Xiaofeng Wu

(484) 294-8134 | shizukanaskytree@gmail.com | LinkedIn | GitHub

Education

The University of Texas at Arlington | Department of Computer Science and Engineering **Arlington, Texas**
Doctor of Science: Computer Science, Deep Learning Framework, GPU Computing Aug. 2017 - Now

Waseda University | Department of Information, Production and Systems **Fukuoka, Japan**
Master of Science: 3D Integrated Circuit Design Aug. 2014 - Jun. 2016

Southeast University | School of Electronic Science & Engineering **Nanjing, China**
Bachelor of Engineering: Electronic Science Aug. 2011 - Jun. 2014

Publications

- **Wu, Xiaofeng**, Jia Rao, Wei Chen, Hang Huang, Chris Ding, and Heng Huang. SwitchFlow: preemptive multitasking for deep learning. In *Proceedings of the 22nd International Middleware Conference*, 2021 Best Paper Award (1 out of 107 submissions)
- **Wu, Xiaofeng**, Kun Suo, Yong Zhao, and Jia Rao. A Side-channel Attack on HotSpot Heap Management. In *10th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud'18)*
- Hang Huang, Jia Rao, Song Wu, Hai Jin, Hong Jiang, Hao Che, and **Wu, Xiaofeng**. Towards exploiting cpu elasticity via efficient thread oversubscription. In *Proceedings of the 30th International Symposium on High-Performance Parallel and Distributed Computing*, 2021
- Hang Huang, Jia Rao, Song Wu, Hai Jin, Kun Suo, and **Wu, Xiaofeng**. Adaptive Resource Views for Containers. In *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, pages 243–254. ACM, 2019
- Yong Zhao, Kun Suo, **Wu, Xiaofeng**, Jia Rao, Song Wu, and Hai Jin. Preemptive Multi-Queue Fair Queuing. In *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, pages 147–158. ACM, 2019

Internship Experience

ProtagoLabs **Vienna, Virginia (Remote)**, June. 2021 - Dec. 2021

- The task is to build a robust elastic distributed deep learning system to support large-scale NLP model training.

ByteDance MLSys Laboratory **Seattle, Washington**, June. 2020 - Aug. 2020

- Auto-tuning various models by TVM to improve model inference performance, e.g., OCR Transformer model.
- Integrate TVM graph runtime into a system which is used for other mobile apps with GPU devices.

Hitachi Research Laboratory **Hitachi, Japan**, Aug. 2015 - Sep. 2015

- Collected the speed and tire rotation speed of Ropits (a self-driving car devised by Hitachi), and processed the data of Ropits, discovered the relationship between vehicle speed and tire rotation speed in order to delimit safety scope.

The Institute of Electronics of Chinese Academy of Sciences **Beijing, China**, Aug. 2014 - Sep. 2014

- Implementation of an object detection algorithm accelerated by GPU.

Projects

Deep Learning Framework and GPU Architecture Research Mar. 2018 - Now

- Study Tensorflow and PyTorch for heterogeneous system to accelerate deep learning workloads on GPUs.
- We proposed SwitchFlow built upon TensorFlow for preemptive deep learning multitasking. Designs: (1) SwitchFlow schedules subgraphs and prevents subgraphs from different models to run simultaneously on a GPU, resulting in less interference and the elimination of out-of-memory errors. Subgraphs running on different devices can overlap with each other, leading to a more efficient execution pipeline. (2) SwitchFlow maintains multiple versions of each subgraph, allowing subgraphs to be migrated across devices at a low cost, thereby enabling low-latency preemption.

Research on HotSpot Heap Management of Java Virtual Machine and propose a side-channel attack. Jun. 2017 - Mar. 2018

Skills

Programming Languages: C/C++, Python, CUDA, Java

Frameworks and Tools: Tensorflow, PyTorch, Huggingface NLP, Keras, TVM, Docker, Ray(Hyperparameter tuning), Bazel, CMake, gRPC