IEEE/ACM

2024 INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN

ICCAD

43rd Edition

IEEE/ACM ICCAD CONFERENCE 2024

CONFERENCE PROGRAM

SPONSORS AND ORGANIZERS

IEEE

CAS
IEEE CIRCUITS AND SYSTEMS SOCIETY

acm
Association for Computing Machinery

CEDA
IEEE Council on Electronic Design Automation

sigda
acm

Electron Devices Society

Please visit our website for more information!

2024.iccad.com

Table of Contents

Executive Committee

General Chair

Jinjun Xiong, University at Buffalo, USA

Past Chair

Evangeline Young, The Chinese University of Hong Kong, China

Program Chair

Robert Wille, Technical University of Munich & Software Competence Center Hagenberg GmbH, Germany

Vice Program Chair

Deming Chen, University of Illinois, USA

Tutorial & Special Session Chair

Ismail S. K. Bustany, AMD, USA

Workshop Chair

Tsung-Yi Ho, The Chinese University of Hong Kong, China

CEDA Representative

Jiang Hu, Texas A&M University, USA

ACM SIGDA Representative

Wanli Chang, Hunan University, China

Asian Representative

Wei Zhang, The Hong Kong University of Science and Technology, China

European Representative

Ulf Schlichtmann, Technical University of Munich, Germany

Industry Liaison

Haoxing (Mark) Ren, NVIDIA, USA

Treasurer

Chen Wang, IBM, USA

General Information

Registration | Location: Grand Ballroom Foyer

Sunday, October 27 – CADathlon ONLY	7:30 – 8:00
Monday, October 28	7:00 – 18:00
Tuesday, October 29	7:30 – 15:45
Wednesday, October 30	8:00 – 17:00
Thursday, October 31	7:30 – 17:00

ICCAD Social Media Platforms

Conference website: https://iccad.com/

LinkedIn: https://www.linkedin.com/company/iccad

Twitter: https://twitter.com/ICCAD

Venue Map & Information

Newark Liberty International Airport Marriott

Address: Newark Liberty International Airport, 1 Hotel Rd, Newark, NJ 07114

Shuttle Services from Newark Liberty International Airport

From baggage claim, we've provided you with directions to navigate pick-up from each terminal at Newark Liberty International Airport to the hotel.

Terminal A: Exit the baggage claim area and cross the street to the center island. Walk to pillar #16 where you will see a digital sign for Marriott Hotel Shuttle and ride share.

Terminal B: From baggage claim walk toward the exit and go outside. Cross the street past the taxi pick lane and look for sign #5 that reads Marriott Shared Services.

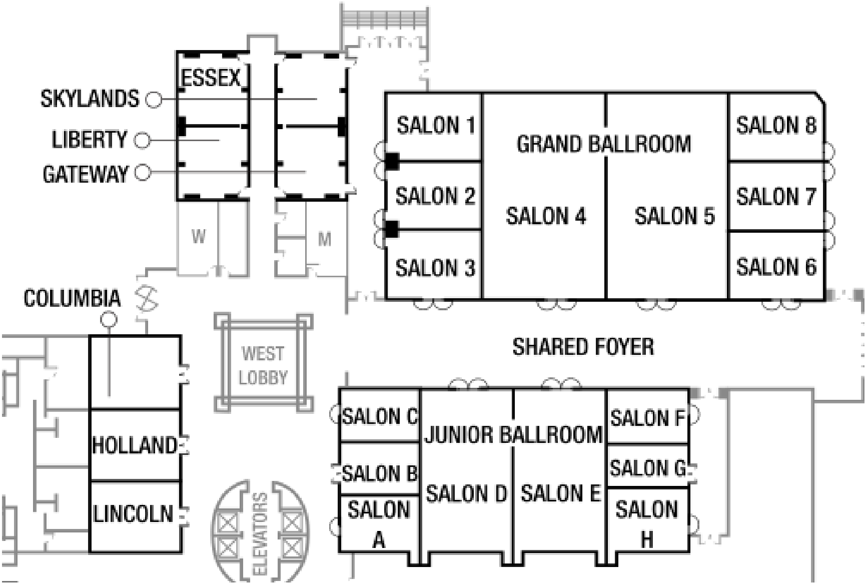
Terminal C: From baggage claim walk toward the exit sign and go outside. Cross the street to the past the taxi pick lane and walk to sign #7 that reads Marriott Shared Services

Parking at the Hotel

On-Site Parking:

- Hourly: $12.00
- Daily: $37.00

Venue Map



Welcome Message

Dear Attendees of the 43rd International Conference on Computer-Aided Design (ICCAD),

It is my immense pleasure to welcome you to this year's ICCAD, taking place for the first time in the vibrant New York/New Jersey metropolitan area. This historic decision reflects, in part, the ICCAD Executive Committee's recognition of New York State's strong commitment to advancing semiconductor technologies, where the ICCAD community will play a pivotal role. We are thrilled to have you join us for this exciting event, uniting thought leaders, innovators, and passionate individuals from around the globe.

This year marks a return to full in-person attendance, offering an immersive experience for attendees to connect with colleagues and friends who share a common interest in advancing cutting-edge computer-aided design techniques for microelectronics, devices, circuits, architectures, systems, and applications in this dynamic era of artificial intelligence.

Jointly sponsored by IEEE and ACM, ICCAD remains the premier venue for researchers and practitioners to share ground-breaking ideas and technological innovations in electronic design automation. Here, you will encounter leading-edge R&D solutions and identify future research directions. We encourage you to engage in dynamic discussions with other attendees and explore new collaborative opportunities.

This year has set a new record for paper submissions, with 1,109 abstract submissions across 19 technical tracks from 29 regions and countries. Out of those, a total of 802 full papers went into a thorough review process, for which we invited 338 exceptional experts in our Technical Program Committee. After a rigorous online evaluation process, 195 papers were recommended for acceptance, reflecting a competitive acceptance rate of 24%. These papers are organized into 60 sessions, complemented by 9 special sessions and 4 embedded tutorials that provide in-depth insights from leading experts.

We feature 4 student contests to further the state-of-the-art in key relevant research areas: (1) RL or logic optimization, (2) power and timing optimization, (2) ML-based gate sizing with GP acceleration, and community-based dataset construction for LLM-Assisted Hardware Code Generation

Keynotes are a highlight of ICCAD, and we are honored to feature five distinguished keynote speakers this year. Monday morning's keynote will be delivered by Dr. Philip Wong from Stanford University on "Design and Design Automation for Future Generations of Chips." On Tuesday morning, Dr. Leon Stok from IBM Corporation will discuss "The Future of Chip Design." Dr. Dilma Da Silva from the U.S. National Science Foundation (NSF) will share her views on "NSF Investments in Semiconductor and Microelectronics" on Wednesday's morning. Our Monday's luncheon keynote will feature Mr. Brandon Wang from Synopsys Corporation on "Connecting Science to Commercials: The Road to One

Trillion Dollar Club," while Dr. Sachin Sapatnekar from the University of Minnesota will be featured during Tuesday's CEDA luncheon keynote on "Enabling Analog Design."

We are also hosting six co-located workshops on a range of compelling topics, such as approximate computing, hardware security, interconnect pathfinding, large circuit models, and quantum computing. These workshops will provide an excellent opportunity for more in-depth interaction. We encourage you to participate and consider extending your stay to engage with these workshops.

ICCAD emphasizes workforce development and promotes diversity within our field. By consolidating the various student-oriented programs, we will have one integrated "Student Scholar Program" for our young scholars. Exemplar activities for them include the SIGDA CADathlon, the ACM Student Research Competition, and the SIGDA Job Fair. Thanks to generous support from our sponsors, particularly Synopsys, we are proud to provide grants to support many young scholars to attend ICCAD, and many of them are the first time ICCAD attendees.

The success of ICCAD hinges on your participation. We hope you will reconnect with old friends and forge new connections. Additionally, we invite you to enjoy an exciting Tuesday night in NYC, including a chance to see the renowned Broadway show "The Lion King." Please make the most of your time here by exploring all the incredible activities NYC has to offer.

Lastly, I want to extend my heartfelt gratitude to our generous sponsors, dedicated supporters, and tireless volunteers that include members of the executive committee, technical program committee, and numerous student activity organizers. Your collective efforts have made this year's ICCAD a reality. Thank you!

Looking forward to an inspiring event!

General Chair of ICCAD 2024

Dr. Jinjun Xiong
Empire Innovation Professor
Department of Computer Science and Engineering
University at Buffalo, New York, USA

Awards

William J. McCalla ICCAD Best Paper Award -

Frontend : "An Agile Framework for Efficient LLM Accelerator Development and Model Inference. "
Lvcheng Chen, Ying Wu, Chenyi Wen, Shizhang Wang, Li Zhang, Bei Yu, QI SUN and Cheng Zhuo

Backend: "A Neural-Ordinary-Differential-Equations Based Generic Approach for Process Modeling in DTCO: A Case Study in Chemical-Mechanical Planarization and Copper Plating"
Yue Qian and Lan Chen

ICCAD 10 Year Retrospective Most Influential Paper Award

Tsung-Wei Huang and Martin D.F. Wong,
"OpenTimer: A high-performance timing analysis tool"
2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)

2024 Ernest S. Kuh Early Career Award

Mahdi Nikdast (Colorado State University)
"for outstanding contributions to design methodologies, optimization, and automation targeting emerging integrated photonic systems-on-chip."

2024 Outstanding Service Recognition Award

Evangeline Young (The Chinese University of Hong Kong)
"for outstanding service to the EDA Community as ICCAD General Chair in 2023"

Best Reviewer Award

Jongeun Lee, Ulsan National Institute of Science and Technology (UNIST)
Leslie Hwang, Arizona State University
Lana Josipovic, ETH Zurich

BPA Candidates

931: Customized Retrieval Augmented Generation and Benchmarking for EDA Tool Documentation QA
Yuan Pu (The Chinese University of Hong Kong); Zhuolun He (The Chinese University of Hong Kong); Tairu Qiu (ChatEDA Tech); Haoyuan WU (Shanghai AI Lab); Bei Yu (The Chinese University of Hong Kong)

662: Robust Implementation of Retrieval-Augmented Generation on Edge-based Computing-in-Memory Architectures
Ruiyang Qin (University of Notre Dame); Zheyu Yan (University of Notre Dame); Dewen Zeng (University of Notre Dame); Zhenge Jia (Shandong University); Dancheng Liu (SUNY Buffalo); Jianbo Liu (University of Notre Dame); Ahmed Abbasi (University of Notre Dame); Zhi Zheng (University of Notre Dame); Ningyuan Cao (University of Notre Dame); Kai Ni (University of Notre Dame); Jinjun Xiong (University at Buffalo); Yiyu Shi (University of Notre Dame)

1005: An Agile Framework for Efficient LLM Accelerator Development and Model Inference
Lvcheng Chen (Zhejiang University); Ying Wu (Zhejiang University); Chenyi Wen (Zhejiang University); Shizhang Wang (Hubei University of Technology); Li Zhang (Hubei University of Technology); Bei Yu (The Chinese University of Hong Kong); QI SUN (Zhejiang University); Cheng Zhuo (Zhejiang University)

1578: eXpect: On the Security Implications of Violations in AXI Implementations
Melisande Zonta (ETH Zürich); Andres Meza (UCSD); Nora Hinderling (ETH); Lucas Deutschmann (University of Kaiserslautern-Landau); Francesco Restuccia (University of California at San Diego); Ryan Kastner (UCSD); Shweta Shinde (ETH Zurich)

1349: Towards Uncertainty-Quantifiable Biomedical Intelligence: Mixed-signal Compute-in-Entropy for Bayesian Neural Networks
Likai Pei (University of Notre Dame); Yifan Qin (University of Notre Dame); Zephan M. Enciso (University of Notre Dame); Boyang Cheng (University of Notre Dame); Jianbo Liu (University of Notre Dame); Steven Davis (University of Notre Dame); Zhenge Jia (Shandong University); Michael Niemier (University of Notre Dame); Yiyu Shi (University of Notre Dame); X. Sharon Hu (University of Notre Dame); Ningyuan Cao (University of Notre Dame)

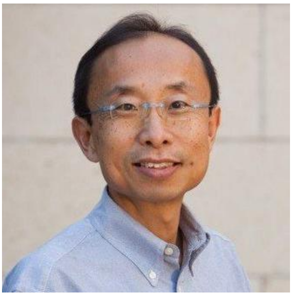
650: Fusion of Global Placement and Gate Sizing with Differentiable Optimization
Yufan Du (Peking University); Zizheng Guo (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

514: A Neural-Ordinary-Differential-Equations Based Generic Approach for Process Modeling in DTCO: A Case Study in Chemical-Mechanical Planarization and Copper Plating
Yue Qian (EDA Center: Institute of Microelectronics: Chinese Academy of Sciences and University of Chinese Academy of Sciences); Lan Chen (EDA Center: Institute of Microelectronics: Chinese Academy of Sciences and University of Chinese Academy of Sciences)

1516: Spiking Transformer Hardware Accelerators in 3D Integration
Boxun Xu (University of California: Santa Barbara); Junyoung Hwang (Georgia Institute of Technology); Pruek Vanna-iampikul (Georgia Institute of Technology); Sung Kyu Lim (Georgia Tech); Peng Li (University of California: Santa Barbara)

795: DDP-Fsim: Efficient and Scalable Fault Simulation for Deterministic Patterns with Two-Dimensional Parallelism
Feng Gu (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Mingjun Wang (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Jianan Mu (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Zizhen Liu (Institute of Computing Technology:Chinese Academy of Sciences); Jiaping Tang (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Hui Wang (CASTEST: Beijing); Yonghao Wang (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Jing Ye (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Huawei Li (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Xiaowei Li (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing)

1080: Efficient Ultra-Dense 3D IC Power Delivery and Cooling Using 3D Thermal Scaffolding
Dennis Rich (Stanford University); Tathagata Srimani (Stanford University); Mohamadali Malakoutian (Stanford University); Srabanti Chowdhury (Stanford University); Subhasish Mitra (Stanford University)



Design and Design Automation for a Future Generation of Chips

Monday | October 28, 2024 | 8:30 – 9:30
Room: Salons 4-8

Philip Wong, Stanford University, USA

Bio: H.-S. Philip Wong is the Willard R. and Inez Kerr Bell Professor in the School of Engineering at Stanford University. He joined Stanford University as Professor of Electrical Engineering in 2004. From 1988 to 2004, he was with the IBM T.J. Watson Research Center. From 2018 to 2020, he was on leave from Stanford and was the Vice President of Corporate Research at TSMC, the largest semiconductor foundry in the world, and since 2020 remains the Chief Scientist of TSMC in a consulting, advisory role.

He is a Fellow of the IEEE and received the IEEE Andrew S. Grove Award, the IEEE Technical Field Award to honor individuals for outstanding contributions to solid-state devices and technology, as well as the IEEE Electron Devices Society J.J. Ebers Award, the society's highest honor to recognize outstanding technical contributions to the field of electron devices that have made a lasting impact.

He is the founding Faculty Co-Director of the Stanford SystemX Alliance – an industrial affiliate program focused on building systems and the faculty director of the Stanford Nanofabrication Facility – a shared facility for device fabrication on the Stanford campus that serves academic, industrial, and governmental researchers across the U.S. and around the globe, sponsored in part by the National Science Foundation. He is the Principal Investigator of the Microelectronics Commons California-Pacific-Northwest AI Hardware Hub, a consortium of over 40 companies and academic institutions funded by the CHIPS Act. He is a member of the US Department of Commerce Industrial Advisory Committee on microelectronics.

Abstract: Three dimensional integration is one of the major technology directions for integrated circuits. Nanosystems of 3D integrated "X" technology (N3XT) is a key concept at the chip level, where X can be memory, photonics, spintronics, power electronics, nanomechanics, sensors and actuators, and RF/mm-wave. We must also go beyond a single chip from a wafer and focus on integrating chips into systems using MOSAIC (MOnolithic Stacked Assembled IC). These future 3D systems require a complete overhaul of the design methodology and design automation. Designs of 3D systems must start from the fundamental assumption that interconnections are in 3D and the interconnection density in all three dimensions are equally dense and efficient. This presents tremendous opportunities for research on system architecture and design automation.

The Future of Chip Design

Tuesday | October 29, 2024 | 8:30 – 9:30
Room: Salons 4-8

Leon Stok, IBM, USA

Bio: Leon Stok studied electrical engineering at Eindhoven University of Technology, The Netherlands, where he graduated with honors in 1986. He obtained his PhD degree from Eindhoven University in 1991. He worked at IBM's Thomas J. Watson Research Center as part of the team that developed BooleDozer, the IBM logic synthesis tool. Subsequently, he managed IBM's logic synthesis group and initiated the development of the first physical synthesis system: PDS, IBM's Placement Driven Synthesis tool. From 1999 to 2004, he led all of IBM's design automation research as the Senior Manager Design Automation at IBM Research. He is currently Vice President, Electronic Design Automation at IBM. He entered the field of Design Automation 25 years ago intrigued by the type of problems being posed by Moore's law. He has enjoyed working on problems from high-level synthesis to prescriptive layout design and DFM. In these 25 years, he attended most of the Design Automation Conferences, as a presenter of his original work in papers, a reviewer of the state of the art in tutorials or as a panelist to give his opinion on current issues. He served in many roles as a member of the DAC executive committee and as the chair of the 48th DAC. He is a Fellow of the IEEE.

Abstract: In 2030 we will be able to produce 200B transistor monolithic chips and over 1T 3D integrated multi-chiplet solutions. These will get generated from a few million lines of RTL. A massive generation effort. But how much of this will be done with generative AI? How much will it be done by classical optimization methods?

As with any new technology there are a lot of open questions. Why can't we create a L*M to take a processor written in Verilog and generate 1nm DRC clean and timing optimized layout and by-pass the entire EDA stack? Or can we? What is the place of generative AI in Hardware design? This keynote will give some food for thought to ponder these questions.

Keynote Speakers (cont.)

Enabling Analog Design

Tuesday | October 29, 2024 | 11:30 – 13:00
Room: Salons 4-8

Sachin Sapatnekar, University of Minnesota, USA

Bio: Sachin S. Sapatnekar received the Ph.D. degree from the University of Illinois at Urbana-Champaign in 1992. He was on the faculty at Iowa State University for five years, and since 1997, he has been teaching at the University of Minnesota, where he holds the Distinguished McKnight University Professorship and the Henle Chair Professorship in ECE. His research is related to developing CAD techniques for the analysis and optimization of circuit performance, currently focused on analog and digital CMOS circuits as well as post-CMOS technologies and novel computational models. He has served as Editor-in-Chief of the IEEE Transactions on CAD and General Chair for the ACM/IEEE Design Automation Conference (DAC), and was lead PI of the ALIGN analog EDA project. He is a recipient of ten conference Best Paper Awards, two ICCAD Ten Year Retrospective Best Paper Awards, the Semiconductor Research Corporation's Technical Excellence Award, the Semiconductor Industry Association University Research Award, a Fulbright award, and the UMN ECE Penrose Teaching Award. He is a Fellow of the IEEE and the ACM.

Abstract: Today's computing models see increasing analog content for two reasons: ever-increasing interaction with an analog real world, and the growing realization that computing on analog substrates can, under certain circumstances, be more efficient than traditional digital computing. To address the new frontiers opened up by this change, it is vital to explore ways in which EDA and systems design can meet the needs of the design community. This talk will address the problem of building an ecosystem that supports the full integration of analog design into computing substrates. Components of this ecosystem range from high-level algorithmic support to back-end design and optimization at the transistor and layout levels. A summary of the current state of the art, and pointers towards future directions, will be presented.

Keynote Speakers (cont.)

NSF investments in Semiconductor and Microelectronics

Wednesday | October 30, 2024 | 9:00 – 9:00
Room: Salons 4-8

Dilma Da Silva, US National Science Foundation, USA

Bio: Dilma Da Silva is a Regent Professor and Holder of the Ford Design Professorship II at the Department of Computer Science and Engineering at Texas A&M University. She serves as Division Director for the Division of Computing and Communication Foundations at the Directorate of Computer and Information Science and Engineering (CISE) at the US National Science Foundation and served as Acting Lead for CISE from December 2023 to May 2024. Her previous roles at Texas A&M include Department Head (2014-2019), Associate Dean (2019-2020), interim director of the Texas A&M Institute of Data Science, and interim director of the Texas A&M Cybersecurity Center. Her primary research interests are distributed systems, operating systems, and computer science education. Before joining Texas A&M, she worked at Qualcomm Research (2012-2014), IBM Research (2000-2012), and the University of São Paulo (1996-2000).

Dilma is an ACM Distinguished Scientist and a Latinas in Computing group co-founder. She was a member of the board of CRA-WP (Computer Research Association's Widening Participation Committee) from 2008 to 2022. She has chaired over 35 conferences/workshops and participated in over 100 program committees. She has published more than 100 technical papers and has 15 patents.

Dilma received her doctoral degree in computer science from Georgia Tech in 1997 and her bachelor's and master's degrees from the University of São Paulo, Brazil. She is passionate about enabling the next generation of talent.

Abstract: The talk provides information on investments in semiconductors, microelectronics, and quantum computing across the US National Science Foundation. It also presents an overview of the Computer and Information Science and Engineering (CISE) directorate, highlighting its programs related to semiconductors, microelectronics, design automation, and emerging technologies.

Synopsys Invited Speaker

Connecting Science to Commercials: The Road to One Trillion Dollar Club

Monday | October 28, 2024 | 12:30 - 13:30
Room: Salons 4-8

Brandon Wang, Synopsys, USA

Bio: Brandon Wang is a Vice President, VP of Technology Strategy at Office of the CEO at Synopsys. Brandon oversees corporate level technology roadmaps and strategies for growth, including global strategic and ecosystem partnerships, and M&A/investments for new horizons. He also heads the Chief Innovation Office, championing organic innovations and worldwide academic and research partnerships.

Prior to joining Synopsys in 2018, Brandon served executive and senior roles at Cadence, ARM, Qualcomm, and Lattice in chief strategy office, product marketing, and R&D organizations for over two decades.

Brandon is currently serving at the board of Efabless corporation, and as a limited partner/Investment Council member at Imec.Xpand, an affiliated VC to Imec. He is also a Limited Partner at AIX ventures, a leading venture capital firm focusing on AI investment and an LP to Black Opal ventures, who invests in life science and healthcare.

