Vijay Janapa Reddi, Ph.D.

Gordon McKay Professor of Electrical & Computer Engineering
Harvard John A. Paulson School of Engineering and Applied Sciences

Harvard Science & Engineering Complex, Room #5.305, 150 Western Avenue, Boston, MA 02134
PHONE +1 (408) 390–2790 EMAIL vj@eecs.harvard.edu
WEB vijay.seas.harvard.edu

Gordon McKay Professor of Electrical & Computer Engineering at Harvard SEAS, Vice President of MLCommons, and (from July 2026) Visiting Professor at ETH Zürich. Co-architect of the MLPerf Inference benchmark—the industry-standard ML benchmarking suite adopted by Google, Microsoft, NVIDIA, Meta, AMD, and Intel—and founder of the open-source *Machine Learning Systems* curriculum used at 50+ universities across 5 continents, with 100,000+ learners reached through the TinyML edX certificate. Two MIT Press textbooks forthcoming (2026 and 2027). Ph.D., Harvard 2010; previously Associate Professor at UT Austin.

Google Scholar: 21,322 citations · h-index 61 · i10-index 162.

Current Position

Gordon McKay Professor of Electrical & Computer Engineering
John A. Paulson School of Engineering and Applied Sciences
Harvard University

Primary Research Areas

Computer Architecture; Machine Learning Systems; Autonomous Agents/Robotics

Education

2010 Ph.D. in Computer Science, Harvard University

2006 M.S. in Electrical and Computer Engineering, University of Colorado—Boulder

2003 B.S. in Computer Engineering, Santa Clara University

Current and Previous Academic Positions

2026/– Visiting Professor, ETH Zürich (starting July 2026)

2025/– Gordon McKay Professor of Electrical & Computer Engineering, Harvard University

2019–2024 John L. Loeb Associate Professor of Engineering and Applied Sciences, Harvard University

2018/– Adjunct Associate Professor, The University of Texas at Austin

2017–2018 Associate Professor, The University of Texas at Austin

2011–2017 Assistant Professor, The University of Texas at Austin

Other Industry Experience

2024–2025 Visiting Researcher, Google (Research)

2022–2023 Visiting Researcher, Google (DeepMind)

2020–2021 Visiting Researcher, Google (Tensor Flow Ecosystem)

2019 Visiting Researcher, Facebook (AR Silicon Team)
2017–2018 Visiting Researcher, Google (gChips)
2015–2016 Consultant, Intel
2015–2016 Consultant, Advanced Micro Devices (AMD)
2014 Consultant, Intel
2010–2011 Senior Design Engineer, Advanced Micro Devices (AMD)
2009 Research Intern, Microsoft Research
2007–2009 Research Intern, VMware
2003–2006 Research Intern, Intel

Broader Impacts

Machine Learning Systems (MLSysBook.AI)

Open-access textbook and curriculum ecosystem for AI engineering education.

Reach 50+ universities across 5 continents teach from the curriculum; 22,800+ GitHub stars; 95+ contributors. Goal: 1M learners by 2030.

Scope 32 chapters (Vol. I: Foundations, Vol. II: At Scale), 35+ lecture decks, 20 TinyTorch modules, and 9,000+ interview questions in the StaffML bank.

Books Two MIT Press hardcover volumes forthcoming (2026, 2027).

Global TinyML4D Academic Network with ICTP—curriculum, mentorship, and hardware kits delivered to universities across Latin America, Africa, and Asia. Annual SciTinyML workshop series.

MLPerf & MLCommons (mlcommons.org)

Co-founded and serve as Vice President of MLCommons, the industry-standard ML benchmarking consortium adopted by Google, Microsoft, NVIDIA, Meta, AMD, Intel, and others.

2019/– Vice President and Board of Directors, MLCommons.
2019/– Research Co-Chair, MLCommons.

Impact Co-led the MLPerf Inference benchmark, now the industry standard across datacenter, edge, mobile, and IoT. MLPerf Inference selected for inclusion in the ISCA@50 25-Year Retrospective.

TinyML on edX

Founded and led the TinyML Professional Certificate on HarvardX/edX in partnership with Google and the tinyML Foundation, with **100,000+ learners** reached worldwide. Recognized as one of ClassCentral's 100 Most Popular Free Online Courses (2021) and winner of the CogX Best Course in AI award (2021).

Advisory & Board Roles

2024/– Board Member, EDGE AI Foundation (edgeaifoundation.org)
2020–2023 Member, International Roadmap for Devices and Systems (IRDS)
2019/– Vice President & Board of Directors, MLCommons (mlcommons.org)
2019/– Research Co-Chair, MLCommons
2019–2020 MLPerf Tiny Co-Chair, MLPerf (mlperf.org)

2018–2020 MLPerf Inference Co-Chair, MLPerf

2017–2020 Associate Editor, SIGARCH Blog (sigarch.org/blog)

Honors and Awards

2026 IEEE Micro Top Picks in Computer Architecture

2025 Best of Computer Architecture Letters (CAL), Editorial Board of IEEE CAL

2025 IEEE Micro Top Picks in Computer Architecture

2024 Best Paper Award, Vail Computer Elements Workshop (VCEW)

2023 MLPerf Inference selected for inclusion in ISCA@50 25-Year Retrospective

2023 IEEE Micro Top Picks in Computer Architecture (Honorable Mention)

2022 IEEE Micro Top Picks in Computer Architecture (Honorable Mention)

2021 BenchCouncil Rising Star Award, International Open Benchmark Council

2021 Deploying TinyML on HarvardX/edX: 100 Most Popular Free Online Courses, ClassCentral

2021 Best Course in AI: Tiny Machine Learning (TinyML) on HarvardX/edX, CogX Awards

2021 IEEE Micro Top Picks in Computer Architecture

2021 Best of Computer Architecture Letters (CAL), Editorial Board of IEEE CAL

2020 Programming Languages Software Award, ACM SIGPLAN

2020 Best Research Paper Award, Design Automation Conference (DAC)

2020 Google Faculty Research Award, Google

2019 Best Paper Nominee, IEEE International Symposium on Perf. Analysis of Systems and Software (ISPASS)

2019 Intl Symp. on High-Performance Computer Architecture (HPCA) Hall of Fame

2018 International Symp. on Microarchitecture (MICRO) Hall of Fame

2018 ACM SIGARCH CS TCCA Outstanding Dissertation Award (Advisee: Yuhao Zhu)

2017 IEEE Micro Top Picks in Computer Architecture

2017 Best Paper Nominee, Design Automation Conference (DAC)

2017 Google Faculty Research Award

2016 IEEE TCCA Young Computer Architect Award

2016 IEEE Micro Top Picks in Computer Architecture (Honorable Mention)

2016 Gilbreth Lectureship Honor, National Academy of Engineering (NAE)

2015 ACM SIGPLAN Most Influential PLDI Paper Award

2015 Google Faculty Research Award

2014 Best of Computer Architecture Letters (CAL) Award

2014 Best Paper Nominee, IEEE International Symposium on Microarchitecture Low Power Electronics and Design (ISLPED)

2014 Indo-American Frontiers of Engineering, National Academy of Engineering (NAE)

2013 Google Faculty Research Award

2013 Intel Early Career Award

2012 Google Faculty Research Award
2011 IEEE Micro Top Picks in Computer Architecture
2010 IEEE Micro Top Picks in Computer Architecture
2009 Best Paper Award, International Symposium on High-Performance Computer Architecture (HPCA)
2008 John A. and Elizabeth S. Armstrong Fellowship, Harvard University
2007 Best Student Presentation, International Symposium on Code Generation and Optimization (CGO)
2006 IEEE Micro Top Picks in Computer Architecture
2005 Best Paper Award, International Symposium on Microarchitecture (MICRO)
2003 Faculty Recognition for Technical Excellence, Santa Clara University
2003 Outstanding Undergraduate (Honorable), Computing Research Association (CRA)

University Committee Assignments

Harvard John A. Paulson School of Engineering and Applied Sciences

2023–2024 Member, Engineering Sciences Committee on Higher Degrees
2023 Member, Generative AI Steering Committee
2020 Member, Quantum Faculty Recruiting Committee
2019 Member, Robotics Faculty Recruiting Committee
2019/– Graduate Student Admissions Committee

The University of Texas at Austin

2016 Member, Faculty Recruiting Committee
2015 Member, Faculty Recruiting Committee
2014 Member, Technology in Teaching
2013 Member, Faculty Recruiting Committee
2011–2016 Graduate Student Admissions Committee

Professional Society Memberships

Institute of Electrical and Electronics Engineers (IEEE)
Association for Computing Machinery (ACM)

Professional Service

General Chair

2023 Tiny Machine Learning Research Symposium (TinyML)
2022 Tiny Machine Learning Research Symposium (TinyML)
2021 Tiny Machine Learning Research Symposium (TinyML)
2017 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)

Program Chair

2019 IEEE International Symposium on Workload Characterization (IISWC)

2014 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)

Program Committee

2025 Conference on Neural Information Processing Systems (NeurIPS)
2025 IEEE/ACM International Symposium on Microarchitecture (MICRO)
2024 IEEE/ACM International Symposium on Microarchitecture (MICRO)
2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)
2023 International Symposium on Computer Architecture (ISCA)
2023 Sixth Conference on Machine Learning and Systems (MLSys)
2023 IEEE/ACM International Symposium on Microarchitecture (MICRO)
2022 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)
2022 International Symposium on Computer Architecture (ISCA)
2022 Fifth Conference on Machine Learning and Systems (MLSys)
2021 ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)
2021 IEEE Micro Top Picks
2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)
2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)
2020 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)
2020 International Symposium on Computer Architecture (ISCA)
2020 IEEE/ACM International Symposium on Microarchitecture (MICRO)
2019 ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS) (ERC)
2019 IEEE Micro Top Picks
2019 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)
2019 International Symposium on Computer Architecture (ISCA)
2019 IEEE/ACM International Symposium on Microarchitecture (MICRO)
2018 International Symposium on Computer Architecture (ISCA) (ERC)
2018 IEEE Micro Top Picks
2018 IEEE/ACM International Symposium on Microarchitecture (MICRO)
2018 IEEE International Symposium on Workload Characterization (IISWC)
2017 IEEE International Symposium on High-Performance Computer Architecture (HPCA) (ERC)
2017 International Symposium on Computer Architecture (ISCA)
2017 IEEE/ACM International Symposium on Microarchitecture (MICRO)
2016 IEEE International Symposium on High-Performance Computer Architecture (HPCA) (ERC)
2016 IEEE Micro Top Picks
2016 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)
2016 IEEE International Symposium on Workload Characterization (IISWC)
2015 International Symposium on Computer Architecture (ISCA) (ERC)

2015 IEEE International Symposium on High-Performance Computer Architecture (HPCA)
2015 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)
2015 IEEE/ACM International Symposium on Microarchitecture (MICRO)
2015 International Symposium on Performance Analysis of Systems and Software (ISPASS)
2015 International Symposium on Principles and Practice of Parallel Computing (PPoPP)
2015 IEEE International Symposium on Workload Characterization (IISWC)
2014 IEEE International Symposium on High-Performance Computer Architecture (HPCA)
2014 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)
2014 IEEE/ACM International Symposium on Microarchitecture (MICRO)
2013 IEEE International Symposium on Workload Characterization (IISWC)

Organizing Committee

2023 Data-centric Machine Learning Research (DMLR) Workshop (ICML)
2020 ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)
2017 Design Automation Conference
2016–2017 Workshop on Cognitive Edge Computing (CogEdge)
2016 Tutorial on Tools for Mobile Computer Architecture (MobiTools)
2015 IEEE/ACM International Symposium on Code Generation and Optimization
2016 IEEE International Symposium on Microarchitecture
2015–2016 Tutorial on Simulation and Analysis Engine (ISCA, ASPLOS, HPCA, ICS, IISWC, ISPASS)
2013 IEEE International Symposium on Workload Characterization
2012 IEEE International Symposium on Performance Analysis of Systems and Software
2013 IEEE International Symposium on Performance Analysis of Systems and Software.

Guest Editor

2023 IEEE Micro Special Issue on Tiny Machine Learning
2016 IEEE Micro Special Issue on Internet of Things
2013 IEEE Micro Special Issue on Reliability-Aware Microarchitecture Design

Local Arrangements Chair

2013 Intl. Symp. on Performance Analysis of Systems and Software (ISPASS)
2015, 2016 Workshop on Silicon Errors in Logic: System Effects (SELSE).

Publications Chair

2013 Intl. Symp. on Workload Characterization (IISWC)

Community Activities

Tiny Machine Learning Open Education Initiative, `https://tinyml.seas.harvard.edu` (Founder)
Hands-on Computer Science (HaCS) for Austin Independent School District (via UT Outreach),

https://outreach.uteach.utexas.edu/uteach-outreach-cs-service-learning-program

PUBLICATIONS

Conference Publications

2026 Shvetank Prakash, Andrew Cheng, Olof Kindgren, Ashiq Ahamed, Graham Knight, Jedrzej Kufel, Francisco Rodriguez, Arya Tschand, David Kong, Mariam Elgamal, Jerry Huang, Emma Chen, Gage Hills, Richard Price, Emre Ozer, and Vijay Janapa Reddi. "Lifetime-Aware Design for Item-Level Intelligence at the Extreme Edge." In *ASPLOS 2026*. [DOI]

2025 David Kong, Shvetank Prakash, Jedrzej Kufel, Georgios Kyriazidis, Yasmine Omri, David Verity, Emre Ozer, Vijay Janapa Reddi, and Gage Hills. "333-eDRAM - 3T Embedded DRAM Leveraging Monolithic 3D Integration of 3 Transistor Types: IGZO, Carbon Nanotube and Silicon FETs." In *DAC 2025*. [DOI]

2025 Subhankar Pal, Aporva Amarnath, Behzad Boroujerdian, Augusto Vega, Alper Buyuktosunoglu, John-David Wellman, Vijay Janapa Reddi, and Pradip Bose. "ARTEMIS: Agile Discovery of Efficient Real-Time Systems-on-Chips in the Heterogeneous Era." In *HPCA 2025*. [DOI]

2025 Zishen Wan, Jiayi Qian, Yuhang Du, Jason Jabbour, Yilun Du, Yang Zhao, Arijit Raychowdhury, Tushar Krishna, and Vijay Janapa Reddi. "Generative AI in Embodied Systems: System-Level Analysis of Performance, Efficiency and Scalability." In *ISPASS 2025*. [DOI]

2025 Arya Tschand, Arun Tejusve Raghunath Rajan, Sachin Idgunji, Anirban Ghosh, Jeremy Holleman, Csaba Király, Pawan Ambalkar, Ritika Borkar, Ramesh Chukka, Trevor Cockrell, Oliver Curtis, Grigori Fursin, Miro Hodak, Hiwot Kassa, Anton Lokhmotov, Dejan Miskovic, Yuechao Pan, Manu Prasad Manmathan, Liz Raymond, Tom St. John, Arjun Suresh, Rowan Taubitz, Sean Zhan, Scott Wasson, David Kanter, and Vijay Janapa Reddi. "MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from \(\mu\)Watts to MWatts for Sustainable AI." In *HPCA 2025*. [DOI]

2025 Peiqing Chen, Minghao Li, Zishen Wan, Yu-Shun Hsiao, Minlan Yu, Vijay Janapa Reddi, and Zaoxing Liu. "OctoCache: Caching Voxels for Accelerating 3D Occupancy Mapping in Autonomous Systems." In *ASPLOS 2025*. [DOI]

2025 Zishen Wan, Yuhang Du, Mohamed Ibrahim, Jiayi Qian, Jason Jabbour, Yang Katie Zhao, Tushar Krishna, Arijit Raychowdhury, and Vijay Janapa Reddi. "ReCA: Integrated Acceleration for Real-Time and Efficient Cooperative Embodied Autonomous Agents." In *ASPLOS 2025*. [DOI]

2024 Jessica Quaye, Alicia Parrish, Oana Inel, Charvi Rastogi, Hannah Rose Kirk, Minsuk Kahng, Erin van Liemt, Max Bartolo, Jess Tsang, Justin White, Nathan Clement, Rafael Mosquera, Juan Ciro, Vijay Janapa Reddi, and Lora Aroyo. "Adversarial Nibbler: An Open Red-Teaming Method for Identifying Diverse Harms in Text-to-Image Generation." In *FAccT 2024*. [DOI]

