2022 IEEE International Parallel and Distributed Processing Symposium
Workshops (IPDPSW)

1ST IEEE International Workshop on

Electronic Design Automation and Machine Learning (EDAML)

June 3, 2022

Virtual Workshop

Machine Learning (ML) has evolved substantially over the past decade and is now an integral component of many applications such as classification and object detection in images and video, speech recognition and language translation, data mining and pattern recognition, and cyber security. However, the design of server, edge, and embedded computing platforms to support these ML and other emerging applications remains a significant challenge. In particular, the design of ML-inspired electronic design automation (EDA) tools to efficiently prototype these platforms, and, conversely, the design of platforms to accelerate ML training and inference in EDA tools, represents an interdependent and important problem. This workshop aims to explore the intersection of ML and EDA and define a roadmap to realize the next generation of parallel and distributed computing systems. This year's program will feature a keynote talk by Prof. David Pan from The University of Texas at Austin and ten invited talks from experts in the field of EDA and ML. We thank the keynote and invited speakers for contributing to a high-quality technical program for this inaugural edition of the EDAML workshop.

Workshop General Chairs:
Sudeep Pasricha, Colorado State University
Muhammad Shafique, New York University, Abu Dhabi

Keynote Speaker
David Z. Pan, The University of Texas at Austin

Invited Speakers:
Ankush Sood, Cadence Design Systems
Deming Chen, University of Illinois at Urbana-Champaign
R. Iris Bahar, Colorado School of Mines
Krishnendu Chakrabarty, Duke University
Laleh Behjat, University of Calgary, Alberta
Muhammad Shafique, New York University, Abu Dhabi
Partha Pande, Washington State University
Sachin S. Sapatnekar, University of Minnesota
Sheldon Tan, University of California at Riverside
Sudeep Pasricha, Colorado State University

Program Schedule
June 3, 2022

Time (ET)	Schedule
9:55-10:00	**Welcome and Opening Remarks** Sudeep Pasricha, Colorado State University Muhammad Shafique, New York University (NYU) Abu Dhabi, UAE
10:00-10:35	**Keynote: Machine Learning for Agile, Intelligent and Open-Source EDA** David Z. Pan, University of Texas at Austin
10:35-10:40	**Break**
10:40-11:00	**Analog and Digital Circuit and Layout Optimization using Machine Learning** Sachin S. Sapatnekar, University of Minnesota
11:00-11:20	**Application of Machine Learning in High Level Synthesis** Ankush Sood, Cadence Design Systems
11:20-11:40	**Combining Optimization and Machine Learning in Physical Design** Laleh Behjat, University of Calgary, Alberta, Canada
11:40-12:00	**Thermal and Power Monitoring and Estimation for Commercial Multicore Processors -- A Machine Learning Perspective** Sheldon Tan, University of California at Riverside
12:00-12:10	**Break**
12:10-12:30	**Fault Criticality Assessment in AI Accelerators** Krishnendu Chakrabarty, Duke University
12:30-12:50	**Scalable ML Architectures for Real-time Energy-efficient Computing** R. Iris Bahar, Colorado School of Mines
12:50-1:10	**Hardware/Software Codesign for Optical Deep Learning Accelerators** Sudeep Pasricha, Colorado State University
1:10-1:20	**Break**
1:20-1:40	**Reliable Processing-in-Memory based Manycore Architectures for Deep Learning: From CNNs to GNNs** Partha Pratim Pande, Washington State University
1:40-2:00	**AI Algorithm and Accelerator Co-design for Computing on the Edge** Deming Chen, University of Illinois at Urbana-Champaign
2:00-2:20	**Dependability and Security of Advanced Machine Learning Systems: Hardware and Software Techniques** Muhammad Shafique, New York University (NYU) Abu Dhabi, UAE
2:20-2:40	**Closing Remarks and Open Forum on Future Directions** All speakers and attendees

EDAML 2022 Keynote Speaker

Machine Learning for Agile, Intelligent and Open-Source EDA

David Z. Pan
Department of Electrical and Computer Engineering
The University of Texas at Austin, TX 78712