An Electrical and Computer Engineer by training, Brandon holds 10 patents, and has published at 20+ IEEE conferences, in journal papers and invited talks in the areas of 3D-IC, Machine Learning, HW security, High-speed Interfaces, Low Power CPU, FPGA and Memory Designs. He also has an MBA degree from the Wharton School at the University of Pennsylvania.

Abstract: Embark on a visionary exploration of the semiconductor industry's path to the coveted One Trillion Dollar Club in our keynote presentation. Delve into three pivotal technology trends shaping this journey:

Novel AI: Uncover the transformative impact of cutting-edge Artificial Intelligence on semiconductor innovation; Quantum Computing: Explore the quantum frontier and its profound influence on semiconductor capabilities.

Sustainability and Cleantech: Navigate the intersection of semiconductor technology and sustainability. These trends are propelled by key enabling technologies: Multi-die Implementation: Unveil the significance of multi-die solutions, revolutionizing semiconductor design to enhance performance, scalability, and efficiency. Software-Defined Vehicle: Explore the transformative impact of semiconductor technologies on the automotive sector, driving the evolution towards software-defined vehicles. [Learn More]

Tutorial Sessions

Security of Quantum Computing Hardware and Architectures

Monday | October 28, 2024 | 10:00 - 12:00
Room: Salons 1-3

Jakub Szefer, Yale University, USA

This tutorial will introduce the audience to the emerging field of quantum computer security, which focuses on research on how to make quantum computing systems secure from attacks. By design, this tutorial will not cover post-quantum cryptography as that is an important, but orthogonal topic. The tutorial focuses on security of quantum computing systems as the rapid advances in quantum computer technologies, quantum computers hold promise to be able to run algorithms for generating novel drugs or material compounds. [Read More]

Hardware Security and Trust Verification

Monday | October 28, 2024 | 13:30 - 15:30
Room: Salons 1-3

System-on-Chip (SoC) is the brain behind computing and communication in a wide variety of systems. Reusable hardware Intellectual Property (IP) based System-on-Chip (SoC) design has emerged as a pervasive design practice in the industry to dramatically reduce design and verification cost while meeting aggressive time-to-market constraints. Growing reliance on these pre-verified hardware IPs, often gathered from untrusted third-party vendors, severely affects the security and trustworthiness of SoC computing platforms. These third-party IPs may come with deliberate malicious implants to incorporate undesired functionality (e.g., hardware Trojans), undocumented test/debug interface working as hidden backdoor, or other integrity issues. [Read More]

Prabhat Mishra
University of Florida, USA

Farimah Farahmandi
University of Florida, USA

Tutorial Sessions (cont.)

Heterogeneous Integration: From physical layer to architecture and packaging

Monday | October 28, 2024 | 16:00 - 18:00
Room: Salons 1-3

The growth of compute- and data-intensive applications has led to a search for new computing architectures. General-purpose architectures such as CPUs, GPUs, often underperform compared to specialized accelerators like TPUs that are tailored for specific tasks like machine learning. Despite the advantages of specialized accelerators, the diversity in behavior among various emerging applications make it difficult to achieve good performance with one type of architecture only. Heterogeneous architectures, that combine multiple types of general-purpose cores and various specialized accelerators, are necessary to ensure that the computing platforms can support a variety of application such as AI, genomics, graph analytics, etc. This embedded tutorial will present some of the challenges and opportunities in heterogeneous architecture design starting from the physical layer, the design process for heterogeneous manycore architectures and packaging techniques. [Read More]

Adrian Evans
CEA/LIST, France

Partha Pratim Pande
Washington State University, USA

Chris Bailey
Arizona State University, USA

Advanced Sparse Linear Solver for Transistor-Level Circuit Simulation

Wednesday | October 30, 2024 | 13:30 - 15:30
Room: Salons 1-3

High-performance sparse linear solvers emerge as pivotal tools to facilitate rapid and accurate transistor-level circuit simulation and verification. Along with the fast development of semiconductor, modern integrated circuits (ICs) have been incredibly complex, consisting of hundreds of millions of components, causing sparse linear solvers to consume more time and memory resources for simulation. Furthermore, circuit matrices frequently exhibit high sparsity and non-uniform distributions of non-zero elements, compounding the chal lenge of achieving efficient acceleration. This tutorial proposes to delve into advanced sparse linear solving methodologies tailored specifically for transistor-level circuit simulation. We aim to explore state-of-the-art algorithms, optimizations, and parallelization strategies across various platforms, including CPUs, GPUs, and heterogeneous clusters to address the challenges posed by modern IC simulation. Practical implementation considerations and real-world case studies will be discussed to provide attendees with actionable insights into enhancing efficiency. [Read More]

Workshops

7th Top Picks in Hardware and Embedded Security Workshop at ICCAD 2024

Wednesday | October 30, 2024 | 10:30 - 17:00
Room: Essex/Liberty

Top Picks recognizes the best of the best in hardware security, spanning the gamut from hardware to microarchitecture to embedded systems. Since it was founded in 2018, we have successfully organized six (6) Top Picks Workshops, all co-located with ICCAD. Top Picks has established itself as a great venue for researchers to share not only their impactful research results, but also their experience in conducting such research. [Learn More]

SUSHI'24 Sustainable Hardware Security -An Interactive Workshop

Thursday | October 31, 2024 | 8:00 - 17:00
Room: Salons 1-3

The desire for digital sovereignty, exacerbated by global semiconductor shortages and geopolitical interests, has led to numerous worldwide initiatives to strengthen semiconductor technology and manufacturing on national and regional scales. Consequently, the pivotal role of hardware security and trustworthiness in computing systems and supply chains has become paramount.

Hardware security is foundational to the integrity of computing systems. Insecure hardware compromises critical system functionality and poses significant risks to society. [Learn More]

Quantum Computing Applications and Systems (QCAS)

Thursday | October 31, 2024 | 8:00 - 17:00
Room: Lincoln/Holland/Columbia

The ICCAD Quantum Computing Applications and Systems (QCAS) Workshop aims to provide a platform for researchers and industry professionals to present and discuss cross-disciplinary research and developments in quantum technologies, with a focus on design automation, computer architecture, and scientific methods. Foster collaboration and networking opportunities among participants from academia and industry to address emerging challenges and explore potential solutions in quantum error correction, quantum algorithms, quantum applications, and quantum control. Facilitate knowledge transfer and technology adoption by showcasing practical applications and case studies in areas such as quantum machine learning, quantum finance, quantum chemistry, and real-world applications in drug discovery and power systems. Promote interdisciplinary interactions and explore the intersection of quantum technologies with machine learning and the sciences, aiming to advance the field of quantum computing and its applications. [Learn More]

Workshops (cont.)

26th ACM/IEEE International Workshop on System-Level Interconnect Pathfinding (SLIP))

Thursday | October 31, 2024 | 8:00 - 17:00
Room: Salons A-C

The 2024 ACM/IEEE International Workshop on System-Level Interconnect Pathfinding (SLIP) is the 26th edition of the Workshop. SLIP, co-located with ICCAD 2024, will bring together researchers and practitioners who have a shared interest in the challenges and futures of system-level interconnect, coming from wide-ranging backgrounds that span system, application, design and technology. [Learn More]

Synergistic Innovations: Leveraging Large Models and EDA for Mutual Advancement

Thursday | October 31, 2024 | 8:00 - 12:30
Room: Skylands/Gateway

Our proposed workshop aims to explore the innovative synergy between large models and Electronic Design Automation (EDA). With the rapid advancements in artificial intelligence (AI) and EDA, these fields are on a converging path, and this workshop will delve into the mutual benefits they can offer each other. By bringing together researchers, practitioners, and industry experts, we aim to exchange ideas, share practical applications, and discuss the latest advancements. This event will serve as a platform to showcase cutting-edge research, foster collaboration, and drive innovation at the intersection of large models and EDA.

AxC'24 The 9th Workshop on Approximate Computing

Thursday | October 31, 2024 | 8:00 - 12:30
Room: Essex/Liberty

Modern computing systems are experiencing an unprecedented growth of data to be processed, since, on one hand, these systems are increasingly used to interact with the physical world and, on the other hand, they process large amounts of data samples coming from all the various sensing sources. This leads to computing systems requiring a tremendous amount of energy, at an increasing rate every year. Power and energy, therefore, became critical requirements in the design of computing systems, especially in pervasive embedded and mobile electronic devices. Additionally, computationally intensive tasks, such as machine-learning applications, have found their way into these power-limited devices, increasing the need for efficient electronics. [Learn More]

ICCAD on Broadway

ICCAD 2024 is thrilled to unveil an exciting update for our attendees! In lieu of the traditional gala dinner, we have something extraordinary planned – a collective journey to the iconic Broadway production of “The Lion King”; renowned for its spectacular performances, awe-inspiring set designs, and a captivating storyline that transcends generations.

Date: Tuesday, October 29, 2024

Address: Minskoff Theatre, 200 W 45th St, New York, NY 10036

Timing:

- 16:00 | Bus pickup at the Newark Liberty International Airport Marriott
- 17:00 | Bus drop off at Minskoff Theatre
- 19:00 | The Lion King show begins
- 21:30 | The Lion King
- 22:15 | Bus leaves Minskoff Theater

Transportation: The bus schedule will be strictly adhered to. If you miss either pick up times you will be responsible for arranging your own transportation.

The following vouchers will be given in addition to your ticket at registration:

- Plush Voucher: small Nala, Timon, Pumbaa, Zazu or Simba
- Single Liquor Drink, Wine or Beer in Souvenir Cup and Snack

To redeem these, please visit the gift shops/concession stands in the theatre.

WE ARE UNABLE TO ISSUE ADDITIONAL TICKETS/VOUCHERS. IF YOU LOSE EITHER WE WILL BE UNABLE TO ASSIST.

Seat Selection: You will not be able to select your seats for the event. Tickets will be assigned once at the conference.

Sponsors & Exhibitors

Platinum Sponsor

SYNOPSYS®

ACADEMIC & RESEARCH
ALLIANCES

Gold Sponsors

cādence®

FUTUREWEI
Technologies

Silver Sponsors

AMD

Empyrean

IBM®

SIEMENS

SUNY

UNIVERSITY
AT ALBANY
STATE UNIVERSITY OF NEW YORK

SAMSUNG

Sponsors & Exhibitors (cont.)

Coffee Break Sponsor

cādence®

Conference Sponsors

CAS
IEEE CIRCUITS AND SYSTEMS SOCIETY

SIGda acm

acm Association for Computing Machinery

CEDA®
IEEE Council on Electronic Design Automation

IEEE®
Advancing Technology for Humanity

ELECTRON DEVICES SOCIETY®

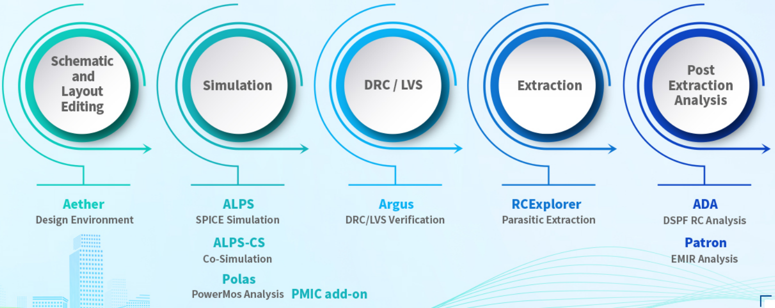
Empyrean

Empyrean Technology, founded in 2009, is an EDA and services provider to the global semiconductor industry. Empyrean Technology offers complete design solutions for analog, memory IC, RF IC, digital SoC, flat panel displays and advanced packaging design solutions.

Accelerating SPICE Simulation with ALPS-GPU

Empyrean ALPS® supports proprietary Smart Matrix Solving technology to provide a zero-accuracy-loss matrix solving and advanced multi-threading algorithms to deliver significant speedup, especially in post-layout simulation over 100M elements. The GPU-accelerated simulator delivers 10x speed improvements.

Complete Solution for Analog and Mixed-Signal designs



Schematic and Layout Editing — **Aether** Design Environment

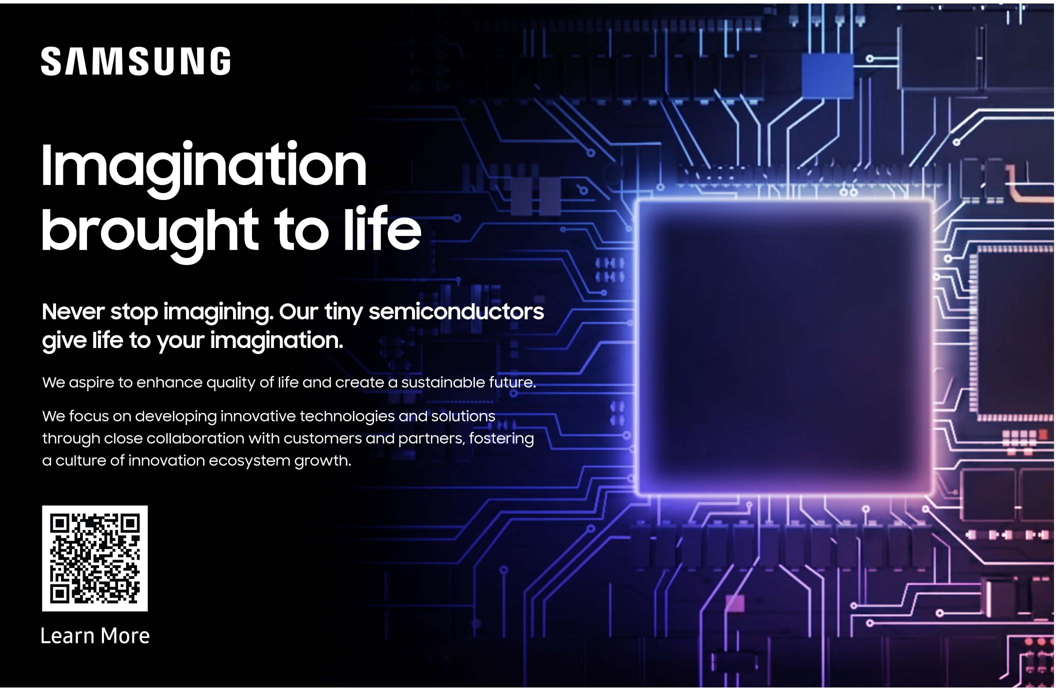
Simulation — **ALPS** SPICE Simulation; **ALPS-CS** Co-Simulation; **Polas** PowerMos Analysis; **PMIC add-on**

DRC / LVS — **Argus** DRC/LVS Verification

Extraction — **RCExplorer** Parasitic Extraction

Post Extraction Analysis — **ADA** DSPF RC Analysis; **Patron** EMIR Analysis

visit our website



SAMSUNG

Imagination brought to life

Never stop imagining. Our tiny semiconductors give life to your imagination.

We aspire to enhance quality of life and create a sustainable future.

We focus on developing innovative technologies and solutions through close collaboration with customers and partners, fostering a culture of innovation ecosystem growth.

Learn More

PROGRAM AT A GLANCE

Sunday | 10/27/2024

7:30	Registration || Room: Salons 1-3 Foyer
8:00	CADathlon Room: Salons 1-3 F
8:30	
9:00	
9:30	
10:00	
10:30	
11:00	
11:30	
12:00	
12:30	
13:00	
13:30	
14:00	
14:30	
15:00	
15:30	
16:00	
16:30	
17:00	

PROGRAM AT A GLANCE

Monday | 10/28/2024

7:00	Registration Room: Grand Ballroom Foyer					
7:30						
8:00	Opening Ceremony & Awards || Room: Salons 4-8					
8:30	Keynote: Design and Design Automation for a Future Generation of Chips Room: Salons 4-8					
9:00						
9:30	Coffee Break | Exhibits || Room: Salons D-E					
10:00	Tutorial: Security of Quantum Computing Hardware and Architectures Room: Salons 1-3	Special Session: The Dawn of Domain-Specific Hardware System for Autonomous Machines Room: Lincoln/Holland/Columbia	LLM4HWDesign Contest Room: Salons A-C	Special Session: Exploring Attack Vectors and Resilient Defense Strategies in Microelectronics A Special Session on Hardware Security Room: Skylands/Gateway	Special Session: Computing over Encrypted Data: Novel Acceleration of Fully Homomorphic Encryption on Hardware Platforms Room: Essex/Liberty	Advanced Partitioning and Floorplanning Room: Salons F-H
10:30						
11:00						State-of-the-Art Placement Room: Salons F-H
11:30						
12:00	Synopsys Invited Speaker Lunch Room: Salons 4-8					
12:30						
13:00						
13:30	Tutorial: Hardware Security Trust and Verification Room: Salons 1-3	Special Session: Advancing AI: Cross-disciplinary Insights into Next-Gen Tools, Tech & Architectures Room: Lincoln/Holland/Columbia	Special Session: AI4HLS: New Frontiers in High-Level Synthesis Augmented with Artificial Intelligence Room: Salons A-C	EDA for Quantum Room: Skylands/Gateway	Techniques for reliability modeling and analys Room: Essex/Liberty	Layout and Cell Optimization Room: Salons F-H
14:00						
14:30				Quantum Simulation and Quantum Cloud Room: Skylands/Gateway	Optimizations in lithography and physical design Room: Essex/Liberty	When Diverse Architectures Meet Diverse Ais Room: Salons F-H
15:00						
15:30	Coffee Break | Exhibits || Room: Salons D-E					
16:00	Tutorial: Heterogeneous Integration: From physical layer to architecture and packaging Room: Salons 1-3	Special Session: Towards Democratized and Reproducible AI for EDA Research: Open Datasets and Benchmarks in Various Aspects Room: Lincoln/Holland/Columbia	Special Session: Exploring Quantum Technologies in Practical Applications Room: Salons A-C	Timing Prediction and Acceleration Room: Skylands/Gateway	How Much ML Can You Squeeze into Your Edge Device? Room: Essex/Liberty	Reliable emerging technologies Room: Salons F-H
16:30						
17:00				Innovative Approaches in Circuit Simulation: High-Fidelity Modeling, Optimization, and Parallelization Room: Skylands/Gateway	Analog, Analog, and More Analog Design using Your Favorite AI Algorithms Room: Essex/Liberty	Emerging Technologies enabling Content Addressable Memories Room: Salons F-H
17:30						
18:00	Welcome Reception & SRC Poster Session Room: Salons D-E					
19:30						

PROGRAM AT A GLANCE

Tuesday | 10/29/2024

7:30	Registration Room: Grand Ballroom Foyer					
8:00						
8:30	Keynote: The Future of Chip Design Room: Salons 4-8					
9:30	Coffee Break | Exhibits || Room: Salons D-E					
10:00	Processor, Memory, and Storage Designs Room: Salons 1-3	Innovating Data Storage: Exploring Adaptive Indexing, Access Pattern Optimization, and Memory Longevity Enhancement for SSDs Room: Lincoln/ Holland/Columbia	Enhancing Simulation Efficiency through Multi-Core/GPU-Acceleration and Instruction-level Fault Injection Room: Salons A-C	Let LLMs Generate Your RTL Code! Room: Skylands/ Gateway	Architectural Mapping Room: Essex/ Liberty	IR Drop and High-speed Link Analysis Room: Salons F-H
10:45	Efficient Machine Learning: from Cloud to Edge Room: Salons 1-3	EdgeML: Efficient and Private ML for the Edge Room: Lincoln/ Holland/Columbia	Advances in Verification through SAT Solving and Machine Learning Room: Salons A-C	New Benchmarks and Understanding Benchmarks using LLMs Room: Skylands/ Gateway	Applications and Architectures Room: Essex/ Liberty	Machine Learning-based Design and Timing Optimization Room: Salons F-H
11:00						
11:30	CEDA Luncheon & Keynote Room: Salons 4-8					
12:00						
12:45						
13:00	Student Research Competition (Oral Presentations) Room: Salons 1-3	Special Session: Co-Designing NVM-based Systems for Machine Learning Applications Room: Lincoln/ Holland/ Columbia	Special Session: Delocalizing AI with Emerging Edge Intelligence (IoT/Internet) Room: Salons A-C	Let AI Power Your Synthesis and Defect Analysis! Room: Skylands/ Gateway	Revolutionizing AI with Low Power Accelerators: Emerging Design Trends Room: Essex/ Liberty	New Techniques in Analog Optimization: Bayesian Sensitivity, Hierarchical Placement, and AI-Driven 2.5D Chiplet Design Room: Salons F-H
13:30						
13:45				Dive into the Deisgn Space for Design Automation Room: Skylands/ Gateway	Machine Learning Innovations for Thermal and Power Optimization Room: Essex/ Liberty	Routing and ECO Routing Room: Salons F-H
14:30	Coffee Break | Exhibits || Room: Salons D-E					
15:00	Application Specific Accelerations Room: Salons 1-3	Enabling Sustainable Next Generation IoT and CPS Room: Lincoln/ Holland/ Columbia	Cycle-Accurate Timing Models, RISC-V Test Failure Analysis, and Low-Power Design Verification Room: Salons A-C	Bayesian Techniques for Software-Hardware Co-Optimization and Routing Room: Skylands/ Gateway	Design Frameworks and Post-place Optimization Room: Essex/ Liberty	Advances in Analog and RF Synthesis: Machine Learning Techniques and Thermal Analysis Room: Salons F-H
15:30						
19:00	Broadway Show (offsite)					
21:30						

