Cong Guo

Visiting Research Scientist, Meta · Menlo Park, CA, USA
congguo.ai@gmail.com · guocong.me · Google Scholar

Research Interests

Computer architecture and hardware–software co-design for efficient, scalable AI systems:

- Hardware-efficient model compression, such as sparsity and quantization.
- Novel accelerator architectures for large language models.
- Systems and infrastructure for large-scale AI serving.

Appointments

08/2026 – Present **Meta** Menlo Park, CA, USA
Visiting Research Scientist, AI and Systems Co-Design

12/2023 – 08/2026 **Duke University** Durham, NC, USA
Postdoctoral Associate, Department of Electrical and Computer Engineering
Mentors: Prof. Yiran Chen and Prof. Hai (Helen) Li

Education

09/2020 – 09/2023 **Shanghai Jiao Tong University** Shanghai, China
Ph.D. in Computer Science
Dissertation: *Software–Hardware Co-Design for Tiny Artificial Intelligence*
Advisor: Prof. Jingwen Leng
Awards: Outstanding Doctoral Dissertation Award (Top 1%, 2024); Outstanding Doctoral Graduate (2024); Ph.D. National Scholarship (2020)

09/2017 – 03/2020 **Shanghai Jiao Tong University** Shanghai, China
M.E. in Computer Technology
Advisor: Prof. Jingwen Leng

09/2012 – 06/2016 **Shenzhen University** Shenzhen, China
B.S. in Computer Science

Honors and Awards

03/2026 **ASPLOS Best Paper Nomination**
ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2026)

01/2026 **HPCA Best Paper Nomination**
IEEE International Symposium on High-Performance Computer Architecture (HPCA 2026); 4 nominees among 119 accepted papers

10/2025 **Nominee for the 2025 Outstanding Postdoctoral Award** (24 nominees university-wide)
Duke University

03/2024 **Outstanding Doctoral Dissertation Award** (Top 1%)
Shanghai Jiao Tong University

03/2024	**Outstanding Doctoral Graduate** *Shanghai Jiao Tong University*
07/2023	**IEEE Micro Top Picks from 2022 Computer Architecture Conferences Honorable Mention** *IEEE Micro* [link]
08/2022	**Excellent Ph.D. Scholarship of Yang Yuanqing Education Fund** (Top 3 of 500+) *Shanghai Jiao Tong University*
11/2020	**Ph.D. National Scholarship** (Top 8 of 500+) *Ministry of Education (China)*
07/2020	**DAC 2020 Richard Newton Young Student Fellow** *Design Automation Conference*

Publications

38 peer-reviewed papers, including 20 at ISCA×6, MICRO×2, HPCA×8, and ASPLOS×4.

* Corresponding author = Equal contribution Underline: student mentored by me

Conference Papers

[C1] **Cong Guo**=*, Xuanzeng Song=, Bowen Duan, Haoxuan Shan, Chiyue Wei, Yuzhe Fu, Adnan Aziz, Hai Li, Yiran Chen. "Duet: Harmonizing Compute and Memory Demands for LLM Decoding." In *IEEE International Symposium on High-Performance Computer Architecture* (**HPCA**), 2027.

[C2] Baijun Zhou, Yintao He, Chenyi Wang, **Cong Guo**, Feng Cheng, Hai Li, Yiran Chen. "DyCache: An Adaptive Context-State Cache for Dynamic Sparse LLM Inference." In *IEEE International Symposium on High-Performance Computer Architecture* (**HPCA**), 2027.

[C3] Weiming Hu, Zihan Zhang, Zihan Liu, Yangjie Zhou, Yu Feng, Shixuan Sun, Haoyan Zhang, **Cong Guo**, Jinyang Liu, Ning Liu, Jing Yang, Yun Lin, Jin Song Dong, Minyi Guo, Jingwen Leng. "Emulating Tensor-Core Numerics for Reproducible Low-Bit LLM Inference." In *IEEE International Symposium on High-Performance Computer Architecture* (**HPCA**), 2027.

[C4] Haoxuan Shan=, **Cong Guo**=*, Bowen Duan, Chiyue Wei, Feng Cheng, Yuzhe Fu, Yintao He, Hai Li, Yiran Chen. "Vortex: Bridging Extreme Compression and Efficient LLM Inference." In *ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (**ASPLOS**), 2027.

