Carl Yang

Curriculum Vitae

1600 Amphitheatre Pkwy
Mountain View, CA 94043
✉ *ctcyang@ucdavis.edu*
www.ece.ucdavis.edu/~ctcyang

Education

2014–2019 **Ph.D., Computer Engineering**, *University of California*, Davis.
- Research Area: Parallel computing, sparse matrix multiplication and graph algorithms
- Advisors: Prof. John D. Owens and Prof. Aydın Buluç (UC Berkeley, LBL)

2006–2010 **B.A.Sc., Electrical Engineering**, *University of British Columbia*, Vancouver.
- Research Area: Mathematical modeling of MEMS

Experience

Jul-2019–Present **Software Engineer**, *Waymo LLC*, Mountain View.
Router team.
- Improve routing for self-driving car.

Jan-2016–Jun-2019 **Graduate Student Research Assistant**, *Lawrence Berkeley National Laboratory*, Berkeley.
Building multi-GPU implementation of sparse matrix multiplication for GraphBLAS.
- Built GraphBLAST: A high-performance linear algebra-based graph framework that erased a 10× performance gap compared to vertex-centric graph frameworks.
- Software available here: https://github.com/gunrock/graphblast

Apr-2018–Sep-2018 **Software Engineer Intern**, *Amazon AWS*, Palo Alto.
Single- and Multi-Machine Training for Deep Learning.
- Integrated Horovod with MXNet deep learning framework to achieve 91% scalability on 256 GPUs
- Collaborated with NVIDIA to train ResNet-50 on ImageNet in 6.3 minutes for first place in MLPerf v0.5 benchmark: https://devblogs.nvidia.com/new-optimizations-accelerate-deep-learning-training-gpu/
- Software available here: https://github.com/uber/horovod/pull/542

Jun-2017–Sep-2017 **Research Intern**, *NVIDIA*, Santa Clara.
Building multi-GPU graph clustering tool.
- Used Markov Chain Monte Carlo (MCMC) sampling to find optimal partition as defined by the stochastic blockmodel, a type of Bayesian mixture model for graphs.
- Used CUDA to write GPU implementation of parallel MCMC.
- Achieved $12\times$ performance speed-up compared to OpenMP reference implementation.

Jul-2015–Aug-2015 **Student Researcher**, *DARPA*, Arlington.
Apply Gunrock to solve big data analytic challenges.
- Investigate how Gunrock can solve problems for US governmental organizations.
- Analyze geodata based on social media to aid UN peacekeeping first response efforts in Yemen.
- Detect patent troll cases for US Patent and Trademark Office using open data.
- Contributor to Gunrock: High-Performance GPU Graph Processing Library available at http://gunrock.github.io

Jan-2015– Dec-2015 **Graduate Student Researcher**, *University of California*, Davis.

Built graph algorithms based on shared-memory parallelism of GPU.

- Investigate linear algebraic formulations of graph algorithms.
- Formulated first sparse matrix-sparse vector (SpMSpV) kernel on GPU that generalizes to arbitrary semi-rings with applications in breadth-first-search, single-source shortest-path and maximal independent set.
- Provides up to 3.3x speed-up over comparable sparse matrix-dense vector (SpMV) kernel.

Professional skills

Proficient C, C++, CUDA, LaTeX, Linux development

Familiar Python, MPI, HTML, Bash, Git, Windows development

Publications

Carl Yang, Aydın Buluç, John D. Owens. GraphBLAST: A High-Performance Linear Algebra-based Graph Framework on the GPU. Currently under review, August 2019.

Carl Yang, Aydın Buluç, John D. Owens. Design Principles for Sparse Matrix Multiplication on the GPU. International European Conference on Parallel and Distributed Computing (Euro-Par), August 2018. Distinguished Paper Award, Best Artifact Award.

Carl Yang, Aydın Buluç, John D. Owens. Implementing Push-Pull Efficiently in GraphBLAS. International Conference on Parallel Processing (ICPP), August 2018.

Tim Mattson, **Carl Yang**, Scott McMillan, Aydın Buluç and José Moreira. GraphBLAS C API: Ideas for future versions of the specification. IEEE High Performance Extreme Computing (HPEC), September 2017.

Yuechao Pan, Yangzihao Wang, Yuduo Wu, **Carl Yang** and John D. Owens. Multi-GPU Graph Analytics. IEEE International Parallel and Distributed Parallel Symposium (IPDPS), May 2017.

Aydın Buluç, Tim Mattson, Scott McMillan, José Moreira and **Carl Yang**. Design of the GraphBLAS API for C. Workshop on Graph Algorithm Building Blocks (GABB), May 2017

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 (TOPC), January 2017.

Leyuan Wang, Yangzihao Wang, **Carl Yang** and John D. Owens. A Comparative Study on Exact Triangle Counting Algorithms on the GPU. In *High Performance Graph Processing Workshop (HPGP)*. May 2016

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

Carl Yang, Yangzihao Wang and John D Owens. Fast sparse matrix and sparse vector multiplication algorithm on the gpu. In *Workshop on Graph Algorithm Building Blocks (GABB)*, May 2015.