Hasan Nazim Genc

Website: hngenc.github.io • *Email:* hngenc@berkeley.edu

Education

Ph.D., Computer Science, *August 2018 – Present*
University of California, Berkeley
GPA: 3.97/4.00

B.S., Electrical and Computer Engineering, *May 2018*
University of Texas at Austin
GPA: 3.86/4.00

Experience

Research Intern, *NVIDIA,* 06/2022 – 09/2022
- Investigated sparsity in transformer attention layers

Software Engineering Intern, *Anyscale*, 12/2020 – 02/2021, 08/2020 – 11/2020
- Added new features to Anyscale product to help customers with cost-management

Graduate Research Assistant, *University of California*, 09/2018 – Present
- Research ML inference and training accelerators for the edge
- Develop Gemmini, an open-source, multi-dataflow, ML hardware accelerator (https://github.com/ucb-bar/gemmini)
- Tape-out chips, performing place-and-route, power and rail analysis, LVS, and DRC
- Investigate how Halide programs can be tuned with algorithms such as MCTS
- Add features, such as first-class enums, to the Chisel hardware description language

Software Development Engineer (SDE) Intern, *Amazon*, 06/2020 – 09/2020
- Investigated low-bitwidth floating point DNN training (down to 8-bits)

Undergraduate Research Assistant, *University of Texas*, 09/2016 – 05/2018
- Investigated optimizations for drone processors and runtime systems
- Developed benchmark suite for autonomous drone applications
- Researched power and performance characterizations of autonomous drones

Software Engineering Intern, *Intel*, 01/2017 – 05/2017
- Researched performance of computing systems for autonomous drones

Software Developer Intern, *CAPSHER Technology*, 06/2016 – 08/2016
- Designed software architecture for a web app that allows users to create slideshows

Publications

Steve Dai, **Hasan Genc**, Rangharajan Venkatesan, Brucek Khailany, "Efficient Transformer Inference with Statically Structured Sparse Attention." *Design Automation Conference (DAC), 2023*

Sehoon Kim, Coleman Hooper, Thanakul Wattanawong, Minwoo Kang, Ruohan Yan, **Hasan Genc**, Grace Dinh, Qijing Huang, Kurt Keutzer, Michael W. Mahoney, Yakun Sophia Shao, Amir Gholami, "Full Stack Optimization of Transformer Inference: a Survey." *arXiv:2302.14017, 2023*

Seah Kim, **Hasan Genc**, Vadim Vadimovich Nikiforov, Krste Asanovic, Borivoje Nikolic, Yakun Sophia Shao, "MoCA: Memory-Centric, Adaptive Execution for Multi-Tenant Deep Neural Networks." *International Symposium on High-Performance Computer Architecture (HPCA), 2023*

Alon Amid, **Hasan Genc**, Jerry Zhao, Krste Asanovic, Borivoje Nikolic, Yakun Sophia Shao, "Accelerating General-Purpose Linear Algebra on DNN Accelerators." *Workshop on Democratizing Domain-Specific Accelerators (WDDSA), 2022*

Yuka Ikarashi, Gilbert Louis Bernstein, Alex Reinking, **Hasan Genc**, Jonathan Ragan-Kelley, "Exocompilation for Productive Programming of Hardware Accelerators." *Programming Language Design and Implementation (PLDI), 2022*

Hasan Genc, Seah Kim, Vadim Vadimovich Nikiforov, Simon Zirui Guo, Borivoje Nikolic, Krste Asanovic, Yakun Sophia Shao, "Gemmini: An Open-Source, Full-System DNN Accelerator Design and Evaluation Platform." *Open-Source Computer Architecture Research Workshop (OSCAR), 2022*

Seah Kim, **Hasan Genc**, Vadim Vadimovich Nikiforov, Krste Asanovic, Borivoje Nikolic, Yakun Sophia Shao, "Memory-centric, Adaptive Execution for Multi-Tenant DNNs." *Architecture, Compiler, and System Support for Multi-model DNN Workloads Workshop (ACSMD), 2022*