2024 Maximilian Lam, Jeff Johnson, Wenjie Xiong, Kiwan Maeng, Udit Gupta, Yang Li, Liangzhen Lai, Ilias Leontiadis, Minsoo Rhu, Hsien-Hsin S. Lee, Vijay Janapa Reddi, Gu-Yeon Wei, David Brooks, and G. Edward Suh. "GPU-based Private Information Retrieval for On-Device Machine Learning Inference." In *ASPLOS 2024*. [DOI]

2024 Jason Jabbour and Vijay Janapa Reddi. "Generative AI Agents in Autonomous Machines: A Safety

Perspective." In *ICCAD 2024*. [DOI]

2024 Biyan Zhou, Pao-Sheng Vincent Sun, Jason Yik, Charlotte Frenkel, Vijay Janapa Reddi, and Arindam Basu. "Grand Challenge on Neural Decoding for Motor Control of non-Human Primates." In *BioCAS 2024*. [DOI]

2024 Sabrina M. Neuman, Brian Plancher, and Vijay Janapa Reddi. "Invited: The Magnificent Seven Challenges and Opportunities in Domain-Specific Accelerator Design for Autonomous Systems." In *DAC 2024*. [DOI]

2024 Vijay Janapa Reddi. "MLSysBook.AI: Principles and Practices of Machine Learning Systems Engineering." In *ISSS 2024*. [DOI]

2024 Zishen Wan, Nandhini Chandramoorthy, Karthik Swaminathan, Pin-Yu Chen, Kshitij Bhardwaj, Vijay Janapa Reddi, and Arijit Raychowdhury. "MulBERRY: Enabling Bit-Error Robustness for Energy-Efficient Multi-Agent Autonomous Systems." In *ASPLOS 2024*. [DOI]

2024 Víctor Mayoral Vilches, Jason Jabbour, Yu-Shun Hsiao, Zishen Wan, Martiño Crespo-Álvarez, Matthew Stewart, Juan Manuel Reina-Muñoz, Prateek Nagras, Gaurav Vikhe, Mohammad Bakhshalipour, Martin Pinzger, Stefan Rass, Smruti Panigrahi, Giulio Corradi, Niladri Roy, Phillip B. Gibbons, Sabrina M. Neuman, Brian Plancher, and Vijay Janapa Reddi. "RobotPerf: An Open-Source, Vendor-Agnostic, Benchmarking Suite for Evaluating Robotics Computing System Performance." In *ICRA 2024*. [DOI]

2024 Brian Plancher, Sebastian Büttrich, Jeremy Ellis, Neena Goveas, Laila D. Kazimierski, Jesús Alfonso López Sotelo, Milan Lukic, Diego Mendez, Rosdiadee Nordin, Andrés Oliva Trevisan, Massimo Pavan, Manuel Roveri, Marcus Rüb, Jackline Tum, Marian Verhelst, Salah Abdeljabar, Segun Adebayo, Thomas Amberg, Halleluyah Oluwatobi Aworinde, José Bagur, Gregg Barrett, Nabil Benamar, Bharat S. Chaudhari, Ronald Criollo, David Cuartielles, José A. Ferreira Filho, Solomon Gizaw, Evgeni Gousev, Alessandro Grande, Shawn Hymel, Peter Ing, Prashant Manandhar, Pietro Manzoni, Boris Murmann, Eric Pan, Rytis Paskauskas, Ermanno Pietrosemoli, Tales C. Pimenta, Marcelo Rovai, Marco Zennaro, and Vijay Janapa Reddi. "TinyML4D: Scaling Embedded Machine Learning Education in the Developing World." In *AAAI 2024*. [DOI]

2023 Srivatsan Krishnan, Amir Yazdanbakhsh, Shvetank Prakash, Jason Jabbour, Ikechukwu Uchendu, Susobhan Ghosh, Behzad Boroujerdian, Daniel Richins, Devashree Tripathy, Aleksandra Faust, and Vijay Janapa Reddi. "ArchGym: An Open-Source Gymnasium for Machine Learning Assisted Architecture Design." In *ISCA 2023*. [DOI]

2023 Vijay Janapa Reddi and Amir Yazdanbakhsh. "Architecture 2.0: Challenges and Opportunities." In *DAC 2023*. [DOI]

2023 Zishen Wan, Nandhini Chandramoorthy, Karthik Swaminathan, Pin-Yu Chen, Vijay Janapa Reddi, and Arijit Raychowdhury. "BERRY: Bit Error Robustness for Energy-Efficient Reinforcement Learning-Based Autonomous Systems." In *DAC 2023*. [DOI]

2023 Shvetank Prakash, Tim Callahan, Joseph Bushagour, Colby R. Banbury, Alan V. Green, Pete Warden, Tim Ansell, and Vijay Janapa Reddi. "CFU Playground: Full-Stack Open-Source Framework for Tiny Machine Learning (TinyML) Acceleration on FPGAs." In *ISPASS 2023*. [DOI]

2023 Shvetank Prakash, Tim Callahan, Joseph Bushagour, Colby R. Banbury, Alan V. Green, Pete Warden, Tim Ansell, and Vijay Janapa Reddi. "CFU Playground: Want a faster ML processor? Do it yourself!."

In *DATE 2023*. [DOI]

2023 Mark Mazumder, Colby R. Banbury, Xiaozhe Yao, Bojan Karlas, William Gaviria Rojas, Sudnya Frederick Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Will Cukierski, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Raje, Max Bartolo, Evan Sabri Eyuboglu, Amirata Ghorbani, Emmett D. Goodman, Addison Howard, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, D. Sculley, Tzu-Sheng Kuo, Jonas W. Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen K. Paritosh, Ce Zhang, James Y. Zou, Carole-Jean Wu, Cody Coleman, Andrew Y. Ng, Peter Mattson, and Vijay Janapa Reddi. "DataPerf: Benchmarks for Data-Centric AI Development." In *NeurIPS 2023*. [Link]

2023 Colby R. Banbury, Vijay Janapa Reddi, Alexander Elium, Shawn Hymel, David Tischler, Daniel Situnayake, Carl Ward, Louis Moreau, Jenny Plunkett, Matthew Kelcey, Mathijs Baaijens, Alessandro Grande, Dmitry Maslov, Arthur Beavis, Jan Jongboom, and Jessica Quaye. "Edge Impulse: An MLOps Platform for Tiny Machine Learning." In *MLSys 2023*. [Link]

2023 Yu-Shun Hsiao, Zishen Wan, Tianyu Jia, Radhika Ghosal, Abdulrahman Mahmoud, Arijit Raychowdhury, David Brooks, Gu-Yeon Wei, and Vijay Janapa Reddi. "MAVFI: An End-to-End Fault Analysis Framework with Anomaly Detection and Recovery for Micro Aerial Vehicles." In *DATE 2023*. [DOI]

2023 Sabrina M. Neuman, Radhika Ghosal, Thomas Bourgeat, Brian Plancher, and Vijay Janapa Reddi. "RoboShape: Using Topology Patterns to Scalably and Flexibly Deploy Accelerators Across Robots." In *ISCA 2023*. [DOI]

2023 Thanh Thi Nguyen, Minh Cuong Nguyen, Thien Huynh-The, Quoc-Viet Pham, Quoc Viet Hung Nguyen, Imran Razzak, and Vijay Janapa Reddi. "Solving Complex Sequential Decision-Making Problems by Deep Reinforcement Learning with Heuristic Rules." In *ICCS 2023*. [DOI]

2023 Yu-Shun Hsiao, Siva Kumar Sastry Hari, Balakumar Sundaralingam, Jason Yik, Thierry Tambe, Charbel Sakr, Stephen W. Keckler, and Vijay Janapa Reddi. "VaPr: Variable-Precision Tensors to Accelerate Robot Motion Planning." In *IROS 2023*. [DOI]

2023 Hyoukjun Kwon, Krishnakumar Nair, Jamin Seo, Jason Yik, Debabrata Mohapatra, Dongyuan Zhan, Jinook Song, Peter Capak, Peizhao Zhang, Peter Vajda, Colby R. Banbury, Mark Mazumder, Liangzhen Lai, Ashish Sirasao, Tushar Krishna, Harshit Khaitan, Vikas Chandra, and Vijay Janapa Reddi. "XRBench: An Extended Reality (XR) Machine Learning Benchmark Suite for the Metaverse." In *MLSys 2023*. [Link]

2022 Srivatsan Krishnan, Zishen Wan, Kshitij Bhardwaj, Paul N. Whatmough, Aleksandra Faust, Sabrina M. Neuman, Gu-Yeon Wei, David Brooks, and Vijay Janapa Reddi. "Automatic Domain-Specific SoC Design for Autonomous Unmanned Aerial Vehicles." In *MICRO 2022*. [DOI]

2022 Zishen Wan, Aqeel Anwar, Abdulrahman Mahmoud, Tianyu Jia, Yu-Shun Hsiao, Vijay Janapa Reddi, and Arijit Raychowdhury. "FRL-FI: Transient Fault Analysis for Federated Reinforcement Learning-Based Navigation Systems." In *DATE 2022*. [DOI]

2022 Brian Plancher, Sabrina M. Neuman, Radhika Ghosal, Scott Kuindersma, and Vijay Janapa Reddi. "GRiD: GPU-Accelerated Rigid Body Dynamics with Analytical Gradients." In *ICRA 2022*. [DOI]

2022 Vijay Janapa Reddi, David Kanter, Peter Mattson, Jared Duke, Thai Nguyen, Ramesh Chukka,

Kenneth Shiring, Koan-Sin Tan, Mark Charlebois, William Chou, Mostafa El-Khamy, Jungwook Hong, Tom St. John, Cindy Trinh, Michael Buch, Mark Mazumder, Relja Markovic, Thomas Atta-Fosu, Fatih Çakir, Masoud Charkhabi, Xiaodong Chen, Cheng-Ming Chiang, Dave Dexter, Terry Heo, Guenther Schmuelling, Maryam Shabani, and Dylan Zika. "MLPerf Mobile Inference Benchmark: An Industry-Standard Open-Source Machine Learning Benchmark for On-Device AI." In *MLSys 2022*. [Link]

2022 Tianyu Jia, En-Yu Yang, Yu-Shun Hsiao, Jonathan J. Cruz, David Brooks, Gu-Yeon Wei, and Vijay Janapa Reddi. "OMU: A Probabilistic 3D Occupancy Mapping Accelerator for Real-time OctoMap at the Edge." In *DATE 2022*. [DOI]

2022 Víctor Mayoral Vilches, Sabrina M. Neuman, Brian Plancher, and Vijay Janapa Reddi. "RobotCore: An Open Architecture for Hardware Acceleration in ROS 2." In *IROS 2022*. [DOI]

2022 Zishen Wan, Ashwin Sanjay Lele, Bo Yu, Shaoshan Liu, Yu Wang, Vijay Janapa Reddi, Cong Hao, and Arijit Raychowdhury. "Robotic Computing on FPGAs: Current Progress, Research Challenges, and Opportunities." In *AICAS 2022*. [DOI]

2022 Srivatsan Krishnan, Zishen Wan, Kshitij Bhardwaj, Ninad Jadhav, Aleksandra Faust, and Vijay Janapa Reddi. "Roofline Model for UAVs: A Bottleneck Analysis Tool for Onboard Compute Characterization of Autonomous Unmanned Aerial Vehicles." In *ISPASS 2022*. [DOI]

2022 William A. Gaviria Rojas, Sudnya Frederick Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. "The Dollar Street Dataset: Images Representing the Geographic and Socioeconomic Diversity of the World." In *NeurIPS 2022*. [Link]

2022 Sabrina M. Neuman, Brian Plancher, Bardienus Pieter Duisterhof, Srivatsan Krishnan, Colby R. Banbury, Mark Mazumder, Shvetank Prakash, Jason Jabbour, Aleksandra Faust, Guido C. H. E. de Croon, and Vijay Janapa Reddi. "Tiny Robot Learning: Challenges and Directions for Machine Learning in Resource-Constrained Robots." In *AICAS 2022*. [DOI]

2022 Brian Plancher and Vijay Janapa Reddi. "TinyMLedu: The Tiny Machine Learning Open Education Initiative." In *SIGCSE 2022*. [DOI]

2022 Yu-Shun Hsiao, Siva Kumar Sastry Hari, Michal Filipiuk, Timothy Tsai, Michael B. Sullivan, Vijay Janapa Reddi, Vasu Singh, and Stephen W. Keckler. "Zhuyi: perception processing rate estimation for safety in autonomous vehicles." In *DAC 2022*. [DOI]

2021 Michael Buch, Zahra Azad, Ajay Joshi, and Vijay Janapa Reddi. "AI Tax in Mobile SoCs: End-to-end Performance Analysis of Machine Learning in Smartphones." In *ISPASS 2021*. [DOI]

2021 Zishen Wan, Aqeel Anwar, Yu-Shun Hsiao, Tianyu Jia, Vijay Janapa Reddi, and Arijit Raychowdhury. "Analyzing and Improving Fault Tolerance of Learning-Based Navigation Systems." In *DAC 2021*. [DOI]

2021 Mark Mazumder, Colby R. Banbury, Josh Meyer, Pete Warden, and Vijay Janapa Reddi. "Few-Shot Keyword Spotting in Any Language." In *22nd Annual Conference of the International Speech Communica 2021*. [DOI]

2021 Maximilian Lam, Gu-Yeon Wei, David Brooks, Vijay Janapa Reddi, and Michael Mitzenmacher. "Gradient Disaggregation: Breaking Privacy in Federated Learning by Reconstructing the User Participant Matrix." In *ICML 2021*. [Link]

2021 Colby R. Banbury, Vijay Janapa Reddi, Peter Torelli, Nat Jeffries, Csaba Király, Jeremy Holleman, Pietro Montino, David Kanter, Pete Warden, Danilo Pau, Urmish Thakker, Antonio Torrini, Jay Cordaro, Giuseppe Di Guglielmo, Javier M. Duarte, Honson Tran, Nhan Tran, Wenxu Niu, and Xuesong Xu. "MLPerf Tiny Benchmark." In *Proceedings of the Neural Information Processing Systems Tra 2021*. [Link]

2021 Colby R. Banbury, Chuteng Zhou, Igor Fedorov, Ramon Matas Navarro, Urmish Thakker, Dibakar Gope, Vijay Janapa Reddi, Matthew Mattina, and Paul N. Whatmough. "MicroNets: Neural Network Architectures for Deploying TinyML Applications on Commodity Microcontrollers." In *MLSys 2021*. [Link]

2021 Mark Mazumder, Sharad Chitlangia, Colby R. Banbury, Yiping Kang, Juan Ciro, Keith Achorn, Daniel Galvez, Mark Sabini, Peter Mattson, David Kanter, Greg Diamos, Pete Warden, Josh Meyer, and Vijay Janapa Reddi. "Multilingual Spoken Words Corpus." In *Proceedings of the Neural Information Processing Systems Tra 2021*. [Link]

2021 Maximilian Lam, Zachary Yedidia, Colby R. Banbury, and Vijay Janapa Reddi. "Precision Batching: Bitserial Decomposition for Efficient Neural Network Inference on GPUs." In *PACT 2021*. [DOI]

2021 James Gleeson, Moshe Gabel, Gennady Pekhimenko, Eyal de Lara, Srivatsan Krishnan, and Vijay Janapa Reddi. "RL-Scope: Cross-stack Profiling for Deep Reinforcement Learning Workloads." In *MLSys 2021*. [Link]

2021 Behzad Boroujerdian, Radhika Ghosal, Jonathan J. Cruz, Brian Plancher, and Vijay Janapa Reddi. "RoboRun: A Robot Runtime to Exploit Spatial Heterogeneity." In *DAC 2021*. [DOI]

2021 Sabrina M. Neuman, Brian Plancher, Thomas Bourgeat, Thierry Tambe, Srinivas Devadas, and Vijay Janapa Reddi. "Robomorphic computing: a design methodology for domain-specific accelerators parameterized by robot morphology." In *ASPLOS 2021*. [DOI]

