

# Keren Zhou

6100 Main ST – Houston, TX – 77005, United States  
☎ +1-281-687-6961    ✉ keren.zhou@rice.edu    🌐 www.jokeren.tech

## EDUCATION BACKGROUND

---

09/2017-07/2022	<b>Department of Computer Science, Rice University</b> <b>Expected Degree:</b> <i>Ph.D. in Computer Science</i> <b>GPA:</b> 3.9/4.0 <b>Advisor:</b> John Mellor-Crummey	<b>Houston, United States</b>
09/2014-07/2017	<b>Institute of Computing Technology, Chinese Academy of Sciences</b> <b>Degree:</b> <i>M.S. in Computer Architecture</i> <b>GPA:</b> 90/100 <b>Advisor:</b> Guangming Tan <b>Thesis:</b> High Performance Deep Learning Algorithms	<b>Beijing, China</b>
09/2010-07/2014	<b>School of Software, Yunnan University</b> <b>Degree:</b> <i>B.E. in Network Engineering</i> <b>GPA:</b> 92/100 (Rank: 1/290) <b>Advisor:</b> Wei Zhou <b>Thesis:</b> A Practical Concurrent Quadtree	<b>Kunming, China</b>

## RESEARCH EXPERIENCE

---

09/2017-NOW	<b>Rice University</b> <i>Research Assistant</i> <b>Scalable GPU Performance Measurement and Analysis Tool</b> <ul style="list-style-type: none"><li>◦ Implemented OpenMP 5.0 OMPT Tool Interface for CUDA backend in llvm-openmp;</li><li>◦ Built a general runtime system to collect GPU activities on NVIDIA, AMD, and Intel GPUs and attributed them back to the corresponding CPU calling context;</li><li>◦ Analyzed GPU binaries to extract GPU functions, recover control flows, and map instructions to source code;</li><li>◦ Associated runtime samples with static GPU program structures to reconstruct calling context on GPUs and estimate instruction throughput and roof-line model;</li><li>◦ Studied HPC and machine learning applications, including PyTorch, Darknet, Quicksilver, Nekbone, Laghos, PeleC, QMCPACK, Nyx, GAMESS, NAMD, and LAMMPS.</li></ul> <b>GPU Performance Advisor</b> <ul style="list-style-type: none"><li>◦ Devised a method to attribute instruction stalls back to the instructions that caused them;</li><li>◦ Built a profile-guided performance advisor based on GPU performance metrics, program structures, and PC samples;</li><li>◦ Derived performance models to estimate speedups of individual suggestions proposed by the advisor;</li><li>◦ Optimized GPU applications by applying advice generated by the advisor to obtain speedups on a V100 GPU with 1.22× on average.</li></ul> <b>GPU Value Redundancy Profiler</b> <ul style="list-style-type: none"><li>◦ Investigated value redundancy problems in HPC and machine learning applications and achieved speedups by up to 1.93×.</li><li>◦ Built the first value profiler for NVIDIA GPUs to explore both temporal and spatial value redundancies in multi-node multi-GPU clusters;</li><li>◦ Devised asynchronous analysis and hierarchical sampling methods to reduce the tool overhead to 7.5× on average for Rodinia benchmarks.</li></ul>	<b>Houston, United States</b>
06/2015-07/2017	<b>Institute of Computing Technology, Chinese Academy of Sciences</b> <i>Research Assistant</i> <b>GPU Performance Model</b> <ul style="list-style-type: none"><li>◦ Decoded Nvidia GPU assembly codes and developed assemblers to generate GPU binaries;</li><li>◦ Built a static performance analysis model to estimate performance bottlenecks in GPU binaries.</li></ul> <b>High Performance Deep Learning Framework</b> <ul style="list-style-type: none"><li>◦ 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;</li><li>◦ Wrote assembly codes to make use of dual issue and avoid bank conflict on GPU, improving convolution performance with up to 1.6× speedup than cuDNN on Kepler architectures.</li></ul>	<b>Beijing, China</b>

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  $1.09\times$  faster than state-of-the-art concurrent trees;

