Yuxin Chen

Personal Website yxxchen@ucdavis.edu
+1-484-790-0788 YuxinxinChen.github.io

ABOUT ME

I have extensive experience in GPU programming, specializing in multi-GPU communication and GPU runtime design. My expertise includes developing high-performance computing frameworks, optimizing deep learning models, and enhancing large-scale distributed systems.

EDUCATION

Ph.D. in Computer Science 2016–2024
University of California, Davis, USA

M.S. in Computer Science 2013–2015
KAUST, Saudi Arabia

PROFESSIONAL EXPERIENCES

Platform Engineer (GPU/CUDA) 2023–Now
Luminary Cloud Inc.

- Developed toolkits for automatically optimizing partial differential equation solvers on GPUs (and multi-GPU systems) using CUDA.
- Increased simulation speed by 2-100x, significantly enhancing the efficiency of customer design iterations.

Deep Learning Architecture Intern 2023 Summer
NVIDIA

- Enhanced Triton to generate multiple device function PTX for the NVIDIA Grafia from Triton Python code, without exposing Grafia technical details.

Research Assistant 2017–2024
University of California, Davis

- *ACCELERATING EMBEDDING TABLE RETRIEVAL FOR DLRM*
- Proposed and implemented a GPU direct one-sided asynchronous communication for multi-GPU embedding retrieval in deep learning recommendation models (DLRM). This approach, integrated into a CUDA backend for PyTorch, achieved 2.63x speedup over the baseline using NCCL.

- *Design and Prototype Multi-GPU PGAS Programming Model*
- Developed a Multi-GPU graph framework: Atos with authentic multi-GPU PGAS programming model. Atos achieved 1–500× on InfiniBand and NVLink distributed GPU systems, surpassing the performance of current state-of-the-art graph frameworks.

- *GPU utilized Apache Spark*
- Developed a GPU version of Spark's Resilient Distributed Datasets (RDDs) and accelerated key operations (map, reduce, logistic regression), achieving up to 5x speedup over CPU-only versions, while maintaining compatibility and efficiency within existing Spark workflows.

Software Intern 2018 Summer
NVIDIA

- Developed a communication aggregator that significantly improves bandwidth utilization by enabling dynamic communication aggregation between GPUs, decoupling communication granularity from computation, and allowing users to send messages in their algorithm's natural logic flow.

SELECTED PUBLICATIONS

- Xiang Cheng, **Yuxin Chen**, Suvrit Sra. **Transformers** Implement Functional Gradient Descent to Learn Non-Linear Functions In Context. In Proceedings of the International Conference on Machine Learning (**ICML**) 2024
- **Yuxin Chen**, Aydın Buluç, Katherine Yelick, and John D. Owens. Accelerating Multi-GPU Embedding Retrieval with PGAS-Style Communication for **Deep Learning Recommendation Systems**. Submitted to The 7th Annual Parallel Applications Workshop, Alternatives To MPI+X (*PAW-ATM'24*)
- **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*) IEEE Press, Article 50, 1–16.
- **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*). Association for Computing Machinery, New York, NY, USA, Article 50, 1–11.

REFERENCES

References available upon request. More details about papers and projects can be found on my personal website.