Yunming Zhang

E-Mail: zhangyunming1990@gmail.com; yunmingz@google.com, Personal Website: https://yunmingzhang17.github.io/

Overview

I am currently a Senior Software Engineer at Google working on the XLA Compiler for TPUs.

- **High-Performance Compiler:**
  - Worked on the XLA Compiler for next-gen TPUs with a focus on large-scale distributed training of embedding-based workloads. Designed and implemented the distributed execution infrastructure, scalable collective algorithms, and scheduling of HLOs to overlap the communications with computations. Contributed to the and fusion of HLOs. I worked with HLOs, MLIR, and LLVM-based Intermediate Representations.
  - Led the development of an open-source domain-specific language and compiler, *GraphIt*, used by Nvidia, UW, and Cornell for CPUs, GPUs, and many-core accelerators.
- **High-Performance Kernels / Performance Engineering:**
  - Optimizing the performance of ML workloads on TPUs with a specialization in embedding-based workloads. I also worked on improving the performance of Neural Radiance Fields (NeRF), Graph Neural Networks (GNN), and approximate Top-K.
  - Optimizing performance of sparse workloads, such as graph algorithms, sparse linear algebra, and approximate nearest neighbor search (e.g. Hierarchical Navigable Small World Graphs) on CPUs.
- **Programming Languages:** Experienced with C++, C, Python, and Java. Familiar with Go and MATLAB.

Education

Massachusetts Institute of Technology **June 2014 – July 2020**
Doctor of Philosophy in Computer Science, Cumulative GPA: **5.0/5.0**
Advisors: Prof. Saman Amarasinghe, Julian Shun
Focus: High-Performance Systems and Compilers for Large-Scale Graph Analytics and Sparse Computations

Rice University, Houston, Texas **May 2013 - May 2014**
Master of Science in Computer Science, Cumulative GPA: **4.0/4.33**
Advisors: Prof. Vivek Sarkar, Alan L. Cox
Focus: Optimizing Multi-Core Performance for Distributed MapReduce Runtime Systems

Rice University, Houston, Texas **Aug 2009 - May 2013**
Bachelor of Science in Computer Science
Cumulative GPA: **3.99/4.33, Magna Cum Laude, Distinction in Research and Creative Work**

Research and Work Experience

Google **Aug 2020 – Present**
Senior Software Engineer, XLA TPU Compiler Team **Manager: Arpith Jacob**

- Worked on the XLA Compiler for next-gen TPUs with a focus on large-scale distributed training of embedding-based workloads. Designed and implemented the distributed execution infrastructure, optimized scalable collective algorithms, and scheduling of HLOs to overlap of communication and computation. Contributed to the pipelining of inputs and fusion of HLOs to reduce HBM traffic. I worked with HLOs, MLIR, and LLVM based Intermediate Representations.
- Optimizing the performance of ML workloads on TPUs with a specialization in embedding-based workloads. I also worked on improving the performance of Neural Radiance Fields (NeRF), Graph Neural Networks (GNN), and approximate Top-K.

Massachusetts Institute of Technology Computer Science Department **June 2014 – July 2020**
Research Assistant, COMMIT group **Advisors: Prof. Saman Amarasinghe, Julian Shun**

- Created and led the design and implementation of ***GraphIt***, a domain-specific language for writing high-performance graph analytics. GraphIt achieved up to 4.8x speedup over the fastest CPU and GPU graph

processing frameworks for ordered and unordered graph algorithms. GraphIt is currently used by University of Washington, Cornell University, NVIDIA for the development of domain-specific accelerators.
- Led the development of ***priority-based extensions to GraphIt*** for supporting high-performance ordered parallelism for applications and the development of new cache optimizations.
- Worked on using GraphIt to generate high-performance GPU implementations of graph algorithms.
- Worked on high-performance Sparse Linear Algebra kernels for SpMV on multi-core CPUs.