Abstract

This talk will present some recent results and trends toward agile, intelligent and open-source design automation for digital/analog/mixed-signal ICs, in particular leveraging AI/machine learning with domain-specific customizations. Placement is a fundamental EDA problem. I will first show how we leverage deep learning hardware and software to develop an open-source VLSI placement engine, DREAMPlace [DAC'19 and TCAD'21 Best Paper Awards], which is around 40x faster than the previous state-of-the-art academic global placer. DREAMPlace has been further developed to tackle detailed placement acceleration and region constraints, and used together with reinforcement learning for macro placement to achieve superhuman results. I will then present the DARPA-funded project MAGICAL which leverages both machine and human intelligence to produce fully automated analog layout from netlists to GDSII, including automatic layout constraint generation, placement, and routing. MAGICAL 1.0 has been open-sourced, and validated with silicon tape-outs. The talk will conclude with some future research directions.

Biography

David Z. Pan is a Professor and Silicon Laboratories Endowed Chair at the Department of Electrical and Computer Engineering, The University of Texas at Austin. His research interests include bidirectional AI and IC interactions, electronic design automation, design for manufacturing, hardware security, and CAD for analog/mixed-signal ICs and emerging technologies. He has published over 420 refereed journal/conference papers and 8 US patents. He has served in many journal editorial boards and conference committees, including various leadership roles such as DAC 2022 Panel Chair, ICCAD 2019 General Chair, ASP-DAC 2017 TPC Chair, and ISPD 2008 General Chair. He has received many awards, including SRC Technical Excellence Award, 20 Best Paper Awards (from TCAD, DAC, ICCAD, DATE, ASP-DAC, ISPD, HOST, etc.), DAC Top 10 Author Award in Fifth Decade, ASP-DAC Frequently Cited Author Award, ACM/SIGDA Outstanding New Faculty Award, NSF CAREER Award, IBM Faculty Award (4 times), and many international CAD contest awards. He has graduated 41 PhD students and postdocs who have won many awards, including the First Place of ACM Student Research Competition Grand Finals (twice, in 2018 and 2021), ACM/SIGDA Student Research Competition Gold Medal (thrice), ACM Outstanding PhD Dissertation in EDA Award (twice), EDAA Outstanding Dissertation Award (thrice), etc. He is a Fellow of ACM, IEEE and SPIE.

EDAML 2022 Invited Speaker

Application of Machine Learning in High Level Synthesis

Ankush Sood
Cadence Design Systems

Abstract

Traditionally high-level optimizations like CSA and sharing have been done technology independent. Doing word level optimizations PPA aware require accurate models for power, timing and area and that needs mapping to technology library and iterations which are runtime intensive and not suitable for multi-million instance designs. In this talk, we discuss how machine learning could help make the right power/area/delay tradeoffs early in the synthesis flow not sacrificing on turnaround time.

Biography

Ankush Graduated with a BTech in Electrical Engineering and a MTech in Microelectronics from IIT Bombay. He has been working with Cadence Design Systems for the past 18 years focused on synthesis. He is currently a Synthesis Fellow leading the frontend and infrastructure teams for Cadence Digital Implementation. Over the years he has led all aspects of synthesis tool development. His current areas of interests are High Level Optimizations, distributed synthesis, Machine Learning and early physical and congestion aware synthesis.

EDAML 2022 Invited Speaker

AI Algorithm and Accelerator Co-design for Computing on the Edge

Deming Chen
University of Illinois at Urbana-Champaign

Abstract

In a conventional top-down design flow, deep-learning algorithms are first designed concentrating on the model accuracy, and then accelerated through hardware accelerators trying to meet various system design targets on power, energy, speed, and cost. However, this approach often does not work well because it ignores the physical constraints that the hardware architectures themselves would have towards the deep neural network (DNN) algorithm design and deployment, especially for the DNNs that will be deployed unto edge devices. Thus, an ideal scenario is that algorithms and their hardware accelerators are developed simultaneously. In this talk, we will present our DNN/Accelerator co-design and co-search methods. Our results have shown great promises for delivering high-performance hardware-tailored DNNs and DNN-tailored accelerators naturally and elegantly. One of the DNN models coming out of this co-design method, called SkyNet, won a double championship in the competitive DAC System Design Contest for both the GPU and the FPGA tracks for low-power object detection.

