Lihan Hu

lihan-hu@uiowa.edu · (+1) 319-473-3260 · LinkedIn Homepage

5 Years of Experiences on GPU, AI Infra | Green Card Holder | Available from April 2027

♡ Research Interest

Efficient Machine Learning Algorithm, GPU Programming, Compiler Optimization, Code Generation, Hardware-software Co-design, LLM Serving

🎓 Education

The University of Iowa, Iowa, USA 2022 – 2027

Ph.D. Candidate in Computer Science **Advisor: Peng Jiang**

Hohai University, Nanjing, Jiangsu, China 2019 – 2022

M.S. in Pattern Recognition and Intelligent Systems **Advisor: Lixin Han**

Nanjing Audit University, Nanjing, Jiangsu, China 2014 – 2018

B.S. in Computer Science and Technology **Advisor: Zhenyuan Xu**

ℹ Professional Experience

The University of Iowa Iowa City, IA Sep. 2022 – Apr. 2027

Research Assistant, Parallel Computing and Compiler Optimization Advisor: Peng Jiang

SK Hynix America San Jose, CA Nov. 2025 – Jan. 2026 & May. 2026 – Aug. 2026

Research Intern, Memory-Centric AI Infra Team Supervisor: Jongryool Kim

The City University of Hong Kong Kowloon, Hong Kong Sep. 2020 – Aug. 2021

Research Assistant, Medical Imaging Processing Advisor: Hong Yan

Hohai University Nanjing, China Sep. 2019 – Jun. 2022

Research Assistant, Disk Failure Detection and Imaging Processing Advisor: Lixin Han

Huazhong Univerisity of Science and Technology Wuhan, China Oct. 2016 – Feb. 2018

Research Intern, Disk Failure Detection Advisor: Ke Zhou

💬 Projects

Dynamic State Pruning for State Space Models Nov. 2025 – Present

- Developed a dynamic state pruning framework for accelerating Mamba without post-training. The framework summarizes offline state selections into a compact template bank and uses lightweight runtime routing to select execution templates for different input chunks. Co-designed a sparse selective scan kernel that maintains a fixed number of active states per warp segment, preserving balanced GPU workloads and regular memory access while reducing state-update overhead.

Distributed Model Serving with vLLM-Omni and CXL Memory May. 2026 – Aug. 2026

- Studied the vLLM-Omni architecture and deployed multimodal models, including Cosmos 3 and Qwen3-Omni, across heterogeneous GPU servers. Explored CXL-based pooled memory for sharing intermediate representations, latent features, and KV caches across serving stages. Investigated pipeline scheduling and data-transfer optimizations to reduce redundant computation and inter-server communication.

Near-HBM Acceleration for Structured SpMM (CAL'26) Nov. 2025 – Jan. 2026

- Developed TC-AIA, a processing-near-HBM architecture for accelerating fine-grained structured sparse matrix multiplication. The design reorganizes sparse metadata and offloads indirect data gathering and input packing to near-memory logic. It converts irregular memory accesses into contiguous inputs for Tensor Core computation, reducing GPU-side memory overhead and improving hardware utilization.

GPU Optimization for Block-Sparse Attention (PACT'26) Oct. 2024 – May. 2026

- Developed Ziggs, a reuse-aware scheduling and GPU execution framework for block-sparse attention. Formulated intra-group block scheduling as a Markov Decision Process and combined offline learned schedules with a lightweight fallback for unseen patterns. Implemented optimized SDDMM and SpMM kernels that use scheduling metadata and shared memory to exploit inter-block reuse and reduce redundant global memory accesses.

Hardware-Software Co-Design for Sparse Attention (IPDPS'26) Oct. 2024 – May. 2026

- Designed a heterogeneous architecture for long-sequence LLM inference with hybrid sparse attention. Mapped memory-intensive sparse attention operations to processing-in-memory hardware while executing compute-intensive fully connected layers on conventional processors. Explored workload partitioning and dataflow optimizations to leverage PIM bandwidth while minimizing communication between heterogeneous components.

Compiler Optimization for Subgraph Matching (IPDPS'25) Nov. 2023 – Jul. 2024

- Extended the cuKE compiler to support backtracking-based subgraph matching and implemented an apply operator for flexible computation expressions. Designed compiler transformations for translating high-level matching operations into optimized GPU kernels. Applied kernel fusion, warp-level optimizations, and shared-memory techniques to reduce execution and intermediate-data overhead.

