Xianwei ZHANG

✉ zhangxw79@mail.sysu.edu.cn

Associate Professor, CSE, Sun Yat-sen University ☞ xianweiz.github.io ✍ 422 NSCC-gz, Guangzhou, China 510006

Research Interests

◇ Computer architecture and system

◇ High-performance computing, Intelligent computing

◇ GPU, Compilation, Memory, Hardware and software co-design

Employment

Sun Yat-sen University Guangzhou, China
`Associate Professor`, School of Computer Science and Engineering 2020.10 - present

AMD® Inc. Seattle/Austin, USA
`Researcher / Engineer`, AMD Research / RTG 2017.08 - 2020.09

NVIDIA® Corporation Austin, USA
`Research Intern`, NVIDIA Research 2016.05 - 2016.08

Education

Ph.D. in Computer Science 2011.08 - 2017.08
University of Pittsburgh, Pittsburgh, USA

- `Thesis`: *"Addressing Prolonged Restore Challenges in Further Scaling DRAMs"*
- `Advisors`: *Youtao Zhang*, *Bruce R. Childers* and *Jun Yang*

B.E. in Software Engineering 2007.09 - 2011.07
Northwestern Polytechnical University, Xi'an, China

Honors & Awards

Sci&Tech Young Talent Program *China-CAST*'2022
- foster the next generation of science and technology think tank professionals

Special Prize for Scientific and Technological Progress *GD-gov*'2022
- in recognition of outstanding contributions to scientific and technological innovation

AMD® Spotlight Award *AMD*'2019
- recognize individuals of extraordinary achievements and significant contributions

Andrew Mellon Fellowship *University of Pittsburgh*'2016
- awarded to Phd students of exceptional achievement and promise

Best Paper Award *ISLPED*'2013
- out of 167 submissions, based on the rating of anonymous reviewers and a panel of judges

Grants

[NSFC] `Award#: 62472462` `National Science Foundation of China`, 2025.01-2028.12
Project: Research on application driven fine-grained GPU resource management optimizations.
Role: Principal Investigator

[MOST] `Award#: 2023YFB3002202` `National Key R&D Program of China`, 2023.12-2026.11
Project: Global storage architecture and data resource management for supercomputing internet.
Role: Principal Investigator

[NSFC] `Award#: 62102465` `National Science Foundation of China`, 2022.01-2024.12
Project: Optimizing GPU memory management with hardware and software co-designs.
Role: Principal Investigator

[Huawei] `Award#: CCF-HuaweiSY202409` `Populus Grove Fund®`, 2024.09-2025.09
Project: Research on cross-platform multi-target container image compilation and construction technology for high-performance computing.
Role: Principal Investigator

[Tencent] `Award#: CCF-TencentRAGR20240102` `Rhino-Bird Open Research Fund®`, 2024.10-2025.12
Project: Research on LLM long-sequence pipeline parallel inference for weakly connected devices.
Role: Principal Investigator

[Phytium] `Award#: CCF-Phytium202204` `Phytium Fund®`, 2022.11-2024.07
Project: Loop unrolling compilation based on machine learning.
Role: Principal Investigator

[NSFC] Award#: 62461146204 — National Science Foundation of China, 2024.08-2029.07
Project: Data sharing infrastructure across Northeast-Asia supercomputing centers for open science.
Role: Key Member

Research — Patent / Tutorial / Publication

Patents

[P4] *X. Zhang*, J. Kalamatianos and B. Beckmann — US 11,487,671 B2
- GPU Cache Management based on Lightweight Locality Type Detection.

[P3] M. Seyedzadeh, *X. Zhang*, B. Beckmann and S. Das — US 7,714,747 B2
- Base Value Sharing in Data Compression Algorithms.

[P2] S. Puthoor, K. Punniya O. Kayiran, *X. Zhang*, Y. Eckert, J. Alsop and B. Beckmann — US 11,507,522 B2
- A Memory Request Priority Assigning Technique for GPUs.

[P1] A. Gutierrez, S. Blagodurov, S. Moe, *X. Zhang*, J.Yin, M. Sinclair — US 11,150,899 B2
- Selecting a Precision Level for Executing a Workload in an Electronic Device.

Tutorial

[T1] A. Gutierrez, *X. Zhang*, T. Ta and B. Beckmann — ISCA'2018
- AMD gem5 APU Simulator: Modeling GPUs using the Machine ISA.

