Haojun Xia

(+61) 493 413 722
xhjustc@gmail.com
www.linkedin.com/in/xhjustc

Education

2022–2026 **University of Sydney (USYD), Sydney, Australia**,
Ph.D. in Computer Science, Advisor: Shuaiwen Leon Song,
Thesis: Compression-Driven Memory-Efficient and High-Throughput GPU Systems for LLM Inference.

2018–2021 **University of Science and Technology of China (USTC), Hefei, China**,
Master's degree in Computer Architecture, First Class Scholarship,
Thesis: The design and implementation of a lightweight Automata Processor.

2014–2018 **University of Science and Technology of China (USTC), Hefei, China**,
Bachelor's degree in Computer Science and Technology, Talented Program,
Thesis: FPGA Based CNN Accelerator Design.

Research Experience

Sept 2024 – May 2026 **Accurate and Efficient 2-bit KV Cache Quantization for Efficient LLM Inference.**, MLSYS'26.
- First Author & Lead Researcher, collaborated with Together AI Team.
- Algorithm-System co-design for accurate 2-bit KV cache quantization, reducing GPU memory consumption and increasing inference throughput during LLM inference.

Sept 2023 – April 2024 **Acceleration LLM Inference on GPUs with FP6 Quantization.**, USENIX ATC'24.
- First Author & Lead Researcher, collaborated with Microsoft DeepSpeed Team.
- Designed and implemented GPU kernel with unified Tensor Core support for various quantization bit-width.
- Developed end-to-end support for quantized inference, $1.69\times$–$2.65\times$ throughput improvement on LLaMA-70B.
- Adopted by the industrial open-source ML frameworks, e.g. Microsoft/DeepSpeed and Pytorch/AO.

Aug 2021 – Sept 2023 **LLM Inference Acceleration Exploiting Unstructured Sparsity**, VLDB'24.
- First Author & Lead Researcher, collaborated with Alibaba Research.
- Identified and analyzed the key bottleneck (HBM bandwidth) during LLM inference.
- Creating highly efficient Large Language Model (LLM) acceleration framework, providing runtime support for LLM inference with unstructured sparsity. (e.g., reducing the inference cost up to 50% for OPT-175B model).

Jan 2021 – Jul 2021 **Hardware Accelerator Design and Prototyping for Efficient BNN Training**, MICRO'21.
- Primary contributor to the implementation (Verilog HDL) and the evaluation (resource/energy estimation on FPGA) of the proposed hardware accelerator.
- We designed and prototyped an efficient hardware accelerator for highly-efficient Bayesian Neural Network (BNN) training for server and edge devices;

Oct 2020 – Feb 2021 **Software-Hardware Co-Design for Efficient Large-scale LSTM Training**, ISCA'21.
- Primary contributor to the implementation and evaluation (resource/energy/latency estimation) of the proposed hardware accelerator; Primary designer of the Omni-PE (Processing Element).
- We prototyped an efficient hardware architecture for large-scale LSTM network training, utilizing intermediate variable compression and cell skipping to reduce memory footprint and data transfer.

Nov 2019 – May 2021 **Automata Processor Designing and Implementation**, DATE'21.
- First Author & Lead Researcher.
- Designed a lightweight hardware processor for large-scale pattern-matching (multi-string matching) tasks;
- Implemented this hardware accelerator using Verilog HDL and instantiated it on real FPGA boards;
- Integrated this processor core with ARM CPUs for efficient heterogeneous computing.

Open-source Repos

Kitty: algorithm-system co-design for accurate and efficient 2-bit KV cache quantization., .
- https://github.com/Summer-Summer/Kitty.

fp6_llm: Efficient GPU support for LLM inference with 6-bit (FP6) quantization., .
- https://github.com/usyd-fsalab/fp6_llm.

flash_llm: Enabling Cost-Effective LLM Inference with Unstructured Sparsity. , .
- https://github.com/AlibabaResearch/flash-llm.

Working Experiences

Nov 2025 – Present **Research Software Engineer @Google Research**, (Full-time, On-site).
- Optimizing production-level TPU-based LLM inference systems.

Mar 2024 – Sept 2024 **Research Consultant @Together AI**, (Part-time, Remote).
- Optimizing the LLM inference system, e.g. identifying and mitigating the performance issue in LoRA inference, developing the FP8 MHA Decoding GPU kernel using OpenAI Triton.

Aug 2021 – Jan 2022 **Research Intern@Alibaba Cloud**, (On-site).
- Investigating SOTA system support for LLMs, and R&D a novel large-scale ML model acceleration framework.
- Writing and publishing a research paper.

Feb 2022 – Aug 2023 **Research Intern@Alibaba Cloud**, (Remote, Un-paid).
- Extension of the former research project.
- Part of the Alibaba Innovative Research (AIR) program.

Publications

[MLSys-26] **"Kitty: Accurate and Efficient 2-bit KV Cache Quantization with Dynamic Channel-wise Precision Boost"**, *Haojun Xia, Xiaoxia Wu, Jisen Li, Tsai-chuan Wu, Junxiong Wang, Jue Wang, Chenxi Li, Aman Singhal, Alay Dilipbhai Shah, Alpay Ariyak, Donglin Zhuang, Zhongzhu Zhou, Ben Athiwaratkun, Zhen Zheng, Shuaiwen Song*, Annual Conference on Machine Learning and Systems, 2026. (Top ML+System Conference).

