Shail Dave

sites.google.com/view/shail

RESEARCH INTERESTS

- **Design Automation:** Design space formulation and exploration; Hardware/software codesign; Sustainable and modular accelerator stack development; ML-based automation.
- **Computer Architecture:** Hardware accelerators; Execution modeling; Microarchitecture; Near-data processing.
- **Embedded Systems:** Resource-efficient execution; TinyML; Compiler, runtime for distributed, time-sensitive apps.
- **ML for Systems:** Learning the design space; Hardware/software execution modeling; Design space exploration.

Envisioned Research & Development Objective: "Enable Agile, Sustainable, and Learning-Assisted Automated Exploration of Efficient Domain-Specific Accelerator Systems through Research, Development, and Demonstration at Scale" [Brief Position Paper] [Short Talk]

EDUCATION

Ph.D. in Computer Engineering (Computer Systems) **2017–Present**
Ira A. Fulton School of Engineering, Arizona State University, Tempe, Arizona CGPA: 4.00/4.00
Dissertation: Agile and Sustainable Methodology for Designing Efficient Accelerators.
Advisor: Prof. Aviral Shrivastava, Graduate Program Chair, Computer Science and Engineering.
Committee Members: Prof. Aviral Shrivastava, Prof. Tony Nowatzki, Prof. Jae-sun Seo, Prof. Fengbo Ren, and Prof. Baoxin Li.

Master of Science, Computer Engineering (Electrical Engineering) **December 2016**
Ira A. Fulton School of Engineering, Arizona State University, Tempe, Arizona CGPA: 4.00/4.00
Thesis: Scalable Register File Architecture and Compiler for CGRA Accelerators
Committee: Prof. Aviral Shrivastava, Prof. Umit Ogras, and Prof. Fengbo Ren.

Bachelors of Engineering, Electronics and Communication Engineering **July 2014**
L. D. College of Engineering, Gujarat Technological University, Gujarat CGPA: 8.80/10.00
Final-year Project: End-to-end System Development for Precise Temperature Control of Cryogenic Cooler Systems of Next Generation Geostationary Satellites.

PROFESSIONAL EXPERIENCE

Research Associate, Compiler Microarchitecture Lab, *Tempe, AZ* **January 2016 – Present**

- Research techniques and develop tools for efficiently executing important applications such as machine learning on hardware accelerators in resource-efficient manner and compiler-aware hardware codesign.
- Developed state-of-the-art techniques and open-source tools for (1) modeling hardware accelerators and analyzing their execution costs (power, performance, area), (2) comprehensive design space formulation for finding efficient designs, and (3) quickly exploring pareto-optimal hardware/software designs from vast space (in seconds/minutes).
- Prospectus Overview Video

Research Intern, *MathWorks*, *Natick, MA* **Summer 2018**

- Project: Loop Optimizations for Target-Aware Code Generation (Group: Code Efficiency, Embedded Coder)
- Established target-aware code generation research infrastructure. Developed compiler transformations to realize performance improvements by embedding target-platform-specific features in the code generation environment, e.g., for more aggressive loop optimizations. Evaluations were performed successfully at scale for avionics and automotive apps, demonstrating up to an order-of-magnitude higher performance some testcases.

ASIC Verification Engineer (Intern), *SanDisk*, *Milpitas, CA* **Summer 2015**

- Worked with architect and technical leads of ASIC team to define and develop an automated tool that helped to make the development of Enterprise Solid State Drive Controller ASIC more efficient.
- Successfully verified some of the key modules of novel ARM-based SoC controller architecture, including developing verification plan, improving functional coverage, developing constrained random test-cases, and verifying interesting corner cases in UVM-based verification environment.

Research Intern, *Space Application Centre, Indian Space Research Organization,* *India* **Spring 2014**

- Collaborated with a team of senior research scientists and group director and achieved precise temperature control of cryogenic coolers (±0.1^{0}C at -123^{0}C) for next-generation geostationary imaging satellites.
- Developed end-to-end system for plant modeling and precise temperature control of cryogenic cooler systems.
- Implemented and prototyped digital PID controller on FPGAs for precise temperature control.
- Demonstrated the end-to-end system through two real-world system prototypes in the real-time environment under various electro-mechanical disturbances.

