Guray Ozen

URL: Home Page | Google Scholar Page   EMAIL: guray.ozen@gmail.com

Affiliation

Google Research, Switzerland — *Compiler Researcher*
August 2022 - PRESENT

MLIR and OpenXLA's IREE GPU Compiler Team , Supervisor: Nicolas Vasilache
- Working on NVIDIA Hopper Tensor Core support in NVGPU and NVVM dialects in MLIR [slides] [video]
- Combined the IREE compiler's tiling and fusion with CUTLASS's optimized GEMM hit peak A100 performance.

Research Interest

Current research centers on optimizing compilers and programming languages for GPU utilization in ML and HPC. Key contributions to production compilers including clang, flang, MLIR, IREE and NVIDIA HPC (formerly PGI).

Previously involved in designing parallel programming languages such as OpenMP and OpenACC. Served as a voting member in the OpenMP Language Committee for NVIDIA and contributed actively to OpenACC language.

Education

Ph.D. — Universitat Politècnica de Catalunya (UPC), 2014 - 2018, Spain
Title: "Compiler and Runtime Based Parallelization and Optimization", Advisors: Prof. Jesús Labarta and Prof. Eduard Ayguadé
Degree: Excellent Cum Laude

MS.c. — Universitat Politècnica de Catalunya (UPC), 2012 - 2014, Spain

B.S. — Dokuz Eylul University, 2006 - 2010, Turkey

Past Working Experience

NVIDIA, Germany — *Senior Compiler Engineer*
March 2018 - August 2022

NVIDIA HPC GPU Compiler Team (formerly PGI compilers), Mentor: Michael Wolfe
- Worked on C, C++, and Fortran to PTX code generation compiler that accelerated programs on top500 supercomputers
- Lead engineer of OpenMP GPU Offload Project. Fastest compiler for SPEC ACCEL® benchmarks
- Implemented Standard Parallelism in Fortran for GPUs

NVIDIA, USA— *Research Intern*
March 2017 - October 2017

NVIDIA HPC GPU Compiler Team (formerly PGI compilers), Mentor: Michael Wolfe
- Developed a novel software-based dynamic loop scheduling for Thread Blocks

IBM T.J. Watson Research Center, USA — *Research Intern*
March 2016 - October 2016

Advanced Compiler Technologies Team, Supervisor: Kevin O'Brien
- Contributed to CORAL project with the Clang OpenMP compiler used on Summit (2nd fastest) and Sierra (3rd fastest) supercomputer

Barcelona Supercomputing Center, Spain— *Research Fellow*
March 2013 - October 2018

Programming Models Team, Advisors: Prof. Jesús Labarta and Prof. Eduard Ayguadé
- Made OmpSs compiler to generate CUDA and OpenCL from OpenMP in C, C++, and Fortran
- Developed GPU compiler optimization techniques for SpMV, outperformed cuSPARSE

Veripark/Akbank, Turkey— *Web Developer*
July 2010 - June 2012
- Developed distributed system for one-time password message sending over million clients

Selected Publications

1. **G. Ozen**, M. Wolfe "Performant Portable OpenMP", *ACM SIGPLAN 2022 International Conference on Compiler Construction (CC),* 2022.

2. C. Daley, A. Southwell, R. Gayatri, S. Biersdorff, C. Toepfer, **G. Ozen**, N. Wright "Non-Recurring Engineering (NRE) Best Practices: A Case Study with the NERSC/NVIDIA OpenMP Contract", *Supercomputing (SC),* 2021.

3. **G. Ozen**, S. Atzeni, M. Wolfe, A. Southwell, G. Klimowicz. "OpenMP GPU Offload in Flang and LLVM" *LLVM-HPC, Supercomputing (SC),* 2018.

4. **G. Ozen**, E. Ayguade, J. Labarta. "Collective Dynamic Parallelism for Directive Based GPU Programming Languages and Compilers", short paper, *Parallel Architectures and Compilation (PACT),* 2016.

5. S. Antao, A. Bataev, A. Jacob, A. Eichenberger, G. Rokos, M. Martineau, T. Jin, **G. Ozen**, Z. Sura, T. Chen, H. Sung, C. Bertolli, and K. O'Brien. "Offloading Support for OpenMP in Clang and LLVM", *LLVM-HPC, Supercomputing (SC),* 2016.

Patent

1. **G.Ozen**, M. Wolfe, Accurate alias analysis in logarithmic time, filed by NVIDIA 2022

Talks and Other Presentations

1. "Targeting H100 with NVGPU and NVVM Dialects", MLIR Open Design Meeting, 2023 [slides] [video]

2. "Thinking OpenMP with NVIDIA HPC Compilers", NASA AMS Seminars, 2021 [video]

3. "Accelerating Applications for the NERSC Perlmutter Supercomputer Using OpenMP", GPU Technology Conference (GTC) 2021 [video]

4. NVIDIA representative in OpenMP Birds of a Feather (BoF) session in *Supercomputing Conference (SC)*, 2019.

5. "Compiler and Runtime Based Parallelization and Optimization", Doctoral Showcase, Ph.D. Forum, *Supercomputing Conference (SC)*, 2018.

Research Activities

1. PC at Accelerators and Hybrid Emerging Systems workshop (AsHES) at IPDPS, 2019, 2020, 2021, 2022, 2023

2. PC at International Workshop on Extreme Heterogeneity Solutions (ExHET) at PPoPP 2022, 2023, 2024

3. PC at ISC High Performance (ISC) at 2019, 2024

4. PC at Workshop on Accelerator Programming Using Directives (WACCPD) at SC19

5. Peer reviewer for Parallel Computing Journal (Parco), IEEE Access Journal