SITAO HUANG

sitaohuang.com ⋄ +1 (217) 418-5166 ⋄ sitaoh@uci.edu
Assistant Professor, EECS, University of California Irvine
3225 Engineering Hall, Irvine, CA 92697, USA

EMPLOYMENT

University of California, Irvine, CA, USA *January 2022 – present*
Assistant Professor in the Department of Electrical Engineering and Computer Science

EDUCATION

University of Illinois at Urbana-Champaign, Urbana, IL, USA *August 2014 – August 2021*
M.S. (2017), Ph.D. (2021) in Electrical and Computer Engineering
Advisors: Prof. Deming Chen and Prof. Wen-mei Hwu.

Tsinghua University, Beijing, China *August 2010 – June 2014*
B.Eng. in Electronic Engineering

RESEARCH INTERESTS

- Hardware architecture and system optimization for high efficiency computing;
- Programming languages, compilers, and high-level synthesis for hardware accelerators;
- Design automation and optimizations for heterogeneous systems.

AWARDS AND HONORS

- **NSF CAREER Award**, 2025
- UCI Samueli School of Engineering CORCL Research Award, 2025
- UCI Samueli School of Engineering REVAMP Award, 2024
- UCI Samueli School of Engineering CORCL Research Award, 2024
- IPC Michael Carano Teacher Excellence Award, 2023
- **DARPA Riser**, Class of 2022
- Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2021) ***Best Paper Award***.
- IEEE/ACM Asia and South Pacific Design Automation Conference (ASP-DAC 2021) ***Best Paper Candidate***.
- IEEE HPEC Graph Challenge 2019 ***Honorable Mention***.
- IEEE/ACM Design Automation Conference 2019 (DAC 2019) System Design Contest ***First Place***.
- UIUC ECE 2019 ***Sundaram Seshu International Student Fellowship***.
- IEEE HPEC Graph Challenge 2018 ***Student Innovation Award***.
- Design Automation Conference 2018 (DAC 2018) System Design Contest ***Third Place***.
- UIUC ECE 2018 ***Rambus Computer Engineering Fellowship***.
- Tsinghua University Department of Electronic Engineering 2013 ***Academic Innovation Scholarship***.

WORK EXPERIENCE

- **University of California, Irvine**, EECS. *Assistant Professor.* *January 2022 – present*
- **Xilinx Research Labs**, Research Labs. *Compiler Intern.* *June 2020 - August 2020*
  Mentor: Dr. Stephen Neuendorffer
  Topic: MLIR-based compiler optimization passes for mapping arbitrary loops onto the AI engines in Xilinx Versal devices.
- **Microsoft Research**, AI Infrastructures. *Research Intern.* *May 2018 - August 2018*
  Mentor: Dr. Yuan Yu
  Topic: ONNX runtime and compiler infrastructures for FPGA platforms in the Microsoft Azure cloud.
- **Microsoft Research**, Deep Learning Group. *Research Intern.* *May 2017 - August 2017*
  Mentors: Dr. Po-Sen Huang, Dr. Chong Wang, Dr. Dengyong Zhou
  Topic: Accelerating the training of neural phrase-based language models using multiple GPUs.
- **Synopsys**, ZeBu Team. *Technical Intern.* *May 2016 - August 2016*
  Mentor: Dr. Lingyi Liu
  Topic: Compiler optimization passes for clock tree optimization in large-scale FPGA-based emulation system.
- **Microsoft Research Asia**, System Algorithm Group. *Research Intern.* *December 2013 - May 2014*
  Mentor: Dr. Thomas Moscibroda
  Topic: Design and optimization of scheduling algorithms in operating systems.
- **University of California, Los Angeles**, VAST Lab. *Research Intern.* *July 2013 - September 2013*
  Mentors: Prof. Jason Cong and Dr. Peng Zhang
  Topic: High-level synthesis and optimizations of dynamic time warping on FPGA systems.

SELECTED RESEARCH PROJECTS

CoVA: Compilatoin Flow for Versatile Architectures 2024–present

· Python-based compilation and optimizations for heterogeneous platforms with CPUs, GPUs, and FPGAs.

Energy-Efficient Binary Transformer Accelerator 2024–present

· Low-power FPGA-based accelerator for binary Transformer models delivering SOTA computational efficiency.