HONORS AND AWARDS

Teaching and Research Awards

- Silver Medal, ACM SIGBED Student Research Competition – 2022-2023
- Outstanding Research Award, Graduate and Professional Students Association, Arizona State University – 2022 (About 10-15 students awarded annually from 27,000+ graduate students and ~4,800 doctoral students.)
- Outstanding Teaching Assistant Award, School of Computing and AI (SCAI), ASU – 2018

Select Projects Funded

- Lead Research Scholar, Semiconductor Research Corporation AI Hardware Program – 2023-
  Visioned/Proposed Funded Project: Agile and Explainable AI Hardware/Software Codesign with AI.
  Preliminary Industry Liaison: Dr. Michael Kishinevsky, Intel Strategic CAD Labs.
- Student lead, ASU team, NSF/Intel joint research center for Computer Assisted Programming for Heterogeneous Architectures (CAPA) – 2018–2021 *(Our group was among total three hubs supported nation-wide. Hubs: MIT/Stanford/University of Washington; UCLA/Cornell; ASU/Penn State/Lehigh.)*
  Industrial Collaboration and Liaison: Intel Parallel Computing Lab.

Additional Honors

- Research Spotlight, IEEE Eta Kappa Nu (HKN) – 2022 (Also featured in IEEE Bridge Volume 118 and social media).
- Topical Expert on the Panel for Architectures Track, NSF Workshop on Sustainable Computing – 2022
  (Authored report on designing sustainable architectures. Report is under submission/revision.)
- Research works regularly invited and featured in several premier industrial forums.
- Invited for Inaugural Google Systems Innovation Summit – 2021
  (*Invitations limited to top global PhD students in computing systems.*)
- Certificate of Appreciation, ACM/IEEE Embedded Systems Week – 2021, 2022
  Recognized for Outstanding Service as Social Media Chair. *Impact:* Established, diversified, and continuously doubled social media presence with ~800/230/380 followers on LinkedIn/Twitter/YouTube, leading to many first-time attendees + more student/minority involvement in design automation and embedded systems communities. Note: ESWEEK encompasses three leading conferences in embedded systems – CASES, CODES+ISSS, EMSOFT; two symposia – NOCS and MEMOCODE; several tutorials and workshops; 10+ educational classes; industry paper sessions and talks; student competitions; panel; planned networking sessions; student recruitment events.
- PhD Forum Presentations at ACM/IEEE DAC and IPDPS – 2022 (~30% acceptance rate at DAC)

Fellowships and Scholarships

- Graduate College Completion Fellowship, Arizona State University – 2022-2023
  (Merit-based university-wide selection, 1-year fund for top graduating students – ~0.5% of all graduate students at ASU. Honored at ASU's Celebrating Excellence of ASU's national-award winners and ASU fellows for 2022-2023)
- Doctoral Fellowship, School of Computing and Augmented Intelligence (SCAI), ASU – 2019, 2023
- Richard Newton Young Student Fellowship, 53rd Annual Design Automation Conference (DAC) – 2016
- Engineering Grad Fellowship, Ira A. Fulton Schools of Engineering, ASU – 2018, 2019, 2020

Competitive Travel Grant Awards:

- ACM Special Interest Group on Design Automation (SIGDA) – 2018, 2022
- National Science Foundation + ACM Special Interest Group on Embedded Systems (SIGBED) – 2019
- IEEE Computer Society Technical Committee on Parallel Processing (TCPP) – 2022
- Graduate College, Arizona State University – 2019
- Graduate and Professional Student Association (GPSA), Arizona State University –2016, 2018, 2019
- School of Computing and Augmented Intelligence, Arizona State University – 2018, 2019