Biography

Dr. Deming Chen obtained his BS in computer science from University of Pittsburgh, Pennsylvania in 1995, and his MS and PhD in computer science from University of California at Los Angeles in 2001 and 2005 respectively. He joined the ECE department of University of Illinois at Urbana-Champaign in 2005. His current research interests include reconfigurable computing, hybrid cloud, machine learning and cognitive computing, system-level and high-level synthesis, and hardware security. He has given more than 130 invited talks sharing these research results worldwide. He has received 10 Best Paper Awards, a TCFPGA Hall-of-Fame paper award, a few Best Poster Awards, and numerous other research and service-related awards. He is the Abel Bliss Professor of the Grainger College of Engineering, an IEEE Fellow, an ACM Distinguished Speaker, and the Editor-in-Chief of ACM Transactions on Reconfigurable Technology and Systems (TRETS).

EDAML 2022 Invited Speaker

Scalable ML Architectures for Real-time Energy-efficient Computing

R. Iris Bahar
Colorado School of Mines

Abstract

Technological advancements have led to a proliferation machine learning systems to assist humans in a wide range of tasks. However, we are still far from accurate, reliable, and resource-efficient operations for many of these systems. Despite the strengths of convolutional neural networks (CNNs) for object recognition, these discriminative techniques have several shortcomings that leave them vulnerable to exploitation from adversaries. In addition, the computational cost incurred to train these discriminative models can be quite significant. Discriminative-generative approaches offers a promising avenue for robust perception and action. Such methods combine inference by deep learning with sampling and probabilistic inference models to achieve robust and adaptive understanding. In this talk, I will present our work on implementing a scalable, computationally efficient generative inference algorithm in hardware that can achieve real-time results in an energy efficient manner. I will also discuss future directions in designing scalable and efficient ML algorithms in hardware more broadly.

Biography

R. Iris Bahar received the B.S. and M.S. degrees in computer engineering from the University of Illinois, Urbana-Champaign, and the Ph.D. degree in electrical and computer engineering from the University of Colorado, Boulder. She recently joined the faculty at the Colorado School of Mines in January 2022 and serves at Department Head of Computer Science. Before joining Mines, she was on the faculty at Brown University since 1996 and held dual appointments as Professor of Engineering and Professor of Computer Science. Her research interests focus on energy-efficient and reliable computing, from the system level to device level. She is the 2019 recipient of the Marie R. Pistilli Women in Engineering Achievement Award and the Brown University School of Engineering Award for Excellence in Teaching in Engineering. She is an IEEE fellow and an ACM Distinguished Scientist.

EDAML 2022 Invited Speaker

Fault Criticality Assessment in AI Accelerators

Krishnendu Chakrabarty
Duke University

Abstract

The ubiquitous application of deep neural networks (DNN) has led to a rise in demand for AI accelerators. DNN-specific functional criticality analysis identifies faults that cause measurable and significant deviations from acceptable requirements such as the inferencing accuracy. This talk will examine the problem of classifying structural faults in the processing elements (PEs) of systolic-array accelerators. The speaker will first present a two-tier machine-learning (ML) based method to assess the functional criticality of faults. The problem of minimizing misclassification will be addressed by utilizing generative adversarial networks (GANs). The two- tier ML/GAN-based criticality assessment method leads to less than 1% test escapes during functional criticality evaluation of structural faults. While supervised learning techniques can be used to accurately estimate fault criticality, it requires a considerable amount of ground truth for model training. The speaker will therefore present a neural-twin framework for analyzing fault criticality with a negligible amount of ground-truth data. A recently proposed misclassification-driven training algorithm will be used to sensitize and identify biases that are critical to the functioning of the accelerator for a given application workload. The proposed framework achieves up to 100% accuracy in fault-criticality classification in 16-bit and 32-bit PEs by using the criticality knowledge of only 2% of total faults in a PE.