PyAIE: A Python-based Programming Framework for Versal ACAP AI Engines 2021–2023

· Python-based code generation and optimizations for AI engine accelerators in Versal ACAP platforms.
· Supports dataflow optimizations, vector optimizations, memory optimizations, etc.

PyLog: A High-Level Programming and Synthesis Flow for FPGAs 2020–2022

· Customized high-level parallel language designed to simpify FPGA programming.
· Separate algorithm specification and implementation details, automatic performance tuning

Optimizing DNN Inference on ReRAM Accelerators with Approximate Computing 2020

· The first work that proposes partial sum quantization with reduced ADC precision.
· Combines weight quantization, input quantization and partial sum quantization for ReRAM accelerators.
· Use learning methods to effectively search for optimal design point in the design space.

Accelerating Sparse DNN Inference with FPGAs 2019

· Sparse DNN inference engine built on FPGA using Vivado high-level synthesis (HLS).
· Won 2019 HPEC Graph Challenge ***Honorable Mention***.

Real-Time Object Detection with DNNs on Low-Power FPGAs and GPUs 2018

· Real-time high-accuracy object detection design on small FPGAs and GPUs for embedded scenarios.

· Won **1st *place*** in DAC 2019 System Design Contest.

Tangram: A High-Level Language for Heterogeneous Computing 2018

· Tangram is a high-level programming language and compiler that targets multi-core CPUs and GPUs.
· Given target hardware, Tangram automatically generates optimal code from high-level input code.

Collaborative Computing on CPU-FPGA and CPU-GPU Platforms 2018

· Formalize and model the CPU-accelerator collaboration patterns
· Optimize the workload distribution between CPU and accelerators

Hardware Acceleration of the Pair-HMM algorithm for DNA Variant Calling 2017

· Achieves state-of-the-art acceleration of Pair-HMM algorithm with PE-ring structures on FPGA

PUBLICATIONS

Book Chapters

[B1] **Compilation and Optimizations for Efficient Machine Learning on Embedded Systems**.
Xiaofan Zhang, Yao Chen, Cong Hao, ***Sitao Huang***, Yuhong Li, Deming Chen
Embedded Machine Learning for Cyber-Physical, IoT, and Edge Computing, Springer Nature. 2023.

Journal Papers

[J5] **A Sparsity-Aware Autonomous Path Planning Accelerator with HW/SW Co-Design and Multi-Level Dataflow Optimization**.
Yifan Zhang, Xiaoyu Niu, Hongzheng Tian, Yanjun Zhang, Bo Yu, Shaoshan Liu, ***Sitao Huang***.
ACM Transactions on Architecture and Code Optimization, volume 22, issue 3, pp. 1-25, 2025.

[J4] **HyperDetect: A Real-Time Hyperdimensional Solution for Intrusion Detection in IoT Networks**.
Junyao Wang, Haocheng Xu, Yonatan G. Achamyeleh, ***Sitao Huang***, Mohammad Al Faruque.
IEEE Internet of Things Journal, vol. 11, no. 8, pp. 14844-14856, 2024.

[J3] **Sgap: Towards Efficient Sparse Tensor Algebra Compilation for GPU**.
Genghan Zhang, Yuetong Zhao, Yanting Tao, Zhongming Yu, Guohao Dai, ***Sitao Huang***, Yuan Wen, Pavlos Petoumenos, Yu Wang.
CCF Transactions on High Performance Computing, **5**, pp. 210–227, 2023.

[J2] **PyLog: An Algorithm-Centric Python-Based FPGA Programming and Synthesis Flow**.
Sitao Huang, Kun Wu, Hyunmin Jeong, Chengyue Wang, Deming Chen, Wen-mei Hwu.
IEEE Transactions on Computers (TC), vol. 70, no. 12, pp. 2015–2028, 1 Dec. 2021.

[J1] **Large Graph Convolutional Network Training with GPU-Oriented Data Communication Architecture**.
Seung Won Min, Kun Wu, ***Sitao Huang***, Mert Hidayetoglu, Jinjun Xiong, Eiman Ebrahimi, Deming Chen, Wen-mei Hwu.
Proceedings of the VLDB Endowment, Volume 14, Issue 11, July 2021, pp. 2087–2100, 2021.

