

Keren Zhou

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

☎ +1-281-687-6961

✉ kerezhou@outlook.com

🌐 www.jokeren.tech

EDUCATION BACKGROUND

- 09/2017-07/2023 **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 Analysis Tool
 - Extended HPCToolkit to support measurement and analysis of accelerated OpenMP and CUDA GPU programming models in a large-scale heterogeneous environment;
 - Built a profile view of GPU program executions and attributed runtime samples to the corresponding calling context.
- 06/2015-07/2017 **Nvidia-Sugon-ICT Deep Learning Joint Laboratory** **Beijing, China**
Research Assistant
GPU Performance Model
 - Decoded Nvidia GPU assembly codes, developed assemblers to generate cuBINs, and built a static performance analysis model that estimates performance bottlenecks;
 - Published two related papers: *A Performance Analysis Framework for Exploiting GPU Microarchitectural Capability* and *Understanding GPU Microarchitecture to Achieve Bare-Metal Performance Tuning*.**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 60% 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 much faster than traditional fine-grained lock versions, and published two technical reports: *Parse Concurrent Data Structures: BST as an Example* and *Quadboost: A Scalable Concurrent Quadtree*;
 - Surveyed concurrent data structures, concluded a general method for development and verification, and published a paper: *Study on Multi-Core Data Structure in Shared-Memory*;
 - Adopted a specialized skiplist in a p2p indexing system and published a paper: *Concurrent Skiplist Based Double-Layer Index Framework for Cloud Data Processing*.

INDUSTRY EXPERIENCE

- 06/2018-08/2018 **Facebook Inc.** **Menlo Park, United States**
 - 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. <i>Research and Development Intern</i> <ul style="list-style-type: none"> ○ 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. | Beijing, China |
| 10/2013-02/2014 | Baidu Inc. <i>Research and Development Intern</i> <ul style="list-style-type: none"> ○ 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. | Beijing, China |

SELECTED PUBLICATIONS

- | | |
|-----|--|
| [1] | Keren, Zhou; John, Mellor-Crummey: A tool for performance analysis of GPU-accelerated applications. In: <i>A tool for performance analysis of GPU-accelerated applications (CGO)</i> , 2019 |
| [1] | Keren, Zhou; Guangming, Tan; Wei, Zhou: Quadboost: A Scalable Concurrent Quadtree. In: <i>IEEE Transactions on Parallel and Distributed Systems (TPDS)</i> , 2018 |
| [2] | Keren Zhou; Guangming Tan; Xiuxia Zhang; Chaowei Wang; Ninghui Sun: A Performance Analysis Framework for Exploiting GPU Microarchitectural Capability. In <i>26th ACM International Conference on Supercomputing (ICS)</i> , 2017 |
| [3] | Xiuxia, Zhang; Guangming, Tan; Shuangbai, Xue; Jiajia, Li; Keren, Zhou; Mingyu, Chen: Understanding GPU Microarchitecture to Achieve Bare-Metal Performance Tuning. In: <i>22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPOPP)</i> , 2017 |
| [4] | Wei, Zhou; Keren, Zhou; Zhongzhi, Luan; Shaowen, Yao; Depei, Qian: Study on Multi-Core Data Structure in Shared-Memory. In: <i>Journal of Software</i> (2016), Nr. 4, S. 1009–1025 |
| [5] | Zilong, Tan; Keren, Zhou; Hao, Zhang; Wei, Zhou: BF-MapReduce: A Bloom Filter Based Efficient Lightweight Search. In: <i>International Conference on Collaboration and Internet Computing (CIC) on IEEE</i> , 2015 |
| [6] | Qiang, Li; Maojie, Gu; Keren, Zhou; Xiaoming, Sun: Mining User Features for Purchase Prediction in M-Commerce. In: <i>Data Mining Workshop (ICDMW), 2015 IEEE International Conference on IEEE</i> , 2015 |
| [7] | Wei, Zhou; Jin, Lu; Keren, Zhou; Shipu, Wang; Shaowen, Yao: Concurrent Skiplist Based Double-Layer Index Framework for Cloud Data Processing. In: <i>Journal of Computer Research and Development</i> (2015) |
| [8] | Keren, Zhou; Guocheng, Niu; Wuzhao, Zhang; Xueqi, Li; Wenqin, Liu: Parse Concurrent Data Structures: BST as an Example. In: <i>arXiv preprint arXiv:1505.03759</i> (2015) |
| [9] | Keren, Zhou; Qian, Yu; Zhenwei, Zhu; Wenjia, Liu: Dynamic Vegas: A Competitive Congestion Control Strategy. In: <i>Proceedings of International Conference on Computer Science and Information Technology Springer</i> , 2014, S. 333–340 |

AWARDS & HONORS

- | | |
|------|--|
| 2019 | Second Place, ACM CGO Student Research Competition |
| 2017 | Ken Kennedy Institute Andrew Ladd Fellowship |
| 2017 | Ken Kennedy Institute CS&E Fellowship |
| 2016 | National Scholarship (2%) |
| 2016 | Merit Student of Chinese Academy of Sciences |
| 2016 | Schlumberger Scholarship (3%) |
| 2015 | Top 10, Alibaba 1st Middleware Engineering Contest |
| 2014 | Bronze Medal, The 2014 ACM-ICPC Asia Anshan 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 | National Scholarship |
| 2011&2012 | Merit Student of Yunnan Province |

SKILLS

| | |
|--------------------|--|
| Languages | C, C++, Java, Python, Bash, JavaScript |
| Parallelism | Pthread, OpenMP, MPI, CUDA |