SITAO HUANG

sitaohuang.com ⋄ +1 (217) 418-5166 ⋄ sitaoh@uci.edu
Assistant Professor, EECS, University of California Irvine
3215 Engineering Hall, Irvine, CA 92697, USA

POSITION

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.S. in Electronic Engineering
Advisors: Prof. Yu Wang and Prof. Rong Luo.

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

- **DARPA Riser**, Class of 2022
- IPC Michael Carano Teacher Excellence Award
- 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***.
- The 6$^{\text{th}}$ International Conference on Learning Representations (ICLR 2018) ***Travel Award***.
- Tsinghua University Department of Electronic Engineering 2013 ***Academic Innovation Scholarship***.
- The 28$^{\text{th}}$ National Competition in Physics for University Students (Non-Major) ***First Prize***.
- The 31$^{\text{st}}$ "Challenge Cup" Competition of Science and Technology in Tsinghua University ***Third Prize***.
- The 26$^{\text{th}}$ Chinese Physics Olympiad (CPhO) in Provinces ***First Prize***.

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

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

· 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 (*to appear*).

Journal Papers

[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

[C19] **Chimera: A Hybrid Machine Learning-Driven Multi-Objective Design Space Exploration Tool for FPGA High-Level Synthesis**.
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 paper).
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**.
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

[W10] **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.

[W9] **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.

[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.

Dissertation

[D1] **High-Efficiency and High-Usability Heterogeneous Hardware Acceleration with FPGAs**
Sitao Huang. University of Illinois at Urbana-Champaign, 2021.

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

Towards High-Efficiency and High-Usability Hardware Acceleration

- Department of Computer Science, University of California Los Angeles *Aug., 2022*
- School of Electronics Engineering and Computer Science, Peking University *June, 2022*

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

- Google Research, Los Angeles *Jan. 2022*
- The 12$^{\text{th}}$ 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*

Accelerating Sparse Deep Neural Network on FPGA

- IEEE High Performance Extreme Computing Conference (HPEC) *Sept. 2019*
- IBM AI Research Week and AI Horizons Colloquium *Sept. 2019*

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

- *Xilinx, Research Gift*, $30,000 hardware and license donation, 2022.
- *(more to be announced)*

PROFESSIONAL SERVICES

- *Technical Program Committee Member* for DATE 2023
- *Technical Program Committee Member* for FPT 2023
- *Technical Program Committee Member* for ICCAD 2022
- *Artifact Evaluation Committee Member* for ASPLOS 2021, 2023
- *Reviewer* for IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)
- *Reviewer* for ACM Transactions on Reconfigurable Technology and Systems (TRETS)
- *Reviewer* for ACM Transactions on Architecture and Code Optimization (TACO)
- *Reviewer* for ACM Transactions on Embedded Computing Systems (TECS)
- *Reviewer* for IEEE Internet Computing
- *Reviewer* for Springer Neural Processing Letters (NPL)
- *Reviewer* for Integration, the VLSI Journal
- *Reviewer* for conferences, DAC, ICCAD, ASP-DAC, FPGA, FCCM, ASPLOS, ISCA, MICRO, etc.
- *Session Chair* for the 15$^{\text{th}}$ IEEE High Performance Computing and Communications conference (HPCC)

STUDENTS SUPERVISED

Haocheng Xu	Ph.D. Student	Fall 2022 – present
Hongzheng Tian	Ph.D. Student	Fall 2022 – present
Ye Qiao	Ph.D. Student	Fall 2022 – present
Yoonha Cha	Ph.D. Student	Fall 2022 – present
Yifan Zhang	M.S. Student	Winter 2022 – present
Xiaofang Zhang	M.S. Student	Winter 2022 – present
Zeqiang Zheng	M.S. Student	Winter 2022 – present
Shining Yang	M.S. Student	Spring 2022 – present
Jinwen Wu	M.S. Student	Fall 2022 – present
Yuhui Li	Undergraduate Student	Winter 2022 – present
DongHwan Seong	Undergraduate Student	Fall 2022 – present
Jiawei Li	Undergraduate Student	Fall 2022 – present
Mingyu Tang	UCInspire Visiting Student	Summer 2022

TEACHING EXPERIENCE

- **(UCI) EECS 112: Organization of Digital Computers** (Instructor, Spring 2022)
- **(UCI) EECS 221: Languages and Compilers for Hardware Accelerators** (Instructor, Winter 2022)
- **(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)