Publications

Note: supervised student — Links: Google Scholar, DBLP, ORCID

◇ Conference

[C23] Xuanteng Huang, Jiangsu Du, Nong Xiao and *Xianwei Zhang*,
- PaSK: Cold Start Mitigation for Inference with Proactive and Selective Kernel Loading on GPUs, The 62nd ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, United States, June 2025.

[C22] Kan Wu, Zejia Lin, Mengyue Xi, Zhongchun Zheng, Wenxuan Pan, *Xianwei Zhang* and Yutong Lu,
- GoPTX: Fine-grained GPU Kernel Fusion by PTX-level Instruction Flow Weaving, The 62nd ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, United States, June 2025.

[C21] Yuhao Gu, Chunyu Chen, Jiangsu Du, Xiaoxi Zhang and *Xianwei Zhang*,
- ORFA: Exploring WebAssembly as a Turing Complete Query Language for Web APIs, The ACM Web Conference (WWW), Sydney, Australia, April 2025.

[C20] Mengyue Xi, Tianyu Guo, Xuanteng Huang, Zejia Lin and *Xianwei Zhang*,
- Mpache: Interaction Aware Multi-level Cache Bypassing on GPUs, The 30th Asia and South Pacific Design Automation Conference (ASP-DAC), Tokyo Odaiba Miraikan, Japan, January 2025.

[C19] Tianyu Guo, Xuanteng Huang, Kan Wu, *Xianwei Zhang* and Nong Xiao,
- SMILE: LLC-based Shared Memory Expansion to Improve GPU Thread Level Parallelism, The 61st ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, United States, June 2024.

[C18] Yuanxin Wei, Jiangsu Du, Jiangzhi Jiang, Xiao Shi, *Xianwei Zhang*, Dan Huang, Nong Xiao and Yutong Lu,
- APTMoE: Affinity-aware Pipeline Tuning for MoE Models on Bandwidth-constrained GPU Nodes, The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), Atlanta, GA, United States, November 2024.

[C17] Zejia Lin, Aoyuan Sun, *Xianwei Zhang* and Yutong Lu,
- MixPert: Optimizing Mixed-precision Floating-point Emulation on GPU Integer Tensor Cores, The 25th ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems (LCTES), Copenhagen, Denmark, June 2024.

[C16] Zhaowen Shan, Xuanteng Huang, Zheng Zhou and *Xianwei Zhang*,
- openLG: A Tunable and Efficient Open-source LSTM on GPUs, The International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, June 2024.

[C15] Zhongchun Zheng, Yuan Wu and *Xianwei Zhang*,
- mLOOP: Optimize Loop Unrolling in Compilation with a ML-based Approach, The 17th International Conference on Networking, Architecture, and Storage (NAS), Guangzhou, China, November 2024.

[C14] Zejia Lin, Zewei Mo, Xuanteng Huang, *Xianwei Zhang* and Yutong Lu,
- KeSCo: Compiler-based Kernel Scheduling for Multi-task GPU Applications, The IEEE 41st International Conference on Computer Design (ICCD), Washington DC, United States, November 2023.

[C13] Tianao Ge, Zewei Mo, Kan Wu, *Xianwei Zhang* and Yutong Lu,
- RollBin: Reducing Code-size via Loop Rerolling at Binary Level, The 23rd ACM SIGPLAN /SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems (LCTES), San Diego, California, United States, June 2022.

[C12] Zewei Mo, Zejia Lin, *Xianwei Zhang* and Yutong Lu,
- moTuner: A Compiler-based Auto-tuning Approach for Mixed-precision Operators, The 19th ACM International Conference on Computing Frontiers (CF), Turin, Italy, May 2022.

[C11] Yue Weng, Tianao Ge, Xi Zhang, *Xianwei Zhang* and Yutong Lu,
- RAISE: Efficient GPU Resource Management via Hybrid Scheduling, The 22nd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGrid), Taormina (Messina), Italy, May 2022.

[C10] Tuan Ta, *Xianwei Zhang*, Anthony Gutierrez and Brad Beckmann,
- Autonomous Data-Race-Free GPU Testing, IEEE International Symposium on Workload Characterization (IISWC), Orlando, Florida, United States, November 2019.

[C9] *Xianwei Zhang*, Rujia Wang, Youtao Zhang and Jun Yang,
- Boosting Chipkill Capability under Retention Error Induced Reliability Emergency, The 24th Asia and South Pacific Design Automation Conference (ASP-DAC), Tokyo, Japan, January 2019.