Conference Papers

[C31] **COBRA: Algorithm-Architecture Co-optimized Binary Transformer Accelerator for Edge Inference**.
(🏆 ***Best Paper Candidate***)
Ye Qiao, Zhiheng Chen, Yian Wang, Yifan Zhang, Yunzhe Deng, ***Sitao Huang***. *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2025.

[C30] **MONAS: Efficient Zero-Shot Neural Architecture Search for MCUs**.
Ye Qiao, Haocheng Xu, Yifan Zhang, ***Sitao Huang***. *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2025.

[C29] **RSEND: Retinex-based Squeeze and Excitation Network with Dark Region Detection for Efficient Low Light Image Enhancement**.
Jingcheng Li, Ye Qiao, Haocheng Xu, ***Sitao Huang***. *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2025.

[C28] **SSDTrain: An Activation Offloading Framework to SSDs for Faster Large Language Model Training**.
Kun Wu, Jeongmin Brian Park, Xiaofan Zhang, Mert Hidayetoğlu, Vikram Sharma Mailthody, ***Sitao Huang***, Steve Lumetta, Wen-mei Hwu. *ACM/IEEE Design Automation Conference (DAC)*, 2025.

[C27] **HeteroBench: Multi-kernel Benchmarks for Heterogeneous Systems**.
Hongzheng Tian, Alok Mishra, Zhiheng Chen, Rolando Pablo Hong Enriquez, Dejan Milojicic, Eitan Frachtenberg, ***Sitao Huang***. *ACM/SPEC International Conference on Performance Engineering (ICPE)*, 2025.

[C26] **A Sparsity-Aware Autonomous Path Planning Accelerator with Algorithm-Architecture Co-Design**.
Yanjun Zhang, Xiaoyu Niu, Yifan Zhang, Hongzheng Tian, Bo Yu, Shaoshan Liu, ***Sitao Huang***. *2024 International Conference on Computer-Aided Design (ICCAD)*, 2024.

[C25] **MicroNAS: Zero-Shot Neural Architecture Search for MCUs**. (short paper)
Ye Qiao, Haocheng Xu, Yifan Zhang, ***Sitao Huang***. *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2024.

[C24] **Hector: An Efficient Programming and Compilation Framework for Implementing Relational Graph Neural Networks in GPU Architectures**.
Kun Wu, Mert Hidayetoğlu, Xiang Song, ***Sitao Huang***, Da Zheng, Israt Nisa, and Wen-mei Hwu. *ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2024.

[C23] **CARMA: Context-Aware Runtime Reconfiguration for Energy-Efficient Sensor Fusion**.
Yifan Zhang, Arnav Vaibhav Malawade, Xiaofang Zhang, Yuhui Li, DongHwan Seong, Mohammad Abdullah Al Faruque, ***Sitao Huang***.
ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), 2023.

[C22] **DistHD: A Learner-Aware Dynamic Encoding Method for Hyperdimensional Classification**.
Junyao Wang, ***Sitao Huang***, Mohsen Imani.
IEEE/ACM Design Automation Conference (DAC), 2023.

[C21] **PyAIE: A Python-based Programming Framework for Versal ACAP Platforms**.
Hongzheng Tian, Shining Yang, Yoonha Cha, Jinwen Wu, Yuqi He, ***Sitao Huang***.
IEEE/ACM Design Automation Conference (DAC) Late Breaking Results, 2023.

[C20] **Scalable and Efficient Hyperdimensional Computing for Network Intrusion Detection**.
Junyao Wang, Hanning Chen, Mariam Issa, ***Sitao Huang***, Mohsen Imani.
IEEE/ACM Design Automation Conference (DAC) Late Breaking Results, 2023.

[C19] **Chimera: A Hybrid Machine Learning-Driven Multi-Objective Design Space Exploration Tool for FPGA High-Level Synthesis**.
(🏆 ***Best Paper Award***)
Mang Yu, ***Sitao Huang***, Deming Chen.
International Conference on Intelligent Data Engineering and Automated Learning (IDEAL), pp. 524–536, 2021.

[C18] **A Python-based High-Level Programming Flow for CPU-FPGA Heterogeneous Systems** (Invited Paper).
Sitao Huang, Kun Wu, Sai Rahul Chalamalasetti, Izzat El Hajj, Cong Xu, Paolo Faraboschi, Deming Chen.
IEEE/ACM Programming Environments for Heterogeneous Computing (PEHC), pp. 20–26, 2021.