PROGRAM AT A GLANCE

Wednesday | 10/30/2024

8:00	Registration Room: Grand Ballroom Foyer					
9:00	Keynote: NSF investments in Semiconductor and Microelectronics Room: Salons 4-8					
10:00	Coffee Break | Exhibits || Room: Salons D-E					
10:30	Microarchitecture Support for Security Room: Salons 1-3	New Research Developments in Synthesis Room: Lincoln/ Holland/Columbia	CTS and FPGA Routing Room: Salons A-C	PIM PIM PIM Room: Skylands/ Gateway	Top Picks Workshop Room: Essex/ Liberty	Real-Time AI: Co-Designing for the Edge Room: Salons F-H
11:00						
11:15	Security by Design and Pre-silicon Security Assurance Room: Salons 1-3	A New Life to Logic Synthesis Room: Lincoln/ Holland/Columbia	Machine Learning for P&R and Post-P&R Room: Salons A-C	Bringing Device Flavours Room: Skylands/ Gateway		IP, Side-Channels, and Acceleration Room: Salons F-H
12:00	Lunch Room: Salons 4-8					
13:00						
13:30	Tutorial: Advanced Sparse Linear Solver for Transistor-Level Circuit Simulation Room: Salons 1-3	Special Session: 2024 CAD Contests at ICCAD Room: Lincoln/ Holland/Columbia	Swift LLMs: Easier Design, Faster Inference Room: Salons A-C	Time is Limited: Fast and Secure Neural Network Accelerators Room: Skylands/ Gateway	Top Picks Workshop Room: Essex/ Liberty	Private Machine Learning Inference Room: Salons F-H
14:00						
14:30			CIM is on the Run: Sparser and More Robust Designs Room: Salons A-C	Treasures in the Graphs: Efficient Designs for GNNs Room: Skylands/ Gateway		More than Matrix Multiplication: Efficient Designs for Neural Networks Room: Salons F-H
15:00						
15:30	Coffee Break | Exhibits Room: Salons D-E					
16:00	Side Channels and Trojans Room: Salons 1-3	Advances in High-Level Synthesis and Optimized Components Room: Lincoln/ Holland/Columbia	Innovations in Neuromorphic Hardware and 3D Integration Room: Salons A-C	Precision Matters: Improving the Robustness and Reconfigurability Room: Skylands/ Gateway	Top Picks Workshop Room: Essex/ Liberty	Sparsity Matters: Sparse Computing Engines for Different Platforms Room: Salons F-H
16:30						
17:00	Job Fair || Room: Salons D-E					
18:00	ACM SIGDA Dinner || Room: Salons 4-8					
19:30						

PROGRAM AT A GLANCE

Thursday | 10/31/2024

7:30	Registration Room: Grand Ballroom Foyer				
8:00	SUSHI Workshop Room: Salons 1-3	Quantum Computing Applications and Systems (QCAS) Room: Lincoln/Holland/Columbia	26th ACM/IEEE International Workshop on System-Level Interconnect Pathfinding (SLIP) Room: Salons A-C	Synergistic Innovations: Leveraging Large Models and EDA for Mutual Advancement Room: Skylands/Gateway	AxC'24 The 9th Workshop on Approximate Computing Room: Essex/Liberty
8:30					
9:00					
9:30					
10:00					
10:30	Coffee Break Room: Grand Ballroom Foyer				
11:00	SUSHI Workshop Room: Salons 1-3	Quantum Computing Applications and Systems (QCAS) Room: Lincoln/Holland/Columbia	26th ACM/IEEE International Workshop on System-Level Interconnect Pathfinding (SLIP) Room: Salons A-C	Synergistic Innovations: Leveraging Large Models and EDA for Mutual Advancement Room: Skylands/Gateway	AxC'24 The 9th Workshop on Approximate Computing Room: Essex/Liberty
11:30					
12:00					
12:30	Lunch Room: Salons D-H				
13:00					
13:30					
14:00	SUSHI Workshop Room: Salons 1-3	Quantum Computing Applications and Systems (QCAS) Room: Lincoln/Holland/Columbia	26th ACM/IEEE International Workshop on System-Level Interconnect Pathfinding (SLIP) Room: Salons A-C		
14:30					
15:00					
15:30					
16:00					
16:30					

Technical Program – October 27, 2024

7:30 – 8:00
CADathlon ONLY Registration
Room: Salons 1-3 Foyer

8:00 – 9:00
CADathlon Breakfast
Room: Salons 1-3

9:00 – 12:00
CADathlon
Room: Salons 1-3

12:00 – 13:00
Lunch
Room: Salons 1-3

13:00 – 15:00
CADathlon
Room: Salons 1-3

15:00 – 15:30
Coffee Break
Room: Salons 1-3

15:30 – 17:00
CADathlon
Room: Salons 1-3

The CADathlon is a challenging, all-day programming competition focusing on practical problems at the forefront of Computer-Aided Design and Electronic Design Automation in particular. The contest emphasizes the knowledge of algorithmic techniques for CADapplications, problem-solving and programming skills, and teamwork.

As the “Olympic games of EDA,” the contest brings together the best and the brightest of the next generation of CAD professionals. It gives academia and the industry a unique perspective on challenging problems and rising stars, and it also helps attract top graduate students to the EDA field.

Technical Program – October 28, 2024

7:00 – 8:00
Registration
Room: Grand Ballroom Foyer

8:00 – 8:30
Opening Ceremony & Awards
Room: Salons 4-8

8:30 – 9:30
Keynote: Design and Design Automation for a Future Generation of Chips
Philip Wong, Stanford University
Room: Salons 4-8
Session Chair(s): Jinjun Xiong

9:30 – 10:00
Coffee Break | Exhibits
Room: Salons D-E

10:00 – 12:00
Tutorial: Security of Quantum Computing Hardware and Architectures
Room: Salons 1-3

This session introduces the emerging field of quantum computer security, focusing on safeguarding quantum systems from attacks. Topics include classical security concepts such as threat modeling, side-channel attacks, and vulnerability analysis. It discusses prototyped attacks on cloud-based quantum computers and strategies to secure future quantum computing platforms.

Technical Program – October 28, 2024

10:00 – 12:00
Special Session: The Dawn of Domain-Specific Hardware System for Autonomous Machines
Room: Lincoln/Holland/Columbia
Session Chair(s): Arijit Raychowdhury

This session highlights the transformative role of the Autonomous Machine Computing (AMC) paradigm across various sectors, from intelligent vehicles to drones. The session will cover AMC's evolution, focusing on critical components (e.g. sensing, computing, and communication technologies), along with key design factors such as performance, resilience, and reconfigurability. It will explore how AMC drives the transition to an Autonomy Economy, presenting significant challenges and opportunities for the hardware design community.

10:00
Imaging, Computing, and Human Perception: Three Agents to Usher in the Autonomous Machine Computing Era
Yuhao Zhu (University of Rochester)

10:30
Thinking and Moving: An Efficient Computing Approach for Integrated Task and Motion Planning in Cooperative Embodied AI Systems
Zishen Wan (Georgia Institute of Technology); Yuhang Du (University of Minnesota, Twin Cities); Mohamed Ibrahim (Georgia Institute of Technology); Yang Zhao (University of Minnesota, Twin Cities); Tushar Krishna (Georgia Institute of Technology); Arijit Raychowdhury (Georgia Institute of Technology)

11:00
Generative AI Agents in Autonomous Machines: A Safety Perspective
Jason Jabbour (Harvard University); Vijay Janapa Reddi (Harvard University)

11:30
Dataflow Accelerator Architecture for Autonomous Machine Computing
Shaoshan Liu (Shenzhen Institute of Artificial Intelligence and Robotics for Society); Yuhao Zhu (University of Rochester); Bo Yu (Shenzhen Institute of Artificial Intelligence and Robotics for Society); Jean-Luc Jean-Luc (University of California, Irvine); Guangrong Gao (University of Delaware); Arijit Raychowdhury (Georgia Institute of Technology)

10:00 – 12:00
LLM4HWDesign Contest
Room: Salons A-C

10:00
LLM4HWDesign Contest: Constructing a Comprehensive Dataset for LLM-Assisted Hardware Code Generation with Community Efforts
Zhongzhi Yu (Georgia Institute of Technology); Chaojian Li (Georgia Institute of Technology); Yongan Zhang (Georgia Institute of Technology); Mingjie Liu (Nvidia Corporation); Nathaniel Pinckney (Nvidia Corporation); Wenfei Zhou (Nvidia Corporation); Haoyu Yang (Nvidia Corporation); Rongjian Liang (Nvidia Corporation); Haoxing Ren (Nvidia Corporation); Yingyan (Celine) Lin (Georgia Institute of Technology)

10:00 – 12:00
Special Session: Exploring Attack Vectors and Resilient Defense Strategies in Microelectronics A Special Session on Hardware Security
Room: Skylands/Gateway
Session Chair(s): Hassan Salmani
Ujjwal Guin

This session tackles hardware security challenges. Topics include quantum RAM attacks, securing NAND Flash supply chains, using permissioned blockchains, detecting hardware Trojans, and mitigating physical attacks with random self-reducibility. These insights aim to strengthen microelectronics systems against emerging security threats in critical environments.

10:00
Exploration of Timing and Higher-Energy Attacks on Quantum Random Access Memory
Yizhuo Tan (Yale University); Chuanqi Xu (Yale University); Jakub Szefer (Yale University)

10:24
Fortifying the NAND Flash Supply Chain with Innovative Security Primitives
Matchima Buddhanoy (Colorado State University); Biswajit Ray (Colorado State University

10:48
Optimizing Supply Chain Management using Permissioned Blockchains
Aritri Priya Saha (Auburn University); Ujjwal Guin (Auburn University)

11:12
Detecting Hardware Trojans in Manufactured Chips without Reference: A GMM-Based Approach
Mahsa Tahghigh (Howard University); Hassan Salmani (Howard University)

11:36

Systematic Use of Random Self-Reducibility in Cryptographic Code against Physical Attacks

Ferhat Erata (Yale University); TingHung Chiu (Virginia Tech); Anthony Etim (Yale University); Srilalith Nampally (Virginia Tech); Tejas Raju (Virginia Tech); Rajashree Ramu (Virginia Tech); Ruzica Piskac (Yale University); Timos Antonopoulos (Yale University); Wenjie Xiong (Virginia Tech) (Presenter); and Jakub Szefer (Yale University)

10:00 – 12:00

Special Session: Computing over Encrypted Data: Novel Acceleration of Fully Homomorphic Encryption on Hardware Platforms

Room: Essex/Liberty

This session explores advancements in hardware acceleration for Fully Homomorphic Encryption (FHE). Topics include optimizing encryption and bootstrapping algorithms for secure computing and automated tools for high-performance Number Theoretic Transform accelerators. These approaches aim to improve FHE's computational efficiency, scalability, and security, to enable privacy-preserving computing in real-time applications and resource-constrained environments.

10:00

Enhancing Privacy-Preserving Computing with Optimized CKKS Encryption: A Hardware Acceleration Approach

Tianyou Bao (Villanova University); Pengzhou He (Villanova University); Jiafeng Xie (Villanova University)

10:30

Efficient Design of TFHE Bootstrapping Implementation with Scalable Parameters

Ming-Chien Ho (Inventec Corporation); Yu-Te Ku (Inventec Corporation); Yu Xiao (Inventec Corporation); Feng-Hao Liu (Washington State University); Chih-Fan Hsu (Inventec Corporation); Ming-Ching Chang (Inventec Corporation); Shih-Hao Hung (National Taiwan University); Wei-Chao Chen (Inventec Corporation)

11:00

OpenNTT: An Automated Toolchain for Compiling High-Performance NTT Accelerators in FHE

Florian Krieger (Graz University of Technology); Florian Hirner (Graz University of Technology); Ahmet Can Mert (Graz University of Technology); Sujoy Sinha Roy (Graz University of Technology)

11:30

HERME: Homomorphic Encryption over Residual Number System for Multi-level Evaluation

Antian Wang (Purdue University); Kaiyuan Zhang (Tufts University); Keshab K. Parhi (University of Minnesota); Yingjie Lao (Tufts University)

10:00 - 11:00

Advanced Partitioning and Floorplanning

Room: Salons F-H

Session Chair(s): Jiang Hu
Ing-Chao Lin

In this session, we explore cutting-edge techniques for partitioning and floorplanning. In the first paper, authors introduce GenPart that introduces a hypergraph partitioner using a generative model, outperforming traditional methods. Then in the second paper, TopoOrderPart is introduced that focuses on scheduling-driven partitioning, balancing topological order while minimizing cut size. The third paper delves into analytical-based approaches for rectilinear floorplanning, allowing shape-adjustable modules, and finally the fourth paper introduces JigsawPlanner, that handles complex-shaped rectilinear modules. Join us to discover the latest innovations in VLSI design planning.

10:00

1338: A Hypergraph Partitioner Utilizing a Novel Graph Generative Model

Magi Chen (National Tsing Hua University); Ting-Chi Wang (National Tsing Hua University)

10:15

1554: TopoOrderPart: a Multi-level Scheduling-Driven Partitioning Framework for Processor-Based Emulation

shunyang bi (Xidian University); jing tang (Xidian University); Hailong You (Xidian University); haonan wu (Xidian University); Cong Li (Xidian University); richard sun (S2C Inc.)

10:30

1440: Modern Fixed-Outline Floorplanning with Rectilinear Soft Modules

Yu-Yang Chen (National Taiwan University); Yi-Chen Lin (National Taiwan University); Tzu-Han Hsu (National Taiwan University); Iris Hui-Ru Jiang (National Taiwan University); Tung-Chieh Chen (Synopsys); Tai-Chen Chen (Synopsys, Inc.); Hua-Yu Chang (Synopsys, Inc.)

10:45

805: JigsawPlanner: Jigsaw-like Floorplanner for Eliminating Whitespace and Overlap among Complex Rectilinear Modules

Xingbo Du (Shanghai Jiao Tong University); Ruizhe Zhong (Shanghai Jiao Tong University); Shixiong Kai (Huawei Noah's Ark Lab); Zhentao Tang (Huawei Noah's Ark Lab); Siyuan Xu (Huawei Noah's Ark Lab); Jianye Hao (Tianjin University); Mingxuan Yuan (Huawei Noah's Ark Lab); Junchi Yan (Shanghai Jiao Tong University)

11:00 - 12:00
State-of-the-Art Placement
Room: Salons F-H
Session Chair(s): Hung-Ming Chen

In this session, we explore innovative techniques for the placement problem. The first paper introduces a novel design flow that integrates gate sizing with global placement, leveraging differentiable timing and leakage power objectives. In the second paper, an approach named SysMix to tailor a mixed-size placement for systolic array in hardware accelerator design. It has scalable speedup and wirelength reduction.

Then, in the third paper, the authors delve into a multilevel framework addressing fence region constraints during cell placement by eliminating inappropriate clustering and gradually refining placement regions. Finally, in the fourth paper, a timing-driven global placement framework is introduced that considers both graph and path information.

11:00
650: Fusion of Global Placement and Gate Sizing with Differentiable Optimization
Yufan Du (Peking University); Zizheng Guo (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

11:15
734: SysMix: Mixed-Size Placement for Systolic-Array-Based Hierarchical Designs
Donghao Fang (Texas A&M University); Hailiang Hu (Texas A&M University); Wuxi Li (AMD); Bo Yuan (Rutgers University); Jiang Hu (Texas A&M University)

11:30
728: An Effective Analytical Placement Approach to Handle Fence Region Constraint
Jai-Ming Lin (Department of Electrical Engineering, National Cheng Kung University); Wei-Yuan Lin (Department of Electrical Engineering, National Cheng Kung University); Yung-Chen Chen (Department of Electrical Engineering, National Cheng Kung University); Pin-Yu Chen (Academy of Innovative Semiconductor and Sustainable Manufacturing Program on Integrated Circuit Design, National Cheng Kung University); Chen-Fa Tsai (Design Service Division of GUC); De-Shiun Fu (Design Service Division of GUC); Che-Li Lin (Design Service Division of GUC)

11:45
1329: Hybrid Modeling and Weighting for Timing-driven Placement with Efficient Calibration
Bangqi Fu (The Chinese University of Hong Kong); Lixin Liu (The Chinese University of Hong Kong); Martin Wong (The Chinese University of Hong Kong); Evangeline Young (The Chinese University of Hong Kong)

12:00 – 13:30
Synopsys Invited Speaker Lunch
Room: Salons 4-8

13:30 – 15:30
Tutorial: Hardware Security Trust and Verification
Room: Salons 1-3

This session provides an overview of key challenges and recent advancements in hardware security. Topics include security vulnerability analysis, automated test generation, formal verification methods like model checking and theorem proving, AI-assisted security verification, and pre-silicon side-channel analysis. The session aims to integrate security verification into the functional validation flow, enhancing the trustworthiness and resilience of SoC designs.

ISLU: Indexing-Efficient Sparse LU factorization for Circuit Simulation on GPUs
Dan Niu (School of Automation, Southeast University); Yiyang Tao (School of Automation, Southeast University); Zhou Jin (SSSLab, Dept. of CST, China University of Petroleum-Beijing); Yichao Dong (School of Automation, Southeast University); Chao Wang (School of Automation, Southeast University); Changyin Sun (School of Artificial Intelligence, Anhui Universit)

13:30 – 15:30
Special Session: Advancing AI: Cross-disciplinary Insights into Next-Gen Tools, Tech & Architectures
Room: Lincoln/Holland/Columbia
Session Chair(s): Rajendra Bishnoi

This session explores advancements in AI hardware design driving next generation AI systems and hardware solutions, focusing on LLM's in EDA for AI accelerators, emerging technologies (e.g. photonics and memories), and innovative architectures such as processor-in-memory and spiking neural networks.

13:30
Hardware-Aware Quantization for Accurate Memristor-Based Neural Networks
Sumit Diware (Delft University of Technology); Mohammad Yaldagard (Delft University of Technology); Rajendra Bishnoi (Delft University of Technology)

14:00
LLM-AID: Leveraging Large Language Models for Rapid Domain-Specific Accelerator Development
Farshad Firouzi (Arizona State University); Sri Sai Rakesh Nakkilla (Arizona State University); Chenghao Fu (Chinese University of Hong Kong); Sanmitra Banerjee (Nvidia); Jonti Talukdar (IEEE); Krishnendu Chakrabarty (Arizona State University

14:30
Neuromorphic Computing for Graph Analytics
Anup Das (Drexel University)

15:00

Shedding Light on LLMs: Harnessing Photonic Neural Networks for Accelerating LLMs

Salma Afifi (Colorado State University); Febin Sunny (AMD); Sudeep Pasricha (Colorado State University); Mahdi Nikdast (Colorado State University)

13:30 – 15:30

Special Session: AI4HLS: New Frontiers in High-Level Synthesis Augmented with Artificial Intelligence

Room: Salons A-C

Session Chair(s): Jun "Jeff" Zhang
Antonino Tumeo

This session discusses the transformative impact of AI/ML on high-level synthesis. Topics include AI's role in generating register-transfer level (RTL) designs, improving design space exploration with ML, and using LLM's to generate hardware from high-level descriptions. These advances promise to enhance both the speed and quality of the hardware design process, pushing the boundaries of HLS methodologies.

13:30

The Promise and Challenges of Designing Digital Logic though Backpropagation

Mihailo Isakov (BoolSi)

14:00

Extending High-Level Synthesis with AI/ML Methods

Nicolas Bohm Agostini (PNNL & Northeastern University); Giovanni Gozzi (Politecnico di Milano); Michele Fiorito (Politecnico di Milano); Serena Curzel (Politecnico di Milano); Ankur Limaye (Pacific Northwest National Laboratory); Marco Minutoli (Pacific Northwest National Laboratory); Vito Giovanni Castellana (Pacific Northwest National Laboratory); Joseph Manzano (Pacific Northwest National Laboratory); Fabrizio Ferrandi (Politecnico di Milano); Antonino Tumeo (Pacific Northwest National Laboratory)

14:30

Are LLMs Any Good for High-Level Synthesis?

Yuchao Liao (University of Arizona); Tosiron Adegbija (University of Arizona); Roman Lysecky (University of Arizona)

15:00

High(er) Level Synthesis: Can HLS Tools Benefit from LLMs?

Siddharth Garg (New York University)

13:30 - 14:30
EDA for Quantum
Room: Skylands/Gateway
Session Chair(s): Liang Zhiding

In this session, we primarily discuss several recent developments in the design automation flows for Quantum Computing pertaining to various types of technologies and constraints generating therefrom.

13:30
730: Barber: Balancing Thermal Relaxation Deviations of NISQ Programs by Exploiting Bit-Inverted Circuits
Enhyeok Jang (Yonsei University); Seungwoo Choi (Yonsei University); Youngmin Kim (Yonsei University); Jeewoo Seo (Yonsei University); Won Woo Ro (Yonsei University)

13:45
510: Quantum State Preparation Circuit Optimization Exploiting Don't Cares
Hanyu Wang (ETH Zurich); Daniel Bochen Tan (University of California: Los Angeles); Jason Cong (UCLA)

14:00
765: ReCon: Reconfiguring Analog Rydberg Atom Quantum Computers for Quantum Generative Adversarial Networks
Nicholas S. DiBrita (Rice University); Daniel Leeds (Rice University); Yuqian Huo (Rice University); Jason Ludmir (Rice University); Tirthak Patel (Rice University)

14:15
1444: SMT-based Layout Synthesis for Silicon-based Quantum Computing with Crossbar Architecture
Sheng-Tan Huang (National Taiwan University of Science and Technology); Ying-Jie Jiang (National Taiwan University of Science and Technology); Shao-Yun Fang (National Taiwan University of Science and Technology); Chung-Kuan Cheng (UCSD)

13:30 - 14:30
Techniques for reliability modeling and analysis
Room: Essex/Liberty
Session Chair(s): Takashi Sato

This session focuses on modeling and analysis techniques aimed at improving modeling accuracy and reliability. The first paper demonstrates improved process modeling using a combination of neural networks and ordinary differential equations. The second paper proposes a customized large-scale multimodal model: called FabGPT: designed to improve IC manufacturing processes. The third paper proposes a rigorous analysis framework for yield estimation based on variational importance sampling: and the last paper proposes a robust ASIC placer that can handle hybrid region constraints via a unified multi-electrostatic formulation.

13:30
🏅 William J. McCalla ICCAD Best Paper Award - Backend
514: A Neural-Ordinary-Differential-Equations Based Generic Approach for Process Modeling in DTCO: A Case Study in Chemical-Mechanical Planarization and Copper Plating
Yue Qian (EDA Center: Institute of Microelectronics: Chinese Academy of Sciences and University of Chinese Academy of Sciences); Lan Chen (EDA Center: Institute of Microelectronics: Chinese Academy of Sciences and University of Chinese Academy of Sciences)

13:45
999: FabGPT: An Efficient Large Multimodal Model for Complex Wafer Defect Knowledge Queries
Yuqi Jiang (Zhejiang University); Xudong Lu (Zhejiang University); Qian Jin (Zhejiang University); QI SUN (Zhejiang University); Hanming Wu (Zhejiang University); Cheng Zhuo (Zhejiang University)

14:00
656: Beyond the Yield Barrier: Variational Importance Sampling Yield Analysis
Yanfang Liu (Beihang University); Lei He (University of California: Los Angeles); Wei W. Xing (The University of Sheffield)

14:15
803: BPINN-EM: Fast Stochastic Analysis of Electromigration Damage using Bayesian Physics-Informed Neural Networks
Subed Lamichhane (University of California: Riverside); Mohammadamir Kavousi (University of California: Riverside); Sheldon Tan (University of California at Riverside)

13:30 - 14:30
Layout and Cell Optimization
Room: Salons F-H
Session Chair(s): Ting-Chi Wang

In this session: we explore cutting-edge techniques for VLSI design and cell optimization. In the first paper: the authors discuss an optimal method for synthesizing area-optimal multi-row standard cells: integrating transistor folding: row partitioning: and transistor placement. The second work introduces ATPlace2.5D: an analytical thermal-aware chiplet placement framework for large-scale 2.5D-ICs. It balances wirelength and temperature. In the third paper: the authors present novel approaches for 3D SRAM arrays: wordline and bitline folding. These designs achieve remarkable reductions in footprint: improved speed: and energy efficiency. Finally: the fourth paper proposes MAXCell: a PPA-directed standard cell layout optimization framework using anytime MaxSAT: surpassing wire length optimization studies.