2021 Bardienus Pieter Duisterhof, Shushuai Li, Javier Burgués, Vijay Janapa Reddi, and Guido C. H. E. de Croon. "Sniffy Bug: A Fully Autonomous Swarm of Gas-Seeking Nano Quadcopters in Cluttered Environments." In *IROS 2021*. [DOI]

2021 Robert David, Jared Duke, Advait Jain, Vijay Janapa Reddi, Nat Jeffries, Jian Li, Nick Kreeger, Ian Nappier, Meghna Natraj, Tiezhen Wang, Pete Warden, and Rocky Rhodes. "TensorFlow Lite Micro: Embedded Machine Learning for TinyML Systems." In *MLSys 2021*. [Link]

2021 Daniel Galvez, Greg Diamos, Juan Torres, Keith Achorn, Juan Felipe Cerón, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. "The People's Speech: A Large-Scale Diverse English Speech Recognition Dataset for Commercial Usage." In *Proceedings of the Neural Information Processing Systems Tra 2021*. [Link]

2021 Bardienus Pieter Duisterhof, Srivatsan Krishnan, Jonathan J. Cruz, Colby R. Banbury, William Fu, Aleksandra Faust, Guido C. H. E. de Croon, and Vijay Janapa Reddi. "Tiny Robot Learning (tinyRL) for Source Seeking on a Nano Quadcopter." In *ICRA 2021*. [DOI]

2021 Muhammad Shafique, Theocharis Theocharides, Vijay Janapa Reddi, and Boris Murmann. "TinyML: Current Progress, Research Challenges, and Future Roadmap." In *DAC 2021*. [DOI]

2020 Thierry Tambe, En-Yu Yang, Zishen Wan, Yuntian Deng, Vijay Janapa Reddi, Alexander M. Rush, David Brooks, and Gu-Yeon Wei. "Algorithm-Hardware Co-Design of Adaptive Floating-Point

Encodings for Resilient Deep Learning Inference." In *DAC 2020*. [DOI]

2020 Jingwen Leng, Alper Buyuktosunoglu, Ramon Bertran, Pradip Bose, Quan Chen, Minyi Guo, and Vijay Janapa Reddi. "Asymmetric Resilience: Exploiting Task-Level Idempotency for Transient Error Recovery in Accelerator-Based Systems." In *HPCA 2020*. [DOI]

2020 Vijay Janapa Reddi, Christine Cheng, David Kanter, Peter Mattson, Guenther Schmuelling, Carole-Jean Wu, Brian Anderson, Maximilien Breughe, Mark Charlebois, William Chou, Ramesh Chukka, Cody Coleman, Sam Davis, Pan Deng, Greg Diamos, Jared Duke, Dave Fick, J. Scott Gardner, Itay Hubara, Sachin Idgunji, Thomas B. Jablin, Jeff Jiao, Tom St. John, Pankaj Kanwar, David Lee, Jeffery Liao, Anton Lokhmotov, Francisco Massa, Peng Meng, Paulius Micikevicius, Colin Osborne, Gennady Pekhimenko, Arun Tejusve Raghunath Rajan, Dilip Sequeira, Ashish Sirasao, Fei Sun, Hanlin Tang, Michael Thomson, Frank Wei, Ephrem Wu, Lingjie Xu, Koichi Yamada, Bing Yu, George Yuan, Aaron Zhong, Peizhao Zhang, and Yuchen Zhou. "MLPerf Inference Benchmark." In *ISCA 2020*. [DOI]

2020 Peter Mattson, Christine Cheng, Gregory F. Diamos, Cody Coleman, Paulius Micikevicius, David A. Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debo Dutta, Udit Gupta, Kim M. Hazelwood, Andy Hock, Xinyuan Huang, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St. John, Carole-Jean Wu, Lingjie Xu, Cliff Young, and Matei Zaharia. "MLPerf Training Benchmark." In *MLSys 2020*. [Link]

2020 Daniel Richins, Dharmisha Doshi, Matthew Blackmore, Aswathy Thulaseedharan Nair, Neha Pathapati, Ankit Patel, Brainard Daguman, Daniel Dobrijalowski, Ramesh Illikkal, Kevin Long, David Zimmerman, and Vijay Janapa Reddi. "Missing the Forest for the Trees: End-to-End AI Application Performance in Edge Data Centers." In *HPCA 2020*. [DOI]

2019 Yazhou Zu, Daniel Richins, Charles Lefurgy, and Vijay Janapa Reddi. "Fine-Tuning the Active Timing Margin (ATM) Control Loop for Maximizing Multi-core Efficiency on an IBM POWER Server." In *HPCA 2019*. [DOI]

2019 Mark D. Hill and Vijay Janapa Reddi. "Gables: A Roofline Model for Mobile SoCs." In *HPCA 2019*. [DOI]

2019 Dimitris Gizopoulos, George Papadimitriou, Athanasios Chatzidimitriou, Vijay Janapa Reddi, Behzad Salami, Osman S. Unsal, Adrián Cristal Kestelman, and Jingwen Leng. "Modern Hardware Margins: CPUs, GPUs, FPGAs Recent System-Level Studies." In *IOLTS 2019*. [DOI]

2019 Matthew Halpern, Behzad Boroujerdian, Todd W. Mummert, Evelyn Duesterwald, and Vijay Janapa Reddi. "One Size Does Not Fit All: Quantifying and Exposing the Accuracy-Latency Trade-Off in Machine Learning Cloud Service APIs via Tolerance Tiers." In *ISPASS 2019*. [DOI]

2019 Wenzhi Cui, Daniel Richins, Yuhao Zhu, and Vijay Janapa Reddi. "Tail latency in node.js: energy efficient turbo boosting for long latency requests in event-driven web services." In *VEE 2019*. [DOI]

2018 Daniel Richins, Tahrina Ahmed, Russell M. Clapp, and Vijay Janapa Reddi. "Amdahl's Law in Big Data Analytics: Alive and Kicking in TPCx-BB (BigBench)." In *HPCA 2018*. [DOI]

2018 An Zou, Jingwen Leng, Xin He, Yazhou Zu, Vijay Janapa Reddi, and Xuan Zhang. "Efficient and reliable power delivery in voltage-stacked manycore system with hybrid charge-recycling

regulators." In *DAC 2018*. [DOI]

2018 Behzad Boroujerdian, Hasan Genc, Srivatsan Krishnan, Wenzhi Cui, Aleksandra Faust, and Vijay Janapa Reddi. "MAVBench: Micro Aerial Vehicle Benchmarking." In *MICRO 2018*. [DOI]

2018 An Zou, Jingwen Leng, Xin He, Yazhou Zu, Christopher D. Gill, Vijay Janapa Reddi, and Xuan Zhang. "Voltage-Stacked GPUs: A Control Theory Driven Cross-Layer Solution for Practical Voltage Stacking in GPUs." In *MICRO 2018*. [DOI]

2017 An Zou, Jingwen Leng, Yazhou Zu, Tao Tong, Vijay Janapa Reddi, David M. Brooks, Gu-Yeon Wei, and Xuan Zhang. "Ivory: Early-Stage Design Space Exploration Tool for Integrated Voltage Regulators." In *DAC 2017*. [DOI]

2016 Yuxi Liu, Zhibin Yu, Lieven Eeckhout, Vijay Janapa Reddi, Yingwei Luo, Xiaolin Wang, Zhenlin Wang, and Cheng-Zhong Xu. "Barrier-Aware Warp Scheduling for Throughput Processors." In *ICS 2016*. [DOI]

2016 Yuhao Zhu and Vijay Janapa Reddi. "GreenWeb: language extensions for energy-efficient mobile web computing." In *PLDI 2016*. [DOI]

2016 Matthew Halpern, Yuhao Zhu, and Vijay Janapa Reddi. "Mobile CPU's rise to power: Quantifying the impact of generational mobile CPU design trends on performance, energy, and user satisfaction." In *HPCA 2016*. [DOI]

2016 Mikhail Kazdagli, Vijay Janapa Reddi, and Mohit Tiwari. "Quantifying and improving the efficiency of hardware-based mobile malware detectors." In *MICRO 2016*. [DOI]

2016 Nadav Chachmon, Daniel Richins, Robert S. Cohn, Magnus Christensson, Wenzhi Cui, and Vijay Janapa Reddi. "Simulation and Analysis Engine for Scale-Out Workloads." In *ICS 2016*. [DOI]

2016 Yazhou Zu, Wei Huang, Indrani Paul, and Vijay Janapa Reddi. "Ti-states: Processor power management in the temperature inversion region." In *MICRO 2016*. [DOI]

2015 Yazhou Zu, Charles R. Lefurgy, Jingwen Leng, Matthew Halpern, Michael S. Floyd, and Vijay Janapa Reddi. "Adaptive guardband scheduling to improve system-level efficiency of the POWER7+." In *MICRO 2015*. [DOI]

2015 Yuhao Zhu, Matthew Halpern, and Vijay Janapa Reddi. "Event-based scheduling for energy-efficient QoS (eQoS) in mobile Web applications." In *HPCA 2015*. [DOI]

2015 Jingwen Leng, Yazhou Zu, and Vijay Janapa Reddi. "GPU voltage noise: Characterization and hierarchical smoothing of spatial and temporal voltage noise interference in GPU architectures." In *HPCA 2015*. [DOI]

2015 Yuhao Zhu, Daniel Richins, Matthew Halpern, and Vijay Janapa Reddi. "Microarchitectural implications of event-driven server-side web applications." In *MICRO 2015*. [DOI]

2015 Matthew Halpern, Yuhao Zhu, Ramesh Peri, and Vijay Janapa Reddi. "Mosaic: cross-platform user-interaction record and replay for the fragmented android ecosystem." In *ISPASS 2015*. [DOI]

2015 Jingwen Leng, Alper Buyuktosunoglu, Ramon Bertran, Pradip Bose, and Vijay Janapa Reddi. "Safe limits on voltage reduction efficiency in GPUs: a direct measurement approach." In *MICRO 2015*. [DOI]

2014 Jingwen Leng, Yazhou Zu, Minsoo Rhu, Meeta Sharma Gupta, and Vijay Janapa Reddi. "GPUVolt: modeling and characterizing voltage noise in GPU architectures." In *ISLPED 2014*. [DOI]

2014 Mikhail Kazdagli, Ling Huang, Vijay Janapa Reddi, and Mohit Tiwari. "Morpheus: benchmarking computational diversity in mobile malware." In *HASP 2014*. [DOI]

2014 Yuhao Zhu and Vijay Janapa Reddi. "WebCore: Architectural support for mobile Web browsing." In *ISCA 2014*. [DOI]

2013 Jingwen Leng, Tayler H. Hetherington, Ahmed ElTantawy, Syed Zohaib Gilani, Nam Sung Kim, Tor M. Aamodt, and Vijay Janapa Reddi. "GPUWattch: enabling energy optimizations in GPGPUs." In *ISCA 2013*. [DOI]

2013 Yuhao Zhu and Vijay Janapa Reddi. "High-performance and energy-efficient mobile web browsing on big/little systems." In *HPCA 2013*. [DOI]

2012 Simone Campanoni, Timothy M. Jones, Glenn H. Holloway, Vijay Janapa Reddi, Gu-Yeon Wei, and David M. Brooks. "HELIX: automatic parallelization of irregular programs for chip multiprocessing." In *CGO 2012*. [DOI]

2012 Vijay Janapa Reddi. "Hardware and software co-design for robust and resilient execution." In *CTS 2012*. [DOI]

2012 Vijay Janapa Reddi, David Z. Pan, Sani R. Nassif, and Keith A. Bowman. "Robust and resilient designs from the bottom-up: Technology, CAD, circuit, and system issues." In *ASP-DAC 2012*. [DOI]

2011 Peter Bailis, Vijay Janapa Reddi, Sanjay Gandhi, David M. Brooks, and Margo I. Seltzer. "Dimetrodon: processor-level preventive thermal management via idle cycle injection." In *DAC 2011*. [DOI]

2010 Vijay Janapa Reddi, Svilen Kanev, Wonyoung Kim, Simone Campanoni, Michael D. Smith, Gu-Yeon Wei, and David M. Brooks. "Voltage Smoothing: Characterizing and Mitigating Voltage Noise in Production Processors via Software-Guided Thread Scheduling." In *MICRO 2010*. [DOI]

2010 Vijay Janapa Reddi, Benjamin C. Lee, Trishul M. Chilimbi, and Kushagra Vaid. "Web search using mobile cores: quantifying and mitigating the price of efficiency." In *ISCA 2010*. [DOI]

2009 Meeta Sharma Gupta, Vijay Janapa Reddi, Glenn H. Holloway, Gu-Yeon Wei, and David M. Brooks. "An event-guided approach to reducing voltage noise in processors." In *DATE 2009*. [DOI]

2009 Vijay Janapa Reddi, Simone Campanoni, Meeta Sharma Gupta, Michael D. Smith, Gu-Yeon Wei, and David M. Brooks. "Software-assisted hardware reliability: abstracting circuit-level challenges to the software stack." In *DAC 2009*. [DOI]

2009 Vijay Janapa Reddi, Meeta Sharma Gupta, Glenn H. Holloway, Gu-Yeon Wei, Michael D. Smith, and David M. Brooks. "Voltage emergency prediction: Using signatures to reduce operating margins." In *HPCA-15 2009*. [DOI]

2007 Vijay Janapa Reddi, Dan Connors, Robert Cohn, and Michael D. Smith. "Persistent Code Caching: Exploiting Code Reuse Across Executions and Applications." In *CGO 2007*. [DOI]

2007 Tipp Moseley, Alex Shye, Vijay Janapa Reddi, Dirk Grunwald, and Ramesh Peri. "Shadow Profiling: Hiding Instrumentation Costs with Parallelism." In *CGO 2007*. [DOI]

2007 Alex Shye, Tipp Moseley, Vijay Janapa Reddi, Joseph Blomstedt, and Daniel A. Connors. "Using Process-Level Redundancy to Exploit Multiple Cores for Transient Fault Tolerance." In *DSN 2007*. [DOI]

2005 Qiang Wu, Margaret Martonosi, Douglas W. Clark, Vijay Janapa Reddi, Dan Connors, Youfeng Wu, Jin Lee, and David M. Brooks. "A Dynamic Compilation Framework for Controlling Microprocessor Energy and Performance." In *MICRO-38 2005*. [DOI]

2005 Tipp Moseley, Alex Shye, Vijay Janapa Reddi, Matthew Iyer, Dan Fay, David Hodgdon, Joshua L. Kihm, Alex Settle, Dirk Grunwald, and Daniel A. Connors. "Dynamic run-time architecture techniques for enabling continuous optimization." In *Proceedings of the Second Conference on Computing Frontiers 2005*. [DOI]

2005 Chi-Keung Luk, Robert S. Cohn, Robert Muth, Harish Patil, Artur Klauser, P. Geoffrey Lowney, Steven Wallace, Vijay Janapa Reddi, and Kim M. Hazelwood. "Pin: building customized program analysis tools with dynamic instrumentation." In *SIGPLAN 2005*. [DOI]

2005 Silvia M. Figueira and Vijay Janapa Reddi. "Topology-Based Hypercube Structures for Global Communication in Heterogeneous Networks." In *Euro-Par 2005*. [DOI]

Journal Publications

2025 Vijay Janapa Reddi and Amir Yazdanbakhsh. "Architecture 2.0: Foundations of Artificial Intelligence Agents for Modern Computer System Design." In *Computer, 2025*. [DOI]

2025 Vijay Janapa Reddi. "Generative AI at the Edge: Challenges and Opportunities: The next phase in AI deployment." In *ACM Queue, 2025*. [DOI]

2025 Mona Sloane, Emanuel Moss, Susan Kennedy, Matthew Stewart, Pete Warden, Brian Plancher, and Vijay Janapa Reddi. "Materiality and risk in the age of pervasive AI sensors." In *Nat. Mac. Intell., 2025*. [DOI]

2025 Shvetank Prakash, Andrew Cheng, Jason Yik, Arya Tschand, Radhika Ghosal, Ikechukwu Uchendu, Jessica Quaye, Jeffrey Jian Ma, Shreyas Grampurohit, Sofia Giannuzzi, Arnav Balyan, Fin Amin, Aadya Pipersenia, Yash Choudhary, Ankita Nayak, Amir Yazdanbakhsh, and Vijay Janapa Reddi. "QuArch: A Question-Answering Dataset for AI Agents in Computer Architecture." In *IEEE Comput. Archit. Lett., 2025*. [DOI]

