Yuxin Chen

yxxchen@ucdavis.edu

EDUCATION

University of California, Davis (UC Davis), The United State
Ph.D in Computer Science — Sep 2016 – Present

King Abdullah University of Science and Technology (KAUST), Saudi Arabia
M.S in Computer Science — Sep 2013 – May 2015

Xiamen University, China
Bachelor of Management — Sep 2009 – May 2013
Bachelor of Economy — Sep 2011 – May 2013

PUBLICATIONS

Yuxin Chen, Benjamin Brock, Serban Porumbescu, Aydın Buluç, Katherine Yelick, and John D. Owens. Atos: A Task-Parallel GPU Scheduler for Graph Analytics. In Proceedings of the International Conference on Parallel Processing, ICPP 2022, August/September 2022.

Yuxin Chen, Benjamin Brock, Serban Porumbescu, Aydın Buluç, Katherine Yelick, and John D. Owens. Scalable irregular parallelism with GPUs: Getting CPUs out of the way. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, SC ’22, November 2022.

Taylor Groves, Ben Brock, **Yuxin Chen**, Khaled Z. Ibrahim, Lenny Oliker, Nicholas J. Wright, Samuel Williams, and Katherine Yelick. 2020. Performance Trade-offs in GPU Communication: A Study of Host and Device-initiated Approaches. In 2020 IEEE/ACM Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS). 126–137.

Benjamin Brock, **Yuxin Chen**, Jiakun Yan, John D. Owens, Aydın Buluç, and Katherine Yelick. RDMA vs. RPC for Implementing Distributed Data Structures. In Proceedings of the IEEE/ACM 9th Workshop on Irregular Applications: Architectures and Algorithms, IA3 2019, pages 17–22, November 2019.

Yuxin Chen, David Keyes, Kody JH Law, and Hatem Ltaief. ”Accelerated dimension-independent adaptive Metropolis.” SIAM Journal on Scientific Computing 38, no. 5 (2016): S539-S565.

RESEARCH EXPERIENCE

Global Address Programming with GPUs,
UC Davis — Jan 2018 – Present

- I developed a programming environment for GPU-based HPC systems that integrates GPUs into Partitioned Global Address Space (PGAS) model. I built a set of high-performance, open-source data structures and algorithm implementations to support irregular patterns of communication, notably those that arise in biology, graph analytics, and sparse linear algebra. These not only are directly useful for end users but also demonstrate how to design and engineer primitives for accelerator-equipped distributed-memory machines.

Accelerated Hash Tables for Membership Queries
Lawrence Livermore National Laboratory, Livermore — Feb 2019 – May 2019

- I accelerated the hash tables member queries by maintaining the same-size fingerprint table with each fingerprint slot takes only 8/16 bits. In the case of insertion fails, instead of recreating a larger hash table, I used an overflow buffer to hold the failed insertions, where the key can be searched efficiently using GPU multi-threading. This hash table can efficiently filter out false queries and had overflow ratio less than 0.1% with varies load factor when tested on RMAT datasets.

Auto-Tuning Hyperparameters of DNNs,
Adobe, San Jose — Sep 2017 – Jun 2017

- I used Bayesian Optimization to auto-tune hyperparameters and system configurations for training deep neural networks. For Inception model on Imagenet, the auto-tuner can reduce the search space by about half.

Enhanced Apache Spark to utilize GPUs,
UC Davis — Jan 2017 – Jun 2017

- Enhancing Apache Spark to utilize GPUs. Following the design philosophy of Apache Spark, I extended the basic data structure (RDD) to utilize GPUs in Spark. The accessed data is automatically transferred between CPU and GPU, and this process is transparent to the users.
- Our GPU-utilized map and reduce operations are twice faster than the one use only CPU. Our GPU-utilized logistic regression is five times faster than the CPU implementation.