13:30
1232: Optimal Layout Synthesis of Multi-Row Standard Cells for Advanced Technology Nodes
Sehyeon Chung (Seoul National University); Hyunbae Seo (Seoul National University); Handong Cho (Seoul National University); Kyumyung Choi (Seoul National University); Taewhan Kim (Seoul National University)

13:45
520: ATPlace2.5D: Analytical Thermal-Aware Chiplet Placement Framework for Large-Scale 2.5D-IC
Qipan Wang (Peking University); Xueqing Li (Peking University); Tianyu Jia (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

14:00
1386: Multi-Tier 3D SRAM Module Design: Targeting Bit-Line and Word-Line Folding
Aditya S. Iyer (Georgia Institute of Technology); Daehyun Kim (Georgia Institute of Technology); Saibal Mukhopadhyay (Georgia Institute of Technology); Sung Kyu Lim (Georgia Tech)

14:15
804: MAXCell: PPA-Directed Multi-Height Cell Layout Routing Optimization using Anytime MaXSAT with Constraint Learning
Jiun-Cheng Tsai (Mediatek); Wei-Min Hsu (Mediatek); Yun-Ting Hsieh (Mediatek); Yu-Ju Li (Mediatek); Wei Huang (Mediatek); CN Ho (Mediatek); Hsuan-Ming Huang (Mediatek); Jen-Hang Yang (Mediatek); Heng-Liang Huang (Mediatek); Aaron C. -W. Liang (National Yang Ming Chiao Tung University); Charles H. -P. Wen (National Yang Ming Chiao Tung University)

14:30 - 15:30
Quantum Simulation and Quantum Cloud
Room: Skylands/Gateway
Session Chair(s): Zhiding Liang

Large-scale, cloud-based Quantum circuit execution brings forth challenges of task scheduling, preventing malicious activities and circuit partitioning for execution speed-up. The papers in this session delve deeper into these topics

14:30
827: Accelerating Quantum Circuit Simulation with Symbolic Execution and Loop Summarization
Tian-Fu Chen (Graduate School of Advanced Technology: National Taiwan University); Yu-Fang Chen (Academia Sinica); Jie-Hong Roland Jiang (National Taiwan University); Sara Jobranova (Brno University of Technology); Ondrej Lengal (Brno University of Technology)

14:45
1387: Detecting Fraudulent Services on Quantum Cloud Platforms via Dynamic Fingerprinting
Jindi Wu (William & Mary); Tianjie Hu (William & Mary); Qun Li (William & Mary)

15:00
682: On Reducing the Execution Latency of Superconducting Quantum Processors via Quantum Job Scheduling
Wenjie Wu (Shanghai Jiao Tong University); Yiquan Wang (Shanghai JiaoTong University); Ge Yan (Shanghai Jiao Tong University); Yuming Zhao (Shanghai Jiao Tong University); Bo Zhang (Shanghai Artificial Intelligence Laboratory); Junchi Yan (Shanghai Jiao Tong University)

15:15
862: A Hardware-Aware Gate Cutting Framework for Practical Quantum Circuit Knitting
Xiangyu Ren (University of Edinburgh); Mengyu Zhang (Tencent Quantum Laboratory); Antonio Barbalace (University of Edinburgh)

14:30 - 15:30

Optimizations in lithography and physical design

Room: Essex/Liberty

Session Chair(s): Keren Zhu

This session will showcase the latest developments relevant to the ever-important topics of lithography and physical design. The first paper proposes a multi-objective mask design optimization considering process variations. The second paper presents a differentiable edge-based OPC framework that combines the manufacturability of EBOPC with the performance of ILT. The third paper proposes a co-optimization framework to reduce the number of design rule violations in the legalization and filler insertion stages. The last paper proposes a robust ASIC placer that can handle hybrid region constraints via a unified multi-electrostatic formulation.

14:30

1098: Enabling Robust Inverse Lithography with Rigorous Multi-Objective Optimization

Yang Luo (The Hong Kong University of Science and Technology (GuangZhou)); Xiaoxiao Liang (The Hong Kong University of Science and Technology (Guangzhou)); Yuzhe Ma (The Hong Kong University of Science and Technology (Guangzhou))

14:45

1107: Differentiable Edge-based OPC

Guojin Chen (The Chinese University of HongKong); Haoyu Yang (NVIDIA Corp.); Haoxing Ren (NVIDIA Corporation); Bei Yu (The Chinese University of Hong Kong); David Z. Pan (University of Texas at Austin)

15:00

851: A Co-optimization Framework with Multi-layer Constraints for Manufacturability

Guohao Chen (Fudan University); Chang Liu (Fudan University); Xingyu Tong (Fudan University); Peng Zou (Shanghai LEDA Technology Co.: Ltd); Jianli Chen (Fudan University)

15:15

986: MORPH: More Robust ASIC Placement for Hybrid Region Constraint Management

Jing Mai (Peking University); Zuodong Zhang (School of Integrated Circuits: Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

14:30 - 15:30
When Diverse Architectures Meet Diverse Ais
Room: Salons F-H
Session Chair(s): Iraklis Anagnostopoulos

This session explores the essential strategies for optimizing deployment on various backendsIncluding CIM architectures: MCUs: and edge GPUs: catering to a wide array of applications such as RAG: MoE: BERT: and CNNs. The importance of this topic lies in its potential to enhance performance: efficiency: and scalability in diverse computing environments. Attendees will benefit from key papers presenting cutting-edge research and practical solutions for deploying these advanced models effectively. Join us to gain insights into the latest techniques and innovations driving the optimization of deployment across multiple hardware platforms.

14:30
662: Robust Implementation of Retrieval-Augmented Generation on Edge-based Computing-in-Memory Architectures
Ruiyang Qin (University of Notre Dame); Zheyu Yan (University of Notre Dame); Dewen Zeng (University of Notre Dame); Zhenge Jia (Shandong University); Dancheng Liu (SUNY Buffalo); Jianbo Liu (University of Notre Dame); Ahmed Abbasi (University of Notre Dame); Zhi Zheng (University of Notre Dame); Ningyuan Cao (University of Notre Dame); Kai Ni (University of Notre Dame); Jinjun Xiong (University at Buffalo); Yiyu Shi (University of Notre Dame)

14:45
977: AdapMoE: Adaptive Sensitivity-based Expert Gating and Management for Efficient MoE Inference
Shuzhang Zhong (Peking University); LING LIANG (Peking University); Yuan Wang (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits: Peking University)

15:00
995: MCUBERT: Memory-Efficient BERT Inference on Commodity Microcontrollers
Zebin Yang (Peking University); Renze Chen (Peking University); Taiqiang Wu (The University of Hong Kong); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits: Peking University); Ngai Wong (The University of Hong Kong); Yun (Eric) Liang (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

15:15
1467: TSB: Tiny Shared Block for Efficient DNN Deployment on NVCIM Accelerators
Yifan Qin (University of Notre Dame); Zheyu Yan (University of Notre Dame); Zixuan Pan (University of Notre Dame); Wujie Wen (North Carolina State University); X. Sharon Hu (University of Notre Dame); Yiyu Shi (University of Notre dame)

15:30 – 16:00
Coffee Break | Exhibits
Room: Salons D-E

16:00 – 18:00

Tutorial: Heterogeneous Integration: From physical layer to architecture and packaging

Room: Salons 1-3

This session explores the challenges and opportunities in designing heterogeneous architectures for compute- and data-intensive applications like AI, genomics, and graph analytics. Topics include physical layer considerations, in-memory computing architectures, advanced packaging techniques such as 2.5D integration, and open-source platforms for advancing heterogeneous system designs.

OpenSource Heterogeneous Chiplet-based Computing Architectures

Adrian Evans (Université Grenoble Alpes: CEA/LIST); César Fuguet (Université Grenoble Alpes: CEA/LIST); Davy Million (Université Grenoble Alpes: CEA/LIST)

Heterogeneous Manycore In-Memory Computing Architectures

Chukwufumnanya Ogbogu (Washington State University); Gaurav Narang (Washington State University); Biresh Joardar (University of Houston); Janardhan Rao Doppa (Washington State University); Partha Pratim (Washington State University)

Package Modeling and Analysis for Heterogeneous Integration

Christopher Bailey (Arizona State University); Leslie Hwang (Arizona State University); Pallavi Praful (Arizona State University)

16:00 – 18:00

Special Session: Towards Democratized and Reproducible AI for EDA Research: Open Datasets and Benchmarks in Various Aspects

Room: Lincoln/Holland/Columbia

Session Chair(s): Zhiyao Xie

This session tackles the challenge of limited access to high-quality design data. It introduces four open-source datasets and benchmarks across digital, analog design, LLM-aided IC design, and AI-assisted circuit data generation. These resources aim to democratize EDA research, fostering fair comparisons, reproducibility, and broader access for innovation in AI-EDA.

16:00

Generative Methods in EDA: Innovations in Dataset Generation and EDA Tool Assistants

Vidya Chhabria (Arizona State University); Bing-Yue Wu (Arizona State University); Utsav Sharma (New York University); Kishor Kunal (University of Minnesota); Austin Rovinski (New York University); Sachin Sapatnekar (University of Minnesota)

16:30
EDALearn: A Comprehensive RTL-to-Signoff EDA Benchmark for Democratized and Reproducible ML for EDA Research
Jingyu Pan (Duke University); Chen-Chia Chang (Duke University); Zhiyao Xie (Hong Kong University of Science and Technology); Yiran Chen (Duke University); Hai (Helen) Li (Duke University)

17:00
AnalogGym: An Open and Practical Testing Suite for Analog Circuit Synthesis
Jintao Li (University of Electronic Science and Technology of China); Haochang Zhi (Southeast University); Ruiyu Lyu (Fudan University); Wangzhen Li (Fudan University); Zhaori Bi (Fudan University); Keren Zhu (Fudan University); Yanhan Zeng (Guangzhou University); Weiwei Shan (National Center of Technology Innovation for EDA); Changhao Yan (Fudan University); Fan Yang (Fudan University); Yun Li (University of Electronic Science and Technology of China); Xuan Zeng (Fudan University)

17:30
OpenLLM-RTL: Open Dataset and Benchmark for LLM-Aided Design RTL Generation
Shang Liu (Hong Kong University of Science and Technology); Yao Lu (Hong Kong University of Science and Technology); Wenji Fang (Hong Kong University of Science and Technology); Mengming Li (Hong Kong University of Science and Technology); Zhiyao Xie (Hong Kong University of Science and Technology)

16:00 – 18:00
Special Session: Exploring Quantum Technologies in Practical Applications
Room: Salons A-C
Session Chair(s): Zhiding Liang

This session focuses on enhancing quantum circuit optimization and parameter setting for Quantum Approximate Optimization Algorithms, for solving combinatorial optimization problems. Topics include gate-level optimizations for improving circuit depth and fidelity, parameter setting techniques like concentration and transfer, and using GNNs for performance prediction of quantum optimization in maximum independent set problems.

16:00
Compiler Optimizations for QAOA
Yuchen Zhu (Rensselaer Polytechnic Institute); Yidong Zhou (Rensselaer Polytechnic Institute); Jinglei Cheng (Purdue University); Yuwei Jin (Rutgers University); Boxi Li (Forschungszentrum Jülich); Siyuan Niu (Lawrence Berkeley National Laboratory); Zhiding Liang (Rensselaer Polytechnic Institute)

16:30

GNN-Based Performance Prediction of Quantum Optimization of Maximum Independent Set

Atefeh Sohrabizadeh (University of Californiia at Los Angeles); Wan-Hsuan Lin (University of California at Los Angeles); Daniel Bochen Tan (University of California at Los Angeles); Madelyn Cain (Department of Physics: Harvard University); Sheng-Tao Wang (QuEra Computing Inc.); Mikhail D. Lukin (Department of Physics: Harvard University); Jason Cong (University of California atLos Angeles)

17:00

Parameter Setting Heuristics Make the Quantum Approximate Optimization Algorithm Suitable for the Early Fault-Tolerant Era

Zichang He (Global Technology Applied Research: JPMorganChase); Ruslan Shaydulin (Global Technology Applied Research: JPMorganChase); Dylan Herman (Global Technology Applied Research: JPMorganChase); Changhao Liu (Global Technology Applied Research: JPMorganChase); Rudy Raymond (Global Technology Applied Research: JPMorganChase); Shree Hari Sureshbabu (Global Technology Applied Research: JPMorganChase); Marco Pistoia (Global Technology Applied Research: JPMorganChase)

17:30

A Comparison on Constrain Encoding Methods for Quantum Approximate Optimization Algorithm

Yiwen Liu (FinQ Tech Inc.); Qingyue Jiao (University of Notre Dame); Yiyu Shi (University of Notre Dame); Ke Wan (FinQ Tech Inc.); Shangjie Guo (FinQ Tech Inc.)

16:00 - 17:00

Timing Prediction and Acceleration

Room: Skylands/Gateway

Session Chair(s): Tsung-Wei Huang

This session discusses the acceleration of timing analysis and the enhancement of timing prediction accuracy. The first paper in this session presents a CPU-GPU heterogeneous STA engine that handles generalized timing exceptions. The second paper presents a learning-based cross-corner timing signoff framework requiring only one RC corner. The last two papers address delay prediction and timing prediction problems leveraging deep learning and network models.

16:00

526: HeteroExcept: A CPU-GPU Heterogeneous Algorithm to Accelerate Exception-aware Static Timing Analysis

Zizheng Guo (Peking University); Zuodong Zhang (School of Integrated Circuits: Peking University); Wuxi Li (AMD); Tsung-Wei Huang (University of Wisconsin at Madison); Xizhe Shi (Peking University); Yufan Du (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

16:15

557: One-for-All: An Unified Learning-based Framework for Efficient Cross-Corner Timing Signoff

Linyu Zhu (Shanghai Jiao Tong University); Yichen Cai (Shanghai Jiao Tong University); Xinfei Guo (Shanghai Jiao Tong University)

16:30
630: CircuitSeer: RTL Post-PnR Delay Prediction via Coupling Functional and Structural Representation
Sanjay Gandham (University of Central Florida); Joe Walston (Synopsys); Sourav Samanta (Synopsys); Lingxiang Yin (University of Central Florida); Hao Zheng (University of Central Florida); Mingjie Lin (University of Central Florida); Stelios Diamantidis (Synopsys)

16:45
1590: Explainable and Layout-Aware Timing Prediction
Zhengyang Lyu (University of Science and Technology of China; Institute of Computing Technology: CAS); Xiaqing Li (Institute of Computing Technology: CAS); Zidong Du (Institute of Computing Technology: CAS); Qi Guo (Institute of Computing Technology: CAS); Huaping Chen (University of Science and Technology of China); Yunji Chen (Institute of Computing Technology: CAS)

16:00 - 17:00
How Much ML Can You Squeeze into Your Edge Device?
Room: Essex/Liberty
Session Chair(s): Hanbin Hu

Applying ML in edge devices has many important applications such as autonomous driving and medical diagnosis. This track includes four presentations considering how to efficiently and effectively apply ML in the edge. The first paper considers cooperative inferences on multiple edge devices. The second paper presents efficient HDR generation with the help of the user's visual attention. The third paper applies deep ensemble in edge devices with constrained resources. The last paper proposes a hardware-friendly softmax for power saving when applying ML in edge devices.

16:00
1551: RACI: A Resource-Aware Cooperative Inference Framework on Heterogeneous Edge Devices
zhenyu wang (Chongqing University); Ao Ren (Chongqing University); Duo Liu (Chongqing University); Haining Fang (Chongqing University); Jiaxing Shi (Chongqing University); Yujuan Tan (Chongqing University); Xianzhang Chen (Chongqing University)

16:15
927: Foveated HDR: Efficient HDR Content Generation on Edge Devices Leveraging User's Visual Attention
Ziyu Ying (Penn State); Sandeepa Bhuyan (The Pennsylvania State University); Yingtian Zhang (The Pennsylvania State University); Yan Kang (The Pennsylvania State University); Mahmut Taylan Kandemir (Penn State); Chita R. Das (Penn State University)

16:30
1337: Tiny Deep Ensemble: Uncertainty Estimation in Edge AI Accelerators via Ensembling Normalization Layers with Shared Weights
Soyed Tuhin Ahmed (KIT - Karlsruhe Institute of Technology: Karlsruhe: Germany); Mehdi Tahoori (Karlsruhe Institute of Technology)

16:45
1117: ConSmax: Hardware-Friendly Alternative Softmax with Learnable Parameters
Shiwei Liu (Google Research); Guanchen Tao (Department of Electrical Engineering and Computer Sciences: University of Michigan); Yifei Zou (Department of Electrical Engineering and Computer Sciences: University of Michigan); Derek Chow (Google Research); Zichen Fan (Department of Electrical Engineering and Computer Sciences: University of Michigan); Kauna Lei (Department of Electrical Engineering and Computer Sciences: University of Michigan); Bangfei Pan (Google Research); Dennis Sylvester (Department of Electrical Engineering and Computer Sciences: University of Michigan); Gregory Kielian (Google Research); Mehdi Saligane (Department of Electrical Engineering and Computer Sciences: University of Michigan)

16:00 - 17:00
Reliable emerging technologies
Room: Salons F-H
Session Chair(s): Shaahin Angizi

This session presents emerging technologies from 3D integration to wavelength routed optical network on chip (NOC) and NAND Flash: and the reliability issues associated with them.

16:00
1080: Efficient Ultra-Dense 3D IC Power Delivery and Cooling Using 3D Thermal Scaffolding
Dennis Rich (Stanford University); Tathagata Srimani (Stanford University); Mohamadali Malakoutian (Stanford University); Srabanti Chowdhury (Stanford University); Subhasish Mitra (Stanford University)

16:15
1512: Three Guides for Efficient Automatic Post-Fabrication Optimization of Modern NAND Flash Memory
Earl Kim (Samsung Electronics); Hyunuk Cho (POSTECH); Sungjun Cho (POSTECH); Myungsuk Kim (Kyungpook National University); Jisung Park (POSTECH (Pohang University of Science and Technology)); Jaeyong Jeong (Samsung Electronics); Eunkyoung Kim (Samsung Electronics); Sunghoi Hur (Samsung Electronics)

16:30
1323: Minimizing Worst-Case Data Transmission Cycles in Wavelength-Routed Optical NoC through Bandwidth Allocation
Liaoyuan Cheng (Technical University of Munich); Mengchu Li (Technical University of Munich); Tsun-Ming Tseng (Technical University of Munich); Ulf Schlichtmann (Technical University of Munich)

16:45

847: REMNA: Variation-Resilient and Energy-Efficient MLC FeFET Computing-in-Memory Using NAND Flash-Like Read and Adaptive Control

Taixin Li (Tsinghua University); Hongtao Zhong (Tsinghua University); Yixin Xu (The Pennsylvania State University); Vijaykrishnan Narayanan (Penn State University); Kai Ni (University of Notre Dame); Huazhong Yang (Tsinghua University); Thomas Kämpfe (Fraunhofer IPMS); Xueqing Li (Tsinghua University)

17:00 - 18:00

Innovative Approaches in Circuit Simulation: High-Fidelity Modeling: Optimization: and Parallelization

Room: Skylands/Gateway

Session Chair(s): Chen Quan

This session highlights groundbreaking advancements in circuit simulation. The first paper introduces a high-fidelity 2D warpage model for advanced packaging: achieving significant speedups and accuracy. The second paper presents an optimization technique that accelerates SPICE simulations using neural ODEs and graph convolution networks. The third paper extends spectral sparsification to nonlinear circuits: enhancing solver performance. The final paper offers a parallel-in-time exponential integrator method for faster transient simulations. Together: these works represent the forefront of high-fidelity modeling: optimization: and parallelization in circuit simulation.

17:00

1006: Efficient High-Fidelity Two-Dimensional Warpage Modeling for Advanced Packaging Analysis

Shao-Yu Lo (National Taiwan University); MaoZe Liu (National Taiwan University); Yao-Wen Chang (National Taiwan University)

17:15

1245: Pseudo Adjoint Optimization: Harnessing the Solution Curve for SPICE Acceleration

Jiatai Sun (China University of Petroleum: Beijing); Xiaru Zha (China University of Petroleum: Beijing); Chao Wang (Southesat University); Xiao Wu (Huada Empyrean Software Co. Ltd); Dan Niu (Southeast University); Wei W. Xing (The University of Sheffield); Zhou Jin (Super Scientific Software Laboratory: China University of Petroleum-Beijing)

17:30

1266: CSP: Comprehensive Sparsification Preconditioning for Nonlinear Circuit Simulation

Yuxuan Zhao (Super Scientific Software Laboratory: China University of Petroleum:Beijing); Xiaoyu Yang (Super Scientific Software Laboratory: China University of Petroleum-Beijing); Yinuo Bai (Super Scientific Software Laboratory: China University of Petroleum-Beijing); Lijie Zeng (Super Scientific Software Laboratory: China University of Petroleum-Beijing); Dan Niu (Southeast University); Weifeng Liu (China University of Petroleum-Beijing); Zhou Jin (Super Scientific Software Laboratory: China University of Petroleum-Beijing)

17:45
1236: EI-PIT: A Parallel-in-Time Exponential Integrator Method for Transient Linear Circuit Simulation
Hang Zhou (School of Microelectronics: Southern University of Science and Technology); Quan Chen (School of Microelectronics: Southern University of Science and Technology)

17:00 - 18:00
Analog: Analog: and More Analog Design using Your Favorite AI Algorithms
Room: Essex/Liberty
Session Chair(s): Orlando Arias

Applying AI in analog design has great potential for efficiency and effectiveness. This track includes four papers in this domain. The first paper combines Bayesian optimization with a large language model to utilize domain-specific knowledge for analog circuit design. The second paper applies an invertible graph generative model in the design of operational amplifiers. The third paper introduces a new physics-inspired NN model that can be implemented in low-power analog circuits. The last paper proposes new physics-informed neural networks for TSV electromigration analysis.