Behzad Boroujerdian, **Hasan Genc**, Srivatsan Krishnan, Bardienus Pieter Duisterhof, Brian Plancher, Kayvan Mansoorshahi, Marcelino Almeida, Wenzhi Cui, Aleksandra Faust, Vijay Janapa Reddi, "The Role of Compute in Autonomous Micro Aerial Vehicles: Optimizing for Mission Time and Energy Efficiency." *ACM Transactions on Computer Systems (TOCS), 2022*

Abraham Gonzalez, Jerry Zhao, Ben Korpan, **Hasan Genc**, Colin Schmidt, John Wright, Ayan Biswas, Alon Amid, Farhana Sheikh, Anton Sorokin, Sirisha Kale, Mani Yalamanchi, Ramya Yarlagadda, Mark Flannigan, Larry Abramowitz, Elad Alon, Yakun Sophia Shao, Krste Asanovic, Borivoje Nikolic, "A 16mm^2 106.1 GOPS/W Heterogeneous RISC-V Multi-Core Multi-Accelerator SoC in Low-Power 22nm FinFET." *European Solid-State Circuits Conference (ESSCIRC), 2021*

Hasan Genc, Seah Kim, Alon Amid, Ameer Haj-Ali, Vighnesh Iyer, Pranav Prakash, Jerry Zhao, Daniel Grubb, Harrison Liew, Howard Mao, Albert Ou, Colin Schmidt, Samuel Steffl, John Wright, Ion Stoica, Jonathan Ragan-Kelley, Krste Asanovic, Borivoje Nikolic, Yakun Sophia Shao, "Gemmini: Enabling Systematic Deep-Learning Architecture Evaluation via Full-Stack Integration." *Design Automation Conference (DAC), 2021* **(Best Paper)**

Seah Kim, **Hasan Genc**, "Gemmini: Enabling Systematic Deep-Learning Architecture Evaluation via Full-Stack Integration." *IBM IEEE CAS/EDS AI Compute Symposium, 2020*

Ameer Haj-Ali, **Hasan Genc**, Qijing Huang, William Moses, John Wawrzynek, Krste Asanović, Ion Stoica, "ProTuner: Tuning Programs with Monte Carlo Tree Search." *arXiv:2005.13685, 2020*

Hasan Genc*, Ameer Haj-Ali*, Vighnesh Iyer*, Alon Amid*, Howard Mao, John Wright, Colin Schmidt, Jerry Zhao, Albert Ou, Max Banister, Yakun Sophia Shao,

Borivoje Nikolic, Ion Stoica, and Krste Asanovic, “Gemmini: An Agile Systolic Array Generator Enabling Systematic Evaluations of Deep-Learning Architectures.” *arXiv:1911.09925, 2019*

Hasan Genc*, Behzad Boroujerdian*, Srivatsan Krishnan, Wenzhi Cui, Aleksandra Faust, and Vijay Janapa Reddi, “MAVBench: Micro Aerial Vehicle Benchmarking.” *International Symposium on Microarchitectures (MICRO), 2018*

Ting-Wu Chin, Chia-Lin Yu, Matthew Halpern, **Hasan Genc**, Shiao-Li Tsao, and Vijay Janapa Reddi, “Domain Specific Approximation for Object Detection.” *IEEE Micro 2017, Special Issue: Approximate Computing*

Hasan Genc, Yazhou Zhu, Ting-Wu Chin, Matthew Halpern, and Vijay Janapa Reddi, “Flying IoT: Toward Low-Power Vision in the Sky.” *IEEE Micro 2017, Special Issue: Ultra-Low-Power Processors*

Hasan Genc, Ting-Wu Chin, Matthew Halpern, Vijay Janapa Reddi, “Optimizing Sensor-Cloud Architectures for Real-time Autonomous Drone Operation.” *Sensors to Cloud Architectures Workshop (SCAW), 2017*

* Authors with equal contribution

Skills

Languages (Software): C, C++, Python, Scala, Java, CUDA, HTML, JavaScript, CSS
Languages (Hardware): Chisel, Verilog
Tools: PyTorch, OpenCV, ROS