Knowledge Graph Embedding Optimization (IPDPS'24, IPDPS'25) Sep. 2022 – Oct. 2024

- Developed cuKE, a GPU code-generation framework for knowledge graph embedding score functions. Profiled computation, warp scheduling, synchronization, and memory-access bottlenecks using NVIDIA Nsight Systems and Nsight Compute, and applied operator fusion and runtime inspection to optimize generated GPU kernels. Further designed frequency-aware parameter placement, duplication, precaching, and segmented synchronization techniques for efficient and accurate multi-GPU training.

Structured Dynamic Sparse Training (NeurIPS'22) Mar. 2022 – Sep. 2022

- Investigated fine-grained block structures for efficient dynamic sparse training. Designed shuffled block patterns that preserve the flexibility and accuracy of unstructured sparsity while enabling hardware-friendly structured execution. Evaluated different block configurations and developed optimized CUDA kernels to balance model accuracy, training overhead, and GPU efficiency.

C. elegans Cell Image Processing (Quantitative Biology) Sep. 2020 – Mar. 2022

- Developed CShaperApp, an automated segmentation and analysis application for *C. elegans* embryonic cell microscopy images using TensorFlow and PyQt5. Built an end-to-end workflow for image preprocessing, cell segmentation, visualization, and morphological analysis. Implemented geometry-based methods, including Delaunay triangulation, to identify cellular boundaries and spatial relationships among cells.

Disk Failure Prediction in Data Centers (KES'20, ICPP'19) Oct. 2016 – May. 2020

- Developed an LSTM-based disk failure prediction framework using time-series SMART data. The framework models temporal changes in disk health and captures behavioral differences among individual disks to improve prediction reliability. Also contributed to a transfer learning approach that transfers failure knowledge from data-rich disk models to minority disk models in heterogeneous data centers.

SKILLS

- Programming Languages: Python, C++, CUDA C++, Java, SQL
- Libraries: PyTorch, Triton, vLLM, vLLM-Omni, TVM, CUTLASS, cuBLAS, NumPy, Pandas, PyQt5
- Profiling and Development Tools: NVIDIA Nsight Systems, NVIDIA Nsight Compute, Git, CMake

Publications

PACT'26 *Improving Data Reuse across Blocks for Efficient Block-sparse Transformers on GPUs.*
Lihan Hu, Xiaoyang Lu, Xian-He Sun, Peng Jiang.
The International Conference on Parallel Architectures and Compilation Techniques (PACT)

CAL'26 *Near-HBM Tensor Core Acceleration for Fine-Grained Sparse Matrix-Matrix Multiplication.*
Lihan Hu, Shiju Li, Hoshik Kim, Jongryool Kim.
IEEE Computer Architecture Letters

IPDPS'26 *IO-Aware PIM Acceleration for Long-Sequence LLM Inference with Hybrid Sparse Attention.*
Xiaoyang Lu*, **Lihan Hu***, Hongrui Huang, Peng Jiang, Xian-He Sun.
40th IEEE International Parallel & Distributed Processing Symposium (IPDPS)

IPDPS'25 *Matcha: A Language and Compiler for Subgraph Matching.*
Yihua Wei, **Lihan Hu**, Peng Jiang.
39th IEEE International Parallel & Distributed Processing Symposium (IPDPS)

IPDPS'25 *Improving Accuracy and Efficiency of Graph Embedding Training with Fine-Grained Parameter Management.*
Lihan Hu, Peng Jiang.
39th IEEE International Parallel & Distributed Processing Symposium (IPDPS)

IPDPS'24 *cuKE: An Efficient Code Generator for Score Function Computation in Knowledge Graph Embedding.*
Lihan Hu, Jing Li, Peng Jiang.
38th IEEE International Parallel & Distributed Processing Symposium (IPDPS)

Quantitative Biology *CShaperApp: Segmenting and analyzing cellular morphologies of the developing Caenorhabditis elegans embryo.*
Jianfeng Cao, **Lihan Hu**, Guoye Guan, Zelin Li, Zhongying Zhao, Chao Tang, Hong Yan.

NeurIPS'22 *Exposing and Exploiting Fine-Grained Block Structures for Fast and Accurate Sparse Training.*
Peng Jiang, **Lihan Hu**, Shihui Song.
The Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS)

KES'20 *A disk failure prediction method based on LSTM network due to its individual specificity.*
Lihan Hu, Lixin Han, Zhenyuan Xu, Tianming Jiang, Huijun Qi.
24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES)