5th IBM IEEE CAS/EDS AI Compute Symposium (AICS’22)

The 5th IBM IEEE CAS/EDS AI Compute Symposium, known as (AICS’22), was held for two days (Oct 11-Oct 12, 2022) at the IBM T. J. Watson Research Center. The event was very well attended and received great responses from the audience all over the world. The symposium was also an initiative supported by the IBM Academy of Technology (https://www.ibm.com/blogs/academy-of-technology/). Dr. Joshi has been the interface for CAS and EDS in organizing this successful event. This year’s symposium was a hybrid event with both in-person and virtual attendees. For the third straight year, the symposium provided a virtual attendance option, allowing an increase in attendance. **Close to 1000 viewers for two days, participation from 50 countries, over 30 student posters, best poster awards, excellent panel discussions, and 11 distinguished speakers from industry and academia were the salient features of this symposium. There were more than 2600 views on the LinkedIn post about the symposium. The theme of the symposium was “Scalability to Sustainability”. In short, the symposium covered a range of topics from device technology, to circuits, architecture, algorithms, and sustainability to make innovations for the cloud with an emphasis on green AI.**

Dr. Rajiv Joshi, lead organizer, and IEEE Life Fellow, gave welcoming remarks, a short history, progress, and the impact of the symposium.

Then Robert Muchsel, Analog Devices fellow, opened the symposium with an excellent presentation related to **“Improving Privacy and Energy Usage by Pushing AI Inference to the Edge of the IoT Frontier”.** Although artificial intelligence dominates the tech news, most AI solutions are expensive, big, and energy hungry. The connected nature of these systems also leads to significant concerns relating to privacy and system autonomy. Mr. Muchsel described ADI’s true edge AI accelerators, which employ many low-power innovations to enable AI inference on a battery while improving privacy through local computing at the edge.

Prof. Aaron Thean, the Dean of the College of Design and Engineering at the National University of Singapore (NUS), followed up with an exciting talk related to “**Novel Material-System Co-Design Opportunities for Analog-Non-Volatile In-Memory Computing and Reconfigurable Edge-AI”.** Ultra-low energy and area-efficient electronic systems are required to enable untethered computing at the edge of IoT. To realize self-learning edge-AI systems, conventional solely software-driven deep-learning neural networks become a major roadblock due to the excessive energy expense of training. Hence, fundamental hardware change is likely needed. In this talk, Dr. Thean reviewed how recent material innovations (e.g. Ferroelectric oxides and 2D Material) coupled with new micro-architecture innovations (e.g. novel memory physical layout and monolithic 3D IC) may significantly accelerate in-memory computation. His talk covered wafer-level solution-processed CMOS-compatible use of 2D material (MoS2/WSe2) to enable high-endurance memristors that can have properties superior to conventional oxide RRAMs. Through material-device-aware data encoding, error correction, and novel physical memory layout (staggered + Manhattan arrays), this work aims to simplify the in-memory data process. Dr. Thean showed how one can significantly manage variabilities while accelerating convolution deep neural network operations and offering substantial low-energy opportunities for Edge-AI systems.

Subsequently, Prof. Tsu-Jae King Liu, Dean of the College of Engineering, University of California, Berkeley gave a very interesting talk about **"Technology Co-Design and Innovation for the Age of Ambient Intelligence"**. As practical limits for transistor miniaturization are reached, alternative approaches for improving integrated-circuit functionality and energy efficiency at acceptable cost will be necessary to meet the growing demand for information and communication technology. This presentation showcased how technology co-design and innovation can achieve dramatic improvements in computing performance to usher in the "Age of Ambient Intelligence."

Then Bill Luan, Senior Program Manager, Coral team at Google talked about **"Using Coral for Scalable and Sustainable AI at the Edge"**. With the advancement in AI research over the past decade, AI/ML technology has expanded from being only available on cloud-based data centers to becoming available on IoT and edge devices, opening huge opportunities for innovations. Leading this change is the Coral platform from Google, making deploying AI at the edge on a large scale not only possible but also sustainable. This presentation covered the Coral platform in detail, including product features and applications by businesses around the world that are leveraging Coral for deploying innovative edge AI solutions at scale.