[C17] **Improved GPU Implementations of the Pair-HMM Forward Algorithm for DNA Sequence Alignment**.
Enliang Li, Subho S Banerjee, ***Sitao Huang***, Ravishankar K Iyer, Deming Chen.
IEEE 39th International Conference on Computer Design (ICCD), pp. 299–306, 2021.

[C16] **Mind Mappings: Enabling Efficient Algorithm-Accelerator Mapping Space Search**.
Kartik Hegde, Po-An Tsai, ***Sitao Huang***, Vikas Chandra, Angshuman Parashar, Christopher Fletcher.
International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2021.

[C15] **Mixed Precision Quantization for ReRAM-based DNN Inference Accelerators**.
(🏆 ***Best Paper Candidate***)
Sitao Huang, Aayush Ankit, Plinio Silveira, Rodrigo Antunes, Sai Rahul Chalamalasetti, Izzat El Hajj, Dong-Eun Kim, Glaucimar Aguiar, Pedro Bruel, Sergey Serebryakov, Cong Xu, Can Li, Paolo Faraboschi, John Paul Strachan, Deming Chen, Kaushik Roy, Wen-mei Hwu, Dejan Milojicic.
The 26th Asia and South Pacific Design Automation Conference (ASP-DAC), 2021.

[C14] **Accelerating Sparse Deep Neural Networks on FPGAs**.
(🏆 ***Graph Challenge Honorable Mention***)
Sitao Huang, Carl Pearson, Rakesh Nagi, Jinjun Xiong, Deming Chen, Wen-mei Hwu.
IEEE High Performance Extreme Computing Conference (HPEC), 2019.

[C13] **Analysis and Optimization of I/O Cache Coherency Strategies for SoC-FPGA Device**.
Seung Won Min, ***Sitao Huang***, Mohamed Aly, Jinjun Xiong, Deming Chen, Wen-mei Hwu.
The International Conference on Field-Programmable Logic and Applications (FPL), 2019.

[C12] **Near-Memory and In-Storage FPGA Acceleration for Emerging Cognitive Computing Workloads**. (invited)
Ashutosh Dhar, ***Sitao Huang***, Jinjun Xiong, Damir Jamsek, Bruno Mesnet, Jian Huang, Nam Sung Kim, Wen-mei Hwu, Deming Chen.
IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2019.

[C11] **FPGA/DNN Co-Design: An Efficient Design Methodology for IoT Intelligence on the Edge**.
Cong Hao, Xianfan Zhang, Yuhong Li, ***Sitao Huang***, Jinjun Xiong, Kyle Rupnow, Wen-mei Hwu, Deming Chen.
The 56th Annual Design Automation Conference (DAC), 2019.

[C10] **Analysis and Modeling of Collaborative Execution Strategies for Heterogeneous CPU-FPGA Architectures**.
Sitao Huang, Li-Wen Chang, Izzat El Hajj, Simon Garcia de Gonzalo, Juan Gómez Luna, Sai Rahul Chalamalasetti, Mohamed El-Hadedy, Dejan Milojicic, Onur Mutlu, Deming Chen, Wen-mei Hwu.
The 10th ACM/SPEC International Conference on Performance Engineering (ICPE), 2019.

[C9] **Automatic Generation of Warp-Level Primitives and Atomic Instructions for Fast and Portable Parallel Reduction on GPUs**.
Simon Garcia De Gonzalo, ***Sitao Huang***, Juan Gómez-Luna, Simon Hammond, Onur Mutlu, Wen-mei Hwu.
The International Symposium on Code Generation and Optimization (CGO), 2019.

[C8] **Hardware-Software Co-Design for an Analog-Digital Accelerator for Machine Learning**.
Joao Ambrosi, Rodrigo Antunes, Aayush Ankit, Sai Rahul Chalamalasetti, Soumitra Chatterjee, Izzat El Hajj, Guilherme Fachini, Paolo Faraboschi, Martin Foltin, ***Sitao Huang***, Wen-mei Hwu, Gustavo Knuppe, Sunil Vishwanathpur Lakshminarasimha, Dejan Milojicic, Mohan Parthasarathy, Filipe Ribeiro, Lucas Rosa, Kaushik Roy, Plinio Silveira, John Paul Strachan.
IEEE International Conference on Rebooting Computing (ICRC), 2018.

