SITAO HUANG

421 Coordinated Science Lab, 1308 W Main Street, Urbana, IL 61801, USA
sitaohuang.com ⋄ +1 (217) 417-4152 ⋄ shuang91@illinois.edu

EDUCATION

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

Tsinghua University, Beijing, China *August 2010 - June 2014*
B.S. in Electronic Engineering

INTERNSHIPS

Microsoft Research, AI Infrastructures. *Research Intern.* *May 2018 - August 2018*
Microsoft Research, Deep Learning Group. *Research Intern.* *May 2017 - August 2017*
Synopsys, ZeBu Team. *Technical Intern.* *May 2016 - August 2016*
Microsoft Research Asia, System Algorithm Group. *Research Intern.* *December 2013 - May 2014*

AWARDS

IEEE HPEC Graph Challenge 2019 ***Honorable Mention***.
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 6th 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***.

PUBLICATIONS

Accelerating Sparse Deep Neural Networks on FPGAs. ***HPEC 2019*** (***Graph Challenge Honorable Mention***). ***Sitao Huang***, Carl Pearson, Rakesh Nagi, Jinjun Xiong, Deming Chen, Wen-mei Hwu.

Near-Memory and In-Storage FPGA Acceleration for Emerging Cognitive Computing Workloads. ***ISVLSI 2019***. Ashutosh Dhar, ***Sitao Huang***, Jinjun Xiong, Damir Jamsek, Bruno Mesnet, Jian Huang, Nam Sung Kim, Wen-mei Hwu, Deming Chen.

Analysis and Optimization of I/O Cache Coherency Strategies for SoC-FPGA Device. ***FPL 2019***. Seung Won Min, ***Sitao Huang***, Mohamed Aly, Jinjun Xiong, Deming Chen, Wen-mei Hwu.

FPGA/DNN Co-Design: An Efficient Design Methodology for IoT Intelligence on the Edge. ***DAC 2019***. Cong Hao, Xianfan Zhang, Yuhong Li, ***Sitao Huang***, Jinjun Xiong, Kyle Rupnow, Wen-mei Hwu, Deming Chen.

Analysis and Modeling of Collaborative Execution Strategies for Heterogeneous CPU-FPGA Architectures. ***ICPE 2019***. ***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.

Automatic Generation of Warp-Level Primitives and Atomic Operations for Fast-Portable GPU Reductions. ***CGO 2019***. Simon Garcia De Gonzalo, ***Sitao Huang***, Juan Gómez-Luna, Simon Hammond, Onur Mutlu, Wen-mei Hwu.

Triangle Counting and Truss Decomposition using FPGA. ***HPEC 2018*** (***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.

Hardware-Software Co-Design for an Analog-Digital Accelerator for Machine Learning. ***ICRC 2018***. 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, *et al.*

Towards Neural Phrase-based Machine Translation. ***ICLR 2018***. Po-Sen Huang, Chong Wang, ***Sitao Huang***, Dengyong Zhou, Li Deng.

Collaborative Computing for Heterogeneous Integrated Systems. ***ICPE 2017***. Li-Wen Chang, Juan Gómez Luna, Izzat El Hajj, ***Sitao Huang***, Deming Chen, Wen-Mei Hwu.

Hardware Acceleration of the Pair-HMM Algorithm for DNA Variant Calling. ***FPGA 2017***. ***Sitao Huang***, Gowthami Jayashri Manikandan, Anand Ramachandran, Kyle Rupnow, Wen-Mei Hwu, Deming Chen.

Accelerating Frequent Item Counting with FPGA. ***FPGA 2014***. Yuliang Sun, Zilong Wang, ***Sitao Huang***, Lanjun Wang, Yu Wang, Rong Luo, Huazhong Yang.

DTW-Based Subsequence Similarity Search on AMD Heterogeneous Computing Platform. ***HPCC 2013***. ***Sitao Huang***, Guohao Dai, Yuliang Sun, Zilong Wang, Yu Wang, Huazhong Yang.

Accelerating Subsequence Similarity Search Based on Dynamic Time Warping Distance with FPGA. ***FPGA 2013***. Zilong Wang, ***Sitao Huang***, Lanjun Wang, Hao Li, Yu Wang, Huazhong Yang.

PATENT

Sequence Modeling via Segmentations. *US Patent US15/903,942.*
Chong Wang, Yining Wang, Po-Sen Huang, Abdelrahman Samir Abdelrahman Mohamed, Dengyong Zhou, Li Deng, ***Sitao Huang***

SERVICES

Rewiwer for IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)
Rewiwer for ACM Transactions on Reconfigurable Technology and Systems (TRETS)
External Reviewer for ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA)
Session Chair for the 15$^{\text{th}}$ IEEE High Performance Computing and Communications conference (HPCC)

TEACHING

ECE 498 ICC: IoT and Cognitive Computing (Spring 2019, ***head TA***)
ECE 527: System-on-Chip Design (Fall 2017, Fall 2018, ***head TA***)
ECE 425: Introduction to VLSI Design (Fall 2016)
ECE 385: Digital Systems Laboratory (Spring 2017, Spring 2018)

SELECTED RESEARCH PROJECTS

High-Level Programming Language for FPGAs 2019

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

Accelerating Sparse DNN Inference with FPGAs 2019

- Sparse DNN inference engine built on FPGA using Vivado high-level synthesis (HLS).

· Targets large and deep DNNs with many fully connected layers.
· 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

TECHNICAL STRENGTHS

Programming Languages	C/C++, Python, Verilog, CUDA, OpenCL, MATLAB, Java, OCaml, etc.
Frameworks & Softwares	LLVM, LaTeX, Vivado, Quartus, ModelSim, etc.

SELECTED COURSES

Advanced Compiler Construction (A+); SoC Design (A); Computer Architecture (A); Manycore Parallel Algorithms (A); Algorithms (A); Programming Languages and Compilers (A); Random Processes (A); Distributed Algorithms (A-); MDPs, Reinforcement Learning (A-).