17:00
1432: ADO-LLM: Analog Design Bayesian Optimization with In-Context Learning of Large Language Models
Yuxuan Yin (University of California: Santa Barbara); Yu Wang (University of California: Santa Barbara); Boxun Xu (University of California: Santa Barbara); Peng Li (University of California: Santa Barbara)

17:15
753: TSO-Flow: A Topology Synthesis and Optimization Workflow for Operational Amplifiers with Invertible Graph Generative Model
Jinglin Han (Beihang University); Yuhao Leng (Beihang University); Xiuli Zhang (Beihang University); Peng Wang (Beihang University)

17:30
598: KirchhoffNet: A Scalable Ultra Fast Analog Neural Network
Zhengqi Gao (Dept. of EECS: MIT); Fan-Keng Sun (EECS: MIT); Ron Rohrer (CMU); Duane Boning (MIT)

17:45
942: Enforcing hard constraints in physics-informed learning for transient TSV electromigration analysis
Xiaoman Yang (Shanghai Jiao Tong University); Hai-Bao Chen (Department of Micro/Nano-electronics: Shanghai Jiao Tong University); Wenjie Zhu (Shanghai Jiao Tong University); Yuhan Zhang (University of Michigan - Ann Arbor); Yongkang Xue (Shanghai Jiao Tong University); Pengpeng Ren (Shanghai Jiao Tong University); Runsheng Wang (Peking University); Zhigang Ji (Shanghai Jiaotong University); Ru Huang (Peking University)

17:00 - 18:00

Emerging Technologies enabling Content Addressable Memories

Room: Salons F-H

Session Chair(s): Mehdi Tahoori

This session delves into the efficient design of Content Addressable Memories (CAM) using a computation in memory style using emerging technologies.

17:00

1263: TReCiM: Lower Power and Temperature-Resilient Multibit 2FeFET-1T Compute-in-Memory Design

Yifei Zhou (Zhejiang University); Thomas Kämpfe (Fraunhofer IPMS); Kai Ni (University of Notre Dame); Hussam Amrouch (Technical University of Munich (TUM)); Cheng Zhuo (Zhejiang University); Xunzhao Yin (Zhejiang University)

17:15

763: CAMSHAP: Accelerating Machine Learning Model Explainability with Analog CAM

John Moon (Hewlett Packard Labs); Giacomo Pedretti (Hewlett Packard Enterprise); Pedro Bruel (Hewlett Packard Labs); Sergey Serebryakov (Hewlett Packard Labs); Omar Eldash (Hewlett Packard Labs); Luca Buonanno (Hewlett Packard Labs); Catherine E. Graves (Hewlett Packard Labs); Paolo Faraboschi (Hewlett Packard Labs); Jim Ignowski (Hewlett Packard Labs)

17:30

1314: ShiftCAM: A Time-Domain Content Addressable Memory Utilizing Shifted Hamming Distance for Robust Genome Analysis

Peiyi He (The University of Hong Kong); Ruibin Mao (The University of Hong Kong); Keyi Shan (The University of Hong Kong); Yunwei Tong (The University of Hong Kong); Zhicheng XU (The University of Hong Kong); Muyuan Peng (The University of Hong Kong); Ruibang Luo (The University of Hong Kong); Can Li (The University of Hong Kong)

17:45

782: TAP-CAM: A Tunable Approximate Matching Engine based on Ferroelectric Content Addressable Memory

Chenyu Ni (Zhejiang University); Sijie Chen (Zhejiang University); Liu Liu (University of Notre Dame); Mohsen Imani (University of California: Irvine); Thomas Kämpfe (Fraunhofer IPMS); Kai Ni (University of Notre Dame); Michael Niemier (University of Notre Dame); Xiaobo Sharon Hu (University of Notre Dame); Cheng Zhuo (Zhejiang University); Xunzhao Yin (Zhejiang University)

18:00 – 19:30

Welcome Reception & SRC Poster Session

Room: Salons D-E

Technical Program – October 29, 2024

7:30 – 8:30
Registration
Room: Grand Ballroom Foyer

8:30 – 9:30
Keynote: The Future of Chip Design
Leon Stok: IBM
Room: Salons 4-8
Session Chair(s): Jinjun Xiong

9:30 – 10:00
Coffee Break | Exhibits
Room: Salons D-E

10:00 - 10:45
Processor: Memory: and Storage Designs
Room: Salons 1-3
Session Chair(s): Xiang Chen

This session contains three papers on system-level design and exploration: a classic problem that always needs new solutions for emerging devices and applications. The first one investigates the sustainability-performance trade-off of the dynamic instruction selection logic in superscalar processors. The second paper presents a tunable framework for generating memory-centric workloads that can be used to evaluate and validate memory subsystem designs. Finally: the third paper presents an approach for coupling heterogeneous management of modern high-density SSDs into the zone management at the host.

10:00
1058: Sustainable High-Performance Instruction Selection for Superscalar Processors
Saeideh Sheikhpour (Ghent University); David Christoph Metz (Norwegian University of Science and Technology); Erling Jullum (Norwegian University of Science and Technology (NTNU)); Magnus Själander (Norwegian University of Science and Technology); Lieven Eeckhout (Ghent University)

10:15
1350: A Framework for Explainable: Comprehensive: and Customizable Memory-Centric Workloads
Mohamed Abuelala (McMaster University); Mohamed Hassan (McMaster University)

10:30
529: ZnH2: Augmenting ZNS-based Storage System with Host-managed Heterogeneous Zones
Yingjia Wang (The Chinese University of Hong Kong); Lok Yin Chow (The Chinese University of Hong Kong); Xirui Nie (The Chinese University of Hong Kong); Yuhong Liang (The Chinese University of Hong Kong); Ming-Chang Yang (The Chinese University of Hong Kong)

10:00 - 10:45
Innovating Data Storage: Exploring Adaptive Indexing: Access Pattern Optimization: and Memory Longevity Enhancement for SSDs
Room: Lincoln/Holland/Columbia
Session Chair(s): Tsung-Yi Ho

This session presents three innovative methods for enhancing solid-state drives (SSDs)Including an adaptive learned index structure for spatial data: an access pattern-aware hybrid learning-based and conventional mapping scheme: and a data recovery strategy to rescue aging 3D TLC NAND flash cells.

10:00
679: ALISA: An Adaptive Learned Index Structure for Spatial Data on Solid-State Drives
Che-Wei Lin (National Yang Ming Chiao Tung University); Chun-Feng Wu (National Yang Ming Chiao Tung University)

10:15
1114: An Access Pattern-aware Hybrid Learning-based and Conventional Mapping for Solid-State Drives
Qian Wei (Shandong University); Xiaosu Guo (University of Texas at Dallas); Jie Wang (Shandong University); Zhaoyan Shen (Shandong University); Dongxiao Yu (Shandong University); Zhiping Jia (Shandong University); Bingzhe Li (University of Texas at Dallas)

10:30
799: CellRejuvo: Rescuing the Aging of 3D NAND Flash Cells with Dense-Sparse Cell Reprogramming
Han-Yu Liao (National Taiwan University of Science and Technology); Yi-Shen Chen (Department of Electronic and Computer Engineering: National Taiwan University of Science and Technology); Jen-Wei Hsieh (National Taiwan University of Science and Technology); Yuan-Hao Chang (Academia Sinica); Hung-Pin Chen (Innodisk Corporation)

10:00 - 10:45

Enhancing Simulation Efficiency through Multi-Core/GPU-Acceleration and Instruction-level Fault Injection

Room: Salons A-C

Session Chair(s): Nektarios Tsoutsos

This session focuses on simulation efficiency. The first paper introduces a powerful fault simulator for multi-core systems that first parallelizes simulation in the fault dimension and subsequently in the pattern dimension. The second paper introduces a GPU-accelerated logic simulator that uses two-phase simulation to enhance parallelism: achieving two to five times speed-ups over the state-of-the-art. The third paper explores innovative techniques for accurately and efficiently simulating processor faults by dynamically shifting the simulation from a register transfer level to a higher instruction level.

10:00

795: DDP-Fsim: Efficient and Scalable Fault Simulation for Deterministic Patterns with Two-Dimensional Parallelism

Feng Gu (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Mingjun Wang (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Jianan Mu (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Zizhen Liu (Institute of Computing Technology:Chinese Academy of Sciences); Jiaping Tang (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Hui Wang (CASTEST: Beijing); Yonghao Wang (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Jing Ye (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Huawei Li (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing); Xiaowei Li (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST: Beijing)

10:15

668: GL0AM: GPU Logic Simulation Using 0-Delay and Re-simulation Acceleration Method

Yanqing Zhang (NVIDIA); Haoxing Ren (NVIDIA Corporation); Brucek Khailany (NVIDIA)

10:30

1075: Accelerating Fault Injection for Validating Processor RTL Implementations

Yi Yuan (University of Texas at Austin); Derek Chiou (Microsoft/UT Austin)

10:00 - 10:45
Let LLMs Generate Your RTL Code!
Room: Skylands/Gateway
Session Chair(s): Sumit Jha

This session showcases papers that aim to enhance the productivity and efficiency of the design automation flows leveraging LLMs. The first two papers address code generation and the final paper addresses debugging. The first paper in the lineup utilizes techniques such as multi-modal program analysis: a search engine: and a cost-aware search algorithm to optimize RTL in terms of area-delay while incurring lower synthesis and verification time. The second paper focuses on establishing a dataset that can most effectively support code generation. OriGen: an open-source system with self-reflection and dataset augmentation is presented to improve the quality of open-source RTL datasets. The final paper of the session introduces the Make Each Iteration Count (MEIC) framework. Syntax and functional errors are detected and mitigated within this framework via the underlying LLM structure.

10:00
1188: RTLRewriter: Methodologies for Large Models aided RTL Code Optimization
Xufeng Yao (Chinese University of HongKong); Yiwen Wang (Huawei); Xing Li (Huawei); Yingzhao Lian (Huawei); Ran Chen (Huawei); Lei Chen (Huawei); Mingxuan Yuan (Huawei Noah's Ark Lab); Hong Xu (CUHK); Bei Yu (The Chinese University of Hong Kong)

10:15
1539: OriGen: Enhancing RTL Code Generation with Code-to-Code Augmentation and Self-Reflection
Fan Cui (Peking University); Chenyang Yin (Peking University); Kexing Zhou (Peking University); Youwei Xiao (School of Integrated Circuits: Peking University); Guangyu Sun (Peking University); Qiang Xu (The Chinese University of Hong Kong); Qipeng Guo (Shanghai Artificial Intelligence Laboratory); Demin Song (Shanghai Artificial Intelligence Laboratory); Dahua Lin (Shanghai Artificial Intelligence Laboratory); Xingcheng Zhang (Shanghai Artificial Intelligence Laboratory); Yun (Eric) Liang (Peking University)

10:30
1315: MEIC: Re-thinking RTL Debug Automation using LLMs
Ke Xu (Southeast University); Jialin Sun (Southeast University); Yuchen HU (Southeast University); Xinwei Fang (University of York); Weiwei Shan (Southeast University); Xi Wang (Tsinghua University); Zhe Jiang (South East University)

10:00 - 10:45
Architectural Mapping
Room: Essex/Liberty
Session Chair(s): Xiaofan Zhang

This session introduces cutting-edge techniques that push the boundaries of reconfigurable hardware based designs: addressing critical challenges in logic verification: irregular workload acceleration: and high-performance matrix computations. The presented papers showcase novel methodologies to the above-mentioned challenges and deliver significant improvements in performance: efficiency: and scalability across a range of applications.

10:00
962: DISC: Exploiting Data Parallelism of Non-Stencil Computations on CGRAs via Dynamic Iteration Scheduling
Yue Liang (Chongqing University); Di Mou (Chongqing University); Dajiang Liu (Chongqing University)

10:15
1220: MatFactory: A Framework for High-performance Matrix Factorization on FPGAs
Mingzhe Zhang (Tsinghua University); Xiaochen Hao (Peking University); Hongbo Rong (Intel Parallel Computing Lab); Wenguang Chen (Tsinghua University)

10:30
687: EasyPart: An Effective and Comprehensive Hypergraph Partitioner for FPGA-based Emulation
Shengbo Tong (Tsinghua University); Haoyuan Li (Tsinghua University); Jiahao Xu (Tsinghua University); Chunyan Pei (Tsinghua University); Wenjian Yu (Tsinghua University); Shengjun Liu (HyperSilicon); Jian Shen (HyperSilicon)

10:00 - 10:45
IR Drop and High-speed Link Analysis
Room: Salons F-H
Session Chair(s): Masanori Hashimoto

This session discusses modeling and estimation methods for addressing waveform distortion in high-speed links and dynamic/static IR drops. The first paper proposes a semi-vector-based assessment flow aimed at providing a more accurate estimation of worst-case peak power and IR drop. The second paper presents a CNN-based approach: along with comprehensive feature extraction: to speed up static IR drop estimation. The session concludes with LiTformer: a more accurate transformer-based model for high-speed link transmitters.

10:00
1121: Peak Power and Dynamic IR-drop Assessment via Waveform Augmenting
Yihan Wen (Beijing University of Technology); Juan Li (Beijing University of Technology); Bei Yu (The Chinese University of Hong Kong); Xiaoyi Wang (Unaffiliated Scholar)

10:15
1034: CFIRSTNET: Comprehensive Features for Static IR Drop Estimation with Neural Network
Yu-Tung Liu (National Yang Ming Chiao Tung University); Yu-Hao Cheng (National Yang Ming Chiao Tung University); Shao-Yu Wu (National Yang Ming Chiao Tung University); Hung-Ming Chen (National Yang Ming Chiao Tung University)

10:30
1212: LiTformer: Efficient Modeling and Analysis of High-Speed Link Transmitters Using Non-Autoregressive Transformer
Songyu Sun (Zhejiang University); Xiao Dong (Zhejiang University); YanLiang Sha (School of Microelectronics: Southern University of Science and Technology); Quan Chen (School of Microelectronics: Southern University of Science and Technology); Cheng Zhuo (Zhejiang University)

10:45 - 11:30
Efficient Machine Learning: from Cloud to Edge
Room: Salons 1-3
Session Chair(s): Caiwen Ding

As machine learning continues to advance: its deployment in different computation platform becomes increasingly complex. This session combines three interesting papers on this topic. The first one proposes an automated optimization framework to support the concurrent execution of multiple neural network models on GPUs. The second work manages multiple DNN workloads on heterogenous embedded devices: aiming to co-optimize the throughput and power efficiency. The third paper aims to achieve energy-efficient Federated Learning (FL): considering the energy constraints in both heterogeneous IoT devices and heterogeneous deep learning models.

10:45
856: GACER: Granularity-Aware ConcurrEncy Regulation for Multi-Tenant Deep Learning
Yongbo Yu (George Mason University); Fuxun Yu (Microsoft); Zhi Tian (George Mason University); Xiang Chen (George Mason University)

11:00
893: MapFormer: Attention-based multi-DNN manager for throughout & power co-optimization on embedded devices
Andreas Karatzas (Southern Illinois University Carbondale); Iraklis Anagnostopoulos (Southern Illinois University Carbondale)

11:15
1428: Towards Energy-Aware Federated Learning via MARL: A Dual-Selection Approach for Model and Client
Jun Xia (University of Notre Dame); Yi Zhang (east china normal university); Yiyu Shi (University of Notre dame)

10:45 - 11:30
EdgeML: Efficient and Private ML for the Edge
Room: Lincoln/Holland/Columbia
Session Chair(s): Bo Yuan

This session showcases three novel methods for efficient and private machine learning on edge devicesIncluding an adaptive private inference solution that allows a model to perform well across edge devices with diverse energy budgets: a pipeline inference framework that supports high-quality offline planning in heterogeneous edge environments: and a fog computing-based communication-efficient on-device learning framework that utilizes implicit neural representation to compress images/videos into neural network weights.

10:45
755: AdaPI: Facilitating DNN Model Adaptivity for Efficient Private Inference in Edge Computing
Tong Zhou (Northeastern University); Jiahui Zhao (University of Connecticut); Yukui Luo (University of Massachusetts Dartmouth); Xi Xie (University of Connecticut); Wujie Wen (North Carolina State University); Caiwen Ding (University of Connecticut); Xiaolin Xu (Northeastern University)

11:00
1218: EPipe: Pipeline Inference Framework with High-quality Offline Parallelism Planning for Heterogeneous Edge Devices
Yi Xiong (university of science and technology of China); Weihong Liu (University of Science and Technology of China); Rui Zhang (School of Software Engineering: Suzhou Institute for Advanced Research: University of Science and Technology of China); Yulong Zu (School of Software Engineering: Suzhou Institute for Advanced Research: University of Science and Technology of China); Zhu Zongwei (University of Science and Technology of China (USTC)); Xuehai Zhou (University of Science and Technology of China)

11:15
658: Residual-INR: Communication Efficient On-Device Learning Using Implicit Neural Representation
Hanqiu Chen (Georgia Institute of Technology); Xuebin Yao (Samsung Semiconductor Inc); Pradeep Subedi (Samsung Semiconductor Inc); Cong "Callie" Hao (Georgia Institute of Technology)

10:45 - 11:30

Advances in Verification through SAT Solving and Machine Learning

Room: Salons A-C

Session Chair(s): Daniel Grosse

This session presents several advancements in automated verification. The first paper introduces a new framework for checking the equivalence of flow-based computing circuits for in-memory processing: using helper variables to transform undirected graphs into directed ones for more effective satisfiability (SAT)-based verification. The second paper presents a novel approach for SAT solving by transforming it into a minimization problem and iteratively refining variable assignments using gradient descent. The third paper focuses on improving formal verification by integrating reinforcement learning to automatically generate conjectures from simulation traces.

10:45

880: Equivalence Checking for Flow-Based Computing using Iterative SAT Solving

Sven Thijssen (University of Central Florida); Muhammad Rashedul Haq Rashed (University of Central Florida); Md Rubel Ahmed (University of Central Florida); Suraj S. Singireddy (University of Texas at San Antonio); Sumit K. Jha (Florida International University); Rickard Ewetz (University of Central Florida)

11:00

996: DiffSAT: Differential MaxSAT Layer for SAT Solving

Yu Zhang (The Chinese University of Hong Kong); Hui-Ling Zhen (Huawei); Mingxuan Yuan (Huawei Noah's Ark Lab); Bei Yu (The Chinese University of Hong Kong)

11:15

715: Word-Level Augmentation of Formal Proof by Learning from Simulation Traces

Zhiyuan Yan (The Hong Kong Univeristy of Science and Technology(Guangzhou)); Hongce Zhang (The Hong Kong Univeristy of Science and Technology(Guangzhou))

10:45 - 11:30
New Benchmarks and Understanding Benchmarks using LLMs
Room: Skylands/Gateway
Session Chair(s): Siddharth Garg

This session introduces works in benchmark and dataset generation with AI technologies. The first paper in the session aims to improve the documentation process with a customized retrieval augmented generation and benchmarking tool to improve Q&A functionality: The second paper presents a benchmarking tool to assess Verilog generation A multi-modal generative model for code generation leverages both image and language to generate Verilog code benchmarks. The final paper in this session introduces benchmarking datasets for ML-driven floorplanning. These datasets reflect the complexities and hard constraints of SoCs effectively.

10:45
931: Customized Retrieval Augmented Generation and Benchmarking for EDA Tool Documentation QA
Yuan Pu (The Chinese University of Hong Kong); Zhuolun He (The Chinese University of Hong Kong); Tairu Qiu (ChatEDA Tech); Haoyuan WU (Shanghai AI Lab); Bei Yu (The Chinese University of Hong Kong)

11:00
685: Natural language is not enough: Benchmarking multi-modal generative AI for Verilog generation
Kaiyan Chang (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences: Beijing; University of Chinese Academy of Sciences); Zhirong Chen (Zhejiang University); Yunhao Zhou (Shanghai Jiao Tong University); Wenlong Zhu (Institute of Computing Technology:Chinese Academy of Sciences); kun wang (UCAS); Haobo Xu (State Key Laboratory of Computer Architecture: Institute of Computing Technology: Chinese Academy of Sciences; University of Chinese Academy of Sciences); Cangyuan Li (State Key Laboratory of Computer Architecture: Institute of Computing Technology: Chinese Academy of Sciences); Mengdi Wang (Institute of Computing Technology: Chinese Academy of Sciences); Shengwen Liang (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences: Beijing; University of Chinese Academy of Sciences); Huawei Li (Institute of Computing Technology: Chinese Academy of Sciences); yinhe han (Institute of Computing Technology:Chinese Academy of Sciences); Ying Wang (State Key Laboratory of Computer Architecture: Institute of Computing Technology: Chinese Academy of Sciences)

11:15
1410: FloorSet - a VLSI Floorplanning Dataset with Design Constraints of Real-World SOCs.
Uday Mallappa (Intel Labs); Hesham Mostafa (Intel Labs); Mikhail Galkin (Intel Labs); Mariano J. Phielipp (Intel Labs); Somdeb Majumdar (Intel Labs)

10:45 - 11:30

Applications and Architectures

Room: Essex/Liberty

Session Chair(s): Xiaofan Zhang

This session delves into state-of-the-art hardware acceleration techniques to address the performance: efficiency: and security demands of modern computing. The papers presented will explore innovative approaches for FPGA- and CGRA-based accelerations: sustainable hardware specialization: and security in multi-tenant FPGA environments. This session offers valuable insights into the evolving landscape of reconfigurable computing for emerging workloads.

