Xiuhong Li

Email: lixiuhong@infini-ai.com
Mobile: +86 18519136939

Education

- **The Chinese University of Hong Kong** — Hong Kong, China
  Postdoctoral Researcher - MM Lab, Information Engineering — *March 2020 - July 2021*
  Research Area: *Deep Learning Compiler, High Performance Computing*
- **Peking University** — Beijing, China
  Ph.D. - Center for Energy-Efficient Computing and Applications, Computer Science — *September 2014 - June 2019*
  Research Area: *Application, Algorithm, and Architecture on GPUs*
- **Peking University** — Beijing, China
  Bachelor of Science - Microelectronics — *September 2010 - June 2014*
  Course: *Programming Language, Basic Mathematics, Physics, Analog Integrated Circuit, Digital Integrated Circuit*

Experience

- **Infinigence-AI** — Beijing, China
  Vice President — *Feb. 2025 -*
  Duty: *Cost-efficiency LLM Inference Infrastructure for MaaS.*
- **Peking University** — Beijing, China
  Rsearch Assistant Professor — *June 2023 - Feb. 2025*
  Research Area: *Training and Inference Infrastructure for Large Language Models.*
- **SenseTime SenseCore** — Beijing, China
  Associate Director — *May 2022 - May 2023*
  Duty: *Design interface between deep learning framework and deep learning accelerators, including operator-level APIs and deep learning compilation path.*
- **SenseTime Research** — Beijing/Hong Kong, China
  Senior Researcher — *July 2019 - April 2022*
  Duty: *Design and develop in-house deep learning compiler based on TVM and MLIR.*

Selected Publications

- [**TPDS'2024**]: Jiangfei Duan, **Xiuhong Li,** Ping Xu, Xingcheng Zhang, Shengen Yan, Yun Liang, Dahua Lin, "Proteus: Simulating the performance of distributed DNN training", in the IEEE Transactions on Parallel and Distributed Systems (TPDS).**(Corresponding Author)**
- [**DAC'2024**]: Jinming Ma, **Xiuhong Li**, Zihan Wang, Xingcheng Zhang, Shengen Yan, Yuting Chen, Yueqian Zhang, Minxi Jin, Lijuan Jiang, Yun Liang, Chao Yang, Dahua Lin, "A Holistic Functionalization Approach to Optimizing Imperative Tensor Programs in Deep Learning", in Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC).**(Corresponding Author)**
- [**ASPLOS'2024**]: Chang Chen, **Xiuhong Li**, Qianchao Zhu, Jiangfei Duan, Peng Sun, Xingcheng Zhang, Chao Yang, "Centauri: Enabling efficient scheduling for communication-computation overlap in large model training via communication partitioning", in Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS). **(Best Paper Award.) (Corresponding Author)**
- [**IISWC'2022**]: **Xiuhong Li**, Shengen Yan, Lijuan Jiang, Ping Xu, Jinming Ma, Xingcheng Zhang, Dahua Lin, "LongTail-Bench: A Benchmark Suite for Domain-Specific Operators in Deep Learning", in the IEEE International Symposium on Workload Characterization (IISWC).
- [**ICPP'2022**]: Lijuan Jiang, Ping Xu, Qianchao Zhu, **Xiuhong Li**, Shengen Yan, Dahua Lin, Wenjing Ma, Zhouyang Li, Jun Liu, Jinmin Ma, Minxi Jin, and Chao Yang, "EasyView: Enabling and Scheduling Tensor Views in Deep Learning Compilers", in the proceeding of 51st International Conference on Parallel Processing (ICPP). **(Corresponding Author)**
- [**PPoPP'2019**]: **Xiuhong Li**, Yun Liang, Shengen Yan, Liancheng Jia and Yinghan Li, "A Coordinated Tiling and Batching Framework for Efficient GEMM on GPUs", in the proceedings of ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, 2019: 229-241. **(Best Paper Award Nomination.)**
- [**ICS'2018**]: **Xiuhong Li**, Yun Liang, Wentai Zhang, Taide Liu, Haochen Li, Guojie Luo, and Ming Jiang, "cuMBIR: An Efficient Framework for Low-dose X-ray CT Image Reconstruction on GPUs", ACM International Conference on Supercomputing, 2018: 184-194.
- [**TECS'2017**]: Yun Liang, **Xiuhong Li**. "Efficient Kernel Management on GPUs," ACM Transactions on Embedded Computing Systems, 2017, Volume 16, Issue 4, pp 1-24.
- [**DATE'2016**]: **Xiuhong Li**, Yun Liang. "Efficient Kernel Management on GPUs," in the proceedings of the Design Automation and Test in Europe, IEEE, 2016: 85-90.
- [**ASPDAC'2015**]: Shuo Wang, Yun Liang, Chao Zhang, Xiaolong Xie, Guangyu Sun, Yongpan Liu, Yu Wang, and **Xiuhong Li**, "Performance-centric Register File Design for GPUs using Racetrack Memory," in the proceedings of the Asia and South Pacific Design Automation Conference, IEEE, 2016: 25-30. **(Best Paper Award Nomination.)**
- [**MICRO'2015**]: Xiaolong Xie, Yun Liang, **Xiuhong Li**, Yudong Wu, Guangyu Sun, Tao Wang, and Dongrui Fan. "Enabling Coordinated Register Allocation and Thread-level Parallelism Optimization for GPUs," in the proceedings of the 48th Annual IEEE/ACM International Symposium on Microarchitecture, IEEE, 2015: 395-406.