Jian Weng

404 Westwood Plaza, EVI 468 – 90095, Los Angeles, CA, USA
✉ jian.weng@ucla.edu • were.github.io

Bio

The goal of my research is to enable accessible, affordable, and diversified hardware acceleration by revolutionizing the full-stack reconfigurable accelerator design.

The exponential speedup of computers in the past decades has driven the flourishing of Computer Science, but it also imposes unsustainability on both hardware vendors and customers: Hardware innovations are monopolized by several extremly large hardware vendors by owning large teams, and customers have to deprecate their devices yearly for performance improvements. To prolong the chip lifetime and democratized hardware innovations, my research interests span designing specialized hardware mechanisms for program behaviors of interests, software/hardware abstraction, and their associated compilation techniques. By unleasing the synergies across these aspects, my work develops a full-stack framework for reconfigurable accelerator. This work has attracted computer architecture researchers from 7 universities across the world to contribute to democratized accelerator design.

I am a sixth-year PhD Candidate at UCLA. My work has been accepted by multiple top-tier conferences in computer architecture area, including ISCA, HPCA, and MICRO. My works have been awarded as IEEE Micro Top Picks, Horable Mentions, as well as Best Paper Runner-up.

Education

University of California, Los Angeles — **Los Angeles, CA**
6th year Ph.D Candidate, Advisor: Tony Nowatzki — *Aug. 2017 – Present*

Shanghai Jiao Tong University — **Shanghai, China**
B.Eng. of Computer Science, ACM Honored Class — *May 2013 – July 2017*
Thesis: Compiling Domain Specific Language to Reconfigurable Accelerators

Selected Awards

MICRO 2022 Best Paper Runner-up	*Oct. 2022*
UCLA Disseration Year Fellowship	*Sep. 2022*
QualComm Fellowship Finalist	*Aug. 2022*
Facebook Fellowship Finalist	*April 2022*
UCLA Graduate Student Fellowship	*Nov. 2017, Mar. 2021*
IEEE Micro Top Picks Honorable Mention	*2021*
HPCA Best Paper Runner-up	*2021*
IEEE Micro Top Picks	*2020*

Conference Publications

- OverGen: Improving FPGA Usability through Domain-specific Overlay Generation.
  Sihao Liu=, **Jian Weng=**, Dylan Kupsh, Atefeh Sohrabizadeh, Zhengrong Wang, Licheng Guo, Jiuyang Liu, Maxim Zhulin, Lucheng Zhang, Jason Cong, Tony Nowatzki.
  International Symposium on Microarchitecture (MICRO), 2022, **Best Paper Runner-up**
- Near-Stream Computing: General and Transparent Near-Cache Acceleration.
  Zhengrong Wang, **Jian Weng**, Sihao Liu, Tony Nowatzki.
  International Symposium on High-Performance Computer Architecture (HPCA), 2022
- UNIT: Unifying Tensorized Instruction Compilation.
  Jian Weng, Animesh Jain, Jie Wang, Leyuan Wang, Yida Wang, and Tony Nowatzki.
  International Symposium on Code Generation and Optimization (CGO), 2021
- Stream Floating: Enabling Proactive and Decentralized Cache Optimizations.
  Zhengrong Wang, **Jian Weng**, Jason Lowe-Power, Jayesh Gaur, Tony Nowatzki. *International Symposium on High-Performance Computer Architecture (HPCA), 2021*, **Best Paper Runner-up**

- DSAGEN: Synthesizing Programmable Spatial Accelerators.
  Jian Weng=, Sihao Liu=, Vidushi Dadu , Zhengrong Wang , Preyas Shah, and Tony Nowatzki.
  International Symposium on Computer Architecture (ISCA), 2020, **IEEE Micro Honorable Mentions**
- A Hybrid Systolic-Dataflow Architecture for Inductive Matrix Algorithms.
  Jian Weng, Sihao Liu, Zhengrong Wang, Vidushi Dadu, and Tony Nowatzki.
  International Symposium on High-Performance Computer Architecture (HPCA), 2020
- Towards General Purpose Acceleration by Exploiting Common Data-Dependence Forms.
  Vidushi Dadu, **Jian Weng**, Sihao Liu, and Tony Nowatzki.
  International Symposium on Microarchitecture (MICRO), 2019 **IEEE Micro Top Picks**
- Hybrid Optimization/Heuristic Instruction Scheduling for Programmable Accelerator Codesign.
  Tony Nowatzki, Newsha Ardalani, Karthikeyan Sankaralingam, and **Jian Weng**.
  Parallel Architectures and Compilation Techniques (PACT), 2018