RESEARCH IMPACT

Citations: They are more than 250[1] with exponential increase at every year (nearly doubled) and almost all are for my first-authored works. My h-index and i-10 index are 7. My research is regularly refereed or evaluated by some of the top global experts (including from MIT, Stanford University, UC Berkeley, Carnegie Mellon University, National University of Singapore, ETH Zurich, Tsinghua University, Harvard University, University of Illinois Urbana-Champaign, Georgia Tech, Peking University, University of Toronto, Hong Kong University of Science & Technology, UCLA, USC, New York University, University of British Columbia, Seoul National University, Shanghai Jiao Tong University, Chinese Academy of Sciences, KU Leuven, Purdue University, KAIST, University of Amsterdam, University of Tokyo, National Taiwan University, Ghent University) in their papers published at topmost venues in machine learning, computer architecture, design automation, embedded systems, parallel and high-performance computing, code optimization, and circuits, including conferences like *ICLR, ICML, NeurIPS, IJCNN, ISCA, MICRO, ASPLOS, HPCA, SIGMETRICS, DAC, ICCAD, ESWEEK, DATE, CGO, PACT*, journals like *IEEE TPDS, JMLR, PMLR, ACM CSUR, Proceedings of the IEEE, Future Generation Computer Systems, IEEE TCAS-I, IEEE TCAD, IEEE Sensors*, *ACM TECS, IEEE TVLSI, IEEE D&T*, *Journal of Systems Architecture, ACM TACO*, books like *Morgan Claypool synthesis lectures,* US patents, and master's and PhD dissertations at top universities, including *National University of Singapore, Stanford, ETH Zurich, Georgia Tech, University of British Columbia, Harvard, UCLA, Purdue, ASU, and UC Santa Barbara*.

Research Usability: Citations of my papers by new works describe my techniques as state-of-the-art prior work and several following works build upon them or compare against them quantitatively. Some citations of my papers are from industry, including NVIDIA Research, Google Brain, Samsung Research, Facebook/Meta AI Research, Intel AI, IBM Research, Xilinx, AMD, Microsoft, Huawei R&D, Robert Bosch, IMEC, Alibaba, PNNL, Sandia National Labs, LBNL, RIKEN, Neural Magic, SambaNova, Graphcore, EdgeCortix, SimpleMachines, Cambricon Technologies, etc. In addition, some of my research works are studied in courses of top universities (e.g., University of Utah, UT Austin). My first-authored article at Proceedings of the IEEE (flagship IEEE publication and leading venue for more than 110 years) described most recent advances in efficient processing of recent machine/deep learning models on hardware accelerators, which has been well received and widely studied by various communities (computer hardware, computing systems, and machine learning), including senior researchers. Further, my research frameworks/tools are downloaded and used by tens of groups world-wide, including in their research development and thesis/dissertations of the students. My frameworks on GitHub receive more than 30–60 stars in a few years and several downloads every month. My research has led to development of a new topical course at ASU – CSE/CEN 598: Topics in Machine Learning Accelerator Design.

SELECT INVITED TALKS

In addition to my conference/workshop talks, my research has been presented into various talks from my collaborators at top industrial research groups and different universities. Most importantly, my research on the tools and methodology for "accelerating computing" and "agile and explainable AI hardware/software codesign" has been presented at the following premier forums:

- NSF/Intel Annual CAPA research center days – 2018, 2019, 2020 (closed events)
  Audience: NSF program managers, Intel research managers and fellows, researchers from top universities
- Renegotiating the Levels of Abstraction for the Post Moore's Law Era, a Workshop at the Arm Research Summit, Austin, US – 2019
- Third Annual Future Chips Forum, "Reconfigurable Computing in Golden Age", Beijing Innovation Center for Future Chips and the Institute of Microelectronics, Tsinghua University – 2018
- ARM Research Summit, Cambridge, UK – 2018
- DAC-ROAD4NN: Annual International Workshop on Research Open Automatic Design for Neural Networks, co-located with the 59th Design Automation Conference (DAC 2022) – 2022 (usually 100s of attendees)
- Annual IBM and IEEE CAS/EDS AI Compute Symposium (AICS) – 2022
- Intel Labs Seminar (Virtual) – 2023

[1] Based on my profiles and citation notifications on Google Scholar, Semantics Scholar, ResearchGate, etc.

SELECT WORK FEATURED

My work has been posted, highlighted, or discussed by various experts on different forums and social media channels, including Twitter, Reddit, Hacker News, DeepAI, and arXiv-based platforms. In addition, my work has been showcased on following public pages:

- Intel Labs Select Publications
- IEEE HKN Graduate Research Spotlight in IEEE Bridge, volume 119 and on various social media
- insideHPC article on Energy-efficient Acceleration of Residual Neural Nets using CGRAs
- ARM Research Summit Talk on ARM Research YouTube Channel
- DAC recommended sessions from industry experts (session "The Art of Mapping for Accelerators")

PUBLICATIONS

My research areas include design automation, computer architecture, and embedded systems. My research is regularly published in (and referred by publications in) the top venues in these areas[2]. My most of the publications are first authored. More information (author's version of the PDFs, presentation slides, posters, recorded YouTube talks, summaries, or open-source tools/artifacts) is available at: https://sites.google.com/view/shail/publications.