Next Arun Venkatachar, Vice President AI, Cloud, and Central Engineering Synopsys Inc., gave a wonderful talk related to the "**Confluence of AI & Cloud with EDA".** Investments into AI/ML/Cloud/Big Data to solve EDA problems with ever-increasing complexities of chip design are starting to come to light. Recent product announcements like the DSO.ai from Synopsys revolutionizes chip design by massively scaling the exploration of options in design workflows. Similarly, there are many applications in production solving different challenges in areas like verification, place & route, manufacturing, etc., that have harnessed the power of AI/ML, Big-data and computing to provide a new set of tools and techniques for EDA to address both existing and new challenges. The next step is for companies to build good data strategies and compute utilization approaches to harness these benefits across their organizations. They will also need to invest in good data and AI/ML and cloud infrastructures to expedite building these solutions.

The final talk on day one is given by Prof. Jason Cong, Director of the Center for Domain-Specific Computing (CDSC) and Director of the VLSI Architecture, Synthesis, and Technology (VAST) Laboratory at UCLA. The talk focused on **"Automated Synthesis and Architecture Optimization for Deep Learning Accelerator Designs"**. AutoSA is an automated compilation framework for generating systolic arrays, and a core acceleration engine for most deep-learning applications. AutoSA is based on the polyhedral framework and incorporates a set of techniques for both computation and communication optimizations. Based on AutoSA, an automated, efficient, and comprehensive design space exploration is performed to achieve optimal systolic array designs for deep learning applications. Dr. Cong's study revealed that a number of widely used heuristics based on “common sense” often lead to sub-optimal solutions, such as limiting to loop count divisors for tiling and pruning based on off-chip data movement minimization. Finally, he showed AutoSA integrated into an end-to-end acceleration framework for deep learning using a flexible and composable architecture called FlexCNN. This approach can deliver high computation efficiency for different types of convolution layers using techniques such as dynamic tiling and data layout optimization. AutoSA and FlexCNN are both open-source projects.

Prof. Susan Trolier-McKinstry, an Evan Pugh University Professor and Steward S. Flaschen Professor of Ceramic Science and Engineering at The Pennsylvania State University, opened the second day with developments in **“New Materials for Three Dimension al Ferroelectric Microelectronics”**. In the last decade, there have been major changes in the families of ferroelectric materials available for integration with CMOS electronics. These new materials, including $Hf_{1-x}Zr_xO_2$, $Al_{1-x}Sc_xN$, $Al_{1-x}B_xN$, and $Zn_{1-x}Mg_xO$, offer the possibility of new functionalities. This talk discussed the possibility of exploiting the 3rd dimension in microelectronics for functions beyond interconnects, enabling 3D non-von Neumann computer architectures exploiting ferroelectrics for local memory, logic in memory, digital/analog computation, and neuromorphic functionality. This approach circumvents the end of Moore’s law in 2D scaling, while simultaneously overcoming the “von Neumann bottleneck” in moving instructions and data between separate logic and memory circuits. Computing accounts for 5 – 15% of worldwide energy consumption. In the U.S., data centers alone are projected to consume approximately 73 billion kWh in 2020. While recent efficiency gains in hardware have partially mitigated the rising energy consumption of computing, major gains are achievable in a paradigm shift to 3D computing systems, especially those that closely couple memory and logic. Dr. Trolier-McKinstry’s talk covered the relevant materials, their deposition conditions, and what is known about the wake-up, fatigue, and retention processes.