[C5] Feng Cheng, **Cong Guo***, Junyao Zhang, Haoxuan Shan, Chiyue Wei, Hong Wang, Hai Li, Yiran Chen. "Gossamer: A Utility-Driven Architecture for Constant-Budget KV Cache Compression in Reasoning LLMs." In *IEEE/ACM International Symposium on Microarchitecture* (**MICRO**), 2026.

[C6] Bowen Duan=, **Cong Guo**=*, Chiyue Wei, Haoxuan Shan, Yuzhe Fu, Xinhua Chen, Yifan Xu, Ziyue Zhang, Changchun Zhou, Hai Li, Yiran Chen. "EVA: Accelerating LLM Decoding via an Efficient Vector Quantization Architecture." In *ACM/IEEE International Symposium on Computer Architecture* (**ISCA**), 2026.

[C7] Chiyue Wei=, **Cong Guo**=*, Junyao Zhang, Haoxuan Shan, Yifan Xu, Ziyue Zhang, Yudong Liu, Qinsi Wang, Changchun Zhou, Hai Li, Yiran Chen. "Focus: A Streaming Concentration Architecture for Efficient Vision-Language Models." In *IEEE International Symposium on*

High-Performance Computer Architecture (**HPCA**), 2026. **(Best Paper Nomination)**

[C8] Yuzhe Fu, Changchun Zhou*, Hancheng Ye, Bowen Duan, Qiyu Huang, Chiyue Wei, **Cong Guo***, Hai Li, Yiran Chen. "FractalCloud: A Fractal-Inspired Architecture for Efficient Large-Scale Point Cloud Processing." In *IEEE International Symposium on High-Performance Computer Architecture* (**HPCA**), 2026.

[C9] Weiming Hu, Zihan Zhang, Haoyan Zhang, Chen Zhang, **Cong Guo**, Yu Feng, Tianchi Hu, Guanglin Li, Guipeng Hu, Junsong Wang, Jingwen Leng. "M2XFP: A Metadata-Augmented Microscaling Data Format for Efficient Low-bit Quantization." In *ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (**ASPLOS**), 2026. **(Best Paper Nomination)**

[C10] Xinhua Chen=, Sitao Huang=, **Cong Guo**=*, Chiyue Wei, Yintao He, Jianyi Zhang, Hai Li, Yiran Chen. "DPad: Efficient Diffusion Language Models with Suffix Dropout." In *International Conference on Learning Representations* (**ICLR**), 2026.

[C11] Haoxuan Shan, **Cong Guo***, Chiyue Wei, Feng Cheng, Junyao Zhang, Hai Li, Yiran Chen. "Platinum: Path-Adaptable LUT-Based Accelerator Tailored for Low-Bit Weight Matrix Multiplication." In *Asia and South Pacific Design Automation Conference* (**ASP-DAC**), 2026.

[C12] Yuzhe Fu, Hancheng Ye, **Cong Guo***, Junyao Zhang, Qinsi Wang, Yueqian Lin, Changchun Zhou, Hai Li, Yiran Chen. "FlashFPS: Efficient Farthest Point Sampling for Large-Scale Point Clouds via Pruning and Caching." In *ACM/IEEE Design Automation Conference* (**DAC**), 2026.

[C13] Jiale Xu=, Rui Zhang=, Yi Xiong=, **Cong Guo***, Zihan Liu, Yangjie Zhou, Weiming Hu, Hao Wu, Changxu Shao, Ziqing Wang, Yongjie Yuan, Junping Zhao*, Minyi Guo, Jingwen Leng*. "eLLM: Elastic Memory Management Framework for Efficient LLM Serving." In *ACM/IEEE Design Automation Conference* (**DAC**), 2026.

[C14] **Cong Guo***, Chiyue Wei, Jiaming Tang, Bowen Duan, Song Han, Hai Li, Yiran Chen. "Transitive Array: An Efficient GEMM Accelerator with Result Reuse." In *ACM/IEEE International Symposium on Computer Architecture* (**ISCA**), 2025.