[C7] **Triangle Counting and Truss Decomposition using FPGA**.
(🏆 ***Graph Challenge Student Innovation Award***)
Sitao Huang, Mohamed El-Hadedy, Cong Hao, Qin Li, Vikram S. Mailthody, Ketan Date, Jinjun Xiong, Deming Chen, Rakesh Nagi, Wen-mei Hwu.
IEEE High Performance Extreme Computing Conference (HPEC), 2018.

[C6] **Towards Neural Phrase-based Machine Translation**.
Po-Sen Huang, Chong Wang, ***Sitao Huang***, Dengyong Zhou, Li Deng.
The Sixth International Conference on Learning Representations (ICLR), 2018.

[C5] **Collaborative Computing for Heterogeneous Integrated Systems**.
Li-Wen Chang, Juan Gómez Luna, Izzat El Hajj, ***Sitao Huang***, Deming Chen, Wen-Mei Hwu.
The Eighth ACM/SPEC International Conference on Performance Engineering (ICPE), 2017.

[C4] **Hardware Acceleration of the Pair-HMM Algorithm for DNA Variant Calling**.
Sitao Huang, Gowthami Jayashri Manikandan, Anand Ramachandran, Kyle Rupnow, Wen-Mei Hwu, Deming Chen.
ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA), 2017.

[C3] **Accelerating Frequent Item Counting with FPGA**. (short paper)
Yuliang Sun, Zilong Wang, ***Sitao Huang***, Lanjun Wang, Yu Wang, Rong Luo, Huazhong Yang.
ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA), 2014.

[C2] **DTW-Based Subsequence Similarity Search on AMD Heterogeneous Computing Platform**.
Sitao Huang, Guohao Dai, Yuliang Sun, Zilong Wang, Yu Wang, Huazhong Yang.
International Conferences on High Performance Computing and Communications (HPCC), 2013.

[C1] **Accelerating Subsequence Similarity Search Based on Dynamic Time Warping Distance with FPGA**.
Zilong Wang, ***Sitao Huang***, Lanjun Wang, Hao Li, Yu Wang, Huazhong Yang.
ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA), 2013.

Workshop Papers, Preprints, Technical Reports

[W20] **Q-ROAR: Outlier-Aware Rescaling for RoPE Position Interpolation in Quantized Long-Context LLMs**.
Ye Qiao, Haocheng Xu, Xiaofan Zhang, ***Sitao Huang***. *arXiv preprint arXiv:2509.14391*, 2025.

[W19] **TeLLMe: An Energy-Efficient Ternary LLM Accelerator for Prefilling and Decoding on Edge FPGAs**.
Ye Qiao, Zhiheng Chen, Yifan Zhang, Yian Wang, ***Sitao Huang***. *arXiv preprint arXiv:2504.16266*, 2025.

[W18] **Optimized Spatial Architecture Mapping Flow for Transformer Accelerators**.
Haocheng Xu, Faraz Tahmasebi, Ye Qiao, Hongzheng Tian, Hyoukjun Kwon, ***Sitao Huang***.
arXiv preprint arXiv:2410.07407, 2024.

[W17] **MONAS: Efficient Zero-Shot Neural Architecture Search for MCUs**.
Ye Qiao, Haocheng Xu, Yifan Zhang, ***Sitao Huang***.
arXiv preprint arXiv:2408.15034, 2024.

[W16] **TBA: Faster Large Language Model Training Using SSD-Based Activation Offloading**.
Kun Wu, Jeongmin Brian Park, Xiaofan Zhang, Mert Hidayetoğlu, Vikram Sharma Mailthody, ***Sitao Huang***, Steven Sam Lumetta, Wen-mei Hwu.
arXiv preprint arXiv:2408.10013, 2024.

[W15] **Optimizing High-Level Synthesis Designs with Retrieval-Augmented Large Language Models**.
Haocheng Xu, Haotian Hu, ***Sitao Huang***.
2024 IEEE LLM Aided Design Workshop (LAD), 2024.

