Keren Zhou

3180 18th St – San Francisco, CA – 94110, United States
+1-281-687-6961 keren.zhou@rice.edu www.jokeren.tech

RESEARCH INTERESTS

High Performance Computing
Parallel Algorithms
Program Analysis
Tools for Machine Learning Systems

EDUCATION

08/2017-05/2022 **Department of Computer Science, Rice University** **Houston, United States**
Degree: *Ph.D. in Computer Science*
Advisor: John Mellor-Crummey
Thesis: Performance Measurement, Analysis, and Optimization of GPU-accelerated Applications

08/2014-08/2017 **Institute of Computing Technology, Chinese Academy of Sciences** **Beijing, China**
Degree: *M.S. in Computer Architecture*
Advisor: Guangming Tan **Thesis:** High Performance Deep Learning Algorithms

09/2010-08/2014 **School of Software, Yunnan University** **Kunming, China**
Degree: *B.E. in Network Engineering* **Rank:** 1/290
Advisor: Wei Zhou **Thesis:** A Practical Concurrent Quadtree

AWARDS & HONORS

2022 ASPLOS Distinguished Artifact Award
2020 ACM–IEEE-CS George Michael Memorial HPC Fellowship
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
2017 PPoPP Best Artifact Award
2016 Merit Student of Chinese Academy of Sciences
2016 Schlumberger Scholarship
2015 Top 10, Alibaba 1st Middleware Engineering 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&2012&2016 National Scholarship

PROFESSIONAL EXPERIENCE

06/2022-current *Member of Technical Staff* at **OpenAI** **San Francisco, United States**
08/2017-05/2022 *Research Assistant* at **Rice University** **Houston, United States**
05/2021-08/2021 *Intern* at Deep Learning Profiler Team, **Nvidia** **Dallas, United States**
05/2020-08/2020 *Intern* at C++ Performance Optimization Team, **Google** **Houston, United States**
06/2018-08/2018 *Intern* at PyTorch Team, **Facebook** **Menlo Park, United States**
06/2015-07/2017 *Research Assistant* at **Chinese Academy of Sciences** **Beijing, China**
04/2017-07/2017 *Intern* at Devtech Team, **Nvidia** **Beijing, China**
10/2013-02/2014 *Intern* at **Baidu** **Beijing, China**

PUBLICATIONS

JOURNALS

[1] Binqian Yin, Qinhong Hu, Yingying Zhu, Chen Zhao, and **Keren Zhou**. Paw-Net: Stacking ensemble deep learning for segmenting scanning electron microscopy images of fine-grained shale samples. In: *Computers & Geosciences*, 2022

[2] **Keren Zhou**, Laksono Adhianto, Jonathon Anderson, Aaron Cherian, Dejan Grubisic, Mark. Krentel, Yumeng Liu, Xiaozhu Meng, John Mellor-Crummey. Measurement and Analysis of GPU-accelerated Applications with HPCToolkit. In: *Parallel Computing* (PARCO), 2021

[3] Ryuichi Sai, John Mellor-Crummey, Xiaozhu Meng, **Keren Zhou**, Mauricio Araya-Polo, Jie Meng. Accelerating High-Order Stencils on GPUs. In: *Concurrency and Computation: Practice and Experience, 2021*

[4] **Keren Zhou**, Xiaozhu Meng, Ryuichi Sai, Dejan Grubisic, and John Mellor-Crummey. An Automated Tool for Analysis and Tuning of GPU-accelerated Code in HPC Applications. In: *IEEE Transactions on Parallel and Distributed Systems* (TPDS), 2021

[5] **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**, Jonathon Anderson, Xiaozhu Meng, and John Mellor-Crummey. Low Overhead and Context Sensitive Profiling of GPU-accelerated Applications. In: *ACM International Conference on Supercomputing* (ICS), 2022

[2] **Keren Zhou***, Yueming Hao*, John Mellor-Crummey, Xiaozhu Meng, and Xu Liu. ValueExpert: Exploring Value Patterns in GPU-accelerated Applications. In: *Conference on Architectural Support for Programming Languages and Operating Systems* (ASPLOS), 2022

[3] Aaron Thomas Cherian, **Keren Zhou**, Dejan Grubisic, Xiaozhu Meng, and John Mellor-Crummey. Measurement and Analysis of GPU-Accelerated OpenCL Computations on Intel GPUs. In: *Workshop on Programming and Performance Visualization Tools* (ProTools), 2021

[4] Barbara Chapman, Buu Pham, Charlene Yang, Christopher Daley, Colleen Bertoni, Dhruva Kulkarni, Dossay Oryspayev, Ed D'Azevedo, Gabriele Jost, Johannes Doerfert, **Keren Zhou**, Kiran Ravikumar, Mark Gordon, Mauro Del Ben, Meifeng Lin, Melisa Alkan, Michael Kruse, Oscar Hernandez, P.K. Yeung, Paul Lin, Peng Xu, Swaroop Pophale, Tosaporn Sattasathuchana, Vivek Kale, William Huhn, and Helen He. Outcomes of OpenMP Hackathon: OpenMP Application Experiences with the Offloading Model. In: *International Workshop on OpenMP* (IWOMP), 2021

[5] **Keren Zhou**, Xiaozhu Meng, Ryuichi Sai, and John Mellor-Crummey. GPA: A GPU Performance Advisor Based on Instruction Sampling. In: *International Symposium on Code Generation and Optimization* (CGO), 2021

[6] **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

[7] **Keren Zhou**, Mark Krentel, and John Mellor-Crummey. Tools for top-down performance analysis of GPU-accelerated applications. In: *ACM International Conference on Supercomputing* (ICS), 2020