2025 Yanjing Li and Vijay Janapa Reddi. "Robust Methods for Deep-Learning Training." In *IEEE Des. Test, 2025*. [DOI]

2024 Emre Ozer, Jedrzej Kufel, Shvetank Prakash, Alireza Raisiardali, Olof Kindgren, Ronald Wong, Nelson Ng, Damien Jausseran, Feras Alkhalil, David Kong, Gage Hills, Richard Price, and Vijay Janapa Reddi. "Bendable non-silicon RISC-V microprocessor." In *Nat., 2024*. [DOI]

2024 Luis Oala, Manil Maskey, Lilith Bat-Leah, Alicia Parrish, Nezihe Merve Gürel, Tzu-Sheng Kuo, Yang Liu, Rotem Dror, Danilo Brajovic, Xiaozhe Yao, Max Bartolo, William Gaviria Rojas, Ryan Hileman, Rainier Aliment, Michael W. Mahoney, Meg Risdal, Matthew Lease, Wojciech Samek, Debojyoti Dutta, Curtis G. Northcutt, Cody Coleman, Braden Hancock, Bernard Koch, Girmaw Abebe Tadesse, Bojan Karlas, Ahmed Alaa, Adji Bousso Dieng, Natasha F. Noy, Vijay Janapa Reddi, James Zou, Praveen K. Paritosh, Mihaela van der Schaar, Kurt Bollacker, Lora Aroyo, Ce Zhang, Joaquin Vanschoren, Isabelle Guyon, and Peter Mattson. "DMLR: Data-centric Machine Learning

Research - Past, Present and Future." In *J. Data-centric Mach. Learn. Res., 2024*. [Link]

2024 Yu-Shun Hsiao, Zishen Wan, Tianyu Jia, Radhika Ghosal, Abdulrahman Mahmoud, Arijit Raychowdhury, David Brooks, Gu-Yeon Wei, and Vijay Janapa Reddi. "Silent Data Corruption in Robot Operating System: A Case for End-to-End System-Level Fault Analysis Using Autonomous UAVs." In *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 2024*. [DOI]

2024 Max Lam, Michael Mitzenmacher, Vijay Janapa Reddi, Gu-Yeon Wei, and David Brooks. "Tabula: Efficiently Computing Nonlinear Activation Functions for Secure Neural Network Inference." In *Trans. Mach. Learn. Res., 2024*. [Link]

2023 Cansu Demirkiran, Furkan Eris, Gongyu Wang, Jonathan Elmhurst, Nick Moore, Nicholas C. Harris, Ayon Basumallik, Vijay Janapa Reddi, Ajay Joshi, and Darius Bunandar. "An Electro-Photonic System for Accelerating Deep Neural Networks." In *ACM J. Emerg. Technol. Comput. Syst., 2023*. [DOI]

2023 Thanh Thi Nguyen and Vijay Janapa Reddi. "Deep Reinforcement Learning for Cyber Security." In *IEEE Trans. Neural Networks Learn. Syst., 2023*. [DOI]

2023 Behzad Boroujerdian, Ying Jing, Devashree Tripathy, Amit Kumar, Lavanya Subramanian, Luke Yen, Vincent Lee, Vivek Venkatesan, Amit Jindal, Robert Shearer, and Vijay Janapa Reddi. "FARSI: An Early-stage Design Space Exploration Framework to Tame the Domain-specific System-on-chip Complexity." In *ACM Trans. Embed. Comput. Syst., 2023*. [DOI]

2023 Alexandros Karargyris, Renato Umeton, Micah J. Sheller, Alejandro Aristizabal, Johnu George, Anna Wuest, Sarthak Pati, Hasan Kassem, Maximilian Zenk, Ujjwal Baid, Prakash Narayana Moorthy, Alexander Chowdhury, Junyi Guo, Sahil S. Nalawade, Jacob Rosenthal, David Kanter, Maria Xenochristou, Daniel J. Beutel, Verena Chung, Timothy Bergquist, James A. Eddy, Abubakar Abid, Lewis Tunstall, Omar Sanseviero, Dimitrios Dimitriadis, Yiming Qian, Xinxing Xu, Yong Liu, Rick Siow Mong Goh, Srini Bala, Victor Bittorf, Sreekar Reddy Puchala, Biagio Ricciuti, Soujanya Samineni, Eshna Sengupta, Akshay Chaudhari, Cody Coleman, Bala Desinghu, Gregory F. Diamos, Debo Dutta, Diane Feddema, Grigori Fursin, Xinyuan Huang, Satyananda Kashyap, Nicholas D. Lane, Indranil Mallick, Pietro Mascagni, Virendra Mehta, Cassiano Ferro Moraes, Vivek Natarajan, Nikola Nikolov, Nicolas Padoy, Gennady Pekhimenko, Vijay Janapa Reddi, G. Anthony Reina, Pablo Ribalta, Abhishek Singh, Jayaraman J. Thiagarajan, Jacob Albrecht, Thomas Wolf, Geralyn Miller, Huazhu Fu, Prashant Shah, Daguang Xu, Poonam Yadav, David Talby, Mark M. Awad, Jeremy P. Howard, Michael Rosenthal, Luigi Marchionni, Massimo Loda, Jason M. Johnson, Spyridon Bakas, and Peter Mattson. "Federated benchmarking of medical artificial intelligence with MedPerf." In *Nat. Mac. Intell., 2023*. [DOI]

2023 Shvetank Prakash, Matthew Stewart, Colby R. Banbury, Mark Mazumder, Pete Warden, Brian Plancher, and Vijay Janapa Reddi. "Is TinyML Sustainable?." In *Commun. ACM, 2023*. [DOI]

2023 Pete Warden, Matthew Stewart, Brian Plancher, Sachin Katti, and Vijay Janapa Reddi. "Machine Learning Sensors." In *Commun. ACM, 2023*. [DOI]

2023 Vijay Janapa Reddi and Boris Murmann. "Special Issue on TinyML." In *IEEE Micro, 2023*. [DOI]

2022 Sathwika Bavikadi, Abhijitt Dhavlle, Amlan Ganguly, Anand Haridass, Hagar Hendy, Cory E. Merkel, Vijay Janapa Reddi, Purab Ranjan Sutradhar, Arun Joseph, and Sai Manoj Pudukotai Dinakarrao. "A Survey on Machine Learning Accelerators and Evolutionary Hardware Platforms." In *IEEE Des. Test, 2022*. [DOI]

2022 Srivatsan Krishnan, Max Lam, Sharad Chitlangia, Zishen Wan, Gabriel Barth-Maron, Aleksandra Faust, and Vijay Janapa Reddi. "QuaRL: Quantization for Fast and Environmentally Sustainable Reinforcement Learning." In *Trans. Mach. Learn. Res., 2022*. [Link]

2021 Brian Plancher, Sabrina M. Neuman, Thomas Bourgeat, Scott Kuindersma, Srinivas Devadas, and Vijay Janapa Reddi. "Accelerating Robot Dynamics Gradients on a CPU, GPU, and FPGA." In *IEEE Robotics Autom. Lett., 2021*. [DOI]

2021 Mark D. Hill and Vijay Janapa Reddi. "Accelerator-level parallelism." In *Commun. ACM, 2021*. [DOI]

2021 Srivatsan Krishnan, Behzad Boroujerdian, William Fu, Aleksandra Faust, and Vijay Janapa Reddi. "Air Learning: a deep reinforcement learning gym for autonomous aerial robot visual navigation." In *Mach. Learn., 2021*. [DOI]

2021 Jingwen Leng, Alper Buyuktosunoglu, Ramon Bertran, Pradip Bose, Yazhou Zu, and Vijay Janapa Reddi. "Erratum to "Predictive Guardbanding: Program-Driven Timing Margin Reduction for GPUs"." In *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 2021*. [DOI]

2021 Jingwen Leng, Alper Buyuktosunoglu, Ramon Bertran, Pradip Bose, Yazhou Zu, and Vijay Janapa Reddi. "Predictive Guardbanding: Program-Driven Timing Margin Reduction for GPUs." In *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 2021*. [DOI]

2021 An Zou, Huifeng Zhu, Jingwen Leng, Xin He, Vijay Janapa Reddi, Christopher D. Gill, and Xuan Zhang. "System-level Early-stage Modeling and Evaluation of IVR-assisted Processor Power Delivery System." In *ACM Trans. Archit. Code Optim., 2021*. [DOI]

2021 Behzad Boroujerdian, Hasan Genc, Srivatsan Krishnan, Bardienus Pieter Duisterhof, Brian Plancher, Kayvan Mansoorshahi, Marcelino Almeida, Wenzhi Cui, Aleksandra Faust, and Vijay Janapa Reddi. "The Role of Compute in Autonomous Micro Aerial Vehicles: Optimizing for Mission Time and Energy Efficiency." In *ACM Trans. Comput. Syst., 2021*. [DOI]

2021 Vijay Janapa Reddi, Christine Cheng, David Kanter, Peter Mattson, Guenther Schmuelling, and Carole-Jean Wu. "The Vision Behind MLPerf: Understanding AI Inference Performance." In *IEEE Micro, 2021*. [DOI]

2020 Peter Mattson, Hanlin Tang, Gu-Yeon Wei, Carole-Jean Wu, Vijay Janapa Reddi, Christine Cheng, Cody Coleman, Greg Diamos, David Kanter, Paulius Micikevicius, David A. Patterson, and Guenther Schmuelling. "MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance." In *IEEE Micro, 2020*. [DOI]

2020 Srivatsan Krishnan, Zishen Wan, Kshitij Bhardwaj, Paul N. Whatmough, Aleksandra Faust, Gu-Yeon Wei, David Brooks, and Vijay Janapa Reddi. "The Sky Is Not the Limit: A Visual Performance Model for Cyber-Physical Co-Design in Autonomous Machines." In *IEEE Comput. Archit. Lett., 2020*. [DOI]

2020 An Zou, Jingwen Leng, Xin He, Yazhou Zu, Christopher D. Gill, Vijay Janapa Reddi, and Xuan Zhang. "Voltage-Stacked Power Delivery Systems: Reliability, Efficiency, and Power Management." In *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 2020*. [DOI]

2019 Daniel Richins, Dharmisha Doshi, Matthew Blackmore, Aswathy Thulaseedharan Nair, Neha Pathapati, Ankit Patel, Brainard Daguman, Daniel Dobrijalowski, Ramesh Illikkal, Kevin Long, David

Zimmerman, and Vijay Janapa Reddi. "AI Tax: The Hidden Cost of AI Data Center Applications." In *ACM Trans. Comput. Syst., 2019*. [DOI]

2019 Jingwen Leng, Alper Buyuktosunoglu, Ramon Bertran, Pradip Bose, and Vijay Janapa Reddi. "Asymmetric Resilience for Accelerator-Rich Systems." In *IEEE Comput. Archit. Lett., 2019*. [DOI]

2018 Ting-Wu Chin, Chia-Lin Yu, Matthew Halpern, Hasan Genc, Shiao-Li Tsao, and Vijay Janapa Reddi. "Domain-Specific Approximation for Object Detection." In *IEEE Micro, 2018*. [DOI]

2018 Vijay Janapa Reddi, Hongil Yoon, and Allan Knies. "Two Billion Devices and Counting." In *IEEE Micro, 2018*. [DOI]

2017 Yuhao Zhu, Vijay Janapa Reddi, Robert Adolf, Saketh Rama, Brandon Reagen, Gu-Yeon Wei, and David M. Brooks. "Cognitive Computing Safety: The New Horizon for Reliability / The Design and Evolution of Deep Learning Workloads." In *IEEE Micro, 2017*. [DOI]

2017 Hasan Genc, Yazhou Zu, Ting-Wu Chin, Matthew Halpern, and Vijay Janapa Reddi. "Flying IoT: Toward Low-Power Vision in the Sky." In *IEEE Micro, 2017*. [DOI]

2017 Yuhao Zhu and Vijay Janapa Reddi. "Optimizing General-Purpose CPUs for Energy-Efficient Mobile Web Computing." In *ACM Trans. Comput. Syst., 2017*. [DOI]

2017 Peter Bailis, Jean Yang, Vijay Janapa Reddi, and Yuhao Zhu. "Research for practice: web security and mobile web computing." In *Commun. ACM, 2017*. [DOI]

2017 Yazhou Zu, Wei Huang, Indrani Paul, and Vijay Janapa Reddi. "Ti-States: Power Management in Active Timing Margin Processors." In *IEEE Micro, 2017*. [DOI]

2016 Vijay Janapa Reddi and Hyesoon Kim. "On the Internet of Things." In *IEEE Micro, 2016*. [DOI]

2016 Jean Yang, Vijay Janapa Reddi, Yuhao Zhu, and Peter Bailis. "Research for Practice: Web Security and Mobile Web Computing." In *ACM Queue, 2016*. [DOI]

2015 Yuhao Zhu, Matthew Halpern, and Vijay Janapa Reddi. "The Role of the CPU in Energy-Efficient Mobile Web Browsing." In *IEEE Micro, 2015*. [DOI]

2014 Yuhao Zhu, Aditya Srikanth, Jingwen Leng, and Vijay Janapa Reddi. "Exploiting Webpage Characteristics for Energy-Efficient Mobile Web Browsing." In *IEEE Comput. Archit. Lett., 2014*. [DOI]

2013 Vijay Janapa Reddi. "Reliability-Aware Microarchitecture Design." In *IEEE Micro, 2013*. [DOI]

2011 Vijay Janapa Reddi, Benjamin C. Lee, Trishul M. Chilimbi, and Kushagra Vaid. "Mobile processors for energy-efficient web search." In *ACM Trans. Comput. Syst., 2011*. [DOI]

2011 Vijay Janapa Reddi and David M. Brooks. "Resilient Architectures via Collaborative Design: Maximizing Commodity Processor Performance in the Presence of Variations." In *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 2011*. [DOI]

2011 Vijay Janapa Reddi, Svilen Kanev, Wonyoung Kim, Simone Campanoni, Michael D. Smith, Gu-Yeon Wei, and David M. Brooks. "Voltage Noise in Production Processors." In *IEEE Micro, 2011*. [DOI]

2010 Vijay Janapa Reddi, Simone Campanoni, Meeta Sharma Gupta, Michael D. Smith, Gu-Yeon Wei, David M. Brooks, and Kim M. Hazelwood. "Eliminating voltage emergencies via software-guided

code transformations." In *ACM Trans. Archit. Code Optim., 2010*. [DOI]

2010 Vijay Janapa Reddi, Meeta Sharma Gupta, Glenn H. Holloway, Michael D. Smith, Gu-Yeon Wei, and David M. Brooks. "Predicting Voltage Droops Using Recurring Program and Microarchitectural Event Activity." In *IEEE Micro, 2010*. [DOI]

2009 Alex Shye, Joseph Blomstedt, Tipp Moseley, Vijay Janapa Reddi, and Daniel A. Connors. "PLR: A Software Approach to Transient Fault Tolerance for Multicore Architectures." In *IEEE Trans. Dependable Secur. Comput., 2009*. [DOI]

2006 Qiang Wu, Margaret Martonosi, Douglas W. Clark, Vijay Janapa Reddi, Dan Connors, Youfeng Wu, Jin Lee, and David M. Brooks. "Dynamic-Compiler-Driven Control for Microprocessor Energy and Performance." In *IEEE Micro, 2006*. [DOI]

2005 Vijay Janapa Reddi, Dan Connors, and Robert S. Cohn. "Persistence in dynamic code transformation systems." In *SIGARCH Comput. Archit. News, 2005*. [DOI]

ArXiV Preprints

2026 Charvi Rastogi, Mukul Bhutani, Minsuk Kahng, Shamsuddeen Hassan Muhammad, Evgeniia Razumovskaia, Priyanka Suresh, Ibrahim Said Ahmad, Charu Kalia, Yaaseen Mahomed, Madhurima Maji, Minjae Lee, Alicia Parrish, Jessica Quaye, Vijay Janapa Reddi, Aishwarya Verma, and Lora Aroyo. "Going PLACES: Participatory Localized Red Teaming for Text-to-Image Safety in the Global South." In *arXiv:2605.19190*. [Link]

