Mike Wilkins

HPC/AI Researcher

wilkins@anl.gov
linkedin.com/in/mjwilkins418
github.com/mjwilkins18

EDUCATION

Ph.D. Computer Engineering
Northwestern University 2023

M.S. Computer Engineering
Northwestern University 2021

B.S. Computer Engineering
Rose-Hulman Institute of Technology 2019

SKILLS

- **Software/Scripting Languages:** C, C++, Python, Standard/Parallel ML, C#, LabVIEW, Java, SQL, Bash
- **Parallel Programming/Communication:** MPI, Libfabric, NCCL, CUDA, PyTorch, Parallel ML
- **Simulators/Tools:** Sniper, gem5, ZSim, Xilinx Vivado, Xilinx ISE, Quartus II
- **Hardware Description Languages:** Chisel, VHDL, Verilog, SPICE

Experiences

Maria Goeppert Mayer Fellow Oct 2024 - Present
Argonne National Laboratory
- Directed a $1M+ lab-funded research project, "Holistic Machine Learning Autotuning for Massive-Scale Artificial Intelligence for Science," supervising a team of 3 students
- Developed and deployed ACCLAiM, an autotuner for collective communication, on production supercomputers including Aurora, delivering widespread speedups of up to 35x
- Contributed major enhancements to MPICH, the leading open-source MPI implementation, with a focus on optimizing collective communication for high-performance computing environments.

Software Engineer Jan-Sep 2024
Cornelis Networks
- Optimized the OPX libfabric provider, achieved a 5x bandwidth improvement for GPU communication among other advancements
- Led the development of the reference libfabric provider for the Ultra Ethernet Consortium
- Created developer productivity tooling, including an OPX performance profiler and a runtime parameter autotuner

AI Research Intern Summer 2023
Meta
- Designed and implemented an application-aware communication (NCCL) autotuner for large-scale AI workloads
- Developed an AI application emulation tool that mimics production models by overlapping communication and genericized compute kernels

Research Aide/Visiting Student 2020 - 2023
Argonne National Laboratory
- Founded the MPI collective algorithm/machine learning project, initially under the supervision of Dr. Min Si and Dr. Pavan Balaji, later Dr. Yanfei Guo and Dr. Rajeev Thakur
- Earned perpetual external funding from ANL for the remainder of my Ph.D

Engineering Leadership Program Intern Summer 2018
National Instruments
- Engaged with technical leaders through field presentations to multiple companies in the Seattle area
- Assisted customers to design and troubleshoot data-acquisition applications using NI platforms

Trailblazer Intern Summer 2017
Flexware Innovation
- Designed an innovative RFID tracking solution to repair a malfunctioning inventory locating system
- Produced a full-stack BI database solution analyzing internal employee and revenue data

Director of Tool Services Summer 2016
Power Solutions International
- Organized and managed the company's inventory of CNC machining tools, valued at more than $500,000
- Trained company technicians on new processes and managed tool services employees

Sample Research Projects

Here is a high-level description of some of my active and former research projects.

- **Holistic Online Autotuning for Large-Scale Artificial Intelligence** (Ongoing)
  - Creating a new autotuner for long-running distributed AI training workloads that improves model quality and workload efficiency during execution
- **ML Autotuning for Generalized MPI Collective Algorithms** (2021-2024)
  - Created new MPI collective algorithms and a machine-learning autotuner (ACCLAiM) that automatically selects and optimizes the best algorithm
  - Invented multiple optimizations to make ML-based MPI autotuning feasible on large-scale systems
- **High-Level Parallel Languages for HPC** (2019-2023)
  - Developing a new hardware/software co-design for the Standard ML language targeted at HPC systems and applications, including AI
  - Created a new version of the NAS benchmark suite using MPL (a parallel compiler for Standard ML) to enable direct comparison between HLPLs and lower-level languages for HPC
- **Cache Coherence for High-Level Parallel Languages** (2019-2022)
  - Identified a low-level memory property called WARD that can be introduced by construction in high-level parallel programs
  - Implemented a custom cache coherence protocol in the Sniper architectural simulator and found an average speedup of 1.46x across the PBBS benchmark suite.

Publications

- **On Transparent Optimizations for Communication in Highly Parallel Systems**
  Michael Wilkins • Ph.D. Thesis
- **Generalized Collective Algorithms for the Exascale Era**
  Michael Wilkins, Hanming Wang, Peizhi Liu, Bangyen Pham, Yanfei Guo, Rajeev Thakur, Nikos Hardavellas, and Peter Dinda • CLUSTER'23
- **Evaluating Functional Memory-Managed Parallel Languages for HPC using the NAS Parallel Benchmarks**
  Michael Wilkins, Garrett Weil, Luke Arnold, Nikos Hardavellas, Peter Dinda • HIPS'23 Workshop
- **WARDen: Specializing Cache Coherence for High-Level Parallel Languages**
  Michael Wilkins, Sam Westrick, Vijay Kandiah, Alex Bernat, Brian Suchy, Enrico Armenio Deiana, Simone Campanoni, Umut Acar, Peter Dinda, Nikos Hardavellas • CGO'23
- **Program State Element Characterization**
  Enrico Deiana, Brian Suchy, Michael Wilkins, Brian Homerding, Tommy McMichen, Katarzyna Dunajewski, Nikos Hardavellas, Peter Dinda, Simone Campanoni • CGO'23
- **ACCLAiM: Advancing the Practicality of MPI Collective Communication Autotuning Using Machine Learning**
  Michael Wilkins, Yanfei Guo, Rajeev Thakur, Peter Dinda, Nikos Hardavellas • CLUSTER'22
- **A FACT-Based Approach: Making Machine Learning Collective Autotuning Feasible on Exascale Systems**
  Michael Wilkins, Yanfei Guo, Rajeev Thakur, Nikos Hardavellas, Peter Dinda, Min Si • ExaMPI'21 Workshop