Next Marian Verhelst, Professor at the MICAS laboratories of KU Leuven and a research director at IMEC, gave a talk related to tinyML, **“Heterogeneous Multi-Core tinyML”**. She described approaches for powerful machine inference in resource-scarce distributed devices. Developing intelligent applications at ultra-low energy and low latency requires compact compute and memory structures that have at very high utilization. This has resulted in a wide variety of proposed state-of-the-art accelerator designs. However, it becomes increasingly clear that intelligent edge devices will need to be equipped with a diverse set of many heterogeneous co-processors, which allow running every workload on the most compatible (combination of) accelerators. Moreover, by using multiple cores in parallel and streaming data between the cores, the required amount of on-chip memory and IO bandwidth can be reduced, leading to area, energy, and latency savings. Dr. Verhelst’s talk explained the benefits and challenges of such heterogeneous ML systems, and how they allow scaling up performance at low budgets.

Dr. Steve Teig, CEO of Perceive, enlightened the audience with **“Machine Learning for Real: Thinking More Carefully About Efficiency, Loss Functions, and GANs”**.

Deep learning seems to touch every discipline these days, but behind its startling magic tricks, it is surprisingly primitive. It is concerning to note the extent to which today's deep learning relies on folklore: on recipes and anecdotes, rather than on scientific principles and explanatory mathematics. Think of how much more trustworthy, robust, compact, and power-efficient our models would be if we designed them more rigorously. Dr. Teig's assertions were accompanied by some motivating (and occasionally humorous) examples.

Then Dr. Stefanie Chiras, Senior Vice President, Partner Ecosystem Success, Red Hat, gave a very interesting talk about **"AI at the Open Edge"**. Complex use cases and game-changing potential collide when AI is delivered at the edge. This creates a perfect petri dish for innovation, not only at the technology level but in how different skills and disciplines collaborate. Building frameworks, architectures, and services can reduce the complexity and enable businesses to extract actionable insights by processing data closer to devices, sensors, and other sources. Red Hat sees this opportunity as an extension of the open hybrid cloud, bringing capability all the way out to the far edge…even as far as the International Space Station.

Finally, Dr. Tamar Eilam, IBM Fellow, gave a talk on **"The Road to Sustainable Computing"**. She presented IBM's initiative related to sustainable and responsible computing. Dr. Eilam described how to design a sustainable data center. Key important steps include finding pathways to achieve multi-DC sustainability goals incorporating exogenous factors, natural resources, cooling, water, energy for IT, and other constraints within a holistic DC Model. She also emphasized the importance of explainable AI, and contra-factual analysis with a focus on capital and operational cost with environmental impacts, such as, capitalized carbon footprint emissions, operational carbon footprint, and operational water use. Dr. Eilam pointed out the need for utilizing renewable energy, such as solar, wind, wave, etc., energy storage mediums, software platforms, infrastructure, and water usage and heat wastage. The key lies in utilizing sustainable computing by predicting the power consumption, co-optimization of software/system-based behavior, and coupling renewable energy.

The Poster Session followed by the keynote talks on day 1 of the symposium. Out of 32 posters, the best posters were awarded from each of 3 tracks. The list of winners is given on the symposium website:
https://www.zurich.ibm.com/thinklab/AIcomputesymposium.html

The symposium closed with a panel discussion on sustainability, with five distinguished panelists including Mr. Robert Muchsel (ADI), Dr. Tamar Eilam (IBM), Prof. Christopher Hill (MIT), Prof. Prashant Shenoiy (University of Massachusetts), and Prof. Aaron Thean (National University of Singapore). Panelists discussed sustainability at all levels, including algorithmic and architecture techniques, data center carbon footprint reduction, product development, and applications where AI can help.