[W14] **RSEND: Retinex-based Squeeze and Excitation Network with Dark Region Detection for Efficient Low Light Image Enhancement**.
Jingcheng Li, Ye Qiao, Haocheng Xu, ***Sitao Huang***.
arXiv preprint arXiv:2406.09656, 2024.

[W13] **TG-NAS: Leveraging Zero-Cost Proxies with Transformer and Graph Convolution Networks for Efficient Neural Architecture Search**.
Ye Qiao, Haocheng Xu, ***Sitao Huang***.
arXiv preprint arXiv:2404.00271, 2024.

[W12] **Pigeon: Optimizing cuda code generator for end-to-end training and inference of relational graph neural networks**.

Kun Wu, Mert Hidayetoğlu, Xiang Song, ***Sitao Huang***, Da Zheng, Israt Nisa, and Wen-mei Hwu.
arXiv preprint arXiv:2301.06284, 2023.

[W11] **Sgap: Towards Efficient Sparse Tensor Algebra Compilation for GPU**.
Genghan Zhang, Yuetong Zhao, Yanting Tao, Zhongming Yu, Guohao Dai, ***Sitao Huang***, Yuan Wen, Pavlos Petoumenos, Yu Wang.
arXiv preprint arXiv:2209.02882, 2022.

[W10] **Compilation and Optimizations for Efficient Machine Learning on Embedded Systems**.
Xiaofan Zhang, Yao Chen, Cong Hao, ***Sitao Huang***, Yuhong Li, Deming Chen.
arXiv preprint arXiv:2206.03326, 2022.

[W9] **A Python-based High-Level Programming Flow for CPU-FPGA Heterogeneous Systems**.
Sitao Huang, Kun Wu, Sai Rahul Chalamalasetti, Izzat El Hajj, Cong Xu, Paolo Faraboschi, Deming Chen.
2021 IEEE/ACM Programming Environments for Heterogeneous Computing (PEHC@SC), 2021.

[W8] **Large Graph Convolutional Network Training with GPU-Oriented Data Communication Architecture**.
Seung Won Min, Kun Wu, ***Sitao Huang***, Mert Hidayetoğlu, Jinjun Xiong, Eiman Ebrahimi, Deming Chen, Wen-mei Hwu.
arXiv preprint arXiv:2103.03330, 2021.

[W7] **PyTorch-Direct: Enabling gpu centric data access for very large graph neural network training with irregular accesses**.
Seung Won Min, Kun Wu, ***Sitao Huang***, Mert Hidayetoğlu, Jinjun Xiong, Eiman Ebrahimi, Deming Chen, Wen-mei Hwu.
arXiv preprint arXiv:2101.07956, 2021.

[W6] **PyLog: An Algorithm-Centric Python-Based FPGA Programming and Synthesis Flow**.
Sitao Huang, Kun Wu, Chengyue Wang, Hyunmin Jeong, Anirudh Kashyap, Deming Chen, Wen-mei Hwu.
Semiconductor Research Corporation Technical Conference (SRC TECHCON), 2020.

[W5] **Design Space Exploration as a Service (DSEaaS)**.
Pedro Bruel, ***Sitao Huang***, Vipin K. Kukkala, Daniel Dauwe, Darel Emmot, Cong Xu,Rodrigo Antunes, Plinio Silveira, Gisani Sackser, Sai R. Chalamalasetti, Dejan Milojicic
Hewlett Packard Enterprise Technical Conference (HPE TechCon), 2020.

[W4] **Automatic Generation of Warp-Level Primitives and Atomic Instructions for Fast and Portable Parallel Reduction on GPUs**.
Simon Garcia De Gonzalo, ***Sitao Huang***, Juan Gómez-Luna, Simon Hammond, Onur Mutlu, Wen-mei Hwu.
Semiconductor Research Corporation Technical Conference (SRC TECHCON), 2019.

[W3] **Optimizing DNN Inference on Crossbars with Approximate Computing**.
Sitao Huang, Aayush Ankit, Rodrigo Antunes, Plinio Silveira, Glaucimar Aguiar, Izzat El Hajj, Martin Foltin, Dejan Milojicic.
Hewlett Packard Enterprise Technical Conference (HPE TechCon), 2019. (Oral Presentation, 2% Acceptance)

[W2] **Tangram: A Performance Portable Code Synthesis Framework**.
Simon Garcia De Gonzalo, ***Sitao Huang***, Li-Wen Chang, Izzat El Hajj, Christopher Rodrigues, Juan Gomez-Luna, Wen-mei Hwu.
Semiconductor Research Corporation Technical Conference (SRC TECHCON), 2018.