2026 Srinivas Sridharan, Theodor-Adrian Badea, Andy Balogh, Bradford M. Beckmann, Brian Coutinho, Louis Feng, Sheng Fu, Sanshan Gao, Mehryar Garakani, Taekyung Heo, David Kanter, Josh Ladd, Ziwei Li, Winston Liu, Changhai Man, Dan Mihailescu, Spandan More, Joongun Park, Ashwin Ramachandran, Vinay Ramakrishnaiah, Saeed Rashidi, Vijay Janapa Reddi, Puneet Sharma, Phio Tian, William Won, Hanjiang Wu, Huan Xu, Jinsun Yoo, and Tushar Krishna. "MLCommons Chakra: Advancing Performance Benchmarking and Co-design using Standardized Execution Traces." In *arXiv:2605.11333*. [Link]

2026 Arya Tschand, Chenyu Wang, Zishen Wan, Andrew Cheng, Ioana Cristescu, Kevin He, Howard Huang, Alexander Ingare, Akseli Kangaslahti, Sara Kangaslahti, Theo Lebryk, Hongjin Lin, Jeffrey Jian Ma, Alexandru Meterez, Clara Mohri, Depen Morwani, Sunny Qin, Roy Rinberg, Paula Rodriguez Diaz, Alyssa Mia Taliotis, Pernille Undrum Fathi, Rosie Zhao, Todd Zhou, and Vijay Janapa Reddi. "GenAI for Systems: Recurring Challenges and Design Principles from Software to Silicon." In *arXiv:2602.15241*. [DOI]

2026 Ikechukwu Uchendu, Swati Goel, Karly Hou, Ebrahim M. Songhori, Kuang-Huei Lee, Joe Wenjie Jiang, Vijay Janapa Reddi, and Vincent Zhuang. "See it to Place it: Evolving Macro Placements with Vision-Language Models." In *arXiv:2603.28733*. [DOI]

2026 Vijay Janapa Reddi. "TinyTorch: Building Machine Learning Systems from First Principles." In *arXiv:2601.19107*. [DOI]

2025 Ikechukwu Uchendu, Jason Jabbour, Korneel Van den Berghe, Joel Runevic, Matthew Stewart, Jeffrey Jian Ma, Srivatsan Krishnan, Izzeddin Gur, Austin Huang, Colton Bishop, Paige Bailey, Wenjie Jiang, Ebrahim M. Songhori, Sergio Guadarrama, Jie Tan, Jordan K. Terry, Aleksandra Faust, and Vijay Janapa Reddi. "A2Perf: Real-World Autonomous Agents Benchmark." In *arXiv:2503.03056*. [DOI]

2025 Gregor von Laszewski, Wesley Brewer, Jeyan Thiyagalingam, Juri Papay, Armstrong Foundjem, Piotr Luszczek, Murali Emani, Shirley V. Moore, Vijay Janapa Reddi, Matthew D. Sinclair, Sebastian Lobentanzer, Sujata S. Goswami, Benjamin Hawks, Marco Colombo, Nhan Tran, Christine R. Kirkpatrick, Abdulkareem Alsudais, Gregg Barrett, Tianhao Li, Kirsten N. Morehouse, Shivaram Venkataraman, Rutwik Jain, Kartik Mathur, Victor Lu, Tejinder Singh, Khojasteh Z. Mirza, Kongtao Chen, Sasidhar Kunapuli, Gavin Farrell, Renato Umeton, and Geoffrey C. Fox. "AI Benchmark Democratization and Carpentry." In *arXiv:2512.11588*. [DOI]

2025 Korneel Van den Berghe, Stein Stroobants, Vijay Janapa Reddi, and Guido C. H. E. de Croon. "Adaptive Surrogate Gradients for Sequential Reinforcement Learning in Spiking Neural Networks." In *arXiv:2510.24461*. [DOI]

2025 Aditi Raju, Jared Ni, William Won, Changhai Man, Srivatsan Krishnan, Srinivas Sridharan, Amir Yazdanbakhsh, Tushar Krishna, and Vijay Janapa Reddi. "COSMIC: Enabling Full-Stack Co-Design and Optimization of Distributed Machine Learning Systems." In *arXiv:2505.15020*. [DOI]

2025 Jason Jabbour, Dong-Ki Kim, Max Olan Smith, Jay Patrikar, Radhika Ghosal, Youhui Wang, Ali Agha, Vijay Janapa Reddi, and Shayegan Omidshafiei. "Don't Run with Scissors: Pruning Breaks VLA Models but They Can Be Recovered." In *arXiv:2510.08464*. [DOI]

2025 Jessica Quaye, Charvi Rastogi, Alicia Parrish, Oana Inel, Minsuk Kahng, Lora Aroyo, and Vijay Janapa Reddi. "From Seed to Harvest: Augmenting Human Creativity with AI for Red-teaming Text-to-Image Models." In *arXiv:2507.17922*. [DOI]

2025 Zishen Wan, Jiayi Qian, Yuhang Du, Jason Jabbour, Yilun Du, Yang Katie Zhao, Arijit Raychowdhury, Tushar Krishna, and Vijay Janapa Reddi. "Generative AI in Embodied Systems: System-Level Analysis of Performance, Efficiency and Scalability." In *arXiv:2504.18945*. [DOI]

2025 Shvetank Prakash, Andrew Cheng, Olof Kindgren, Ashiq Ahamed, Graham Knight, Jedrzej Kufel, Francisco Rodriguez, Arya Tschand, David Kong, Mariam Elgamal, Jerry Huang, Emma Chen, Gage Hills, Richard Price, Emre Ozer, and Vijay Janapa Reddi. "Lifetime-Aware Design of Item-Level Intelligence." In *arXiv:2509.08193*. [DOI]

2025 Arissa Wongpanich, Tayo Oguntebi, Jose Baiocchi Paredes, Yu Emma Wang, Phitchaya Mangpo Phothilimthana, Ritwika Mitra, Zongwei Zhou, Naveen Kumar, and Vijay Janapa Reddi. "Machine Learning Fleet Efficiency: Analyzing and Optimizing Large-Scale Google TPU Systems with ML Productivity Goodput." In *arXiv:2502.06982*. [DOI]

2025 Jason Yik, Walter Gallego Gomez, Andrew Cheng, Benedetto Leto, Alessandro Pierro, Noah Pacik-Nelson, Korneel Van den Berghe, Vittorio Fra, Andreea Danielescu, Gianvito Urgese, and Vijay Janapa Reddi. "Modeling and Optimizing Performance Bottlenecks for Neuromorphic Accelerators." In *arXiv:2511.21549*. [DOI]

2025 Srivatsan Krishnan, Jason Jabbour, Dan Zhang, Natasha Jaques, Aleksandra Faust, Shayegan Omidshafiei, and Vijay Janapa Reddi. "Multi-Agent Reinforcement Learning for Sample-Efficient Deep Neural Network Mapping." In *arXiv:2507.16249*. [DOI]

2025 Shvetank Prakash, Andrew Cheng, Arya Tschand, Mark Mazumder, Varun Gohil, Jeffrey Jian Ma, Jason Yik, Zishen Wan, Jessica Quaye, Elisavet Lydia Alvanaki, Avinash Kumar, Chandrashis Mazumdar, Tuhin Khare, Alexander Ingare, Ikechukwu Uchendu, Radhika Ghosal, Abhishek Tyagi, Chenyu Wang, Andrea Mattia Garavagno, Sarah Gu, Alice Guo, Grace Hur, Luca P. Carloni, Tushar Krishna, Ankita Nayak, Amir Yazdanbakhsh, and Vijay Janapa Reddi. "QuArch: A Benchmark for

Evaluating LLM Reasoning in Computer Architecture." In *arXiv:2510.22087*. [DOI]

2025 Shvetank Prakash, Andrew Cheng, Jason Yik, Arya Tschand, Radhika Ghosal, Ikechukwu Uchendu, Jessica Quaye, Jeffrey Jian Ma, Shreyas Grampurohit, Sofia Giannuzzi, Arnav Balyan, Fin Amin, Aadya Pipersenia, Yash Choudhary, Ankita Nayak, Amir Yazdanbakhsh, and Vijay Janapa Reddi. "QuArch: A Question-Answering Dataset for AI Agents in Computer Architecture." In *arXiv:2501.01892*. [DOI]

2025 Jeffrey Jian Ma, Milad Hashemi, Amir Yazdanbakhsh, Kevin Swersky, Ofir Press, Enhui Li, Vijay Janapa Reddi, and Parthasarathy Ranganathan. "SWE-fficiency: Can Language Models Optimize Real-World Repositories on Real Workloads?." In *arXiv:2511.06090*. [DOI]

2025 Chenyu Wang, Zishen Wan, Hao Kang, Emma Chen, Zhiqiang Xie, Tushar Krishna, Vijay Janapa Reddi, and Yilun Du. "Slm-mux: Orchestrating small language models for reasoning." In *arXiv:2510.05077*. [DOI]

2025 Jason Jabbour, Kai Kleinbard, Olivia Miller, Robert Haussman, and Vijay Janapa Reddi. "SocratiQ: A Generative AI-Powered Learning Companion for Personalized Education and Broader Accessibility." In *arXiv:2502.00341*. [DOI]

2025 Arya Tschand, Muhammad A. Awad, Ryan Swann, Kesavan Ramakrishnan, Jeffrey Jian Ma, Keith Lowery, Ganesh Dasika, and Vijay Janapa Reddi. "SwizzlePerf: Hardware-Aware LLMs for GPU Kernel Performance Optimization." In *arXiv:2508.20258*. [DOI]

2024 Jessica Quaye, Alicia Parrish, Oana Inel, Charvi Rastogi, Hannah Rose Kirk, Minsuk Kahng, Erin van Liemt, Max Bartolo, Jess Tsang, Justin White, Nathan Clement, Rafael Mosquera, Juan Ciro, Vijay Janapa Reddi, and Lora Aroyo. "Adversarial Nibbler: An Open Red-Teaming Method for Identifying Diverse Harms in Text-to-Image Generation." In *arXiv:2403.12075*. [DOI]

2024 Jeffrey Jian Ma, Alan Tu, Yiling Chen, and Vijay Janapa Reddi. "FedStaleWeight: Buffered Asynchronous Federated Learning with Fair Aggregation via Staleness Reweighting." In *arXiv:2406.02877*. [DOI]

2024 Jason Jabbour and Vijay Janapa Reddi. "Generative AI Agents in Autonomous Machines: A Safety Perspective." In *arXiv:2410.15489*. [DOI]

2024 Arya Tschand, Arun Tejusve Raghunath Rajan, Sachin Idgunji, Anirban Ghosh, Jeremy Holleman, Csaba Király, Pawan Ambalkar, Ritika Borkar, Ramesh Chukka, Trevor Cockrell, Oliver Curtis, Grigori Fursin, Miro Hodak, Hiwot Kassa, Anton Lokhmotov, Dejan Miskovic, Yuechao Pan, Manu Prasad Manmathan, Liz Raymond, Tom St. John, Arjun Suresh, Rowan Taubitz, Sean Zhan, Scott Wasson, David Kanter, and Vijay Janapa Reddi. "MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from Microwatts to Megawatts for Sustainable AI." In *arXiv:2410.12032*. [DOI]

2024 Matthew Stewart, Emanuel Moss, Pete Warden, Brian Plancher, Susan Kennedy, Mona Sloane, and Vijay Janapa Reddi. "Materiality and Risk in the Age of Pervasive AI Sensors." In *arXiv:2402.11183*. [DOI]

2024 Sabrina M. Neuman, Brian Plancher, and Vijay Janapa Reddi. "The Magnificent Seven Challenges and Opportunities in Domain-Specific Accelerator Design for Autonomous Systems." In *arXiv:2407.17311*. [DOI]

2024 Jacob Huckelberry, Yuke Zhang, Allison Sansone, James Mickens, Peter A. Beerel, and Vijay Janapa Reddi. "TinyML Security: Exploring Vulnerabilities in Resource-Constrained Machine Learning Systems." In *arXiv:2411.07114*. [DOI]

2024 Colby R. Banbury, Emil J. Njor, Matthew Stewart, Pete Warden, Manjunath Kudlur, Nat Jeffries, Xenofon Fafoutis, and Vijay Janapa Reddi. "Wake Vision: A Large-scale, Diverse Dataset and Benchmark Suite for TinyML Person Detection." In *arXiv:2405.00892*. [DOI]

2023 Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Max Bartolo, Oana Inel, Juan Ciro, Rafael Mosquera, Addison Howard, Will Cukierski, D. Sculley, Vijay Janapa Reddi, and Lora Aroyo. "Adversarial Nibbler: A Data-Centric Challenge for Improving the Safety of Text-to-Image Models." In *arXiv:2305.14384*. [DOI]

2023 Srivatsan Krishnan, Amir Yazdanbakhsh, Shvetank Prakash, Jason Jabbour, Ikechukwu Uchendu, Susobhan Ghosh, Behzad Boroujerdian, Daniel Richins, Devashree Tripathy, Aleksandra Faust, and Vijay Janapa Reddi. "ArchGym: An Open-Source Gymnasium for Machine Learning Assisted Architecture Design." In *arXiv:2306.08888*. [DOI]

2023 Zishen Wan, Nandhini Chandramoorthy, Karthik Swaminathan, Pin-Yu Chen, Vijay Janapa Reddi, and Arijit Raychowdhury. "BERRY: Bit Error Robustness for Energy-Efficient Reinforcement Learning-Based Autonomous Systems." In *arXiv:2307.10041*. [DOI]

2023 Luis Oala, Manil Maskey, Lilith Bat-Leah, Alicia Parrish, Nezihe Merve Gürel, Tzu-Sheng Kuo, Yang Liu, Rotem Dror, Danilo Brajovic, Xiaozhe Yao, Max Bartolo, William Gaviria Rojas, Ryan Hileman, Rainier Aliment, Michael W. Mahoney, Meg Risdal, Matthew Lease, Wojciech Samek, Debojyoti Dutta, Curtis G. Northcutt, Cody Coleman, Braden Hancock, Bernard Koch, Girmaw Abebe Tadesse, Bojan Karlas, Ahmed M. Alaa, Adji Bousso Dieng, Natasha F. Noy, Vijay Janapa Reddi, James Zou, Praveen K. Paritosh, Mihaela van der Schaar, Kurt Bollacker, Lora Aroyo, Ce Zhang, Joaquin Vanschoren, Isabelle Guyon, and Peter Mattson. "DMLR: Data-centric Machine Learning Research - Past, Present and Future." In *arXiv:2311.13028*. [DOI]

2023 Matthew Stewart, Pete Warden, Yasmine Omri, Shvetank Prakash, Joao Santos, Shawn Hymel, Benjamin Brown, Jim MacArthur, Nat Jeffries, Brian Plancher, and Vijay Janapa Reddi. "Datasheets for Machine Learning Sensors." In *arXiv:2306.08848*. [DOI]

2023 Maximilian Lam, Jeff Johnson, Wenjie Xiong, Kiwan Maeng, Udit Gupta, Yang Li, Liangzhen Lai, Ilias Leontiadis, Minsoo Rhu, Hsien-Hsin S. Lee, Vijay Janapa Reddi, Gu-Yeon Wei, David Brooks, and G. Edward Suh. "GPU-based Private Information Retrieval for On-Device Machine Learning Inference." In *arXiv:2301.10904*. [DOI]

2023 Shvetank Prakash, Matthew Stewart, Colby R. Banbury, Mark Mazumder, Pete Warden, Brian Plancher, and Vijay Janapa Reddi. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." In *arXiv:2301.11899*. [DOI]

2023 Cansu Demirkiran, Rashmi Agrawal, Vijay Janapa Reddi, Darius Bunandar, and Ajay Joshi. "Leveraging Residue Number System for Designing High-Precision Analog Deep Neural Network Accelerators." In *arXiv:2306.09481*. [DOI]

2023 Jason Yik, Soikat Hasan Ahmed, Zergham Ahmed, Brian Anderson, Andreas G. Andreou, Chiara Bartolozzi, Arindam Basu, Douwe den Blanken, Petrut Bogdan, Sander M. Bohté, Younes Bouhadjar, Sonia M. Buckley, Gert Cauwenberghs, Federico Corradi, Guido de Croon, Andreea Danielescu, Anurag Reddy Daram, Mike Davies, Yigit Demirag, Jason Eshraghian, Jeremy Forest, Steve B.