Select Publications

[ASPLOS] EXPLAINABLE-DSE: An Agile and Explainable Exploration of Hardware/Software Codesigns of Deep Learning Accelerators, Shail Dave, Tony Nowatzki, and Aviral Shrivastava. In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS). [Teaser]

- Key topics: Enabling Explainability in hardware/software codesign exploration of domain-specific computing architectures, which can characterize obtained designs and why optimization selected those designs; Formalizing specification of bottleneck analysis for domain-specific design space exploration; Explainable DSE using bottleneck analysis leads to several-fold efficient solutions under design constraints, while taking orders-of-magnitude lower search time (minutes-hour vs. days—weeks).

- **Silver Medal at ACM Student Research Competition, Host: ACM SIGBED**

- Invited presentations at top industrial and academic forums.

[CODES+ISSS] dMazeRunner: Executing Perfectly Nested Loops on Dataflow Accelerators. **Shail Dave**, Youngbin Kim, Sasikanth Avancha, Kyoungwoo Lee, Aviral Shrivastava, in ACM Transactions on Embedded Computing Systems (TECS), Vol. 18, No. 5s, 2019 [Special Issue on ESWEEK 2019 - ACM/IEEE International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)]. **(55+ citations in 3 years; 2nd most cited CODES+ISSS 2019 paper and 2nd most downloaded ESWEEK 2019 paper on ACM digital library; 35+ GitHub stars and several monthly downloads)**

- Collaboration with authors from Yonsei, Intel Parallel Computing Lab, and ASU.
- *Key topics: Defining comprehensive hardware/software design space for loop nests; Accelerator cost model for evaluating execution metrics for variations in hardware architecture, mappings, deep learning model layers; Search-space reduction techniques for getting efficient mappings in a few seconds; Generic algorithms for obtaining all unique data reuse scenarios for loop-orderings.*

[Proceedings of the IEEE] Hardware Acceleration of Sparse and Irregular Tensor Computations of ML Models: A Survey and Insights. **Shail Dave**, Riyadh Baghdadi, Tony Nowatzki, Sasikanth Avancha, Aviral Shrivastava, Baoxin Li, in Proceedings of the IEEE, volume 109, issue 10, 2021, pages 1706-1752. **(45+ citations) (journal impact factor: 14.91)**

- Collaboration with authors from NYU/MIT, UCLA, Intel Parallel Computing Lab, and ASU.
- *Key topics: Compact deep/machine learning models, their need, and sources of tensor compression; Implications of irregular or structured sparsity on acceleration; Accelerator-aware DNN model pruning; circuit/architecture/mapping/model level techniques and their qualitative/quantitative analysis for performance, energy efficiency, storage, or chip area; Impact of varying sparsity and tensor shapes of different DNN operators on data reuse and storage efficiency; Quantitatively understanding achievable speedups for*

[2] Google Scholar publication metrics for computer hardware design

recent DNNs; Techniques for data extraction and load balancing of effectual computations; Sparsity-aware dataflows; Leveraging value similarity and approximation in temporal and spatial data of computer vision and speech processing applications; Trends and directions for accelerator/model codesigns for DNNs.

[DAC] RAMP: Resource-Aware Mapping for CGRAs. **Shail Dave**, Mahesh Balasubramanian, Aviral Shrivastava, in Proceedings of the 55th Annual Design Automation Conference (DAC), 2018. **(65+ citations in 4 years; regarded as a state-of-the-art technique for mapping general-purpose loops on CGRAs with comprehensive P&R strategies. In top 5% highly cited papers as per Google Scholar among 1000+ published papers at DAC from 2017-2022)**

- One of the industry experts-recommended DAC 2018 session.
- *Key topics: Mapping optimizations for accelerating loops of general-purpose computing through software pipelining; Failure analysis-driven and resource-aware routing of the data dependencies during mapping.*

[Demonstrations @ DATE, DAC] CCF: CGRA Compilation and Simulation Framework. Shail Dave, Aviral Shrivastava, in University Booth Demonstration at the 21st International Conference on Design Automation and Test in Europe (DATE), 2018 (also selected for university demonstration at 55th annual Design Automation Conference). **(25+ total citations, 65+ GitHub stars, and even more direct downloads and research usages.)**

Under Progress/Submission

- Automating Execution Modeling and Characterization for Domain-Specific Accelerators.