Journal Publications

- Unifying Spatial Accelerator Compilation with Idiomatic and Modular Transformations.
  Jian Weng, Sihao Liu, Dylan Kupsh, Tony Nowatzki.
  IEEE Micro Special Issue on Compiling for Accelerators, 2022
- DAEGEN: A Modular Compiler for Exploring Decoupled Access Execte Accelerators.
  Jian Weng, Sihao Liu, Vidushi Dadu, and Tony Nowatzki.
  Computer Architecture Letters (CAL), 2019

Workshops and Tutorials

- OverGen: Decoupled-Spatial Architecture Framework
  Sihao Liu, **Jian Weng**, Dylan Kupsh, and Tony Nowatzki.
  International Symposium on Microarchitecture (@MICRO'22)
- A Full-Stack Infrastructure for Automating Spatial Architecture Research
  Jian Weng, Sihao Liu, Dylan Kupsh, and Tony Nowatzki.
  Workshop on Democratizing Domain-Specific Accelerators (@MICRO'22)
- Generality is the Key Dimension in Accelerator Design
  Jian Weng, Sihao Liu, Vidushi Dadu, and Tony Nowatzki.
  Workshop on Languages, Tools, and Techniques for Accelerator Design (@ASPLOS'21)
- DSAGEN: Democratizing Decoupled Spatial Architecture Research
  Jian Weng, Sihao Liu, Vidushi Dadu, and Tony Nowatzki.
  International Symposium on Microarchitecture (@MICRO'20)

Invited Talks

Synthesizing Programmable Accelerators: A Full-Stack Perspective

- University of Illinois, Urbana Champaign, ILLIXR Consortium — Nov. 9th, 2022
- Institute of Computing Technology, Chinese Acadamy of Sciences — Nov. 19th, 2022
- Georgia Institute of Technology, EIC Lab — Nov. 23rd, 2022

Synthesizing Programmable Accelerators: A Compiler Perspective

- University of Michigan, Computer Engineering Lab — Feb. 23rd, 2022
- University of Illinois, Urbana Champaign, CAP Seminar — Feb. 8th, 2022
- University of Washington, SAMPL Research Group — Jan. 27th, 2022

Unit: Unifying Tensorized Instruction Compilation

- TVM Conference — Dec. 2nd, 2020

Hybrid Script: A Text Format for Halide IR & A Python-TVM Hybrid Frontend

- TVM Conference — Dec. 12th, 2018

Open-Source Projects and Infrastructures

Democratizing Spatial Accelerator Design ↗
Co-first Author *Aug. 2017 – Now*

- Develop the full-stack infrastructure for spatial architecture design.

TVM ↗ **Apache**
Committer *Aug. 2017 – Now*

- Implement the `tvm.te.hybrid` module to enhance the expressiveness of the framework.
- Write tutorials on performance tuning and customized compilation pass.

Professional Experiences

Amazon Web Service **East Palo Alto**
Applied Scientist Intern, Mentor: Yida Wang *June 2019 – Sep. 2019, Jan. 2020 – Sep. 2020*

- Compile the newly emerging tensorized instructions to diverse hardware platforms (CGO 2021).
- Automatically integrate cuDNN to a next-generation machine learning framework ↗ .

Teaching

CS33: Introduction to Computer Organization **University of California, Los Angeles**
Teaching Assistant, w/ Prof. Glenn Reinman *Mar. 24th — Jun. 11th, 2021*

MS109: Compiler Design and Implementation **Shanghai Jiao Tong University**
Teaching Assistant, w/ Prof. Yong Yu *Mar. 22nd — Jun. 26th, 2016*

MS105: Data Structure **Shanghai Jiao Tong University**
Teaching Assistant, w/ Prof. Huiyu Weng *Sep. 14th, 2015 — Jan. 17th, 2016*

Research Mentoring

Rishabh Mani **University of California, Los Angeles**
Develop memory access pattern analysis; coauthor of OverGen. *Mar. 23rd, 2022 — Now*

Dylan Kupsh ↗ **University of California, Los Angeles**
Study spatial architecture mapping techniques; coauthor of OverGen. *Sep. 20th, 2021 — Now*

Siyuan Ma **University of Texas, Austin**
Study in-memory processing; submitted to ISCA'23. *Mar. 23rd, 2022 — Now*

Shuo Wang **University of California, Los Angeles**
Study concurrent application mapping to DSAGEN. *Mar. 23rd, 2022 — Now*

Community Services

DAC 2022: Invited Reviewer

MICRO 2022: Student Reviewer