Furber, P. Michael Furlong, Aditya Gilra, Giacomo Indiveri, Siddharth Joshi, Vedant Karia, Lyes Khacef, James C. Knight, Laura Kriener, Rajkumar Kubendran, Dhireesha Kudithipudi, Gregor Lenz, Rajit Manohar, Christian Mayr, Konstantinos P. Michmizos, Dylan R. Muir, Emre Neftci, Thomas Nowotny, Fabrizio Ottati, Ayça Özcelikkale, Noah Pacik-Nelson, Priyadarshini Panda, Pao-Sheng Sun, Melika Payvand, Christian Pehle, Mihai A. Petrovici, Christoph Posch, Alpha Renner, Yulia Sandamirskaya, Clemens JS Schaefer, André van Schaik, Johannes Schemmel, Catherine D. Schuman, Jae-sun Seo, Sumit Bam Shrestha, Manolis Sifalakis, Amos Sironi, Kenneth Michael Stewart, Terrence C. Stewart, Philipp Stratmann, Guangzhi Tang, Jonathan Timcheck, Marian Verhelst, Craig M. Vineyard, Bernhard Vogginger, Amirreza Yousefzadeh, Biyan Zhou, Fatima Tuz Zohora, Charlotte Frenkel, and Vijay Janapa Reddi. "NeuroBench: Advancing Neuromorphic Computing through Collaborative, Fair and Representative Benchmarking." In *arXiv:2304.04640*. [DOI]

2023 Víctor Mayoral Vilches, Jason Jabbour, Yu-Shun Hsiao, Zishen Wan, Alejandra Martínez-Fariña, Martiño Crespo-Álvarez, Matthew Stewart, Juan Manuel Reina-Muñoz, Prateek Nagras, Gaurav Vikhe, Mohammad Bakhshalipour, Martin Pinzger, Stefan Rass, Smruti Panigrahi, Giulio Corradi, Niladri Roy, Phillip B. Gibbons, Sabrina M. Neuman, Brian Plancher, and Vijay Janapa Reddi. "RobotPerf: An Open-Source, Vendor-Agnostic, Benchmarking Suite for Evaluating Robotics Computing System Performance." In *arXiv:2309.09212*. [DOI]

2023 Yu-Shun Hsiao, Siva Kumar Sastry Hari, Balakumar Sundaralingam, Jason Yik, Thierry Tambe, Charbel Sakr, Stephen W. Keckler, and Vijay Janapa Reddi. "VaPr: Variable-Precision Tensors to Accelerate Robot Motion Planning." In *arXiv:2310.07854*. [DOI]

2022 Shvetank Prakash, Tim Callahan, Joseph Bushagour, Colby R. Banbury, Alan V. Green, Pete Warden, Tim Ansell, and Vijay Janapa Reddi. "CFU Playground: Full-Stack Open-Source Framework for Tiny Machine Learning (tinyML) Acceleration on FPGAs." In *arXiv:2201.01863*. [Link]

2022 Mark Mazumder, Colby R. Banbury, Xiaozhe Yao, Bojan Karlas, William Gaviria Rojas, Sudnya Frederick Diamos, Greg Diamos, Lynn He, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Sabri Eyuboglu, Amirata Ghorbani, Emmett D. Goodman, Tariq Kane, Christine R. Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen K. Paritosh, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Y. Ng, Peter Mattson, and Vijay Janapa Reddi. "DataPerf: Benchmarks for Data-Centric AI Development." In *arXiv:2207.10062*. [DOI]

2022 Shawn Hymel, Colby R. Banbury, Daniel Situnayake, Alexander Elium, Carl Ward, Matthew Kelcey, Mathijs Baaijens, Mateusz Majchrzycki, Jenny Plunkett, David Tischler, Alessandro Grande, Louis Moreau, Dmitry Maslov, Arthur Beavis, Jan Jongboom, and Vijay Janapa Reddi. "Edge Impulse: An MLOps Platform for Tiny Machine Learning." In *arXiv:2212.03332*. [DOI]

2022 Behzad Boroujerdian, Ying Jing, Amit Kumar, Lavanya Subramanian, Luke Yen, Vincent Lee, Vivek Venkatesan, Amit Jindal, Robert Shearer, and Vijay Janapa Reddi. "FARSI: Facebook AR System Investigator for Agile Domain-Specific System-on-Chip Exploration." In *arXiv:2201.05232*. [Link]

2022 Zishen Wan, Aqeel Anwar, Abdulrahman Mahmoud, Tianyu Jia, Yu-Shun Hsiao, Vijay Janapa Reddi, and Arijit Raychowdhury. "FRL-FI: Transient Fault Analysis for Federated Reinforcement Learning-Based Navigation Systems." In *arXiv:2203.07276*. [DOI]

2022 Javier M. Duarte, Nhan Tran, Benjamin Hawks, Christian Herwig, Jules Muhizi, Shvetank Prakash, and Vijay Janapa Reddi. "FastML Science Benchmarks: Accelerating Real-Time Scientific Edge

Machine Learning." In *arXiv:2207.07958*. [DOI]

2022 Pete Warden, Matthew Stewart, Brian Plancher, Colby R. Banbury, Shvetank Prakash, Emma Chen, Zain Asgar, Sachin Katti, and Vijay Janapa Reddi. "Machine Learning Sensors." In *arXiv:2206.03266*. [DOI]

2022 Srivatsan Krishnan, Natasha Jaques, Shayegan Omidshafiei, Dan Zhang, Izzeddin Gur, Vijay Janapa Reddi, and Aleksandra Faust. "Multi-Agent Reinforcement Learning for Microprocessor Design Space Exploration." In *arXiv:2211.16385*. [DOI]

2022 Tianyu Jia, En-Yu Yang, Yu-Shun Hsiao, Jonathan J. Cruz, David Brooks, Gu-Yeon Wei, and Vijay Janapa Reddi. "OMU: A Probabilistic 3D Occupancy Mapping Accelerator for Real-time OctoMap at the Edge." In *arXiv:2205.03325*. [DOI]

2022 Víctor Mayoral Vilches, Sabrina M. Neuman, Brian Plancher, and Vijay Janapa Reddi. "RobotCore: An Open Architecture for Hardware Acceleration in ROS 2." In *arXiv:2205.03929*. [DOI]

2022 Zishen Wan, Ashwin Sanjay Lele, Bo Yu, Shaoshan Liu, Yu Wang, Vijay Janapa Reddi, Cong Hao, and Arijit Raychowdhury. "Robotic Computing on FPGAs: Current Progress, Research Challenges, and Opportunities." In *arXiv:2205.07149*. [DOI]

2022 Srivatsan Krishnan, Zishen Wan, Kshitij Bhardwaj, Ninad Jadhav, Aleksandra Faust, and Vijay Janapa Reddi. "Roofline Model for UAVs: A Bottleneck Analysis Tool for Onboard Compute Characterization of Autonomous Unmanned Aerial Vehicles." In *arXiv:2204.10898*. [DOI]

2022 Maximilian Lam, Michael Mitzenmacher, Vijay Janapa Reddi, Gu-Yeon Wei, and David Brooks. "Tabula: Efficiently Computing Nonlinear Activation Functions for Secure Neural Network Inference." In *arXiv:2203.02833*. [DOI]

2022 Sabrina M. Neuman, Brian Plancher, Bardienus Pieter Duisterhof, Srivatsan Krishnan, Colby R. Banbury, Mark Mazumder, Shvetank Prakash, Jason Jabbour, Aleksandra Faust, Guido C. H. E. de Croon, and Vijay Janapa Reddi. "Tiny Robot Learning: Challenges and Directions for Machine Learning in Resource-Constrained Robots." In *arXiv:2205.05748*. [DOI]

2022 Hyoukjun Kwon, Krishnakumar Nair, Jamin Seo, Jason Yik, Debabrata Mohapatra, Dongyuan Zhan, Jinook Song, Peter Capak, Peizhao Zhang, Peter Vajda, Colby R. Banbury, Mark Mazumder, Liangzhen Lai, Ashish Sirasao, Tushar Krishna, Harshit Khaitan, Vikas Chandra, and Vijay Janapa Reddi. "XRBench: An Extended Reality (XR) Machine Learning Benchmark Suite for the Metaverse." In *arXiv:2211.08675*. [DOI]

2022 Yu-Shun Hsiao, Siva Kumar Sastry Hari, Michal Filipiuk, Timothy Tsai, Michael B. Sullivan, Vijay Janapa Reddi, Vasu Singh, and Stephen W. Keckler. "Zhuyi: Perception Processing Rate Estimation for Safety in Autonomous Vehicles." In *arXiv:2205.03347*. [DOI]

2021 Cansu Demirkiran, Furkan Eris, Gongyu Wang, Jonathan Elmhurst, Nick Moore, Nicholas C. Harris, Ayon Basumallik, Vijay Janapa Reddi, Ajay Joshi, and Darius Bunandar. "An Electro-Photonic System for Accelerating Deep Neural Networks." In *arXiv:2109.01126*. [Link]

2021 Zishen Wan, Aqeel Anwar, Yu-Shun Hsiao, Tianyu Jia, Vijay Janapa Reddi, and Arijit Raychowdhury. "Analyzing and Improving Fault Tolerance of Learning-Based Navigation Systems." In *arXiv:2111.04957*. [Link]

2021 Srivatsan Krishnan, Thierry Tambe, Zishen Wan, and Vijay Janapa Reddi. "AutoSoC: Automating

Algorithm-SOC Co-design for Aerial Robots." In *arXiv:2109.05683*. [Link]

2021 Vijay Janapa Reddi, Greg Diamos, Pete Warden, Peter Mattson, and David Kanter. "Data Engineering for Everyone." In *arXiv:2102.11447*. [Link]

2021 Mark Mazumder, Colby R. Banbury, Josh Meyer, Pete Warden, and Vijay Janapa Reddi. "Few-Shot Keyword Spotting in Any Language." In *arXiv:2104.01454*. [Link]

2021 Brian Plancher, Sabrina M. Neuman, Radhika Ghosal, Scott Kuindersma, and Vijay Janapa Reddi. "GRiD: GPU-Accelerated Rigid Body Dynamics with Analytical Gradients." In *arXiv:2109.06976*. [Link]

2021 Maximilian Lam, Gu-Yeon Wei, David Brooks, Vijay Janapa Reddi, and Michael Mitzenmacher. "Gradient Disaggregation: Breaking Privacy in Federated Learning by Reconstructing the User Participant Matrix." In *arXiv:2106.06089*. [Link]

2021 Yu-Shun Hsiao, Zishen Wan, Tianyu Jia, Radhika Ghosal, Arijit Raychowdhury, David Brooks, Gu-Yeon Wei, and Vijay Janapa Reddi. "MAVFI: An End-to-End Fault Analysis Framework with Anomaly Detection and Recovery for Micro Aerial Vehicles." In *arXiv:2105.12882*. [Link]

2021 Colby R. Banbury, Vijay Janapa Reddi, Peter Torelli, Jeremy Holleman, Nat Jeffries, Csaba Király, Pietro Montino, David Kanter, Sebastian Ahmed, Danilo Pau, Urmish Thakker, Antonio Torrini, Pete Warden, Jay Cordaro, Giuseppe Di Guglielmo, Javier M. Duarte, Stephen Gibellini, Videet Parekh, Honson Tran, Nhan Tran, Wenxu Niu, and Xuesong Xu. "MLPerf Tiny Benchmark." In *arXiv:2106.07597*. [Link]

2021 Srivatsan Krishnan, Zishen Wan, Kshitij Bhardwaj, Paul N. Whatmough, Aleksandra Faust, Sabrina M. Neuman, Gu-Yeon Wei, David Brooks, and Vijay Janapa Reddi. "Machine Learning-Based Automated Design Space Exploration for Autonomous Aerial Robots." In *arXiv:2102.02988*. [Link]

2021 Alexandros Karargyris, Renato Umeton, Micah J. Sheller, Alejandro Aristizabal, Johnu George, Srini Bala, Daniel J. Beutel, Victor Bittorf, Akshay Chaudhari, Alexander Chowdhury, Cody Coleman, Bala Desinghu, Gregory F. Diamos, Debo Dutta, Diane Feddema, Grigori Fursin, Junyi Guo, Xinyuan Huang, David Kanter, Satyananda Kashyap, Nicholas D. Lane, Indranil Mallick, Pietro Mascagni, Virendra Mehta, Vivek Natarajan, Nikola Nikolov, Nicolas Padoy, Gennady Pekhimenko, Vijay Janapa Reddi, G. Anthony Reina, Pablo Ribalta, Jacob Rosenthal, Abhishek Singh, Jayaraman J. Thiagarajan, Anna Wuest, Maria Xenochristou, Daguang Xu, Poonam Yadav, Michael Rosenthal, Massimo Loda, Jason M. Johnson, and Peter Mattson. "MedPerf: Open Benchmarking Platform for Medical Artificial Intelligence using Federated Evaluation." In *arXiv:2110.01406*. [Link]

2021 James Gleeson, Srivatsan Krishnan, Moshe Gabel, Vijay Janapa Reddi, Eyal de Lara, and Gennady Pekhimenko. "RL-Scope: Cross-Stack Profiling for Deep Reinforcement Learning Workloads." In *arXiv:2102.04285*. [Link]

2021 Behzad Boroujerdian, Radhika Ghosal, Jonathan J. Cruz, Brian Plancher, and Vijay Janapa Reddi. "RoboRun: A Robot Runtime to Exploit Spatial Heterogeneity." In *arXiv:2108.13354*. [Link]

2021 Bardienus Pieter Duisterhof, Shushuai Li, Javier Burgués, Vijay Janapa Reddi, and Guido C. H. E. de Croon. "Sniffy Bug: A Fully Autonomous Swarm of Gas-Seeking Nano Quadcopters in Cluttered Environments." In *arXiv:2107.05490*. [Link]

2021 Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter,

Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. "The People's Speech: A Large-Scale Diverse English Speech Recognition Dataset for Commercial Usage." In *arXiv:2111.09344*. [Link]

2021 Vijay Janapa Reddi, Brian Plancher, Susan Kennedy, Laurence Moroney, Pete Warden, Anant Agarwal, Colby R. Banbury, Massimo Banzi, Matthew Bennett, Benjamin Brown, Sharad Chitlangia, Radhika Ghosal, Sarah Grafman, Rupert Jaeger, Srivatsan Krishnan, Maximilian Lam, Daniel Leiker, Cara Mann, Mark Mazumder, Dominic Pajak, Dhilan Ramaprasad, J. Evan Smith, Matthew Stewart, and Dustin Tingley. "Widening Access to Applied Machine Learning with TinyML." In *arXiv:2106.04008*. [Link]

2020 Daniel Richins, Dharmisha Doshi, Matthew Blackmore, Aswathy Thulaseedharan Nair, Neha Pathapati, Ankit Patel, Brainard Daguman, Daniel Dobrijalowski, Ramesh Illikkal, Kevin Long, David Zimmerman, and Vijay Janapa Reddi. "AI Tax: The Hidden Cost of AI Data Center Applications." In *arXiv:2007.10571*. [Link]

2020 Colby R. Banbury, Vijay Janapa Reddi, Max Lam, William Fu, Amin Fazel, Jeremy Holleman, Xinyuan Huang, Robert Hurtado, David Kanter, Anton Lokhmotov, David A. Patterson, Danilo Pau, Jae-sun Seo, Jeff Sieracki, Urmish Thakker, Marian Verhelst, and Poonam Yadav. "Benchmarking TinyML Systems: Challenges and Direction." In *arXiv:2003.04821*. [Link]

2020 George Papadimitriou, Athanasios Chatzidimitriou, Dimitris Gizopoulos, Vijay Janapa Reddi, Jingwen Leng, Behzad Salami, Osman S. Unsal, and Adrián Cristal Kestelman. "Exceeding Conservative Limits: A Consolidated Analysis on Modern Hardware Margins." In *arXiv:2006.01049*. [Link]

