

TONG ZHOU

(865) 347-8154 \diamond tz@gatech.edu

SUMMARY

My work has focused on building programming tools that improve the code performance (make them run faster and/or with less energy consumption) and the productivity of the programmer, especially for domains such as scientific computing, data science and machine learning. In the past few years I mostly worked on compiling Python programs from scientific and machine learning domains for efficient CPU and GPU execution. On the CPU front, I have led the development of Intrepydd, a Python to C++ compiler which compiles a broad class of Python language constructs and NumPy array operators to sequential and parallel C++ code. On the GPU front, I have been developing a GPU programming model and compiler, APPy (Annotated Parallelism for Python), which enables users to annotate loops and tensor operators in Python with compiler directives akin to OpenACC. APPy automatically compiles the annotated code to Triton GPU kernels. Additionally, for programs consisting of sparse tensor operators, I have introduced techniques to fuse sparse/dense operators together that achieve greater redundancy elimination and higher performance than the state of the art.

EDUCATION

- | | |
|--|--------------------|
| Georgia Institute of Technology
Ph.D., Computer Science. Advisor: Vivek Sarkar | <i>2018 - 2024</i> |
| University of Tennessee, Knoxville
M.S., Computer Science. Advisor: Michael R. Jantz | <i>2015 - 2018</i> |
| Beijing University of Posts and Telecommunications (China)
B.S., Electronic Information Engineering | <i>2011 - 2015</i> |

PROFESSIONAL EXPERIENCE

- | | |
|---|------------------------------|
| Software Engineer Intern, Meta, PyTorch Compiler Team
Designed and prototyped approaches in TorchDynamo compiler towards automatic GPU kernel generation using Triton for sparse tensor computations in transformers. Supervisor: Animesh Jain. | <i>May 2022 - Aug. 2022</i> |
| PhD Intern, Pacific Northwest National Laboratory, HPC Team
Designed and prototyped efficient code generation approaches for sparse tensor algebra in COMET compiler framework. Supervisors: Roberto Gioiosa and Gokcen Kestor. | <i>Jun. 2021 - Dec. 2021</i> |

PUBLICATIONS

- **APPy: Annotated Parallelism for Python on GPUs.**
Tong Zhou, Jun Shirako, Vivek Sarkar.
33rd ACM SIGPLAN International Conference on Compiler Construction (CC '24), March 2024.
- **ReACT: Redundancy-Aware Code Generation for Tensor Expressions.**
Tong Zhou, Ruiqin Tian, Rizwan Ashraf, Gokcen Kestor, Roberto Gioiosa, Vivek Sarkar.
The 31st International Conference on Parallel Architectures and Compilation Techniques (PACT), Oct. 2022, Chicago.
- **Intrepydd: Performance, Productivity, and Portability for Data Science Application Kernels.**

Tong Zhou, Jun Shirako, Anirudh Jain, Srieshan Srikanth, Thomas Conte, Richard Vuduc, Vivek Sarkar.

19th ACM SIGPLAN Onward!, co-located with SPLASH OOPSLA, November 2020.

- **Valence: Variable Length Calling Context Encoding.**

Tong Zhou, Michael R. Jantz, Prasad A. Kulkarni, Kshitij A. Doshi, Vivek Sarkar.

28th International Conference on Compiler Construction (CC '19), February 2019.

- **MemBrain: Automated Application Guidance for Hybrid Memory Systems.**

M. Ben Olson, Tong Zhou, Michael R. Jantz, Kshitij A. Doshi, M. Graham Lopez, and Oscar Hernandez.

*13th IEEE International Conference on Networking, Architecture, and Storage (NAS '18). **Best Paper Award**, October 2018.*

- **On Automated Feedback-Driven Data Placement in Hybrid Memories.**

Chad Effler, Adam P. Howard, Tong Zhou, Michael R. Jantz, Kshitij A. Doshi, and Prasad A. Kulkarni.

In the International Conference on Architecture of Computing Systems (ARCS '18), ser. Lecture Notes in Computer Science, April 2018.

POSTERS

- **Efficient Block-Sparse GPU Kernel Generation.**

Tong Zhou, Animesh Jain, Vivek Sarkar.

Poster at the 31st International Conference on Parallel Architectures and Compilation Techniques (PACT), Oct. 2022, Chicago.

- **Redundancy-Avoiding Fusion for Tensor Algebra.**

Tong Zhou, Ruiqin Tian, Gokcen Kestor, Roberto Gioiosa, Vivek Sarkar.

Poster at Georgia Tech's Advanced Research Computing (ARC) Virtual Symposium, held jointly with SC21 hybrid conference, Nov. 2021.

TEACHING ACTIVITIES

- Teaching Assistant, CS4240: Compilers and Interpreters (Spring 2022 at GT)
- Teaching Assistant, COSC580: Algorithms (Spring 2018 at UTK)
- Teaching Assistant, COSC340: Software Engineering (Spring 2016 at UTK)

PROGRAMMING EXPERIENCE

- Proficient in Python, C++, and Java.
- Experience with the following open source projects: Python AST, Triton Language, LLVM, and COMET (built upon MLIR),