[C8] Anthony Gutierrez, Bradford M. Beckmann, Alexandru Dutu, Joseph Gross, Michael LeBeane, John Kalamatianos, Onur Kayiran, Matthew Poremba, Brandon Potter, Sooraj Puthoor, Matthew D. Sinclair, Mark Wyse, Jieming Yin, *Xianwei Zhang*, Akshay Jain and Timothy Rogers,
- Lost in Abstraction: Pitfalls of Analyzing GPUs at the Intermediate Language Level, The 24th IEEE International Symposium on High-Performance Computer Architecture (HPCA), Vienna, Austria, February 2018.

[C7] *Xianwei Zhang*, Youtao Zhang, Bruce Childers and Jun Yang,
- DrMP: Mixed Precision-aware DRAM for High Performance Approximate and Precise Computing, The 26th International Conference on Parallel Architectures and Compilation Techniques (PACT), Portland, Oregon, September 2017.

[C6] *Xianwei Zhang*, Youtao Zhang, Bruce Childers and Jun Yang,
- Restore Truncation for Performance Improvement in Future DRAM Systems, The 22nd IEEE Symposium on High Performance Computer Architecture (HPCA), Barcelona, Spain, March 2016.

[C5] *Xianwei Zhang*, Youtao Zhang, Bruce Childers and Jun Yang,
- Exploiting DRAM Restore Time Variations in Deep Sub-micron Scaling, The IEEE Conference on Design, Automation and Test in Europe (DATE), Grenoble, France, March 2015.

[C4] *Xianwei Zhang*, Youtao Zhang and Jun Yang,
- DLB: Dynamic Lane Borrowing for Improving Bandwidth and Performance in Hybrid Memory Cube, The 33rd IEEE International Conference on Computer Design (ICCD), New York City, New York, United States, October 2015.

[C3] *Xianwei Zhang*, Youtao Zhang and Jun Yang,
- TriState-SET: Proactive SET for Improved Performance in MLC Phase Change Memories, The 33rd IEEE International Conference on Computer Design (ICCD), New York City, New York, United States, October 2015.

[C2] *Xianwei Zhang*, Lei Zhao, Youtao Zhang and Jun Yang,
- Exploit Common Source-Line to Construct Energy Efficient Domain Wall Memory based Caches, The 33rd IEEE International Conference on Computer Design (ICCD), New York City, New York, United States, October 2015.

[C1] *Xianwei Zhang*, Lei Jiang, Youtao Zhang, Chuanjun Zhang and Jun Yang,
- WoM-SET: Lowering Write Power of Proactive-SET based PCM Write Strategy Using WoM Code, The International Symposium on Low Power Electronics and Design (ISLPED), Beijing, China, September 2013.

◇ Journal

[J5] Hengzhong Liang, Han Huang and *Xianwei Zhang*,
- SuCL: Supply Unified Communication Layer to Improve SYCL-based Heterogeneous Computing, CCF Transactions on High Performance Computing (THPC), 2025.

[J4] Xuanteng Huang, *Xianwei Zhang*, Panfei Yang and Nong Xiao,
- Benchmarking GPU Tensor Cores on General Matrix Multiplication Kernels through CUTLASS, Applied Sciences, 13(24), 13022, December 2023.

[J3] Xi Zhang, Xiaohu Guo, Yue Weng, *Xianwei Zhang*, Yutong Lu and Zhong Zhao,
- Hybrid MPI and CUDA Paralleled Finite Volume Unstructured CFD Simulations on a Multi-GPU System, Future Generation Computer Systems (FGCS), Volume 139, Issue C, 2023.

[J2] Yue Weng, Xi Zhang, Xiaohu Guo, *Xianwei Zhang*, Yutong Lu and Yang Liu,
- Effects of Mesh Loop Modes on Performance of Unstructured Finite Volume GPU Simulations, Advances in Aerodynamics (AIA), 3(21), 2021.

[J1] *Xianwei Zhang*, Youtao Zhang, Bruce Childers and Jun Yang,
- On the Restore Time Variations of Future DRAM Memory, ACM Transaction on Design Automation of Electronic Systems (TODAES), 22(2), 2017.

◇ Short/WIP

[W7] Tianyi Zhang, Guanyi Chen, Wenxuan Pan, Zhongchun Zheng, Gaojin Sun and *Xianwei Zhang*,
- REFIT: Improve Code Efficiency via Binary Level Loop Optimization, The 10th International Conference on Computer and Communication Systems (ICCCS), Chengdu, China, April 2025.