2020 Vijay Janapa Reddi, David Kanter, Peter Mattson, Jared Duke, Thai Nguyen, Ramesh Chukka, Kenneth Shiring, Koan-Sin Tan, Mark Charlebois, William Chou, Mostafa El-Khamy, Jungwook Hong, Michael Buch, Cindy Trinh, Thomas Atta-Fosu, Fatih Çakir, Masoud Charkhabi, Xiaodong Chen, Jimmy Chiang, Dave Dexter, Woncheol Heo, Guenther Schmuelling, Maryam Shabani, and Dylan Zika. "MLPerf Mobile Inference Benchmark: Why Mobile AI Benchmarking Is Hard and What to Do About It." In *arXiv:2012.02328*. [Link]

2020 Colby R. Banbury, Chuteng Zhou, Igor Fedorov, Ramon Matas Navarro, Urmish Thakker, Dibakar Gope, Vijay Janapa Reddi, Matthew Mattina, and Paul N. Whatmough. "MicroNets: Neural Network Architectures for Deploying TinyML Applications on Commodity Microcontrollers." In *arXiv:2010.11267*. [Link]

2020 Maximilian Lam, Zachary Yedidia, Colby R. Banbury, and Vijay Janapa Reddi. "Quantized Neural Network Inference with Precision Batching." In *arXiv:2003.00822*. [Link]

2020 Robert David, Jared Duke, Advait Jain, Vijay Janapa Reddi, Nat Jeffries, Jian Li, Nick Kreeger, Ian Nappier, Meghna Natraj, Shlomi Regev, Rocky Rhodes, Tiezhen Wang, and Pete Warden. "TensorFlow Lite Micro: Embedded Machine Learning on TinyML Systems." In *arXiv:2010.08678*. [Link]

2019 Mark D. Hill and Vijay Janapa Reddi. "Accelerator-level Parallelism." In *arXiv:1907.02064*. [Link]

2019 Thierry Tambe, En-Yu Yang, Zishen Wan, Yuntian Deng, Vijay Janapa Reddi, Alexander M. Rush, David Brooks, and Gu-Yeon Wei. "AdaptivFloat: A Floating-point based Data Type for Resilient Deep Learning Inference." In *arXiv:1909.13271*. [Link]

2019 Srivatsan Krishnan, Behzad Boroujerdian, William Fu, Aleksandra Faust, and Vijay Janapa Reddi. "Air Learning: An AI Research Platform for Algorithm-Hardware Benchmarking of Autonomous

Aerial Robots." In *arXiv:1906.00421*. [Link]

2019 Thanh Thi Nguyen and Vijay Janapa Reddi. "Deep Reinforcement Learning for Cyber Security." In *arXiv:1906.05799*. [Link]

2019 Ethan Shaotran, Jonathan J. Cruz, and Vijay Janapa Reddi. "GLADAS: Gesture Learning for Advanced Driver Assistance Systems." In *arXiv:1910.04695*. [Link]

2019 Bardienus Pieter Duisterhof, Srivatsan Krishnan, Jonathan J. Cruz, Colby R. Banbury, William Fu, Aleksandra Faust, Guido C. H. E. de Croon, and Vijay Janapa Reddi. "Learning to Seek: Autonomous Source Seeking with Deep Reinforcement Learning Onboard a Nano Drone Microcontroller." In *arXiv:1909.11236*. [Link]

2019 Behzad Boroujerdian, Hasan Genc, Srivatsan Krishnan, Wenzhi Cui, Marcelino Almeida, Kayvan Mansoorshahi, Aleksandra Faust, and Vijay Janapa Reddi. "MAVBench: Micro Aerial Vehicle Benchmarking." In *arXiv:1905.06388*. [Link]

2019 Vijay Janapa Reddi, Christine Cheng, David Kanter, Peter Mattson, Guenther Schmuelling, Carole-Jean Wu, Brian Anderson, Maximilien Breughe, Mark Charlebois, William Chou, Ramesh Chukka, Cody Coleman, Sam Davis, Pan Deng, Greg Diamos, Jared Duke, Dave Fick, J. Scott Gardner, Itay Hubara, Sachin Idgunji, Thomas B. Jablin, Jeff Jiao, Tom St. John, Pankaj Kanwar, David Lee, Jeffery Liao, Anton Lokhmotov, Francisco Massa, Peng Meng, Paulius Micikevicius, Colin Osborne, Gennady Pekhimenko, Arun Tejusve Raghunath Rajan, Dilip Sequeira, Ashish Sirasao, Fei Sun, Hanlin Tang, Michael Thomson, Frank Wei, Ephrem Wu, Lingjie Xu, Koichi Yamada, Bing Yu, George Yuan, Aaron Zhong, Peizhao Zhang, and Yuchen Zhou. "MLPerf Inference Benchmark." In *arXiv:1911.02549*. [Link]

2019 Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David A. Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debojyoti Dutta, Udit Gupta, Kim M. Hazelwood, Andrew Hock, Xinyuan Huang, Bill Jia, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Guokai Ma, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St. John, Carole-Jean Wu, Lingjie Xu, Cliff Young, and Matei Zaharia. "MLPerf Training Benchmark." In *arXiv:1910.01500*. [Link]

2019 Matthew Halpern, Behzad Boroujerdian, Todd W. Mummert, Evelyn Duesterwald, and Vijay Janapa Reddi. "One Size Does Not Fit All: Quantifying and Exposing the Accuracy-Latency Trade-off in Machine Learning Cloud Service APIs via Tolerance Tiers." In *arXiv:1906.11307*. [Link]

2019 Srivatsan Krishnan, Sharad Chitlangia, Maximilian Lam, Zishen Wan, Aleksandra Faust, and Vijay Janapa Reddi. "Quantized Reinforcement Learning (QUARL)." In *arXiv:1910.01055*. [Link]

2019 Behzad Boroujerdian, Hasan Genc, Srivatsan Krishnan, Bardienus Pieter Duisterhof, Brian Plancher, Kayvan Mansoorshahi, Marcelino Almeida, Wenzhi Cui, Aleksandra Faust, and Vijay Janapa Reddi. "The Role of Compute in Autonomous Aerial Vehicles." In *arXiv:1906.10513*. [Link]

2018 Ting-Wu Chin, Chia-Lin Yu, Matthew Halpern, Hasan Genc, Shiao-Li Tsao, and Vijay Janapa Reddi. "Domain Specific Approximation for Object Detection." In *arXiv:1810.02010*. [Link]

2016 Mikhail Kazdagli, Ling Huang, Vijay Janapa Reddi, and Mohit Tiwari. "EMMA: A New Platform to Evaluate Hardware-based Mobile Malware Analyses." In *arXiv:1603.03086*. [Link]

Workshop Papers

2017 Jayashree Mohan, Dhathri Purohith, Matthew Halpern, Vijay Chidambaram, and Vijay Janapa Reddi. "Storage on Your SmartPhone Uses More Energy Than You Think." In *HotStorage 2017*. [Link]

2005 Alex Shye, Matthew Iyer, Tipp Moseley, David Hodgdon, Dan Fay, Vijay Janapa Reddi, and Daniel A. Connors. "Analysis of path profiling information generated with performance monitoring hardware." In *INTERACT-9 2005*. [DOI]

2005 Alex Shye, Matthew Iyer, Vijay Janapa Reddi, and Daniel A. Connors. "Code coverage testing using hardware performance monitoring support." In *AADEBUG 2005*. [DOI]

2004 Vijay Janapa Reddi, Alex Settle, Daniel A. Connors, and Robert S. Cohn. "PIN: a binary instrumentation tool for computer architecture research and education." In *ISCA 2004*. [DOI]

Books and Chapters

2013 Vijay Janapa Reddi and Meeta Sharma Gupta. "Resilient Architecture Design for Voltage Variation." In *Synthesis Lectures on Computer Architecture, 2013*. [DOI]

Talks

10/2024 "MLSysBook.AI: Principles and Practices of Machine Learning Systems Engineering." Embedded Systems Week (ESWEEK), Virtual.

09/2024 "Architecture 2.0." International Symposium on Workload Characterization (IISWC). (Keynote).

9/2024 "Architecture 2.0: Challenges and Opportunities for tinyML." tinyAI Forum on Generative AI on the Edge, Virtual.

6/2024 "Architecture 2.0." Design Automation Conference (DAC).

6/2024 "The Magnificent Seven Challenges and Opportunities in Domain-Specific Accelerator Design for Autonomous Systems." Design Automation Conference (DAC).

6/2024 "Architecture 2.0: From Concept to Breaking Ground." CogArch Workshop at (ISCA). (Keynote).

6/2024 "The Magnificent Seven Challenges and Opportunities in Domain-Specific Accelerator Design for Autonomous Systems." EPOCHS Workshop at ISCA.

4/2024 "The State of TinyML Benchmarking: Current Landscape, Challenges, and Emerging Trends." tinyML Summit.

1/2024 "Architecture 2.0." Free and Open Source Silicon (FOSSI).

10/2023 "Architecture 2.0." Open Compute Project (OCP), Virtual.

09/2023 "Architecture 2.0." MIT Industry-Academia Partnership.

07/2023 "Architecture 2.0." Design Automation Conference (DAC), Virtual.

06/2023 "The Parameter and Chip Wars." Vail Computer Elements Workshop, Virtual.

06/2023 "Adopting AI: With Power Comes Responsibility." Panel, FDCAI, Virtual.

04/2023 "TinyML." IEEE International Symposium on Low-Power and High-Speed Chips (COOL Chips), Virtual.

04/2023 "The Parameter and Chip Wars: Shifting the Focus from Model-centric to Data-centric AI." MICRON, Virtual.

04/2023 "NeuroBench: Advancing Neuromorphic Computing through Collaborative and Rigorous Benchmarking." NICE, Virtual.

11/2022 "TinyML." Urban Sensor Networks Workshop Panel, Virtual.

10/2022 "ML Metrics: The Past, Present, and Future of Benchmarking ML Systems, Datasets, and Use Cases." Specialization with Benchmarks for Emerging Applications (MICRO), Virtual.

09/2022 "Tiny Machine Learning." Chips & Compilers Symposium, MLSys '22, Virtual.

09/2022 "MLPerf & DataPerf." The Autonomous Vehicle Computing Consortium (AVCC), Virtual.

09/2022 "Benchmarking FastML Systems." Fast ML for Science Workshop.

08/2022 "Tiny Machine Learning: A System-level Perspective." ACM/IEEE International Symposium on Low Power Electronics and Design, Virtual. (Keynote).

08/2022 "The Vision Behind MLPerf and DataPerf." Monterey Data Conference, Virtual.

08/2022 "DataPerf: Benchmarks for Data-centric AI Development." The Future of Data-Centric AI, Snorkel.ai, Virtual.

07/2022 "Tiny Machine Learning: Challenges and Opportunities." Design Automation Conference, ROAD4NN Workshop, Virtual. (Keynote).

07/2022 "The Future of Smart Cities is Tiny and Bright." ACM International Conference on Future Energy Systems (ACM e-Energy), Virtual. (Keynote).

06/2022 "Machine Learning Metrics." HiPEAC AccML Workshop.

05/2022 "Democratizing TinyML." Rutgers Efficient AI (REFAI) Seminar, Virtual.

04/2022 "Tiny Machine Learning (TinyML) for Domain-Specific Systems." International Workshop on Domain Specific System Architecture (DOSSA-4), Virtual.

03/2022 "IoT 2.0: The Era of Intelligence on Things." Design Automation and Test in Europe, Virtual.

01/2022 "Tiny Machine Learning." Accelerated Machine Learning Workshop, co-located with HiPEAC 2022, Virtual.

12/2021 "Machine Learning's Future is Tiny & Bright." AICON GWANGJU, Virtual. (Keynote).

12/2021 "Democratizing TinyML." Globecom Workshop on Sustainable Environmental Sensing Systems, Virtual. (Keynote).

11/2021 "Tiny Machine Learning." EdukCircle International Convention on Engineering and Computer Technology, Virtual.

11/2021 "Tiny Machine Learning (TinyML) for Robotics." Conference on Robot Learning, Virtual. (Keynote).

11/2021 "Democratizing TinyML: Generalization, Standardization and Automation." Workshop on Hardware and Algorithms for Learning On-a-chip (HALO) workshop, ICCAD conference, Virtual. (Keynote).

11/2021 "Democratizing TinyML: Generalization, Standardization and Automation." Multi-DNN Workshop, co-located with MICRO, Virtual. (Keynote).

11/2021 "Data for TinyML." Data for AI Summit @Google (internal), Virtual. (Keynote).

10/2021 "Widening Access to Applied Machine Learning with TinyML." IEEE Global Humanitarian Technology Conference, Virtual. (Keynote).

10/2021 "The Vision Behind MLPerf." SiFive Engineering Forum, Virtual.

10/2021 "The Vision Behind MLPerf." Samsung AI Cambridge, Santa Clara.

10/2021 "Democratizing TinyML." MICRO 2021 Workshop on Systems for Multi-DNN Workloads, Virtual.

(Keynote).

09/2021 "The Vision Behind MLPerf." Tensorrent, Virtual.

07/2021 "Tiny Machine Learning." Workshop on Artificial Intelligence, Machine Learning, & Computational Intelligence, Virtual.

07/2021 "The Vision Behind MLPerf." STMicroelectronics, Virtual.

03/2021 "tinyMLPerf: Benchmarking Ultra-low-power Systems." Tiny Machine Learning Summit, San Francisco.

03/2021 "tinyMLPerf: Benchmarking Ultra-low-power Systems." "Machine Learning at the Edge," Workshop co-located with Design Automation Conference.

06/2020 "The Vision Behind MLPerf." AMD Tech Talk, Austin.

03/2020 "tinyMLPerf: Benchmarking Ultra-low-power Systems." Tiny Machine Learning Summit.

02/2020 "The Vision Behind MLPerf." International Solid-State Circuits Conference (ISSCC), San Francisco.

02/2020 "MLPerf Inference." Machine Learning Systems Workshop, Santa Clara.

09/2019 "The Vision Behind MLPerf: A Broad ML Benchmark Suite for Measuring the Performance of ML Software Frameworks, ML Hardware Accelerators in Cloud and Edge Computing." Taiwan Semiconductor Manufacturing Company (TSMC).

09/2019 "The Vision Behind MLPerf: A Broad ML Benchmark Suite for Measuring the Performance of ML Software Frameworks, ML Hardware Accelerators in Cloud and Edge Computing." Taiwan AI Labs. (Keynote).

09/2019 "The Vision Behind MLPerf: A Broad ML Benchmark Suite for Measuring the Performance of ML Software Frameworks, ML Hardware Accelerators in Cloud and Edge Computing." Synopsis SNUG, Taiwan.

04/2019 "Ten Commandments for Mobile Computer Architecture." Workshop on Infrastructure and Methodology for SoC-level Performance and Power Modeling, co-located with ASPLOS.

03/2019 "The Vision Behind MLPerf (mlperf.org)." Intel VSSAD.

03/2019 "Evaluating Resiliency in End-to-end Learning for Autonomous Machines." The 15th Workshop on Silicon Errors in Logic – System Effects.

03/2019 "Autonomous Aerial Computing Machines." International Workshop on Performance Analysis of Machine Learning Systems.

12/2018 "The Vision Behind MLPerf: A Broad ML Benchmark Suite for Measuring the Performance of ML Software Frameworks, ML Hardware Accelerators, and ML Cloud and Edge Platforms." IEEE BigBench co-located with the IEEE Big Data Conference. (Keynote).

12/2018 "The Vision Behind MLPerf: A Broad ML Benchmark Suite for Measuring the Performance of ML Software Frameworks, ML Hardware Accelerators in Cloud and Edge Computing." The Forum of Turing Centers, Shanghai Jiao Tong University. (Keynote).

12/2018 "The Vision Behind MLPerf: A Broad ML Benchmark Suite for Measuring the Performance of ML Software Frameworks, ML Hardware Accelerators in Cloud and Edge Computing." Boston Area Computer Architecture Workshop. (Keynote).

11/2018 "The Vision Behind MLPerf: A Broad ML Benchmark Suite for Measuring the Performance of ML Software Frameworks, ML Hardware Accelerators, and ML Cloud and Edge Platforms." Samsung Advanced Computing Lab (ACL).