10:45

1357: DoS-FPGA: Denial of Service on Cloud FPGAs via Coordinated Power Hammering

Hassan Nassar (Karlsruher Institut für Technologie); Philipp Machauer (KIT); Lars Bauer (Karlsruhe Institute of Technology); Dennis R. E. Gnad (Karlsruhe Institute of Technology); Mehdi Tahoori (Karlsruhe Institute of Technology); Joerg Henkel (KIT)

11:00

693: HG-PIPE: Vision Transformer Acceleration with Hybrid-Grained Pipeline

Qingyu Guo (School of Integrated Circuits: Peking University); Jiayong Wan (School of Integrated Circuits: Peking University); Songqiang Xu (School of Integrated Circuits: Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits: Peking University); Yuan Wang (Peking University)

11:15

1196: Sustainable Hardware Specialization

Pranav Dangi (National University of Singapore); Thilini Kaushalya Bandara (National University of Singapore); Saeideh Sheikhpour (Ghent University); Tulika Mitra (National University of Singapore); Lieven Eeckhout (Ghent University)

10:45 - 11:30

Machine Learning-based Design and Timing Optimization

Room: Salons F-H

Session Chair(s): Seokhyeong Kang

This session discusses design and timing optimization powered by machine leaning techniques. The first paper in this session proposes RankTuner: which is a ranking-based tool parameter tuning framework learning the dominant relationship between parameters. The second paper presents a new gate sizer based on LeakGAN which incorporates GAN with reinforcement learning. The last one addresses the challenges in timing-driven placement during the physical design flow of VLSI circuits with GNN.

10:45

1223: RankTuner: When Design Tool Parameter Tuning Meets Preference Bayesian Optimization

Peng Xu (The Chinese University of Hong Kong); Su Zheng (The Chinese University of Hong Kong); Yuyang Ye (Southeast University); Chen BAI (The Chinese University of Hong Kong); Siyuan Xu (Huawei Noah's Ark Lab); Hao Geng (ShanghaiTech University); Tsung-Yi Ho (The Chinese University of Hong Kong); Bei Yu (The Chinese University of Hong Kong)

11:00

1296: LAG-Sizer: A Novel Gate Sizer Based on Leak Generative Adversarial Network with Feature Fusion

Zhanhua Zhang (Southeast University); Wenjie Ding (Southeast university); Guoqing He (Southeast University); Peng Cao (Southeast University)

11:15

1166: A Physical and Timing Aware Placement Optimization Framework Based on Graph Neural Network

Wenjie Ding (Southeast university); Zhanhua Zhang (Southeast University); Guoqing He (Southeast University); Peng Cao (Southeast University)

11:30 – 13:00

CEDA Luncheon & Keynote

Sachin Sapatnekar: University of Minnesota

Room: Salons 4-8

Session Chair(s): Jinjun Xiong

13:00 – 14:30

Student Research Competition

Room: Salons 1-3

13:00 – 14:30

Special Session: Co-Designing NVM-Based Systems for Machine Learning Applications

Room: Lincoln/Holland/Columbia

Session Chair(s): Jörg Henkel

This session explores how non-volatile memory (NVM) technologies can enhance performance of embedded ML systems. Topics include using phase change memory (PCM) and memristors to optimize ML performance at the edge, accelerating large-scale search tasks with NVM-based in-memory computing, and overcoming challenges posed by memory reliability and latency to improve energy efficiency in resource-constrained environments.

13:00

Co-Designing NVM-based Systems for Machine Learning and In-memory Search Applications

Jörg Henkel (Karlsruhe Institute of Technology); Lokesh Siddhu (Karlsruhe Institute of Technology); Hassan Nassar (Karlsruhe Institute of Technology); Lars Bauer (-); Jian-Jia Chen (TU Dortmund); Christian Hakert (TU Dortmund); Tristan Seidl (TU Dortmund); Kuan Hsun Chen (University of Twente); Xiaobo Sharon Hu (University of Notre Dame); Mengyuan Li (University of Notre Dame); Chia-Lin Yang (National Taiwan University); Ming-Liang Wei (National Taiwan University);

13:22

Non-volatile Memory Technologies for Edge AI

Xiaoyu Sun (TSMC Corporate Research); Win-San Khwa (TSMC Corporate Research); Xiaochen Peng (TSMC Corporate Research); Meng-Fan Chang (TSMC Corporate Research); Kerem Akarvardar (TSMC Corporate Research)

13:44

Challenges and opportunities in accelerating large-scale search problems using NVM-based in-memory computing

X. Sharon Hu (University of Notre Dame)

14:06

Memory-Centric Deployment of Machine Learning Models

Chia-Lin Yang (National Taiwan University); Ming-Liang Wei (National Taiwan University)

13:00 – 14:30

Special Session: Delocalizing AI with Emerging Edge Intelligence (IoT/Internet)

Room: Salons A-C

Session Chair(s): Dharanidhar Dang
Haitong Li

This session addresses the challenges posed by the growing computational demands of AI algorithms and the rise of trillions of intelligent edge devices. It highlights innovations in circuit, system, and network-level designs, such as 2.5D photonic architectures, delocalized federated learning, and error correction techniques aimed at achieving secure, robust, and energy-efficient AI computing for the next generation of edge intelligence systems.

13:00

Large Scale Delocalized Federated Learning Over a Huge Diversity of Devices in Emerging Next-Generation Edge Intelligence Environments

Mahdi Morafah (University of California at San Diego); Hojin Chang (University of California at San Diego); Bill Lin (University of California at San Diego)

13:18

Error Correction and Detection for Analog AI Computing at the Edge

Anxiao Jiang (Texas A&M University College Station)

13:36

Codesigning 2.5D Photonic Accelerator for Distributed Transformer at the Edge

Dharanidhar Dang (University of Texas at San Antonio); Priyabrata Dash (University of Texas at San Antonio); Luqi Zheng (Purdue University); Haitong Li (Purdue University)

13:54

A Materials- and Devices-Centric Approach Toward Neuromorphic Computing

Shaloo Rakheja (University of Illinois)

14:12

Fault Tolerant In-Memory Computing based on Emerging Technologies for Ultra-Low Precision Edge AI Accelerators

Akul Malhotra (Purdue University); Sumeet Kumar Gupta (Purdue University)

13:00 - 13:45

Let AI Power Your Synthesis and Defect Analysis!

Room: Skylands/Gateway

Session Chair(s): Bei Yu

The advancement in AI is now boosting the EDA design flow. This track has three presentations that apply AI in chip placement: in HLS DSE: and in defect detection respectively.

13:00

632: The Power of Graph Signal Processing for Chip Placement Acceleration

Yiting Liu (Fudan University); Hai Zhou (Northwestern University); Jia Wang (Illinois Institute of Technology); Fan Yang (Fudan University); Xuan Zeng (Fudan University); Li Shang (fudan university)

13:15

885: Efficient Task Transfer for HLS DSE

Zijian Ding (University of California: Los Angeles); Atefeh Sohrabizadeh (University of California Los Angeles); Weikai Li (University of California: Los Angeles); Zongyue Qin (UCLA); Yizhou Sun (University of California: Los Angeles); Jason Cong (UCLA)

13:30

1004: SEM-CLIP: Precise Few-Shot Learning for Nanoscale Defect Detection in Scanning Electron Microscope Image

Qian Jin (Zhejiang University); Yuqi Jiang (Zhejiang University); Xudong Lu (Zhejiang University); Yumeng Liu (Zhejiang University); yining chen (Zhejiang University); Dawei Gao (Zhejiang University); QI SUN (Zhejiang University); Cheng Zhuo (Zhejiang University)

13:00 - 13:45
Revolutionizing AI with Low Power Accelerators: Emerging Design Trends
Room: Essex/Liberty
Session Chair(s): Shaahin Angizi
Dimitrios Soudris

This session includes three papers focusing on different aspects of low-power approximate and stochastic computing designs. The first paper introduces a novel mixed-signal Compute-in-Entropy hardware primitive aimed at addressing the substantial resource demands of Bayesian Neural Networks hardware implementations. The second paper presents an innovative stochastic computing-based method for neural network acceleration: which resolves critical issues found in existing accelerators. The third paper offers a comprehensive exploration of printed neural network accelerators: beginning with the analog-to-digital interface—an important area and major power sink in sensor processing applications—and extending to networks of ternary neurons and their implementation.

13:00
1349: Towards Uncertainty-Quantifiable Biomedical Intelligence: Mixed-signal Compute-in-Entropy for Bayesian Neural Networks
Likai Pei (University of Notre Dame); Yifan Qin (University of Notre Dame); Zephan M. Enciso (University of Notre Dame); Boyang Cheng (University of Notre Dame); Jianbo Liu (University of Notre Dame); Steven Davis (University of Notre Dame); Zhenge Jia (Shandong University); Michael Niemier (University of Notre Dame); Yiyu Shi (University of Notre Dame); X. Sharon Hu (University of Notre Dame); Ningyuan Cao (University of Notre Dame)

13:15
527: OSCA: End-to-end Serial Stochastic Computing Neural Acceleration with Fine-grained Scaling and Piecewise Activation
Yixuan Hu (Peking University); Yikang Jia (Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits: Peking University); Yuan Wang (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

13:30
923: Evolutionary Approximation of Ternary Neurons for On-sensor Printed Neural Networks
Vojtech Mrazek (Brno University of Technology); Argyris Kokkinis (Aristotle University of Thessaloniki); Panagiotis Papanikolaou (University of Michigan); Zdenek Vasicek (Brno University of Technology); Kostas Siozios (Department of Physics: Aristotle University of Thessaloniki); Georgios Tzimpragos (University of Michigan); Mehdi Tahoori (Karlsruhe Institute of Technology); Georgios Zervakis (University of Patras)

13:00 - 13:45

New Techniques in Analog Optimization: Bayesian Sensitivity: Hierarchical Placement: and AI-Driven 2.5D Chiplet Design

Room: Salons F-H

Session Chair(s): Sheldon Tan

This session highlights new advancements in analog optimization methods. The first paper is focused on optimization of analog sizing using Bayesian optimization combined with very accurate adjoint sensitivity analysis. The second paper presents an new approach for optimization of joint placement which exploits hierarchical information for analog/mixed-sigmal designs. The final paper uses AI to evaluate and optimize 2.5D chiplet bump pitch effects. These papers each present novel approaches to the optimized design of analog circuits.

13:00

966: Revisiting sensitivity-based analog sizing with derivative-aware Bayesian optimization and error-suppressed adjoint analysis

Ruiyu Lyu (School of Microelectronics: State Key Laboratory of Integrated Chips & System: Fudan University); Aidong Zhao (Fudan University); Yuan Meng (School of Microelectronics: State Key Laboratory of Integrated Chips & System: Fudan University); Zhaori Bi (Fudan University); Keren Zhu (The Chinese University of Hong Kong); Changhao Yan (Fudan University); Fan Yang (Fudan University); Dian Zhou (Fudan University); Xuan Zeng (Fudan University)

13:15

602: Joint Placement Optimization for Hierarchical Analog/Mixed-Signal Circuits

Xiaohan Gao (Peking University); Haoyi Zhang (Peking University); Bingyang Liu (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

13:30

1436: AI-Driven Evaluation and Optimization of Bump Pitch Effects on Chiplet and Interposer Design Quality

Seungmin Woo (Georgia Institute of Technology); Pruek Vanna-iampikul (Georgia Institute of Technology); Sung Kyu Lim (Georgia Tech)

13:45 - 14:30

Dive into the Design Space for Design Automation

Room: Skylands/Gateway

Session Chair(s): Iraklis Anagnostopoulos

This session delves into the critical methodologies for Design Space Exploration (DSE) in ASIC and FPGA design: focusing on innovative techniques like Bayesian Optimization: Reinforcement Learning: and Design Space Mining. The discussion will highlight the significance of these approaches in achieving efficient and effective DSE: which is paramount for optimizing performance and resource utilization in semiconductor design. Key papers to be presented include groundbreaking research on these topics: showcasing their practical applications and impact on the future of hardware design. Attendees will gain valuable insights into the latest advancements and best practices in DSE.

13:45

988: Is Vanilla Bayesian Optimization Enough for High-Dimensional Architecture Design Optimization?

Yuanhang Gao (Zhejiang University); Donger Luo (Shanghaitech University); Chen BAI (The Chinese University of Hong Kong); Bei Yu (The Chinese University of Hong Kong); Hao Geng (ShanghaiTech University); QI SUN (Zhejiang University); Cheng Zhuo (Zhejiang University)

14:00

998: TransLib: An Extensible Graph-Aware Library Framework for Automated Generation of Transformer Operators on FPGA

Yang Liu (Fudan University); Tianchen Wang (Fudan University); Yuxuan Dong (Fudan University); Zexu Zhang (Fudan University); Shun Li (Fudan University); Jun Yu (Fudan University); Kun Wang (Fudan University)

14:15

1084: MapTune: Advancing ASIC Technology Mapping via Reinforcement Learning Guided Library Tuning

Mingju Liu (University of Maryland College Park); Daniel Robinson (University of Utah); Yingjie Li (University of Maryland: College Park); Cunxi Yu (University of Maryland: College Park)

13:45 - 14:30

Machine Learning Innovations for Thermal and Power Optimization

Room: Essex/Liberty

Session Chair(s): Yukai Chen
Vojtech Mrazek

As the complexity and integration density of modern 3D-ICs and chiplet designs continue to grow: efficient thermal and power management becomes crucial for maintaining performance and reliability. This session not only delves into the latest ML advancements that address these challenges but also highlights their practical benefits. The session features pioneering research that leverages ML techniques to revolutionize the thermal analysis and power management processes. These sophisticated ML models significantly accelerate transient thermal prediction for full-chip designs: achieving speedups of several orders of magnitude over traditional simulators while ensuring long-term stability and minimal prediction errors. Additionally: this session showcases hierarchical power management frameworks designed for LLM chiplet designs: which integrate scalable simulation techniques with dynamic power delivery networks to adapt power strategies in real-time. This results in substantial energy savings and enhanced power efficiency: addressing the unique demands of chiplet-based systems. Join us to explore the intersection of machine learning: thermal management: and power optimization and discover how these innovations can be practically applied in your work: shaping the future of EDA.

13:45

972: FaStTherm: Fast and Stable Full-Chip Transient Thermal Predictor Considering Nonlinear Effects

Tianxiang Zhu (Peking University); Qipan Wang (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

14:00

974: Hierarchical Power Co-Optimization and Management for LLM Chiplet Designs

Yanchi Dong (Peking University); Xueping Liu (Peking University); Xiaochen Hao (Peking University); Yun (Eric) Liang (Peking University); Ru Huang (Peking University); Le Ye (Peking University); Tianyu Jia (Peking University)

14:15

830: ARO: Autoregressive Operator Learning for Transferable and Multi-fidelity 3D-IC Thermal Analysis With Active Learning

Mingyue Wang (Beihang University); Yuanqing Cheng (Beihang University); Weiheng Zeng (Beihang University); Zhenjie Lu (Shenzhen University); Vasilis F. Pavlidis (Aristotle University of Thessaloniki); Wei W. Xing (The University of Sheffield)

13:45 - 14:30

Routing and ECO Routing

Room: Salons F-H

Session Chair(s): Dirk Stroobandt
Shao-Yun Fang

Routing takes effort. This session starts with competing efforts for Global Routing in the first two presentations: HeLEM-GR and InstantGR. Routing takes even more effort: the third paper details a routing effort at ECO for DRV mitigation.

13:45

523: HeLEM-GR: Heterogeneous Global Routing with Linearized Exponential Multiplier Method

Chunyuan Zhao (Peking University); Zizheng Guo (Peking University); Rui Wang (Southwestern University of Finance and Economics); Zaiwen Wen (Peking University); Yun (Eric) Liang (Peking University); Yibo Lin (Peking University)

14:00

1239: InstantGR: Scalable GPU Parallelization for Global Routing

Shiju Lin (The Chinese University of Hong Kong); Liang Xiao (The Chinses University of Hong Kong); Jinwei Liu (The Chinese University of Hong Kong); Evangeline Young (The Chinese University of Hong Kong)

14:15

1501: An Effective ECO Methodology for Reducing Back-side Design Rule Violations in Double-sided Signal Routing

Che-Ping Tsai (National Tsing Hua University); Fang-Yu Hsu (National Tsing Hua University); Wai-Kei Mak (National Tsing Hua University); Ting-Chi Wang (National Tsing Hua University)

14:30 - 15:00

Coffee Break | Exhibits

Room: Salons D-E

15:00 - 15:45
Application Specific Accelerations
Room: Salons 1-3
Session Chair(s): Chengmo Yang

The proliferation of AI applications is pushing on accelerating AI on several hardware architectures and platforms. This session combines three interesting papers on this topic. The first paper proposes an efficient approach to accelerate the inference of Large Language Models on modern GPUs. The second work proposes an architecture to accelerate K-nearest neighbor graph construction in vector search. The third paper introduces an innovative approach to accelerate partial differential equation by exploiting value similarities.

15:00
1280: Fast and Efficient 2-bit LLM Inference on GPU: 2/4/16-bit in a Weight Matrix with Asynchronous Dequantization
Jinhao Li (Shanghai Jiao Tong University); Jiaming Xu (Shanghai Jiao Tong University); Shiyao Li (Tsinghua University); Shan Huang (Shanghai Jiao Tong University); Jun Liu (Shanghai Jiao Tong University); Yaoxiu Lian (Shanghai Jiao Tong University); Guohao Dai (Shanghai Jiao Tong University)

15:15
1140: AGC: A Unified Architecture for Accelerating K-Nearest Neighbor Graph Construction in Vector Search
Lei Dai (SKLP:Institute of Computing Technology:Chinese Academy of Sciences;University of Chinese Academy of Sciences); Ziming Yuan (State Key Laboratory of Processors: Institute of Computing Technology: Chinese Academy of Sciences: Beijing; University of Chinese Academy of Sciences); Shengwen Liang (State Key Lab of Processors: Institute of Computing Technology: Chinese Academy of Sciences: Beijing; University of Chinese Academy of Sciences); Wen Li (School of Computer and Information Technology: Shanxi University); Kaiwei Zou (Tsinghua University); Ying Wang (State Key Laboratory of Computer Architecture: Institute of Computing Technology: Chinese Academy of Sciences); Cheng Liu (Institute of Computing Technology: Chinese Academy of Sciences); Huawei Li (Institute of Computing Technology: Chinese Academy of Sciences); Xiaowei Li (ICT: Chinese Academy of Sciences)

15:30
965: Partial Differential Equation Acceleration by Exploiting Value Similarity
Zehua Li (Xi'an Jiaotong University); Kaisheng Ma (Tsinghua University)

15:00 - 15:45
Enabling Sustainable Next Generation IoT and CPS
Room: Lincoln/Holland/Columbia
Session Chair(s): Shaoyi Huang

This session presents three methods for next generation IoT and CPS applications Including an FPGA-based acceleration framework for path planning: an adversarial attack technique targeting hyperdimensional computing: and a hybrid power failure recovery scheme for intermittent computation.

15:00
789: A Sparsity-Aware Autonomous Path Planning Accelerator with Algorithm-Architecture Co-Design
Yanjun Zhang (Beijing Institute of Technology); Xiaoyu Niu (Beijing Institute of Technology); Yifan Zhang (University of California: Irvine); Hongzheng Tian (University of California: Irvine); Bo Yu (Shenzhen Institute of Artificial Intelligence and Robotics for Society); Shaoshan Liu (Shenzhen Institute of Artificial Intelligence and Robotics for Society); Sitao Huang (University of California: Irvine)

15:15
907: HDXpose: Harnessing Hyperdimensional Computing's Explainability for Adversarial Attacks
Fatemeh Asgarinejad (University of California at San Diego); Flavio Ponzina (University of California at San Diego); Onat Gungor (University of California at San Diego); Tajana Rosing (Univeristy of California at San Diego); Baris Aksanli (San Diego State University)

15:30
546: Hybrid Power Failure Recovery for Intermittent Computing
Gan Fang (Purdue University); Jongouk Choi (University of Central Florida); Changhee Jung (Purdue University)

15:00 - 15:45

Cycle-Accurate Timing Models: RISC-V Test Failure Analysis: and Low-Power Design Verification

Room: Salons A-C

Session Chair(s): Mohamed Hassan

The papers from this session address several challenges in simulation and verification. The first paper introduces a technique for generating cycle-accurate timing models for hardware accelerators from their register transfer level descriptions based on dependency analysis and constraint solving. The second paper details a modular: open-source framework designed to automate verification for RISC-V extensions: which reduces the need for manual result analysis by isolating failing instructions. The third paper introduces automatic algorithms to verify the partial set of retention registers identified to maintain correct design functionality when low-power circuits are powered off and on.

15:00

568: Automatic Generation of Timing Models from RTL for Hardware Accelerators

Yu Zeng (Princeton University); Aarti Gupta (Princeton University); Sharad Malik (Princeton University)

15:15

1013: Single Instruction Isolation for RISC-V Vector Test Failures

Manfred Schlaegl (Johannes Kepler University); Daniel Grosse (Johannes Kepler University Linz)

15:30

1061: Automatic Verification and Identification of Partial Retention Register Sets for Low-Power Designs

Yu-An Shih (Princeton University); Sharad Malik (Princeton University)

15:00 - 15:45

Bayesian Techniques for Software-Hardware Co-Optimization and Routing

Room: Skylands/Gateway

Session Chair(s): Sercan Aygun

This session focus on the application of Bayesian techniques in optimizing software-hardware co-design and routing strategies. Presentations will address multi-objective optimization for High-Density Processing In-Memory systems: reliability-aware routing in flow-based microfluidics: and the use of Bayesian-informed hyperdimensional learning for efficient data processing. The session aims to showcase how Bayesian methods can enhance the reliability and performance of complex computing systems.

15:00

699: Multi-Objective Software-Hardware Co-Optimization for HD-PIM via Noise-Aware Bayesian Optimization

Chien-Yi Yang (University of California at San Diego); Minxuan Zhou (University of California at San Diego); Flavio Ponzina (University of California at San Diego); Suraj Sathya Prakash (University of California at San Diego); Raid Ayoub (Intel corporation); Pietro Mercati (Intel Labs); Mahesh Subedar (Intel Labs); Tajana Rosing (University of California at San Diego)

15:15

985: RABER: Reliability-Aware Bayesian-Optimization-based Control Layer Escape Routing for Flow-based Microfluidics

Siyuan Liang (The Chinese University of Hong Kong); Rongliang Fu (The Chinese University of Hong Kong); Mengchu Li (Technical University of Munich); Tsun-Ming Tseng (Technical University of Munich); Ulf Schlichtmann (Technical University of Munich); Tsung-Yi Ho (The Chinese University of Hong Kong)

15:30

705: Bayesian-Informed Hyperdimensional Learning for Intelligent and Efficient Data Processing

Hamza Errahmouni Barkam (University Of California Irvine); Tamoghno Das (University of California: Irvine); Prathyush P. Poduval (ICS: UCI); SungHeon Jeong (UCI); Calvin Yeung (University of California Irvine); Mostafa A. Solitan (Alexandria university); Mohsen Imani (University of California Irvine)

15:00 - 15:45

Design Frameworks and Post-place Optimization

Room: Essex/Liberty

Session Chair(s): Hui-Ru Jiang

In this session: we explore two cutting-edge topics. In the first paper: the authors introduce SeGen: an automated framework for generating sequencing elements (such as flip-flops) in digital integrated circuits. SeGen covers a wide range of designs: outperforming human-crafted solutions. In the second paper: the authors discuss a novel reinforcement learning-based post-place optimization framework that simultaneously optimizes power: performance: and area. By dynamically selecting effective actions: this approach surpasses traditional co-optimization methods: achieving improved Pareto-frontier sets.