[W6] Chunyu Chen, Haoquan Chen, Yunhao Han, Yuhao Gu and *Xianwei Zhang*,
- CWL-Bubble: Extending CWL for Dynamic Scientific Workflows, The 10th International Conference on Computer and Communication Systems (ICCCS), Chengdu, China, April 2025.

[W5] Lianghong Huang, Zejia Lin, Wei Liu and *Xianwei Zhang*,
- Hay: Enhancing GPU Sharing Performance with Two-Level Scheduling for Ray, The 29th IEEE International Conference on Parallel and Distributed Systems (ICPADS), Hainan, China, December 2023.

[W4] *Xianwei Zhang* and Evgeny Shcherbakov,
- DELTA: Validate GPU Memory Profiling with Microbenchmarks, The International Symposium on Memory Systems (MEMSYS), Washington DC, United States, October 2020.

[W3] Johnathan Alsop, Matthew D. Sinclair, Srikant Bharadwaj, Alexandru Dutu, Anthony Gutierrez, Onur Kayiran, Michael LeBeane, Brandon Potter, Sooraj Puthoor, *Xianwei Zhang*, Tsung Tai Yeh and Bradford M. Beckmann,
- Optimizing GPU Cache Policies for MI Workloads, IEEE International Symposium on Workload Characterization (IISWC), Orlando, Florida, United States, November 2019.

[W2] *Xianwei Zhang*, Youtao Zhang, Bruce Childers and Jun Yang,
- AWARD: Approximation-aWAre Restore in Further Scaling DRAM, The International Symposium on Memory Systems (MEMSYS), Washington DC, United States, October 2016.

[W1] *Xianwei Zhang*, Youtao Zhang and Jun Yang,
- Adaptive Lane Borrowing of Hybrid Memory Cube, The 52nd ACM/IEEE Design Automation Conference (DAC), San Francisco, California, United States, June 2015.

Talks

Workshop of A3 Foresight Program — *Dec 2024*, Guangzhou, China
[T15] - Computing Environment Support for HPC Application Porting and Deploying

CCF China Storage — *Nov 2024*, Guangzhou, China
[T14] - System Construction and Application Support based on Fused Data Space

Huawei® Connect — *Sep 2024*, Shanghai, China
[T13] - Compilation-based GPU Computing Acceleration and Heterogeneous Support

International Supercomputing Conference (ISC) — *May 2024*, Hamburg, Germany
[T12] - Scalable Data Processing and Management for Converged High Performance and Intelligent Computing

Southern University of Science and Technology — *Apr 2024*, Shenzhen, China

[T11] - GPU-based Computing and Cases of Software/Hardware Optimizations

The Hong Kong University of Science & Technology (Guangzhou) *Oct 2023*, Guangzhou, China

[T10] - GPU-based Computing and Hardware/Software Optimizations

HPC China *Sep 2023*, Qingdao, China

[T9] - Compilation-based GPU Mixed-precision and Task Parallel Computing Optimization

China National Computer Congress (CNCC) *Oct 2020*, Beijing, China

[T8] - GPU Exploration and Optimization towards Next Generation Computing

Department of Energy/AMD® *Oct 2019*, Austin, USA

[T7] - Improving Reuse and Reducing Overheads in GPU Cache Hierarchies

Alibaba® DAMO Academy *Feb 2019*, Sunnyvale, USA

[T6] - Architectural Studies and Modeling of Memory and GPU Systems.

AMD® Research *Apr 2018*, Bellevue, USA

[T5] - Improve High-performance Computing and Deep Learning via GPU Memory System Optimizations.

NVIDIA® Research *Aug 2016*, Austin, USA

[T4] - Understanding and Mitigating the Impact of Long-latency Memory Systems on GPUs.

HPCA Symposium *Mar 2016*, Barcelona, Spain
Swanson School of Engineering *Apr 2016*, Pittsburgh, USA

[T3] - Restore Truncation for Performance Improvement in Future DRAM Systems.

MEMSYS Symposium *Oct 2016/2015*, Washington DC, USA

[T2] - Mitigate Restore Issues in Further Scaling DRAM (Refresh-based and Approx-based).

[T1] - Achieving Yield, Density and Performance Effective DRAM at Extreme Technology Sizes.