[C15] Chiyue Wei, Bowen Duan, **Cong Guo***, Jingyang Zhang, Qingyue Song, Hai Li, Yiran Chen. "Phi: Leveraging Pattern-based Hierarchical Sparsity for High-Efficiency Spiking Neural Networks." In *ACM/IEEE International Symposium on Computer Architecture* (**ISCA**), 2025.

[C16] Feng Cheng, **Cong Guo***, Chiyue Wei, Junyao Zhang, Changchun Zhou, Edward Hanson, Jiaqi Zhang, Xiaoxiao Liu, Hai Li, Yiran Chen. "Ecco: Improving Memory Bandwidth and Capacity for LLMs via Entropy-aware Cache Compression." In *ACM/IEEE International Symposium on Computer Architecture* (**ISCA**), 2025.

[C17] Chiyue Wei, **Cong Guo***, Feng Cheng, Shiyu Li, Hao Yang, Hai Li, Yiran Chen. "Prosperity: Accelerating Spiking Neural Networks via Product Sparsity." In *IEEE International Symposium on High-Performance Computer Architecture* (**HPCA**), 2025.

[C18] Weiming Hu, Haoyan Zhang, **Cong Guo**, Yu Feng, Renyang Guan, Zhendong Hua, Zihan Liu, Yue Guan, Minyi Guo, Jingwen Leng. "MANT: Efficient Low-bit Group Quantization for LLMs via Mathematically Adaptive Numerical Type." In *IEEE International Symposium on High-Performance Computer Architecture* (**HPCA**), 2025.

[C19] Zihan Liu, Xinhao Luo, Junxian Guo, Wentao Ni, Yangjie Zhou, Yue Guan, **Cong Guo**, Weihao Cui, Yu Feng, Minyi Guo, Yuhao Zhu, Minjia Zhang, Jingwen Leng, Chen Jin. "VQ-LLM: High-performance Code Generation for Vector Quantization Augmented LLM Inference." In *IEEE International Symposium on High-Performance Computer Architecture* (**HPCA**), 2025.

[C20] Yangjie Zhou, Honglin Zhu, Qian Qiu, Weihao Cui, Zihan Liu, Peng Chen, Mohamed Wahib, **Cong Guo**, Siyuan Feng, Jintao Meng, Haidong Lan, Jingwen Leng, Yun Lin, Jin Song Dong, Wenxi Zhu, Minwen Deng. "A Sample-Free Compilation Framework for Efficient Dynamic Tensor Computation." In *International Conference for High Performance Computing, Networking, Storage and Analysis* (**SC**), 2025.

[C21] Linshen Liu, Boyan Su, Junyue Jiang, Guanlin Wu, **Cong Guo**, Ceyu Xu, Hao Frank Yang. "Towards Accurate and Efficient 3D Object Detection for Autonomous Driving: A Mixture of Experts Computing System on Edge." In *IEEE/CVF International Conference on Computer Vision* (**ICCV**), 2025.

[C22] **Cong Guo**, Rui Zhang, Jiale Xu, Jingwen Leng, Zihan Liu, Ziyu Huang, Minyi Guo, Hao Wu, Shouren Zhao, Junping Zhao, Ke Zhang. "GMLake: Efficient and Transparent GPU Memory Defragmentation for Large-scale DNN Training with Virtual Memory Stitching." In *ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (**ASPLOS**), 2024.

[C23] Zihan Liu, Wentao Ni, Jingwen Leng, Yu Feng, **Cong Guo**, Quan Chen, Chao Li, Minyi Guo, Yuhao Zhu. "JUNO: Optimizing High-Dimensional Approximate Nearest Neighbour Search with Sparsity-Aware Algorithm and Ray-Tracing Core Mapping." In *ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (**ASPLOS**), 2024.

[C24] **Cong Guo**[=], Jiaming Tang[=], Weiming Hu, Jingwen Leng, Chen Zhang, Fan Yang, Yunxin Liu, Minyi Guo, Yuhao Zhu. "OliVe: Accelerating Large Language Models via Hardware-Friendly Outlier-Victim Pair Quantization." In *ACM/IEEE International Symposium on Computer Architecture* (**ISCA**), 2023.