15:00

609: SeGen: Automatic Topology Generator for Sequencing Elements

Kyounghun Kang (Korea Advanced Institute of Science and Technology); Wanyeong Jung (KAIST)

15:15

796: Improving Timing & Power Trade-off in Post-place Optimization Using Multi-agent Reinforcement Learning

Jaemin Seo (Pohang University of Science and Technology (POSTECH)); Sejin Park (POSTECH); Seokhyeong Kang (Pohang University of Science and Technology)

15:00 - 15:45

Advances in Analog and RF Synthesis: Machine Learning Techniques and Thermal Analysis

Room: Salons F-H

Session Chair(s): Subed Lamichghane

This session presents the latest advances in analog and RF synthesis and modeling techniques. The first paper introduces a new machine learning-based synthesis flow: transitioning from S-parameters to layout for RF passive connector design. The second paper demonstrates a comprehensive analog hierarchical synthesis flow: supported by reinforcement learning in Python environments: applicable to both standard and low-temperature designs: and available as open source. The third paper discusses recent developments in machine learning-based full-chip thermal analysis: accounting for FinFET self-heating effects in RF power amplifiers.

15:00

1130: PulseRF: Physics Augmented ML Modeling and Synthesis for High-Frequency RFIC Design

Hyunsu Chae (University of Texas at Austin); Hao Yu (University of Texas at Austin); Sensen Li (University of Texas at Austin); David Z. Pan (University of Texas at Austin)

15:15

1500: Reinforcement Learning-Enhanced Cloud-Based Open Source Analog Circuit Generator for Standard and Cryogenic Temperatures in 130-nm and 180-nm OpenPDKs

Ali Hammoud (University of Michigan); Anhang Li (University of Michigan - Ann Arbor); Ayushman Tripathi (University of Michigan - Ann Arbor); Wen Tian (University of Michigan - Ann Arbor); Harsh Khandeparkar (University of Michigan - Ann Arbor); Ryan Wans (University of Michigan - Ann Arbor); Gregory Kielian (Google AI); Boris Murmann (University of Hawai'i); Dennis Sylvester (University of Michigan - Ann Arbor); Mehdi Saligane (University of Michigan - Ann Arbor)

15:30

1534: Analyzing the Impact of FinFET Self-Heating on the Performance of RF Power Amplifiers

Nibedita Karmokar (University of Minnesota); Sai-Wang Tam (NXP); Thanh Viet Dinh (NXP); Vidya A. Chhabria (Arizona State University); Ramesh Harjani (University of Minnesota); Sachin S. Sapatnekar (University of Minnesota)

19:00 - 21:45

ICCAD on Broadway

*Transportation will be provided

Technical Program – October 30, 2024

8:00 – 9:00
Registration
Room: Grand Ballroom Foyer

9:00 – 10:00
Keynote: NSF investments in Semiconductor and Microelectronics
Dilma Da Silva: US National Science Foundation
Room: Salons 4-8
Session Chair(s): Jinjun Xiong

10:00 – 10:30
Coffee Break | Exhibits
Room: Salons D-E

10:30 – 12:00
Top Picks Workshop | Part I
Room: Essex/Liberty

10:30 - 11:15
Microarchitecture Support for Security
Room: Salons 1-3
Session Chair(s): Michael Zuzak

This session focuses on security features and vulnerabilities at an architectural level. The first paper delves into security vulnerabilities in MRAM-based in-memory computing architectures and explores how these could be exploited to model extraction attacks. The second paper presents a key-switching accelerator architecture that delivers high-throughput for fully homomorphic encryption solution. The third paper in this session introduces an LLM-assistant that can detect and mitigate microarchitectural attacks in real-time.

10:30
707: On the Security Vulnerabilities of MRAM-based In-Memory Computing Architectures against Model Extraction Attacks
Saion K. Roy (University of Illinois at Urbana-Champaign); Naresh Shanbhag (University of Illinois at Urbana-Champaign)

10:45

777: An FPGA-based Key-Switching Accelerator with Ultra-High Throughput for FHE

Zhaojun Lu (School of Cyber Science and Engineering: Huazhong University of Science and Technology); Peng Xu (School of Cyber Science and Engineering: Huazhong University of Science and Technology); Yijie Wang (Huazhong University of Science and Technology); Yifan Yang (School of Cyber Science and Engineering: Huazhong University of Science and Technology); Qidong Chen (School of Cyber Science and Engineering: Huazhong University of Science and Technology); Weizong Yu (School of Cyber Science and Engineering: Huazhong University of Science and Technology); Gang Qu (University of Maryland: College Park)

11:00

1581: µLAM: A LLM-Powered Assistant for Real-Time Micro-architectural Attack Detection and Mitigation

Upasana Mandal (Indian Institute of Technology: Kharagpur); Shubhi Shukla (Indian Institute of Technology Kharagpur); Ayushi Rastogi (Indian Institute of Technology Kharagpur); Sarani Bhattacharya (Indian Institute of Technology Kharagpur); Debdeep Mukhopadhyay (Department of Computer Science and Engineering: Indian Institute of Technology Kharagpur)

10:30 - 11:15

New Research Developments in Synthesis

Room: Lincoln/Holland/Columbia

Session Chair(s): Victor Kravets

This session presents novel methods to optimize technology mapping and physical synthesis to create better chips. The first paper presents an improvement to cut choices in technology mapping based on machine learning. The second paper introduces a physical synthesis infrastructure based on a flexible intermediate representation. The third paper embraces technology mapping and physical synthesis with gate placement based on a novel wirelength-driven mapping algorithm.

10:30

1288: A Machine Learning Guided Cut Choices for ASIC Technology Mapping

Chandan Karfa (Indian Institute of Technology Guwahati); Chandrabhushan Reddy Chigarapall (IIT Guwahati); Harshwardhan Bhakkad (IIT Guwahati); Sukanta Bhattacharjee (Iitg); Animesh Basak Chowdhury (New York University)

10:45

522: RapidIR: A Practical Infrastructure for FPGA High-Level Physical Synthesis

Jason Lau (UCLA); Yuanlong Xiao (RapidStream Design AutomationInc.); Yutong Xie (RapidStream Design AutomationInc.); Yuze Chi (RapidStream Design AutomationInc.); Linghao Song (UCLA); Sihao Liu (UCLA); Shaojie Xiang (Cornell University); Michael Lo (UCLA); Zhiru Zhang (Cornell University); Jason Cong (UCLA); Licheng Guo (RapidStream Design AutomationInc.)

11:00
670: Physically Aware Synthesis Revisited: Guiding Technology Mapping with Primitive Logic Gate Placement
Hongyang Pan (Fudan university); Cunqing Lan (Fudan university); Yiting Liu (Fudan University); Zhiang Wang (University of California at San Diego); Li Shang (fudan university); Xuan Zeng (Fudan University); Fan Yang (Fudan University); Keren Zhu (The Chinese University of Hong Kong)

10:30 - 11:15
CTS and FPGA Routing
Room: Salons A-C
Session Chair(s): Ting-Chi Wang
Baris Taskin

The session starts with the presentation of a unique CTS methodology: OCTS: that synthesizes optical clock trees with EDA. The StarRoute and Potter papers compete in a virtual FPGA routing contest: enamored with novelties in space and parallelism exploration.

10:30
980: OCTS: An Optical Clock Tree Synthesis Methodology for 2.5D Systems
Aristotelis Tsekouras (Aristotle University of Thessaloniki); Georgios Kyriazidis (Harvard University); Gage Hills (Harvard University); Vasilis Pavlidis (University of Manchester)

10:45
808: AceRoute: Adaptive Compute-Efficient FPGA Routing with Pluggable Intra-Connection Bidirectional Exploration
Xinming Wei (Peking University); Ziyun Zhang (Peking University); Sunan Zou (School of Computer Science: Peking University); Kaiwen Sun (Deepoly); Jiahao Zhang (Peking University); Jiaxi Zhang (Peking University); Ping Fan (DeePoly Technology Inc.); Guojie Luo (Peking University)

11:00
1226: Potter: A Parallel Overlap-Tolerant Router for UltraScale FPGAs
Xinshi Zang (The Chinese University of Hong Kong); Wenhao Lin (The Chinese University of Hong Kong); Jinwei Liu (The Chinese University of Hong Kong); Evangeline Young (The Chinese University of Hong Kong)

10:30 - 11:15
PIM PIM PIM
Room: Skylands/Gateway
Session Chair(s): Sharon Hu

This session studies the challenges and recent developments in processing-in-memory (PIM) technologies, covering new devices, architecture designs and efficient mapping of arithmetic kernels.

10:30
614: NAND-Tree: A 3D NAND Flash Based Processing In Memory Accelerator for Tree-Based Models on Large-Scale Tabular Data
Hongtao Zhong (Tsinghua University); Taixin Li (Tsinghua University); Yiming Chen (Tsinghua University); Wenjun Tang (Tsinghua University); Juejian Wu (Tsinghua University); Huazhong Yang (Tsinghua University); Xueqing Li (Tsinghua University)

10:45
1156: A Processing-using-Memory Architecture for Commodity DRAM Devices with Enhanced Compatibility and Reliability
Hoon Shin (Seoul National University: Samsung Electronics); Rihae Park (Seoul National University); Jae W. Lee (Seoul National University)

11:00
1195: Towards Floating Point-Based Attention-Free LLM: Hybrid PIM with Non-Uniform Data Format and Reduced Multiplications
Lidong Guo (Tsinghua University); Zhenhua Zhu (Tsinghua University); Tengxuan Liu (Tsinghua University); Xuefei Ning (Tsinghua University); Shiyao Li (Tsinghua University); Guohao Dai (Shanghai Jiao Tong University); Huazhong Yang (Tsinghua University); Wangyang Fu (Tsinghua University); Yu Wang (Tsinghua University)

10:30 - 11:15

Real-Time AI: Co-Designing for the Edge

Room: Salons F-H

Session Chair(s): Yiyu Shi

This session dives into methods for optimizing real-time object detection and transformer network inference on edge devices. It emphasizes automated deployment strategies: software-hardware co-design: and the integration of binarized neural networks with FPGA technology to enhance computational efficiency and accuracy. The papers discuss end-to-end solutions and co-design approaches that significantly improve the performance and deployment of AI applications on edge platforms.

10:30

550: AyE-Edge: Automated Deployment Space Search Empowering Accuracy yet Efficient Real-Time Object Detection on the Edge

Chao Wu (Northeastern University); Yifan Gong (Northeastern University); Liangkai Liu (University of Michigan); Mengquan Li (Hunan University); Yushu Wu (Northeastern University); Xuan Shen (Northeastern University); Zhimin Li (Northeastern University); Geng Yuan (University of Georgia); Weisong Shi (Wayne State University); Yanzhi Wang (Northeastern University)

10:45

615: Edge-BiT: Software-Hardware Co-design for Optimizing Binarized Transformer Networks Inference on Edge FPGA

Shuai Zhou (Fudan University); Sisi Meng (Fudan University); Huinan Tian (Fudan University); Jun Yu (Fudan University); Kun Wang (Fudan University)

11:00

949: Co-Designing Binarized Transformer and Hardware Accelerator for Efficient End-to-End Edge Deployment

Yuhao Ji (Nanjing University); Chao Fang (Nanjing University); Shaobo Ma (Nanjing University); Haikuo Shao (Nanjing University); Zhongfeng Wang (Nanjing University)

11:15 - 12:00

Security by Design and Pre-silicon Security Assurance

Room: Salons 1-3

Session Chair(s): Yuntao Liu

This session explores the security design and security assurance at the microarchitectural level. The first paper introduces HybriDIFT, a dynamic information flow tracking approach that is memory aware. The second paper proposes a formal verification approach of CHERI at the register transfer level. The final paper in this session, eXpect, presents a systematic approach to analyze AXI implementations and detects functional and security violations.

11:15

571: HybriDIFT: Scalable Memory-Aware Dynamic Information Flow Tracking for Hardware

Flavien Solt (ETH Zurich); Kaveh Razavi (ETH Zurich)

11:30

852: VeriCHERI: Exhaustive Formal Security Verification of CHERI at the RTL

Anna Lena Duque Antón (RPTU Kaiserslautern-Landau); Johannes Müller (RPTU Kaiserslautern-Landau); Philipp Schmitz (RPTU Kaiserslautern-Landau); Tobias Jauch (RPTU Kaiserslautern-Landau); Alex Wezel (RPTU Kaiserslautern-Landau); Lucas Deutschmann (University of Kaiserslautern-Landau); Mohammad Rahmani Fadiheh (Technische Universität Kaiserslautern); Dominik Stoffel (TU Kaiserslautern); Wolfgang Kunz (TU Kaiserslautern)

11:45

1578: eXpect: On the Security Implications of Violations in AXI Implementations

Melisande Zonta (ETH Zürich); Andres Meza (UCSD); Nora Hinderling (ETH); Lucas Deutschmann (University of Kaiserslautern-Landau); Francesco Restuccia (University of California at San Diego); Ryan Kastner (UCSD); Shweta Shinde (ETH Zurich)

11:15 - 12:00

A New Life to Logic Synthesis

Room: Lincoln/Holland/Columbia

Session Chair(s): Yu-Guang Chen

This session presents innovative methods and applications for logic synthesis. The first paper presents a new architecture for circuit representation learning based on transformers. The second paper improves the understanding of circuit functionality for AIGs. The third paper applies rarity-reducing logic synthesis for hardware security.

11:15

1249: DeepGate3: Towards Scalable Circuit Representation Learning

Zhengyuan Shi (The Chinese University of Hong Kong); Ziyang Zheng (The Chinese University of Hong Kong); Sadaf Khan (The Chinese University of Hong Kong); Jianyuan Zhong (The Chinese University of Hong Kong); Min Li (Huawei Noah's Ark Lab); Qiang Xu (The Chinese University of Hong Kong)

11:30

1560: PolarGate: Breaking the Functionality Representation Bottleneck of And-Inverter Graph Neural Network

Jiawei Liu (Beijing University of Posts and Telecommunications); Jianwang Zhai (Beijing University of Posts and Telecommunications); Mingyu Zhao (Beijing University of Posts and Telecommunications); Zhe Lin (Sun Yat-sen University); Bei Yu (The Chinese University of Hong Kong); Chuan Shi (Beijing University of Posts and Telecommunications)

11:45

947: RareLS: Rarity-Reducing Logic Synthesis for Mitigating Hardware Trojan Threats

Chang Meng (EPFL); Mingfei Yu (EPFL); Hanyu Wang (ETH Zurich); Wayne Burleson (U Massachusetts Amherst); Giovanni De Micheli (École Polytechnique Fédérale de Lausanne (EPFL))

11:15 - 12:00

Machine Learning for P&R and Post-P&R

Room: Salons A-C

Session Chair(s): Evangeline Young
Tinghuan Chen

What does machine learning have to do the P&R and post-layout optimization? This session brings together latest approaches in GNNs: attention networks and reinforcement learning to solve some of the hardest problems: old and new: in these domains.

11:15

702: GAT-Steiner: Rectilinear Steiner Minimal Tree Prediction Using GNNs

Bugra Onal (University of California Santa Cruz); Eren Dogan (University of California: Santa Cruz); Muhammad Hadir Khan (University of California Santa Cruz); Matthew Guthaus (UC Santa Cruz)

11:30

1530: Placement Tomography-Based Routing Blockage Generation for DRV Hotspot Mitigation

Andrew Kahng (UCSD); Sayak Kundu (University of California: San Diego); Dooseok Yoon (University of California: San Diego)

11:45

1278: RL-Fill: Timing-Aware Fill Insertion Using Reinforcement Learning

Jinoh Cho (Pohang University of Science and Technology (POSTECH)); Seonghyeon Park (POSTECH); Jakang Lee (Pohang University of Science and Technology); Sung-Yun Lee (Pohang University of Science and Technology (POSTECH)); Jinmo Ahn (Postech); Seokhyeong Kang (Pohang University of Science and Technology)

11:15 - 12:00
Bringing Device Flavours
Room: Skylands/Gateway
Session Chair(s): Vidya Chhabria

Accurate device modelling and novel circuits unlock efficient architecture and design automation. The papers in this session focus on such contributions.

11:15
867: Multi-phase Coupled CMOS Ring Oscillator based Potts Machine
Yilmaz Ege Gonul (Drexel University); Baris Taskin (Drexel University)

11:30
1521: Accurate: Yet Scalable: A SPICE-based Design and Optimization Framework for eNVM based Analog In-memory Computing
S M Mojahidul Ahsan (The University of Kansas); Muhammad Sakib Shahriar (Ulkasemi Inc.); Mrittika Chowdhury (University of Mississippi); Tanvir Hossain (The University of Kansas); Md Sakib Hasan (University of Mississippi); Tamzidul Hoque (The University of Kansas)

11:15 - 12:00
IP: Side-Channels: and Acceleration
Room: Salons F-H
Session Chair(s): Xiaolong Guo
Michel Kinsey

This session presents a set of hardware approaches to protect static design information as well as dynamic application data. The first paper describes the use of encryption to protect the IP of a Spiking Neural Network implementation. The second paper presents a logic synthesis approach to improve resilience against side-channel attacks. The final paper proposes a hardware accelerator to efficiently compute a hash function used for Zero-Knowledge proofs.

11:15
915: SNNGX: Securing Spiking Neural Networks with Genetic XOR Encryption on RRAM-based Neuromorphic Accelerator
Kwun Hang WONG (University of Hong Kong); Songqi Wang (University of Hong Kong); Wei Huang (University of Hong Kong); Xinyuan Zhang (University of Hong Kong); Yangu He (University of Hong Kong); Karl M.H. Lai (University of Hong Kong); Yuzhong Jiao (AI Chip Center for Emerging Smart Systems (ACCESS)); Ning Lin (The University of Hong Kong); Xiaojuan Qi (University of Hong Kong); Xiaoming Chen (Institute of Computing Technology: Chinese Academy of Sciences); Zhongrui Wang (University of Hong Kong)

11:30
1472: ASCENT: Amplifying Power Side-Channel Resilience via Learning & Monte-Carlo Tree Search
Jitendra Bhandari (New York University); Animesh Basak Chowdhury (New York University); Ozgur Sinanoglu (New York University Abu Dhabi); Siddharth Garg (New York University); Ramesh Karri (NYU); Johann Knechtel (New York University Abu Dhabi)

11:45
1361: AMAZE: Accelerated MiMC Hardware Architecture for Zero-Knowledge Applications on the Edge
Anees Ahmed (Arizona State University); Nojan Sheybani (University of California at San Diego); Davi De Almeida (Arizona State University); Tengkai Gong (University of California at San Diego); Nges Njungle (Arizona State University); Michel Kinsy (Arizona State University); Farinaz Koushanfar (University of California at San Diego)

12:00 – 13:30
Lunch
Room: Salons 4-8

13:30 – 15:30
Tutorial: Advanced Sparse Linear Solver for Transistor-Level Circuit Simulation
Room: Salons 1-3

This session highlights GPU-based sparse LU factorization for circuit simulation. It covers GLU 3.0, which improves kernel efficiency and resource allocation, significantly enhancing performance for large-scale circuit simulations using GPUs.

13:30 – 15:30
Special Session: 2024 CAD Contests at the ICCAD
Room: Lincoln/Holland/Columbia
Session Chair(s): Shao-Yun Fang
Yi-Yu Liu
Chung-Kuan Cheng
Tsun-Ming Tseng

13:30
Overview of 2024 CAD contest at ICCAD
Shao-Yun Fang (National Taiwan University of Science and Technology); Yi-Yu Liu (National Taiwan University of Science and Technology); Chung-Kuan Cheng (University of California at San Diego); Tsun-Ming Tseng (Technical University of Munich)

13:40

2024 ICCAD CAD Contest Problem A: Reinforcement Logic Optimization for a General Cost Function Chung-Han Chou (Cadence TaiwanInc.); Chih-Jen (Jacky) Hsu (Cadence Design SystemInc.); Chi-An (Rocky) Wu (Cadence TaiwanInc.); Kuan-Hua Tu (Cadence TaiwanInc.); Kwangsoo Han (Cadence Design SystemInc.); Zhou Li (Cadence Design SystemInc.)

14:10

2024 ICCAD CAD Contest Problem B: Power and Timing Optimization Using Multibit Flip-Flop

Sheng-Wei Yang (SynopsysInc.); Jhih-Wei Hsu (SynopsysInc.); Ting Wei Li (SynopsysInc.); Tzu-Hsuan Chen (SynopsysInc.); Chin-Fang Cindy Shen (SynopsysInc.)

14:40

2024 ICCAD CAD Contest Problem C: Scalable Logic Gate Sizing using ML Techniques and GPU Acceleration

Bing-Yue Wu (Arizona State University); Rongjian Liang (NVIDIA); Geraldo Pradipta (NVIDIA); Anthony Agnesina (NVIDIA); Haoxing Ren (NVIDIA); Vidya A. Chhabria (Arizona State University)

15:10

Strengthening the Foundations for IC Physical Design and ML EDA Research

Vidya A. Chhabria (Arizona State University); Vikram Gopalakrishnan (Arizona State University); Andrew B. Kahng (University of California San Diego); Sayak Kundu (University of California San Diego); Zhiang Zhiang (University of California San Diego); Bing-Yue Wu (Arizona State University); Dooseok Yoon (University of California San Diego)

13:30 – 15:30

Top Picks Workshop | Part II

Room: Essex/Liberty

13:30 - 14:30

Swift LLMs: Easier Design: Faster Inference

Room: Salons A-C

Session Chair(s): Jun Xia

This session explores various innovations in hardware design tailored for large language model (LLM) optimization and acceleration. The topics discussed include developing agile frameworks for LLM accelerator creation: speculative scheduling to enhance LLM serving: and dynamic approaches for efficient token processing in parallel decoding. Additionally: there's an emphasis on integrating FPGA-based solutions to handle unstructured sparsity in large language models: showcasing advancements in both theoretical and practical aspects of LLM deployment.