## SELECTED PUBLICATIONS

---

### JOURNALS.....

- [1]        **Keren Zhou**, Guangming Tan, and Wei Zhou. Quadboost: A Scalable Concurrent Quadtree. In: *IEEE Transactions on Parallel and Distributed Systems* (TPDS), 2018

### CONFERENCES.....

- [1]        **Keren Zhou**, Xiaozhu Meng, Ryuichi Sai, John Mellor-Crummey. GPA: A GPU Performance Advisor Based on Instruction Sampling. In: *International Symposium on Code Generation and Optimization* (CGO), 2021
- [2]        **Keren Zhou**, Yueming Hao, John Mellor-Crummey, Xiaozhu Meng, and Xu Liu. GVProf: A Value Profiler for GPU-based Clusters. In: *The International Conference for High Performance Computing, Networking, Storage, and Analysis* (SC), 2020
- [3]        **Keren Zhou**, Mark Krentel, and John Mellor-Crummey. Tools for top-down performance analysis of GPU-accelerated applications. In: *34th ACM International Conference on Supercomputing* (ICS), 2020
- [4]        **Keren Zhou**, Guangming Tan, Xiuxia Zhang, Chaowei Wang, and 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**, and 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

### POSTERS.....

- [1]        **Keren Zhou**, Mark Krentel, and John Mellor-Crummey. A tool for top-down performance analysis of GPU-accelerated applications. In: *25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (PPoPP), 2020
- [2]        **Keren Zhou** and John Mellor-Crummey. A tool for performance analysis of GPU-accelerated applications. In: *International Symposium on Code Generation and Optimization* (CGO), 2019

## INDUSTRY EXPERIENCE

---

- 05/2020-08/2020   **C++ Performance Optimization Team, Google Inc.**   **Houston, United States**  
*Software Engineering Intern*
  - Developed AutoDiff, a performance regression analysis tool that locates performance difference ranges, provide line and call site information, and allows flexible queries;
  - Proposed a method to recover high-resolution calling context with minimal overhead by augmenting call stacks for each instruction in LBR entries;
  - Reference: Software Engineer Wei Mi, wmi@google.com.
- 06/2018-08/2018   **PyTorch Team, 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   **Devtech Team, 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.

<b>10/2013-02/2014</b>	<b>Baidu Inc.</b> <i>Software Engineering Intern</i>	<b>Beijing, China</b>
	<ul style="list-style-type: none"> <li>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;</li> <li>Developed a Hadoop workflow monitoring system that can display multiple workflow status and report exceptions;</li> <li>Reference: Senior Engineer Jing Li, lijing16@baidu.com.</li> </ul>	

## AWARDS & HONORS

---

<b>2020</b>	ACM–IEEE-CS George Michael Memorial HPC Fellowship
<b>2019</b>	Ken Kennedy Institute ExxonMobil Fellowship
<b>2019</b>	Second Place, ACM CGO Student Research Competition
<b>2017</b>	Ken Kennedy Institute Andrew Ladd Fellowship
<b>2017</b>	Ken Kennedy Institute CS&E Fellowship
<b>2016</b>	Schlumberger Scholarship (3%)
<b>2015</b>	Top 10, Alibaba 1st Middleware Engineering Contest
<b>2014</b>	Bronze Medal, The 2014 ACM-ICPC Asia Regional Contest
<b>2014</b>	Outstanding B.E. Degree Thesis of Yunnan University
<b>2013</b>	Best Creative Award, Baidu Future Search Engine Contest
<b>2013</b>	Meritorious Winner, Mathematical Contest in Modeling
<b>2011</b>	Second Prize, China Undergraduate Mathematical Contest in Modeling
<b>2011&amp;2012&amp;2016</b>	National Scholarship (2%)

## ACADEMIC SERVICES

---

<b>ICS'21</b>	Conference paper subreviewer
<b>ICDCS'21</b>	Conference paper subreviewer
<b>IPDPS'21</b>	Conference paper subreviewer
<b>PPoPP'21</b>	Artifact Evaluation Committee member

## SKILLS

---

<b>Languages</b>	C, C++, Java, Python, Bash, Go, JavaScript
<b>Parallelism</b>	Pthread, OpenMP, OpenCL, MPI, CUDA/HIP, DPCPP, RAJA/Kokkos