

ZHENGRONG WANG

404 Westwood Plaza, EVI 468 – 90095, Los Angeles, CA, USA

seanzw@ucla.edu <https://seanzw.github.io> Google Scholar

BIO

My research aims to build general, automatic and end-to-end near-data acceleration by revolutionizing the orchestration between data and computation throughout the entire system.

Conventional von Neumann architectures draw a clear boundary between computation and data, in which centralized compute units process the data provided by memory units. However, two forces dramatically changed the landscape: 1. Emerging modern applications, e.g. large language models, graph neural networks, recommend systems, etc., scale rapidly with the data size (e.g. GPT-4 has 170 trillion parameters), putting extremely high pressure on the memory system; 2. The widening gap between the compute and memory throughput (known as the memory wall), as well as the upscaling in system size together make the data movement an increasingly bottleneck. To continue the performance and energy efficiency scaling, my research takes the data as a first-class citizen in system design and spans across extensive aspects of data/computation orchestration including microarchitecture designs, ISA abstractions, compiler optimizations, and codesigning data structures.

I am currently a sixth-year PhD candidate at UCLA. My open source work has been accepted by multiple top-tier conferences in computer architecture, including ISCA, MICRO, ASPLOS, HPCA, and awarded Best Paper Runner-Ups as well as IEEE Micro Top Pick Honorable Mentions. I am also a maintainer of gem5, a widely used cycle accurate simulator in computer architecture.

EDUCATION

University of California, Los Angeles, Department of Computer Science Los Angeles, USA
Ph.D. Candidate in Computer Science, Advisor: Tony Nowatzki Aug. 2018 - Jul. 2024 (Expected)

University of California, Los Angeles, Department of Computer Science Los Angeles, USA
Master of Science in Computer Science Sep. 2016 - Jul. 2018
Thesis: An LLVM-IR Datagraph-Based Simulator for Flexible Design Space Exploration over Accelerator Architectures

Tsinghua University, Department of Electronic Engineering Beijing, China
Bachelor of Engineering in Electronic Engineering, GPA: 91/100 Aug. 2012 - Jul. 2016
Thesis: Optimizing Convolutional Neural Network on FPGA under Heterogeneous Computing Framework with OpenCL

ETH Zürich, Department of Information Technology Zürich, Switzerland
Exchange Student, International Academic Program, GPA: 5.50/ 6.00 Sept. 2014 - Feb. 2015

PUBLICATION

Infinity Stream: Portable and Programmer-Friendly In-/Near-Memory Fusion
Zhengrong Wang, Christopher Liu, Aman Arora, Lizy John, Tony Nowatzki
ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2023, Vancouver, Canada.

Infinity Stream: Enabling Transparent and Automated In-Memory Computing
Zhengrong Wang, Christopher Liu, Tony Nowatzki
IEEE Computer Architecture Letters, Vol. 21, No. 2, 2022.

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
IEEE/ACM International Symposium on Microarchitecture (MICRO), 2022, Chicago, USA.

Best Paper Runner-Up

Near-Stream Computing: General and Transparent Near-Cache Acceleration
Zhengrong Wang, Jian Weng, Sihao Liu, Tony Nowatzki
IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2022, Seoul, South Korea.

Stream Floating: Enabling Proactive and Decentralized Cache Optimizations
Zhengrong Wang, Jian Weng, Jason Lowe-Power, Jayesh Gaur, Tony Nowatzki
IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2021, Seoul, South Korea.

Best Paper Runner-Up

DSAGEN: Synthesizing Programmable Spatial Accelerators

Jian Weng, Sihao Liu, Vidushi Dadu, **Zhengrong Wang**, Preyas Shah, Tony Nowatzki
ACM International Symposium on Computer Architecture (ISCA), 2020, virtual.

IEEE Micro Top Picks Honorable Mention

A Hybrid Systolic-Dataflow Architecture for Inductive Matrix Algorithms

Jian Weng, Sihao Liu, **Zhengrong Wang**, Vidush Dadu, Tony Nowatzki

IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2020, San Diego, USA.

