Younghyun Cho

Assistant Professor of Computer Science and Engineering
Curriculum Vitae

Santa Clara University
Santa Clara, CA, USA
✉ *younghyun.cho@scu.edu*
⌂ *https://younghyunc.github.io*

Research Overview

My research focuses on system software and performance tools for parallel and high-performance computing (HPC). I am interested in improving performance of applications, ranging from traditional linear algebra operations (e.g. BLAS/LAPACK routines) to emerging machine learning workloads (e.g. randomized algorithms), running on various platforms, ranging from conventional CPUs/GPUs to domain-specific accelerators to DOE's leadership supercomputers, based on compiler/runtime system, resource management, performance profiling, and automatic performance tuning. I am one of the core developers of GPTune which is a machine learning-based, open-source performance autotuner for HPC applications and has been used for many applications in ECP (Exascale Computing Project) and other projects, across various institutions in the U.S. and abroad.

Experience

Assistant Professor, Santa Clara University, Santa Clara, CA, USA, **Sept. 2023 – present**
Department of Computer Science and Engineering

Postdoctoral Scholar, University of California, Berkeley, Berkeley, CA, USA, **Aug. 2020 – Aug. 2023**
Department of Electrical Engineering and Computer Sciences
Supervisor: Professor James Demmel

Visiting Researcher, ETH Zürich, Zürich, Switzerland, **March – May 2018 & June – Aug. 2016**
Host: Professor Thomas R. Gross

Education

Doctor of Philosophy, Seoul National University, Seoul, South Korea, **Aug. 2020**
Department of Electrical Engineering and Computer Science
Advisor: Professor Bernhard Egger

Bachelor of Science in Engineering, University of Seoul, Seoul, South Korea, **Feb. 2013**
School of Computer Science

Activities

Organizing committee
- Int'l Conference on Languages, Compilers, Tools and Theory of Embedded Systems (LCTES), 2023 (publicity chair and artifact evaluation co-chair)

Program committee
- Int'l Conference on Languages, Compilers, Tools and Theory of Embedded Systems (LCTES), 2023
- International Parallel & Distributed Processing Symposium (IPDPS) 2022

Artifact evaluation committee
- Principles and Practice of Parallel Programming (PPoPP), 2018, 2019
- Int'l Conference on Languages, Compilers, Tools and Theory of Embedded Systems (LCTES), 2019

Guest talk
- UC Santa Cruz Hardware Group Seminar (VLSI/CAD Seminar), "Autotuning HPC Applications with Gaussian Process Regression and Crowd-based Transfer Learning", Feb. 2023.

Tutorial

- ECP Annual Meeting 2023, "Autotuning ECP codes using the GPTune package"
- ECP Annual Meeting 2022, "Performance Autotuning of ECP Applications with Gaussian Process-Based and Cloud Database-Enhanced GPTune Package"
- ECP Annual Meeting 2021, "GPTune: Performance Autotuner for ECP Applications"

Award

- "SnuMAP: SNU Many-core Profiler for Big-data", 2nd Prize in Open Source Software World Challenge, Seoul, South Korea, 2016.

▬ Teaching

At Santa Clara University

- COEN20, Section 1, Introduction to Embedded Systems, Fall 2023

Teaching assistant roles at Seoul National University

- 4190.570, Advanced Compiler Construction, Fall 2018
- 4190.409, Compilers, Fall 2017, Fall 2014
- 4190.203, System Programming, Fall 2013, Spring 2013

▬ Publications

Dissertation

Ph.D. Dissertation	Younghyun Cho. Parallelism Management for Co-Located Parallel Applications. Seoul National University, August 2020.

Full Papers in Conferences and Journals

IPDPS 2023	Younghyun Cho, James W. Demmel, Jacob King, Xiaoye S. Li, Yang Liu, Hengrui Luo. Harnessing the Crowd for Autotuning High-Performance Computing Applications. In *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, May 2023.
PPoPP 2022	Younghyun Cho, Jiyeon Park, Florian Negele, Changyeon Jo, Thomas R. Gross, and Bernhard Egger. Dopia: Online Parallelism Management for Integrated CPU/GPU Architectures. In *Principles and Practice of Parallel Programming (PPoPP)*, April 2022.
MCSoC 2021	Younghyun Cho, James W. Demmel, Xiaoye S. Li, Yang Liu, Hengrui Luo. Enhancing Autotuning Capability with a History Database. In *International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, December 2021.
JPDC 2020	Reza Entezari-Maleki, Younghyun Cho, and Bernhard Egger. Evaluation of memory performance in NUMA architectures using Stochastic Reward Nets. In *Journal of Parallel and Distributed Computing (JPDC)*, October 2020.
TPDS 2020	Younghyun Cho, Surim Oh, and Bernhard Egger. Performance Modeling of Parallel Loops on Multi-Socket Platforms using Queueing Systems. In *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, February 2020.
PACT 2018A	Younghyun Cho, Camilo A. Celis Guzman, and Bernhard Egger. Maximizing System Utilization via Parallelism Management for Co-Located Parallel Applications. In *Proceedings of the 27th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, November 2018.
PACT 2018B	Younghyun Cho, Florian Negele, Seohong Park, Bernhard Egger, and Thomas R. Gross. On-The-Fly Workload Partitioning for Integrated CPU/GPU Architectures. In *Proceedings of the 27th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, November 2018.
PACT 2016	Younghyun Cho, Surim Oh, and Bernhard Egger. Online Scalability Characterization of Data-Parallel Programs on Many Cores. In *Proceedings of the 25th International Conference on Parallel Architecture and Compilation Techniques (PACT)*, September 2016.

