Yangzihao Wang

E slashspirit@gmail.com
W yzhwang.github.io

Education

Ph.D., Computer Science, University of California, Davis. 2011–2016
Advisor: Prof. John D. Owens
dissertation: Gunrock: A Programming Model and Implementation for Graph Analytics on Graphics Processing Units

M.E., Software Engineering, Beihang University. 2008–2011

B.E., Computer Science, Beihang University. 2003–2007

Honors and Awards

2016: Distinguished Paper Award, ACM SIGPLAN PPoPP

2014: NVIDIA Graduate Fellowship Finalist

Experience Highlights

Founding Research Engineer, SEA AI Lab. 2021/04–2023/08
AI for System Research, Neural Graphics Research, Game Agent System Research

Senior Software Engineer, WeChat at Tencent Beijing. 2019/09–2021/04
PlatoDeep:
Designed and implemented GPU module of Tencent-wide graph neural network framework.
Achieved **10x speedup at kernel level and 2-3x end-to-end speedups** for GNN training.
Company-wide open-sourced.
PlatoEmbedding:
Led a 3 people team on design and implementation of Wechat-wide multi-GPU graph embedding system.
Trained on a 300 billion graph within 3 minutes on five 8-GPU nodes.
Company-wide open-sourced, and one SIGMOD'21 paper in submission.
Currently working on distributed GPU random-walk and sampling library.
WeChat 视频号 Backend:
Extended TensorFlow with dynamic sparse embedding module to support training and deployment of several distributed short-video recommendation models (FM, DeepFM, and DIN).
To be merged into TensorFlow. Published one paper on SIGIR'20.

.

Senior Software Engineer, Tencent Technology (Beijing) Co., Ltd.. 2018/07–2019/08
Taichi Platform:
Lead architect of Tencent-wide deep learning training platform (5K internal users).
Co-developed hierarchical AllReduce algorithm and tensor-fusion algorithm that helped **break the world record of ImageNet Training in 2018 (trained ImageNet on 1024 GPUs within 4 minutes)**.
Published one paper on NeurIPS 2018 System for ML workshop.
Tianfeng AutoML platform:
Lead architect and developer of Tencent-wide AutoML platform. Designed and developed first version of the platform. Integrated most SOTA AutoML methods.
Used for Honor of Kings (HoK) Game AI project. Helped increase the winning rate for several 1v1 models from 50% to over 70%.
Won company-wide business breakthrough prize of 2019 with HoK Game AI team.

.

Software Engineer, Google Brain. 2017/01–2018/05
TensorFlow Performance:
Developed fused Conv2D kernel for Waymo. Increased inference performance by 50%.
Developed dilated Convolution kernel for DeepMind. Used in WaveNet.
Performance optimization for matmul, transpose, and FFT kernels.
Implemented several optimization strategies such as topology-aware AllReduce and convolution/matmul autotune features for TensorFlow's official CNN benchmark.

.

Intern, Google. 2015/08–2015/10
web page classifier for search infrastructure team

Intern, AMD Research. **2012/06–2012/09**
software rasterizer using OpenCL

Graduate Student Researcher, Institute of Data Analysis and Visualization, UC Davis. **2011/08–2016/12**

Gunrock: Created the fastest GPU Graph Processing Library.
Used as one backend in NVIDIA's cuGraph. Open-sourced on github: https://github.com/gunrock/gunrock
XDATA: From 2014 to 2016, applied Gunrock to several practical research problems by Defense Advanced Research Projects Agency (DARPA).
Projects include anomaly detection/subgraph matching on social networks and centrality analysis on bitcoin transaction networks.

Professional Skills

Proficient:: C/C++, CUDA, TensorFlow, Python, MPI, LaTeX, git, Linux development

Familiar:: OpenGL, Spark

Professional Service

Journal/Conference Reviewer:: TC, TPDS, JDPC, PeerJ, PLDI'18, SC'20, WWW'21

Conference Program Committee Member::

The 1st GPUTech Workshop at ICCSA 2016

Graph Algorithms Building Blocks Workshop at IPDPS 2018

Workshop on Graphs, Architectures, Programming, and Learning at IPDPS 2019

SuperComputing 2020, Machine Learning and HPC Track

Selected Publications

Haidong Rong, Yangzihao Wang, Feihu Zhou, Junjie Zhai, Haiyang Wu, Rui Lan, Fan Li, Han Zhang, Yuekui Yang, Zhenyu Guo, and Di Wang. Distributed equivalent substitution training for large-scale recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 911–920, New York, NY, USA, 2020. Association for Computing Machinery.

Wanjing Wei, Yangzihao Wang, Pin Gao, Shijie Sun, and Donghai Yu. A distributed multi-gpu system for large-scale node embedding at tencent, 2020.

Xianyan Jia, Shutao Song, Wei He, Yangzihao Wang, Haidong Rong, Feihu Zhou, Liqiang Xie, Zhenyu Guo, Yuanzhou Yang, Liwei Yu, Tiegang Chen, Guangxiao Hu, Shaohuai Shi, and Xiaowen Chu. Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes. In *Workshop on Systems for ML and Open Source Software*, NeurIPS 2018, December 2018.

Oh Chin-Yang, Kunhao Zheng, Bingyi Kang, Xinyi Wan, Zhongwen Xu, Shuicheng Yan, Min Lin, and Yangzihao Wang. Hloenv: A graph rewrite environment for deep learning compiler optimization research, December 2022.

Yangzihao Wang, Yuechao Pan, Andrew Davidson, Yuduo Wu, Carl Yang, Leyuan Wang, Muhammad Osama, Chenshan Yuan, Weitang Liu, Andy T. Riffel, and John D. Owens. Gunrock: GPU graph analytics. *ACM Transactions on Parallel Computing*, 2017.

Yuechao Pan, Yangzihao Wang, Yuduo Wu, Carl Yang, and John D. Owens. Multi-GPU graph analytics. In *Proceedings of the 31st IEEE International Parallel and Distributed Processing Symposium*, IPDPS 2017, May/June 2017.

Yangzihao Wang, Andrew Davidson, Yuechao Pan, Yuduo Wu, Andy Riffel, and John D. Owens. Gunrock: A high-performance graph processing library on the GPU. In *Proceedings of the 21st ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, PPoPP 2016, March 2016. Distinguished Paper.

Yuduo Wu, Yangzihao Wang, Yuechao Pan, Carl Yang, and John D. Owens. Performance characterization for high-level programming models for GPU graph analytics (best paper finalist). In *IEEE International Symposium on Workload Characterization*, IISWC 2015, October 2015. Best Paper finalist.