Stream-Based Memory Access Specialization for General Purpose Processors

Zhengrong Wang, Tony Nowatzki

ACM International Symposium on Computer Architecture (ISCA), 2019, Phoenix, USA.

The Gem5 Simulator: Version 20.0+

Jason Lowe-Power, Abdul Mutaal Ahmad, Ayaz Akram, ..., **Zhengrong Wang**, et al.

arXiv:2007.03152v2, 2020.

Optimizing Convolutional Neural Network on FPGA under Heterogeneous Computing Framework with OpenCL


Zhengrong Wang, Fei Qiao, Zhen Liu, Yuxiang Shan, Xunyi Zhou, Li Luo, Huazhong Yang

IEEE Region 10 Conference (TENCON), 2016, Singapore.

AWARDS AND HONORS

Dissertation Year Fellowship, <i>UCLA</i>	June. 2023
Best Paper Runner-Up (OverGen, in <i>MICRO '22</i>), <i>IEEE</i>	Oct. 2022
Best Paper Runner-Up (Stream Floating, in <i>HPCA '21</i>), <i>IEEE</i>	Feb. 2021
IEEE Micro Top Picks 2020 Honorable Mention (DSAGEN, in <i>ISCA '20</i>), <i>IEEE</i>	Jan. 2021
2021 Dongguan Entrepreneur Scholarship, <i>Dongguan Entrepreneurs Federation</i>	Nov. 2021
Second-class Scholarship for Excellent Freshmen, <i>Tsinghua University</i>	Oct. 2012
Wang Zhaosheng Scholarship for Excellent Student from Dongguan, <i>Wang Zhaosheng Foundation</i>	Oct. 2012

OPEN SOURCE PROJECTS & INFRASTRUCTURES

Stream-Specialized Near-Data Acceleration Framework 	Jan. 2018 - Present
First Author & Maintainer	

- Full-stack implementation of stream-specialized near-data acceleration.
- Include LLVM-based compiler transformation and end-to-end simulation in gem5.
- Results published in ISCA' 19, HPCA' 21, HPCA' 22 and ASPLOS '23. More in submission.

Gem5-AVX 	Jan. 2019 - Present
First Author & Maintainer	

- Add AVX-512 support to gem5 simulator, extensively used in research.
- Faithfully model the microarchitecture of vectorized instructions, including microops.
- Detailed tutorials on how to support new instructions.

Gem5 Simulator 	Jan. 2019 - Present
Committer & Maintainer	

- Contribute various bug fixes for instruction decoding, microarchitecture deadlock.
- Review pull requests.

PROFESSIONAL EXPERIENCES

Nvidia Research	Jun. 2017 - Sep. 2017
Research Scientist, Mentor: Neal Crago, Manager: Steve Keckler	

- Examine memory bottleneck in GPU for key machine learning kernels.
- Build a prototype of an enhanced tensor memory accelerator (TMA).
- Evaluted with state of the art point cloud applications.

TEACHING

CS33: Introduction to Computer Organization

Mar. 2022 - June. 2022

Teaching Assistant w/ Prof. Glenn Reinman

- Lead two-hour discussion every week, lab grading.
- Overall evaluation: 8/9.
- “Very knowledgeable, helpful, and encouraging. He ensured that I understood the content.”

RESEARCH MENTORING

Christopher Liu

CS PhD Student, UCLA

Develop the compiler support for Infinity Stream (Published in ASPLOS '23)

Dec. 2021 - Present

Nick Makaha

CS Undergraduate, UCLA

Improve stream support for high-performance join.

Jun. 2023 - Sep. 2023

Arteen Abrishami

CS Undergraduate, UCLA

Explore near-data acceleration on chiplet architectures.

Jan. 2023 - Now

Shyandeep Das

EE Master, UCLA

Support reuse for near-stream computing. Now Graduate Research Architect at ARM.

Jan. 2023 - Jun. 2023

COMMUNITY SERVICES

ISCA '23 uArch Workshop

Jun. 2023

Mentor for 4 talented undergraduate computer architects: Shujuan Chen (UC Davis), Viansa Schmulbach (UC Berkeley), Eden Alem (Washington University in St. Louis) and Evan Cheng (Stanford).

HPCA '21 Student Reviewer