Replays of the entire two-day symposium are available on the symposium website:
https://www.zurich.ibm.com/thinklab/AIcomputesymposium.html

Report by Dr. Rajiv Joshi, Kaoutar El Maghraoui, Arvind Kumar, Matthew Ziegler
Affiliation – T. J. Watson Research Center, Yorktown Heights, NY 10598
Executive Sponsor – Dr. Mukesh Khare
Sponsors – IEEE CAS and EDS
AoT Sponsor – Dr. John "Boz" Handy Bosma
Committee- Rajiv Joshi, Matthew Ziegler, Arvind Kumar, Xin Zhang, Krishnan Kailas, Kaoutar El Maghraoui, Jin-Ping Han, Anna Topol, John Rozen

General Chair

Raijiv Joshi

Program Committee

Matthew M. Ziegler

Arvind Kumar

Kaoutar El Maghraoui

John Rozen

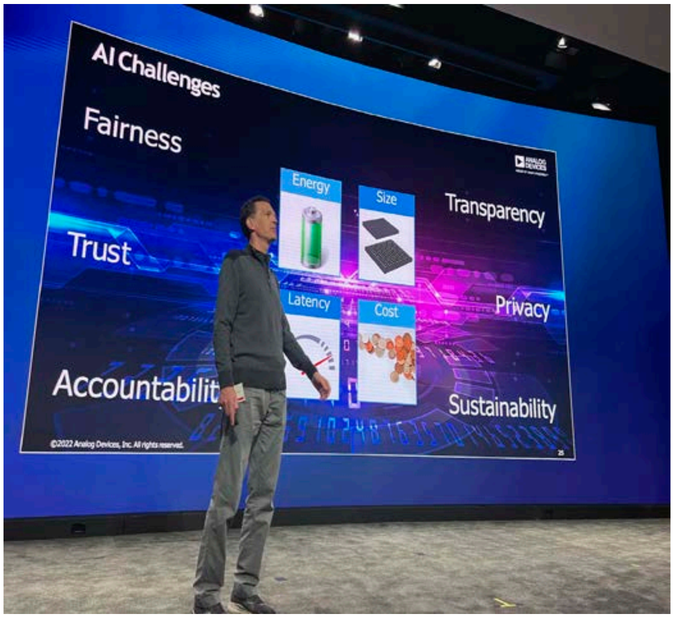
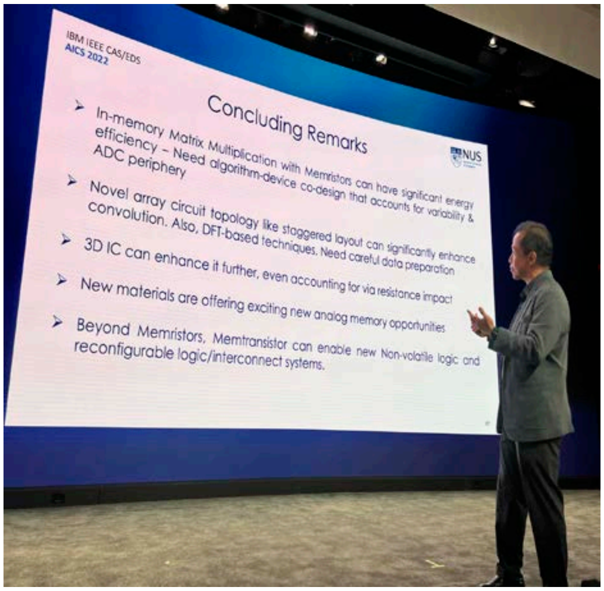
Krishnan Kailas