[ATC-24] **"FP6-LLM: Efficiently Serving Large Language Models Through FP6-Centric Algorithm-System Co-Design"**, *Haojun Xia, Zhen Zheng, Xiaoxia Wu, Shiyang Chen, Zhewei Yao, Stephen Youn, Arash Bakhtiari, Michael Wyatt, Donglin Zhuang, Zhongzhu Zhou, Olatunji Ruwase, Yuxiong He, Shuaiwen Leon Song*, USENIX Annual Technical Conference, 2024. (Top Computer System Conference: A).

[VLDB-24] **"Flash-LLM: Enabling Cost-Effective and Highly-Efficient Large Generative Model Inference with Unstructured Sparsity"**, *Haojun Xia, Zhen Zheng, Yuchao Li, Donglin Zhuang, Zhongzhu Zhou, Xiafei Qiu, Yong Li, Wei Lin, Shuaiwen Leon Song*, International Conference on Very Large Databases, 2024. (Top Computer System Conference: A*).

[OSDI-24] **"MonoNN: Enabling a New Monolithic Optimization Space for Neural Network Inference Tasks on Modern GPU-Centric Architectures"**, *Donglin Zhuang, Zhen Zheng, Haojun Xia, Xiafei Qiu, Junjie Bai, Wei Lin, Shuaiwen Leon Song*, USENIX Symposium on Operating Systems Design and Implementation, 2024. (Top Computer System Conference: A*).

[TC] **"Enabling Fast and Memory-efficient Acceleration for Pattern Matching Workloads: the Lightweight Automata Processing Engine"**, *Lei Gong, Chao Wang, Haojun Xia, Xianglan Chen, Xi Li, Xuehai Zhou*, IEEE Transactions on Computers, 2022. (Top Computer System Journal: A*).

[MICRO-54] **"Shift-BNN: Highly-Efficient Probabilistic Bayesian Neural Network Training via Memory-Friendly Pattern Retrieving"**, *Qiyu Wan, Haojun Xia, Xingyao Zhang, Lening Wang, Shuaiwen Leon Song, Xin Fu*, In Proceedings of the 54th IEEE/ACM International Symposium on Microarchitecture (MICRO), 2021, (Top Computer Architecture Conference: A*).

[ISCA-48] **"η-LSTM: Co-Designing Highly-Efficient Large LSTM Training via Exploiting Memory-Saving and Architectural Design Opportunities"**, *Xingyao Zhang, Haojun Xia, Donglin Zhuang, Hao Sun, Xin Fu, Michael Taylor, Shuaiwen Leon Song*, In Proceedings of International Symposium on Computer Architecture (ISCA), 2021, (Top Computer Architecture Conference: A*).

[DATE] **"LAP: A Lightweight Automata Processor for Pattern Matching Tasks"**, *Haojun Xia, Lei Gong, Chao Wang, Xianglan Chen and Xuehai Zhou*, In Proceedings of Design, Automation and Test in Europe Conference (DATE), 2021.

[Arxiv] **"ZeroQuant (4+ 2): Redefining LLMs Quantization with a New FP6-Centric Strategy for Diverse Generative Tasks"**, *Xiaoxia Wu, Haojun Xia, Stephen Youn, Zhen Zheng, Shiyang Chen, Arash Bakhtiari, Michael Wyatt, Yuxiong He, Olatunji Ruwase, Leon Song, Zhewei Yao*.

Achievements

2022 – 2026 **Faculty of Engineering Research Scholarship**, *PhD study*, USYD.
- Established to encourage and support outstanding research students at the Faculty of Engineering.

2021 **Outstanding Graduates**, *Master's study*, USTC.
- Top 15% among all graduates.

2020 **Suzhou Park Scholarship**, *Master's study*, USTC.
- Awarding postgraduate students at USTC who have excellent grades and have made contributions in research.

2018-2021 **First Class Academic Scholarship**, *Master's study*, USTC.
- Awarded to first-class postgraduate students.

2018 **Yang Yuanqing Education Fund - Top Research Scholarship**, *Bachelor's study*, USTC.
- Awarding 8 undergraduates and 4 postgraduates major in computer science and math each year.

2018 **Outstanding Graduates**, *Bachelor's study*, USTC.
- Top 15% among all graduates.

2014-2018 **Talent program in computer science and technology**, *Bachelor's study*, USTC.
- Taking more advanced courses and receiving extra scholarship.

Interests & Technical Skills

Interests Performance Optimization, Machine Learning System, Runtime Systems
Computer Architecture, Demain Specific Architectures, GPU/TPU Kernel Design.

Languages JAX, Pallas (TPU), C/C++, Python, CUDA(GPU), Triton(GPU), Verilog HDL (FPGA)

Software Development Machine Learning Frameworks (e.g. PyTorch, Huggingface Transformers, FasterTransformer), Embedded System Design and Implementation (e.g. Xilinx FPGA + ARM CPUs)

Extra Curricular

Mar 2019 – Jun 2019 **Teaching Assistant**, *USTC*, Computer Architecture 2019.

Mar 2018 – Jun 2018 **Teaching Assistant**, *USTC*, Computer Architecture 2018.

Sep 2017 – Dec 2017 **Teaching Assistant**, *USTC*, Digital Circuit Theory 2017.

Sep 2016 – Dec 2016 **Teaching Assistant**, *USTC*, Digital Circuit Experiments 2016.