Rice University Computer Science Department **Aug 2012 – May 2014**
Research Assistant, Habanero Multi-Core Software Group **Advisor: Prof. Vivek Sarkar**

- Designed and implemented the ***HJ-Hadoop*** MapReduce runtime, which integrates Habanero Java's shared memory model into Hadoop MapReduce's distributed memory model. HJ-Hadoop enables efficient data sharing among different tasks and improves the performance of data analytics applications by up to 3x.

IBM Research Lab, Austin **May 2013 – Aug 2013**
Research Intern, Distributed High performance Key-Value Store **Mentor: Dr. Juan Rubio**

- Designed and implemented a query API for a distributed key-value store and integrated it into the application.
- Worked on integrating consistent hashing into the distributed key-value store.

Microsoft, Redmond **May 2012 – Aug 2012**
Software Developer Engineering Intern, Azure Data Market Team **Manager: David Shiflet**

- Improved search functionalities on the website with NLP libraries to better match user interest with data or application offered by Azure Data Market.

Publications

Taming the Zoo: The Unified GraphIt Compiler Framework for Novel Architectures
Ajay Brahmakshatriya, Emily Furst, Victor A. Ying, Claire Hsu, Changwan Hong, Max Ruttenberg, **Yunming Zhang,** Tommy Jung**,** Dustin Richmond, Michael Taylor, Julian Shun, Mark Oskin, Daniel Sanchez, and Saman Amarasinghe
- International Symposium on Computer Architecture (***ISCA***) 2021

Efficient Stepping Algorithms and Implementations for Parallel Shortest Paths
Xiaojun Dong, Yan Gu, Yihan Sun, **Yunming Zhang**
- ACM Symposium on Parallelism in Algorithms and Architectures (***SPAA***) 2021

Compiling Graph Algorithms for GPUs with GraphIt
Ajay Brahmakshatriya, **Yunming Zhang**, Changwan Hong, Shoaib Kamil, Julian Shun, and Saman Amarasinghe
- International Symposium on Code Generation and Optimization (***CGO***) 2021 ***Best Pape*****r**

Evaluation of Graph Analytics Frameworks Using the GAP Benchmark Suite
A. Azad, M. M.Aznaveh, S. Beamer, M. Blanco, J. Chen, L. D'Alessandro, R. Dathathri, T. Davis, K. Deweese, J. Firoz, H. A. Gabb, G. Gill, B. Hegyi, S. Kolodziej, T. M. Low, A. Lumsdaine, T. Manlaibaatar, T. G. Mattson, S. McMillan, R. Peri, K. Pingali, U. Sridhar, G. Szarnyas, **Yunming Zhang**, and, Y. Zhang **(ordered alphabetically)**
- IEEE International Symposium on Workload Characterization (***IISWC***) 2020

Optimizing Ordered Graph Algorithms with GraphIt
Yunming Zhang, Ajay Brahmakshatriya, Xinyi Chen, Laxman Dhulipala, Shoaib Kamil, Saman Amarasinghe, Julian Shun
- International Symposium on Code Generation and Optimization (***CGO***) 2020

Tiramisu: A Polyhedral Compiler for Expressing Fast and Portable Code
Riyadh Baghdadi, Jessica Ray, Malek Ben Romdhane, Emanuele Del Sozzo, Abdurrahman Akkas, **Yunming Zhang**, Patricia Suriana, Shoaib Kamil, Saman Amarasinghe

- International Symposium on Code Generation and Optimization (***CGO***) 2019

GraphIt – A High-Performance DSL for Graph Analytics
Yunming Zhang, Mengjiao Yang, Riyadh Baghdadi, Shoaib Kamil, Julian Shun, Saman Amarasinghe

- ACM SIGPLAN Conference on Object-oriented Programming, Systems, Languages, and Applications (***OOPSLA***) 2018
- **Project Page:** https://graphit-lang.org/, **Github:** https://github.com/GraphIt-DSL/graphit

