Keren Zhou

3180 18th St – San Francisco, CA – 94110, United States
+1-281-687-6961 ✉ kerenzhou@outlook.com www.jokeren.tech

EDUCATION

09/2017-05/2022 **Department of Computer Science, Rice University** **Houston, United States**
Degree: *Ph.D. in Computer Science* **Advisor:** John Mellor-Crummey

09/2014-07/2017 **Institute of Computing Technology, Chinese Academy of Sciences** **Beijing, China**
Degree: *M.S. in Computer Architecture* **Advisor:** Guangming Tan

09/2010-07/2014 **School of Software, Yunnan University** **Kunming, China**
Degree: *B.E. in Network Engineering* **Advisor:** Wei Zhou

RESEARCH EXPERIENCE

09/2017-05/2022 **Rice University** **Houston, United States**
Performance Profiling, Analysis, and Optimization of GPU-accelerated Applications

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

INDUSTRY EXPERIENCE

06/2022-current **OpenAI LP** **San Francisco, United States**
- Developing deep learning compilers and profilers;
- Reference: Head of Compute, Chris Berner, christopher.berner@openai.com.

05/2021-08/2021 **Nvidia Inc.** **Dallas, United States**
- Profiled input and output of deep learning operators to enhance performance reports and recover the computation graph;
- Reference: Software Manager, Timothy Gerdes, tgerdes@nvidia.com.

05/2020-08/2020 **Google Inc.** **Houston, United States**
- Performance Regression Analysis of Feedback-direct Optimization (FDO) Based Programs;
- Reference: Software Engineer, Wei Mi, wmi@google.com.

06/2018-08/2018 **Facebook Inc.** **Menlo Park, United States**
- Accelerated neural networks on ARM CPUs using auto-tuning methods;
- Reference: Research Scientist, Hao Lu, hlu@fb.com.

04/2017-07/2017 **Nvidia Inc.** **Beijing, China**
- Developed quantization tools on emerging GPUs to utilize INT8 capabilities;
- Reference: Technical Manager, Julien Lai, julienlai@nvidia.com.

10/2013-02/2014 **Baidu Inc.** **Beijing, China**
- Optimized Hadoop workflow and improved its performance by 30%;
- Reference: Senior Engineer, Jing Li, lijing16@baidu.com.

SELECTED PUBLICATIONS

[1] **Keren, Zhou**; Mark, Krentel; John, Mellor-Crummey: Tools for top-down performance analysis of GPU-accelerated applications. In: *34th ACM International Conference on Supercomputing* (ICS), 2020

[2] **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

AWARDS & HONORS

2020 ACM–IEEE-CS George Michael Memorial HPC Fellowship

2019 Ken Kennedy Institute ExxonMobil Fellowship

2017 Ken Kennedy Institute Andrew Ladd Fellowship

2017 Ken Kennedy Institute CS&E Fellowship