[W1] **Thoughts on massively-parallel heterogeneous computing for solving large problems** (invited paper).
Wen-mei Hwu, Mert Hidayetogglu, Weng Cho Chew, Carl Pearson, Simon Garcia, ***Sitao Huang***, Abdul Dakkak.
Computing and Electromagnetics International Workshop (CEM), 2017.

Dissertations

[D2] **High-Efficiency and High-Usability Heterogeneous Hardware Acceleration with FPGAs**
Sitao Huang. Ph.D. Dissertaion. University of Illinois at Urbana-Champaign, 2021.

[D1] **Hardware Acceleration of the Pair HMM Algorithm for DNA Variant Calling**
Sitao Huang. M.S. Thesis. University of Illinois at Urbana-Champaign, 2017.

Patents

[P1] **Sequence Modeling via Segmentations**.
US Patent US 2019/0266246 A1.
Chong Wang, Yining Wang, Po-Sen Huang, Abdelrahman Samir Abdelrahman Mohamed, Dengyong Zhou, Li Deng, ***Sitao Huang***

TALKS AND POSTERS

High-level Programming and Compilation for Heterogeneous Systems

· Department of Computer Science and Engineering, University of California Santa Cruz *Nov., 2024*

Optimizing High-Level Synthesis Designs with Retrieval-Augmented Large Language Models

· Leveraging Large Models and EDA for Mutual Advancement, Workshop at ICCAD 2024 *Oct., 2024*

Efficient Intelligent Computing via Automatic Search, Tuning, and Compilation

· Qing Yuan Research Institute, Shanghai Jiao Tong University *Sept., 2023*
· Department of Electronic Engineering, Tsinghua University *July, 2023*
· ROAD4NN workshop at DAC 2023 *July, 2023*

Towards High-Efficiency and High-Usability Hardware Acceleration

· ROTO High School Visit Day at UC Irvine *May, 2023*
· Department of Computer Science, University of California, Irvine *March, 2023*
· Department of Computer Science, University of California Los Angeles *Aug., 2022*
· School of Electronics Engineering and Computer Science, Peking University *June, 2022*

Accelerating Sparse Deep Neural Network on FPGAs

· Sparsity in ML Workshop at MLSys 2022 *Aug., 2022*
· IEEE High Performance Extreme Computing Conference (HPEC) *Sept., 2019*
· IBM AI Research Week and AI Horizons Colloquium *Sept., 2019*

PyLog: An Algorithm-Centric Python-Based FPGA Programming and Synthesis Flow

· Open-Source Source-to-Source Transformation for High-Level Synthesis Workshop at ISFPGA *Feb. 2022*
· Google Research, Los Angeles *Jan. 2022*
· The 12^{th} Joint Laboratory for Extreme Scale Computing (JLESC) Workshop *Feb. 2021*
· University of California, Santa Cruz Hardware Systems Collective (HSC) Seminar *Feb. 2021*
· Applications Driving Architectures Center (ADA) Liaison Meeting *Jan. 2021*
· Center for Research in Intelligent Storage and Processing in Memory (CRISP) Annual Review *Nov. 2020*
· Circuit IR Compilers and Tools (CIRCT) Weekly Meeting *Oct. 2020*
· Applications Driving Architectures Center (ADA) Liaison Meeting *Sept. 2020*
· Semiconductor Research Corporation Technical Conference (SRC TECHCON) *Sept. 2020*
· Applications Driving Architectures Center (ADA) Annual Review *May 2020*

Extending HLS with High-Level Descriptive Language for Configurable Algorithm-Level Spatial Structure Design

· 2021 IEEE 29th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), poster *May 2021*

Compiler and Hardware Design for In-Memory and Near-Memory Acceleration of Cognitive Applications

· Center for Research in Intelligent Storage and Processing in Memory (CRISP) Seminar *Feb. 2019*