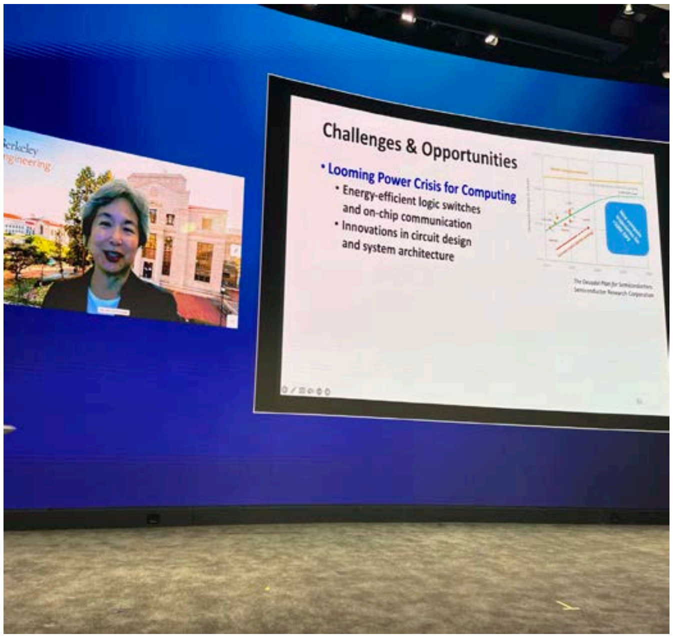
Jin-Ping Han

Xin Zhang

Linda Rudin
Web Support

Anna Topol

 AI Challenges Fairness Trust Accountability Energy Size Latency Cost Transparency Privacy Sustainability ©2022 Analog Devices, Inc. All rights reserved. **Mr. Robert Muchsel, ADI highlighting AI Challenges.**	 IBM IEEE CAS/EDS AICS 2022 Concluding Remarks ➢ In-memory Matrix Multiplication with Memristors can have significant energy efficiency – Need algorithm-device co-design that accounts for variability & ADC periphery ➢ Novel array circuit topology like staggered layout can significantly enhance convolution. Also, DFT-based techniques. Need careful data preparation ➢ 3D IC can enhance it further, even accounting for via resistance impact ➢ New materials are offering exciting new analog memory opportunities ➢ Beyond Memristors, Memtransistor can enable new Non-volatile logic and reconfigurable logic/interconnect systems. **Prof. Aaron Thean, National University of Singapore presenting concluding remarks.**

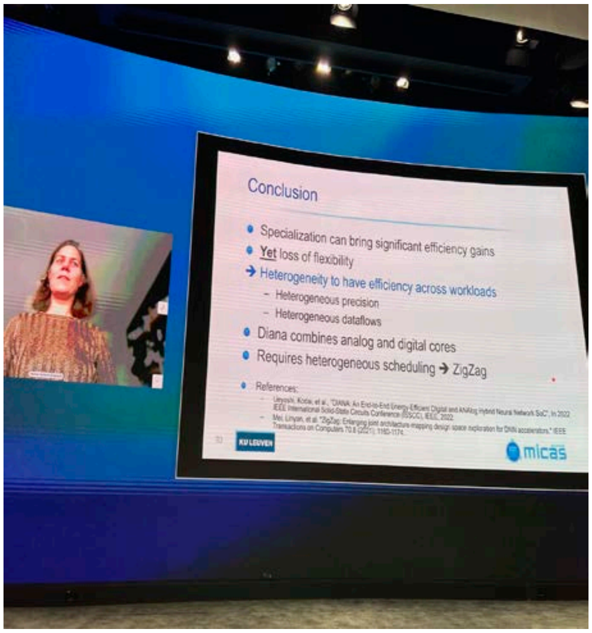


Challenges & Opportunities

- Looming Power Crisis for Computing
  - Energy-efficient logic switches and on-chip communication
  - Innovations in circuit design and system architecture

The Decadal Plan for Semiconductors
Semiconductor Research Corporation

Prof. Tsu-Jae King Liu, University of California, Berkely, showcasing challenges and opportunities in her talk Technology and Innovation.

