2024/08–2024/12)
- Doctoral Dissertation Committee Member for:
  Xin Xin (2021/07–2023/06)
  Yuqi Li (2022/02–Now)
  Yue Tang (2022/04–2024/07)
  Yanan Guo (2022/11–2024/06)
  Noah Perryman (2022/11–2024/07)
  Luke Kljucaric (2022/12–2023/10)
  Tyler Garrett (2022/12–2024/07)
  Vivswan Shah (2023/02–Now)
  Jieru Zhao (2023/04–2024/07)
  Calvin Gealy (2023/08–Now)
  Michael Cannizzaro (2023/09–Now)
  Kai Ye (2023/11–Now)
  Siyuan Dai (2023/12–Now)
  Daniel Stumpp (2023/12–Now)
  Jefferson Booth (2024/03–Now)
  Kai Huang (2024/03–Now)
  Richard Gibbons (2024/04–Now)
  Jianing Deng (2024/04–Now)
  Qiang Zhou (2024/04–Now)
- Undergraduate Researcher Advisor for:
  Devon Kear-Leng (2025/09–Now)
  Gannon Lemaster (2025/09–Now)

COURSES TAUGHT

Courses taught at Brown University

- 2024-2025 Fall: ENGN2911X Reconfigurable Computing (Graduate Course, attendees: 24)
- 2024-2025 Spring: ENGN1600 Design and Implementation of Digital Integrated Circuits (Undergraduate Course, attendees: 33)
- 2025-2026 Fall: ENGN2911X Reconfigurable Computing (Graduate Course)

Courses taught at University of Pittsburgh

- 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
- 2023-2024 Fall: ECE1110 Computer Organization and Architecture (Undergraduate Course, attendees: 60), Teaching Score: 4.16 / 5.00
- 2023-2024 Spring: ECE1110 Computer Organization and Architecture (Undergraduate Course, attendees: 17), Teaching Score: 4.78 / 5.00

TALKS

- 2025/10/07, Distinguished Speaker for the IEEE Lafayette Section YP Affinity Group "IEEE Day Tech Chat" event, Invited Talk: "Efficient Programming on Heterogeneous Accelerators: Architecture and Compiler Insights"

- 2025/09/12, New York Scientific Data Summit 2025: Powering the Future of Science with Artificial Intelligence, Invited Talk: "Efficient Programming on Heterogeneous Accelerators"
  YouTube: https://www.youtube.com/watch?v=1Wo4FOTsTsM
- 2025/05/09, University of Delaware, Computer & Information Sciences Seminar, Invited Talk: "Efficient Programming on Heterogeneous Accelerators"
- 2025/02/06, University of North Texas, Data Science Talk Series, Invited Talk: "Efficient Programming on Heterogeneous Accelerators"
- 2024/12/03, Wayne State University, Computer Science Department Seminar, Invited Talk: "Efficient Programming on Heterogeneous Accelerators"
- 2024/11/26, Rutgers Efficient AI (REFAI) Seminar, Rutgers University, Invited Talk: "Efficient Programming on Heterogeneous Accelerators"
- 2024/10/25, ECE Colloquium, University of Connecticut, Invited Talk: "Efficient Programming on Heterogeneous Accelerators"
- 2024/07/03, ISVLSI 2024 Special Session on Sustainable Computing from Edge to Data Center, Invited Talk: "SCARIF: Towards Carbon Modeling of Cloud Servers with Accelerators"
  Slides: https://peipeizhou-eecs.github.io/publication/2024_isvlsi_scarif/
- 2024/04/19, Open Compute Project (OCP) AI-HW-SW-CoDesign Working Group, Invited Talk: "Efficient Programming on Heterogeneous Accelerators"
- 2023/10/18, San Jose Convention Center, 2023 Open Compute Project (OCP) Global Summit, Future Technology Symposium, Invited Talk: "Architectural Challenges and Innovation for Compute Infrastructure Co-Design"
  YouTube: https://www.youtube.com/watch?v=Lp5Al9vbqLg

Slides: https://peipeizhou-eecs.github.io/talk/architectural-challenges-and-innovation-for-compute-infrastructure-co-design/OCP23G_SYM-PeipeiZhou.pdf

- 2023/08/30, Temple University, Department of Electrical and Computer Engineering Seminars, Invited Talk: "CHARM: Composing Heterogeneous Accelerators for Matrix Multiply on Versal ACAP Architecture"
https://events.temple.edu/fall-seminar-series
- 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: https://youtu.be/Hl__s5WqSZA
Slides: https://peipeizhou-eecs.github.io/publication/fpga23/FPGA2023_CHARM.pdf
- 2022/10, 18th EMBEDDED SYSTEMS WEEK, Invited Tutorial: "Practice on Performance Autotuning in AI Compute Chip"
https://esweek.org/wp-content/uploads/2022/10/ESWEEK_2022_booklet.pdf
- 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.