Triangle Counting and Truss Decomposition using FPGA

· IEEE High Performance Extreme Computing Conference (HPEC) *Sept. 2018*
· IBM AI Research Week and AI Horizons Colloquium *Sept. 2018*

Collaborative Computing on Heterogeneous CPU-FPGA Systems

· IBM AI Research Week and AI Horizons Colloquium *Sept. 2017*

Acceleration of the Pair-HMM Algorithm for DNA Variant Calling

· International Symposium on Field-Programmable Custom Computing Machines (FCCM) *May 2016*

RESEARCH GRANTS

- NSF CAREER Award (single PI, Award Number 2443992), $650,000, 2025
- NSF CCF SHF Medium Award (co-PI, Award Number 2504809), $900,000 total (co-PI share $300,000), 2025
- Hewlett Packard Labs, $60,000 research fund, 2025
- UCI Samueli School of Engineering CORCL Research Award, $3,050, 2025
- UCI Samueli School of Engineering REVAMP Award, $4,000, 2024
- UCI Samueli School of Engineering CORCL Research Award, $3,100, 2024
- IPC Michael Carano Teacher Excellence Award, $1,000, 2023
- AMD-Xilinx Research Gift, $30,000 hardware and license donation, 2022

PROFESSIONAL SERVICES

- *Associate Editor* for ACM Transactions on Reconfigurable Technology and Systems (TRETS), 2024–present
- *Publication Chair* for IEEE International Symposium on Workload Characterization (IISWC), 2025
- *Co-chair* for DAC Early Career Workshop, 2025
- *Artifact Evaluation Co-chair* for ISFPGA conference, 2024–present
- *Technical Program Committee Member* for DAC, ICCAD, DATE, FPGA, FPT, CODES+ISSS
- *Reviewer* for journals: TCAD, TRETS, TACO, TECS, IEEE Internet Computing, Neural Processing Letters

STUDENTS SUPERVISED

Yifan Zhang	Ph.D. Student	2023 – present
Haocheng Xu	Ph.D. Student	2022 – present
Hongzheng Tian	Ph.D. Student	2022 – present
Ye Qiao	Ph.D. Student	2022 – present
Daniel Hsu	M.S. Student	2024 – present
Yian Wang	M.S. Student	2022 – present
Dhananjay Vijayaraghavan	M.S. Student	2022 – 2024
Bhavya Chunchu	M.S. Student	2023 – 2023
Karim Barada	M.S. Student	2022 – 2024
Jinwen Wu	M.S. Student	2022 – 2024
Zeqiang Zheng	M.S. Student	2022 – 2024
Shining Yang	M.S. Student	2022 – 2024
Xiaofang Zhang	M.S. Student	2022 – 2023
Micheal Al Mosharafieh	Undergraduate Student	2024 – present
Jiwon Youn	Undergraduate Student	2024 – present
Samarth Vilesh Shah	Undergraduate Student	2024 – present
Yuhui Li	Undergraduate Student	2022 – 2024
DongHwan Seong	Undergraduate Student	2022 – 2024
Jiawei Li	Undergraduate Student	2022 – 2023
Sai Sandeep Reddy Baddam	Undergraduate Student	2022 – 2023

TEACHING EXPERIENCE

- **(UCI) EECS 298: System-on-Chip (SoC) Design** (Fall 2025, Fall 2024, Fall 2023)
- **(UCI) EECS 221: Languages and Compilers for Hardware Accelerators** (W2025, W2024, W2023, W2022)
- **(UCI) ECPS 295: GPU Parallel Programming** (Fall 2024)
- **(UCI) ECPS 209: CPS Case Study: GPU Parallel Computing** (Fall 2023, Fall 2025)
- **(UCI) EECS 112: Organization of Digital Computers** (Spring 2023, Spring 2022)
- **(UCI) EECS 31L: Introduction to Digital Logic Laboratory** (Spring 2025, Spring 2024)
- **(UIUC) ECE 498 ICC: IoT and Cognitive Computing** (Head TA, Spring 2019)
- **(UIUC) ECE 527: System-on-Chip Design** (Head TA, Fall 2017, Fall 2018)
- **(UIUC) ECE 425: Introduction to VLSI Design** (TA, Fall 2016)
- **(UIUC) ECE 385: Digital Systems Laboratory** (TA, Spring 2017, Spring 2018)