Biography

Krishnendu Chakrabarty received the B. Tech. degree from the Indian Institute of Technology, Kharagpur, in 1990, and the M.S.E. and Ph.D. degrees from the University of Michigan, Ann Arbor, in 1992 and 1995, respectively. He is now the John Cocke Distinguished Professor of Electrical and Computer Engineering at Duke University. Prof. Chakrabarty is a recipient of the National Science Foundation CAREER award, the Office of Naval Research Young Investigator award, the Humboldt Research Award from the Alexander von Humboldt Foundation, Germany, the IEEE TCAD Donald O. Pederson Best Paper Award (2015), the IEEE TVLSI Systems Prize Paper Award (2021), the ACM TODAES Best Paper Award (2017), multiple IBM Faculty Awards and HP Labs Open Innovation Research Awards, and over a dozen best paper awards at major conferences. He is also a recipient of the IEEE Computer Society Technical Achievement Award (2015), the IEEE Circuits and Systems Society Charles A. Desoer Technical Achievement Award (2017), the IEEE Circuits and Systems Society Vitold Belevitch Award (2021), the Semiconductor Research Corporation Technical Excellence Award (2018), the IEEE-HKN Asad M. Madni Outstanding Technical Achievement and Excellence Award (2021), and the IEEE Test Technology Technical Council Bob Madge Innovation Award (2018). He is a 2018 recipient of the Japan Society for the Promotion of Science (JSPS) Invitational Fellowship in the “Short Term S: Nobel Prize Level” category. Prof. Chakrabarty’s current research projects include: design-for-testability of 3D integrated circuits; AI accelerators; microfluidic biochips; hardware security; AI for healthcare; neuromorphic computing systems. He is a Fellow of ACM, IEEE, and AAAS, and a Golden Core Member of the IEEE Computer Society.

EDAML 2022 Invited Speaker

Combining Optimization and Machine Learning in Physical Design

Laleh Behjat
University of Calgary, Alberta, Canada

Abstract

The exponential increase in computing power and the availability of big data have ignited innovations in EDA. The most recent trend in innovations has involved using machine learning algorithms for solving problems of scale. Machine learning techniques can solve large-scale problems efficiently once they are trained. However, their training takes a large amount of computing power and might not translate well from one type of problem to another. On the other hand, many of the existing algorithms in physical design take advantage of mathematical optimization techniques to improve their solution quality. These techniques can find optimal or near-optimal solutions using fast heuristics. These techniques do not require a large amount of data but need some level of insight into the nature of the problem by the designer. The mathematical optimization techniques rely heavily on the developed models. In this talk, we will discuss how machine learning can be used to develop better models for optimization problems and how optimization techniques can then use the models to generate more data to improve the accuracy and robustness of machine learning techniques. We will first discuss the algorithm-driven nature of the optimization techniques and compare that to the data-driven nature of the machine learning techniques. We will use examples of physical design placement and routing. Then, we will discuss how optimization and ML can be used to solve the problems of scale both in numbers and transistor sizes. We will also discuss how reinforcement learning can be used to come up with new heuristics for solving the problems encountered in physical design. The talk will end with some practical suggestions on how to improve the quality and speed of the design.