TC 2016	Bernhard Egger, <u>Younghyun Cho</u>, Changyeon Jo, Eunbyung Park, and Jaejin Lee. Efficient Checkpointing of Live Virtual Machines. In *IEEE Transactions on Computers (TC)*, January 2016.

Non peer-reviewed pre-prints and miscellaneous documents

Arxiv 2023	<u>Younghyun Cho</u>, James W. Demmel, Micha Dereziski, Haoyun Li, Hengrui Luo, Michael W. Mahoney, Riley J. Murray. Surrogate-based Autotuning for Randomized Sketching Algorithms in Regression Problems, arXiv:2308.15720, August 2023.
UG 2022	<u>Younghyun Cho</u>, James W. Demmel, Grace Dinh, Xiaoye S. Li, Yang Liu, Hengrui Luo, Osni Marques, Wissam M. Sid-Lakhdar. GPTune User Guide. In `https://github.com/gptune/GPTune/blob/master/Doc/GPTune_UsersGuide.pdf`, October 2022.
Arxiv 2022	Hengrui Luo, <u>Younghyun Cho</u>, James W. Demmel, Xiaoye S. Li, Yang Liu. Hybrid Models for Mixed Variables in Bayesian Optimization. arXiv:2206.01409, June 2022.
Arxiv 2021	Hengrui Luo, James W. Demmel, <u>Younghyun Cho</u>, Xiaoye S. Li, Yang Liu. Non-smooth Bayesian Optimization in Tuning Problems. arXiv:2109.07563, September 2021.

Workshop/Poster/Other Presentations

MLArchSys 2023	Grace Dinh, Iniyaal Kannan Jegadesan Valsala, Hengrui Luo, Charles Hong, <u>Younghyun Cho</u>, James Demmel, Sherry Li, Yang Liu, Sample-Efficient Mapspace Optimization for DNN Accelerators with Bayesian Learning, In Workshop on ML for Computer Architecture and Systems (MLArchSys @ISCA 2023)
SC 2022 (Poster)	Mohammad Zaeed, Tanzima Islam, <u>Younghyun Cho</u>, Xiaoye S. Li, Hengrui Luo, Yang Liu. Analysis and Visualization of Important Performance Counters To Enhance Interpretability of Autotuner Output. In *The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, November 2022.
PACT 2017 (Poster)	<u>Younghyun Cho</u>, Camilo A. Celis Guzman, and Bernhard Egger. POSTER: Improving NUMA System Efficiency with a Utilization-based Co-scheduling. In *Proceedings of the 26th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, September 2017.
Multiprog 2017A	<u>Younghyun Cho</u>, Surim Oh, and Bernhard Egger. Cooperative Parallel Runtimes for Multicores. *Presented at the 10th International Workshop on Programmability and Architectures for Heterogeneous Multicores*, January 2017.
Multiprog 2017B	Camilo A. Celis Guzman, <u>Younghyun Cho</u>, and Bernhard Egger. SnuMAP: an Open-source Trace Profiler for Manycore Systems. *Presented at the 10th International Workshop on Programmability and Architectures for Heterogeneous Multicores*, January 2017.
CATC 2016	Surim Oh, <u>Younghyun Cho</u>, and Bernhard Egger. Efficient Resource Management for Manycores with Centralized L2 Caches using Distributed Control Processors. *Presented at the 7th Compiler, Architectures and Tools Conference*, September 2016 (CATC 2016 was held as a workshop co-located with PACT 2016).
JSSPP 2016	<u>Younghyun Cho</u>, Surim Oh, and Bernhard Egger. Adaptive Space-shared Scheduling for Shared-memory Parallel Programs. *Presented at the 20th Workshop on Job Scheduling Strategies for Parallel Processing*, May 2016. *Also published in Lecture Notes in Computer Science (LNCS), Volume 10353*, July 2016.

Patents

US 11,301,016	Bernhard Egger, <u>Younghyun Cho</u>, Surim Oh, and Donghoon Yoo. Computing devices and methods of allocating power to plurality of cores in each computing device, 2022.
US 10,503,557	Bernhard Egger, Surim Oh, <u>Younghyun Cho</u>, and Donghoon Yoo. Method of processing OpenCL Kernel and Computing Device Therefor. 2019.
US 10,409,351	Bernhard Egger, <u>Younghyun Cho</u>, Surim Oh, and Donghoon Yoo. Computing devices and methods of allocating power to plurality of cores in each computing device, 2019.

Last update: Sept. 25, 2023.