Other Publications (Chronological Order)

[DATE] Learning-Oriented Reliability Improvement of Computing Systems from Transistor to Application Level (Invited Special Session). Behnaz Ranjbar, Florian Klemme, Paul R. Genssler, Hussam Amrouch, Jinhyo Jung, **Shail Dave**, Hwisoo So, Kyongwoo Lee, Aviral Shrivastava, Ji-Yung Lin, Pieter Weckx, Subrat Mishra, Francky Catthoor, Dwaipayan Biswas, Akash Kumar, in Proceedings of the 26th International Conference on Design Automation and Test in Europe (DATE), 2023.

[VTS] Towards an Agile Design Methodology for Efficient, Reliable, and Secure ML Systems. **Shail Dave**, Alberto Marchisio, Muhammad Abdullah Hanif, Amira Guesmi, Aviral Shrivastava, Ihsen Alouani, Muhammad Shafique, in Proceedings of the 40th IEEE VLSI Test Symposium (VTS), 2022 (Invited special session).
Key topics: Agile design methodology for developing efficient neural processing unit architectures (NPUs); Mitigating reliability challenges such as soft errors, process variations, aging, and manufacturing defects; Securing AI systems with robustness to faults, adversarial attacks, and privacy challenges.

[TACO] SPX64: A Scratchpad Memory for General-Purpose Microprocessors. Abhishek Singh, **Shail Dave**, PanteA Zardoshti, Robert Brotzman, Chao Zhang, Xiaochen Guo, Aviral Shrivastava, Gang Tan, Michael Spear, in ACM Transactions on Architecture and Code Optimization (TACO), Vol. 18, No. 1, 2021 [Presented by Invitation at HiPEAC 2021 - 16th International Conference on High-Performance Embedded Architectures and Compilers].
Key topics: Software-managed cache; Securing execution against side-channel attacks; Accelerating persistent transactions for non-volatile memory.

[ICASSP] dMazeRunner: Optimizing Convolutions and GEMMs on Dataflow Accelerators. **Shail Dave**, Aviral Shrivastava, Youngbin Kim, Sasikanth Avancha, Kyoungwoo Lee, in Proceedings of the 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020.

[DATE] URECA: A Compiler Solution to Manage Unified Register File for CGRAs. **Shail Dave**, Mahesh Balasubramanian, Aviral Shrivastava, in Proceedings of the 21st International Conference on Design Automation and Test in Europe (DATE), 2018.

[DATE] LASER: A Hardware/Software Approach to Accelerate Complicated Loops on CGRAs. Mahesh Balasubramanian, **Shail Dave**, Aviral Shrivastava, Reiley Jeyapaul, in Proceedings of the 21st International Conference on Design Automation and Test in Europe (DATE), 2018.
Key topics: Mapping optimizations for accelerating imperfectly nested loops and loops with conditionals from general-purpose computing applications.

[INROADS] Derivation of transfer function model based on Miniaturized cryocooler behavior. Jiten Bhatt, **Shail Dave**, Manish M. Mehta, and Nitin Upadhyay. INROADS, volume 5, no. 1s (2016): 336-340.