[C25] Yangjie Zhou, Yaoxu Song, Jingwen Leng, Zihan Liu, Weihao Cui, Zhendong Zhang, **Cong Guo**, Quan Chen, Li Li, Minyi Guo. "AdaptGear: Accelerating GNN Training via Adaptive Subgraph-Level Kernels on GPUs." In *ACM International Conference on Computing Frontiers* (**CF**), 2023.

[C26] **Cong Guo**, Chen Zhang, Jingwen Leng, Zihan Liu, Fan Yang, Yunxin Liu, Minyi Guo, Yuhao Zhu. "ANT: Exploiting Adaptive Numerical Data Type for Low-bit Deep Neural Network Quantization." In *IEEE/ACM International Symposium on Microarchitecture* (**MICRO**), 2022. **(IEEE Micro Top Picks Honorable Mention)**

[C27] **Cong Guo**, Yuxian Qiu, Jingwen Leng, Xiaotian Gao, Chen Zhang, Yunxin Liu, Fan Yang, Yuhao Zhu, Minyi Guo. "SQuant: On-the-Fly Data-Free Quantization via Diagonal Hessian Approximation." In *International Conference on Learning Representations* (**ICLR**), 2022.

[C28] **Cong Guo**, Yuxian Qiu, Jingwen Leng, Chen Zhang, Ying Cao, Quanlu Zhang, Yunxin Liu, Fan Yang, Minyi Guo. "Nesting Forward Automatic Differentiation for Memory-Efficient Deep Neural Network Training." In *IEEE International Conference on Computer Design* (**ICCD**), 2022.

[C29] Mustafa Tarik Sanic, **Cong Guo**, Jingwen Leng, Minyi Guo, Weiyin Ma. "Towards Reliable AI Applications via Algorithm-Based Fault Tolerance on NVDLA." In *International Conference on Mobility, Sensing and Networking* (**MSN**), 2022.

[C30] Yang Wang, Chen Zhang, Zhiqiang Xie, **Cong Guo**, Yunxin Liu, Jingwen Leng. "Dual-side Sparse Tensor Core." In *ACM/IEEE International Symposium on Computer Architecture* (**ISCA**), 2021.

[C31] Yangjie Zhou, Mengtian Yang, **Cong Guo**, Jingwen Leng, Yun Liang, Quan Chen, Minyi Guo, Yuhao Zhu. "Characterizing and Demystifying the Implicit Convolution Algorithm on Commercial Matrix-Multiplication Accelerators." In *IEEE International Symposium on Workload Characterization* (**IISWC**), 2021.

[C32] **Cong Guo**, Bo Yang Hsueh, Jingwen Leng, Yuxian Qiu, Yue Guan, Zehuan Wang, Xiaoying Jia, Xipeng Li, Minyi Guo, Yuhao Zhu. "Accelerating Sparse DNN Models without Hardware-Support via Tile-Wise Sparsity." In *International Conference for High Performance Computing, Networking, Storage and Analysis* (**SC**), 2020.

[C33] **Cong Guo**, Yangjie Zhou, Jingwen Leng, Yuhao Zhu, Zidong Du, Quan Chen, Chao Li, Bin Yao, Minyi Guo. "Balancing Efficiency and Flexibility for DNN Acceleration via Temporal GPU-Systolic Array Integration." In *ACM/IEEE Design Automation Conference* (**DAC**), 2020.

[C34] Yuxian Qiu, Jingwen Leng, **Cong Guo**, Quan Chen, Chao Li, Minyi Guo, Yuhao Zhu. "Adversarial Defense Through Network Profiling Based Path Extraction." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (**CVPR**), 2019.

Journal and Magazine Articles

[J1] **Cong Guo**, Feng Cheng, Zhixu Du, James Kiessling, Jonathan Ku, Shiyu Li, Ziru Li, Mingyuan Ma, Tergel Molom-Ochir, Benjamin Morris, Haoxuan Shan, Jingwei Sun, Yitu Wang, Chiyue Wei, Xueying Wu, Yuhao Wu, Hao Frank Yang, Jingyang Zhang, Junyao Zhang, Qilin Zheng, Guanglei Zhou, Hai Li, Yiran Chen. "A Survey: Collaborative Hardware and Software Design in the Era of Large Language Models." *IEEE Circuits and Systems Magazine* (**CAS-M**), 2025.