[8] **Keren Zhou**, Guangming Tan, Xiuxia Zhang, Chaowei Wang, and Ninghui Sun. A Performance Analysis Framework for Exploiting GPU Microarchitectural Capability. In *ACM International Conference on Supercomputing* (ICS), 2017

[9] Xiuxia Zhang, Guangming Tan, Shuangbai Xue, Jiajia Li, **Keren Zhou**, and Mingyu Chen. Understanding GPU Microarchitecture to Achieve Bare-Metal Performance Tuning. In: *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (PPOPP), 2017

[10] Zilong Tan, **Keren Zhou**, Hao Zhang, and Wei Zhou. BF-MapReduce: A Bloom Filter Based Efficient Lightweight Search. In: *International Conference on Collaboration and Internet Computing on IEEE* (CIC), 2015

[11] Qiang Li, Maojie Gu, **Keren Zhou**, Xiaoming Sun. Mining User Features for Purchase Prediction in M-Commerce. In: *Data Mining Workshop, IEEE International Conference on IEEE* (ICDMW), 2015

PRESENTATIONS

12/2022 **Invited Talk**, *UC Merced*, Towards Agile Development of Efficient Deep Learning Operators

05/2022 **Invited Talk**, *ThirdAI*, Practical Performance Optimization for Deep Learning Applications

03/2022 **Conference Talk**, *Conference on Architectural Support for Programming Languages and Operating Systems* (ASPLOS), ValueExpert: Exploring Value Patterns in GPU-accelerated Applications

11/2021 **Conference Talk**, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (SC), Performance Measurement, Analysis, and Optimization of GPU-accelerated Applications

04/2021 **Invited Talk**, *Nvidia GPU Technology Conference* (GTC), Measurement and Analysis of GPU-accelerated Applications with HPCToolkit

04/2021 **Tutorial**, *ECP Annual Meeting*, Using HPCToolkit for performance analysis on GPU-accelerated applications

03/2021 **Tutorial**, *NERSC*, Using HPCToolkit to Measure and Analyze the Performance of GPU-accelerated Applications

03/2021 **Conference Talk**, *IEEE/ACM International Symposium on Code Generation and Optimization* (CGO), GPA: A GPU Performance Advisor Based on Instruction Sampling

11/2020 **Conference Talk**, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (SC), GVProf: A Value Profiler for GPU-Based Clusters

07/2020 **Conference Talk**, *Proceedings of the ACM International Conference on Supercomputing* (ICS), Tools for Top-down Performance Analysis of GPU-Accelerated Applications

02/2020 **Tutorial**, *ECP Annual Meeting*, Using HPCToolkit to Measure and Analyze the Performance of GPU-Accelerated Applications

10/2019 **Invited Talk**, *BP*, Measurement and Analysis of GPU-computations Using HPCToolkit

08/2019 **Invited Talk**, *Intel Performance Brown Bag*, HPCToolkit—A tool for performance analysis for GPU-accelerated applications

08/2019 **Invited Talk**, *ECP/NERSC OpenMP Hackathon*, HPCToolkit + OpenMP

07/2019 **Conference Talk**, *Scalable Tools Workshop*, Optimizing GPU-accelerated Applications with HPCToolkit

06/2017 **Conference Talk**, *Proceedings of the International Conference on Supercomputing* (ICS), A performance analysis framework for exploiting GPU microarchitectural capability

ACADEMIC SERVICES

Reviewer ASPLOS'23, SC'22, ICS'21, ICDCS'21, IPDPS'21, CLUSTER'21, PPoPP'21, TPDS, JPDC, TECS, TJSC

AE Committee EuroSys'22, PPoPP'22, PPoPP'21, LCTES'21, SOSP'21

Session Chair CLUSTER'21

Projects

09/2017-05/2022 Rice University **Houston, United States**

Scalable GPU Performance Measurement and Analysis Tool

- Built a general *context-sensitive profiling tool* that collects and analyzes GPU activities on Nvidia, AMD, and Intel GPUs;
- Studied HPC and machine learning applications, including TensorFlow, PyTorch, Darknet, Quicksilver, Nekbone, Laghos, PeleC, QMCPACK, Nyx, Castro, GAMESS, NAMD, SUPERLU, and LAMMPS.

GPU Performance Advisor

- Built a *profile-guided performance advisor* based on GPU performance metrics, program structure, instruction counts, and PC samples;
- Optimized GPU applications by applying advice generated by the advisor to obtain speedups on V100 and A100 GPUs with 1.19× on average.

GPU Value Profiler

- Developed the first *value profiler* for Nvidia GPUs to explore inefficient value patterns in applications running on multi-node multi-GPU clusters;
- Devised innovative instrumentation callbacks, sampling methods, and on-the-fly data processing GPU kernels to reduce the profiling overhead.

06/2015-07/2017 Institute of Computing Technology, Chinese Academy of Sciences **Beijing, China**

High Performance Deep Learning Framework

- Devised a coarse-grained parallelism strategy with fine-grained vectorization and blocking, making CNNs 5-12 times faster than Caffe on a 16-core E5-2670;
- Wrote assembly code to make use of dual issue and avoid bank conflict on GPUs, improving convolution performance with up to 1.6× speedup than cuDNN on Kepler architectures.

GPU Performance Model

- Decoded Nvidia GPU assembly codes and developed assemblers to generate GPU binaries;
- Built a static performance analysis model to estimate performance bottlenecks in GPU binaries.

01/2013-07/2014 Intelligent Web Laboratory, Yunnan University **Kunming, China**

Concurrent Data Structures

- Designed several concurrent multi-dimensional trees, including the first lock-free quadtree and k-d tree that are 2.09× faster than state-of-the-art concurrent trees;