Biography

Laleh Behjat is a professor in the Department of Electrical and Software Engineering at the University of Calgary and the Natural Sciences and Engineering Council of Canada Chair for Women in Science and Engineering (Prairie Region). She received her Ph.D. from the University of Waterloo in Canada in 2002. Her research interests include developing mathematical programming and optimization techniques for solving large-scale optimization problems especially related to the physical design of integrated circuits. She has won several awards for her research, including first place in the ISPD placement contest in 2014 and second place in 2015. She was the general chair for GLSVLSI and ISPD. She is also an associate editor for the IEEE Transactions on Computer-Aided Design and ACM Transactions on Design and Automation of Electronic Circuits. Dr. Behjat's other research interests include design, innovation, creativity and their relationship with diversity. She has conducted several workshops related to diversity and creativity. In 2015, she won third place in Design Automation Perspective Challenge. She believes that interdisciplinary research is the key to performing innovative research. Dr. Behjat has won several teaching awards, especially for teaching graduate courses. She has also taught optimization in international summer schools and has presented her work in applying optimization techniques to CAD problems in the prestigious Fields Institute for Mathematics and Banff International Research Station.

EDAML 2022 Invited Speaker

Reliable Processing-in-Memory based Manycore Architectures for Deep Learning: From CNNs to GNNs

Partha Pratim Pande
Washington State University

Abstract

Resistive random-access memory (ReRAM)-based processing-in-memory (PIM) architectures have recently become a popular architectural choice for deep-learning applications. ReRAM-based architectures can accelerate inferencing and training of deep learning algorithms and are more energy efficient compared to traditional GPUs. However, these architectures have various limitations that affect the model accuracy and performance. Moreover, the choice of the deep-learning application also imposes new design challenges that must be addressed to achieve high performance. In this talk, we present the advantages and challenges associated with ReRAM-based PIM architectures by considering Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs) as important application domains. We also outline methods that can be used to address these challenges.

Biography

Partha Pratim Pande (M'05-SM'11-F'20) is a professor and holder of the Boeing Centennial Chair in computer engineering at the school of Electrical Engineering and Computer Science, Washington State University, Pullman, USA. He is currently the director of the school. His current research interests are novel interconnect architectures for manycore chips, on-chip wireless communication networks, and heterogeneous architectures. Dr. Pande currently serves as the Editor-in-Chief (EIC) of IEEE Design and Test (D&T). He is on the editorial boards of IEEE Transactions on VLSI (TVLSI) and ACM Journal of Emerging Technologies in Computing Systems (JETC) and IEEE Embedded Systems letters. He was/is the technical program committee chair of IEEE/ACM Network-on-Chip Symposium 2015 and CASES (2019-2020).

EDAML 2022 Invited Speaker

Analog and Digital Circuit and Layout Optimization using Machine Learning

Sachin S. Sapatnekar
University of Minnesota

Abstract

Machine learning (ML) has opened new doors for a wide range of techniques that can increase the efficiency of the design cycle -- ranging from methods that traditionally require manual input from expert designers to methods that perform fast analyses and optimizations. This talk will overview some of our recent research on applying ML methods to problems in electronic design automation (EDA). We will describe methods that are used for circuit recognition, power grid analysis and optimization, and realistic benchmark generation, describing the ability of ML to bring new capabilities to EDA that were previously unavailable.

Biography

Sachin S. Sapatnekar is the Henle Chair in ECE and Distinguished McKnight University Professor at the University of Minnesota. His research interests include design automation methods for analog and digital circuits, circuit reliability, and algorithms and architectures for machine learning. He is a recipient of the the SRC Technical Excellence Award, the Semiconductor Industry Association University Research Award, and 11 Best Paper awards. He has served as Editor-in-Chief of the IEEE Transactions on CAD and General Chair for the ACM/IEEE Design Automation Conference (DAC). He is a Fellow of the IEEE and the ACM.

EDAML 2022 Invited Speaker

Dependability and Security of Advanced Machine Learning Systems: Hardware and Software Techniques