- Collaboration with senior researchers and group directors from Space Application Center, Indian Space Research Organization (undergraduate final project).
- *Key topics: Achieving precise temperature control of cryogenic cooler systems for next-gen geostationary satellites; End-to-end system that automatically fine-tunes the controller in presence of variations from mechanical and electrical components of the system.*

Workshop Papers and Presentations

[LATTE at ASPLOS] Design Space Description Language for Comprehensive Exploration of Next-Gen Hardware Accelerators. **Shail Dave**, Aviral Shrivastava, in Second Workshop on Languages, Tools, and Techniques for Accelerator Design (LATTE), co-located with ACM 27th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2022. **(Vision/Position paper)**

- *Key topics: Overcoming limitations of existing architecture design frameworks (that use template specified in architecture description language) with design space description language; Improving: (1) Comprehensiveness of design space (2) Reusability of design tools (3) Explainability of designs and their exploration, and (4) Exploration speed for making dynamic deployment feasible.*
- *Led to new project fundings, including from Semiconductor Research Corporation AI Hardware Program.*

[SNN, co-located with ICML] Efficient Processing of Sparse and Compact DNN Models on Hardware Accelerators. Survey and Insights. In annual Sparsity in Neural Networks workshop, 2022. Shail Dave and Aviral Shrivastava [Summary Poster]

- Joint work with Riyadh Baghdadi (MIT CSAIL and NYU), Tony Nowatzki (UCLA), Sasikanth Avancha (Intel Labs). and Baoxin Li (ASU).

Patents[3]

1. Hybrid and efficient approach to accelerate complicated loops on coarse-grained reconfigurable arrays (CGRA) accelerators, Mahesh Balasubramanian, Shail Dave, Aviral Shrivastava, and Reiley Jeyapaul. (Application number: US20200133672A1; Provisional patent filed in 2018).
2. Systems And Methods for Dynamic and Efficient Hardware/Software Codesigns of Deep Learning Accelerators. Provisional patent filed by ASU in 2022.
3. Explainable Computing Systems Designs and Explainable Optimizations for Efficiency and Productivity. Provisional patent filed by ASU in 2022.

TECHNICAL SKILLS

- **Programming Languages & HDL:** C, C++, Python, MIPS Assembly, Shell Scripting, Verilog, System Verilog.
- **Parallel Programming Libraries:** OpenMP, POSIX, MPI, OpenCL, CUDA.
- **Tools and Frameworks:** Gem5 Processor Simulator, LLVM Compiler (Front-End/Middle End), PyTorch, Apache TVM, Xilinx Vivado and Vivado HLS, Deep Learning Model Compression Tools, Altera ModelSim, Cadence and Synopsys ASIC design tools, MATLAB, GDB, Version management tools.

[3] My recent research was supported through NSF/Intel program, which required the work to be public, so no patents could be filed from my research group during the 3–4-year timespan. For my most recent works on cutting-edge design methodologies for novel architectures, some are under filing process, and I plan to file a few more patents.

- **Relevant Courses:** Deep Learning, Hardware Acceleration, High Performance Computing, Computer Architecture, Distributed and Multi-Processing Operating Systems, VLSI Design, Hardware Design and Verification Languages, Statistical Machine Learning, Algorithms, Topics in Machine Learning Accelerator Design.

RESEARCH PROJECTS

- Design Space Language and Framework for Exploring Next-Gen Hardware Accelerators — Spring 2022–Present
- Automated Cost Modeling for Arbitrary Dataflow Accelerators — Spring 2021, Fall 2022–Present
- Making Computer Architecture Development/Simulation in Gem5 Easier with LLMs (Mentoring a master's student on thesis) — Fall 2022–Present
- Explainable Design Space Exploration of Hardware/Software Codesigns — Fall 2020, Summer–Fall 2021
- Analyzing Sparse Tensor Computations of Machine Learning Models on Accelerators — Spring 2020–Spring 2021
- General-Purpose Scratchpad Memory for Secure, Persistent Execution — Fall 2019–Fall 2020
- Design Modeling and Mapping Optimizations for Deep Learning Accelerators — Fall 2018–Fall 2019
- Compiler and Simulator Infrastructure for — Spring 2016–Fall 2018
- Mapping General-purpose Loops on Coarse-Grain Reconfigurable Architectures (CGRAs) — Spring 2016–Spring 2018

PROJECT GRANT PROPOSALS WRITING

In addition to the writing, I have assisted senior researchers (PIs) for conceptualization of novel projects and analyzing/laying challenges and research tasks for the following proposals.

