

Keren Zhou

6100 Main ST – Houston, TX – 77005, United States

☎ +1-281-687-6961

✉ kerezhou@outlook.com

🌐 www.jokeren.tech

EDUCATION BACKGROUND

- 09/2017-07/2022 **Department of Computer Science, Rice University** **Houston, United States**
Expected Degree: *Ph.D. in Computer Science* **GPA:** 4.0/4.0
Advisor: John Mellor-Crummey
- 09/2014-07/2017 **Institute of Computing Technology, Chinese Academy of Sciences** **Beijing, China**
Degree: *M.S. in Computer Architecture* **GPA:** 90/100
Advisor: Guangming Tan **Thesis:** High Performance Deep Learning Algorithms
- 09/2010-07/2014 **School of Software, Yunnan University** **Kunming, China**
Degree: *B.E. in Network Engineering* **GPA:** 92/100 (Rank: 1/290)
Advisor: Wei Zhou **Thesis:** A Practical Concurrent Quadtree

RESEARCH EXPERIENCE

- 09/2017-NOW **Rice University** **Houston, United States**
Research Assistant
GPU Performance Measurement and Analysis Tool
 - Implemented OpenMP Tool Interface for CUDA backend in llvm-openmp;
 - Built a runtime system to collect GPU activities in a heterogeneous environment and attributed them back to the corresponding CPU calling context;
 - Analyzed GPU binaries to extract GPU functions, recover control flows, and map instructions to source code;
 - Associated runtime samples with static GPU program structures to reconstruct calling context on GPUs and estimate instruction throughput and roof-line model;
 - Optimized large-scale GPU-accelerated applications including Laghos, QMCPACK, Nyx, and LAMMPS;
 - Building a profile-guided performance advisor based on GPU performance metrics, program structures, and PC samples.
- 06/2015-07/2017 **Institute of Computing Technology, Chinese Academy of Sciences** **Beijing, China**
Research Assistant
GPU Performance Model
 - Decoded Nvidia GPU assembly codes, developed assemblers to generate cuBINs, and wrote accelerated kernels that outperform cuDNN by 40%;
 - Built a static performance analysis model that estimates performance bottlenecks.**High Performance Deep Learning Framework**
 - Devised a coarse-grained parallelism strategy with fine-grained vectorization and blocking effects on CPU, making CNNs 5-12 times faster than Caffe on a 16-core E5-2670;
 - Wrote assembly codes to make full use of dual issue and avoid bank conflict on GPU, improving convolution performance with up to 160% speedup than cuDNN on Kepler architectures.
- 01/2013-07/2014 **Intelligent Web Laboratory, Yunnan University** **Kunming, China**
Research Assistant
Concurrent Data Structures
 - Designed several concurrent multi-dimensional trees, including the first lock-free quadtree and k-d tree that are 109% faster than state-of-the-art concurrent trees;
 - Surveyed concurrent data structures, concluded a general method for development and verification;
 - Adopted a specialized skiplist in a p2p indexing system.

INDUSTRY EXPERIENCE

- 06/2018-08/2018 Facebook Inc. Menlo Park, United States**
Research Intern
- Accelerated neural networks on ARM CPUs using auto-tuning methods;
 - Analyzed Winograd algorithm's complexities of various convolution configurations;
 - Reference: Research Scientist Hao Lu, hlu@fb.com.
- 04/2017-07/2017 Nvidia Inc. Beijing, China**
Research Intern
- Developed quantization tools on emerging GPUs to utilize INT8 capabilities;
 - Evaluated the precision and speed of different quantization modes on Pascal Titan X;
 - Reference: Technical Manager Julien Lai, julienlai@nvidia.com.
- 10/2013-02/2014 Baidu Inc. Beijing, China**
Software Engineering Intern
- Optimized Hadoop workflow with its performance improved by 30%, making it capable of extracting thousands of features from raw text files and loading them into data warehouse;
 - Developed a Hadoop workflow monitoring system that can display multiple workflow states and report exception handling;
 - Reference: Senior Engineer Jing Li, lijing16@baidu.com.

SELECTED PUBLICATIONS

- [1] **Keren, Zhou;** Mark Krentel; John, Mellor-Crummey: A tool for top-down performance analysis of GPU-accelerated applications. In: *Principles and Practice of Parallel Programming (PPoPP)*, 2020
- [2] **Keren, Zhou;** John, Mellor-Crummey: A tool for performance analysis of GPU-accelerated applications. In: *International Symposium on Code Generation and Optimization (CGO)*, 2019
- [3] **Keren, Zhou;** Guangming, Tan; Wei, Zhou: Quadboost: A Scalable Concurrent Quadtree. In: *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, 2018
- [4] **Keren Zhou;** Guangming Tan; Xiuxia Zhang; Chaowei Wang; Ninghui Sun: A Performance Analysis Framework for Exploiting GPU Microarchitectural Capability. In *26th ACM International Conference on Supercomputing (ICS)*, 2017
- [5] Xiuxia, Zhang; Guangming, Tan; Shuangbai, Xue; Jiajia, Li; **Keren, Zhou;** Mingyu, Chen: Understanding GPU Microarchitecture to Achieve Bare-Metal Performance Tuning. In: *22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPOPP)*, 2017

AWARDS & HONORS

- 2019** Ken Kennedy Institute ExxonMobil Fellowship
- 2019** Second Place, ACM CGO Student Research Competition
- 2017** Ken Kennedy Institute Andrew Ladd Fellowship
- 2017** Ken Kennedy Institute CS&E Fellowship
- 2016** Schlumberger Scholarship (3%)
- 2015** Top 10, Alibaba 1st Middleware Engineering Contest
- 2014** Bronze Medal, The 2014 ACM-ICPC Asia Regional Contest
- 2014** Outstanding B.E. Degree Thesis of Yunnan University
- 2013** Best Creative Award, Baidu Future Search Engine Contest
- 2013** Meritorious Winner, Mathematical Contest in Modeling
- 2011** Second Prize, China Undergraduate Mathematical Contest in Modeling
- 2011&2012&2016** National Scholarship (2%)

SKILLS

- Languages** C, C++, Java, Python, Bash, JavaScript
- Parallelism** Pthread, OpenMP, MPI, CUDA/HIP, RAJA/Kokkos