[J2] Yangjie Zhou, Zhihui Zhang, Shuwen Lu, **Cong Guo**, Jingwen Leng, Feng Zhang, Yufei Ma, Yun Liang, Minyi Guo. "A Full-Stack Framework for GNN Acceleration via Partition-Compiler-Architecture Co-Design." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (**TCAD**), 2025.

[J3] **Cong Guo**, Fengchen Xue, Jingwen Leng, Yuxian Qiu, Yue Guan, Weihao Cui, Quan Chen, Minyi Guo. "Accelerating Sparse DNNs Based on Tiled GEMM." *IEEE Transactions on Computers* (**TC**), 2024.

[J4] Chen Zhang, Yang Wang, Zhiqiang Xie, **Cong Guo**, Yunxin Liu, Jingwen Leng, Guangyu Sun, Zhigang Ji, Runsheng Wang, Yuan Xie, Ru Huang. "DSTC: Dual-Side Sparsity Tensor Core for DNNs Acceleration on Modern GPU Architectures." *IEEE Transactions on Computers* (**TC**), 2024.

Patents

[P1] Jingwen Leng, **Cong Guo**, Jiaming Tang, Minyi Guo. "Quantization Method for Language Models and Electronic Device." Chinese invention patent (granted), Application No. CN202310395701.9, filed April 2023. Applicant: Shanghai Jiao Tong University.

[P2] Jingwen Leng, Yuhao Zhu, **Cong Guo**, Bin Yao, Minyi Guo. "Reconfigurable Single-Instruction-Multiple-Data Systolic Array Architecture, Processor, and Electronic Terminal." Chinese invention patent (granted), Application No. CN201810344156.X, filed April 2018. Applicant: Shanghai Jiao Tong University.

Research Internships

04/2023 – 09/2023	**Ant Group** *Research Intern* GPU memory management for DNN training and LLM serving (GMLake [C22], eLLM [C13])	Beijing, China
06/2021 – 12/2023	**Shanghai Qi Zhi Institute** *Research Intern* Data-free and outlier-aware quantization (SQuant [C27], OliVe [C24])	Shanghai, China
06/2020 – 05/2021	**Microsoft Research Asia** *Research Intern* Adaptive numerical data types for DNN accelerators (ANT [C26])	Beijing, China
05/2019 – 12/2019	**NVIDIA** *Research Intern* Structured sparsity on GPU Tensor Cores (Tile-Wise Sparsity [C32])	Shanghai, China

Grant Proposal Experience

2026 **NSF AI Institute for Edge Computing and Systems (Athena), Phase II**
$15M over five years, renewed 09/2026 (Director: Yiran Chen) [link]

- Contributed to writing the renewal proposal.
- Helped organize the NSF annual review and presented research at the review.

Open-Source Software and Artifacts

- **GMLake** [C22]: GPU memory defragmentation via virtual memory stitching; open-sourced as part of Ant Group's GLake and deployed in production large-scale DNN training.
- **ANT and OliVe** [C26, C24]: adaptive numerical data types and outlier-victim pair quantization for low-bit DNNs and LLMs. github.com/clevercool/ANT-Quantization
- **SQuant** [C27]: on-the-fly data-free quantization. github.com/clevercool/SQuant
- **Tile-Wise Sparsity** [C32]: structured sparsity for DNN acceleration on GPU Tensor Cores. github.com/clevercool/TileSparsity
- *Artifacts released with mentored students:* EVA [C6], Focus [C7], DPad [C10], FractalCloud [C8], FlashFPS [C12], Prosperity [C17].