- FMSG: Cyber: Co-design of Microstructure Models and Hardware Accelerator for Real-time Additive Control in Additive Manufacturing (ASU Multiple PIs) (NSF Future Manufacturing Seed Grant)
- *Explainable and Agile AI Hardware/Software Optimization with AI* (ASU) (SRC AIHW proposal)
- *Hardware-software Co-design for Delegated Homomorphic Computation* (ASU, Multiple PIs) (SRC proposal)
- *Sustainable Full System Stack Development for Next-Generation AI Hardware* (Meta AI research)
- *Concertina-A Language & Framework for Time-Sensitive Machine Learning Applications*. (CMU, ASU) (NSF proposal)
- *Automated Accident Detection Using Machine Learning* (ASU) (Arizona MCDOT proposal)
- PPoSS: *Towards Wide-scale, Time-sensitive, Federated Learning Applications*. (ASU multiple PIs) (NSF proposal)
- FoMR: *System-wide Reconfigurable Acceleration for Near-Data Computing*. (ASU and Lehigh) (NSF proposal)
- *Execute DNNs on Programmable Dataflow Accelerators: Where's your System Stack?* (ASU) (Facebook AI research)
- *Acceleration Beyond GPUs for High-Performance Computing*. (ASU) (SRC, NSF proposal)

PROFESSIONAL SERVICE

Based on my research expertise and accomplishments, I am regularly invited to serve as an external/expert reviewer for the top/flagship conferences and journals in my research areas, including design automation, computer architecture, and embedded systems. Note that serving on the program committee (PC) of top conferences typically requires PhD, and therefore external experts including PhD students like me are invited as an external reviewer for a few papers (e.g., review load of 2-4 papers vs 10-15 for a PC member). My reviews are usually regarded by area editors or PC members as detailed and thorough.

Program Committee Member

- Second Workshop on Languages, Tools, and Techniques for Accelerator Design (LATTE), co-located with ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS) – 2022, 2023

External Reviewer

- ACM Design Automation Conference (DAC) – 2018, 2019, 2020, 2022
- IEEE Real-Time Systems Symposium (RTSS) – 2019
- ACM/IEEE International Conference on Compilers, Architecture & Synthesis for Embedded Systems (CASES) – 2019, 2023
- IEEE/ACM International Conference on Hardware-Software Codesign & System Synthesis (CODES+ISSS) – 2016, 2017

- IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM) – 2019
- IEEE International Conference on Design Automation and Test in Europe (DATE) – 2016, 2017
- IEEE International Conference on Computer Design (ICCD) – 2019
- IEEE Asia South Pacific Design Automation Conference (ASP-DAC) – 2022
- IEEE International Conference on VLSI Design and Embedded Systems (VLSID) – 2017, 2018

Journal Referee

- IEEE Transactions on Circuits and Systems I: Regular Papers – 2023
- IEEE Sensors Journal – 2022
- ACM Transactions on Embedded Computing Systems (TECS) – 2017, 2019, 2020, 2022
- ACM Transactions on Design Automation of Electronic Systems (TODAES) – 2019, 2020
- IEEE Transactions on Multi-Scale Computing Systems (TMSCS) – 2016, 2017
- Springer International Journal on Design Automation for Embedded Systems (DAES) – 2017

Panelist

- NSF Workshop on Sustainable Computing (SUSCOMP) - 2022 (Topical Expert and Scribe)
- IEEE HKN GradLab: Graduate Student Podcast – 2022 [YouTube] [Organizer: IEEE Eta Kappa Nu (HKN)]
- Early Student Career Panel, ASU 101 – 2016, 2017
- Invited Speaker and Panelist, ASU International Student Orientation – 2016, 2017
  (Joint planning with ASU GPSA – Graduate and Professional Student Association and ISSC – International Students and Scholars Center)

Organization

- Social Media Chair, ACM/IEEE Embedded Systems Week (ESWEEK) –2021, 2022
  Impact: Established, diversified, and doubled social media presence with ~800/230/380 followers on LinkedIn/Twitter/YouTube. In addition, I helped General Chairs in various ways, including (a) suggesting and helping establish several new calls, initiatives, and awards; (b) conference planning (overall program organization, activities for virtual and in-person attendance, etc.); (c) website maintenance.
- Vice Chair, IEEE Eta Kappa Nu, ASU Chapter, 2016–2017
- Logistics management, Distinguished Speaker Seminars, School of Computing, Informatics, and Decisions Systems Engineering, ASU – 2017, 2018
- Special Session Organization, IEEE 40th VLSI Test Symposium – 2022
- Webmaster, Compiler Microarchitecture Lab / Make Programming Simple Lab, 2017–2022