10/2018 "The Vision Behind MLPerf: A Broad ML Benchmark Suite for Measuring the Performance of ML

Software Frameworks, ML Hardware Accelerators, and ML Cloud and Edge Platforms." Samsung Austin Research Center (SARC).

10/2018 "Mobile Robotics for Computer Architects." First Annual Workshop on Domain Specific System Architecture co-located with International Symposium on Microarchitecture (MICRO). (Keynote).

04/2018 "Aerial Computing: Challenges and Opportunities for Hardware and Software Architects Designing Flying Systems." IBM T. J. Watson.

03/2018 "Architecting for Big Data Analytics: Think Dubai rather than Venice." Workshop on BigData Benchmarks, Performance, Optimization and Emerging Hardware (co-located with ASPLOS).

02/2017 "Architecture Support for Scripting from Mobile to Cloud." Stanford University, Palo Alto.

05/2016 "Watt-Wise-Web://Architecting for Responsiveness and Energy-Efficiency." The University of Chicago, Chicago–IL.

05/2016 "Mobile CPU Evolution: The Past, the Present, and the Future." Rice University – TexasWISE Keynote, Houston.

05/2016 "Microarchitectural Implications of Event-driven Programming." Northwestern, Chicago–IL.

05/2016 "Microarchitectural Implications of Event-driven Programming." Intel Santa Clara–CA.

05/2016 "Microarchitectural Implications of Event-driven Programming." AMD Austin–TX.

03/2016 "Programming the Web of Things: Why Architects Should Care." Sensors to Cloud Architectures Workshop, Barcelona. (Keynote).

02/2016 "From Moore's Law to Moore's Crawl: Architecting the Next-Generation of Mobile Computing Devices." University of Washington, Seattle–WA.

02/2016 "From Moore's Law to Moore's Crawl: Architecting the Next-Generation of Mobile Computing Devices." National Academy of Engineering (NAE) Annual Event, Irvine–CA.

12/2015 "Programming the Web of Things." Workshop on Internet of Things (IoT) held in conjunction with International Symposium on Microarchitecture, Hawaii.

12/2015 "End of the Road for My CAREER." Workshop on Negative Outcomes, Post-mortems, and Experiences (NOPE) held in conjunction with International Symposium on Microarchitecture, Hawaii.

11/2015 "Watt-Wise Web: Architecting for a Responsive and Energy-Efficient Mobile Web." Texas A&M University.

10/2015 "Watt-Wise Web: Architecting for a Responsive and Energy-Efficient Mobile Web." Google Faculty Summit.

10/2015 "Watt-Wise Web: Architecting for a Responsive and Energy-Efficient Mobile Web." Georgia Tech University.

09/2015 "What Users Want and What Hardware Provides: Bridging the Gap Between User Quality of Experience (QoE) and Mobile Device Trends." National Taiwan University, Taiwan.

09/2015 "What Users Want and What Hardware Provides: Bridging the Gap Between User Quality of Experience (QoE) and Mobile Device Trends." Mediatek, Taiwan.

09/2015 "What Users Want and What Hardware Provides: Bridging the Gap Between User Quality of Experience (QoE) and Mobile Device Trends." Academia Sinica, Taiwan.

09/2015 "Mobile CPU Evolution: The Past, the Present, and the Future." Taiwan Application Processor Union – Mobile SoC Summer Course, Taiwan.

06/2015 "What Users Want and What Hardware Provides: Bridging the Gap Between User Quality of Experience (QoE) and Mobile Device Trends." Duke University, Raleigh–NC.

06/2015 "GPU Voltage Guardband Management to Achieve Exascale Energy-Efficiency." AMD Austin–TX.

05/2015 "Voltage Noise in Multicore Processors." Intel, Portland.

05/2015 "GPU Voltage Guardband Management to Achieve Exascale Energy-Efficiency." Intel, Portland.

04/2015 "What Users Want and What Hardware Provides: Bridging the Gap Between User Quality of Experience (QoE) and Mobile Device Trends." Qualcomm, Raleigh–NC.

04/2015 "Mobile CPU Evolution: The Past, the Present, and the Future." Microsoft, Seattle–WA.

03/2015 "What Users Want and What Hardware Provides: Bridging the Gap Between User Quality of Experience (QoE) and Mobile Device Trends." Facebook, Menlo Park–CA.

02/2015 "Mobile CPU Evolution: The Past, the Present, and the Future." Intel Santa Clara–CA.

11/2014 "Watt-Wise Web: Architecting for a Responsive and Energy-Efficient Mobile Web." Univ. of Michigan.

06/2014 "Simulators are Perfect, Authors are Oracles, Users are Innocent." Workshop on Duplicating, Deconstructing and Debunking (WDDD) held in conjunction with International Symposium on Computer Architecture.

06/2014 "Architecting for the Mobile Web: Where We've Been, Where We're Heading, and What We Need to Address." Parallelism in Mobile Platforms (PRISM) held in conjunction with International Symposium on Computer Architecture.

05/2014 "Mobile Processor Architectures: Design Implications and Challenges for Energy Efficiency." Indo-American Frontiers of Engineering (IAFOE), Mysore–India.

05/2014 "Hardware and Software Co-Design for Robust and Resilient Execution." International Conference on Integrated Circuit Design and Technology (ICICDT), Austin.

03/2014 "Architectural Support for the Interactive Mobile Web." Samsung Austin–TX.

03/2014 "Architectural Support for the Interactive Mobile Web." ARM Austin–TX.

02/2014 "Robust and Resilient Systems from the Bottom-Up: Circuits, Architecture and Software Integration." ISSCC Forum, San Francisco–CA.

02/2014 "Architectural Support for the Interactive Mobile Web." Intel Austin–TX.

02/2013 "Toward High-Performance and Energy-Efficient Mobile Web Browsing." Qualcomm, Santa Clara–MA.

08/2012 "Toward High-Performance and Energy-Efficient Mobile Web Browsing." Intel Austin–TX.

08/2012 "Toward High-Performance and Energy-Efficient Mobile Web Browsing." AMD Austin–TX.

10/2010 "Web Search Using Small Cores." AMD, Boxborough–MA.

07/2010 "Web Search Using Small Cores." SeaMicro Santa Clara–CA.

07/2010 "Web Search Using Small Cores." Intel Hudson–MA.

07/2010 "Web Search Using Small Cores." IBM T. J. Watson Labs, Hawthorne–NY.

07/2010 "Web Search Using Small Cores." HP Labs, Palo Alto–CA.

07/2010 "Web Search Using Small Cores." Google, Palo Alto–CA.

07/2010 "Web Search Using Small Cores." Facebook, Palo Alto–CA.

07/2010 "Software-Assisted Hardware Reliability." Intel, Portland.

07/2010 "Software-Assisted Hardware Reliability." IBM T. J. Watson Labs, Yorktown–NY.

06/2010 "Web Search Using Small Cores." Amazon, Seattle–WA.

06/2010 "Software-Assisted Hardware Reliability." Microsoft Research, Redmond–WA.

03/2010 "Software-Assisted Hardware Reliability." Intel Santa Clara–CA.

03/2010 "Software-Assisted Hardware Reliability." AMD Austin–TX.

03/2007 "Persistent Code Caching." Intel Santa Clara–CA.

Books

2027 V. Janapa Reddi. *Machine Learning Systems at Scale*. MIT Press. *Forthcoming.* (Volume II of the MLSysBook series.)

2026 V. Janapa Reddi. *Introduction to Machine Learning Systems*. MIT Press. *Forthcoming.* (Volume I of the MLSysBook series; companion site mlsysbook.ai.)

2013 V. Janapa Reddi and Meeta Sharma Gupta. *Resilient Architecture Design for Voltage Variation*. Synthesis Lectures on Computer Architecture. Morgan & Claypool Publishers. ISBN: 9781608456376.

Technical Reports

2010 V. Janapa Reddi, B. Lee, T. Chilimbi, K. Vaid. "Web Search Using Small Cores: Quantifying the Price of Efficiency," in Microsoft Research Tech. Report, June.

Theses

2010 V. Janapa Reddi. "Software-Assisted Hardware Reliability: Enabling Aggressive Timing Speculation Using Run-Time Feedback from Hardware and Software," Ph.D. Thesis, School of Engineering and Applied Sciences, Harvard University.

2005 V. Janapa Reddi. "Deploying Dynamic Code Transformation in Modern Computing Environments," M.S. Thesis, Department of Electrical and Computer Engineering, University of Colorado.

2003 V. Janapa Reddi. "Heterogeneous Networks of Workstations Across Wide Area Networks," B.S. Thesis, Department of Electrical and Computer Engineering, Santa Clara University.

Patents

2005 R. Cohn, T. Moseley, and V. Janapa Reddi. "System and method to instrument references to shared memory." U.S. Patent Application 11/143,130, filed June 1, 2005.

2012 N. Kim, J. O'Connor, M. Schulte, and V. Janapa Reddi. "Method and apparatus for power reduction during lane divergence." U.S. Patent Application 13/605,460, filed September 6, 2012.

2015 V. Janapa Reddi, M. Gupta, G. Holloway, G. Wei, M. D. Smith, and D. Brooks. "Adaptive event-guided system and method for avoiding voltage emergencies." U.S. Patent 8,949,666, issued February 3, 2015.

Teaching

Harvard University

Sp' 2026 ES 50: Introduction to Electrical and Computer Engineering

Sp' 2025 ES 50: Introduction to Electrical and Computer Engineering

Sp' 2024 COMPSCI 141: Computing Hardware

Fa' 2023 COMPSCI 249R: Tiny Machine Learning
Sp' 2023 COMPSCI 141: Computing Hardware
Fa' 2022 COMPSCI 249R: Tiny Machine Learning
Sp' 2021 COMPSCI 141: Computing Hardware
Fa' 2020 COMPSCI 249R: Tiny Machine Learning
Fa' 2019 COMPSCI 249R: Autonomous Machines

The University of Texas at Austin

Sp' 2016 EE 319K: Introduction to Embedded Systems
Sp' 2015 EE 319K: Introduction to Embedded Systems
Fa' 2014 EE 382V: Code Generation and Optimization
Sp' 2014 EE 319K: Introduction to Embedded Systems
Fa' 2013 EE 382V: Code Generation and Optimization
Sp' 2013 EE 319K: Introduction to Embedded Systems
Fa' 2012 EE 382V: Dynamic Compilation
Sp' 2012 EE 382V: Code Generation and Optimization
Fa' 2011 EE 382V: Dynamic Compilation

Students

Current PhD students

2025/– Chenyu Wang
2024/– Zander Ingare
2023/– Jeffrey Ma
2023/– Oishii Banerjee
2022/– Ikechukwu Uchendu
2022/– Jason Jabbour
2022/– Jason Yik
2022/– Jessica Quaye
2021/– Emma Chen
2021/– Mark Mazumder
2020/– Shvetank Prakash
2019/– Radhika Ghosal

Graduated PhD students

2019–2024 **Max Lam**
PhD Thesis: "Systems and Algorithms for Efficient, Secure and Private Machine Learning Inference,"
First Job: Research Scientist, Apple.

2019–2024 **Colby Banbury**
PhD Thesis: "Efficient and Scalable Tiny Machine Learning,"

First Job: Senior Research Scientist, Microsoft Research.

2019–2024 **Yu-shun Hsiao**
PhD Thesis: "Safety-Aware System Optimization for Autonomous Machines,"
First Job: CEO, Founder, Robotics Start-up.

2018–2024 **Srivatsan Krishnan**
PhD Thesis: "Designing Efficient Domain-Specific Architectures for Autonomous Systems,"
First Job: Senior Software Engineer, NVIDIA.

2018–2022 **Brian Plancher**
PhD Thesis: "GPU Acceleration for Real-time, Whole-body, Nonlinear Model Predictive Control,"
First Job: Assistant Professor at Barnard College at University of Columbia.

2014–2022 **Behzad Boroujerdian**
PhD Thesis: "Agile Development of Domain-Specific Solutions for Emerging Mobile Systems,"
First Job: Deep Learning Researcher, NVIDIA.

2014–2022 **Daniel Richins**
PhD Thesis: "Bottlenecks in Big PhD Thesis: Data Analytics and AI Applications and Opportunities for Improvement,"
First Job: Instructor, Brigham Young University.

2013–2018 **Yazhou Zu**
PhD Thesis: "Active Timing Margin Management to Improve Microprocessor Power Efficiency,"
First Job: Software Engineer, Google.

2011–2016 **Jingwen Leng**
PhD Thesis: "Guardband Management in Heterogeneous Architectures,"
First Job: Assistant Professor at Shanghai Jiao Tong University (CSE).

2011–2016 **Yuhao Zhu**
PhD Thesis: "Energy-Efficient Mobile Web Computing,"
First Job: Assistant Professor at Univ. of Rochester (CS).

Postdoctoral Associates

2020–2024 Matthew Stewart

2020–2023 Sabrina Neuman, Assistant Professor at Boston University.

M.S. students

2019–2021 Jonathan Cruz, Harvard

2015–2017 Wenzhi Cui, UT Austin

2014–2017 Matthew Halpern, UT Austin

2011–2013 Aditya Srikanth, UT Austin

2011–2013 Ankita Garg, UT Austin

255 publications · 104 conference · 51 journal · 94 arXiv · 4 workshop

Bio

Vijay Janapa Reddi is the Gordon McKay Professor of Electrical & Computer Engineering at Harvard University, Vice President and co-founder of MLCommons (mlcommons.org), and (from July 2026) Visiting Professor at ETH Zürich. As Vice President of MLCommons, he oversees the MLCommons Research organization and serves on the Board of Directors. He co-led the development of the MLPerf Inference benchmark—now the industry standard for ML benchmarking across datacenter, edge, mobile, and IoT, adopted by Google, Microsoft, NVIDIA, Meta, AMD, and Intel—and the MLPerf Power benchmark for sustainable AI. Before joining Harvard, he was an Associate Professor at the University of Texas at Austin's Department of Electrical and Computer Engineering. Drawing on his expertise in computer architecture, runtime systems, and applied machine learning, he creates innovative solutions at the intersection of mobile, edge, and embedded intelligence.

Dr. Janapa Reddi has earned numerous awards, including the Gilbreth Lectureship Honor from the National Academy of Engineering (NAE) in 2016, the IEEE TCCA Young Computer Architect Award (2016), and the Intel Early Career Award (2013). His research has been recognized with Best Paper Awards at the 2024 Vail Computer Elements Workshop (VCEW), the 2020 Design Automation Conference (DAC), the 2009 International Symposium on High-Performance Computer Architecture (HPCA), and the 2005 International Symposium on Microarchitecture (MICRO). His papers have been selected for IEEE Micro Top Picks in Computer Architecture in 2026, 2025, 2021, 2017, 2011, and 2010, with honorable mentions in 2023, 2022, and 2016, and the MLPerf Inference benchmark was selected for inclusion in the ISCA@50 25-Year Retrospective. He is a recipient of multiple Google Faculty Research Awards (2012, 2013, 2015, 2017, and 2020), the ACM SIGPLAN Programming Languages Software Award (2020), and the BenchCouncil Rising Star Award (2021). He is inducted into the MICRO Hall of Fame (2018) and the HPCA Hall of Fame (2019).

Dr. Janapa Reddi is strongly devoted to expanding access to applied machine learning for STEM, diversity, and the application of AI for social good. He is the founder and lead author of the open-source *Machine Learning Systems* textbook and curriculum ecosystem (mlsysbook.ai), used at 50+ universities across five continents, with two MIT Press hardcover volumes forthcoming in 2026 and 2027. Through the TinyML4D Academic Network (with ICTP), he ships hardware kits and curriculum to universities across Latin America, Africa, and Asia. He also founded and led the TinyML Professional Certificate on HarvardX/edX in partnership with Google and the tinyML Foundation, reaching 100,000+ learners worldwide and recognized as one of ClassCentral's 100 Most Popular Free Online Courses (2021) and the CogX Best Course in AI award (2021). At UT Austin, he previously led the Hands-on Computer Science (HaCS) program for K–12 students in partnership with the Austin Independent School District.

Dr. Janapa Reddi holds degrees in computer science from Harvard University (Ph.D.), electrical and computer engineering from the University of Colorado at Boulder (M.S.), and computer engineering from Santa Clara University (B.S.). His passion is helping individuals and teams succeed while making the world a better place.

Last updated: May 21, 2026