Muhammad Shafique
New York University (NYU) Abu Dhabi, UAE

Abstract

Modern Machine Learning (ML) and Artificial Intelligence (AI) approaches, such as, the Deep Neural Networks (DNNs), have shown tremendous improvement over the past years to achieve a significantly high accuracy for a certain set of tasks, like image classification, object detection, natural language processing, and medical data analytics. These DNNs are deployed in s a wide range of applications from Smart Cyber Physical Systems (CPS) and Internet of Thing (IoT) domains on resource-constrained devices subjected to unpredictable and harsh scenarios, thereby requiring dependable AI solutions. Moreover, in the era of growing cyber-security threats, the intelligent features of a smart CPS and IoT system face new type of attacks, requiring novel design principles for robust ML/AI. In my research labs at New York University and TU Wien, I have been extensively investigating the foundations for the next-generation energy-efficient, dependable, and secure AI/ML computing systems, while addressing the above-mentioned challenges across the hardware and software stacks. This talk will present design challenges and hardware/software techniques for building dependable and secure AI systems, which leverage optimizations at different software and hardware layers, and at different design stages (e.g., design-time vs. run-time approaches). These techniques provide crucial steps towards enabling the wide-scale deployment of dependable and secure embedded AI systems like UAVs, autonomous vehicles, Robotics, IoT-Healthcare / Wearables, Industrial-IoT, etc.

Biography

Muhammad Shafique received his Ph.D. degree in computer science from the Karlsruhe Institute of Technology (KIT), Germany, in 2011. Afterwards, he established and led a highly recognized research group at KIT for several years as well as conducted impactful R&D activities in Pakistan and across the globe. Besides co-founding a technology startup in Pakistan, he was also an initiator and team lead of an ICT R&D project. He has also established strong research ties with multiple universities in Pakistan and worldwide, where he has been actively co-supervising various R&D activities and student/research Theses since 2011, resulting in top-quality research outcome and scientific publications. Before KIT, he was with Streaming Networks Pvt. Ltd. where he was involved in research and development of video coding systems several years. In Oct.2016, he joined the Institute of Computer Engineering at the Faculty of Informatics, Technische Universität Wien (TU Wien), Vienna, Austria as a Full Professor of Computer Architecture and Robust, Energy-Efficient Technologies (CARE-Tech.). Since Sep.2020, he is with the Division of Engineering at New York University (NYU) Abu Dhabi in UAE, and is a Global Network faculty at the NYU's Tandon School of Engineering in New York, USA. He is the director of the eBrain research lab, and is also a Co-PI/Investigator in multiple NYU-AD Centers, including Center of Artificial Intelligence and Robotics (CAIR), Center of Cyber Security (CCS), Center for InTeractIng urban nEtworkS (CITIES), and Center for Quantum and Topological Systems (CQTS).

EDAML 2022 Invited Speaker

Thermal and Power Monitoring and Estimation for Commercial Multicore Processors
-- A Machine Learning Perspective

Sheldon Tan
University of California at Riverside

Abstract

In this talk, I will present our recent work from my VSCLAB at UC Riverside on machine learning based thermal map and power density map estimation methods for commercial multi-core CPUs. I will first present the task-dependent hot spot or heat sources identification for multi-core processors based on measured thermal maps. In our work, instead of using traditional functional unit powers as input, the new models are directly based on the on-chip real-time high level chip utilizations and thermal sensor information of commercial chips without any assumption of additional physical sensors requirement. We first framed the problem as the static or transient mapping between the chip utilizations and thermal maps. To build the transient thermal model, we utilized temporal-aware long-short-term-memory (LSTM) neural networks with system-level variables such as chip frequency, voltage, and instruction counts as inputs. Instead of a pixel-wise heatmap estimation, we used 2D spatial discrete cosine transformation (DCT) on the heatmaps so that they can be expressed with just a few dominant DCT coefficients. Second, we explored generative learning for the full-chip thermal map estimation problem. In our work, we treated the thermal modeling problem as an image-generation problem using the generative neural networks. The resulting thermal map estimation method, called ThermGAN can provide tool-accurate full-chip transient thermal maps from the given performance monitor traces of commercial off-the-shelf multi-core processors. Third, I will present a new full-chip power map estimation method for commercial multi-core processors. We proposed to use a simple first-principle based 2D spatial Laplace method to generate the power maps from the measured or simulated thermal maps. Then I will show how we can obtain the thermal map estimation for multi-processor chips under practical heat sink cooling conditions in which no thermal maps can be measured.