Service

Program Committee

- TPC, CCGrid'2025 (IEEE Int'l Sym. on Cluster, Cloud, and Internet Computing)
- TPC, IJCNN'2025 (Int'l Joint Conf. on Neural Networks)
- TPC, NAS'2024 (IEEE Int'l Conf. on Networking, Architecture, and Storage)
- TPC, NPC'2024 (IFIP Int'l Conf. on Network and Parallel Computing)
- TPC, HiPC'2024/2023/2022 (IEEE Int'l Conf. on HPC, Data, Analytics, and Data Science)
- TPC, ICPADS'2022 (IEEE Int'l Conf. on Parallel and Distributed Systems)
- TPC, PECS'2022 (Int'l Congress on Power, Energy, and Computer Systems)
- ERC, MICRO'2020 (IEEE/ACM Int'l Sym. on Microarchitecture)
- TPC, ICCD'2020/2019/2018 (IEEE Int'l Conf. on Computer Design)

Organizer

- Co-chair, CNCC'2023 - Seminar ("Efficient Large-scale Computing")

Teaching

◇ Courses

Instructor *CSE*, Sun Yat-sen University

- *DCS290* - Compilation Principle (Ug), 2021-2024 Spring.
- *DCS292* - Compiler Construction (Ug), 2021-2025 Spring.
- *DCS3013* - Computer Architecture (Ug), 2022 Fall.
- *DCS5637/6207* - Advanced Computer Architecture (Gr), 2021-2024 Fall.

◇ Service

Technical Committee
2024 NSCSCC-Compiler Design Competition
- *”National Student Computer System Capability Challenge - Compiler Design Competition, Huawei Bisheng Cup”*

◇ Awards

First Prize of Best Teaching Case (2024)
China Computer Education Conference
- *”LLVM Compilation Practice Teaching based on a Developer-friendly Experience”*

First Prize of Compiler Competition (Mentor, 2023)
NSCSCC - Compiler Design Competition
- *”Yat-CC: Self-developed SYsY Language Compiler for RISC-V and ARM CPUs”*

First Prize of Teaching Award (2023)
Sun Yat-sen University
- *”Multicore Parallelism: Practice in Building an Elite Talent Training System for Domestic Computing Ecosystems”*

First Prize of Outstanding Teaching Paper (2022)
China Computer Education Conference
- *”Building a Holistic View of Compilation Practice based on Clang/LLVM Infrastructure”*

Mentoring

◇ Current Members (Note: co-advise$^{\diamond}$)

P.h.D.
- Class of 2025: Xianjie Chen, Mingen Liang
- Class of 2023: Han Huang$^{\diamond}$
- Class of 2022: Xuanteng Huang$^{\diamond}$, Yuhao Gu$^{\diamond}$, Zejia Lin$^{\diamond}$, Tianyu Guo$^{\diamond}$
- Class of 2021: Kan Wu$^{\diamond}$

MS
- Class of 2025: Yunhao Han, Junru Chen, Xin Huang
- Class of 2024: Hongxin Xu, Tengyang Zheng$^{\diamond}$, Gaojin Sun, Lu Wu, Jingyi He, Bingjie Liu
- Class of 2023: Mengyue Xi, Wenyuan Liang, Hengzhong Liang, Wenxuan Pan, Aoyuan Sun, Zhongchun Zheng
- Class of 2022: Chun-yu Chen, Tianyi Zhang, Zhaowen Shan

Ug/RA
- Guanyi Chen, Zheng Zhou, Haoquan Chen, Yipeng Ouyang

◇ Alumni

- Yinchuan Guo (MS, 2021-2024, First placement: *R&D Engr* @ Huawei)
- Lianghong Huang (MS, 2021-2024, First placement: *R&D Engr* @ MetaX)
- Yue Weng (MS, 2020-2023, First placement: *R&D Engr* @ NVIDIA)
- Tianao Ge (MS, 2020-2022, First placement: *PhD* @ HKUST-gz)
- Zewei Mo (MS, 2020-2022, First placement: *R&D Engr* @ Intel)

Misc

GoogleScholar:	https://scholar.google.com/citations?user=k9_kXbQAAAAJ&hl=en
DBLP:	https://dblp.org/pid/135/8227-1.html
ORCID:	https://orcid.org/0000-0003-3507-4299
Github:	https://github.com/arcsysu
Yat Compiler:	https://yatcc.github.io
Linkedin:	https://www.linkedin.com/in/xianweizhang/
Homepage:	https://xianweiz.github.io

(Last updated on 02/2025)