Mentorship (Beyond Preliminary Affiliations)

- Mentor, Computer Architecture Long-Term Mentoring (CALM) – 2022-Present
  Year-long mentoring program by ACM SIGARCH Computer Architecture Student Association (CASA)
- Mentor, Meet a Student Architect (MaSA) – 2022

PROFESSIONAL ASSOCIATIONS

I hold student membership of and regularly participate into events (in various capacity) of the following organizations:

- Association for Computing Machinery (ACM)
- Institute of Electrical and Electronics Engineers (IEEE)
- New York Academy of Sciences (NYAS)
- IEEE Computer Society (IEEE CS)
- IEEE Eta Kappa Nu Honors Society (IEEE HKN)
- ACM Special Interest Group on Computer Architecture (SIGARCH)
- ACM Special Interest Group on Embedded Systems (SIGBED)
- IEEE Council on Design Automation (CEDA)
- IEEE Technical Committee on Computer Architecture (TCCA)
- Computer Architecture Student Association (CASA) (supported by: ACM SIGARCH / IEEE TCCA)

TEACHING

Course Developed, School of Computing and Augmented Intelligence, ASU

- Topics in Machine Learning Accelerator Design – Spring 2023 [Reading List]
  (Total enrollment: 30+ masters and PhD students; students are also encouraged to submit topical surveys. We anticipate development and sharing of new material through upcoming offerings.)

Guest Lecturer, School of Computing, Informatics and Decision Systems Engineering, ASU

- ASU 101 (The ASU Experience) – Fall 2018
- CSE 420/520 (Computer Architecture) – Fall 2018 (on ML accelerators), Fall 2021 (several tutorials on Gem5, including implementing various branch predictors and cache replacement policies that are not in Gem5)

Teaching Assistant, School of Computing, Informatics and Decision Systems Engineering, ASU – 2016–2018, 2021

- CSE 100 (Principles of Programming with C++)
- CSE 330 (Operating Systems)
- CSE 420/520 (Computer Architecture I and II) **(Outstanding TA Award in 2018)**

Instructional Aide, School of Mathematics and Statistical Sciences, Arizona State University – 2014–2016

- MAT 210 (Brief Calculus)
- MAT 142 (College Mathematics)
- MAT 117 (College Algebra)

OUTREACH AND MENTORING

Advising and Mentoring

During my research at ASU, I have regularly collaborated with or guided master's and PhD peer students, including under-represented (about 30%-40% of my advisees) and non-thesis-track master's students, in various capacities for their learning of cutting-edge technology topics and tools and professional growth. They are now senior research engineers or architecture engineers at international companies like Qualcomm, Intel, Apple, Cadence, ON semiconductor, MediaTech, Cirrus Logic, and one is now a PhD student at North Carolina State University.

I have now also signed up as a mentor for CALM: Computer Architecture Long-term Mentoring program and MaSA: Meet a Senior Architect, run by Computer Architecture Student Association (CASA) (supported by: ACM SIGARCH / IEEE TCCA). Through this global program, I plan to regularly mentor non-ASU students every year, especially juniors who may not have accesses to sufficient technical/professional resources, are less privileged, or are under-represented. I also plan to sign up for upcoming mentorship programs e.g., from IEEE HKN or IEEE Young Professionals.

Selected Outreach for Professional Growth

I regularly serve on student seminar series or panels, discussing about the research and professional growth opportunities. I have served as Vice chair of IEEE HKN ASU chapter, where I established "luncheon with faculty" program. It allowed students to meet faculty of their research interest from SECEE or SCAI and have discussions on a broad range of topics, beyond their courses. I also served for a few years as an invited speaker for ASU international student orientation, discussing various opportunities at ASU and how to build a strong professional profile. I have also served on various student panels, including GradLab podcast organized by IEEE HKN, where I discussed navigating graduate school challenges and opportunities, and ASU 101, where I discussed with freshman students about computer science and engineering and their questions on various opportunities and career paths during and after their graduation. I have also assisted in hosting distinguished speakers for SCAI seminar series at ASU, including renowned researchers from Stanford, University of Minnesota, KAIST, UT Austin, etc.