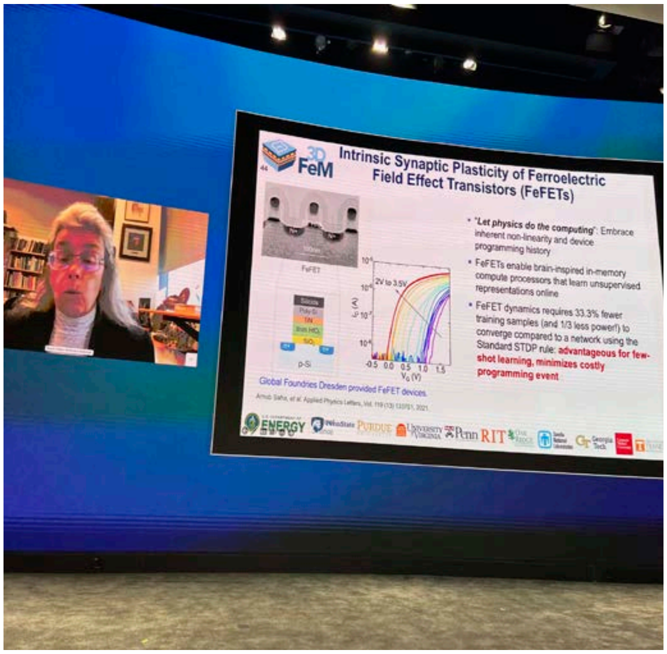


Conclusion

- Specialization can bring significant efficiency gains
- Yet loss of flexibility

→ Heterogeneity to have efficiency across workloads
  - Heterogeneous precision
  - Heterogeneous dataflows
- Diana combines analog and digital cores
- Requires heterogeneous scheduling → ZigZag
- References:
  - Ueyoshi, Kodai, et al., "DIANA: An End-to-End Energy-Efficient Digital and ANAlog Hybrid Neural Network SoC", In 2022 IEEE International Solid-State Circuits Conference (ISSCC), IEEE, 2022.
  - Mei, Linyan, et al. "ZigZag: Enlarging joint architecture-mapping design space exploration for DNN accelerators." IEEE Transactions on Computers 70.8 (2021): 1160-1174.

Prof. Marian Verhelst, KU Leuven concluding her talk.



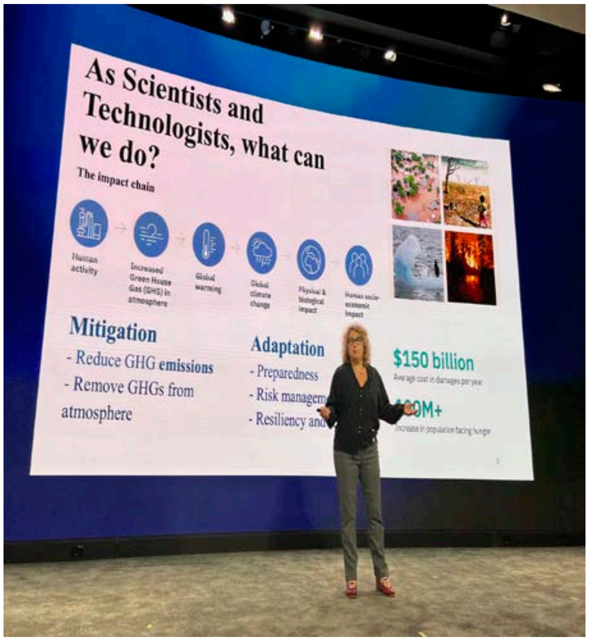
Intrinsic Synaptic Plasticity of Ferroelectric Field Effect Transistors (FeFETs)

- "Let physics do the computing": Embrace inherent non-linearity and device programming history
- FeFETs enable brain-inspired in-memory compute processors that learn unsupervised representations online
- FeFET dynamics requires 33.3% fewer training samples (and 1/3 less power!) to converge compared to a network using the Standard STDP rule: advantageous for few-shot learning, minimizes costly programming event

Global Foundries Dresden provided FeFET devices.

Arnob Saha, et al. Applied Physics Letters, Vol. 119 (13) 133701, 2021

Prof. Susan Trolier-McKinstry, Pennsylvania State University describing Ferro-electric materials and their applications.



As Scientists and Technologists, what can we do?

The impact chain

Human activity → Increased Green House Gas (GHG) in atmosphere → Global warming → Global climate change → Physical & biological impact → Human socio-economic impact

Mitigation
- Reduce GHG emissions
- Remove GHGs from atmosphere

Adaptation
- Preparedness
- Risk managem
- Resiliency and

$150 billion
Average cost in damages per year

M+
Increase in population facing hunger

Dr. Tamar Ellam, IBM, emphasizing the need for sustainability.