13:30

William J. McCalla ICCAD Best Paper Award - Frontend

1005: An Agile Framework for Efficient LLM Accelerator Development and Model Inference

Lvcheng Chen (Zhejiang University); Ying Wu (Zhejiang University); Chenyi Wen (Zhejiang University); Shizhang Wang (Hubei University of Technology); Li Zhang (Hubei University of Technology); Bei Yu (The Chinese University of Hong Kong); QI SUN (Zhejiang University); Cheng Zhuo (Zhejiang University)

13:45

575: ALISE: Accelerating Large Language Model Serving with Speculative Scheduling

Youpeng Zhao (University of Central Florida); Jun Wang (University of Central Florida)

14:00

757: ProPD: Dynamic Token Tree Pruning and Generation for LLM Parallel Decoding

Shuzhang Zhong (Peking University); Zebin Yang (Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits: Peking University); Ruihao Gong (SenseTime); Runsheng Wang (Peking University); Ru Huang (Peking University)

14:15

1082: ChatOPU: An FPGA-based Overlay Processor for Large Language Models with Unstructured Sparsity

Tiandong Zhao (University of California: Los Angeles); Shaoqiang Lu (Shanghai Jiao Tong University: Shanghai: China; Ningbo Institute of Digital Twin: Eastern Institute of Technology: Ningbo: China); Chen Wu (Ningbo Institute of Digital Twin: Ningbo: China); Lei He (Ningbo Institute of Digital Twin: Eastern Institute of Technology: Ningbo: China; University of California: Los Angeles)

13:30 - 14:30
Time is Limited: Fast and Secure Neural Network Accelerators
Room: Skylands/Gateway
Session Chair(s): Peipei Zhou

This session examines strategies for improving the reliability and performance of neural network systems in critical applications like self-driving vehicles. It discusses techniques such as latency-constrained scheduling to ensure timely and reliable perception: eager gradient prediction to enhance training efficiency of attention mechanisms: and methods for reinforcing DNN accelerator integrity through selective and permuted recomputation. These approaches aim to boost both the speed and accuracy of AI applications: ensuring safer and more efficient operations.

13:30
741: LACO: A Latency-Constraint Offline Neural Network Scheduler towards Reliable Self-Driving Perception
Zhanhong Tan (State Key Laboratory of Intelligent Vehicle Safety Technology); Zijian Zhu (Tsinghua University); Mengdi Wu (Tsinghua University); Kaisheng Ma (Tsinghua University)

13:45
1148: OFT: An accelerator with eager gradient prediction for attention training
Miao Wang (Northwestern Polytechnical University); Shengbing Zhang (Northwestern Polytechnical University); Sijia Wang (Northwestern Polytechnical University); Zhao Yang (Chang'an University); Meng Zhang (Northwestern Polytechnical University)

14:00
894: Enhancing DNN Accelerator Integrity via Selective and Permuted Recomputation
Jhon Ordoñez (University of Delaware); Chengmo Yang (University of Delaware)

13:30 - 14:30
Private Machine Learning Inference
Room: Salons F-H
Session Chair(s): Dean Sullivan
Geng Yuan

This session presents a series of approaches to perform machine learning inference privately and efficiently. The first paper reduces communication costs in distributed inference by co-optimizing quantization communication protocol constraints. The second: third: and fourth papers all employ homomorphic encryption to establish private inference. The second paper describes an approach to automatically generate efficient kernels for homomorphic encryption. The solution presented in the third paper combines homomorphic encryption with Garbled Circuits. The fourth paper focuses on reducing the latency induced by homomorphism in the convolution layers of CNNs.

13:30
596: PrivQuant: Communication-Efficient Private Inference with Quantized Network/Protocol Co-Optimization
Tianshi Xu (Peking University); Shuzhang Zhong (Peking University); Wenxuan Zeng (Peking University); Runsheng Wang (Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits: Peking University)

13:45
973: FlexHE: A flexible Kernel Generation Framework for Homomorphic Encryption-Based Private Inference
Jiangrui Yu (Peking University); Wenxuan Zeng (Peking University); Tianshi Xu (Peking University); Renze Chen (Peking University); Yun (Eric) Liang (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits: Peking University)

14:00
1237: APINT: A Full-Stack Framework for Acceleration of Privacy-Preserving Inference of Transformers based on Garbled Circuits
Hyunjun Cho (Korea Advanced Institute of Science and Technology (KAIST)); Jaeho Jeon (KAIST); Jaehoon Heo (KAIST); Joo-Young Kim (KAIST)

14:15
1552: Hyena: Optimizing Homomorphically Encrypted Convolution for Private CNN Inference
Hyeri Roh (Seoul National University); Woo-Seok Choi (Seoul National University)

14:30 - 15:30

CIM is on the Run: Sparser and More Robust Designs

Room: Salons A-C

Session Chair(s): Umamaheswara Rao Tida

This session focuses on architectures and strategies for enhancing in-memory computation for AI applications. Discussions include leveraging hybrid RRAM-SRAM designs for accelerating sparse transformers and eigen-decomposition: along with probabilistic approximation techniques to optimize sparsity-centric computation. Furthermore: the session explores variation-resilient memory solutions and specialized accelerators for processing complex data structures like voxel-based point clouds: pushing the boundaries of efficiency and performance in AI hardware.

14:30

594: AESHA: Accelerating Eigen-decomposition-based Sparse Transformer with Hybrid RRAM-SRAM Architecture

Xuliang Yu (Zhejiang University); Tianwei Ni (Zhejiang University); Xinsong Sheng (Zhejiang University); Yun Pan (Zhejiang University); Lei He (University of California: Los Angeles); Liang Zhao (Zhejiang University)

14:45

802: PACiM: A Sparsity-Centric Hybrid Compute-in-Memory Architecture via Probabilistic Approximation

Wenlun Zhang (Keio University); Shimpei Ando (Keio University); Yung-Chin Chen (National Taiwan University); Satomi Miyagi (Keio University); Shinya Takamaeda-Yamazaki (The University of Tokyo); Kentaro Yoshioka (Keio University)

15:00

1003: ReSCIM: Variation-Resilient High Weight-Loading Bandwidth In-Memory Computation Based on Fine-Grained Hybrid Integration of Multi-Level ReRAM and SRAM Cells

Xiaomeng WANG (The Hong Kong University of Science and Technology); Jingyu HE (The Hong Kong University of Science and Technology); Kunming SHAO (The Hong Kong University of Science and Technology); Jiakun ZHENG (The Hong Kong University of Science and Technology); Fengshi TIAN (The Hong Kong university of Science and Technology); Tim Kwang-Ting CHENG (The Hong Kong University of Science and Technology); Chi-Ying TSUI (The Hong Kong University of Science and Technology)

15:15

838: Voxel-CIM: An Efficient Compute-in-Memory Accelerator for Voxel-based Point Cloud Neural Networks

Xipeng Lin (The Hong Kong University of Science and Technology (Guangzhou)); Shanshi Huang (Hong Kong University of Science and Technology (Guangzhou)); Hongwu Jiang (The Hong Kong University of Science and Technology (Guangzhou))

14:30 - 15:30
Treasures in the Graphs: Efficient Designs for GNNs
Room: Skylands/Gateway
Session Chair(s): Cong Hao

This session delves into innovative approaches for optimizing graph neural network (GNN) computations on FPGA platforms. It features discussions on leveraging advanced dataflow techniques and high-bandwidth memory to accelerate sparse matrix multiplications: algorithm-hardware co-design for efficient large-scale GNN training: and the development of memory-optimized processors tailored for parallel graph mining. These papers highlight the ongoing advancements in hardware configurations and algorithmic strategies to enhance the performance and scalability of GNN applications.

14:30
1169: Leda: Leveraging Tiling Dataflow to Accelerate SpMM on HBM-Equipped FPGAs for GNNs
Enxin Yi (Super Scientific Software Laboratory: China University of Petroleum-Beijing); Jiarui Bai (Super Scientific Software Laboratory: China University of Petroleum-Beijing); Yijie Nie (Super Scientific Software Laboratory: China University of Petroleum-Beijing); Dan Niu (Southeast University); Zhou Jin (Super Scientific Software Laboratory: China University of Petroleum-Beijing); Weifeng Liu (Super Scientific Software Laboratory: China University of Petroleum-Beijing)

14:45
1561: CoCoA: Algorithm-Hardware Co-Design for Large-Scale GNN Training using Compressed Graph
Yunki Han (KAIST); Jaekang Shin (Korea Advanced Institute of Science and Technology (KAIST)); Gunhee Park (Samsung Electronics); Lee-Sup Kim (KAIST)

15:00
968: FLOP: A Flexible Memory-Optimized Processor for Parallel Graph Mining on FPGA
Guoyu Li (Fudan University); Runzhou Zhang (Fudan University); Jun Yu (Fudan University); Kun Wang (Fudan University)

14:30 - 15:30

More than Matrix Multiplication: Efficient Designs for Neural Networks

Room: Salons F-H

Session Chair(s): Meng Li

This session explores in-memory computing architectures designed to enhance the efficiency and throughput of deep neural networks and transformers. Topics include a reprogramming-free RRAM-based approach for optimizing deep neural network computations through basis combination: a transposable digital SRAM architecture tailored for energy-efficient transformer acceleration: and a matrix multiplication accelerator that supports various levels of sparsity. These advancements underscore efforts to reduce energy consumption while boosting the computational performance of AI accelerators.

14:30

828: BasisN: Reprogramming-Free RRAM-Based In-Memory-Computing by Basis Combination for Deep Neural Networks

Amro Eldebiky (Technical University of Munich); Grace Li Zhang (TU Darmstadt); Xunzhao Yin (Zhejiang University); Cheng Zhuo (Zhejiang University); Ing-Chao Lin (National Cheng Kung University); Ulf Schlichtmann (Technical University of Munich); Bing Li (Technical University of Munich)

14:45

1579: TP-DCIM: Transposable Digital SRAM CIM Architecture for Energy-Efficient and High Throughput Transformer Acceleration

Junwoo Park (Korea University); Kyeongho Lee (korea univ.); Jongsun Park (Korea University)

15:00

599: FSMM: An Efficient Matrix Multiplication Accelerator Supporting Flexible Sparsity

Yuxuan Qiao (Fudan University); Fan Yang (Fudan University); Yecheng Zhang (State Key Laboratory of Mobile Network and Mobile Multimedia Technology: ZTE Corporation); Xiankui Xiong (State Key Laboratory of Mobile Network and Mobile Multimedia Technology: ZTE Corporation); Xiao Yao (State Key Laboratory of Mobile Network and Mobile Multimedia Technology: ZTE Corporation); Haidong Yao (State Key Laboratory of Mobile Network and Mobile Multimedia Technology: ZTE Corporation)

16:00 - 17:00

Top Picks Workshop | Part III

Room: Essex/Liberty

16:00 - 17:00

Side Channels and Trojans

Room: Salons 1-3

Session Chair(s): Amin Rezaei

This session presents the latest research in physical attacks and hardware Trojans. The first paper, RandOhm, explores mitigating impedance side channel attacks using randomized configuration of circuits. The second paper presents an interesting approach to prevent insertion of layout-level hardware Trojans at the Physical Design level. The next paper presents a novel approach that is capable for detecting rowhammer, rowpress, and leakage detection in DRAMs. The last paper details how optical probing attacks could be detected and mitigated.

16:00

724: RandOhm: Mitigating Impedance Side-channel Attacks using Randomized Circuit Configurations

Saleh Khalaj Monfared (Worcester Polytechnic Institute (WPI)); Domenic Forte (University of Florida); Shahin Tajik (Worcester Polytechnic Institute)

16:15

806: Layout-level Hardware Trojan Prevention in the Context of Physical Design

Xingyu Tong (Fudan University); Guohao Chen (Fudan University); Min Wei (Fudan University); Zhijie Cai (Fudan University); Peng Zou (Shanghai LEDA Technology Co.: Ltd); Zhifeng Lin (Fuzhou University); Jianli Chen (Fudan University)

16:30

1403: A Built-In Integrated Rowhammer: Rowpress: and Leakage Detection Sensor for DRAM

Nezam Rohbani (Institute for Research in Fundamental Sciences (IPM)); Rouzbeh Pirayadi (Sharif Univerisyt of Technology); Mohammad Arman Soleimani (Sharif University of Technology); Adrian Cristal Kestelman (Barcelona Supercomputing Center); Osman Unsal (Barcelona supercomputing center); Hamid Sarbazi-Azad (Sharif U of Tech)

16:45

1492: LaserEscape: Detecting and Mitigating Optical Probing Attacks

Saleh Khalaj Monfared (Worcester Polytechnic Institute (WPI)); Kyle Mitard (Worcester Polytechnic Institute); Andrew Cannon (University of Florida); Domenic Forte (University of Florida); Shahin Tajik (Worcester Polytechnic Institute)

16:00 - 17:00

Advances in High-Level Synthesis and Optimized Components

Room: Lincoln/Holland/Columbia

Session Chair(s): Shigeru Yamashita

This session presents the latest developments in high-level synthesis and the creation of optimized micro-architectural components. The first paper presents a novel intermediate representation to optimize the dynamic scheduling and optimization of memory operations. The second paper introduces an approach based on LLM for high-level synthesis. The third paper discusses a method to estimate circuit metrics at higher levels of abstraction. The fourth paper presents a framework to optimize the generation of multipliers and MACs.

16:00

652: R-HLS: An IR for Dynamic High-Level Synthesis and Memory Disambiguation based on Regions and State Edges

David Christoph Metz (Norwegian University of Science and Technology); Nico Reissmann (Independent Researcher); Magnus Själander (Norwegian University of Science and Technology)

16:15

1222: HLSPilot: LLM-based High-Level Synthesis

Chenwei Xiong (Institute of Computing Technology: Chinese Academy of Sciences); Cheng Liu (Institute of Computing Technology: Chinese Academy of Sciences); Huawei Li (Institute of Computing Technology: Chinese Academy of Sciences); Xiaowei Li (ICT: Chinese Academy of Sciences)

16:30

1240: Balor: HLS Source Code Evaluator Based on Custom Graphs and Hierarchical GNNs

Emmet Murphy (ETH Zurich); Lana Josipovic (ETH Zurich)

16:45

1265: UFO-MAC: A Unified Framework for Optimization of High-Performance Multipliers and Multiply-Accumulators

Dongsheng Zuo (The Hong Kong University of Science and Technology (Guangzhou)); Jiadong ZHU (The Hong Kong University of Science and Technology (Guangzhou)); Chenglin Li (The Hong Kong University of Science and Technology (Guangzhou)); Yuzhe Ma (The Hong Kong University of Science and Technology (Guangzhou))

16:00 - 17:00

Innovations in Neuromorphic Hardware and 3D Integration

Room: Salons A-C

Session Chair(s): Shaahin Angizi

This session explores the latest advancements in neuromorphic hardware: with a particular focus on 3D integration and novel computational approaches. Presentations will cover topics such as spiking transformer hardware accelerators: neural architecture search for robust printed circuits: and the design of spatiotemporal denoising filters for dynamic vision sensors. The session will also discuss the synergy between Liquid State Machines and RRAM-based accelerators: highlighting the potential of analog-digital computation.

16:00

1516: Spiking Transformer Hardware Accelerators in 3D Integration

Boxun Xu (University of California: Santa Barbara); Junyoung Hwang (Georgia Institute of Technology); Pruek Vanna-iampikul (Georgia Institute of Technology); Sung Kyu Lim (Georgia Tech); Peng Li (University of California: Santa Barbara)

16:15

1397: Neural Architecture Search for Highly Bespoke Robust Printed Neuromorphic Circuits

Priyanjana Pal (Karlsruhe Institute of Technology); Haibin Zhao (Karlsruhe Institute of Technology); Tara Gheshlaghi (Karlsruhe Institute of Technology); Michael Hefenbrock (RevoAI GmbH); Michael Beigl (Karlsruhe Institute of Technology (KIT)); Mehdi Tahoori (Karlsruhe Institute of Technology)

16:30

809: An O(m+n)-Space Spatiotemporal Denoising Filter with Cache-Like Memories for Dynamic Vision Sensors

Qinghang Zhao (Xidian University); Jiaqi Wang (Xidian University); Yixi Ji (Xidian University); Jinjian Wu (Xidian University); Guangming Shi (Xidian University)

16:45

739: LSMR: Synergy Randomness in Liquid State Machine and RRAM-based Analog-digital Accelerator

Ning Lin (The University of Hong Kong); Songqi Wang (The University of Hong Kong); Xinyuan Zhang (The University of Hong Kong); Shaocong Wang (the University of Hong Kong); Yangu He (The University of Hong Kong); Woyu Zhang (Institute of Microelectronics: Chinese Academy of Sciences); Bo Wang (The University of Hong Kong); Jiankun Li (The University of Hong Kong); Mingzi Li (The University of Hong Kong); Binbin Cui (The University of Hong Kong); Yi Li (The University of Hong Kong); Jia Chen (The Hong Kong University of Science and Technology); Chunwei Xia (University of Leeds); Wei Xuan (AI Chip Center for Emerging Smart Systems (ACCESS)); Xiaoming Chen (Institute of Computing Technology: Chinese Academy of Sciences); Dashan Shang (Institute of Microelectronics: Chinese Academy of Sciences); Zhongrui Wang (The University of Hong Kong)

16:00 - 17:00

Precision Matters: Improving the Robustness and Reconfigurability

Room: Skylands/Gateway

Session Chair(s): Sitao Huang

This session dives into the development of specialized architectures and computational formats to optimize neural network inference under constrained conditions. Discussions include a novel number format designed to enhance the robustness of sub-8-bit neural network operations: the design of a reconfigurable accelerator for dynamic adaptation in AI tasks: and the implementation of mixed-precision neural networks through ISA extensions for soft SIMD operations on RISC-V cores. These innovations are aimed at improving efficiency: flexibility: and performance in AI processing environments.

16:00

1248: FlexInt: A New Number Format for Robust Sub-8-Bit Neural Network Inference

Minuk Hong (Ulsan National Institute of Science and Technology (UNIST)); Hyeonuk Sim (Samsung Advanced Institute of Technology); Sugil Lee (Ulsan National Institute of Science and Technology (UNIST)); Jongeun Lee (Ulsan National Institute of Science and Technology (UNIST))

16:15

1293: MARCA: Mamba Accelerator with Reconfigurable Architecture

Jinhao Li (Shanghai Jiao Tong University); Shan Huang (Shanghai Jiao Tong University); Jiaming Xu (Shanghai Jiao Tong University); Jun Liu (Shanghai Jiao Tong University); Li Ding (Shanghai Jiao Tong University); Ningyi Xu (Shanghai Jiao Tong University); Guohao Dai (Shanghai Jiao Tong University)

16:30

1600: Mixed-precision Neural Networks on RISC-V Cores: ISA extensions for Multi-Pumped Soft SIMD Operations

Giorgos Armeniakos (National Technichal University of Athens); Alexis Maras (National Technical University of Athens); Sotirios Xydis (National Technical University of Athens); Dimitrios Soudris (NTUA)

16:00 - 17:00
Sparsity Matters: Sparse Computing Engines for Different Platforms
Room: Salons F-H
Session Chair(s): Hanrui Wang

This session explores hardware designs and algorithmic adaptations to enhance the performance and energy efficiency of AI accelerators. It includes a co-sparse photonic accelerator that integrates algorithmic and circuit design for thermal tolerance and power-efficient light distribution: an approach for exploiting sparsity in feed-forward networks and attention mechanisms in transformers on FPGA: and a dedicated point cloud inference engine optimized for RISC-V processors. These papers highlight the intersection of innovative hardware solutions and algorithmic efficiencies to push the boundaries of AI computational frameworks.

16:00
881: SCATTER: Algorithm-Circuit Co-Sparse Photonic Accelerator with Thermal-Tolerant: Power-Efficient In-situ Light Redistribution
Ziang Yin (Arizona State University); Nicholas Gangi (Rensselaer Polytechnic Institute); Meng Zhang (Rensselaer Polytechnic Institute); Jeff Zhang (Arizona State University); Rena Huang (Rensselaer Polytechnic Institute); Jiaqi Gu (Arizona State University)

16:15
981: FAS-Trans: Fully Exploiting FFN and Attention Sparsity for Transformer on FPGA
Hongji Wang (Fudan University); Yifan Zhang (Fudan University); Jun Yu (Fudan University); Kun Wang (Fudan University)

16:30
1179: RISCSparse: Point Cloud Inference Engine on RISC-V Processor
Shangran Lin (The Chinese University of Hong Kong: Shenzhen); Xinrui Zhu (Chinese University of Hong Kong: Shenzhen); baohui xie (Chinese University of Hong Kong (ShenZhen)); Tinghuan Chen (The Chinese University of Hong Kong: Shenzhen); Cheng Zhuo (Zhejiang University); QI SUN (Zhejiang University); Bei Yu (The Chinese University of Hong Kong)

17:00 - 18:00
Job Fair
Room: Salons D-E

18:00 - 19:30
ACM SIGDA Dinner
Room: Salons 4-8

Technical Program – October 31, 2024

7:30 – 8:00
Registration
Room: Grand Ballroom Foyer

8:00 – 17:00
Workshop: SUSHI (Sustainable Hardware Security -An Interactive Workshop-)
Room: Salons 1-3

8:00 – 17:00
Workshop: Quantum Computing Applications and Systems (QCAS)
Room: Lincoln/Holland/Columbia

8:00 – 12:30
Workshop: ACM/IEEE International Workshop on System-Level Interconnect Pathfinding (SLIP)
Room: Salons A-C

8:00 – 12:30
Workshop: Synergistic Innovations: Leveraging Large Models and EDA for Mutual Advancement
Room: Skylands/Gateway

8:00 – 12:30
Workshop: Workshop on Approximate Computing (AxC)
Room: Essex/Liberty

10:30 – 11:00
Coffee Break
Room: Grand Ballroom Foyer

13:30 – 14:30
Lunch
Room: Salons D-E