Making Caches Work for Graph Analytics
Yunming Zhang, Vladimir Kiriansky, Charith Mendis, Matei Zaharia, Saman Amarasinghe

- IEEE International Conference on Big Data (***BigData***) 2017 ***Best Student Paper***

Optimizing Indirect Memory References with Milk
Vladimir Kiriansky, **Yunming Zhang**, Saman Amarasinghe

- International Conference on Parallel Architectures and Compilation Techniques (***PACT***) 2016

HJ-Hadoop: An Optimized MapReduce Runtime for Multi-core Systems.
Yunming Zhang, Alan Cox, Vivek Sarkar.

- 5th USENIX Workshop on Hot Topics in Parallelism (***HotPar*** '13). June 2013. [poster with paper]

Awards and Honors

- Best Paper, CGO 2021
- Best Student Paper, BigData 2017
- Third place, Undergraduate, ACM Student Research Competition at SPLASH 13 (2013)
- Research Fellowship for Master of Science in Computer Science at Rice University (2013)

Talks

- "Optimizing Ordered Graph Algorithms with GraphIt", CGO 2020
- "Making Graph Computations Fast, Simple, and Portable with GraphIt", Nvidia Research, 2020
- "Making Graph Computations Fast, Simple, and Portable with GraphIt", Microsoft Research, 2020
- "Writing High-Performance Graph Applications with GraphIt", Facebook Boston, 2019
- "Writing High-Performance Graph Applications with GraphIt", Google NY, 2019
- "GraphIt: A Domain-Specific Language for Writing High-Performance Graph Applications", MIT Fast Code Seminar, MIT Graphics Seminar 2019, MIT *Algorithm Engineering (6.886) 2019, 2020*
- "Compiling Sparse Graphs and Tensors", University of Texas at Austin ICES Seminar 2018
- "GraphIt: A DSL for Writing High-Performance Graph Applications", SRC TECHCON 2018
- "Optimizing Cache Performance for High-Performance Graph Analytics", MIT *Graph Analytics (6.886)* 2018
- "GraphIt, a High-Performance Graph DSL". *OOPSLA 2018*
- "Making Caches Work for Graph Analytics". *BigData 2017*

Teaching and Mentorship Experience

Teaching Assistants at MIT and Rice

- MIT: TA for *Performance Engineering of Software Systems (6.172)* in Fall 2016
- Rice: TA for *Fundamentals of Parallel Computing* (COMP 322) for 2 semesters. *Advanced Object-Oriented Computing* (COMP 310), *Computational Thinking* (COMP 140). (From 2010 to 2013)

Mentoring Master and Undergraduate Students at MIT

- Mengjiao Yang, Master of Engineering, (coauthor of GraphIt paper at OOPSLA 2018)
- Xinyi Chen, Undergraduate Researcher, (coauthor of GraphIt extensions paper at CGO 2020, SuperUROP award)
- Tugsbayasgalan Manlaibaatar, Master, Undergraduate Researcher (High-Performance Graph Algorithms)

Service

- Journal of Supercomputing (JOSC) 2021 Reviewer
- ACM Transactions on Computer Systems (**TOCS**) 2020 Reviewer
- Journal of Computer Science and Technology (**JCST**) 2020 Reviewer
- ACM Transactions on Parallel Computing (**TOPC**) 2020 Reviewer
- International Conference on Very Large Data Bases **(VLDB)** 2020 External Reviewer
- IEEE International Parallel & Distributed Processing Symposium (**IPDPS**) 2019 Reviewer
- Transaction on Parallel and Distributed Systems **(TPDS)** 2019 Reviewer
- Symposium on Parallelism in Algorithms and Architectures (**SPAA**) 2018 Reviewer
- ACM Computing Surveys 2017 Reviewer
- International Symposium on Code Generation and Optimization (***CGO***) 2016 Reviewer