DEMOS

- 2024/10/01, ESWEEK 2024 Embedded Systems Software Competition (ESSC), Demo: "AIM: Accelerating Arbitrary-precision Integer Multiplication on Heterogeneous Reconfigurable Computing Platform Versal ACAP", Demo Presenters: My Ph.D. students Zhuoping Yang and Jinming Zhuang
  Paper_EventPhotos
- 2024/07/03, ISVLSI 2024 Special Session on Sustainable Computing from Edge to Data Center, Invited Talk: "SCARIF: Towards Carbon Modeling of Cloud Servers with Accelerators"
  Demo_Video_SCARIF_Usage
- 2024/06/25, DAC 2024 University Demonstration, Demo: "Enabling On-Device Large Language Model Personalization with Self-Supervised Data Selection and Synthesis", Demo Presenter: Collaborator Student Ruiyang Qin (Advisor: Prof. Yiyu Shi).
- 2023/09/26, NSF FABRIC KNIT 7, the third in-person FABRIC Community Workshop, Demo: "CHARM: Composing Heterogeneous AcceleRators for End-to-end Deep Learning Inference on Versal ACAP Architecture", Demo Presenters: My Ph.D. students Jinming Zhuang and Zhuoping Yang
  Poster_EventPhotos
- 2023/07/10, DAC 2023 University Demonstration, Demo: "FPGA HLS acceleration on real-time fiber-optic-based pedestrian recognition system", Demo Presenter: Collaborator Student Yuqi Li (advisor: Prof. Kevin Chen), who I mentored, won 3rd Place Award at DAC 2023 University Demonstration
  News_Certificate_Paper_EventPhotos
- 2023/07/10, DAC 2023 University Demonstration, Demo: "Versal-ViT: A Fully Pipelined Spatial Accelerator Design for Vision Transformers on Versal ACAP", Demo Presenter: My Ph.D. student Jinming Zhuang
  Paper_EventPhotos

- 2022/05/16, FCCM 2022 Demo: "XDSE: A Heterogeneous DSE Framework on Versal Platforms", Demo Presenter: My Ph.D. student Jinming Zhuang
Poster_EventPhotos

ACADEMIC SERVICES

- **Associate Editor**

  - IEEE Transactions on Computers (Area: Architecture Design and Optimization), 2023-Present

  - ACM Transactions on Reconfigurable Technology and Systems, 2024-Present

- **Technical Program Committee**
  - FCCM 2026 (TPC Co-Chair)
  - FPGA 2026
  - HPCA 2026 Industry Track
  - ICCAD 2025 (Track Co-Chair)
  - GLSVLSI 2025 (TPC Co-Chair)
  - Supercomputing 2025
  - FCCM 2025
  - FPGA 2025
  - IEEE COINS 2025 (Track Co-Chair)
  - DATE 2025
  - IGSC 2024 (TPC Co-Chair)
  - ICCAD 2024 (Track Co-Chair)
  - HEART 2024 (TPC Co-Chair)
  - 2024 The 1st IEEE International Workshop on LLM-Aided Design (Founding TPC)
  - ASAP 2024
  - Supercomputing 2024
  - DAC 2024
  - FPGA 2024
  - FCCM 2024
  - DATE 2024
  - 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**
  - GLSVLSI 2026: General Chair
  - IGSC 2026: General Chair
  - FCCM 2026: Technical Program Committee (TPC) Co-Chair
  - FPGA 2026: Publicity and Website Chair
  - GLSVLSI 2025: Technical Program Committee (TPC) Chair
  - IEEE COINS 2025: Education Co-Chair
  - HPCA 2025: Registration Co-Chair
  - FCCM 2025: Publicity and Website Chair
  - IGSC 2024: Technical Program Committee (TPC) Chair
  - HipChips 2024 @ MICRO 2024: HiPChips Academic Chair
  - ICCAD 2024: TPC Track Co-Chair
  - IEEE SecDev 2024: SecDev Cybersecurity Award for Practice committee
  - RAW 2024: Ph.D. Forum Chair
  - FCCM 2024: Publicity and Website Chair
  - HEART 2024: Technical Program Committee (TPC) Chair
  - DAC 2023: University Demonstration Chair
  - HipChips 2023 @ ISCA 2023: HiPChips Academic Chair
  - NOCS 2023: Publicity Co-Chair
  - FCCM 2022: Hybrid Co-Chair
  - DAC 2022: University Demonstration Vice Co-Chair
- **Conference Session Chair**

- DAC 2025
- FCCM 2025
- FPGA 2025
- IGSC 2024
- ICCAD 2024
- ISVLSI 2024
- DAC 2024
- FCCM 2024
- FPGA 2024
- ICCAD 2023
- DAC 2023
- ICCAD 2022
- MLSYS 2022
- DAC 2022
- FCCM 2022
- DAC 2021
- IEEE/ACM CHASE 2021

• **Journal 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 Design Automation of Electronic Systems (TODAES)
- ACM Transactions on Reconfigurable Technology and Systems (TRETS)
- IEEE Transactions on Industrial Informatics (TII)
- IEEE Transactions on Very Large Scale Integration Systems (VLSI)
- 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)

• **U.S. Government Agency Proposal Reviewer**
- National Science Foundation, 2022, 2024, 2025
- Department of Energy, 2024

• **Workshop/Special Session/Panel Organizer**
- NSF Workshop on Algorithm-Hardware Co-design for Medical Applications, 2024

- Special Session "Design for Environmental Sustainability in Computing", Embedded Systems Week (ESWEEK), 2024

- **Subreviewer**
  - FPGA 2019
  - FCCM 2018, FCCM 2017
  - FPL 2019, FPL 2018, FPL 2016
  - HPCA 2018
  - IISWC 2017
  - MICRO 2017
  - PACT 2016

PROFESSIONAL EXPERIENCE

Institute of Electrical and Electronics Engineers (IEEE) 2024/06 – Current
Senior Member
Only 10% of IEEE's more than 450,000 members hold this grade, which requires extensive experience, and reflects professional maturity and documented achievements of significance.

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 six-person 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 (acquired by Xilinx), Los Angeles, CA 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, Redmond, WA 2017/06 – 2017/09
Software Engineer Intern Supervisor: Dr. 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, Redmond, WA 2014/06 – 2014/09
Research Engineer Intern Supervisor: Dr. 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

- **ARIES**: ARIES provides a Python-based frontend, an MLIR-based middle end and can target different backend devices with AMD AI Engines.
  GitHub: https://github.com/arc-research-lab/ARIES
  Publication: ARIES: An Agile MLIR-Based Compilation Flow for Reconfigurable Devices with AI Engines, FPGA 2025
- **SSR**: SSR is an automated framework to compile end-to-end transformer-based deep learning inference models onto AMD Versal adaptive compute acceleration platforms (ACAPs) hardware.
  GitHub: https://github.com/arc-research-lab/SSR
  Publication: SSR: Spatial Sequential Hybrid Architecture for Latency Throughput Tradeoff in Transformer Acceleration, FPGA 2024
- **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 1: CHARM: Composing Heterogeneous Accelerators for Matrix Multiply on Versal ACAP Architecture, FPGA 2023
  Publication 2: High Performance, Low Power Matrix Multiply Design on ACAP: from Architecture, Design Challenges, and DSE Perspectives, DAC 2023
  Publication 3: CHARM 2.0: Composing Heterogeneous Accelerators for Deep Learning on Versal ACAP Architecture, ACM Transactions on Reconfigurable Technology and Systems, 2024
  Publication 4: EQ-ViT: Algorithm-Hardware Co-Design for End-to-End Acceleration of Real-Time Vision Transformer Inference on Versal ACAP Architecture, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2024
- **AIM**: AIM is a compilation framework to compile arbitrary-integer arithmetic applications onto AMD Versal ACAP hardware.
  GitHub: https://github.com/arc-research-lab/AIM
  Publication: AIM: Accelerating Arbitrary-precision Integer Multiplication on Heterogeneous Reconfigurable Computing Platform Versal ACAP, ICCAD 2023
  Recorded Presentation: https://www.youtube.com/watch?v=fpUPCLNd59w

- **SCARIF**: SCARIF is a tool to estimate the embodied carbon emissions of data center servers with accelerator hardware, including GPUs, FPGAs, etc.
  GitHub: https://github.com/arc-research-lab/SCARIF
  Publication: SCARIF: Towards Carbon Modeling of Cloud Servers with Accelerators, ISVLSI 2024
  Demo: https://peipeizhou-eecs.github.io/publication/2024_isvlsi_scarif/
- **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