Peipei Zhou

(+86)13023157641, memoryzpp@cs.ucla.edu
http://vast.cs.ucla.edu/~peipei/
Enflame Technology, Shengxia Road, Pudong New Area, Shanghai, 201203 China

EDUCATION

University of California, Los Angeles Los Angeles, CA
Ph.D., Computer Science, GPA: 4.0/4.0 2014 – 2019/08

- Advisor: Professor Jason Cong
- Dissertation: "Modeling and Optimization for Customized Computing: Performance, Energy and Cost Perspective"

University of California, Los Angeles Los Angeles, CA
M.S., Electrical Engineering, GPA: 3.8/4.0 2012 - 2014

- Advisor: Professor Jason Cong
- Thesis: "A Fully Pipelined and Dynamically Composable Architecture of CGRA"

Southeast University, Chien-Shiung Wu Honors College Nanjing, Jiangsu, China
B.S., Electrical Engineering, GPA: 3.9/4.0 2008 - 2012

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
- **Best Poster Award**, High Throughput Sequencing (HiTSeq 2015)
- **Honeywell Innovator Scholarship**, one of five recipients in Mainland China, 2011
- **Enflame CEO Award,** 6 out of 300 employees in Enflame Technology, 2020

RESEARCH INTERESTS

- Computer architecture: especially for customized accelerator design [1][2][3][4][5][7][8][9], accelerator-rich architecture [10][13][14]
- Performance and energy characterization, modeling and optimization for mobile and embedded systems [9] and distributed systems [6][15]
- Hardware architecture and system optimization for high performance computing [2][9][12], health (precision medicine) [1][6][15] and machine learning (neural network) [3][8][9]

FULL PAPER PUBLICATIONS ([4][6][9][10][1] are the first-author paper)

[1] 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

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

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

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

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

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

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

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

[1] For [2][4][7][10], authors are listed in alphabetical order.

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

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

PREPRINTS & CONFERENCE POSTERS

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

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

[13] ARAPrototyper: Enabling Rapid Prototyping and Evaluation for Accelerator-Rich Architecture
Yu-Ting Chen, Jason Cong, Zhenman Fang, Bingjun Xiao, Peipei Zhou
arXiv:1610.09761 [cs.AR], October 2016

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

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

TALKS

- **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., December 2018

- **Mocha: Multinode Optimization of Cost in Heterogeneous Cloud with Accelerators (Technology Review Talk)**
  Falcon Computing Solutions Technology Review, Los Angeles, February 2019

- **Cost Optimization Engine for Heterogeneous Public Cloud Featuring Genomics Workloads (Technology Review Talk)**
  Falcon Computing Solutions Technology Review, Los Angeles, November 2018

- **Analyzing and Accelerating Genome Analysis Toolkit GATK4**
  CDSC/InTrans Project Annual Review, Los Angeles, June 2017
  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).

- **Quantifying the I/O Impact on Apache Spark: with Application to Computational Genomic**
- **Fully Pipeline or Not? An Energy Perspective**
  CDSC/InTrans Project Annual Review, Los Angeles, October 2016

- **Parallelization of CAVIAR and Identification of Causal Variants in Breast Cancer**
  CDSC/InTrans Project Annual Review, Los Angeles, January 2016

- **Prototype of A Fully Pipelined Configurable Array FPCA**
  CDSC Year 5 Semi-Annual Review, Los Angeles, May 2014

- **FPGA Prototyping of Accelerator-Rich Architectures**
  CDSC Year 4 Semi-Annual Review, Los Angeles, October 2013

POSTERS

- **Mocha: Multinode Optimization of Cost in Heterogeneous Cloud with Accelerators**
  CDSC/InTrans Project Annual Review, Los Angeles, February 2019

- **Modeling for Customized Computing: Performance, Energy and Cost Perspective**
  55th Annual Design Automation Conference PhD Forum, June 2018

- **Latte: Locality Aware Transformation for High-Level Synthesis**
  CDSC/InTrans Project Annual Review, Los Angeles, March 2018

- **Does I/O Still Matter in the Age of In-Memory Computation? Quantitative Analysis and Modeling of Computational Genomics**
- **AutoAccel: Automated Accelerator Generation and Optimization with Composable, Parallel and Pipeline Architecture**
  Center for Future Architecture Research (C-FAR) Annual Review, Ann Arbor, September 2017. C-FAR is one of six centers supported by the STARnet phase of the Focus Center Research Program (FCRP), a Semiconductor Research Corporation program sponsored by MARCO and DARPA. Participating universities are University of Michigan, Harvard University, Princeton University, University of California Los Angeles and etc.

- **On Sale Cost Optimization for GATK4 in Cloud**
  CDSC/InTrans Project Annual Review, Los Angeles, June 2017

- **Cloud Scale Acceleration of Fine-Mapping in Genetic Studies**
- **Fully Pipeline or Not? An Energy Perspective**
  CDSC/InTrans Project Annual Review, Los Angeles, January 2016

- **A Fully Pipelined and Dynamically Composable Architecture of CGRA**
  CDSC Year 5 Semi-Annual Review, Los Angeles, September 2014

ACADEMIC SERVICES

- **Technical Program Committee**, 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

TEACHING EXPERIENCE

- **Tutorial, Introduction to Vivado High-Level Synthesis, CS259 Customized Computing for Big-Data Applications**
  Computer Science Department, University of California, Los Angeles, 2017 Winter Quarter

- **Tutorial, Introduction to GATK4, CS259 Customized Computing for Big-Data Applications**
  Computer Science Department, University of California, Los Angeles, 2017 Winter Quarter

- **Teaching Assistant, EE110L Circuit Measurements Laboratory**
  Electrical Engineering Department, University of California, Los Angeles, 2013 Summer Quarter

- **Teaching Assistant, EE110 Circuit Analysis II**
  Electrical Engineering Department, University of California, Los Angeles, 2013 Fall Quarter, 2014 Spring Quarter

PROFESSIONAL EXPERIENCE

- **Research Scientist**, Enflame Technology, Shanghai, China — Aug. 2019 – Current
- **Software Engineer Intern**, Falcon Computing Solutions, Los Angeles, CA — June 2018 – Aug. 2019
- **Software Engineer Intern**, Microsoft, Redmond, WA — June 2017 – Sept. 2017
- **Research Intern**, Microsoft Research, Redmond, WA — June 2014 – Sept. 2014
- **Research Intern**, Honeywell Automation & Control, Nanjing, China. — July 2011 – Sept. 2011

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

REFERENCE

Professor Jingsheng Jason Cong
University of California, Los Angeles
Computer Science Department
468 Engineering VI, phase 2, Los Angeles, California 90095
Tel: +1-310- 206-2775
E-mail: cong@cs.ucla.edu

Professor Vivek Sarkar
Georgia Institute of Technology
College of Computing
Klaus Advanced Computing Building, Room 2332, 266 Ferst Dr NW Atlanta GA 30332
Tel: +1-404-385-5989
E-mail: vsarkar@gatech.edu

Professor Zhiru Zhang
Cornell University
School of Electrical and Computer Engineering

320 Rhodes Hall, Ithaca, NY 14853
Tel: +1-607-255-5954
E-mail: zhiruz@cornell.edu

Professor Zhenman Fang
Simon Fraser University
Computer Engineering Option School of Engineering Science
8888 University Drive, Burnaby BC, Canada, V5A 1S6
Tel: +1-778-782-4332
Email: zhenman@sfu.ca