Biography

Dr. Sheldon Tan is a Professor in the Department of Electrical Engineering, University of California, Riverside, CA. He is the Associate Director of Compute Engineering Program (CEN) and cooperative faculty member in the Department of Computer Science and Engineering at UCR. Dr. Sheldon Tan received his B.S. and M.S. degrees in electrical engineering from Fudan University, Shanghai, China in 1992 and 1995, respectively and the Ph.D. degree in electrical and computer engineering from the University of Iowa, Iowa City, in 1999. He is a visiting professor of Kyoto University as a JSPS Fellow since Dec. 2017. His research interests include machine and deep learning for VLSI reliability modeling and optimization at circuit and system levels, machine learning for circuit and thermal simulation, thermal modeling, optimization and dynamic thermal management for many-core processors. He has published more than 330 technical papers and has co-authored 6 books on those areas. Dr. Tan received NSF CAREER Award in 2004. He also received Best Paper Awards from ICSICT'18, ASICON’17, ICCD'07, DAC’09. He also received ASPDAC Prolific Author Award in 2020.

EDAML 2022 Invited Speaker

Hardware/Software Codesign for Optical Deep Learning Accelerators

Sudeep Pasricha
Colorado State University

Abstract

The massive data deluge from mobile, IoT, and edge devices, together with powerful innovations in data science and hardware processing, have established machine learning (ML) as the cornerstone of modern medical, automotive, industrial automation, and consumer electronics domains. Domain-specific ML accelerators such as Google's TPU and Apple's Bionic, now dominate CPUs and GPUs for energy-efficient ML processing. However, the evolution of these electronic accelerators is facing fundamental limits due to the slowdown of Moore's law and the reliance on metal wires, which already severely bottleneck computational performance today. Silicon photonics represents a promising post-Moore technological alternative to overcome these limitations. Not only can photonic interconnects fabricated in CMOS-compatible processes provide near speed of light transfers at the chip-scale, but photonic devices can now also perform computations entirely in the optical domain. In this talk, I will present my vision of how silicon photonics can drive an entirely new class of sustainable ML hardware accelerators that can provide orders of magnitude energy improvements over today's accelerators. I will discuss new directions in hardware/software codesign for ML acceleration with silicon photonics, with multi-objective goals related to power and energy minimization, variation tolerance, fault resilience, and secure computing.

Biography

Sudeep Pasricha received the B.E. degree in Electronics and Communication Engineering from Delhi Institute of Technology, India, in 2000, after which he spent several years working for STMicroelectronics, India/France, and Conexant, USA. He received his Ph.D. degree in Computer Science from the University of California, Irvine in 2008. He joined Colorado State University (CSU) in 2008 where he is currently a Walter Scott Jr. College of Engineering Professor in the Department of Electrical and Computer Engineering. He is a former University Distinguished Monfort Professor and Rockwell-Anderson Professor. He is currently also Chair of Computer Engineering and Director of the Embedded, High Performance, and Intelligent Computing (EPIC) Laboratory at CSU. His research focuses on the design of innovative software algorithms, hardware architectures, and hardware-software co-design techniques for energy-efficient, fault-tolerant, real-time, and secure computing, with applications to embedded, IoT, and cyber-physical systems. Prof. Pasricha's contributions have been recognized with various awards, including the George T. Abell Outstanding Research Faculty Award, IEEE-CS/TCVLSI Mid-Career Research Achievement Award, IEEE/TCSC Award for Excellence for a Mid-Career Researcher, AFOSR Young Investigator Award, ACM Technical Leadership Award, and ACM SIGDA Distinguished Service Award. He is currently the Vice Chair of ACM SIGDA and the Steering Committee Chair for the IEEE Transactions on Sustainable Computing. He is also a Senior Associate Editor for the ACM Journal of Emerging Technologies in Computing, and an Associate Editor with several ACM and IEEE journals. He is an IEEE Senior Member, an ACM Distinguished Member, and an ACM Distinguished Speaker.