Xinning Hui

⌂ Homepage ☎ (+1) 984-218-7267 | ✉ xhui@ncsu.com | xinning-hui-9a3b2b99/

Research Interests

Efficient ML/AI on Serverless Computing, ML/AI systems, GPU for ML/AI, Sustainable ML/AI.

Education

North Carolina State University *Raleigh, North Carolina, USA*
Ph.D. in Computer Science; Advisor: Dr. Xipeng Shen *Aug. 2021 - Present*

Tsinghua University *Beijing, China*
Master of Computer Science; Advisor: Dr. Zhihui Du *Aug. 2015 - Jul. 2018*

Qinghai University *Qinghai, China*
Bachelor of Computer Science; Ranking: 1/43 *Sep. 2010 - Jul. 2014*

Research Experience

Department of Computer Science, North Carolina State University *Raleigh, NC, USA*
Research Assistant; Advisor: Dr. Xipeng Shen *Jan. 2022 - Present*

- ***Efficient Scheduling of DNN Workflows on Serverless Platforms with Shareable GPUs (HPDC 24)***
  - Designed the first optimality-guided adaptive scheduling algorithm on serverless, that simultaneously tackles inter-function relations, GPU sharing (Multi-Instance-GPU), batching, and Service Level Objective (SLO).
  - Implemented the proposed design on OpenWhisk, achieving a significant improvement in SLO hit rates of 61%-80% and a cost reduction of 47%-187%.
- ***Exploring Function Granularity for ML Applications on Serverless (under review)***
  - Conducted the first systematic study on the impact of function granularity and revealed the performance and cost heterogeneity from function granularity on serverless computing.
  - Designed the prediction model using machine learning and reinforcement learning methods to select the best function granularity based on current workloads, latency, and cost requirements.
  - Implemented the programming and runtime support to to integrate adaptive function granularity in serverless computing, achieving a 29.2% improvement in SLO hit rates and up to a 24.6% reduction in resource cost.
- ***A Dynamic Pipelined Solution for ML on Serverless Platforms with Multi-Instance GPUs (under review)***
  - Conducted a comprehensive evaluation and identified that the root cause of low GPU utilization with MIG is resource fragmentation and exclusive keep-warm.
  - Designed an automatic pipeline to utilize fragmented resources and eviction-based temporal sharing to improve resource utilization.
  - Implemented the novel programming support, on-the-fly pipeline construction, and GPU-aware function state management , achieving 25-75% improvement in throughput while improving up to 90% in SLO hit rate.

Department of Computer Science, North Carolina State University *Raleigh, NC, USA*
Research Assistant; Advisor: Dr. Guoliang Jin *Aug. 2021 - Jan. 2022*

- ***Schedule Tuning on Stable Synchronization Determinism (PACT 22)***
  - Conducted a systemically categorization of existing Deterministic MultiThreading (DMT) systems to identify totally-ordered synchronization and workload-length imbalance as scheduling-oblivious overheads.
- ***Deep Study of the Effects and Fixes of Server-Side Request Races in Web Applications (MSR 22)***
  - Investigate the external effects of request races resulting in semantics violations, and classified request races as latent and non-latent.

Institute of High Performance Computing, Tsinghua University *Beijing, China*

Research Assistant; Advisor: Dr. Zhihui Du *Aug. 2015 - Jul. 2018*

- ***Inter-Job Scheduling of High-Throughput Material Screening Applications (IPDPS 2020)***
  - Formulated the Material Screening problem as solving a systems of Kohn-Sham (KS) equation.
  - Designed two Inter-Job scheduling algorithms, a qualitative and a quantitative method to explore similarities between simulation runs.
- ***When Good Enough Is Better: Energy-Aware Scheduling for Multicore Servers (IPDPSW 17)***
  - Proposed and implemented approximate computing in job scheduling to significantly reduce the energy consumption, saving 23.9% energy cost with sufficient quality (90%).

Publication

NSDI 2025 (under review) **Xinning Hui**, Yuanchao Xu, Xipeng Shen. "**FluidFaaS: A Dynamic Pipelined Solution for ML on Serverless Platforms with Multi-Instance GPUs**", the 22nd USENIX Symposium on Networked Systems Design and Implementation, Philadelphia, PA, USA, April, 2025.

ASPLOS 2025 (under review) **Xinning Hui**, Yuanchao Xu, Xipeng Shen. "**Exploring Function Granularity for Serverless Machine Learning Applications with GPU Sharing**", the ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Rotterdam, The Netherlands, April, 2025

HPDC 2024 **Xinning Hui**, Yuanchao Xu, ZhiShan Guo, Xipeng Shen. "**ESG: Pipeline-Conscious Efficient Scheduling of DNN Workflows on Serverless Platforms with Shareable GPUs**", The 33rd International Symposium on High-Performance Parallel and Distributed Computing, Pisa, PI, Italy, June 2024. **[Paper]**

PACT 2022 Qi Zhao, Zhengyi Qiu, Shudi Shao, **Xinning Hui**, Hassan Ali Khan, Guoliang Jin. "**Understanding and Reaching the Performance Limit of Schedule Tuning on Stable Synchronization Determinism**", The 31st International Conference on Parallel Architectures and Compilation Techniques, Chicago, IL, USA, October 2022. **[Paper]**

MSR 2022 Zhengyi Qiu, Shudi Shao, Qi Zhao, Hassan Ali Khan, **Xinning Hui**, Guoliang Jin. "**A Deep Study of the Effects and Fixes of Server-Side Request Races in Web Applications**", The 2022 Mining Software Repositories Conference, Pittsburgh, PA, USA, May 2022. **[Paper]**

IPDPS 2020 Zhihui Du, **Xinning Hui**, Yurui Wang, Jun Jiang, Jason Liu, Baokun Lu, Chongyu Wang. "**Inter-Job Scheduling of High-Throughput Material Screening Applications**", The 34th IEEE International Parallel and Distributed Processing Symposium, New Orleans, Louisiana USA, May 2020. **(The first student author)** **[Paper]**

IPDPSW 2017 **Xinning Hui**, Zhihui Du, Jason Liu, Hongyang Sun, Yuxiong He, David A. Bader. "**When Good Enough Is Better: Energy-Aware Scheduling for Multicore Servers**", The 31th IEEE International Parallel and Distributed Processing Symposium Workshops, Orlando, Florida USA, May 2017. **[Paper]**

Honors & Awards

2017	Friends of Tsinghua University - Gemalto Outstanding Woman Engineer Scholarship	*China*
2014	Outstanding Bachelor Thesis Award **(Top 1 in the department)**	*China*
2013	National Scholarships of China **(most honorable scholarship for Chinese undergraduate)**	*China*

Skills

Programming Languages:	Python, C, C++, Shell, MATLAB, JAVA
HPC Programming Models:	MPI, OpenMP, CUDA
Frameworks:	OpenWhisk, Knative, Kubernete, PyTorch, TensorRT, TensorFlow,