Invited Talks

11/2026	"New Architectural Opportunities for LLM Quantization." *ICCAD 2026 Workshop on Memory-Centric Computing for LLM Inference*, San Jose, CA
11/2026	"Where Quantization Architecture Goes Next: Co-Designing Data Types and Hardware for LLM Inference." *MTIA TED Talk*, Meta, Menlo Park, CA
09/2026	"Where Quantization Architecture Goes Next: Co-Designing Data Types and Hardware for LLM Inference." *AI and Systems Co-Design Organization*, Meta, Menlo Park, CA

04/2025 "Unlocking New Opportunities in Quantization and Sparsity Co-Design for Large Language Models."
Athena Seminar Series, NSF AI Institute for Edge Computing and Systems (Athena), Duke University, Durham, NC

03/2025 "Unlocking New Opportunities in Quantization and Sparsity Co-Design for Large Language Models."
Computer Science Seminar, North Carolina State University, Raleigh, NC

08/2023 "Challenges and Opportunities of Large Language Model Quantization for Computer Architecture."
Forum for Outstanding Young Scholars and Ph.D. Students in Computer Architecture, CCF Annual Conference on Computer Architecture (ACA 2023), Nanchang, China

07/2020 "Balancing Efficiency and Flexibility for DNN Acceleration via Temporal GPU-Systolic Array Integration" [C33]
DAC Young Fellows Program, 57th Design Automation Conference (DAC 2020), Virtual

Conference Presentations

06/2026 "EVA: Accelerating LLM Decoding via an Efficient Vector Quantization Architecture" [C6]
ISCA 2026, Raleigh, NC

06/2025 "Transitive Array: An Efficient GEMM Accelerator with Result Reuse" [C14]
ISCA 2025, Tokyo, Japan

04/2024 "GMLake: Efficient and Transparent GPU Memory Defragmentation for Large-scale DNN Training with Virtual Memory Stitching" [C22]
ASPLOS 2024, La Jolla, CA

06/2023 "OliVe: Accelerating Large Language Models via Hardware-Friendly Outlier-Victim Pair Quantization" [C24]
ISCA 2023, Orlando, FL

10/2022 "ANT: Exploiting Adaptive Numerical Data Type for Low-bit Deep Neural Network Quantization" [C26]
MICRO 2022, Chicago, IL

10/2022 "Nesting Forward Automatic Differentiation for Memory-Efficient Deep Neural Network Training" [C28]
ICCD 2022, Olympic Valley, CA

04/2022 "SQuant: On-the-Fly Data-Free Quantization via Diagonal Hessian Approximation" [C27]
ICLR 2022, Virtual

11/2020 "Accelerating Sparse DNN Models without Hardware-Support via Tile-Wise Sparsity" [C32]
SC 2020, Virtual

07/2020 "Balancing Efficiency and Flexibility for DNN Acceleration via Temporal GPU-Systolic Array Integration" [C33]
DAC 2020, Virtual

Research Mentoring

During my senior Ph.D. years at Shanghai Jiao Tong University and my postdoctoral appointment at Duke University, I served as the hands-on, day-to-day research mentor for the students below—guiding problem selection, architecture design, experimental methodology, and paper writing. Together we published 17 papers, including 13 at ISCA, MICRO, HPCA, and ASPLOS; former mentees have gone on to Ph.D. programs at MIT, Princeton, Tsinghua, SJTU, and Duke, and to industry positions at Google and Moonshot AI.

Postdoctoral Period, Duke University (2023–2026)

- **Feng Cheng**, Ph.D. student, Duke University *Now: Ph.D. student, Duke*
  Mentored papers: Ecco (ISCA'25) [C16], Gossamer (MICRO'26) [C5]
- **Chiyue Wei**, Ph.D. student, Duke University *Now: Ph.D. student, Duke*
  Mentored papers: Prosperity (HPCA'25) [C17], Phi (ISCA'25) [C15], Focus (HPCA'26, co-first author, Best Paper Nomination) [C7], Duet (HPCA'27) [C1]
- **Bowen Duan**, visiting undergraduate from Tsinghua University *Now: Ph.D. student, Duke*
  Mentored papers: Transitive Array (ISCA'25) [C14], Phi (ISCA'25) [C15], EVA (ISCA'26, co-first author) [C6], Duet (HPCA'27) [C1]
- **Haoxuan Shan**, Ph.D. student, Duke University *Now: Ph.D. student, Duke*
  Mentored papers: Platinum (ASP-DAC'26) [C11], Vortex (ASPLOS'27, co-first author) [C4], Duet (HPCA'27) [C1]
- **Yuzhe Fu**, Ph.D. student, Duke University *Now: Ph.D. student, Duke*
  Mentored papers: FractalCloud (HPCA'26) [C8], FlashFPS (DAC'26) [C12], Duet (HPCA'27) [C1]
- **Xinhua Chen**, visiting undergraduate from Fudan University *Now: Ph.D. student, Tsinghua*
  Mentored papers: DPad (ICLR'26, co-first author) [C10]
- **Sitao Huang**, visiting undergraduate from University of Oxford *Now: Ph.D. student, Princeton*
  Mentored papers: DPad (ICLR'26, co-first author) [C10]
- **Xuanzeng Song**, visiting undergraduate from Tsinghua University*Now: undergraduate, Tsinghua*
  Mentored papers: Duet (HPCA'27, co-first author) [C1]
- **Junyao Zhang**, Ph.D. student, Duke University *Now: Software Engineer (TPU), Google*
  Mentored papers: Ecco (ISCA'25) [C16], Focus (HPCA'26) [C7], Platinum (ASP-DAC'26) [C11], FlashFPS (DAC'26) [C12], Gossamer (MICRO'26) [C5]

Senior Ph.D. Period, Shanghai Jiao Tong University (2022–2023)

- **Jiaming Tang**, undergraduate, ACM Honors Class, SJTU *Now: Ph.D. student, MIT*
  Mentored papers: OliVe (ISCA'23, co-first author) [C24]
- **Weiming Hu**, M.S. student, ShanghaiTech University *Now: Ph.D. student, SJTU*
  Mentored papers: OliVe (ISCA'23) [C24], MANT (HPCA'25) [C18]
- **Jiale Xu**, M.S. student, SJTU *Now: Engineer, Moonshot AI (Kimi)*
  Mentored papers: GMLake (ASPLOS'24) [C22], eLLM (DAC'26) [C13]

Teaching

04/2024	**Tutorial Lecturer**, "Accelerating Sparse DNNs with Structured Sparsity" ASPLOS 2024 Tutorial: *Amanda: Unified Instrumentation Framework for Deep Neural Networks*, La Jolla, CA
Spring 2019	**Teaching Assistant**, Embedded System Principle and Practice [link] Shanghai Jiao Tong University; Instructor: Prof. Jingwen Leng
Summer 2018	**Teaching Assistant**, Introduction to Computing Summer course taught by Prof. Yale N. Patt (UT Austin), hosted by Prof. Jingwen Leng at Shanghai Jiao Tong University; managed homework, student Q&A sessions, and course logistics
Spring 2018	**Teaching Assistant**, Embedded System Principle and Practice [link] Shanghai Jiao Tong University; Instructor: Prof. Jingwen Leng

Teaching Interests

- *Undergraduate:* Computer Organization and Design; Computer Architecture; Embedded Systems; Introduction to Computing Systems.
- *Graduate:* Advanced Computer Architecture; High-Performance Computing; Parallel Computing and GPU Programming.
- *New course:* **Efficient AI Hardware and Systems** (graduate, open to senior undergraduates), covering sparsity, quantization, and LLM serving systems, with hands-on labs.

Professional Service

Organizing	Publicity Co-Chair, **HPCA 2027**
Program Committee	**ISCA 2027**; **DAC 2027**; **ISPASS 2027**; **DATE 2027**; **MICRO 2026**; **DAC 2026**
Session Chair	**ISCA 2026**; **DAC 2026**
Conf. Reviewer	**HPCA 2027**; **MLSys 2027**; **AAAI 2027**; **ISCAS 2027**; **NeurIPS 2026**; **ISCAS 2026**; **ICCV 2025**
Journal Reviewer	IEEE Transactions on Computers (TC); IEEE Computer Architecture Letters (CAL); ACM Transactions on Architecture and Code Optimization (TACO); ACM Transactions on Embedded Computing Systems (TECS); IEEE Transactions on Very Large Scale Integration (VLSI) Systems (TVLSI); IEEE Transactions on Circuits and Systems for Artificial Intelligence (TCASAI); IEEE Transactions on Circuits and Systems for Video Technology (TCSVT); Journal of Systems Architecture (JSA); Science China Information Sciences (SCIS)

Last updated: November 2026